Exploration of rare-missense variant clustering in Mendelian disease-genes

Adam Waring

Green Templeton College

University of Oxford

Supervised by Professor Martin Farrall

Submitted in Trinity Term 2020

Thesis submission for the Degree of Doctor of Philosophy (DPhil) in Genomic Medicine and Statistics

Abstract

Decades of medical genetics research have yielded great insight into clinically relevant genetic variation. Despite this, the overwhelming complexity of the human genome means much still remains unknown. In this thesis, it is hypothesised that our understanding can be further enhanced by investigating the phenomenon of missense-variant clustering in disease-genes.

Missense-variants clustering is the key concept uniting this thesis. A clustering-detection method (*BIN-test*), a rare-variant association method (*ClusterBurden*), models of mutational hotspots (*hotspot models*), and pathogenicity prediction models (*hotspot+ models*), comprise the bulk of research reported here. Existing statistical methods are adapted to incorporate the clustering signal, solving either gene-discovery or rare-variant interpretation, and their performance is evaluated from multiple angles. The final chapters report auxiliary analyses including investigation into the prevalence of clustering and potential biases impacting rare-variant analyses.

The developed association methods had superior power to published alternatives when missense variants clustered. In a large inherited-heart-disease gene-panel dataset, *BIN-test* confirmed the existence of clustering in many definite positive-control genes and *ClusterBurden* gave theoretical power boosts in all of these. *Hotspot models* then inferred the locations of mutational hotspots in these clustered genes, enhancing variant interpretation. Extension to include pathogenicity scores in *hotspot+ models*, further improved interpretation. A large-scale scan in a database of clinical relevant variation yielded a huge number of significantly clustered genes and identified thousands of missense-variant clusters. Finally, supplementary analyses provide some insight into biases in rare-variant analyses of population controls.

Clustering observed in this thesis, and simultaneous publication of the opposite phenomenon regional constraint, indicate its widespread prevalence. This confirms the value of the statistical methods developed here and urges their continued development and application to increasingly large datasets. The deployment of methods developed here as an R package and web application, support their application by researchers of clinical genetics into the future.

Acknowledgements

I would like to give my appreciation for Professor Martin Farrall who guided the work in this thesis. Martin has a unique style of supervision which values the *philosophy* in DPhil. Martin does not give away answers easily but allows students to find their own way. This encouraged the development of scientific reasoning, permits leadership in all stages of the project cycle and provides the independence to explore new ideas and skills. If a thesis was measured in personal growth, then mine has been a resounding success, and for that I am grateful.

I would like to further acknowledge members of the cardiovascular and clinical genetics community who have supported me, especially Professor Hugh Watkins, Dr Kate Thompson and Dr Andrew Harper. Dr Andrew Harper, with support from Silvia Salatino, led the bioinformatics processing of the large inherited heart disease data analysed in this thesis.

Finally, I would like to thank my wife and three children for providing an essential balance throughout this journey and helping to sustain good mental wellbeing throughout.

Abbreviations and definitions

ACMG = American College of Medical Genetics

AD = Anderson-Darling (test)

AFR = African/African American

AMR = Admixed American

ARVC = Arrhythmogenic right-ventricular cardiomyopathy

ASJ = Ashkenazi Jewish

AUC = area under the curve

B = benign

BAM = Binary alignment map

BrS = Brugada syndrome

BWA = Burrows-Wheeler Aligner

BWS = Baumgartner-Weiss-Schindler (test)

CADD = Combined Annotation Dependent Depletion

CCR = constrained coding region

CV = cross-fold validation

dbNSFP = Database of Nonsynonymous SNVs and their Functional Predictions

DCM = Dilated cardiomyopathy

DN = Dominant-negative

EAS = East Asian

ES = Epps-Singleton (test)

ExAC = Exome aggregation consortium

FE = Fisher's-exact (test)

FIN = Finnish

FNR = false-negative rate

FPR = false-positive rate

FSC = final selection coefficient

GAM = generalized-additive model

GATK = Genome Analysis ToolKit

GLM = generalized-linear model

GnomAD = Genome aggregation database

GoF = gain-of-function

HCM = Hypertrophic cardiomyopathy

HCMR = Hypertrophic cardiomyopathy registry

HI = Haploinsufficiency

KS = Kolmorgorov-Smirnov

LB = likely benign

LoF = Loss-of-function

LP = Likely pathogenic

LQTS = Long QT syndrome

LRT = likelihood-ratio test

MAF = minor allele frequency

MSA = multiple sequence alignment

NFE = n on-Finnish European

NGS = next generation sequencing

OMGL = Oxford Medical Genetics Laboratory

OR = odds-ratio

OTH = Other - population not assigned

P = pathogenic

pCI = Poisson confidence interval

ROC = receiver operating characteristic

RVAT = rare-variant association study

SAS = South Asian

SKAT = Sequence kernel association test

TNR = true-negative rate (specificity)

TPR = true-positive rate (sensitivity)

UE-test = Unconditional-exact test

VCF = variant calling format

VEP = variant effect predictor

VUS = variant of uncertain significance

WST = Weighted sum test

Table of contents

Chapter 1

Introduction

1.1 Background

This thesis occupies the field of genomic medicine and statistics. Genomic medicine (commonly branded personalised medicine), is a branch of medicine with the key objective of using patient genomic data to inform clinical decision making. The promise of the human genome project was to characterise the relationship between genotype and phenotype, transforming diagnosis and treatment of disease (Shendure *et al.* 2019). However, the enormous complexity of this task will keep researchers employed for generations. Not only is discovering genotype-phenotype relationships difficult, but the consequent clinical implementation of new research (translatability) is also a complex task (Manolio *et al.* 2013).

Much work goes into discovering genotype-phenotype correlations. For the association of rare-variants with a specific trait, this may be tackled by; segregation analysis of an affected family, candidate gene sequencing and variant interpretation in affected individuals, exome-wide scans of

genes excessively burdened in a case group (relative to controls), or even looking beyond the exome at non-coding variation. All these approaches heavily rely on robust statistical methods to determine confidence in associations or to automate analyses over vast datasets. Much scope exists to improve these methods, such as the inclusion of new features hypothesised to be useful at separating pathogenic and benign variation.

Genetic discovery is also supported by a vast array of related projects aimed at understanding our genome. This includes huge population genetics studies characterising normal variation (Auton *et al.* 2015; Lek *et al.* 2016; Karczewski *et al.* 2020) and studies aimed at understanding the clinical genome (Caulfield *et al.* 2017; Bycroft *et al.* 2018; Landrum *et al.* 2018). Alongside a huge number of additional tools that aid in genetic analysis from processing to statistical analysis, modern-age statistical genetics involves building on these incredible resources to make improvements in current methodologies or identify novel areas to exploit as they become available with access to new data. In this thesis, variant positional information has been identified as a key signal that can now be exploited in the association of disease-genes and variants, with access to huge online databases of variation and increasingly large clinical sequencing datasets.

The following introduction covers a background into the genetics and statistics relevant to this thesis. It begins with a description of the phenomenon of rare-missense variant clustering which motivates much of the work reported here. Then, some background is given on the five inherited heart diseases analysed throughout the subsequent chapters. This is followed by some context on the clinical interpretation of genetic variants and the relevance of computational evidence in line with the current guidelines. The narrative then turns to statistical concepts related to the developed methods reported in chapters 2 and 3, including rare-variant association methods and *in-silico* pathogenicity prediction algorithms.

1.2 Genetics

1.2.1 Rare-missense variant clustering

Variant clustering is a recurring theme throughout this text and will benefit from a definition at the outset. In a general context, a cluster is loosely defined as a group of observations occurring closely together. In this thesis, where the subject at hand is predominantly missense-variants, "closely together" refers to their proximity in the linear protein sequence. For the most part, when *clusters* are discussed in this thesis, it is usually in reference to a region of the protein where pathogenic variants tend to be observed in diseased individuals, often with a corresponding depletion in control samples (i.e. population-level constraint). The term "rare-variant" is also worth defining and almost always refers to a variant present in less than 1 in 10,000 individuals in the population.

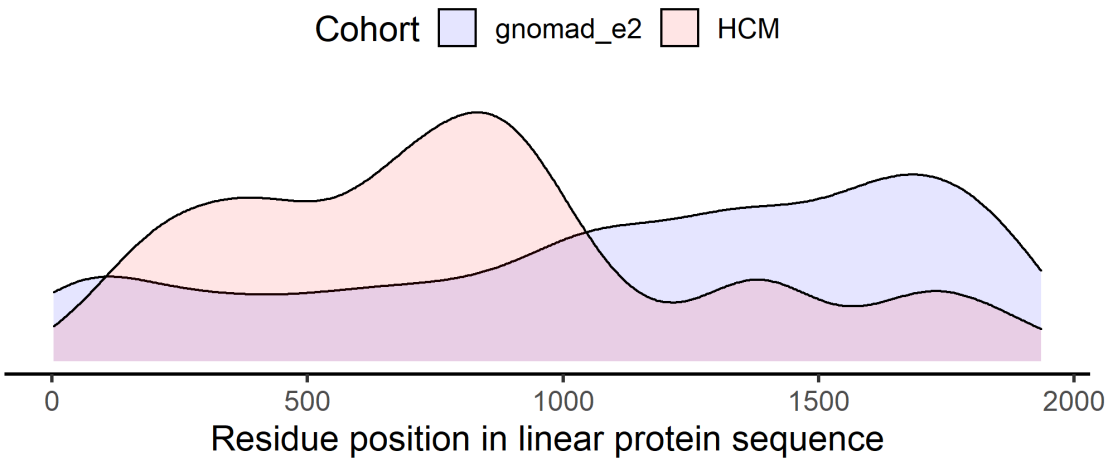

Figure 1.1: Gaussian density plot of rare-missense *MYH7* variants in HCM cases and GnomAD exomes controls (*gnomad_e2*). The missense variant positions are derived from a case-cohort of 5,338 hypertrophic cardiomyopathy cases and up to 125,748 GnomAD controls.

Inspiration for this thesis originated from the hypertrophic cardiomyopathy (HCM) disease-causing gene; myosin heavy chain 7 (*MYH7*). Clinical scientists with knowledge of this gene will be aware that patients are more likely to harbour variants in the N-terminal myosin-head motor domain. As sequencing capacity increased, large datasets, comprised of the aggregation of clinical reports, were generated for this gene. Now this effect can be clearly observed (Figure 1.1). It is noteworthy that not only are case variants enriched in the first half of the protein, but control variants are depleted. This suggests both disease-association and purifying selection.

This is not an isolated occurrence. Many examples of variant clustering in disease-genes have been observed throughout the history of clinical sequencing (Bissler *et al.* 1994; Robertson *et al.* 2003; Hunt *et al*. 2008; Henderson *et al.* 2010; Hoischen *et al.* 2010; Fine *et al.* 2012; Platzer *et al.* 2017; Rucco *et al.* 2019). An exploratory analysis of rare-missense variants available in the deafness variation database (Azaiez *et al.* 2018), revealed many genes with clear differences in the case distribution of missense variants compared to population controls (Figure 1.2). While some signals are weaker than others (e.g. *USH2A*), some demonstrate incredibly clear case clustering (e.g. *C10orf2*, *FGFR2, GATA3*), and some display control depletion in the same regions (e.g. *COL2A1*, *SLC26A4*, *FGFR1*). It seems this phenomenon is common, a topic that will be reconsidered in chapter 4.

A common mechanism mediating the phenotypic impact of a genetic variant in a coding region is haploinsufficiency (HI). This arises when a single allele of a gene does not produce enough protein to achieve the wild-type phenotype. Therefore, a variant that causes loss-of-function of one allele is sufficient to cause the disease phenotype. However, non-HI mechanisms also exist, such as gain of function (GoF) or dominant-negative (DN). GoF variants generate a new protein function, changing the behaviour of the wild-type protein. For example, in the tumour suppressor gene *TP53*, mostly somatic missense mutations can contribute to cancer progression by gain-of-function mechanisms, including resistance to anti-cancer treatments (Oren and Rotter 2010). DN variants

produce a protein that disrupts the function of the wild-type protein and are often described as *poison peptides*. For example, in Rieger Syndrome, a missense variant in the homeodomain gene *PITX2* was demonstrated to suppress the wild-type protein by preventing it from binding to another transcription factor (Saadi *et al.* 2001). It has been noted that variants which confer their effect via non-HI mechanisms tend to cluster in specific regions or domains or proteins (Lelieveld *et al.* 2017). Conversely, variants causing HI, although they may still cluster in regions more likely to cause loss-of-function, will do so less remarkably with a background of pathogenic variants throughout the linear sequence of the protein.

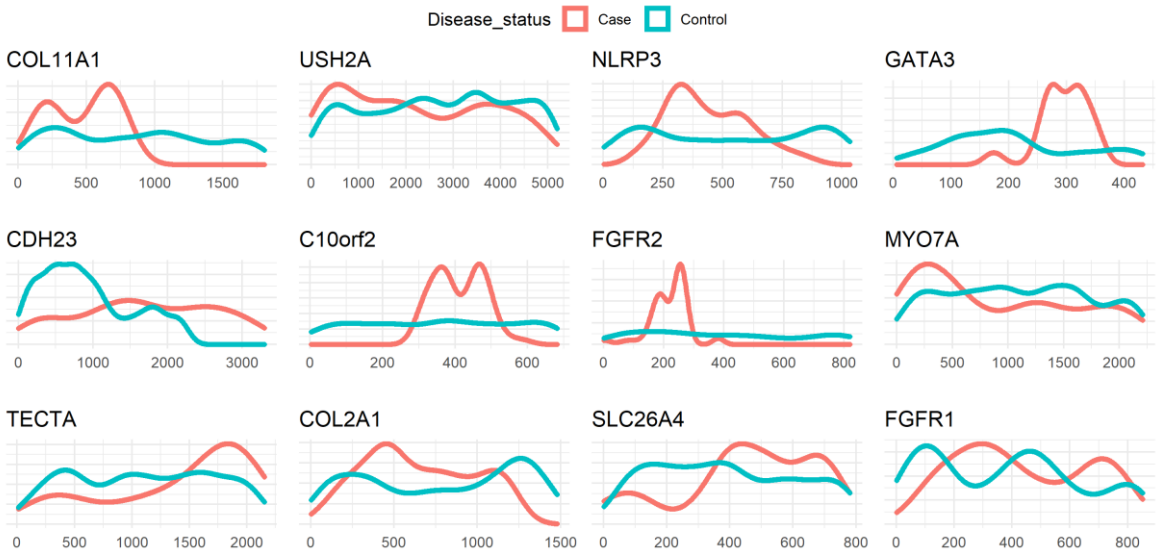

Figure 1.2: Distribution of rare-missense variants extracted from the Deafness variation database (http://deafnessvariationdatabase.org/) relative to GnomAD exomes population controls. This set of twelve genes represents a subset of deafness-associated genes that demonstrate clear variant positional differences against GnomAD controls. Variant distributions are smoothed using a Gaussian kernel.

Exome-wide analysis of this clustering signal has been recently addressed. Pérez-Palma *et al.* (2019) discovered 465 exonic regions enriched for patient-derived variants. These regions heavily overlapped with de-novo variants from patients with neurodevelopmental disorders and previously observed pathogenic variants. Iqbal *et al.* (2020) discovered wide-spread mutational

hotspots mapping to specific protein feature signatures. As previously stated, alongside disease cohort clustering, a depletion may present in controls due to purifying selection at the population-level. Traynelis *et al.* (2017) found that pathogenic variants in epilepsy patients cluster in regions with low missense tolerance, defined by observed to expected ratio of missense variants in population controls. Havrilla *et al.* (2019) developed a map of constrained coding regions (CCR) across the large GnomAD population reference cohort (Karczewski *et al.* 2020). They identified highly constrained regions that overlapped with known disease loci, suggesting novel genes with potential disease association.

1.2.2 Five inherited heart diseases

Much of this thesis focuses on the statistical analysis of inherited heart diseases. These include three structural heart diseases; hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM) and arrhythmogenic right ventricular cardiomyopathy (ARVC), and two non-structural heart diseases; Long QT-syndrome (LQTS) and Brugada syndrome (BrS). Cardiomyopathies are defined by abnormalities in heart muscle structure (Figure 1.3), in the absence of coronary artery disease (Jacoby and McKenna 2011). The group is usually broken down into three dominant forms; HCM, DCM and ARVC. Cardiomyopathy patients often become symptomatic in the fourth decade of life or beyond, but more aggressive paediatric cardiomyopathies do present in 1.5/100,000 individuals, and are associated with a larger number of adverse events and complications (Passantino *et al.* 2018).

HCM is a common cause of sudden cardiac death, and onset typically occurs later in life, at a median age of 45 (Ho *et al.* 2018). It is characterised by left ventricular hypertrophy, the thickening of the left-ventricular muscle. It is associated with a significant burden of heart failure and atrial fibrillation, usually occurring years after diagnosis. DCM is characterised by a left or biventricular dilation (enlargement) with poor contractility (ability to contract), not explained by other heart

conditions (Schultheiss *et al.* 2019). It is a major risk factor of heart failure (McNally and Mestroni 2017). ARVC causes progressive displacement of right ventricular myocardium with fibrofatty tissue. Progression of scar tissue into the endocardium results in wall thinning and is associated with inflammatory infiltrates (Corrado *et al.* 2017).

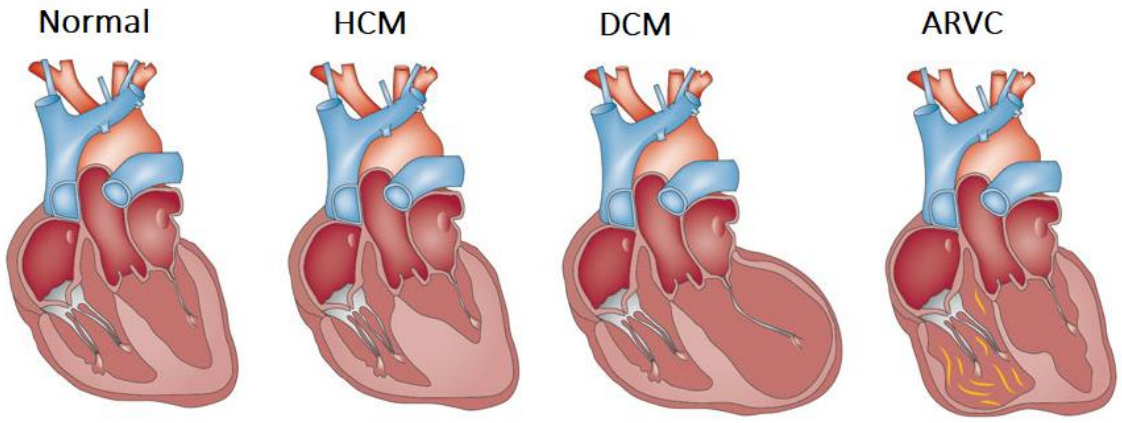

Figure 1.3: Differences in heart muscle structure in the dominant forms of cardiomyopathy. In HCM, left-ventricular hypertrophy is the main structural symptom and can be observed here as the enlargement of the light pink myocardium on the lower right-hand side of the plot. In DCM, in the same location, the ventricle is now clearly enlarged (dilated). In ARVC, the right ventricle (left-hand side of the plot), the myocardium is replaced with fibrous-fatty tissue. (Figure modified from Hershberger *et al.* 2013).

Although each cardiomyopathy subgroup is usually considered a Mendelian disease and is definitely monogenic, as a single mutation in a single gene can cause the disease, they are more prevalent than the average Mendelian disorder, which tends to be attributable to fewer genes (Amberger *et al.* 2019). The estimated prevalence of each cardiomyopathy is 1/500 for HCM (Maron *et al.* 1995), between 1/250 to 1/2700 for DCM (McNally and Mestroni 2017) and 1/1000 for ARVC (Peters *et al.* 2004). The European Union defines a rare disease as one present in less than 1 in 2000 of the general population. According to this definition, none of the three main subgroups

of cardiomyopathy qualify. This puts the cardiomyopathies in the territory between rare and common diseases.

LQTS is a heart rhythm disorder. It is characterised by prolonged ventricular repolarisation, a specific stage in the electric activity cycle of the heart (Figure 1.4; Wallace *et al.* 2019). Symptoms of LQTS include cardiac arrhythmia, loss of consciousness, seizures and possibly sudden death (Waddell-Smith *et al.* 2020). BrS is another heart rhythm disorder, characterised by ST-segment elevation, with a risk of ventricular fibrillation (irregular heartbeat) and sudden death, but with a structurally normal heart. Symptomatic patients may suffer from syncope, seizures and ventricular tachycardia (Brugada *et al.* 2018). In a study of individuals who died of sudden death, but who had a normal autopsy (i.e. no structural pathology), 42% had an inherited cardiac condition and BrS was the most prevalent diagnosis amongst these (Papadakis *et al.* 2018). The prevalence of LQTS is estimated to be 1 in 2000 (Schwartz *et al.* 2009), whereas prevalence estimates for BrS range between 1/5,000 and 1/20,000 (Brugada *et al.* 2018). Therefore, these diseases are less prevalent than the main subgroups of cardiomyopathy.

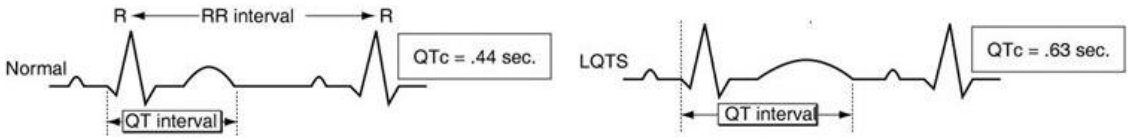

Figure 1.4: Typical electrocardiogram for a normal heart and the heart for an LQTS sufferer. Long QT-interval is associated with an extended time to repolarise after a heartbeat. It lengthens the period of time between contraction and relaxation of the heart (QT-interval). Figure modified from sads.org/Library/Long-QT-Syndrome.

1.2.3 Genetic architecture of the inherited heart diseases

The underlying causal variants for any given trait differ in their number, frequency and effect sizes (Moutsianas *et al*. 2015). Excluding more complicated genetic mechanisms, such as gene

interactions and parent-of-origin effects, these metrics define the genetic architecture of a trait. At a high level, these architectures fall on a spectrum from the infinitesimal model to the rare-allele model (Gibson 2011). The infinitesimal model seems to fit well for common diseases, as incremental increases in power for genome-wide association studies yield further signals of smaller effect. Conversely, the rare allele model considers a trait affected by a small number of low frequency variants with large effect sizes. This model seems to fit for rarer traits and is accompanied by a simpler inheritance pattern. These are the so-called "Mendelian" genes.

The fruit of decades of HCM-research has narrowed down key genes that harbour disease-causing variants. The first genetic association for any cardiomyopathy was found in 1989 (Jarcho *et al.*), by linkage analysis of a large family with familial hypertrophic cardiomyopathy. A locus on chromosome 14 was discovered to be coinherited with the disease and later identified as the gene; β-cardiac myosin heavy chain (*MYH7*; Geisterfer-Lowrance 1990). In the following years, additional genes were rapidly discovered; *TPM1* and *TNNT2* in 1994 (Thierfelder *et al.*), *MYBPC3* in 1995 (Watkins *et al.*) and *MYL2* and *MYL3* in 1996 (Poetter *et al.*). It is now firmly established that pathogenic variants in eight sarcomeric proteins; *ACTC1, MYBPC3, MYH7, MYL2, MYL3, TNNI3, TNNT2, and TPM1*, are the primary Mendelian cause of HCM. Most sarcomere-positive individuals carry a variant in the two most frequently mutated disease-causing genes; *MYH7* and *MYBPC3*. However, even for the strongly associated sarcomere genes, there is substantial incomplete penetrance with many population controls carry a sarcomere variant in the absence of a known HCM phenotype.

Despite the strong association between variants in these sarcomeric genes and HCM, approximately 50% of patients (Alfares *et al.* 2015) do not carry a disease-causing sarcomeric variant (sarcomeric-negative). This suggests that although for many HCM patients, the disease is essentially monogenic, where a single variant in one of the known genes causes the disease, for many the genetic susceptibility is found elsewhere in the genome. This has led to the search for

more causal genes and the discovery of many weakly associated HCM genes. However, Alfares *et al.* (2015) and others (Thomson *et al.* 2019) have observed, that expanding the search beyond their standard 11-gene panel resulted in a negligible increase in diagnostic yield.

As well as a complex genetic architecture, left-ventricular hypertrophy is also a complex phenotype. This is highlighted by the phenocopy genes; *GLA, LAMP2* and *PRKAG2*, which cause hypertrophy of the left ventricle, but as part of a wider metabolic syndrome (Alfares *et al.* 2015). It is crucial to identify patients with pathogenic variants in these genes as prognosis and treatment are vastly different. It has been noted that Fabry's disease, Danon's disease, and Noonan's syndrome can be difficult to distinguish from sarcomere-associated HCM on a clinical basis alone (Jacoby and McKenna 2012). Another phenotypic complexity is known as "burnt-out HCM" whereby long-term hypertrophy produces a phenotype more similar to DCM. These factors provide many barriers to the successful genetic diagnosis of HCM and outline our incomplete understanding of its genetic architecture.

Like HCM, DCM is predominantly autosomal-dominantly inherited with high variable expressivity and penetrance. The pathogenesis of DCM is complex with numerous genetic and environmental risk factors including hypertension, infections and toxins. There are a large number of genes with varying evidence of association with dilated cardiomyopathy resulting in large gene-panels for genetic testing. Genes that have been associated with DCM play a variety of cellular roles, including sarcomeric, desmosomal, cytoskeletal and membrane, indicating that the underlying aetiological pathways are diverse. *TTN* truncating variants are the most common genetic cause of DCM and account for up to 25% of cases (McNally and Mestroni 2017).

A study of families with Naxos disease, a syndromic cause of ARVC, linked the syndrome to chromosome 17 and subsequently a small deletion in *JUP*, a desmosomal protein (McKoy *et al.* 2000). Investigations in more families with Naxos disease, identified recessive LoF mutations in *DSP* (Norgett *et al.* 2000). This hinted that ARVC was a cardiac desmosomal related disease and using a

candidate gene approach *PKP2*, *DSG2*, *DSC2*, *JUP* and *DSP* were associated with an autosomal-dominant form of the disease (Jacoby and McKenna 2012). Later, non-autosomal *TMEM43* and *RYR2* were discovered in multiply affected families (Merner *et al.* 2008; Tiso *et al.* 2001).

Knowledge of the genetics underlying LQTS is strong and allows stratification of patients into different subtypes. The three most common subtypes (type 1-3), explain ~75% of cases and are caused by variants in the genes *KCNQ1, KCNH2* and *SCN5A*, respectively (Waddell-Smith *et al.* 2020). For familial LQTS, genetic testing is the most informative of all inherited cardiac conditions with a diagnostic yield of up to 80%. Gene panels for LQTS have increased in size but now include genes possibly unassociated with the disease, only causal for a small number of families or only risk modifiers (Waddell-Smith *et al.* 2020; Adler *et al.* 2020).

Like the other inherited cardiac disorders, BrS is (generally) inherited in an autosomally dominant pattern. A candidate gene study led to the first gene discovered to harbour pathogenic variants that led to ventricular fibrillation, and the first associated BrS gene; *SCN5A* (Chen *et al.* 1998). Subsequently, the majority of pathogenic variants for this disease have been identified in *SCN5A*, which accounts for approximately 30% of cases, with variants in lesser genes potentially accounting for a further 2-5% of cases. Sex and age-dependent penetrance, incomplete penetrance, variable expressivity and phenocopies all complicate the genetic analysis of BrS (Brugada *et al.* 2018). Although 42 genes have been associated with BrS, a comprehensive analysis of lesser genes associated with the disease identified definitely pathogenic variants in only four additional genes; *SLMAP*, *SEMA3A*, *SCNN1A*, and *SCN2B*, all of which encode either sodium channels or related proteins (Campuzano *et al.* 2019). In another expert reappraisal of definitely associated BrS genes, only *SCN5A* was shown to be conclusively associated (Hosseini *et al.* 2018).

1.2.4 Clinical interpretation of genetic variants

While a genetic diagnosis can lead to a clinical diagnosis for some patients with unknown disorders (Taylor *et al.* 2015), many patients who undergo clinical sequencing are already diagnosed with a specific and known disease based on their symptoms. Even in this scenario, a genetic diagnosis is valuable for a number of reasons. 1) It can help patients come to terms with why they have the disease, the risk of disease for their children, current or future, and other relatives. This is important for effective genetic counselling and cascade screening, whereby patients' relatives are genetically screened if a pathogenic variant is found. 2) In terms of disease research, a better understanding of the genetic aetiology of the disease can lead to breakthroughs in the molecular understanding and identification of novel therapeutic strategies (Chong *et al.* 2015). 3) The genotype underlying a patient's condition can also ascertain a specific diagnosis of disease subtype and aid prognosis, including the identification of cases whose symptoms are part of a syndrome (Alfares *et al.* 2015).

Genetic diagnosis plays an important part in risk stratification of inherited heart disease patients and can play a role in key clinical decisions such as the implantation of a cardioverter-defibrillator to prevent sudden death (Maron *et al.* 2019). In HCM, multiple sarcomeric mutations are associated with increased risk of adverse events (Girolami 2010), pathogenic *MYH7* variants can predict atrial fibrillation (Lee *et al.* 2018) thin filament gene variants strongly predict end-stage HCM (Marstrand *et al.* 2019). Conversely, approximately 40% of HCM patients have a less severe non-familial subtype of the disease that is associated with the reduced presence of sarcomeric variants (Ingles *et al.* 2017). The comparable pattern is observed for most other inherited heart conditions. In ARVC, there is a strong association with multiple variants across desmosomal genes and disease severity (Quarta *et al.* 2011), genetic testing in DCM can advise arrhythmia risk (McNally and Mestroni 2017) and stratifying LQTS into subtypes can inform treatment (Waddell-Smith *et al.* 2020). Furthermore, risk factors, which vary by sex and age, depend on the specific LQTS gene affected (Kutyifa *et al.* 2018).

The American College of Clinical Genetics (ACMG) produces guidelines for the clinical interpretation of genetic variants (Richards *et al.* 2015). The framework was developed to provide a consistent approach for different laboratories to classify variants. Variants are stratified into five categories; benign (B), likely benign (LB), variant of uncertain significance (VUS), likely pathogenic (LP) and pathogenic (P). The fundamental basis of these guidelines is that each classification is based on scientific evidence and weighted according to the specific evidence category (Bean and Hegde 2017). The allele frequency in affected cases and population controls are probably the most crucial pieces of evidence to interpret the pathogenicity of a variant. However, many more lines of evidence can be investigated. The ACMG define 28 evidence categories weighted as either supporting, moderate, strong or very strong, and with a direction of either benign or pathogenic. Combing these individual pieces of evidence, including population frequencies and molecular studies, delivers the final classification.

Evidence criteria, and their strength and direction, are summarised in Figure 1.5 (Strande *et al.* 2018). Odds of pathogenicity are derived from using a Bayesian framework with a prior probability of gene-association of 0.1. The justification for combining different evidence categories into one classification was statistically validated using a naïve Bayes classifier (Tavtigian *et al.* 2018). *Allelic Evidence & Cosegregation Data* refers to family studies determining whether an observed genetic variant cosegregates with disease status in an affected family. *Functional Data* relates to functional arrays assessing a deleterious effect of the variants. Neither of these evidence categories will be considered further within this thesis. *Population Data* and *Computation & Predictive Data* are very relevant to this thesis. In *Population Data*, BA, BS1 and PM2 all relate to the frequency of the variant in population controls. This is implicitly used in many analyses by filtering variants which are too common to be consistent with a rare-disease causing variant. The category *Computational & Predictive Data* assimilates evidence such as predicted loss-of-function (PM4, PVS1), co-occurrence with previously identified pathogenic variants (PM5, PS1), variant located in a mutational hotspot (PM1) and *in-silico* variant prediction algorithms (PP3).

The ACMG evidence category with the most relevance to this thesis is criterion PM1: Variant located in a mutational hot spot and/or critical and well-established functional domain (in the absence of benign variation). Moderate evidence of pathogenicity is attributed to any variant meeting this criterion. This is a vague criterion and is applied as such. In *MYH7,* an expert panel recommends applying it for any variant located between amino acids residues 181–937 (Kelly *et al.* 2018). For other genes, this criterion is applied inconsistently with suggestions of its over-application. Often mutational hotspots are defined qualitatively based on the judgment and experience of clinical scientists. With data to empirically estimate them, a statistical modelling framework should supercede this approach.

The intention of these guidelines was to create a consistent framework to classify variants across multiple Mendelian disorders. However, by attempting to fit a framework for all diseases and genes, it may be suboptimal for some disease-gene areas, especially if they are less well-researched. This has resulted in multiple guidelines for specific genes and/or disease areas including an expert framework for the specific genes *MYH7* (Kelly *et al.* 2018) and *PTEN* (Mester *et al.* 2018), and the diseases rasopathies (Gelb *et al.* 2018) and hearing loss (Oza *et al.* 2018), mostly working under the umbrella of ClinGEN (Rehm *et al.* 2015), a collaborative effort to create public resources to improve understanding of clinically-relevant variation.

		BENIGN CRITERIA		PATHOGENIC CRITERIA			
Strength of evidence		Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Odds of Pathogenicity*		–18.7	–2.08	2.08	4.33	18.7	350.0
Evidence Category and Corresponding ACMG/AMP Codes	Population Data	*BA1*+ BS1 BS2			PM2	PS4	
	Allelic Evidence & Cosegregation Data	BS4	BP2 BP5	PP1 →	PM3 PM6	PS2	
	Computation & Predictive Data		BP1 BP3 BP4 BP7	PP2 PP3	PM1 PM4 PM5	PS1	PVS1
	Functional Data	BS3				PS3	
	Other		BP6	PP4 PP5			

Figure 1.5: ACMG criteria summarised by the strength and direction of evidence. The odds of pathogenicity are calculated based on a prior probability of 0.1 for the association of the gene. Figure duplicated from Strande *et al*. (2018).

Despite these standardised guidelines, there is still substantial discordance in classifications between testing laboratories. Seventeen percent of variants in ClinVAR (*clinvar*; Landrum *et al.* 2018) with multiple submitters had conflicting interpretations (Amendola *et al.* 2016). This is partly due to the complexity of variant interpretation, including difficulties in the functional evaluation of molecular consequences of each variant, and a general dearth of available evidence (Evans *et al.* 2019). This leads to clinically important differences in variant interpretation between different clinical sequencing laboratories. Many differences are reported to be due to inconsistent application of the guidelines and privately-held data (Furqan *et al.* 2017).

1.2.5 Clinical application of in-silico evidence

In-silico prediction algorithms usually refer to any class of algorithms designed to predict pathogenic or functional variants. Generally, these terms are used interchangeably as a functional

variant is assumed to be potentially pathogenic. In the ACMG guidelines, these algorithms are permitted as evidence of variant pathogenicity; criteria PP3: “Multiple lines of computational evidence support a deleterious effect on the gene or gene product”. However, at best, they only provide supporting evidence, the weakest evidence category.

One of the main reasons these methods were developed was to aid in the discovery of pathogenic variants. As the number of candidate variants is usually extremely high, methods to reduce this to a credible set are crucial in this process. However, despite much progress, the methods still exhibit substantial uncertainty. For the clinical laboratories, this means these methods are given low importance when assessing a candidate variant. Furthermore, as the ACMG guidelines do not specify which algorithms to use, they are applied inconsistently across clinical testing laboratories (Bean and Hegde 2017).

Ghosh *et al.* (2017) reported that not specifying which or how many algorithms to use, and the requirement to observe concordance between all algorithms, produces discordance and more VUSs. They made a number of further observations. Firstly, a non-trivial number of “false concordances”, whereby all algorithms supported an incorrect classification. This may be caused by high levels of correlation between predictors whereby a single shared latent feature is driving the classification for each score. Additionally, it was noted that pre-defined thresholds associated with algorithms for classifying variants into benign or pathogenic were often arbitrary and not well-calibrated.

Previously, two types of circularity have been reported as potential confounders in the use of variant prediction algorithms (Grimm *et al.* 2015). The first is type-1 circularity, whereby variants in the validation set are also used to train the algorithms. Type-2 circularity is slightly more complex and results from the joint labelling of multiple variants in a given gene as either benign or pathogenic. This may result in poor generalisation to new genes or poor performance for predicting different variants within the same gene. Ghosh *et al.* (2017) noted that type-2 circularity was a

particular problem for the sequence-based algorithm FATHMM where performance dropped when benign and pathogenic variants were more balanced in a given gene.

The early approaches to pathogenicity prediction used heterogeneous variant data derived from different genes and associated with different diseases. Recently, this approach has been improved using gene and disease-specific approaches. This was not possible until recently as sufficiently large training datasets have become available. With the advent of large population reference databases such as ExAC (Lek *et al.* 2016), multiple frameworks were developed to prioritise putative pathogenic genes based on background rates of variation indicative of constraint (Samocha *et al.* 2014). This gene-level constraint approach was then extended to identify regional constraint within each gene (Havrilla *et al.* 2019). Using this regional constraint metric alongside variant-level features and *disease-specific* training sets; the algorithm PathoPredictor was demonstrated to outperform other variant prediction tools with a modest but significant increase in precision (Evans *et al.* 2019).

1.3 Statistics

1.3.1 Rare variant association

For the vast majority of Mendelian disorders, the associated loci that have been discovered fall in protein-coding regions and affect protein function (Chong *et al.* 2015). In the flagship analysis of variation in the 60,706 sample ExAC dataset (Lek *et al.* 2016), it was found that the majority of variation was rare with 99% of variants having a frequency below 1% and 54% observed only once (singletons). It has been estimated that the majority of these rare-missense variants are at least mildly deleterious in humans (Kryukov *et al*. 2007). An analysis of rates of purifying selection acting on different mutation types, adjusted for mutability, indicated that nonsense and essential splice site variants were the most selected against but stop lost, start lost and missense variants were

also subject to substantial purifying selection. Conversely, downstream, intergenic, upstream and intronic variants had comparatively low constraint.

Rare-variant association tests are a specific class of statistical methods aimed at detecting rare-variants associated with a specific trait. For Mendelian diseases, it is assumed that penetrant causal variants are rare; otherwise, the disease would not be rare. For dominant diseases, the principal barrier to finding an association of rare-variants is the background of thousands of presumably neutral or only mildly deleterious variants. Due to the enormous catalogue of variation present in every human genome, 3.53 million SNPs per genome (Europeans; Auton *et al.* 2015), it is clear that most variants do not cause rare-diseases. Therefore, signals need to be isolated amongst an immense background of noise.

The causative loci underlying many Mendelian disorders have been discovered using linkage mapping and candidate gene sequencing. Subsequently, studies focused on candidate genes similar to previously associated loci in the same or similar diseases. Later, exome-wide association studies were proposed as an alternative approach (Bamshad *et al.* 2011). Exome sequencing comprises of the targeted capture of exonic regions in the human genome followed by massively parallel sequencing to characterise all coding variation. The first example of exome sequencing finding the cause of rare Mendelian disease, was Miller Syndrome, with only four affected individuals (Ng 2009).

A simple approach to rare-variant association is a single-variant analysis, otherwise known as a single-marker test. Here, the frequency of a rare-variants are compared between cases and controls. However, if causal variants are rare, then power to detect using single-marker is limited (Moutsianas *et al.* 2015) and extremely large sample sizes and effect sizes are required to have sufficient power, especially after multiple testing correction. Instead, signals are aggregated over a region of interest. Due to evidence that genes harbour many pathogenic variants, this unit of aggregation is usually a gene.

The most simple approach to gene-based rare-variant association is burden testing. Here, variants in a region are collapsed into one score. For an additive model, minor alleles are counted across the region. For a dominant model, the number of samples with at least one minor allele are counted, such as with the cohort allelic sums test (Morgenthaler and Thilly 2007). This simple approach can be extended to include weights for each variant, either by their allele frequency (Madsen and Browning 2009) or their functional prediction scores (Curtis *et al.* 2018). Limitations of the burden testing approach include the assumption that all variants in a region are causal, have the same effect size and same effect direction (Lee *et al.* 2012).

To overcome the specification of an unknown genetic model, adaptive tests were developed. Most adaptive tests permute over hyperparameters of the genetic model. For example, the variable-threshold test (Price *et al.* 2010) permutes over frequency thresholds and calculates p-values for the region based on the threshold with the highest p-value. The kernel-based adaptive clustering test (Liu *et al.* 2010), weights by multi-allelic genotypes to capture information on gene-interaction and modifiers, alongside adaptive down-weighting of putatively non-causal variants in a region. However, potential problems of these approaches include unstable estimates of effects sizes and directions with sparse data (Lee *et al.* 2014) and computational intensive permutation which may actually have a reduction in power where decent prior information is available on the true genetic model.

An alternative approach, which was the first major deviation from the classical burden framework, are variance-component tests such as C-alpha (Neale *et al.* 2011) and SKAT (Lee *et al.* 2012). Here, using a random-effects model, the variance in frequencies at each segregating site in the region are compared to a null model. These methods are inherently two-sided and test equally for protective and deleterious variants. One problem of these approaches is that the estimation of the null model for variance is unstable when counts are very low. Additionally, for scenarios where most variation

in a region is deleterious, the two-sided nature of these approaches causes power loss compared to a one-sided burden test where protective variants are not considered.

1.3.2 Methods that incorporate clustering

Comparing two empirical distributions was first tackled by two-sample distribution tests. The Kolmogorov-Smirnov test (KS-test) is one of the oldest and most commonly used goodness-of-fit tests. The test assesses the null hypothesis that two samples are drawn from the same underlying distribution. The KS-test does not compare the complete empirical distribution function but instead compares summary statistics, specifically the supremum. This metric is based on the largest difference between the cumulative distribution functions of the samples to be compared. Previous power comparisons of the KS-test revealed it has weak power to detect differences between two distributions compared to other two-sample tests (Razali and Wah 2011).

The two-sample Anderson-Darling test (AD-test) is a non-parametric method that evaluates the hypothesis of k independent samples sharing a common distribution. The test statistic is derived from integrated differences between the empirical distribution functions of the samples under investigation (Scholz and Stephens 1987). The test purposely gives extra weight to the tails of the distributions as opposed to the KS-test which incidentally gives more weight to deviances in the middle of the distributions. Also considered were the Epps-Singleton test (ES-test) and the modified Baumgartner test (BWS-test; Baumgartner *et al.* 1998). ES-test is another two-sided non-parametric test to detect differences in distributions between two samples. BWS-test is another two-sample test, applicable to discrete data and valid for unequal sample sizes. Both tests have been demonstrated to be more powerful than the KS-test (Goerg and Kaiser 2009). These statistical methods, although not specifically designed for rare-variant analyses, may be useful to determine whether positional distributions of variants differ between two cohorts.

For RVAS, multiple approaches have been proposed to incorporate clustering or positional information into the genetic model. These include sliding windows approach where test statistics are calculated for each window across the gene (Ionita-Laza *et al.* 2012; McCallum and Ionita-Laza 2015) or kernel methods where distance matrices between variants are compared for each cohort (Fier *et al.* 2012; Schaid *et al.* 2013; Lin 2014; Bodenhoefer 2015) and combined methods where burden and position statistics are combined together (Chen *et al.* 2013; Persyn *et al.* 2017).

In sliding window approaches, the overall goal is to compute differences in case-control frequency differences in each window, relative to case-control frequencies outside the window. This is done on multiple windows across the region of interest. This solution has been implemented by Ionita Laza *et al.* (2012) using an extension of the Kulldorff statistic which detects circular clusters in geographic epidemiological data using likelihood-ratio based statistics (Kulldorff and Nagarwalla 1995). This approach was extended using a Bayesian framework that permitted inclusion of covariates such as functional prediction scores (McCallum and Ionita-Laza 2015). Potential issues with this strategy are the selection of the hyperparameter window-size and multiple testing burden. The consequence is that the test statistic is calculated over many different window sizes, and p-values are calculated by permutation to limit false positives. Furthermore, where multiple clusters exist and the clustering pattern is more complex, focusing on one window at a time may result in loss of information.

Kernel-based approaches use a combination of variant frequency data in cases and controls and a kernel matrix of pairwise variant distances in each cohort. Here, the kernel is essentially a density smoothing function that determines the weight for different distances between variants. For example, are only variants in very close proximity given weight or do we extend the weighting to variants observed further away? This is equivalent to choosing cluster size. The kernel chosen was based on Tango's statistic (Tango 1984). Like Kulldorf's statistic, it was originally developed for detecting clusters of disease occurrence in epidemiological data. Examining multiple potential

cluster sizes can be achieved by varying the scaling factor of this kernel. Another kernel-based positional weighting approach is CLUSTER (Lin 2014) which extends from the adaptive combination of p-values method (Lin *et al.* 2014) by combining variants that are more likely to be causal. However, like the sliding windows approaches, this results in a multiple testing burden, which alongside asymptotic properties not being met with sparse rare-variant data, means permutation must be used for p-value calculation.

An alternative approach is using the direct combination of burden and positional signals. Persyn *et al.* (2017) developed a framework to combined global average differences in allele frequency (i.e. burden testing) and the difference in area between Gaussian density curves for positional distributions in cases and controls. These two signals were combined into one test statistic and p-values were calculated by permutation. Chen *et al.* (2013) created two likelihood ratio tests to capture both burden and positional information. For their positional component, counts in different windows across the gene were compared against a multinomial distribution. Different window sizes and offsets were used and p-values were calculated by permutation. The resulting statistics were combined into a single log-likelihood ratio to generate a p-value.

1.3.3 In-silico prediction algorithms

Due to the sheer size and complexity of the human genome, all analysis of genetic variants relies on the support of computational algorithms. This starts with overlapping sequence reads, mapping contiguous sequences to a reference genome and finding genomic positions where variation occurs. Temporarily ignoring this non-trivial multifarious bioinformatics processing step, once a list of variants is obtained, computational tools are still necessary before analysis. Without them, you would simply have a list of chromosome positions, reference and alternate bases and a variety of quality scores. This is not enough information for much statistical analysis. Therefore, tools such as Annovar (Wang *et al.* 2010) or VEP (McLaren *et al.* 2016), annotate sequence variants with genetic

regions they overlap such as protein-coding regions. If the variant does overlap with a protein-coding region, then the tools will predict its impact, such as missense, synonymous or loss-of-function.

A massive amount of additional information is available from decades of genetics research. This includes conservation or constraint estimates, protein regions of known function, amino acid substitution probabilities, protein structural information and specific data on previously observed variants such as molecular experiments and related phenotypes. Many algorithms have now been developed to search for patterns within these features that are predictive of pathogenic variants. This has been done in both a hypothesis-driven approach, conserved regions are hypothesised to be functionally important, and a data-driven approach, training algorithms supervised by data on putatively pathogenic and benign variants. Henceforth, when *in-silico* algorithms are discussed, this refers to variant pathogenicity predictors. These methods lead to insights such as disease-causing variants are enriched for features such as hydrogen bond network and salt bridge disruption and are likely to cause destabilisation of proteins and their interactions (Petukh *et al.* 2015).

Evolutionary constraint is a common measure used across these algorithms. The foundation of this approach is to compute multiple sequence alignments (MSAs) across different species and estimating the constraint-based on the observed number of substitutions versus the expected number. The aim is to identify conserved regions where variants are likely to be deleterious based on the hypothesis that these regions have been subject to purifying selection (Ritchie and Flicek 2014). Methods that use this approach include GERP (Cooper *et al.* 2005), PhastCons and phyloP (Siepel *et al.* 2005). The method SIFT (Kumar *et al.* 2005) also uses MSAs, but of protein sequences and further incorporates the empirical distribution of amino acid residues in its estimates. FATHMM uses a similar approach except it uses a hidden-Markov model and can incorporate weights based on putative disease-causing variants.

The main drawback of methods reliant on conservation is that they are agnostic to adaptations that have evolved since the last common ancestor and therefore a lack of conservation does not necessarily mean that the region is not functional (Ritchie and Flicek 2014). Another strategy is to use a supervised modelling approach using features of putative functional and non-functional variants as training data. The widely used method PolyPhen (Adzhubei *et al.* 2010) is one of the first examples of this approach. It uses features such as protein-level constraint, variant position, three-dimensional structure and overlap with Pfam domains. Many more methods exist using this strategy including MutationAssessor (Reva *et al.* 2011) and MutationTaster (Schwarz *et al.* 2010). The approach CADD also using a supervised approach but instead uses fixed variation and simulated unobserved variation as its training data. This avoids bias in annotations of pathogenic variants. As more methods became available, meta-predictor scores were developed using machine learning algorithms to combine predictions of published scores with training datasets usually derived from large public repositories of variants.

While many of these algorithms generate features individually for each individual base substitution in the coding region, some use regional features to support or drive predictions. For example, the variant prediction algorithms FATHMM (Shihab *et al.* 2013) is a sequence-based model where the predictions of each position are somewhat dependent on the properties of the previous positions in the sequence. The VEST4 (Carter *et al.* 2013) algorithm is another example of an algorithm that uses regional features. More recently algorithms that determine intra-species regional constraint estimated from population databases are also capable of making position-dependent predictions; such as CCR (Havrilla *et al.* 2019), MPC (Samocha *et al.* 2017) and MTR (Traynelis *et al.* 2017).

1.4 Hypothesis and objectives

This thesis is founded on a number of hypotheses. 1) Rare-missense variant clustering is a ubiquitous phenomenon in Mendelian disease-genes that is generally the rule rather than the

exception. While it may seem obvious that the position of a variant in the protein somewhat determines its potential pathogenicity, as different protein regions contribute to structure and function in different ways, here it is hypothesised that in many proteins, contiguous residues in the linear protein sequence share this increase in pathogenicity. 2) Missense variant clustering will enhance separation of benign and pathogenic variants. This will improve both the identification of novel associated genes and improve interpretation of missense variants within established genes. 3) Data-driven approaches to variant interpretation will enhance performance relative to qualitative approaches. The qualitative outcome from the current interpretation framework (Richards *et al.* 2015) does not necessarily equate to a specific probability of variant pathogenicity. Here, quantitative approaches are explored to derive greater precision in estimates of pathogenicity. To test these hypotheses, this thesis focuses on the development of statistical methods that incorporate variant position as a feature as well as explore evidence of variant clustering in private and public datasets.

1.5 Outline

In chapter 2, I present the development of statistical methods to enhance rare-variant association. The goal of the first method (*BIN-test*) is to create an efficient and well-calibrated method to test the hypothesis of clustered pathogenic variants in a gene. The principal goal of any hypothesis test is to maximise power and keep false-positives at a nominal threshold. This was assessed in simulated data. This approach was then extended to test the joint hypothesis of variant clustering and an overburdened disease-gene (*ClusterBurden*). Again the properties of this method are assessed in simulated data and its performance is compared with conceptually similar alternatives. A final test (LRT-dbNSFP), is then explored to test the hypothesis that pathogenicity prediction scores differ between case and control variants. These methods are then applied to a large private gene-panel sequence dataset for five inherited heart diseases.

In chapter 3, models are explored to infer mutational hotspots and pathogenicity prediction in well-established inherited heart disease genes. Using generalised additive models, the regional burden of rare-missense variants is estimated across the linear sequence of the proteins. Then, to incorporate further information such as constraint, conservation, structure or predicted functional consequences, *in-silico* algorithms are also included in the model. Mutational hotspots identified by the models of regional burden are then assessed for the clinical utility and overlap with protein features. Pathogenicity prediction models are also assessed against alternatives designed for disease-agnostic pathogenicity prediction.

In chapter 4, the prevalence of missense-variant clustering in Mendelian disease-genes is explored. Using the *clinvar* database of clinical relevant variants as cases and GnomAD population reference cohort as controls. The *BIN-test*, developed in chapter 2, is applied across thousands of genes to detect signals of variant clustering. Properties of clustered genes and clusters themselves are then explored to determine features that may be useful as priors in a *ClusterBurden* exome-scan.

In chapter 5, allele frequency filtering approaches and the implications for ancestry in burden testing are considered. Potential biases incurred from frequency filtering strategy are highlighted and optimal approaches to counter these are investigated. The ancestry-dependent burden is calculated for different subpopulations of gnomad_e2, indicating some populations with a relative reduction in rare-variation.

In the final chapter, conclusion, limitations and future directions are discussed.

Chapter 2

ClusterBurden

2.1 Inherited heart disease datasets

A total of 5,338 cases of hypertrophic cardiomyopathy (HCM) were available by combining two next-generation sequencing (NGS) datasets; the Oxford Medical Genetics Laboratory (OMGL) and the Hypertrophic Cardiomyopathy Registry (HCMR). Up to 2,757 probands sequenced at OMGL, all referred by cardiac specialists between 2013 and 2018 and 2,636 probands with a clinical diagnosis of HCM sequenced as part of the HCMR project, also sequenced between 2013 and 2018. NGS data from OMGL was also available for up to 910 dilated cardiomyopathy (DCM) cases, 266 arrhythmic right-ventricular cardiomyopathy (ARVC) cases, 593 cases with Long QT syndrome (LQTS) and 303 cases with Brugada syndrome (BrS). The publicly accessible GnomAD exomes version 2 (*gnomad_e2)*, consisting of data for up to 125,748 samples, was used as control data (Table 2.1). This dataset is an aggregation of multiple exome sequencing projects, with severe paediatric diseases filtered out and all samples processed on the same pipeline.

Table 2.1: Counts of rare-missense variants in five inherited heart diseases and GnomAD exomes. Genes are present on either the cardiomyopathy or arrhythmia gene-panels with some overlap between the two. The denominator beside each variant count is the number of samples with sequencing data available. Blank cells indicate genes not sequenced for that disease.

Gene	ENST	HCM	DCM	ARVC	LQTS	BrS	*gnomad_e2*
ACTC1	ENST00000290378.4	24/5259	4/910	0/266			63/125443
ACTN2	ENST00000366578.4	36/5279	5/909	2/266			706/125008
AKAP9	ENST00000356239.3				10/591	6/300	2689/12391
ANK2	ENST00000357077.4				15/573	10/29	2744/12515
ANKRD1	ENST00000371697.3	19/5233	0/903	0/266			238/125201
BAG3	ENST00000369085.3	27/4266	4/631	1/146			545/124345
CACNA1C	ENST00000347598.4				8/592	6/301	1160/12001
CACNA2D	ENST00000356860.3				2/591	4/301	445/123773
CACNB2	ENST00000324631.7				4/590	2/297	660/123877
CALM1	ENST00000356978.4				0/592	0/301	7/125605
CALM2	ENST00000272298.7				1/584		8/122893
CALM3	ENST00000291295.9				0/591	0/301	12/123276
CASQ2	ENST00000261448.5				5/591	0/300	318/125343
CAV3	ENST00000343849.2				0/590	0/300	79/125500
CRYAB	ENST00000533475.1	9/5282	1/910	0/266			124/122249
CSRP3	ENST00000533783.1	30/5279	0/910	0/266			199/125647
DES	ENST00000373960.3	19/5245	6/902	2/264	1/586	2/302	293/113297
DLG1	ENST00000346964.2				2/591	1/301	585/122761
DMD	ENST00000357033.4	91/4254	11/62	2/146			1818/90804
DSC2	ENST00000280904.6	32/5225	5/894	7/259	5/565	1/291	743/124507
DSG2	ENST00000261590.8	42/5229	6/874	7/255	4/575	5/286	783/123827
DSP	ENST00000379802.3	103/526	32/90	11/26	9/592	8/302	2503/12468
FHL1	ENST00000394155.2	26/5265	5/910	0/266			126/91311
FHL2	ENST00000358129.4	17/5269	1/905	0/261			245/125156
FLNC	ENST00000325888.8	139/426	17/63	5/146			2324/12219
GLA	ENST00000218516.3	23/5278	1/909	0/266			96/91591
GPD1L	ENST00000282541.5				2/592	0/301	224/124881
HCN4	ENST00000261917.3				1/576	2/295	710/105680
JUP	ENST00000393931.3	45/5229	3/895	1/253	5/580	5/303	625/123602
KCND3	ENST00000315987.2				0/592	0/301	327/124119
KCNE1	ENST00000337385.3				2/593	0/302	135/125648
KCNE1L	ENST00000372101.2				0/592	0/301	67/76977
KCNE2	ENST00000290310.3				0/591	0/299	150/125703
KCNE3	ENST00000310128.4				2/592	0/301	120/125611
KCNH2	ENST00000262186.5				47/574	3/294	815/108958
KCNJ2	ENST00000243457.3				5/593	1/301	204/125651
KCNJ5	ENST00000529694.1				2/592	0/301	319/124649
KCNJ8	ENST00000240662.2				1/509	3/241	142/125462
KCNQ1	ENST00000155840.5				77/585	3/300	479/115319
LAMP2	ENST00000434600.2	14/5239	0/903	0/264			134/89868
LMNA	ENST00000368300.4	22/5207	14/89	2/227	4/592	1/302	474/119600
MYBPC3	ENST00000545968.1	461/525	17/90	5/261			1180/11577
MYH7	ENST00000355349.3	517/524	40/90	2/266			1454/12514

MYL2	ENST00000228841.8	30/5283	1/910	0/266			172/125387
MYL3	ENST00000395869.1	21/5284	1/910	1/266			145/125573
PKP2	ENST00000070846.6	41/5203	12/90	18/25	5/552	2/301	810/122946
PLN	ENST00000357525.5	2/5283	1/910	0/266	0/570		20/125572
PRKAG2	ENST00000287878.4	23/5230	3/899	0/251			382/122945
RBM20	ENST00000369519.3	40/4225	14/61	2/136			550/76104
RYR2	ENST00000366574.2				27/584	7/301	2930/12092
SCN10A	ENST00000449082.2				7/564	3/289	1817/12457
SCN1B	ENST00000415950.3				2/588	0/301	165/110609
SCN2B	ENST00000278947.5				1/592	1/301	226/125572
SCN3B	ENST00000392770.2				1/592	0/301	253/125626
SCN4B	ENST00000324727.4				0/589	0/301	205/125466
SCN5A	ENST00000413689.1	81/5244	13/90	3/262	21/558	33/30	1698/12244
SLMAP	ENST00000295951.3				2/582	1/301	550/123238
SNTA1	ENST00000217381.2				0/584	1/293	327/113045
TAZ	ENST00000299328.5	8/4251	0/631	0/146			69/89583
TMEM43	ENST00000306077.4	17/5282	4/910	2/266	1/593	2/303	395/125467
TNNC1	ENST00000232975.3	7/4270	0/631				56/125128
TNNI3	ENST00000344887.5	77/5271	7/890	0/266			135/119490
TNNT2	ENST00000509001.1	47/5240	14/89	0/266			130/124577
TPM1	ENST00000358278.3	33/5122	5/856	1/216			91/124717
TRDN	ENST00000398178.3				3/576	3/293	391/103787
TRPM4	ENST00000252826.5				6/560	1/283	1010/11955
TTR	ENST00000237014.3	5/4269	1/629	0/146			72/125689
VCL	ENST00000211998.4	29/4262	5/631	1/146			683/124733

Although the gene-panel for cardiomyopathy comprises a set of genes sequenced for clinical referrals with HCM, DCM or ARVC, not all genes are definitely associated with each condition. The same is true for the arrhythmias; LQTS and BrS. Some genes are present in both gene-panels; *DES*, *DSC2*, *DSG2*, *DSP, LMNA*, *PKP2*, *PLN*, *SCN5A* and *TMEM43*. Due to the sparseness of rare-variation, for some gene-disease pairs, zero rare-missense variants were observed. For some highly penetrant or large genes, sequenced in thousands of patients, many rare-missense variants were observed; *MYBPC3* (461/5258) and *MYH7* (517/5245).

Most samples in the OMGL datasets were referred by inherited cardiac condition centres within the UK. No ancestry information (self-reported or genetically determined) was available for these data. For the HCMR cohort, self-reported ancestry and additional genome-wide SNP-array data

permitted the exclusion of individual's 3rd-degree relatives or closer. This was completed using the KING relationship inference software (Manichaikul *et al.* 2010). As comparable data was unavailable for OMGL samples, closely related samples could not be identified for exclusion.

For HCM, DCM and ARVC, most samples were sequenced at the same generic 35-gene cardiomyopathy panel. The genes *BAG3, DMD*, *FLNC*, *RBM20*, *TAZ*, *TNNC1*, *TTR* and *VCL*; were included in the OMGL cardiomyopathy panel later than 2013. This resulted in a reduced sample size for these genes in HCM (5,338 to 4,379), DCM (910 to 614) and ARVC (266 to 149). The arrhythmia gene-panel consisted of 43 genes, all of which had full sample sizes with the exception of *PLN,* which had a reduced sample size in LQTS (593 to 571) and BrS (303 to 291). Both gene panels were liberal and included genes with limited evidence of association with either cardiomyopathy or arrhythmia disorders. Although sequencing data was also available for *TTN*, this was not included in the subsequent analyses due to complexities resulting from its immense size (n residues = 34,350).

Genetic analysis of OMGL cases consisted of; target enrichment with a custom-designed Agilent HaloPlex kit, sequencing on Illumina MiSeq, adaptor trimming using Cutadapt and short or low-quality read removal with Trimmomatic. The same workflow was followed for the HCMR cohort, with the exception of the target enrichment kit, which was a custom-designed Illumina Truseq kit. Joint bioinformatics processing of all disease cohorts proceeded with mapping to the human reference genome (hs37d5 assembly) with BWA and haplotype calling and genotyping with an in-house pipeline adapted from the GATK Best Practices. All OMGL variants were confirmed by Sanger sequencing. All HCMR variants were visually confirmed by manual inspection of the BAM files.

For all genes, only Ensembl canonical transcripts were considered. All datasets were annotated using VEP. This included *gnomad_e2*, which was downloaded in VCF format (https://gnomad.broadinstitute.org/downloads/), then re-annotated using the same VEP pipeline. Subsequent analyses, treat both insertions and deletions, affecting three or fewer amino acid

residues, as *missense* variants. In both cases and controls, only variants with less than 0.01% frequency in all ancestry groups in *gnomad_e2* were included. The standard 'popmax' filter, proposed by Lek *et al.* (2016), does not include Finnish, Ashkenazi Jewish or the heterogeneous "Other" group. This is because due to bottlenecks and other demographic factors, genuine pathogenic variants may rise to higher frequencies in these populations (Kerminen *et al.* 2017; Zlotogora 2018). However, while this seems sensible for Mendelian disease variant prioritization, for aggregation analysis, including these population-specific founders has the potential to add more noise than signal. Therefore, a stricter popmax was implemented by excluding any variant above 0.01% in any *gnomad_e2* sub-population.

Depth of sequencing coverage (henceforth coverage) was calculated using the BAM files for the OMGL and HCMR datasets using 'samtools depth' (Li 2011) and coverage files were downloaded from the GnomAD browser for *gnomad_e2*. The proportion of samples with mean 10X coverage at each residue was assessed in all genes on the cardiomyopathy and arrhythmia gene panels (Figure 2.1). For some genes, low coverage regions existed in the cases. Although these poorly covered regions were filled using Sanger sequencing in the clinical labs, this data could not be mapped to the NGS data here without sample identification. As this could not be obtained for privacy reasons, there was no way to recover these low coverage regions.

Some low coverage regions also existed in *gnomad_e2*. For the cardiomyopathy gene-panel (*gnomad_e2* in blue), this is most evident in the large *DMD, FLNC, MYBPC3* and *SCN5A* genes and the smaller *DES*. Less prominent regions of low coverage were also present in several more genes including *TNNI3* and *LMNA*. For the arrhythmia gene-panel (*gnomad_e2* in green), multiple genes had markedly uneven coverage across the genes. The most notable examples were the two calcium channels *CACNA1C* and *CACNA2D1* as well as *KCNH2*, *RYR2* and *TRDN*. BrS had high coverage across the gene-set, whereas LQTS had specific regions across many genes with very low coverage.

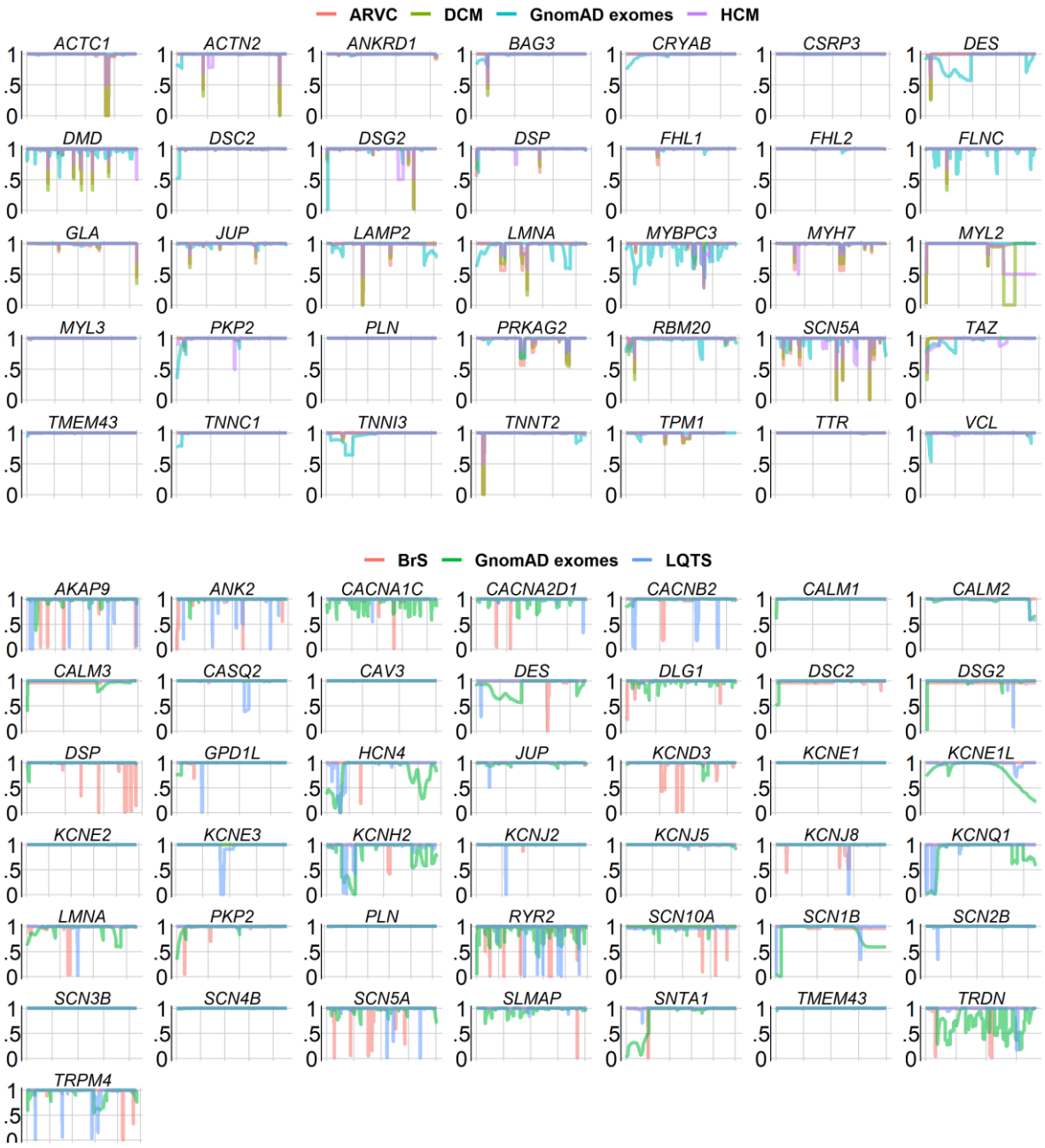

Figure 2.1: Proportion of samples across the 6 datasets with at least 10X coverage for each residue in the 35-cardiomyopathy gene-panel proteins and the 43-gene arrhythmia gene-panel proteins. Each subplot is scaled from 0-1 on y-axis and 1 to the protein length on the invisible x-axis. Downward spikes indicate regions of reduced coverage.

2.2 Development of *ClusterBurden*

2.2.1 Overview

This section reports the development of a rare-variant association test (RVAT), *ClusterBurden*. From the outset, the strategy was to create a method to detect clustering, then combine this with the result of a burden test. Later, a third test comparing differences in i*n-silico* prediction scores is also considered. The section is broken into five subsections; 1) developing a test for clustering 2) dealing with confounding by uneven sequencing depth 3) defining the burden testing approach 4) combining burden and position in *ClusterBurden* and 5) inclusion of *in-silico* predictors.

2.2.2 Detecting variant clustering

Missense variants are usually defined as single nucleotide polymorphisms that result in the substitution of one amino acid for another in the coded protein. The definition used here also includes short inframe insertions and deletions (indels). Inframe indels occur when the number of inserted or deleted bases are in multiples of 3. Thus, they do not impact the downstream reading frame for each gene and are therefore not necessarily loss-of-function variants.

Every protein is made of a linear sequence of amino acid residues, labelled from 1 to n, where n is the length of the protein. Residue 1, which encodes for the starting residue Methionine, cannot be missense as a variant here is a start-loss variant. Similarly, residue n, which encodes a stop codon, cannot be missense as it is a stop-loss variant. Throughout this thesis, the index of a missense-variant in this linear protein sequence (from 2 to n-1) is considered its position.

The hypothesis underpinning the clustering test is that the pathogenicity of missense variants varies across the linear sequence of the protein. Different regions of the protein contribute to protein structure and function in different ways, and particular structural or functional change may be specifically related to a disease phenotype. A detectable consequence of this phenomenon is

that in genes where missense variants cause disease, pathogenic variants may form clusters in specific protein regions. A naive approach to detect the presence of this variant clustering would be to look for a non-uniform distribution of variants in the case-cohort (Walsh *et al.* 2019). However, the background distribution of missense variants in controls is not uniform (Figure 2.2), so this may result in false-positives where clusters present in regions of high mutability or where adjacent regions are more constrained. Therefore, the distribution of observed case variants must be compared against the background distribution represented by the population controls. This approach automatically controls for variation in mutational rates and patterns of intra-species constraint irrelevant to the investigated phenotype.

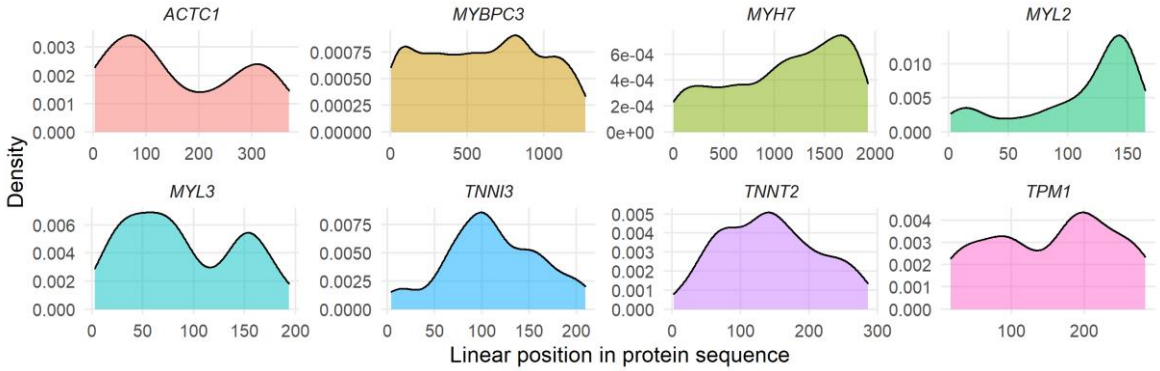

Figure 2.2: Gaussian density plots for population control variants in the eight-core sarcomeric genes definitively associated with HCM. Data include all missense variants occurring at frequencies less than 0.1% (subpopulation maximum) in *gnomad_e2*. Due to the large sample size of the cohort (n~125,748), these control distributions are derived from a large number of variant carriers; *ACTC1* (63), *MYBPC3* (2360), *MYH7* (2908), *MYL2* (172), *MYL3* (145), *TNNI3* (270), *TNNT2* (260) and *TPM1* (91).

2.2.3 The BIN-test

In order to detect differences in clustering between cases and reference controls, a chi-squared two-sample test was performed. The protein sequence was split into *k* bins of equal length, and the total number of rare-missense variants in each bin was counted for each cohort. As an example, counts for the gene *TNNI3* in our HCM-*gnomad_e2* case-control dataset are displayed in Table 2.2.

Splitting the linear sequence of the protein into ten even-sized bins, and counting the rare-missense variants occurring each bin in each cohort, produces this contingency table.

Table 2.2: Example contingency table for a chi-squared two-sample goodness-of-fit test. The columns represent each equal-sized bin across the protein sequence (due to non-integers results from after division of protein length into bins, some bins sizes may differ by one residue).The column headers display the ranges collapsed to form each bin. Counts within each cell, are the number of rare-variants observed in cases (row 1) or controls (row 2).

	1 - 21	*22 - 42*	*43 - 63*	*64 - 84*	*85 - 105*	*106 - 126*	*127 - 147*	*148 - 168*	*169 - 189*	*190 - 210*
Cases	1	1	1	2	1	0	31	19	13	15
Controls	22	6	12	40	58	42	30	32	8	20

A chi-squared two-sample test (hereafter *BIN-test*) was then applied to this contingency table. The *BIN-test* statistic is defined as: $B = \sum_{i=1}^{k} \frac{(K_1 R_i - K_2 S_i)^2}{R_i + S_i}$ where k is the total number of bins, R_i is the frequency of observed variants in the i^{th} bin for cases, S_i is the frequency of observed variants in the ith bin for controls and K_1 and K_2 are scaling constants to adjust for unequal numbers of observations between the groups. K_1 is calculated as the square root of the sum of counts in the second group divided by the sum of counts in the first group, and vice versa for K_2. This is important to adjust for unequal sample sizes. For the above contingency table, the scaling constants K_1 and K_2 are equal to 1.79 and 0.557. The chi-squared test statistic (χ^2) is then calculated according to the above equation for each bin (Table 2.3).

Table 2.3: Chi-squared statistic for each bin in a chi-squared two-sample test. The test statistic is the sum of chi-squared statistics in each bin and has k-1 degrees of freedom where k is equal to the number of bins.

	1 - 21.9	*21.9 - 42.8*	*42.8 - 63.7*	*63.7 - 84.6*	*84.6 - 106*	*106 - 126*	*126 - 147*	*147 - 168*	*168 - 189*	*189 - 210*
χ^2	4.8	0.34	1.8	8.3	16	13	25	5.2	17	7.1

The individual bin statistics are then summed ($\sum_{i=1}^{k} bin\ statistics$) to give the final test statistic *B* which is asymptotically chi-squared distributed with k-1 degrees of freedom (k = number of bins). In the example case, the final test statistic is 98.1 with 9 degrees of freedom producing a p-value of $<3.8 \times 10^{-17}$. As the test statistic *B* is chi-squared distributed, it is calculated analytically and does not require computationally expensive and time-consuming permutation. Essentially, the test looks for scaled differences in the counts in each bin between the cohorts. When the distributions are identical, all bins have a relative difference of zero, leading to a p-value of 1. However, the test statistic increases proportionally to the number of bins with differences in their relative frequencies of variant counts, and by the magnitude of these differences.

2.2.4 Optimal binning procedure

The method for binning the data is not automatically inferred in the typical use case. How many bins and their locations are hyperparameters of the test. The optimal binning procedure is dependent on the latent properties of the data. The ideal procedure would be to bin all variants in pathogenic regions and bin all variants in benign regions. Without knowing where the pathogenic and benign regions are, this is not possible. One approach would be to collapse on known genomic or protein features, such as domains or exons. However, this may be nonoptimal due to the huge heterogeneity in exon sizes, number of exons and number of domains (if any). In general, for most use cases of this test, regions are split into *k* bins of equal size. The choice is then simplified to choosing the number of bins *k*. One factor to consider is that each additional bin adds a degree of

freedom to the test. As a result, the most powerful test results from using the minimum number of bins to capture the latent distribution differences in the data. It has been demonstrated that smaller bins are more powerful to detect smaller clusters (White *et al*. 2009). However, without prior information on the latent clustering of variants, it is difficult to use this information.

The exon-structure of a gene may be an intuitive way to bin the data, as well as a familiar vocabulary for delineating clustered regions of genes in the literature. This was not the chosen method for two reasons. 1) There is a vast inconsistency in exon numbers and exon sizes between genes. For example, in 1,493 (7.5%) of genes, only one exon exists, but the protein lengths in this same gene set range from 24 (e.g. *MTRNR2L10*) to 2452 (*CCDC168*). For these genes the test would not be possible, for many other genes with few exons the test would have very low resolution, and overall it would be inconsistently applied across the exome. 2) Despite commonly reported as units of clustering in pathogenic genes, the observed clusters in the cardiomyopathy genes considered in this thesis, were not contained within exons.

A number of heuristics to automate this decision were explored. Mann and Wald (1942), suggested a formula to decide the optimal number of bins based on the number of observations. A simplification of this can be represented as $k = n^{2/5}$ where k is the number of bins and n is the number of observations. Cochran (1952) proposed general guidelines that each cell count should exceed one and that most should exceed five. Although this may be conservative (Koehler and Larntz 1980), in the spirit of this guidance, we split the genes into bins every five case variants, ensuring each bin had at least five variants in cases. As the controls were so much more numerous than the cases, this almost always meant there were at least five variants in controls. The software Palisade (http://www.palisade.com/stattools/) uses the conditional formula *n/5* for $n < 35$ and $1.88 * n^{2/5}$ for $n >= 35$ to choose bins. A static approach was also considered where the data was split into a static number of bins ranging from 3 to 25. Finally, a permutation method was considered. Here, the χ^2 value of the *BIN-test* was calculated for bin sizes 3 through to 25. The maximum of χ^2

of these 23 tests was then calculated as χ^2max. Next, the data was permuted by shuffling case-control status across all observed variants 1,000 times and calculating χ^2max each time. The proportion of these permuted χ^2max statistics more extreme than empirical χ^2max, was specified as the p-value for this method. Approaches to choosing *k* bins were classified as heuristic, static, or permutation and are summarised in Table 2.4.

Table 2.4: Approaches to choosing optimal bin number *k* for a chi-squared two-sample test given the number of observations *n*. Optimal choice of k is dependent on latent properties of the data at hand but can be approximated by the total number of observed variants n. 'Rule-of-thumbs' have been proposed (Mann-Wald, Cochran), choices can be static (e.g. k=10) or permutation can be implemented to explore multiple values of k.

Name	Type	*n* bins	Notes
Bins 3-25	Static	$k = (3 – 25)$	
Mann-Wald	Heuristic	$k = n^{2/5}$	
Palisade	Heuristic	$K = n/5$ for $n < 35$ and $1.88 * n^{2/5}$	
Cochran 5	Heuristic	$k = n\ (cases) / 5$	Break point every 5 case variants
Cochran 10	Heuristic	$k = n\ (cases) / 10$	Break point every 10 case variants
Permute	Permutation	$k = (3\text{-}25)$	Permuted over all bins

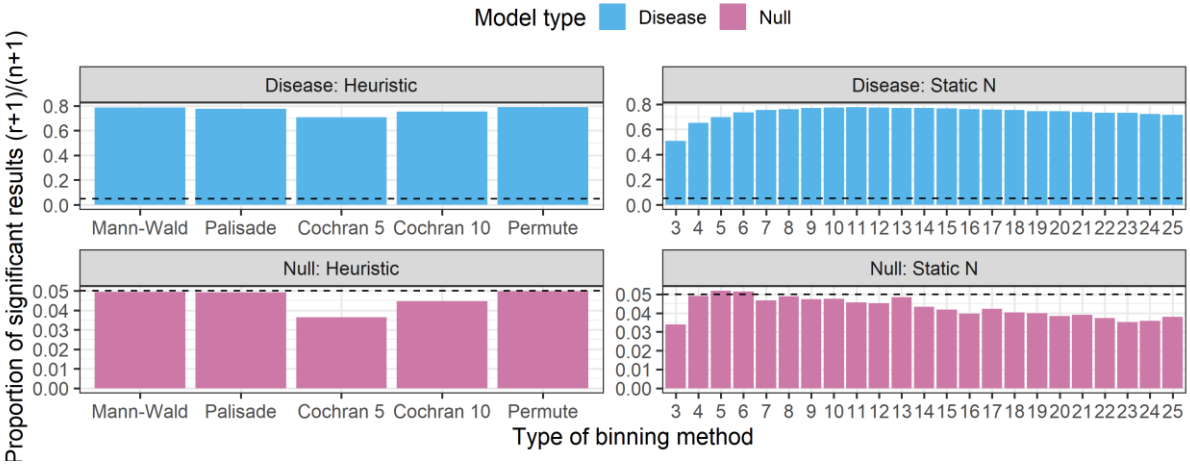

Figure 2.3: Theoretical power and type 1 error estimates for different chi-squared two-sample test data-binning approaches. Metrics are calculated by the proportion of simulated datasets that are significant at alpha = 0.05.

To select the optimal binning method, 10,000 simulations were generated from a simple framework (Figure 2.3). Each parameter of the simulation model was drawn from a distribution to maximise the breadth of scenarios considered. Protein lengths were sampled from the empirical lengths of all proteins in the human proteome with a length greater than 200 (n= 16753, mean=648, UniProt). Samples sizes for each cohort were simulated from a uniform distribution between 2000 and 10,000. The number of variants for each cohort was simulated from a binomial distribution with probability drawn from a uniform distribution between 0.01 and 0.02. For pathogenic clusters, their size was drawn from a uniform distribution between 1/20 and ¼ of the total protein length, and their location was random within the linear sequence. The probability of a variant affecting a specific residue was drawn from an exponential distribution; this caused some residues to harbour multiple variants imitating the site frequency spectrum of real data. To specify pathogenic clusters, cluster odds-ratios (ORs) were simulated from a uniform distribution between 2 and 5. These ORs represented the relative excess of case variants in this region. The reciprocal (multiplied by 1.5) represented the relative depletion for controls. As such, the depletion (population-level constraint) was less extreme than the excess in cases, as witnessed in our real data.

Although average power was similar between approaches, the Mann-Wald heuristic (power=78.7%), was marginally more powerful than the static approaches and other heuristics, equal up to three significant figures to the permutation approach (power=78.7%). Interestingly the differences between static N approaches were also small but choosing a k close to 10 gave the optimal average performance. As the number of bins increased beyond this, the test became conservative and power decreased, presumably due to increased degrees of freedom. Calculating p-values for all *ks* between 3 and 25, then permuting to get a final p-value was only marginally the most powerful approach, but with the added complexity of permutation. Furthermore, when data is sparse, the total number of possible unique permutations is low, producing imprecise results. The heuristics Cochran 5 and Cochran 10 were conservative under the null and gave underpowered

results. In light of these results, we used the $k \sim n^{2/5}$ heuristic (Mann and Wald 1942) to select the number of bins (k) dependent on n, the number of observations.

As not based on a regular grid of parameters, the effect of individual parameters could not be assessed in a conventional way. However, by keeping the randomly generated parameters for each simulation the relationship between p-values and model parameters was assessed. When considering all p-values in the disease model, significant results ($p < 0.05$) were characterised by more case variants, more control variants and larger cluster sizes, while protein length and cluster location were largely similar (Table 2.5).

Table 2.5: Mean parameter values for significant ($p < 0.05$) and insignificant ($p > 0.05$) simulation results. The standard deviation in this mean across different bin selection methods is provided in brackets next to the mean parameter values. P-values indicate differences in mean parameters between cases and controls.

p	**Case variants**	**Control variants**	**Protein length**	**Cluster Size**	**Cluster Location**
P > 0.05	87.54	80.55	646.92	72.27	142.98
P < 0.05	97.63	100.2	649.74	106.42	143.2
T.test p-value	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	0.3578	$<2.2 \times 10^{-16}$	0.7538

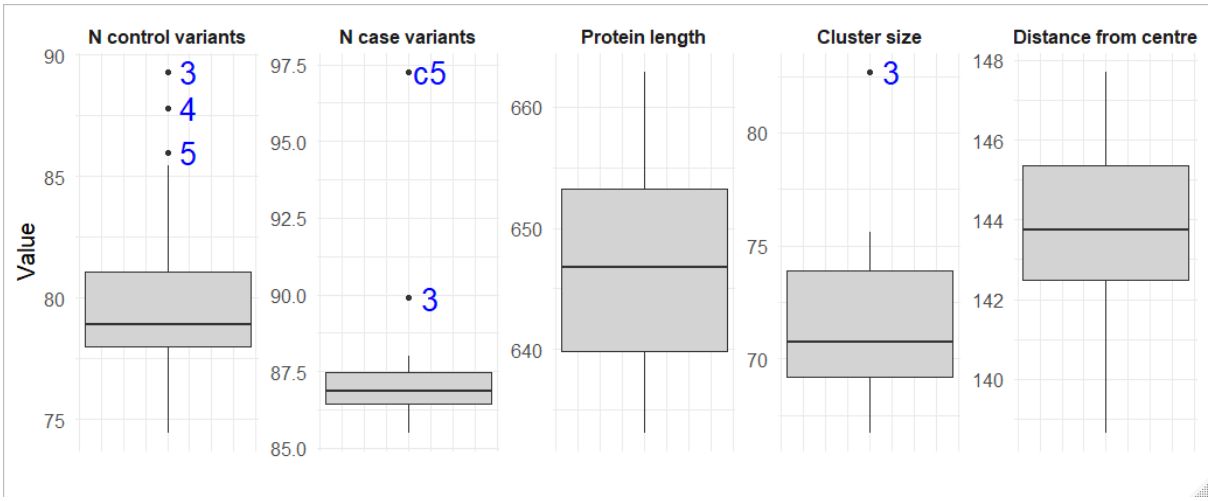

Figure 2.4: Mean parameter values for the significant results under each of the 27 binning approaches. Outliers, determined as values greater than 1.5 interquartile ranges from the first and 3rd quartiles, are labelled and highlighted in blue. The labels *3, 4* and *5* indicate a static number of 3, 4 or 5 bins and *c5* indicates the Cochran 5 approach.

The mean parameter values for significant results, did not significantly vary between bin selection methods for protein length ($p<1$), cluster size ($p<0.64$) or cluster location ($p<0.77$). However, significant differences in the mean number of case ($p<7.6 \times 10^{-11}$) and control ($<2.2 \times 10^{-16}$) variants in significant results was observed. Outlier groups that this may have been driven by are labelled in Figure 2.4. Using three bins gave mean parameter values for significant results that were identified as outliers in 3/5 parameters. One interpretation, consistent with its reduced power, was that more variants and larger cluster sizes were needed to reach significance. The same may be true for the 4 bin, 5 bin and Cochran 5 approach.

2.2.5 Controlling for coverage

Depth of sequencing (coverage), is the number of unique reads that cover a specific nucleotide base in the sequenced DNA. Coverage is an important potential confounder of case-control rare-variant association studies and can substantially impact results (Lee *et al.* 2014). In the case of

burden testing, differences in mean coverage between the case and control group can impact the observed number of rare-variants in each cohort, resulting in an excess of variants in the higher coverage group, even when true frequencies are the same. For the *BIN-test*, and any other position-based test, uneven coverage arguably has even more potential for confoundment. If coverage is uneven across a region, and this unevenness differs between cohorts, a low coverage region in either cohort may present as a cluster in the opposing cohort.

Like all technical artefacts, confounding by coverage is more likely when cases and controls are sequenced under different conditions. The inherited heart disease data from OMGL, the HCM cases from HCMR and *gnomad_e2*, are all sequenced using different capture technologies, sequencing machines, and processed using different bioinformatics pipelines. Even *gnomad_e2* itself is a smorgasbord of different exome-sequencing projects. For this reason, mean coverage differences and uneven coverage distributions can be expected throughout the genes investigated.

Although coverage is a known technical confounder, currently most advice suggests matching case and control cohorts (Guo *et al.* 2016, Lee *et al.* 2014, Neale *et al.* 2011), or excluding data with mismatched coverage, or coverage below a specific threshold, based on a specific metric such as <90% 10X coverage (Guo *et al.* 2018*,* Povysil *et al.* 2019). Few approaches exist to control for this confounder in unmatched analyses without excluding potentially useful data. Due to the limited amount of data in our analysis, a less conservative approach was attempted. Residue positions with lower than 10X coverage in 50% of samples in either cases or controls were removed from the analysis (Table 2.6). For the remaining data, uneven coverage was adjusted for in the *BIN-test* and the burden test.

Table 2.6: Regions with less than 50% of samples with at least 10X coverage. Only genes with at least ten residues excluded are shown.

Dataset	Gene	N excluded	Excluded residues
ARVC	DMD	16/3685	697-710, 1826-1827, 2197-2199

BrS	AKAP9	27/3907	113-115, 193-202, 1504-1506, 2253-2255, 2878-2879, 2886-2886, 3097-3097, 3896-3907
BrS	CACNB2	25/660	41-53, 420-427, 547-553
BrS	CASQ2	15/399	246-261
BrS	HCN4	57/1203	125-179, 201-204
BrS	KCNH2	17/1159	178-180, 209-211, 284-292, 300-305
BrS	KCNQ1	27/676	2-10, 38-57
BrS	RYR2	25/4967	2169-2169, 2172-2185, 2755-2757, 3006-3010, 3691-3693, 3829-3831, 4647-4649
BrS	TRPM4	13/1214	93-95, 390-391, 716-718, 771-775, 781-785
DCM	DMD	16/3685	697-710, 1826-1827, 2197-2199
DCM	MYL2	16/166	1-1, 118-134
gnomad_e2	AKAP9	35/3907	311-346
gnomad_e2	DSG2	14/1118	1-15
gnomad_e2	HCN4	223/1203	1-166, 1021-1079
gnomad_e2	KCNE1L	20/142	122-142
gnomad_e2	KCNH2	147/1159	158-305
gnomad_e2	KCNQ1	69/676	1-70
gnomad_e2	MYBPC3	14/1274	98-110, 910-912
gnomad_e2	RYR2	50/4967	1-16, 2710-2736, 2803-2812
gnomad_e2	SCN1B	12/268	1-13
gnomad_e2	SNTA1	102/505	1-103
gnomad_e2	TRDN	116/729	78-84, 131-161, 379-388, 457-473, 513-522, 533-549, 595-610, 624-639
LQTS	AKAP9	27/3907	113-115, 193-202, 1504-1506, 2253-2255, 2878-2879, 2886-2886, 3896-3907
LQTS	CACNB2	25/660	41-53, 420-427, 547-553
LQTS	CASQ2	15/399	246-261
LQTS	HCN4	57/1203	125-179, 201-204
LQTS	KCNH2	16/1159	178-180, 209-211, 284-291, 300-305
LQTS	KCNQ1	27/676	2-10, 38-57
LQTS	RYR2	13/4967	2169-2170, 2185-2185, 2755-2757, 3006-3010, 3691-3693, 3829-3831, 4647-4649

Relative to the total number of residues analysed across all genes and datasets, few regions were removed from the analysis. Notable exclusions include the two strongest associated LQTS genes *KCNH2* and *KCNQ1,* which have 16/1159 and 27/676 residues excluded respectively. These genes also contained more substantial missing regions in *gnomad_e2* controls. The largest proportion of excluded low-coverage residues was observed in the *gnomad_e2* genes *KCNE1L* (14.1%)*, TRDN* (15.1%)*, HCN4* (18.5%) and *SNTA1* (20.2%).

After removing positions excluded by low-coverage, the Mann-Wald heuristic was used to split the region in *k* bins. For each cohort, the proportion of samples with at least 10X coverage was averaged over all genomic bases spanning the bin. The empirical counts in this bin were then divided by the reciprocal of this averaged metric. Therefore if the metric was 0.5, the minimum possible after excluding variant positions with less coverage less than this, then the cell counts were effectively multiplied by 2 for that cohort. Where cell counts were zero, a continuity correction was applied whereby cell counts were first set to 0.5 if coverage was below 90%. For the burden test, effective sample sizes for each cohort were adjusted by multiplying by the mean number of samples with 10X coverage across the whole region, after excluding positions with coverage below 50%.

The metric used to summarise coverage (the proportion of samples with at least 10X coverage) is widely used to specify low coverage regions where variant calls may be missed (Sims *et al.* 2014, Wang *et al.* 2017, Guo *et al.* 2018, Povysil *et al.* 2020). The proposed adjustment procedure essentially imputes missing data based on the assumption that no variants were called in samples without at least 10X coverage. For OMGL and HCMR data, stringent variant filtering criteria ensured this was true. Similarly, for *gnomad_e2*, adjusted allele counts were used with include only variants with a depth of coverage (DP) greater than or equal to 10.

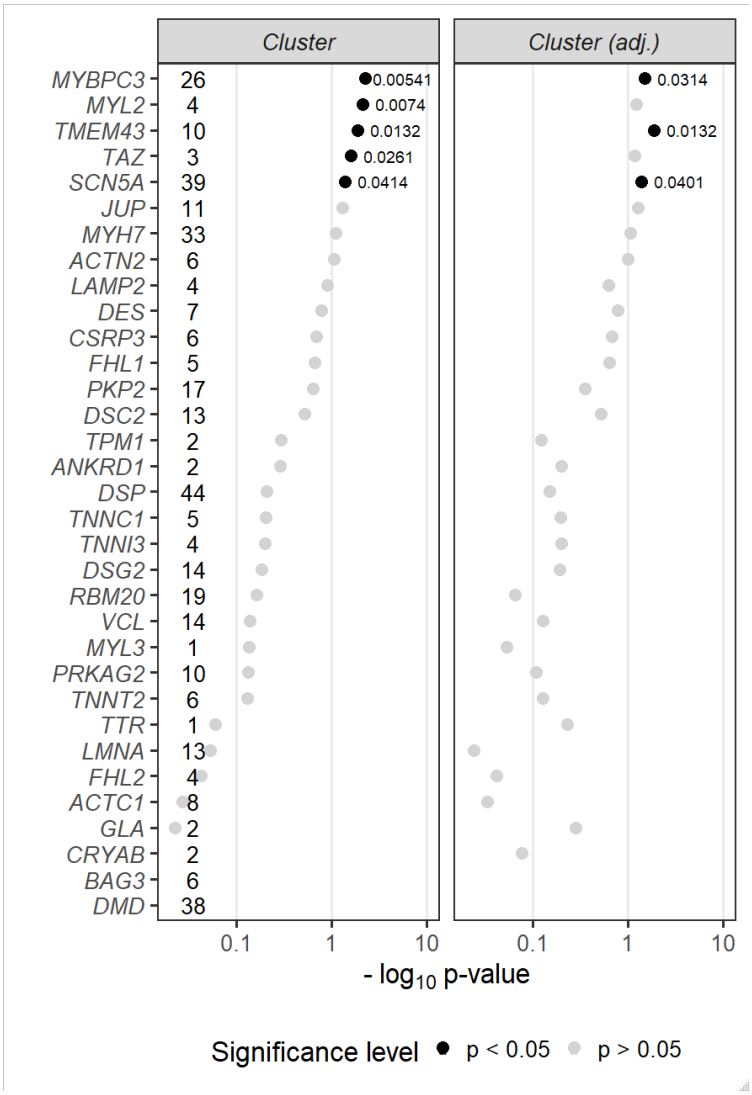

Figure 2.5: *BIN-test* p-values for synonymous variants in HCM cases versus GnomAD exomes controls. Differences in the location of synonymous variants in 5,338 HCM cases and up to 125,748 *gnomad_e2* controls were compared across the cardiomyopathy gene-panel. Most results were insignificant but without coverage adjustment, five genes gave nominally significant results. After up-weighting low coverage regions in each cohort, only three nominally significant genes persisted.

To explore the behaviour of this coverage-control approach, the false-positive rate of the *BIN-test* for synonymous variants was assessed, for HCM versus *gnomad_e2*. This scenario is assumed to represent a null model of no association. P-values were calculated adjusted and unadjusted for coverage (Figure 2.4). Reassuringly, no results were *b-significant* with or without adjustment for coverage. Furthermore, coverage adjustment reduced the significance of 5/6 *n-significant* unadjusted p-values leaving only three, all of which had p-values above 0.01. Interestingly, the most significant gene was *MYBPC3,* which has some previous evidence of pathogenic splice-disrupting synonymous variants (Ito *et al.* 2017), which could be driving the signal here. This coverage-adjustment procedure is explored in more detail in chapter 4.

2.2.6 Detecting excessive variant burden

Burden testing is a common method to assess the association between a gene and a disease (Morganthaler and Thilly 2007). The underlying hypothesis is simple; if variants in a gene cause a disease, then disease cohorts should carry a higher frequency of variants than unaffected controls (on average). Of course, in reality, observing this signal is dependent on a number of factors related to the genetic architecture of the trait, including; penetrance, mode of inheritance, gene length, background mutation rates, and locus heterogeneity (Guo *et al.* 2016). However, for well-powered studies, penetrant genes will present with easily detectable differences in the aggregated frequency of variants between an affected and unaffected group, for a truly causal gene.

To conduct a burden test, the gene of interest must first be sequenced, and variations from the reference sequence must be detected. For rare-variant association, observed variants are filtered based on population control frequencies to exclude putatively benign common variants. Variant classes with a low prior probability of pathogenicity, such as synonymous and intronic variants (Lek *et al.* 2016), are also filtered from the analysis. After filtering, one approach is to count the number of samples carrying a variant in the region for each cohort. Notably, while this is suitable for dominant conditions, for recessive, the number of homozygous and compound heterozygotes would need to be counted. The resulting contingency table of cohort status by carrier status can be assessed for significance using a Fisher's exact test (FE-test). This method essentially compares the frequency of variant carriers in cases versus the frequency of variant carriers in controls. In the absence of sample-level information, it is impossible to determine if two separate variant observations are derived from different samples. Such is the case when using summary-level population control cohorts such as GnomAD. In this scenario, the statistical test approximates an additive genetic model.

In rare-variant analyses, observations are sparse by definition. Most samples do not carry a variant in the region under study and for many genes, the total carrier counts may be fewer than 5 (or even 0). Under these conditions, a FE-test is preferred over a chi-squared test, as the chi-squared test statistic is no longer reliably approximated by a chi-squared distribution. An assumption of FE-test, is that the row and column totals of the contingency table are fixed before the investigation. For a burden test, this means the number of samples and number of observed variants are fixed. This is not correct as the number of observed variants in each cohort are unknown before the study. In this scenario, a singly-conditioned exact test of odds ratios (Ludbrook 2008) is a more appropriate statistical test. However, this approach vastly increases the number of permutations required to derive a p-value from a 2 by 2 contingency table, especially when sample sizes are large. Consequently, it is impractical in the setting of large population control cohorts, or whole-exome burden testing where the number of tests is high.

2.2.7 Unconditional-exact test

It has been previously noted that FE-test is conservative, especially when counts are low (Routledge 1992). To determine whether a computationally expensive unconditional-exact test (UE-test) provided improvements to power, the power and type 1 error of the two approaches was calculated using simulated data. As the UE-test was so slow, a small simulation framework was implemented. For both the disease and null model, the case and control sample sizes were 1000. For the disease model, variant carriers were drawn from a binomial distribution with probability 0.02 for cases and 0.01 for controls, simulating a true OR of 2. For the null model, the probability was 0.02 for both cohorts, simulating a true OR of 1. The proportion of significant results out of 1000 simulations, for each test in the null and disease model, are displayed in Figure 2.6.

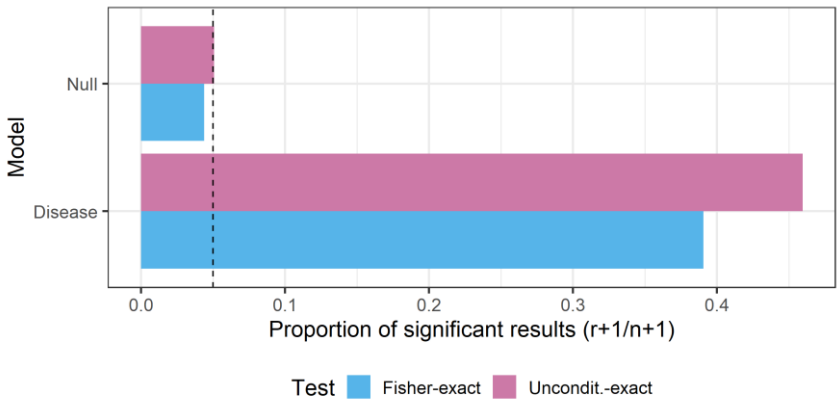

Figure 2.6: Power and type-1 error of Fisher's exact test and an unconditional exact test in a small simulation scenario. A simple simulation scenario, where counts are modelled by a binomial distribution. For the disease, an average of 2% of cases and 1% of controls have a variant in the gene (true OR=2). For the null, 2% of both cases and controls have a variant in the gene (true OR=1).

Under these circumstances, with relatively small sample sizes, FE-test was conservative under the null and was less powerful in the disease model, than the UE-test. However, even with these relatively small contingency tables, it still took 30 minutes for the UE-test over 1000 simulations, whereas the FE-test took less than 1 second. As the relationship between contingency table size and computational time is approximately $O(n^2)$ for the UE-test, this is prohibitively slow. It is also worth noting that the conservative nature of FE-test is dependent on the sample size with smaller sample sizes leading to increased conservativeness. Attempts at simulating at larger samples sizes resulted in memory issues for the UE-test. It was decided that a memory-efficient reimplementation of the UE-test was beyond the scope of this thesis. Therefore, the conservative FE-test was used as a computationally practical alternative.

In inherited heart disease, at least, there are no known examples of a protective burden of rare exonic variants. This situation would arise where genetic variants in the region under consideration have an average protective effect. As no evidence of this occurring could be found in the literature for inherited heart disease genes, this hypothesis was ignored and only one-sided significance tests were considered. This means only an excess burden of case variants produces a significant result.

The p-value is then the proportion of permuted contingency tables with odds ratios *higher* than the observed one. In the example contingency table shown (Table 2.2), this halves the p-value from 0.03757 to 0.1878. It is therefore a more powerful approach to detect an excess of rare-variants in cases.

2.2.8 ClusterBurden – a rare variant association test

To assess the joint significance of both a difference in variant positions and a gene-wide case excess of variants, the p-values of the contributing tests were combined. This was achieved by using Fisher's method (Fisher 1925) and the complete framework was named *ClusterBurden*. Fisher's method was created to combine p-values from independent tests converging on the same null hypothesis (Figure 2.7). This usually involves combining the results from two identical tests calculated using different data, here it is applied to two different tests on the same data with an approximately equivalent null hypothesis. While the alternative hypothesis for each test here is not exactly the same, they converge on the same overarching hypothesis that genetic variation in the gene under investigation is associated with the disease. P-values are combined by summing their natural logarithms and multiplying by minus two. This produces a chi-squared statistic with 4 degrees of freedom. As the approach increases the degrees of freedom, this imposes a penalty when either contributing test is insignificant. This is an essential property to retain type 1 errors at the nominal threshold. However, as a consequence, *ClusterBurden* will be less powerful than a standard burden test when clustering is completely absent. Example p-value combinations using Fisher's method are displayed in Figure 2.6. An assumption of Fisher's method is that the contributing p-values are uncorrelated. This assumption was assessed in data simulated under the null via a Spearman's rank correlation coefficient (Spearman 1904).

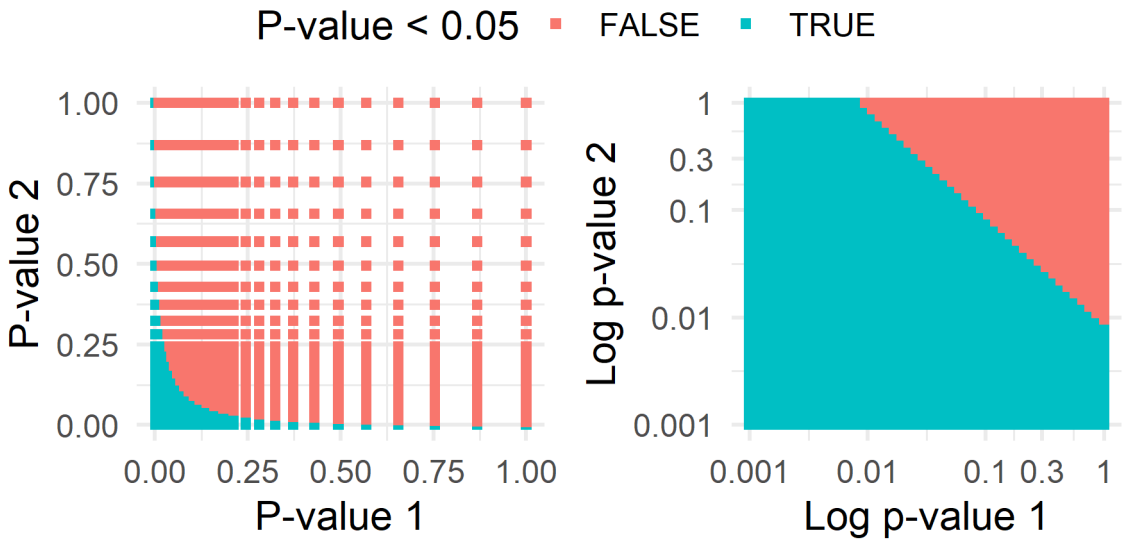

Figure 2.7: Visual explanation of p-value combination using Fisher's method. For each subplot, a sequence of 50 p-values were taken from a log distribution from 0.001 to 1. Each pair of these p-values were combined using Fisher's method and plotted.

2.2.9 In-silico variant pathogenicity prediction

There are many computational models that attempt to predict the pathogenicity of missense variants. We hypothesised that in disease-genes, case-variants would be predicted as more pathogenic than control-variants on average. To test this hypothesis, variants were annotated with twenty-seven *in-silico* prediction scores from the publicly available database dbNSFP. For each prediction score we calculated the Spearman rank correlation between HCM (cohort=1) and *gnomad_e2* (cohort=0) status in nine firmly established sarcomeric HCM disease-genes (Figure 2.8). The majority of the scores had a correlation above 0.1 (Rho) and eleven scores had correlation above 0.2. While not strong correlations, there is indication that these scores have predicted value for variant pathogenicity in HCM. An important point which will be returned to in chapter 3, is that correlation is very unlikely to be 1 with this experimental design. This is because all rare-missense

variants are included in both cases and controls. As a result, incompletely penetrant pathogenic variants may be present in controls and neutral variants may be present in cases.

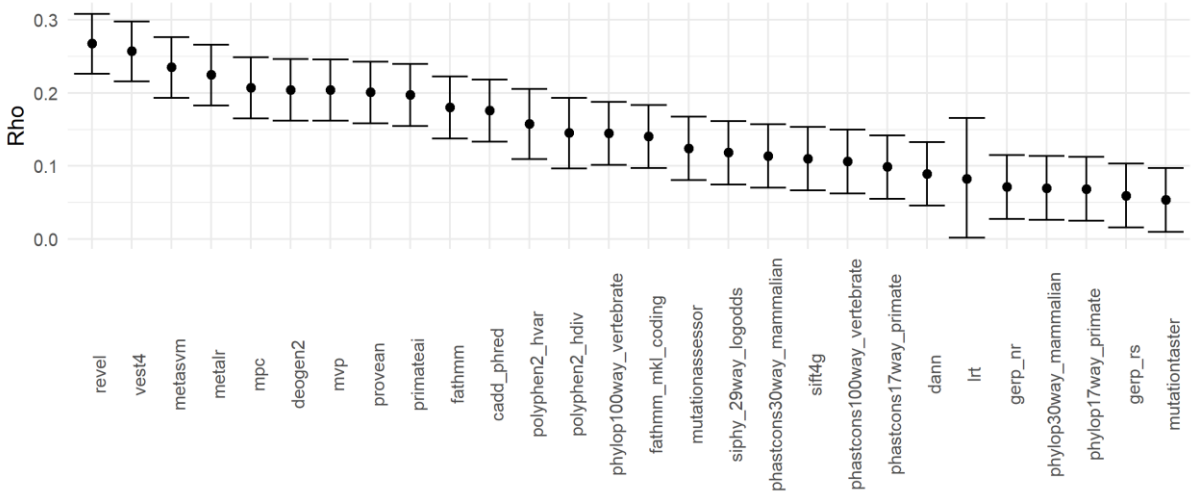

Figure 2.8: Correlation between variant prediction scores and disease status for missense variants in nine firmly-established hypertrophic cardiomyopathy causing genes. Rare-missense variants observed in 5,338 HCM cases and 125,748 were annotated with 27 prediction scores downloaded from dbNSFP. The Spearman rank correlation coefficient (Rho) and confidence intervals were calculated to estimate the correlation between cohort status and numeric pathogenicity score. All p-values were significant except lrt which had high rates of missingness.

To determine gene pathogenicity using these scores, a linear model of dbNSFP features was compared to an intercept-only model. This is equivalent to assessing whether the dbNSFP predictions explain more variance in case-control status than random chance. As models with more predictive features have more degrees of freedom, a feature selection procedure was implemented to reduce the feature-space of the model. First, a model was trained on all pathogenic or likely pathogenic *clinvar* variants and all common GnomAD variants. In the first stage, all 28 dbNSFP features were used. If any feature had a model p-value above 0.05, it was removed and a new model was trained without this feature (backwards elimination). This process was repeated until all features had a p-value above 0.05. The result was a model that included; PolyPhen-2 (HDIV), FATHMM, VEST4, REVEL, MVP, MPC, Deogen2, DANN and SiPhy 29way (logOdds).

An alternative strategy to choosing model predictors would be to dynamically choose the optimal predictors for each gene. While this approach is effective for building powerful predictive models, it fails to keep false positives below the nominal alpha when used in the setting of hypothesis testing without permutation. Therefore for every gene, the set of predictors that performed well in the generic model, trained on *clinvar* pathogenic and GnomAD common variants, were used. A likelihood ratio test (LRT) was then used to assess the goodness-of-fit of the *in-silico* model compared to a null intercept-only model.

2.3 The properties of the proposed methods

2.3.1 Forward-time simulator

In order to assess the type 1 error and power of the *BIN-test* and *ClusterBurden* versus published alternatives, simulated null and disease datasets were generated using a bespoke forward-time simulation engine (ClusterSim). In order to simulate protein sequences with discrete regions of increased pathogenicity potential, the algorithm allows differential selection across the protein sequence to predefine pathogenic regions. Higher relative selection coefficients for these regions resulted in constrained coding regions in the final simulated population, and later in the subset control group. The probability of variant causing disease was based on its stochastic selection coefficient (influenced by location) and again its location. Location was considered twice as population-level constraint is insufficient to specify disease clusters. This is because the constraint is only perfectly correlated with the disease phenotype when it is guaranteed to cause a fitness of zero. Therefore for phenotypes with a fitness above zero, the ratio of case-variant clustering to population-level constraint is higher. These two features, case variant-clustering and complementary control-constraint, give rise to similar discrepancies in variant positions, observed between cases and controls, in the heart disease cases and *gnomad_e2* controls.

ClusterSim was developed in the R programming language using object orientated programming with R's built-in reference classes. Efficient data-wrangling functions from the data.table package (Dowle and Srinivasan 2019) were used to evolve populations rapidly. To further speed things up, variants were ignored after becoming too common and recombination was not considered. These implementation shortcuts allowed the entire demographic and mutational history of a population to be evolved in less than a minute. The effect of recombination was ignored as it was assumed to be of low aggregate importance across simulations for rare-variant association in a single protein-coding region. To achieve the large numbers of populations necessary for accurate type 1 error and power calculations, populations were simulated in parallel on a high performance computing cluster (WTCHG ResComp).

Each population followed the European population history demographic model, proposed by Kryukov *et al.* (2009). The three-stage model begins with a burn-in stage of 10,000 generations with an initial diploid population of 8,100 individuals. This stage is followed by a brief "bottleneck", acknowledging the out of Africa hypothesis, and then a final stage of exponential growth allowing the population to reach a size of 900,000. In terms of rare-variant analyses, this final stage was the most important for our analysis, as exponential growth increases the proportion of rare-variants in a population (Kryukov *et al.* 2009).

Each generation evolved via a three-stage process of mutation, selection and mating. In each generation, the number of new mutations was randomly drawn from a binomial distribution X~B(n,p) where *n* is the total number of mutable sites in the population and *p* is the mutation rate. The mutation rate was calculated by multiplying 1.8×10^{-8}, the estimated per-nucleotide base mutation rate in the human genome (Peng and Liu 2011), by the three bases in a codon and the naïve-empirical probability of a missense mutation 0.619 (with the simplifying assumption that all base changes are equally likely and all amino acids are present in equal proportions). This approximates an average missense mutation rate of 3.34×10^{-8}.

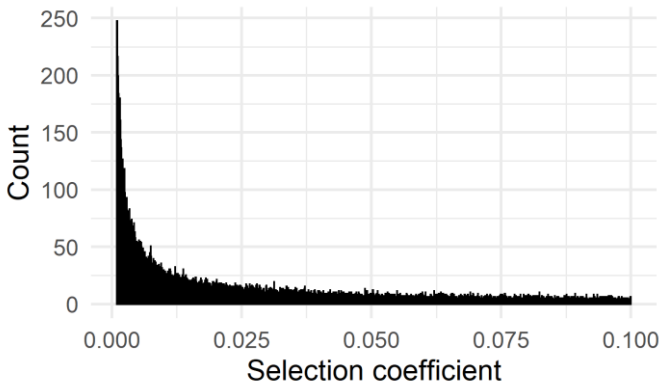

Figure 2.9: The distribution of randomly allocated selection coefficients assigned to new mutations in the simulated populations. The minimum coefficient was 0.0001 and the largest coefficient was 0.1.

Each new mutation was assigned a selection coefficient drawn from an exponential distribution between the range 0.0001 and 0.1 (Figure 2.9). The result of this process is that the majority of simulated variants are assigned a weak random selection coefficient close to 0.0001. Compared to a non-carrier, a selection coefficient of 0.1 gives the carrier a 10% reduced chance of success at the mating stage of the evolutionary cycle. Due to the exponential distribution, fewer variants had such large selection coefficients. In the mating process, the chance of success was dependent on the sum of selection coefficients for each variant carried by an individual (an additive model). For example, if a sample had two variants, the selection coefficients were added together to represent the fitness of the individual. As all variants were detrimental to survival, carriers had reduced fitness compared to non-carriers.

In order to simulate missense-variant clustering, discrete pathogenic regions were assigned in the disease models. The sizes of these pathogenic regions were randomly drawn to span from one-twentieth to one-quarter of the simulated protein length. These regions were randomly located and allowed to overlap, so that each simulated population considered a different pattern of pathogenicity potential across the linear protein. This prevented a bias towards tests that are known to be more powerful in detecting differences in distributions that occur in the middle (e.g. Kolmorgorov-Smirnov test) or the tails (e.g. Anderson-Darling test). These pathogenic regions modify the stochastic selection coefficients of variants that occurred within them. Before evolution of a population begins, a prior static selection coefficient was calculated for each simulated residue,

based on the discrete pathogenic regions in the protein. To calculate this, variants within a pathogenic region were given a coefficient of five and variants outside were given a coefficient of 1. Every residue was then divided by the mean coefficient across the protein to give the final static selection coefficients a mean of 1, for all simulated proteins. During the simulation the random selection coefficient assigned to each new mutation was multiplied by this pre-calculated static selection coefficient to give the final selection coefficient (FSC) used to implement selection.

In order to simulate phenotype (i.e. specify cases and controls), each simulated variant in the final population was allocated an OR. ORs for the final simulated variants were drawn from an L-shaped distribution with mean 2.60 and median 2.35. All variants in the population were then ranked by FSC and assigned an equivalent ranked OR, therefore variants with the highest FSC had the highest ORs. These ORs were then multiplied by 3 or 1/3, depending on whether they fall within or outside a pathogenic region. Therefore, a variant's position is considered twice. The first time to generate population-level constraint. The second time to cause case-variant clustering.

Affection status was then simulated based on these odds-ratios according to the formula**;** $\boldsymbol{P(Affected|X)} = \frac{e^{\alpha+BX}}{1+ e^{\alpha+BX}}$ where *X* is the vector of genotypes, *B* is their respective ORs and alpha was defined by log(prevalence / 1 –prevalence). Prevalence was specified as 1/500 to correspond to the population prevalence of hypertrophic cardiomyopathy (Maron *et al.* 1995). After each sample was assigned a phenotype, 2,500 cases and 123,000 controls were randomly selected from the population. Three different amino-acid residue clustering models were explored; uniform, single-cluster and multiple-clusters. In the uniform model, all variants in a protein were equally likely to be pathogenic. In the clustered models, only specific regions of the protein harboured pathogenic variants. All clustering scenarios were observed in the HCM gene panel data. For the multiple cluster model, two or three clusters were possible and were simulated in a ratio of 70:30. The relationship between protein length (number of amino acid residues) and statistical power was also explored by simulating at two different protein lengths; 500 and 1,000 residues.

2.3.2 *Simulated populations*

Ten thousand simulated populations were generated under the six simulation scenarios characterised by each of three clustering scenarios and two protein lengths in a more conventional grid scenario. This was done for both a null and disease model. To achieve a null model, cases and controls were randomly selected from simulated populations. Under the null model, 10,000 simulations produces precise estimates of type 1 error with a binomial exact 95% confidence interval (500 successes in 10,000 trials) of 0.0458 to 0.0545. Under the disease model, 10,000 simulations should be sufficient to distinguish between the power of two approaches.

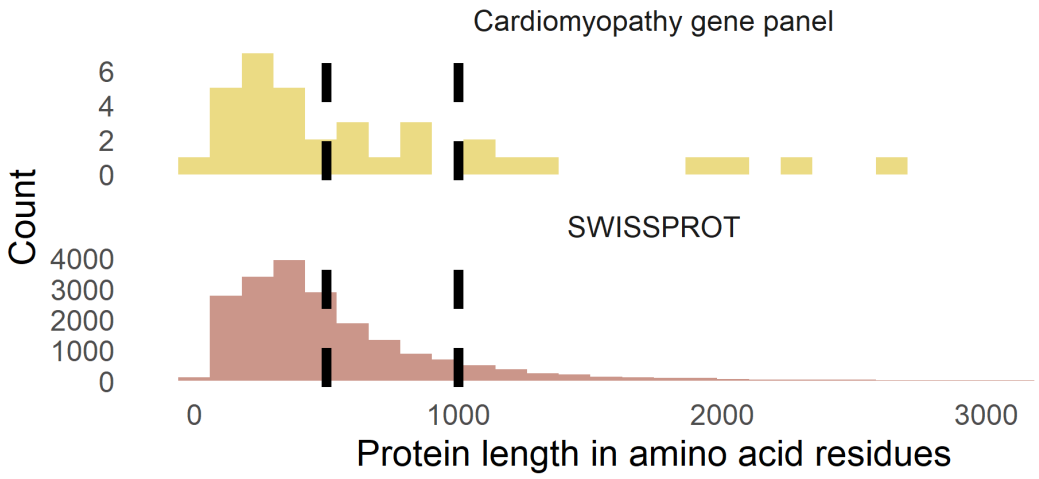

Figure 2.10: Protein lengths in 20,233 SWISSPROT proteins and 35 cardiomyopathy gene-panel genes. Blue dashed vertical lines at 500 and 1,000 denote the selected protein lengths for our simulations. Mean SWISSPROT=560, mean HCM = 688.

The chosen values for protein length; 500 and 1000, fall within the range of observed protein lengths for both SWISSPROT and the cardiomyopathy gene panel (Figure 2.10). The x-axis is truncated at 3,500 residues, excluding 105 very large proteins in SWISSPROT, such as *TTN*, that visually obscure the shape of the distribution. While, the mean protein length in the cardiomyopathy gene-panel (688) was higher than the mean SWISSPROT protein length (560), the median protein length was lower in cardiomyopathy (480) than SWISSPROT (491). This was driven

by some very large proteins including; *DSP* (2271), *FLNC* (2692), *MYH7* (1936) and *SCN5A* (1999). The smallest gene in the cardiomyopathy gene-panel was *PLN* which is only 51 residues long.

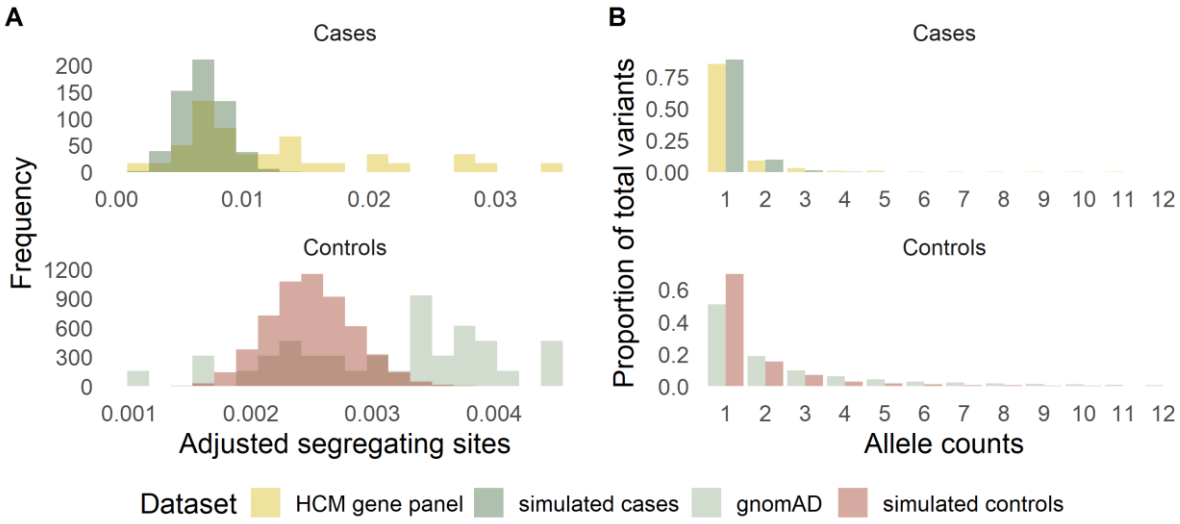

Figure 2.11: The number of segregating sites and the site-frequency spectrum for simulated data and observed data. Segregating sites are adjusted for sample size and total number of sites in the sequence.

To determine whether the simulated data was broadly similar to the observed data, the total number of segregating sites (nSS) and site frequency spectrum (SFS) was compared to the HCM and *gnomad_e2* datasets (Figure 2.11). When nSS was adjusted for sequence length and sample size, simulated nSS was similar to the empirical data but had fewer nSS on average (Figure 2.11A). This was intentional as highly penetrant genes such as *MYH7* and *MYBPC3* which skew the empirical nSS, are so easily detectable that all methods would have had 100% power to detect significance. The SFS were also comparable between the real and simulated data (Figure 2.11B). This is important as repeated observations, have implications for the assumptions of many statistical methods.

To evaluate the selection procedure, FSCs (final selection coefficients) were compared to the final allele counts (FACs) of simulated mutations. There is an inverse relationship between FACs and mean FSC, with high-frequency variants having weaker FSCs (Figure 2.12A). As intended, variants

within pathogenic regions, defined by the clustered models, had higher FSCs than variants outside these regions (Figure 2.12B). This demonstrates that variants in these regions are subject to purifying selection causing constrained regions in the final population.

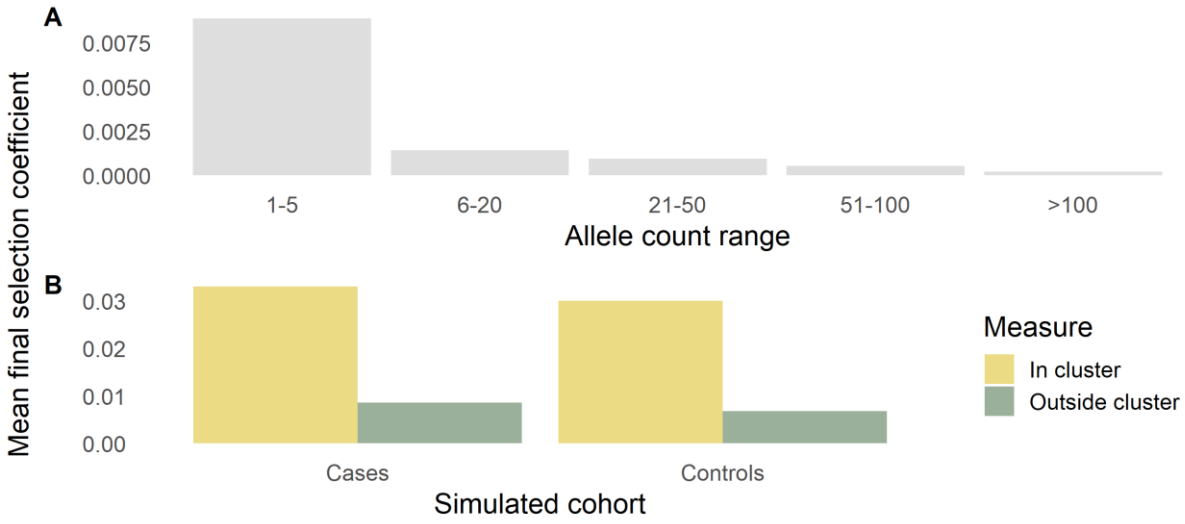

Figure 2.12: Selection coefficients of simulated variants. The upper panel shows the relationship between the final allele count of simulated variants and their static selection coefficients. The lower panel shows the average selection coefficient of variants that fall inside and outside the causal cluster(s).

2.3.3 Comparison of cluster-detection methods

We compared the *BIN-test* with four two-sample distribution tests (Table 2.7): AD-test, KS-test, ES-test and BWS-test. These tests were originally developed as goodness-of-fit tests to determine whether a sample of data is drawn from a given probability distribution. Subsequently, they have been adapted to two sample scenario, testing the null hypothesis that both samples are drawn from the same underlying probability distribution.

Table 2.7: Implementation details for all published statistical methods considered in comparisons of type 1 error and power estimates. All methods were implemented in the R programming language and parameters were kept as their defaults. Position = two-sample distribution test, RVAT = rare-variant association test.

Test name	Type	Implementation (R)	Original paper
Anderson-Darling (AD-test)	Position	kSamples::ad.test	Scholz and Stephens 1987
Kolmorgorov-Smirnov (KS-test)	Position	base::ks.test	Kolmorgorov 1932
Epps-singleton (ES-test)	Position	es.test ilkka.j.leppanen@aalto.fi	Goerg and Kaiser 2009
Baumgartner (BWS-test)	Position	BWStest	Pav 2018
CLUSTER	RVAT	homepage.ntu.edu.tw/~linwy/CLUSTER.r	Cheung *et al* 2012
DoEstRare	RVAT	DoEstRare::DoEstRare	Persyn *et al* 2017
WST	RVAT	assoctesteR::WST	Madsen and Browning 2009
C-alpha	RVAT	assoctesteR::CALPHA	Neale *et al* 2011
SKAT	RVAT	assoctesteR::SKAT	Wu *et al* 2011

To determine which method was the best at detecting variant clustering, the type 1 error and power of each clustering method was assessed in the simulated data (Figure 2.13). *BIN-test* was clearly the most powerful method with a mean power of 50.4% across both clustering scenarios and both protein lengths. The next best method was Anderson-Darling which has a mean power of 31% over the same simulated datasets. KS-test had comparable power to AD-test. Results from the null model clearly show that BWS-test and ES-test are not well calibrated for rare-variant distribution analyses. Multiple factors contribute towards this, including inability to cope with unequal numbers of observations between the groups. Both methods have a very high false-positive rate, hence why their power was not calculated in the disease model. Both *BIN-test* and AD-test kept type 1 errors at the nominal threshold of 0.05 for the null model the burden-only uniform model. KS-test was conservative in all null scenarios.

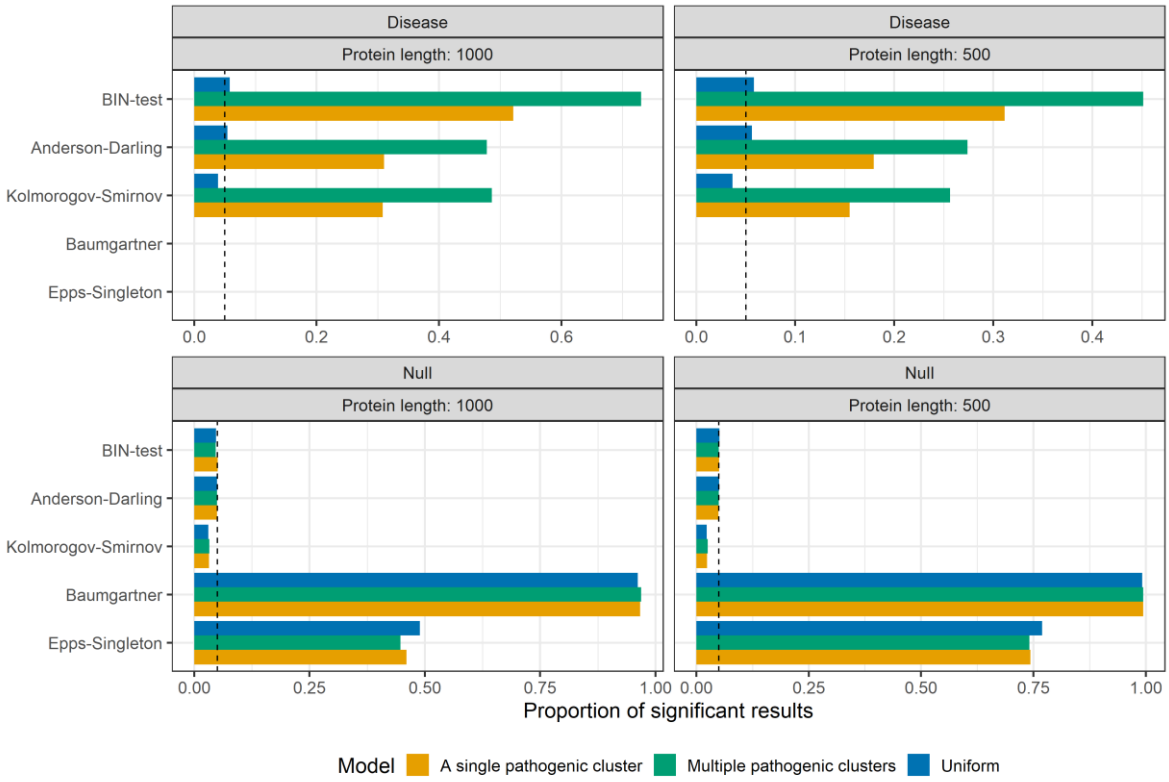

Figure 2.13: Type 1 error and power of two-sample goodness-of-fit tests to detect differences in variant clustering in simulated data. Due to markedly inflated false-positive rates in the null model, Baumgartner and Epps-Singleton were not assessed in the disease model.

2.3.4 Comparison of rare-variant association methods

To determine whether the p-values from FE-test and *BIN-test* were statistically independent, simulated p-values were correlated in a disease and a null model. In the disease model, the Spearman rank correlation coefficient (Rho) was 0.398 across all simulations. This indicates that when the disease is truly associated, power to detect a positional signal covaries with power to detect a burden signal. For the null model, Rho was equal to -0.0065 and the Spearman rank correlation test did not find statistically significant correlation between the two methods. This satisfies the assumption of independence for Fisher's method.

The performance of the ClusterBurden was compared to published RVATs, including the position-agnostic WST, C-alpha and SKAT, and the position-informed DoEstRare and CLUSTER (Figure 2.14).

When clustering was absent, the most powerful RVAT was CLUSTER, with a mean power of 89.4% across both protein lengths. However, when variants clustered with respect to cases or controls, *ClusterBurden* was the most powerful method, with 58.9% power to detect a single cluster and 84.1% power to detect multiple clusters.

DoEstRare was the next most powerful method in the clustered scenarios, with 54.9% and 82.2% power to detect single and multiple clusters respectively. CLUSTER was only marginally better than FE-test with or without clustering indicating that clustering signal had very low weight in the method. DoEstRare placed more weight on the clustering signal than CLUSTER, but less than *ClusterBurden*. This meant it had superior power than a FE-test to detect association of a clustered gene but less power than *ClusterBurden*.

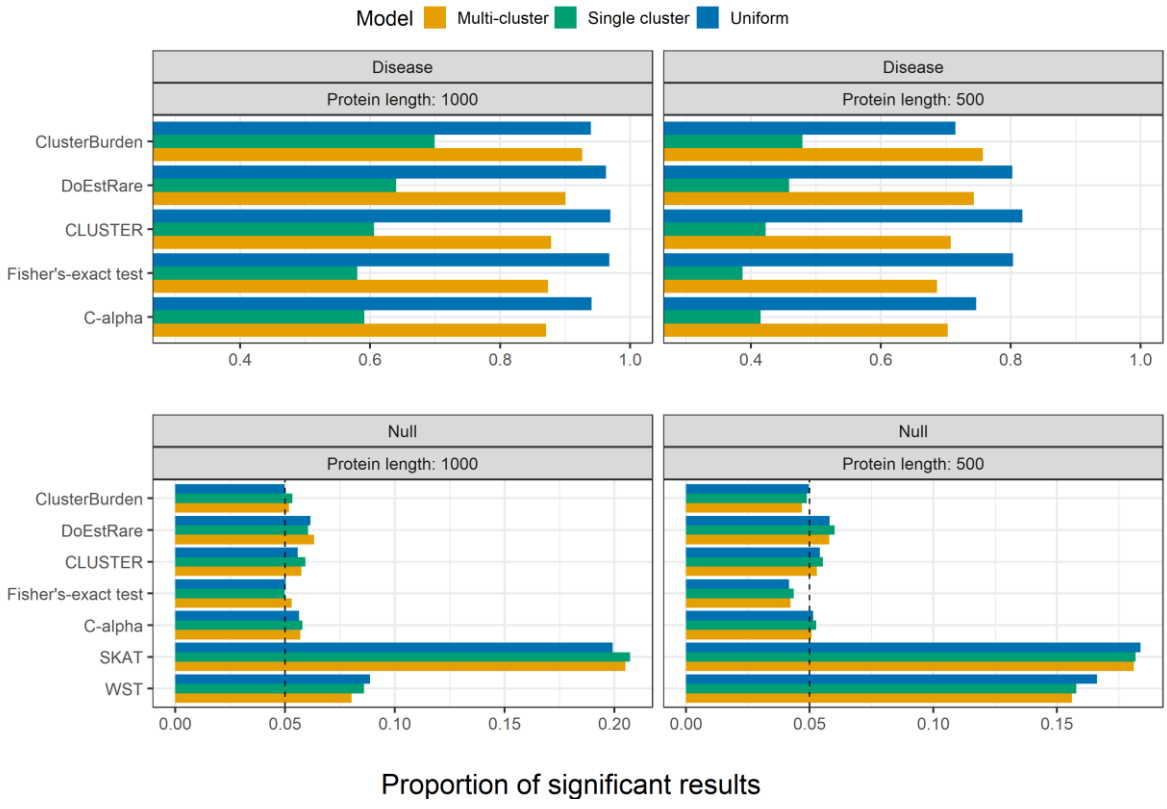

Figure 2.14: Type 1 error and power of rare-variant association tests to detect association of genetic variants in a disease-causing gene. Due to high type-1 errors, SKAT and WST were excluded from the disease model (upper panels).

In the null model, both SKAT and WST had false-positive rates substantially above the nominal threshold of 5% resulting in their exclusion from power comparisons. For the remaining methods, false-positives were kept at a reasonable level. Although p-values were calculated by permutation in DoEstRare and CLUSTER, the type 1 error did not exactly match the nominal significance threshold. This was because the number of permutations was limited to increase the speed of these time-consuming methods.

All methods increased in power when protein length was larger, this was due to the increased number of observed variants in larger proteins. In the simulation model, doubling the protein length was equivalent to doubling the sample size. After accounting for number of observations, there was no difference between the power for proteins with 500 or 1000 residues. There was higher power to detect association in uniformly burdened genes compared to the clustered models and more power in multi-cluster models than single cluster models. This was due to a correlation between power and the total number of potentially pathogenic residues in the protein. Therefore, in simulated, and real data in the same circumstances, more pathogenic variants are observed in these genes. Notably, the power differences relating to single and multiple clusters were similar independent of the method, including different strategies to detect clustering.

2.4 Application to gene-panel data

P-values for FE-test (burden), as well as the proposed methods *BIN-test, ClusterBurden* and LRT-dbNSFP, were calculated for all disease-gene pairs in the inherited heart disease gene-panel datasets using *gnomad_e2* controls. As stated in section 2.2, only missense variants or small inframe indels (<3 residues) occurring at less than 0.01% in any *gnomad_e2* subpopulation were considered for analysis. Table 2.8 displays all results alongside the gene-level ORs and flags for potential coverage or protein size confounding. The results from each test will be reported in

subsequent sections however this table serves as a reference for all numerical results from this analysis.

Table 2.8: P-values from all proposed methods across all disease-gene pairs in our inherited heart disease data. Bonferroni significant results corrected for the number of genes in each disease group are highlighted in gold.

Disease	Gene	Odds	Burden (FE-test)	Clustering (*BIN-test*)	ClusterBurden	LRT-dbNSFP	Flag: PL*	Flag: Cov1**	Flag: Cov2***
HCM	*ACTC1*	11.10	**1.40E-15**	0.591	**2.97E-14**	**0.000717**	Pass	Pass	Pass
HCM	*ACTN2*	1.17	0.234	0.552	0.394	0.298	Pass	Pass	Pass
HCM	*ANKRD1*	1.87	*0.0206*	0.331	*0.0407*	0.217	Pass	Pass	Pass
HCM	*BAG3*	1.58	*0.0184*	0.277	*0.0321*	0.558	Pass	Pass	Pass
HCM	*CRYAB*	1.63	0.152	0.989	4.35E-01	9.22E-01	Pass	Pass	Pass
HCM	*CSRP3*	3.11	**1.07E-05**	0.584	**8.09E-05**	*0.0138*	Pass	Pass	Pass
HCM	*DES*	1.79	*0.0207*	0.953	0.0972	5.31E-01	Pass	Pass	High
HCM	*DMD*	1.58	**8.60E-05**	0.751	**0.000687**	*0.00648*	Vhigh	Pass	Pass
HCM	*DSC2*	1.03	0.458	0.903	0.779	0.975	Pass	Pass	Pass
HCM	*DSG2*	1.35	*0.0441*	0.705	0.139	*0.0445*	Mod.	Pass	Pass
HCM	*DSP*	1.14	0.13	0.17	1.06E-01	2.34E-01	High	Pass	Pass
HCM	*FHL1*	6.31	**1.84E-14**	*0.0265*	**1.77E-14**	*0.00177*	Pass	Pass	Pass
HCM	*FHL2*	1.79	*0.0329*	7.90E-01	1.21E-01	9.62E-02	Pass	Pass	Pass
HCM	*FLNC*	1.87	**8.06E-10**	**7.14E-06**	**1.95E-13**	6.00E-02	High	Pass	Pass
HCM	*GLA*	5.16	**5.56E-08**	*0.0461*	**5.32E-08**	**0.000155**	Pass	Pass	Pass
HCM	*JUP*	1.50	*0.0177*	0.816	0.0757	0.63	Pass	Pass	Pass
HCM	*LAMP2*	2.71	**0.00137**	0.99	*0.0103*	**0.00104**	Pass	Pass	Pass
HCM	*LMNA*	1.25	0.191	8.87E-01	4.71E-01	2.45E-01	Pass	Pass	Pass
HCM	*MYBPC3*	10.40	**1.55E-203**	**4.80E-36**	**4.09E-236**	**4.52E-55**	Mod.	Pass	Pass
HCM	*MYH7*	11.40	**4.27E-292**	**5.50E-58**	**0.00E+00**	**6.90E-52**	Mod.	Pass	Pass
HCM	*MYL2*	4.05	**6.41E-08**	**3.36E-05**	**6.00E-11**	**8.81E-06**	Pass	Pass	Mod.
HCM	*MYL3*	3.33	**6.66E-06**	*0.0101*	**1.18E-06**	*0.012*	Pass	Pass	Pass
HCM	*PKP2*	1.44	*0.0271*	0.0741	*0.0145*	0.548	Pass	Pass	Pass
HCM	*PLN*	2.36	0.225	0.421	0.318	0.402	Pass	Pass	Pass
HCM	*PRKAG2*	1.65	*0.0188*	0.234	*0.0283*	*0.0381*	Pass	Pass	Pass
HCM	*RBM20*	1.40	*0.0327*	0.213	*0.0415*	*0.0418*	Mod.	Pass	Pass
HCM	*SCN5A*	1.40	*0.00479*	0.128	*0.00515*	0.465	High	Pass	Pass
HCM	*TAZ*	1.92	0.169	0.777	0.397	0.119	Pass	Pass	Pass
HCM	*TMEM43*	1.20	0.27	0.706	0.507	6.59E-01	Pass	Pass	Pass
HCM	*TNNC1*	3.73	*0.00433*	8.75E-02	*3.37E-03*	**1.76E-06**	Pass	Pass	Pass
HCM	*TNNI3*	15.00	**2.67E-54**	**1.05E-11**	**4.18E-63**	**1.31E-06**	Pass	Pass	Mod.
HCM	*TNNT2*	10.10	**8.13E-31**	**5.77E-07**	**3.97E-35**	**4.99E-05**	Pass	Pass	Pass
HCM	*TPM1*	13.90	**3.35E-21**	0.224	**3.72E-20**	**8.86E-05**	Pass	Pass	Pass
HCM	*TTR*	2.11	0.0996	0.788	0.278	0.0503	Pass	Pass	Pass
HCM	*VCL*	1.38	0.0664	0.801	0.209	0.321	Mod.	Pass	Pass
DCM	*ACTC1*	10.90	**0.000675**	0.107	**0.000761**	**0.000736**	Pass	Pass	Pass
DCM	*ACTN2*	1.28	0.356	1	0.724	0.0601	Pass	Pass	Pass
DCM	*BAG3*	1.67	0.223	0.786	0.48	0.0969	Pass	Pass	Pass
DCM	*CRYAB*	1.37	0.521	0.785	0.774	0.593	Pass	Pass	Pass
DCM	*DES*	4.36	*0.0015*	0.394	*0.00498*	0.067	Pass	Pass	High
DCM	*DMD*	1.46	0.142	0.903	0.392	0.348	Vhigh	Pass	Pass
DCM	*DSC2*	1.08	0.491	1	0.84	0.99	Pass	Pass	Pass
DCM	*DSG2*	1.20	0.389	0.98	0.749	7.40E-01	Mod.	Pass	Pass
DCM	*DSP*	2.24	**8.47E-05**	0.997	**0.000877**	0.0737	High	Pass	Pass
DCM	*FHL1*	5.78	*0.00219*	0.307	*0.00558*	**0.000541**	Pass	Pass	Pass
DCM	*FHL2*	0.75	0.737	0.405	0.66	0.315	Pass	Pass	Pass
DCM	*FLNC*	1.65	*0.046*	0.841	0.165	0.491	High	Pass	Pass
DCM	*GLA*	1.61	0.466	0.984	0.816	0.517	Pass	Pass	Pass
DCM	*JUP*	1.04	0.541	0.982	8.67E-01	5.86E-01	Pass	Pass	Pass

DCM	*LMNA*	4.95	**2.26E-06**	0.194	**6.85E-06**	*0.0028*	Pass	Pass	Pass
DCM	*MYBPC3*	2.70	**0.000352**	0.918	*2.92E-03*	6.08E-01	Mod.	Pass	Pass
DCM	*MYH7*	4.90	**7.91E-15**	*0.0341*	**9.95E-15**	**4.93E-05**	Mod.	Pass	Pass
DCM	*MYL2*	1.14	0.585	0.255	0.434	0.221	Pass	Pass	Pass
DCM	*MYL3*	0.93	0.661	0.865	0.891	0.216	Pass	Pass	Pass
DCM	*PKP2*	2.90	**0.00133**	0.907	*0.00932*	0.878	Pass	Pass	Pass
DCM	*PLN*	6.91	0.141	0.561	0.279	0.819	Pass	Pass	Pass
DCM	*PRKAG2*	0.84	0.688	0.988	9.42E-01	4.07E-01	Pass	Pass	Pass
DCM	*RBM20*	3.62	**6.59E-05**	0.0926	**7.94E-05**	*0.00282*	Mod.	Pass	Pass
DCM	*SCN5A*	1.51	0.0989	0.996	0.327	0.179	High	Pass	Pass
DCM	*TMEM43*	1.24	0.438	0.987	0.795	9.67E-01	Pass	Pass	Pass
DCM	*TNNI3*	7.85	**4.98E-05**	0.544	**3.12E-04**	1.89E-01	Pass	Pass	Mod.
DCM	*TNNT2*	18.10	**8.27E-15**	0.302	**8.64E-14**	*2.85E-03*	Pass	Pass	Pass
DCM	*TPM1*	13.60	**5.34E-05**	0.371	**0.000235**	*0.00924*	Pass	Pass	Pass
DCM	*TTR*	3.01	0.286	0.941	0.622	0.866	Pass	Pass	Pass
DCM	*VCL*	1.76	0.162	0.989	4.54E-01	9.95E-01	Mod.	Pass	Pass
ARVC	*ACTN2*	1.75	0.318	0.915	0.65	1	Pass	Pass	Pass
ARVC	*BAG3*	1.72	0.443	0.908	0.769	0.109	Pass	Pass	Pass
ARVC	*DES*	4.26	0.0824	0.851	0.257	0.52	Pass	Pass	High
ARVC	*DMD*	1.09	0.549	1	0.878	*0.00385*	Vhigh	Pass	Pass
ARVC	*DSC2*	5.30	**0.000492**	0.905	*0.00388*	0.414	Pass	Pass	Pass
ARVC	*DSG2*	4.16	*0.00397*	0.809	*0.0216*	0.306	Mod.	Pass	Pass
ARVC	*DSP*	2.83	*0.00263*	0.997	*0.0182*	0.835	High	Pass	Pass
ARVC	*FLNC*	2.29	0.0736	0.959	0.258	0.431	High	Pass	Pass
ARVC	*JUP*	0.89	0.678	0.739	0.847	0.104	Pass	Pass	Pass
ARVC	*LMNA*	2.40	0.206	0.648	0.402	0.957	Pass	Pass	Pass
ARVC	*MYBPC3*	1.62	0.287	0.96	0.631	0.574	Mod.	Pass	Pass
ARVC	*MYH7*	0.83	0.695	0.995	0.947	*0.00436*	Mod.	Pass	Pass
ARVC	*MYL3*	3.18	0.272	8.24E-01	5.59E-01	7.47E-01	Pass	Pass	Pass
ARVC	*PKP2*	16.70	**8.27E-17**	**2.76E-07**	**1.21E-21**	**7.15E-11**	Pass	Pass	Pass
ARVC	*RBM20*	2.11	0.248	0.94	0.572	0.266	Mod.	Pass	Pass
ARVC	*SCN5A*	1.19	0.463	0.997	0.818	0.102	High	Pass	Pass
ARVC	*TMEM43*	2.84	0.16	0.494	0.279	0.406	Pass	Pass	Pass
ARVC	*TPM1*	9.30	0.104	0.476	0.198	0.765	Pass	Pass	Pass
ARVC	*VCL*	1.45	0.501	0.857	0.792	0.333	Mod.	Pass	Pass
LQTS	*AKAP9*	0.90	0.668	1	0.937	0.365	Vhigh	Pass	Pass
LQTS	*ANK2*	1.31	0.196	0.986	0.511	*0.0244*	Vhigh	Pass	Pass
LQTS	*CACNA1C*	2.05	*0.0468*	0.19	0.0508	*0.0219*	High	Pass	Pass
LQTS	*CACNA2D1*	1.20	0.497	1	0.844	*0.00758*	Mod.	Pass	Pass
LQTS	*CACNB2*	1.33	0.396	0.799	0.68	*0.0363*	Pass	Pass	Pass
LQTS	*CALM2*	26.60	*0.0415*	0.441	0.0915	0.616	Pass	Pass	Pass
LQTS	*CASQ2*	2.40	0.134	0.785	0.342	0.408	Pass	Pass	Pass
LQTS	*DES*	0.95	0.652	0.851	0.882	0.778	Pass	Pass	High
LQTS	*DLG1*	0.92	0.64	0.992	0.923	0.319	Pass	Pass	Pass
LQTS	*DSC2*	1.33	0.356	1	0.723	0.175	Pass	Pass	Pass
LQTS	*DSG2*	0.30	0.962	0.998	0.999	0.708	Mod.	Pass	Pass
LQTS	*DSP*	1.01	0.533	0.922	0.841	0.221	High	Pass	Pass
LQTS	*GPD1L*	2.26	0.224	0.946	0.541	0.178	Pass	Pass	Pass
LQTS	*HCN4*	0.42	0.906	0.65	0.901	0.104	Mod.	Pass	Mod.
LQTS	*JUP*	1.19	0.462	0.781	0.729	0.143	Pass	Pass	Pass
LQTS	*KCNE1*	3.32	0.125	0.31	0.165	*0.0144*	Pass	Pass	Pass
LQTS	*KCNE3*	3.97	0.0933	9.33E-01	3.00E-01	1.51E-01	Pass	Pass	Pass
LQTS	*KCNH2*	22.90	**1.93E-46**	**4.27E-18**	**1.21E-61**	**2.37E-18**	Mod.	Pass	Mod.
LQTS	*KCNJ2*	7.47	**0.000713**	0.18	**0.00128**	**1.65E-06**	Pass	Pass	Pass
LQTS	*KCNJ5*	1.64	0.347	0.755	0.612	0.719	Pass	Pass	Pass
LQTS	*KCNJ8*	1.75	0.437	7.65E-01	7.01E-01	5.63E-01	Pass	Pass	Pass
LQTS	*KCNQ1*	45.10	**6.79E-94**	**2.02E-07**	**3.16E-98**	**3.11E-21**	Pass	Pass	Mod.
LQTS	*LMNA*	2.15	0.121	0.961	0.366	0.0703	Pass	Pass	Pass
LQTS	*PKP2*	1.85	0.14	0.902	0.388	4.64E-01	Pass	Pass	Pass
LQTS	*RYR2*	2.63	**1.38E-05**	0.832	**0.000142**	0.751	Vhigh	Pass	Pass
LQTS	*SCN10A*	1.03	0.516	1	0.858	0.925	Mod.	Pass	Pass

LQTS	*SCN1B*	3.37	0.122	0.787	0.321	0.528	Pass	Pass	Mod.
LQTS	*SCN2B*	1.04	0.617	0.507	0.676	0.747	Pass	Pass	Pass
LQTS	*SCN3B*	1.57	0.473	0.516	5.88E-01	5.37E-01	Pass	Pass	Pass
LQTS	*SCN5A*	3.65	**1.86E-06**	0.108	**3.31E-06**	*0.00573*	High	Pass	Pass
LQTS	*SLMAP*	0.94	0.626	0.993	0.917	0.635	Pass	Pass	Pass
LQTS	*TMEM43*	0.63	0.795	0.776	0.915	0.137	Pass	Pass	Pass
LQTS	*TRDN*	1.36	0.432	0.919	0.764	0.618	Pass	Pass	High
LQTS	*TRPM4*	1.43	0.249	0.999	0.595	0.7	Mod.	Pass	Mod.
BrS	*AKAP9*	1.06	0.498	1	0.845	0.816	Vhigh	Pass	Pass
BrS	*ANK2*	1.47	0.19	1	0.506	0.981	Vhigh	Pass	Pass
BrS	*CACNA1C*	2.53	0.0523	0.998	0.206	0.315	High	Pass	Pass
BrS	*CACNA2D1*	3.56	0.0549	0.96	0.208	0.113	Mod.	Pass	Pass
BrS	*CACNB2*	1.74	0.321	0.666	0.544	0.155	Pass	Pass	Pass
BrS	*DES*	3.75	0.102	0.791	0.284	0.941	Pass	Pass	High
BrS	*DLG1*	0.90	0.672	1	0.939	0.114	Pass	Pass	Pass
BrS	*DSC2*	0.65	0.785	0.953	0.965	0.207	Pass	Pass	Pass
BrS	*DSG2*	2.41	0.0891	1	0.305	0.298	Mod.	Pass	Pass
BrS	*DSP*	1.78	0.0898	1	0.306	0.451	High	Pass	Pass
BrS	*HCN4*	1.67	0.339	0.54	0.494	0.297	Mod.	Pass	Mod.
BrS	*JUP*	3.95	*0.0101*	0.92	0.0528	0.448	Pass	Pass	Pass
BrS	*KCNH2*	0.85	0.694	0.763	0.866	0.187	Mod.	Pass	Mod.
BrS	*KCNJ2*	2.92	0.292	0.114	0.147	0.524	Pass	Pass	Pass
BrS	*KCNJ8*	6.92	*0.0355*	0.851	0.136	0.334	Pass	Pass	Pass
BrS	*KCNQ1*	2.93	0.0861	1	0.297	0.465	Pass	Pass	Pass
BrS	*LMNA*	1.05	0.615	0.928	0.891	0.263	Pass	Pass	Pass
BrS	*PKP2*	1.45	0.403	0.86	0.714	*0.0154*	Pass	Pass	Pass
BrS	*RYR2*	1.12	0.453	1	0.812	0.264	Vhigh	Pass	Pass
BrS	*SCN10A*	0.87	0.671	1	0.939	0.398	Mod.	Pass	Pass
BrS	*SCN5A*	11.90	**4.58E-22**	0.0933	**2.24E-21**	**3.19E-09**	High	Pass	Pass
BrS	*SLMAP*	0.93	0.661	0.964	0.924	0.116	Pass	Pass	Pass
BrS	*SNTA1*	1.53	0.481	0.702	0.704	1	Pass	Pass	Pass
BrS	*TMEM43*	1.24	0.554	0.369	0.529	0.474	Pass	Pass	Pass
BrS	*TRDN*	2.69	0.173	0.693	0.374	0.176	Pass	Pass	High
BrS	*TRPM4*	0.47	0.883	0.995	0.992	0.271	Mod.	Pass	Mod.

*(Flag: PL) Very high>3000, High>2000, Moderate>1000, Pass < 1000 residues.
**(Flag: Cov1) Very high>50%, High>30%, Moderate>10%, Pass < 10% of *BIN-test* bins with lower than 80% coverage.
***(Flag: Cov2) Very high<80%, High<70%, Moderate<60%, Pass > 80% mean coverage.

2.4.1 Burden signals

The burden of rare-missense variants across all genes was assessed by FE-test (Figure 2.15). In HCM, fourteen genes passed *b-significance* (after correcting for 35 genes). This was higher than any other disease group considered and undoubtedly reflects the larger sample size and associated increase in power. Eight of the eleven most significant genes; *MYH7* ($p < 4.3 \times 10^{-292}$), *MYBPC3* ($p < 1.6 \times 10^{-203}$), *TNNI3* ($p < 2.7 \times 10^{-54}$), *TNNT2* (8.1×10^{-31}), *TPM1* (3.4×10^{-21}), *ACTC1* (1.4×10^{-15}), *MYL2* (6.4×10^{-8}) and *MYL3* (6.7×10^{-6}), are all involved in the cardiac sarcomere and have strong prior evidence

of association with HCM (Morimoto 2007). The two top genes, *MYH7* and *MYBPC3*, are by far the most strongly associated genes for HCM, an expected outcome and consistent with the prevalence of disease-causing variants in these genes. ORs for these genes both exceed 10, indicating that there are at least 10 times as many rare-missense variants observed in these genes than in *gnomad_e2*.

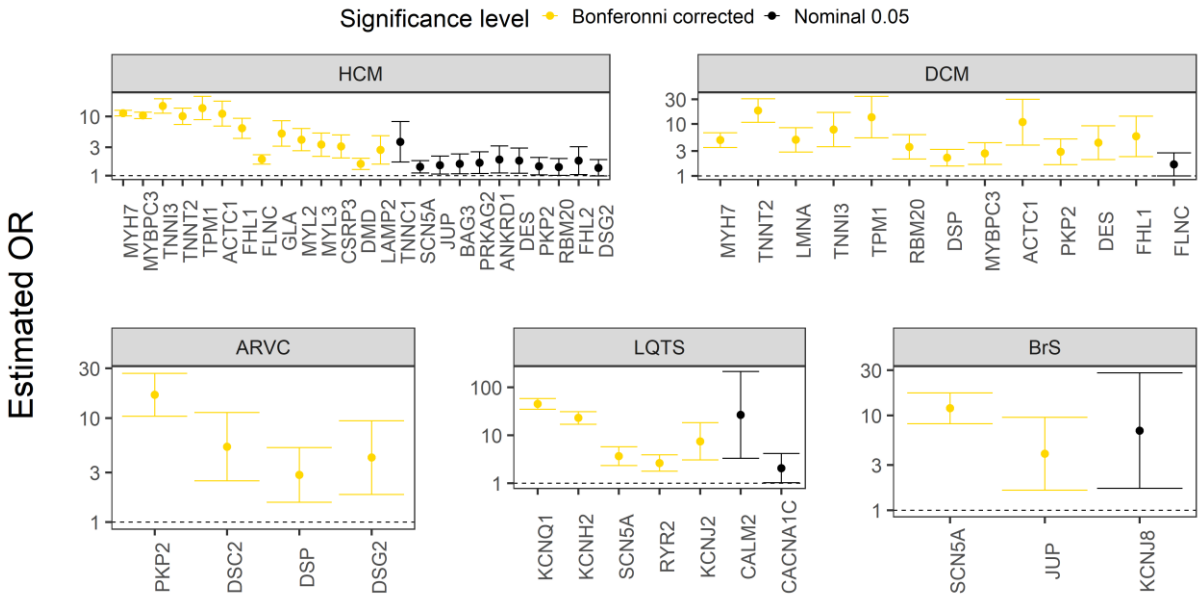

Figure 2.15: OR and CI estimates for all significantly burdened genes. Bonferroni significant genes (p<0.00143) are highlighted in yellow, while the remaining genes in black are only nominally significant.

In DCM, eight genes passed multiple testing correction, and a further two were *n-significant*. Once again, *MYH7* ($p < 7.9 \times 10^{-15}$) is at the top of the list alongside four more sarcomeric proteins; *TNNT2* (8.3×10^{-15}), *TNNI3* (5×10^{-5}), *TPM1* (5.3×10^{-5}) and *ACTC1* ($p < 0.00068$), all with prior evidence of association with DCM. Three genes; *PKP2* (8.3×10^{-17}), *DSC2* ($p < 0.0005$) and *DSP* ($p < 0.0027$), pass the multiple testing significance threshold for ARVC, while a further one gene, *DSG2* (p <0.004), was *n-significant*. All four genes are known genetic causes of ARVC.

In LQTS there were five *b-significant* genes; *KCNQ1* ($p < 6.8 \times 10^{-94}$), *KCNH2* ($p < 1.9 \times 10^{-46}$), *SCN5A* ($p < 1.9 \times 10^{-6}$), *RYR2* ($p < 1.4 \times 10^{-5}$), *KCNJ2* ($p < 0.00071$) and three *n-significant* genes; *CALM2* ($p < 0.042$) and *CACNA1C* (0.047). In a study examining evidence for genes reported to cause LQTS; *KCNQ1, KCNH2, SCN5A* and *CALM2* were assessed as having definitive evidence of association, while KCNJ2 had limited evidence of association (Alder *et al.* 2020). In BrS the top and only *b-significant* gene was *SCN5A* ($p < 4.6 \times 10^{-22}$), consistent with previous findings that suggest that this gene is the only definitive cause of the disease and accounts for approximately 30% of patients (Brugada *et al.* 2018). Nominally significant *KCNJ8* ($p < 0.036$) has some previous evidence of association (Sarquella-Brugada *et al.* 2015), but *JUP* ($p < 0.01$) has no published evidence of association.

Overall using burden testing largely distinguished the known and expected genes from a background of unrelated genes in each heterogeneous gene panel (More discussion of the significantly associated genes here is in Appendix 3). To follow up, as a negative-control, genes were assessed for an excess of synonymous variants. As there are no known examples of synonymous variants playing a role in cardiomyopathy or arrhythmia, with the exception of possible splice-impacting synonymous variants in *MYBPC3* (Rangaraju *et al.* 2017), results here were expected to be null. The only gene with a *b-significant* association was *DMD* in HCM. This provided further evidence that its significance in the missense-variant analysis was spurious. Notably, the *PKP2* gene had *n-significant* synonymous burden results for both HCM, DCM and ARVC. Similarly, *LAMP2* ranked high for both HCM and ARVC. The lambda inflation statistic for the synonymous burden p-values in all groups was 4.0, 3.6, 1.3, 2.2, and 2.9 for HCM, DCM, ARVC, LQTS and BrS respectively. Although small sample sizes in terms of the number of p-values, these are much higher than expected. This likely reflects uncontrolled confounding in the data, possibly due to the use of the heterogeneous population controls unmatched to the inherited heart disease datasets.

2.4.2 *Clustering signals*

The proposed *BIN-test* was used to detect significant missense variant clustering. *B-significant* clustering signals (corrected for 35 genes) were observed in; HCM, ARVC and LQTS. In HCM, *MYH7* ($p < 5.5 \times 10^{-58}$), *MYBPC3* ($p < 4.8 \times 10^{-36}$), *TNNI3* ($p < 1.1 \times 10^{-11}$), *TNNT2* (5.8×10^{-7}), *FLNC* ($p < 7.1 \times 10^{-6}$) and *MYL2* ($p < 3.4 \times 10^{-5}$) passed multiple testing correction (Figure 2.16). All are well-established causes of HCM (Gómez *et al.* 2017; Ingles *et al.* 2019). The most significant position signal was observed in *MYH7*, a long-recognised result (Watkins *et al.* 1992), with most enrichment in the myosin-head motor domain. The next strongest signal, driven by four potential clusters, was observed in *MYBPC3*. The first cluster overlaid the C1 myosin S2 and actin-binding domain, and the last cluster was in the C10 TTN-binding domain (Flashman *et al.* 2004).

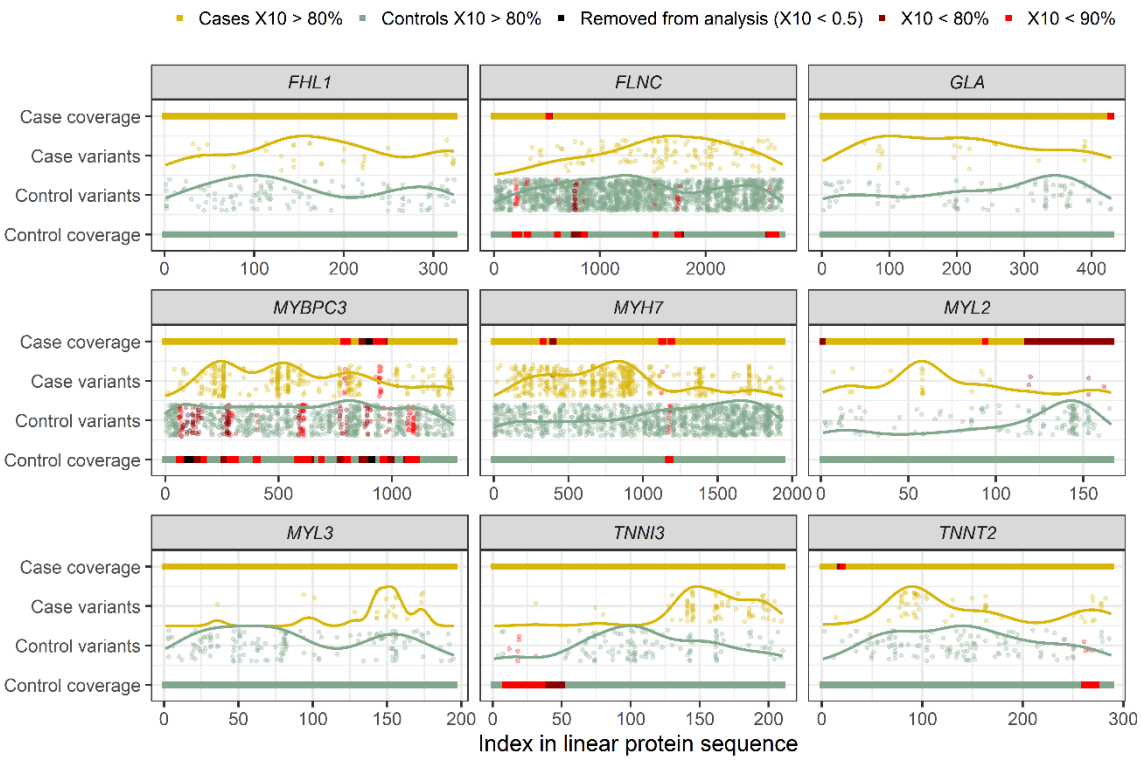

Figure 2.16: Distribution of variants in nine HCM genes with significant clustering. Empirical variant positions (points) are overlaid with a smoothing kernel to indicate relative density. Coverage for each residue, averaged for the three bases of the codon, for HCM cases and *gnomad_e2* are also displayed.

In *TNNI3*, which encodes cardiac troponin I, 91% of 34 case variants mapped to a C-terminal cluster spanning residues 128-209. This accords with previous studies (Mogensen *et al.* 2004). *FLNC* is the only gene present in this list that is not a thin or thick filament cardiac sarcomere protein. An analysis of missense *FLNC* variants in HCM found they clustered in the C-terminus as observed here (Ader *et al.* 2019). Eighty-nine percent of 27 case variants in *TNNT2*, which encodes cardiac troponin T, map to two clusters between residues 67-179 and residues 250-282. Both clusters are previously reported (Palm *et al.* 2001; Moore *et al.* 2014) and the first overlaps with a tropomyosin-binding region (Pearlstone *et al.* 1977). In *MYL2*, 50% of 30 case variants cluster between residues 25 and 100, whereas control variants tended to cluster towards the C-terminus.

Of *n-significant* genes, *MYL3* ($p < 0.01$) is a known causative gene, and approximately 80% of 14 case variants clustered between residues 125 and 175, whereas control variants were more uniformly distributed. *FHL1*, also a known HCM-gene with *b-significant* burden, was also amongst the *n-significant* clustering signals, with a potential increased density of variants in the centre of the protein, relative to *gnomad_e2* controls. The phenocopy gene *GLA* ($p < 0.046$), had an increase of variants in the N-terminal region of the protein whereas this region was clearly constrained in controls.

In DCM, the only significant clustering was found in *MYH7* ($p < 0.034$). Variant clustering in DCM has been previously reported for; *LMNA, MYH7* and *RBM20* (Horvat *et al.* 2018) and though these genes fall at the top of the list, *LMNA* and *RBM20* do not even reach *n-significance*. Previous reports suggest that variant clustering is not as strong in *MYH7* for DCM cohorts with variants more uniformly dispersed (Bollen and van der Velden 2017), this is observed here also (Figure 2.18). There appears to be a substantial cluster of variants in the C-terminus of the ARVC gene *PKP2* ($p < 2.8 \times 10^{-7}$). However, this is strongly driven by the variant c.2197C>G which is observed in 11 cases (and four controls). When this variant is removed in a sensitivity analysis, the p-value is no longer significant ($p < 0.34$), casting doubt on the validity of this signal.

There are two convincing *b-significant* clustering signals in the LQTS genes (Figure 2.17); *KCNH2* ($p < 4.3 \times 10^{-18}$) and *KCNQ1* ($p < 2 \times 10^{-7}$). In *KCNH2*, a central cluster is observed and in *KCNQ1*, the variants are enriched in the N-terminus of the protein. Missense variants in *KCNQ1* have been reported to produce faulty proteins that cause disease in a dominant-negative fashion. Furthermore, variants located in cytoplasmic loops between the transmembrane domains, have been demonstrated to cause higher incidence of cardiac events in carriers (Ruwald *et al.* 2016). In an analysis of *KCNQ1* clustered variants identified by Moss *et al.* (2007), there was an almost identical overlap with the variant clusters observed here. Variants tend to cluster in the intracellular domains the two transmembrane domains S5 and S6.

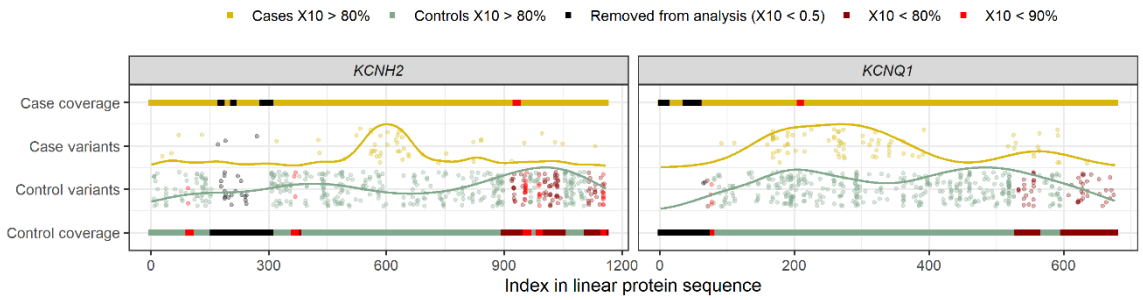

Figure 2.17: Distribution of rare-missense variants and coverage in *KCNH2* and *KCNQ1*. Empirical variant positions (points) are overlaid with a smoothing kernel to indicate relative density. Coverage for each residue, averaged for the three bases of the codon, for HCM cases and *gnomad_e2* are also displayed.

2.4.3 ClusterBurden signals

When using *ClusterBurden*, a total of thirteen genes were *b-significant* in HCM, all of which had a *b-significant* burden signal. Only one gene with a *b-significant* burden signal, *LAMP2*, does not pass multiple testing correction in *ClusterBurden* and is now only *n-significant*. Overall, *ClusterBurden* provided a theoretical increase in power for nine HCM genes, including the six most firmly established disease-causing sarcomeric proteins; *MYH7*, *MYBPC3*, *MYL2*, *MYL3*, *TNNI3*, *TNNT2*. In DCM, two genes are now n-*significant* (*ClusterBurden)* rather than *b-significant* (burden); *MYBPC3*

and *PKP2*. Neither gene is definitively associated with DCM. In ARVC, while *DSC2* was *b-significant* in burden testing ($p < 0.0049$), it is only *n-significant* in *ClusterBurden* ($p < 0.0039$). Whereas the p-value for *PKP2* becomes stronger by several orders of magnitude. In LQTS, all *b-significant* burden signals are *b-significant ClusterBurden* signals. The same is true for BrS.

Post-hoc power calculations were performed for the subset of HCM genes with a significant clustering signal. Both FE-test and *ClusterBurden* were treated as likelihood ratio tests with non-centrality parameters scaling with sample size (Liu 1997, Agresti 1996). This demonstrated enhanced empirical power to detect truly disease-causing genes. This was most noticeable in *FLNC, MYL2* and *MYL3* where the increased power from integrating clustering information was equivalent to increasing the sample size by; 66%, 53% and 52% respectively. It should be noted however, that retrospective power calculations are biased by winner's curse and may not directly inform prospective power.

2.4.4 LRT-dbNSFP signals

For HCM, The LRT-dbNSFP test detected ten genes where the subset of dbNSFP predictors were *b-significantly* better at predicting whether a variant was derived from the cases or *gnomad_e2*, than a model based on carrier counts in cases and controls (intercept-only null model). All genes here also had a *b-significant* burden signal, with the exception of *TNNC1,* which has moderate prior evidence of association with HCM (Ingles *et al.* 2019), and also had an *n-significant* burden signal. For DCM, *b-significant* signals were observed in established causative genes *MYH7* and *ACTC1*, but also in *FHL1* which has limited prior evidence of a true association and was only *n-significant* in the burden test. *N-significant* genes here; *LMNA*, *RBM20, TNNT2* and *TPM1* are all likely to be truly associated with DCM. For ARVC, only one gene, *PKP2*, was significant after multiple testing correction. A further two genes *MYH7* and *DMD* were nominally significant, however, neither of

these genes has strong prior evidence of any association with ARVC. Notably both *MYH7* and *DMD* are large proteins, however it is unclear how this would bias the results of the LRT-dbNSFP test.

For LQTS, the top two genes were the expected *KCNQ1* and *KCNH2* with *KCNJ2* also amongst the *b-significant* results. *SCN5A, CACNA2D1*, *KCNE1, CACNA1C* and *ANK2* were also amongst the *n-significant*. Although prior associations with epilepsy and intellectual disability, *CACNA2D1* is not known to be associated with LQTS (Vergult *et al.* 2014). Historically, a small number of variants in *KCNE1,* sequenced in LQTS patients, have been demonstrated to alter the biophysical properties of the channel and reduced magnitude of QT interval increase. A recent reappraisal of variants in this gene suggest pathogenic variants in this gene are uncommon and cause a weaker form of the disease (Garmany *et al.* 2020). In BrS, only *SCN5A* was Bonferroni significant and only *PKP2* was nominally significant. *PKP2* had a very weak burden signal and although some evidence exists linking it to BrS there is a lack of evidence definitively connecting variants in this gene with the BrS phenotype (Campuzano *et al.* 2016).

2.5 Discussion

In this chapter, multiple gene-association methods were developed to detect disease-associated genes based on rare-missense variants. The *BIN-test* and *ClusterBurden* were assessed in simulated data. Both suitably preserved false-positives at the nominal threshold and fared favourably in terms of power against published alternatives. As expected, *ClusterBurden,* which integrated both the burden and positional signals, has increased power for clustered genes but reduced power for non-clustered genes, relative to alternate RVATs. When applied to inherited heart disease genes, both the *BIN-test* and LRT-dbNSFP discovered signals in many firmly established genes, supporting their usefulness as complementary signals to gene-burden.

2.5.1 Performance of the BIN-test to assess rare-missense variant clustering

BIN-test was vastly more powerful in simulated data compared to other two-sample goodness-of-fit tests (Subsection 2.4.3; Figure 2.13). Alternative two-sample methods collapse distributions to summary statistics, potentially resulting in loss of information for non-monotonic distributions. Conversely, *BIN-test* collapses on each region of the protein separately, limiting data loss. In this setting, power is only reduced when true clusters are substantially smaller than the bins, or occur on the boundaries between two bins. However, avoiding these biases by permuting over bin sizes and locations, increased complexity and gave limited power gain (Figure 2.3). Therefore, the *BIN-test* with Mann-Wald heuristic represented a parsimonious strategy to select clustered genes.

2.5.2 Performance of ClusterBurden for rare-variant association

For simulated genes where variants clustered in regions of increased pathogenicity, *ClusterBurden* was more powerful than previously published RVATs (Subsection 2.4.4; Figure 2.14). Due to the increased degrees of freedom when combining p-values from two tests, it had less power than a simple burden test for un-clustered genes. This may occur for both uniformly pathogenic genes and genes with complicated or subtle clustering that is difficult to detect with limited data (either explanation may apply to *ACTC1* and *TPM1* in HCM). However, in situations where burden testing has not yielded any novel findings in whole-exome screens, application of the alternative position-dependent methods CLUSTER or DoEstRare is unlikely to uncover any additional loci. This is because these tests give very low weight to the clustering signal making them essentially equivalent to a FE-test. In the same scenario, *ClusterBurden* may detect genes with both a modest burden and clustering signal that did not reach the significance threshold based on the burden p-value alone. In this sense, it is a viable alternative follow-up strategy after burden testing. *FLNC* is an exemplar for this. It is a low-penetrance gene with a modest excess of rare-missense variants in cases (OR=1.87) but with strong clustering ($p<7.1 \times 10^{-6}$). It is feasible that for this gene, and others like it, this signal may be the difference in making a discovery.

In the null model, SKAT and WST had unacceptably high false-positive rates (Subsection 2.4.4; Figure 2.14). Although broadly used for both monogenic and complex diseases, the methods were designed to model both rare (MAF < 0.1%) and uncommon (MAF < 1%) variants. In the case of simple Mendelian diseases where only very rare variants are of interest, all common variants are excluded from the analysis. For WST, allele frequency is used to weight a variants contribution to the test statistic and overrepresentation of singleton variants, leads to poor calibration of this test statistic. For SKAT, rare and uncommon variants present in both cohorts calibrate the variance-component of the test. A variant set of mostly singletons and private variants may result in poor estimates of variance leading to high false-positives.

The weighting of the clustering signal seemed to be the primary factor influencing power of the position-informed RVATs in simulated data (Subsection 2.4.4; Figure 2.14). Interestingly, for burden testing of non-clustered genes, the simple and rapid FE-test had almost identical power to the best method CLUSTER. This demonstrates that at least for these simple simulations, and a disease-model of rare-deleterious causal variants, there is no benefit to the increased statistical complexity of these methods. C-alpha, which considers both protective and deleterious variants, had weaker power than the remaining methods (which were all one-sided), as protective variants were not modelled in the simulated data.

Estimates of power in simulated data were lower for clustered models compared to uniform models, lower for single clusters compared to multiple clusters and lower for smaller proteins (Subsection 2.4.4; Figure 2.14). Underlying all these observations is the covariance of power and the total number of potentially pathogenic residues. More pathogenic residues increase the likelihood of a disease-causing mutation to occur in the population. An untested hypothesis logically following from this, is that smaller genes with fewer residues are likely to harbour fewer pathogenic variants, have lower power to be associated, may explain less heterogeneity of traits and may be associated with rarer phenotypes. This may also be the case for the pathogenic

mechanisms; gain-of-function (GoF) or dominant-negative (DN) which are associated with specific functional protein regions resulting in fewer pathogenic residues. Conversely, missense-variants that exhibit their effect via HI may occur almost anywhere in the protein increasing the number of pathogenic residues. Consequently, power to detect these disease mechanisms (GoF, DN), may be reduced relative to HI. As these genes are also genes likely to be clustered (Lelieveld *et al.* 2017), these may represent a subset of pathogenic genes much more likely to be detected using *ClusterBurden*.

Although memory inefficient and time-consuming, the UE-test was potentially more powerful than the FE-test in the small sample simulation considered (Subsection 2.3.7; Figure 2.6). In future implementations, to maximise power, a one-sided, singly-conditioned, memory-efficient implementation of this approach could be developed to boost power in *ClusterBurden*. Alternatively, many other approaches could be used for the burden component of *ClusterBurden,* such as logistic regression or almost any other fully-fledged RVAT, uncorrelated with positional information, including variance-component tests. Thus, *ClusterBurden* represents a flexible framework for rare-variant association.

A potential drawback of the clustering approach identified in this chapter is the reduced power in non-clustered genes and the relatively large sample sizes required to find sufficient power to detect positional signals. A possible way to counter these problems, would be to only use *ClusterBurden* on genes likely to be associated with missense-variant clustering and weighting particular regions of proteins that are more likely to habour clusters. Both these could be achieved by determining prior probabilities for clustered genes and cluster locations, a point that will be discussed further in chapter 5.

2.5.3 Controlling for Coverage

Coverage was adjusted for in both the burden and *BIN-test* (Subsection 2.3.5). For burden testing, the approach is intuitive as the denominator in carrier frequency estimates is adjusted based on the effective allele number for the gene. For the *BIN-test*, a novel approach was adopted whereby counts were imputed relative to the inverse of the coverage for each bin. One potential limitation of both approaches is the assumption that all samples with above 10X coverage have 100% variant calls. A limitation specific to the *BIN-test* coverage control strategy, is the inaccurate adjustment of low count bins, especially in the extreme case of zero count cells, whereby a continuity correction is necessary. However, given the imperfect data, this coverage strategy is an intuitive way to limit data loss and make inference on as much data as possible.

2.5.4 Carrier-level analysis and founder variants

Generally, in burden testing, the number of carriers of a rare-allele are counted in each cohort (Morgenthaler and Thilly 2007). Although widely accepted, it may break the assumption of statistical independence between observations. This has potential implications for both the burden and *BIN-test*. While most variants in the datasets are singletons, a few high allele count variants exist. These may be founder variants, in which case they are definitely not independent observations. However, as founder variants are included in both cases and controls, this may actually lead to better estimates of the pathogenicity and penetrance of these variants, increasing resolution. One potential problem is that these founder variants may be given a lot of weight in the analysis. This is especially problematic for the *BIN-test* where a cluster may be detected based on a single variant site. This necessitates caution in the interpretation of results.

2.5.5 Analysis of inherited heart disease data

Almost all previous analyses of large hypertrophic cardiomyopathy cohorts (Walsh *et al.* 2017, Ingles 2019), use data aggregated from clinical reports. Here, the first analysis of aggregated NGS

data is reported. There are a number of potential advantages to this experimental design. Firstly, *all* variants sequenced in a patient are recorded in the NGS data, whereas in clinical reports, variants predicted as likely benign may be omitted, especially after finding another variant with strong evidence of pathogenicity. Additionally, although low coverage regions are re-sequenced by Sanger sequencing for clinical reports, there is no associated coverage files in aggregated clinical report datasets to ascertain, and control for, regions of uneven coverage. Conversely, the data used here are derived from the BAM files which allow coverage files to be generated. Finally, in sequencing reports, it is not guaranteed that the bioinformatics pipeline for processing the sequence data from each report are consistent. Whereas, when processing thousands of VCFs in one bioinformatics pipeline, it is easier to detect and remove poor quality variants or samples. Hypothetically, these factors improve estimates of the true background variation within the cohort.

2.5.6 Burden testing results

Burden testing of the inherited heart disease cohorts yielded strongly significant signals in all firmly established genes for each cohort (Subsection 2.5.1: Figure 2.15). For some genes, which underlie a high proportion of the genetic heterogeneity of the diseases, effect sizes were large, leading to massively significant p-values. However, many *n-significant* results were observed in genes with limited prior evidence of association. It is unclear how to interpret these results as many alternative hypotheses compete to explain them, including a true association and a variety of confounders or biases. Corroborating evidence pointing towards confoundment are the high estimates of p-value inflation (lambda) for synonymous variants in each cohort. As there is limited evidence of pathogenic synonymous variants in inherited heart diseases, the signals driving these inflation statistics are almost certainly spurious. A possible cause of this, is the use of heterogeneous population controls processed using a different pipeline to the inherited heart disease data. For more accurate estimates of rare-missense variant ORs, this potentially global excess of variant calls in the heart disease data would need to be calibrated to prevent inflation of these estimates.

Alternatively, a better matched cohort would be required, however, this is not readily available in the scenario of aggregated clinical sequence data.

2.5.7 Cluster detection results

An analysis of variant positions across five inherited heart diseases reveals rare-missense variants in numerous established disease-genes (Subsection 2.5.2; Figure 2.16). It seems intuitive that in most disease-causing proteins, the pathogenicity of missense variants would be dependent on their position. For the eight well-established sarcomeric genes in HCM, clear clustering is evident in six. For the remaining two; *TPM1* and *ACTC1*, although significant clustering was not detected with the proposed method (*BIN-test*), there does appear to be a population control depletion in the centres of these proteins, as well as an enrichment in our cases. Perhaps with more data or different methods, these signals would also be significant. In DCM, fewer genes were identified as significantly clustered. This may result from the lower sample size and indicates the need for large sample sizes to detect this signal. In LQTS, strong clustering was successfully detected in the two major gene; *KCNQ1* and *KCNH2*.

ClusterBurden rediscovered *b-significant* signals in all firmly associated genes across the datasets. Post-hoc power calculations indicated a theoretical increase in power for clustered genes, which included many firmly associated genes in this analysis. As expected, for non-clustered genes, the resulting p-value was lower. While a necessary consequence of testing the joint association of both signals, the application of this test may be improved by prior identification of genes likely to be clustered. This will be discussed in more detail in chapter 4.

Saturation mutagenesis offers a source of orthogonal information to validate the pathogenicity of variants. Here, experiments induce substitutions at multiple amino acid positions and observe the functional consequence in a relevant cell line. Of course, while potentially providing useful information, the time and cost of these experiments can be prohibitive for the assessment of all

substitutions with hypothetical disease-relevance. *ClusterBurden*, and clustering analysis in general, may aid implementation of targeted mutagenesis by highlighting specific regions of increased putative importance.

2.5.8 LRT-dbNSFP results

The LRT-dbNSFP test identified strong signals in many true pathogenic genes indicating its potential utility in rare-variant association (Subsection 2.5.4). Future iterations of the approach should focus on the identification of which scores and how many to use in the disease-model. As the model proposed here contains multiple predictors, its complexity makes it non-optimal for genes with few observed variants. An alternative and simpler approach to using pathogenic prediction scores would be to filter variants based on these scores. Which scores to filter on and which threshold to use to define a pathogenic variant still remain as non-trivial barriers to implementing this approach. Another possible enhancement to this approach would be to balance the number of predictors in the disease-model with the number of observations. Finally, in order to be confident in its application, circularity must be avoided by ensuring no variant in the observed data is in the training data used by these variant prediction scores. This could lead to inflated false-positives but is unlikely to be a problem for recent sequenced datasets.

Chapter 3

Modelling mutational hotspots and predicting pathogenicity

3.1 Developing mutational *hotspot models*

3.1.1 Conceptual background

As a relatively common disease (1 in 500 prevalence), HCM is an excellent candidate Mendelian disease for gene-specific data-driven variant interpretation. The large patient cohorts available provide researchers with empirical data to refine and improve current guidelines. These data-driven approaches can be used to estimate the best methods for applying different criteria. PM1 is a criterion within the ACMG Variant Interpretation Guidelines (Richards *et al.* 2015). It allocates moderate evidence of pathogenicity to variants observed in a 'mutational hot spot and/or critical

and well-established functional domain'. Currently, in HCM, this criterion is only applied consistently for *MYH7*. If an observed variant falls in the motor-head region of the protein, it is given moderate evidence in accordance with PM1 (Kelly *et al.* 2017). For other genes, it is unofficially noted by clinical scientists, that some regions have a higher incidence of pathogenic variants. However, the informal nature of this approach results in its inconsistent application, without statistical justification. With access to large training datasets, the application of this criterion may be improved using a robust quantitative framework. After years of clinical sequencing, enough data is now available to apply a data-driven approach. Here, a flexible modelling procedure is explored to estimate these hotspots (henceforth *hotspot models*) in HCM and four other inherited heart diseases.

In order to estimate hotspots, the location of missense variants in both affected cases *and* unaffected controls must be considered. The background (neutral) distribution of rare-variants within a protein are not necessarily expected to be uniformly distributed. Non-uniform mutation rates and purifying selection, unrelated to the disease under investigation, can both cause non-uniform variant densities. As a result of this, an increased density of variants in cases, a 1-dimensional cluster, is not sufficient to determine disease-association. Consequently, to estimate mutational hotspots, specific to the diseases under investigation, the distribution of case variants was compared against *gnomad_e2*. Here the controls represent the null distribution of variants in a region, automatically controlling for both mutation rate and non-relevant population-level constraint. This is important as a region with a dearth of observed variation may be driven by its association with any disease. For the uncontrolled scenario, this region of constraint could mistakenly give the impression of an adjacent region of enrichment. However, if not relevant to the disease under investigation, the constrained region should be present in both cases and controls avoiding a false-positive result.

When considering mutational hotspots, intra-gene differences in variant density is only one part of the picture. Firmly established disease-genes are themselves hotspots for pathogenic variation, in the context of the wider genome. This is apparent from the gene-level ORs, which may substantially exceed 1. For this reason, it is also essential to capture the gene-level burden alongside intra-gene patterns of variation. In a sense, the gene-level burden is equivalent to placing a prior probability on the pathogenicity of each gene. The combination of these two features; distribution of variant residue positions and gene-level burden, produces a model that can be interpreted as the regional burden across the length of the protein. As an added benefit, variants within a gene and between genes can be compared on the same scale.

3.1.2 Generalized-additive models

Hotspot models were generated using generalized-additive models (GAMs; Hastie and Tibshirani 1990). GAMs adaptively model linear and non-linear relationships of unknown complexity, between explanatory variables (e.g. burden, residue position) and the response variable (case-control status). In a generalized linear model (GLM), a single linear coefficient is estimated for each model term: $f(Y) = \beta_0 + \beta x_1 + \beta x_2 + \epsilon$. Sample predictions are the sum of feature values (x_1 to x_n) multiplied by their corresponding (estimated) model coefficients (β_1 to β_n), plus the intercept (β_0). This value is transformed using a link function (f) to generate the final prediction. For GAMs, the process is similar, but the scalar coefficients are replaced by smooth functions. These functions are specified by the linear sum of their underlying basis functions, each of which is described by a scalar coefficient. In this sense, when smooth functions are decomposed into their basis functions, both GAMs and GLMs can be represented as the additive sum of scalar coefficients. The primary difference between them (at least in the context used here), is that GAMs automatically estimate the smooth function complexity to infer the most parsimonious model. Smooth functions are estimated by restricted maximum likelihood, which penalises excessive model complexity (curve wiggliness). GAMs therefore facilitate a flexible model-selection strategy to infer the relationship

between residue position and disease status. Moreover, the additive nature of these models permits the easy inclusion of an arbitrary number of additional covariates.

To understand GAM in more detail, consider this simple example using the gene *FLNC*. The number of rare-missense variants carriers is 152 in our HCM cases and 2,503 in *gnomad_e2* controls. A GAM was created to predict case-control status based on the residue position of these variants. Coefficients for the smooth function and the intercept can be extracted from the model (Table 3.1). Each coefficient corresponds to one of nine basis functions, which are thin-plate regression splines in this example, and can be visualised (Figure 3.1A). Each basis function is then multiplied by their coefficients (Figure 3.1B), added together and exponentiated (logistic link function), to give the predicted OR for each position (Figure 3.1C).

Table 3.1: Basis function coefficients for a GAM residue position smooth function in the *FLNC* gene. The outcome variable in this GAM is case-control status where cases are HCM affected and controls are derived from GnomAD exomes.

Intercept	**s1**	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**	**s8**	**s9**
-2.953	0.137	-5.193	-0.673	-3.082	-0.284	2.790	-0.667	8.731	3.035

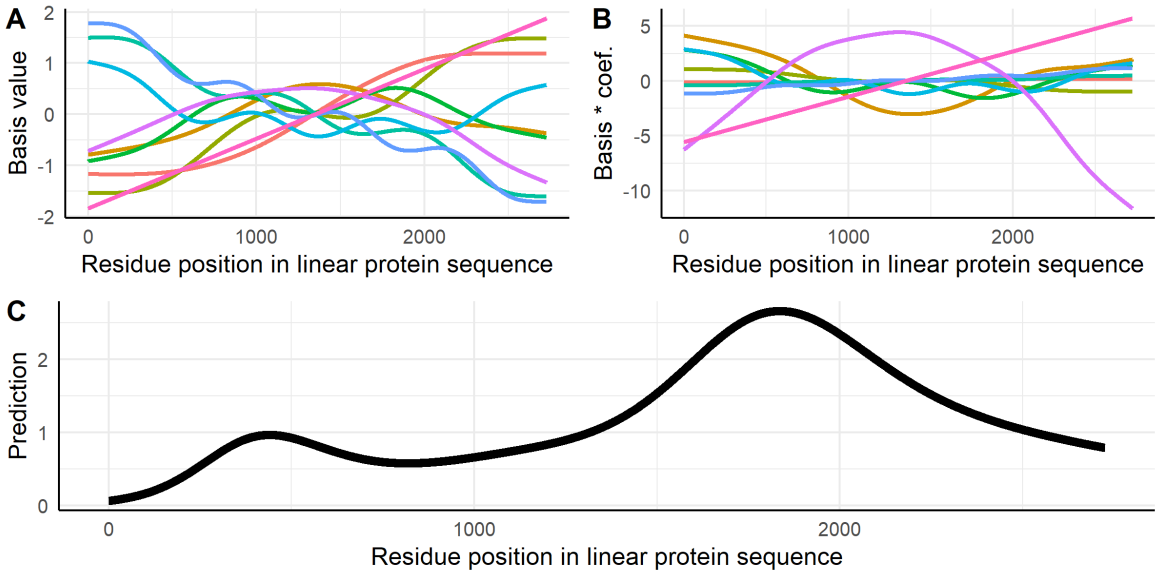

Figure 3.1: Basis functions, scaled basis functions and final predictions for a GAM of *FLNC.* A GAM was generated for the HCM-associated gene *FLNC*, modelling cohort status as a smooth function of missense-variant residue position. A) The nine basis functions for the smooth function. B) Basis functions multiplied by their corresponding model coefficient estimated by GAM. C) Final model predictions in odds-ratios calculated as the exponentiated sum of all basis functions.

The modelling framework was implemented in the R package "mgcv" (Wood 2011). Thin plate regression splines were used as they are isotropic (non-directional), have a specific start and end (non-cyclic) and do not require pre-specified breakpoints (knots). The latter point is important as the latent distribution of mutational hotspots is unknown prior to modelling. As the outcome variable is binary (disease status), it was modelled using a binomial distribution and a logit link function. Smooth functions were fitted using restricted maximum likelihood as this has been shown to reduce optimization issues and provide less variable estimates (Wood 2011).

Both non-carriers and carriers were modelled to include gene-level burden into the model. For non-carriers, i.e. samples with no variants in the gene, no value existed for residue position, or any other feature we will consider later on. This presented a complication. If the position was specified as zero for this continuous variable in non-carriers, the model would assume a high density of observations at the 0^{th} residue position. This is a meaningless result. If they were instead specified as NAs, the model either excludes all non-carriers or cannot fit. To overcome this, a nested model

structure was implemented whereby residue position was included only as an interaction with carrier status. Carrier status then functioned as a binary indicator variable for residue position, and when zero, was multiplied with the model matrix for residue position in non-carriers to exclude this undefined data from the model. Carrier status was then also included as a linear model term. This allowed carrier rates between the two cohorts (i.e. gene-level burden) to be modelled. Without information on sample size, here supplied via the carrier status variable, modelling burden is not possible. The result of combining these two features was the desired regional burden model.

The model structure for the *hotspot model* and code implementation in mgcv are presented below;

$$ln(P/1\text{-}P) = \beta_0 + \beta_1\, carrier + s_1\,(pos, by = carrier) + \varepsilon$$

```
gam(affected ~ carrier + s(pos, by = carrier, k=15, bs="tp"), method="REML", weights=weights,
family="binomial", data = case_control)
```

where ln(P/1-P) is a logistic function specifying the probability of being affected, β_0 is the model intercept, β_1 is the linear coefficient for *carrier* (status), s_1 the smooth function for *pos* (residue position), "by" is used to generate factor-smooth interactions and ε is a binomial random residual error term. In the R mgcv implementation; the first input denotes the model formula, the weights are for coverage adjustment, and the "case_control" dataset consisted of a tabular dataset with three columns; affected (case=1, control=0), carrier (variant in gene=1, variant not observed=0) and pos (residue position of variant). *K=15* indicates the maximum number of basis functions (and effective degrees of freedom) for the model and *bs="tp"* sets the smooth class as thin-plate regression splines.

3.1.3 Adjusting for uneven sequencing depth

For some genes investigated for mutational hotspots, low coverage regions existed in either the OMGL, HCMR or *gnomad_e2* datasets (Figure 2.1). Due to this, the effective sample size was

potentially different for some regions of the gene, as not all samples were covered to sufficient depth for variant-calling. For analysis of burden, this would result in the incorrect denominator for the calculation of frequency of variant carriers, leading to overestimates of carrier frequencies. For positional analyses, this could generate artificial hotspots driven by low coverage regions in one cohort.

Coverage was accounted for in the model by assigning prior weights on the contribution of each variant to the log-likelihood. Each variant was weighted by the reciprocal of the mean 10X coverage in the local region for its cohort (i.e. cases or controls). The local region was specified as the variant residue plus five residue flanks towards both the N and C terminus (11 residues). The result of this weighting procedure is that low coverage regions were up-weighted in the model. As a proof-of-concept, a simple simulation framework was used to assess this weighting approach for coverage adjustment (Figure 3.2). Variants were simulated in two cohorts, for a disease model with one C-terminal hotspot, and a null model of uniform distribution of variants, and either 100 or 200 total variants. In each of these four scenarios, three models were trained. The first was with the original data (Unmodified model). In the second, the region between 100 and 200 residues was then assigned as a low coverage region (10X = 0.5) and a random 50% of observations were removed (LowCov model). The LowCov scenario was then remodelled with the proposed weighting procedure (Adjusted model). Under these scenarios, the adjusted model was similar to the unmodified model, whereas the LowCov model determined an incorrect relationship. As expected, the higher the number of variants, the more successful the adjusted model was at recovering the true relationship.

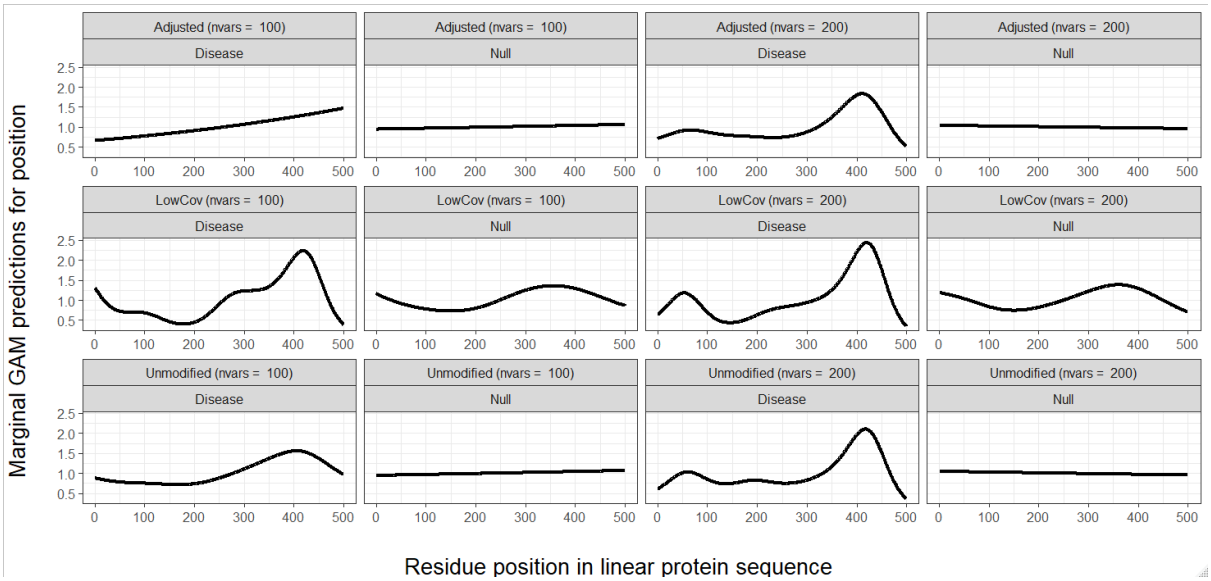

Figure 3.2: The effect of inverse-coverage weighting for simulated low-coverage data. Data is simulated for a disease and null gene. Variants are then systematically removed in a region defined as low coverage. A GAM model is generated for the original data (Unmodified), the low-coverage data without adjustment (LowCov) and the low-coverage data with adjustment (Adjusted).

3.1.4 Filtering genetic variants before modelling

Although both missense and loss-of-function (LoF) variants can cause disease in the genes under investigation in this chapter, only missense are considered for the *hotspot models*. Again, our definition of missense variants is extended to include small inframe insertions and deletions, and only the canonical transcript is considered in each gene. All variants are filtered using the maximum observed frequency in any of the eight ancestry group in *gnomad_e2* or globally in *gnomad_g2* (up to 15,708 samples). The threshold for excluding variants based on this maximum frequency was 0.01% (i.e. 1 in 10,000). Allele frequency filtering was implemented for both cases and controls. This allows interpretation of model results purely in terms of burden differences of rare-variants between the cohorts.

3.1.5 Model evaluation

Hotspot models were trained on the heart disease cases and *gnomad_e2* controls. Raw model outputs in the form of logistic probabilities and standard errors were transformed into odds-ratios and 95% confidence intervals for each residue in the protein. To integrate these models into the ACMG guidelines, model predictions were stratified into evidence categories based on the assumption that model ORs were equivalent to the probability of variant pathogenicity. A similar approach has been proposed before for criterion PM1 (Walsh *et al.* 2019). Here, we use a more conservative approach and apply supporting evidence for ORs greater than 10, moderate evidence for ORs greater than 20 and strong evidence for ORs greater than 100. These OR thresholds roughly correspond to the probabilities; 90%, 95% and 99%, respectively.

Although many methods exist to evaluate the performance of predictive models, the main objective here is the inference of mutational hotspots. Therefore, model performance was primarily assessed by the OR estimates and associated 95% confidence intervals. Nonetheless, traditional evaluation metrics were also calculated to evaluate the predictive power of residue position. Furthermore, model predictions were correlated with ordinal pathogenicity classifications made by the Oxford Medical Genetics Laboratory (using the ACMG variant interpretation framework). Overlap of hotspots and constrained-coding regions (CCRs; Havrilla *et al.* 2019), the distribution of common GnomAD variants, pathogenic *clinvar* variants, and protein feature annotations, was assessed. Inferred regions in HCM were compared to hotspots identified by Walsh *et al.* (2019) using a different framework.

As this is a data-driven approach, it was not possible to generate meaningful models for genes with few observations. A cutoff was implemented, whereby only genes with at least five unique variants were considered for modelling. Consequently, *hotspot models* were generated for 13 disease-gene pairs with at least *n-significant* clustering signals, determined by the *BIN-test*. This includes 9 HCM

genes, a single DCM and ARVC and two LQTS genes. However, in order not to miss genes with the potential for hotspot modelling that were insignificant in the burden test, this was followed by modelling all genes with *b-significant* burden signals and *BIN-test* p-values over 0.05, consisting of; 4 HCM genes, 8 DCM genes, three LQTS genes and one BrS gene.

3.2 Hotspot models

3.2.1 Hotspot models for clustered genes

Hotspot model predictions for each amino acid residue of the investigated proteins are displayed in Figure 3.3. For a protein with uniformly equal burden in cases and controls, predictions are expected to be a flat line at 1 (blue horizontal line on the plot). For all genes, almost all residues had a predicted OR greater than 1, primarily driven by a global excess of rare-missense variants. As hypothesized, predictions varied substantially across the proteins, highlighting mutational hotspots with greater enrichment than the uniform burden OR (plotted as a red horizontal line). No upper confidence intervals are observed below 1, suggesting no protective regions are inferred with high confidence.

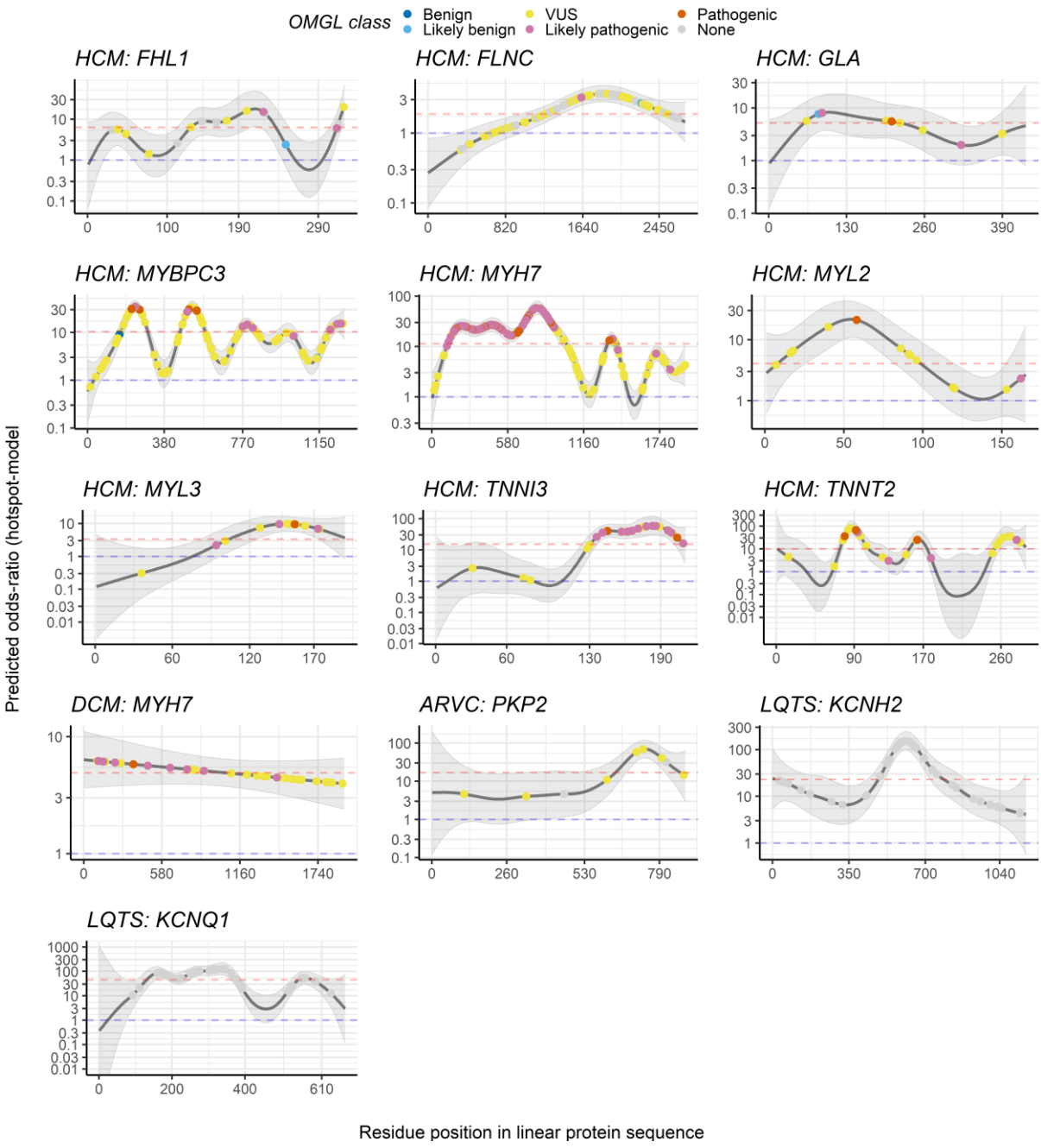

Figure 3.3: Estimation of regional burden and mutational hotspots across multiple inherited-heart disease disease-genes. Red lines indicate the gene-level odds-ratio, blue lines intersect the y-axis at one and grey ribbons indicate the 95% confidence intervals. Training data from the case-cohort are plotted and coloured by their pathogenicity classification.

The models with the highest confidence were the firmly established HCM genes; *MYH7* and *MYBPC3*, each having 517 and 461 rare-missense variant carriers respectively. This is far higher than the next most frequently mutated HCM gene, *FLNC*, which had only 139 variants (primarily

due to its large size rather than high penetrance). In *MYH7*, the estimated burden is substantially higher than the uniform burden (OR=11.4) in the motor-head region of the protein, with predicted ORs reaching a maximum of 59 in the second peak and are consistently over 50 from residues 762 to 850. One previous definition of clustering in *MYH7* defines the mutational hotspot in the motor-head region (Kelly *et al.* 2018) spanning residues 85-778. A more recent data-driven estimate determined a mutational hotspot between residues 167-931 (Walsh *et al.* 2019). For the former, motor-head region definition, moderate evidence was supplied by criterion PM1, for the latter, data-driven approach, strong evidence is proposed. For both, the whole region is assumed to be uniformly enriched with rare-variation. Here, predictions vary within regions proposed by Kelly *et al.* (2018) or Walsh *et al.* (2019). For the motor-head domain, the *hotspot model* predictions range from 6 at position 85, to 55 at position 778. Interestingly, the range from the domain-based approach does not cover the residue with the highest prediction, residue 805 (OR=59). The dearth of variation in the tail region of *MYH7* may inspire a hypothesis of a local protective effect. However, the *hotspot model* analysis makes it clear that while the excess variation is less than the gene-wide average (OR=11.4), it is uniformly greater than 1, discounting this hypothesis. Within the tail region, there are two weakly clustered regions overlapping with observed pathogenic variants in this region. Although not higher than the gene-level burden, it may have relevance to the functional or structural importance of this region.

A confident excess of rare-missense variation was predicted in the C-terminus of the large *FLNC* gene. While the gene had a relatively low average effect (OR=1.87), the predicted odds-ratio exceeded 3 for the C-terminal region ranging from 1563-2164. The lower 95% confidence interval also exceeded the uniform OR of 1.87, giving confidence the effect-size in this region is truly above the average effect. *FLNC* is a recently discovered low-penetrance HCM gene and its molecular relevance to HCM pathology is under investigation.

A complex pattern of burden was seen throughout the small *TNNT2* gene (length=298 residues). This was characterized by a central increase in variant density consisting of two peaks, a C-terminal peak and a modest uplift of predicted burden in the N-terminus. The maximum observed predicted OR in this gene was 84 and observed at residue position 87. Within the central cluster, there are two clusters where predicted ORs exceed 30 across the ranges 79-98 and 262-274. A very clear clustering signal was observed in the other small troponin gene *TNNI3* (length=210 residues), composed of a prominent C-terminal hotspot with relatively tight confidence intervals and an N-terminal region with similar variant density to controls (OR~1) with wide confidence intervals. Confidence intervals are wide in this specific case (and many others), as when observations are sparse in the smaller cohort, the addition or subtraction of a single observation makes a large difference to the point estimates. Predicted ORs exceed 30 within the range 138 to 201, reaching a maximum of 59 in this region and a mean of 45.

There is a distinctive hotspot in the N-terminus of *MYL2* with 95% confidence that variants in this region have a higher OR than the average rare-missense *MYL2* variant (OR=4.05). The region of increased variant density in cases contains a moderate cluster (OR>20), between residues 48-60, with predictions peaking at 21.6 in this region. This is flanked by regions of supporting evidence, with the whole region (OR>10) spanning from 28-78. There is a moderate hotspot in the N-terminus of *MYL3*, and although predicted ORs are substantially higher than the gene-level OR of 3.44, they do not exceed the nominal threshold for supporting evidence (OR>=10). A modest increase in case variant density was observed in the N-terminus of *GLA*. Again, no predicted ORs exceed the threshold for supporting evidence, and confidence intervals are wide. An interesting pattern of regional burden was detected in *FHL1* characterised by an overburdened centre and two modest peaks in each tail. Supporting evidence of pathogenicity is predicted for two ranges; 180-229 and 317-323.

For DCM, the only gene with *n-significant* clustering was *MYH7*. Although a trend of increased case density in the N-terminus was observed, a roughly similar pattern to that in HCM, this did not breach the threshold for supporting evidence. Interestingly, this finding is contrary to previous evidence suggesting DCM-causing variants in *MYH7* are dispersed uniformly across the gene (Walsh *et al.* 2017). In ARVC, the prevalent missense variant in the tail region, p.His733Asp, observed in 11/502 cases, drives a strong hotspot signal. In LQTS, the two genes *KCNH2* and *KCNQ1* were the only genes present in our data with predicted ORs above 100. For *KCNH2*, a strong cluster is observed in the centre of the protein between residues 560 and 664, peaking at an OR of 155. This is flanked by supporting and moderate clusters than altogether span the range 426 to 900. In *KCNQ1*, evidence of pathogenicity is found in the wide N-terminal hotspot between residues 90-405, including a strong cluster between 267-355.

Multiple peaks were observed in *MYBPC3* with predictions falling as low as OR=0.6, at the second residue of protein, and peaking at OR=35, at residue 238, much higher than the uniform gene-level odds-ratio (OR=10.4). Two substantial peaks with odds-ratios higher than 30, were detected at residues 216-258 and 501-538. Two further regions had predicted ORs higher than the gene odds-ratio, between ranges 753-834 and 1199-1275. Peaks overlap with variants clinically classified as pathogenic, including potential founder variants. To assess the impact of these relatively high frequency case-variants, the model was rebuilt, counting these variants only once in cases and excluding from controls (if present). This applied to 6 variants; p.Val219Leu (AC=15), p.Glu258Lys (AC=59), p.Arg495Gln (AC=23), p.Glu542Gln (AC=28), p.Asp770Asn (AC=11) and p.Arg810His (AC=17). Importantly, the most well-known founder and pathogenic variant p.Arg502Trp is not present in the filtered data due to a frequency of 0.00017 in the GnomAD exomes “other” population. This potentially over-zealous filtering will be reconsidered later. For now, the *MYBPC3 hotspot model* was generated with the new training data (Figure 3.4). Although less pronounced, the overall pattern of regional burden is consistent with the previous model. Five of these variants are present in the *gnomad_e2* dataset anyway, up to twelve times for p.Arg810His, presumably

due to incomplete penetrance. As these were present in the training data labelled as controls, this may already dampen the potential over-weighting of these variants, as the predicted OR can be considered the penetrance at that position, estimated based on the frequency of these variants in each cohort. The sensitivity analysis provides some evidence the variants are not unduly driving patterns observed in the *MYBPC3 hotspot model*.

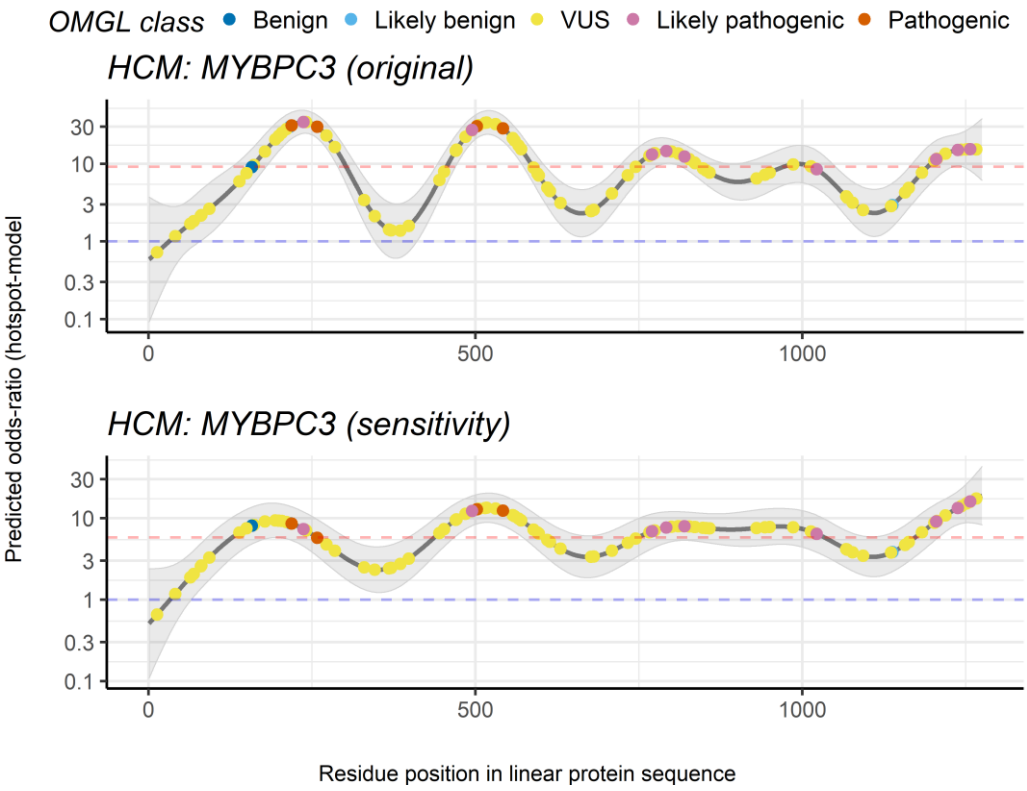

Figure 3.4: *Hotspot model* for *MYBPC3* excluding potential founder variants (sensitivity analysis). Six variants with an allele count above ten were removed from the training set to determine the impact of these potential founders on the estimated regional burden.

Specific regions of supporting, moderate or strong evidence of pathogenicity (based on criterion PM1) are reported in Table 3.2. Only two regions in the analysis exceeded an OR of 100, reaching strong evidence of pathogenicity in line with the cautious evidence-thresholds proposed. These were observed in the LQTS genes *KCNH2* and *KCNQ1*. However, regions of supporting and moderate evidence of pathogenicity were discovered in nine out of thirteen n-significant clustering

genes investigated for hotspots. Unlike previous attempts at hotspot estimation, much higher resolution is provided, dissecting proteins into PM1_supporting, PM1_moderate, PM1_strong or no evidence for specific ranges.

Table 3.2: Specific protein regions with supporting, moderate and strong evidence of overburden, in significantly associated inherited heart disease genes. Max = the maximum OR in this region, N = number of case variants within the region, n P = number of pathogenic variants, n LP = number of likely pathogenic variants, n VUS = number of variants of uncertain significance.

Disease	Gene	Level	Region	Max	N	n P	n LP	n VUS
ARVC	*PKP2*	Mod	647-848	70	14	0	0	13
ARVC	*PKP2*	Sup	601-646, 849-880	20	2	0	0	2
HCM	*FHL1*	Mod	323-323	22.8	0	0	0	0
HCM	*FHL1*	Sup	180-229, 317-322	19.9	10	0	2	4
HCM	*MYBPC3*	Mod	193-278, 481-560	34.7	170	110	34	25
HCM	*MYBPC3*	Sup	163-192, 279-299, 459-480, 561-	19.7	91	0	35	56
HCM	*MYH7*	Mod	168-550, 663-956	59	333	139	124	67
HCM	*MYH7*	Sup	114-167, 551-662, 957-1038,	20	75	34	21	19
HCM	*MYL2*	Mod	48-60	21.6	10	9	1	0
HCM	*MYL2*	Sup	28-47, 61-78	19.9	1	0	0	1
HCM	*TNNI3*	Mod	134-206	59.5	71	21	37	13
HCM	*TNNI3*	Sup	128-133, 207-211	20	4	0	1	3
HCM	*TNNT2*	Mod	77-101, 160-169, 258-281	84	41	21	3	17
HCM	*TNNT2*	Sup	1-1, 74-76, 102-106, 154-159,	19.3	4	0	0	4
LQTS	*KCNH2*	Mod	1-55, 477-559, 665-804	99.7	8	0	0	0
LQTS	*KCNH2*	Strong	560-664	155	26	0	0	0
LQTS	*KCNH2*	Sup	56-201, 426-476, 805-900	20	9	0	0	0
LQTS	*KCNQ1*	Mod	110-266, 356-391, 524-620	99.5	53	0	0	0
LQTS	*KCNQ1*	Strong	267-355	129	22	0	0	0
LQTS	*KCNQ1*	Sup	90-109, 392-405, 509-523, 621-	19.8	4	0	0	0

3.2.2 Hotspot modelling of overburdened genes, not significantly clustered

Modelling of burdened genes generally gave flat results with wide confidence intervals, however, some regions did reach the threshold for supporting and even moderate evidence (Table 3.3; Figure 3.5). For HCM, three important genes; *ACTC1*, *CSRP3* and *TPM1*, did not show significant evidence of clustering in the *BIN-test*. Interestingly, the GAM model detected weak clustering in the centre of *ACTC1*. This clustering was sufficient to propose at least supporting evidence of pathogenicity

throughout the range 73-308 and moderate evidence from 136-245. For *TPM1*, the model predicted a uniform distribution of burden throughout the gene just underneath the level of supporting evidence. For *CSRP3* and the additional HCM genes; *DMD* and *LAMP2*, the predicted ORs were all well below the supporting level (OR=10).

Table 3.3: Regions of supporting and moderate evidence of pathogenicity in significantly overburdened genes without clear evidence of intra-gene clustering. Max = the maximum OR in this region, N = number of case variants within the region, n P = number of pathogenic variants, n LP = number of likely pathogenic variants, n VUS = number of variants of uncertain significance.

Disease	Gene	Evidence	Region	Max	N	n P	n LP	n VUS
ARVC	*DSC2*	Supporting	100-302	14.7	6	0	2	4
BrS	*SCN5A*	Supporting	1-277, 1060-1350, 1570-1800	20	17	4	5	8
BrS	*SCN5A*	Moderate	1350-1570	22.1	7	0	3	4
DCM	*TNNT2*	Supporting	91-112, 208-259	19.8	2	2	0	0
DCM	*TNNT2*	Moderate	113-207	35.8	12	3	6	3
DCM	*TPM1*	Supporting	207-284	13.7	3	1	0	2
HCM	*ACTC1*	Supporting	73-135, 246-308	19.9	10	3	0	7
HCM	*ACTC1*	Moderate	136-245	24.6	7	0	0	7
LQTS	*KCNJ2*	Supporting	33-199	16.4	4	0	0	0

In DCM, for some genes such as *ACTC1*, sparse data led to infinite confidence intervals in some regions. In *TNNT2* and *TPM1*, the *hotspot models* do determine some regions with sufficient burden to provide evidence of pathogenicity. In both cases signals are driven by the large gene-level ORs accompanied by modest clustering. The same is true for *DSC2* in ARVC, *KCNJ2* in LQTS and *SCN5A* in BrS. It should be noted that while supporting or moderate evidence are achieved for some regions in these genes, confidence intervals surrounding these estimates are wide.

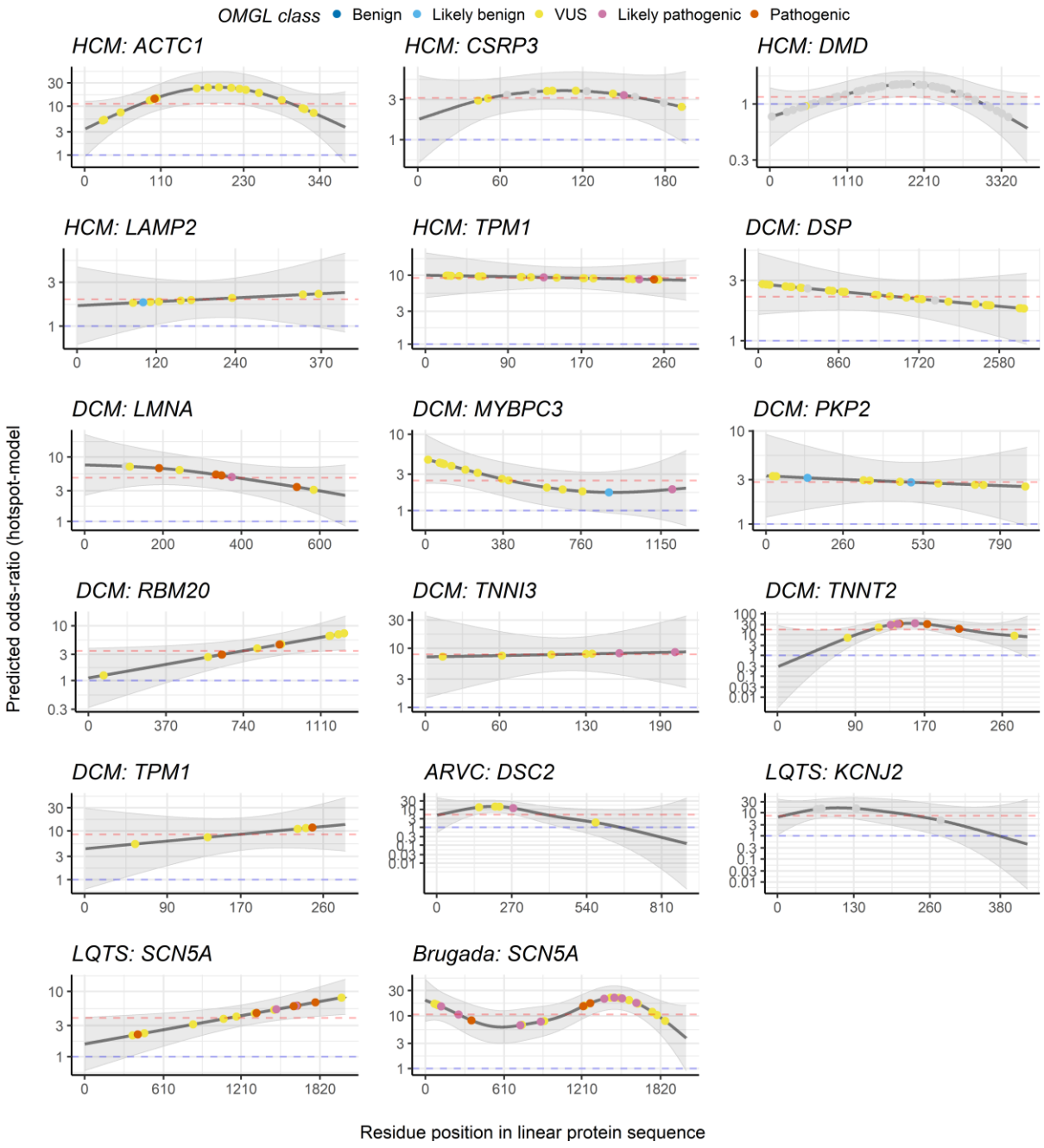

Figure 3.5: *Hotspot models* for overburdened genes without clear evidence of clustering. Red lines indicate the gene-level odds-ratio, blue lines intersect the y-axis at one and grey ribbons indicate the 95% confidence intervals. Training data from the case-cohort are plotted and coloured by their pathogenicity classification.

3.2.3 *Inter-disease hotspots*

There is an overlap of causal genes for HCM and DCM, but it is unclear what drives this, nor how to distinguish variants likely to be associated with either phenotype in the same gene. It is hypothesized, that they are at least in part, separable by their location in the gene (Bollen and van der Velden 2017). To determine if the GAM modelling approach could identify regions in the gene with differential burden across HCM or DCM, inter-disease models were generated for the seven genes with *n-significant* burden signals in both cardiomyopathies; *ACTC1*, *FHL1*, *MYBPC3*, *MYH7*, *TNNI3*, *TNNT2* and *TPM1* (Figure 3.6).

For *ACTC1* and *FHL1*, regions with substantial sparseness in either cohort led to meaningless models. For *MYBPC3*, the model predicted that regions in the centre of the protein were more likely to lead to HCM, whereas at the very beginning of the protein, density is actually higher in DCM, however, confidence intervals do exceed one even in this region. For *MYH7*, the burden of variants is higher in HCM throughout the gene, except in the very tail of the protein, but again confidence intervals are higher than one throughout. *TNNI3* is the first gene where upper confidence intervals reach below one and lower confidence intervals reach above it. This may indicate variants in the N-terminus of *TNNI3* are more likely to cause DCM whereas C-tail variants are more likely to cause HCM. A potentially interesting pattern is also observed in *TNNT2* where a small region in the centre of the protein is overburdened in DCM compared with HCM. The wide confidence intervals of *TPM1* suggest almost uniform pathogenicity in both cohorts.

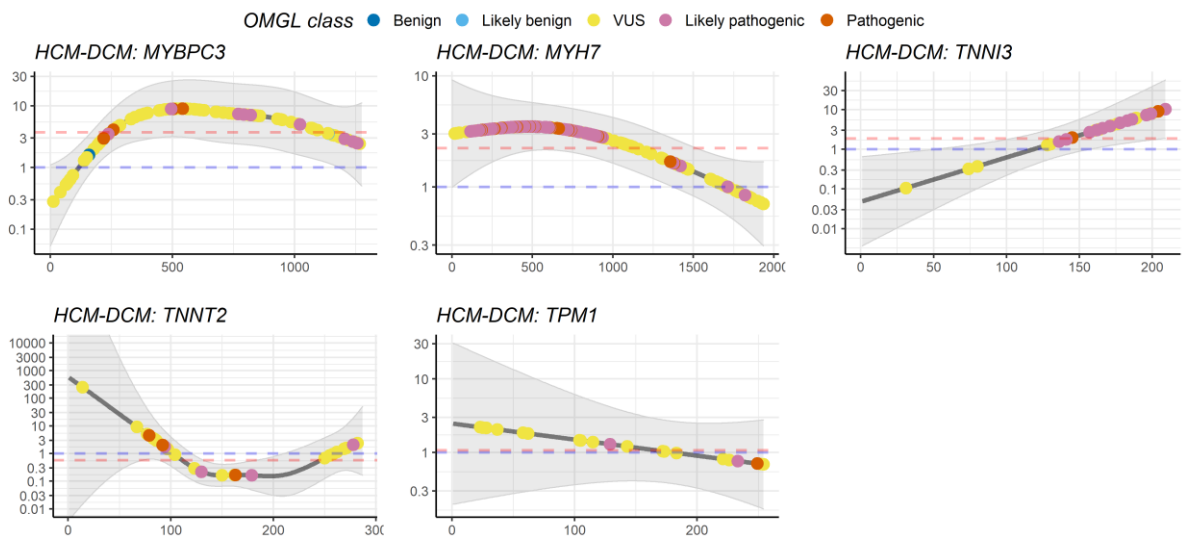

Figure 3.6 Regional burden of rare-missense variants in HCM cases relative DCM cases, in genes known to harbour pathogenic variants for both diseases. Red lines indicate the gene-level odds-ratio (relative to the first disease group), blue lines intersect the y-axis at one and grey ribbons indicate the 95% confidence intervals. Training data from the HCM-cohort are plotted and coloured by their pathogenicity classification.

The wide confidence intervals of these results indicate a lack of power to find true differences. Furthermore, it may be unlikely to discover true differences using this approach. It is possible that the locational differences that drive the phenotype in each direction differ at a much more complex level, requiring more refined detail to uncover them. For the case-control study, variants may cluster in functional regions in cases. For the case-case study, variants may cluster in functional regions in both cohorts, and the difference may lie in their exact position in that region or even the alternative amino acids.

3.2.4 Model evidence ACMG

Predictions from all *hotspots models* with at least one region with an OR>=5 were stratified into supporting (OR>=10), moderate (OR>=20) or strong (OR>=100) evidence of pathogenicity (Figure 3.7). Regions with predicted ORs above five were also plotted to improve resolution. Of the 23 genes that this applied to, only predicted ORs in the LQTS genes *KCNH2* and *KCNQ1* exceeded 100

(plotted as gold), the threshold for strong evidence. A further ten genes have regions of moderate evidence of pathogenicity, and a further two had at least supporting evidence. This results in 14 genes across the gene panels in which the criterion PM1 could be applied using a data-driven statistical approach. This includes several genes where no significant clustering was determined, highlighting the value of gene-level burden as a signal of mutational hotspots.

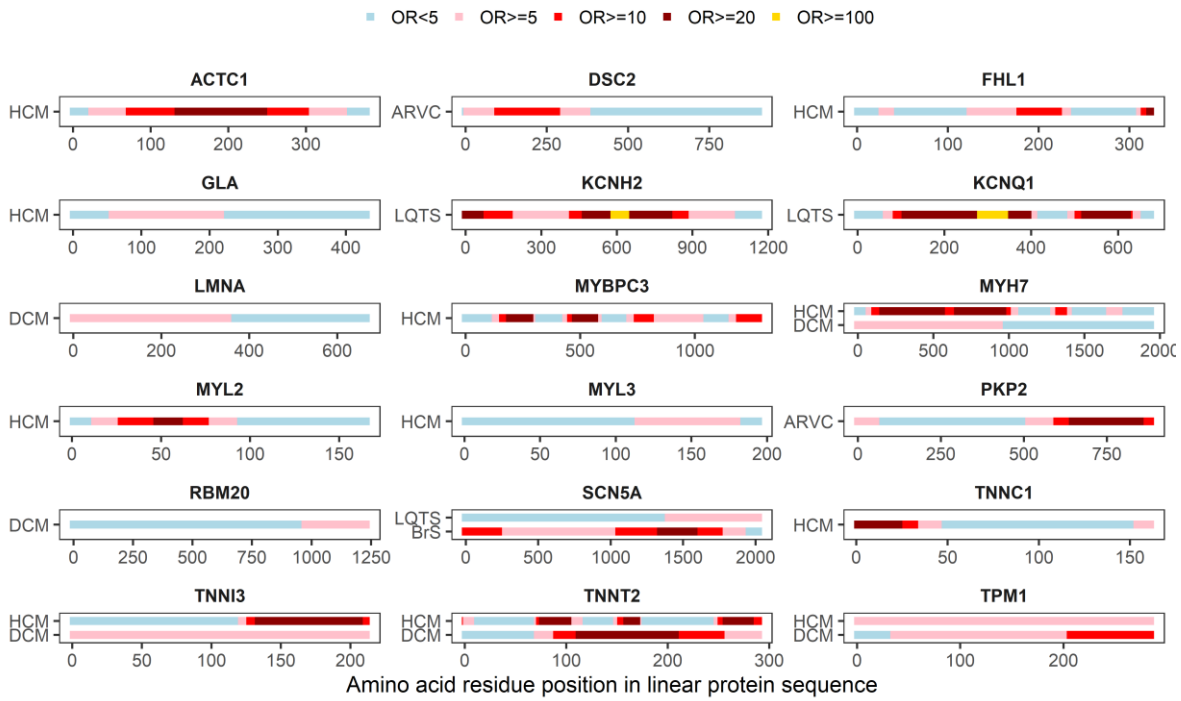

Figure 3.7: Protein regions stratified by PM1 evidence category for all genes with at least one region with a predicted hotspot OR greater than 5. Disease group is labelled on the y-axis with some genes having hotspots in multiple disease; *MYH7*, *SCN5A*, *TNNI3* and *TNNT2*.

3.2.5 Prediction metrics

Generic prediction metrics were calculated for each *hotspot model* to determine their power to predict whether a variant was observed in a case or control (Table 3.4). Variants observed in both cohorts, due to neutral background variation or incomplete penetrance, imposed an upper limit on predictive accuracy. Five confusion matrix measures were calculated for each model; accuracy,

true-positive rate (TPR), true-negative rate (TNR), false-positive rate (FPR) and false-negative rate (FNR). The TPR (sensitivity) is the proportion of cases that carried a missense-variant in the gene that were correctly identified as cases dependent on the position of that variant. The TNR (specificity) is the proportion of controls that carried a missense-variant in the gene, correctly identified as controls. Naturally, the FPR is the proportion of controls classified as cases, and the FNR is the proportion of cases classified as controls. To calculate these performance metrics, a threshold must be specified to split logistic model probabilities into binary classifications. Different probability thresholds can be adopted depending on the purpose of the model, allowing sensitivity or specificity to be favoured. In this analysis, the threshold that produced the closest true-positive rate (specificity) and true-negative rate (sensitivity) was used.

Table 3.4: Metrics to assess *hotspot model* performance ordered by model accuracy. TPR = true-positive rate, TNR = true-negative rate, FPR = false-positive rate, FNR = false-negative rate. The number of variants in cases and controls are denoted by columns, Cases and Controls respectively. The cutoff denotes to the threshold used to stratify logistic probabilities to binary outcomes.

Disease	Gene	Cases	Controls	Cut-off	Accuracy	TPR	TNR	FPR	FNR
ARVC	*DSC2*	7	638	0.0242	0.833	0.857	0.832	0.168	0.143
HCM	*TNNT2*	53	125	0.381	0.826	0.811	0.832	0.168	0.189
HCM	*TNNI3*	78	124	0.606	0.762	0.833	0.718	0.282	0.167
HCM	*MYL3*	21	148	0.238	0.746	0.714	0.75	0.25	0.286
ARVC	*PKP2*	19	580	0.0387	0.741	0.737	0.741	0.259	0.263
HCM	*FHL1*	32	120	0.257	0.737	0.719	0.742	0.258	0.281
HCM	*MYL2*	24	140	0.162	0.732	0.75	0.729	0.271	0.25
LQTS	*KCNH2*	51	508	0.075	0.726	0.725	0.726	0.274	0.275
HCM	*MYH7*	504	1140	0.389	0.704	0.706	0.702	0.298	0.294
HCM	*MYBPC3*	365	884	0.327	0.703	0.704	0.702	0.298	0.296
HCM	*ACTC1*	24	51	0.355	0.693	0.708	0.686	0.314	0.292
HCM	*TNNC1*	7	54	0.126	0.689	0.857	0.667	0.333	0.143
DCM	*TNNT2*	16	125	0.165	0.688	0.688	0.688	0.312	0.312
LQTS	*KCNQ1*	79	429	0.2	0.679	0.684	0.678	0.322	0.316
BrS	*SCN5A*	31	1190	0.0313	0.678	0.677	0.678	0.322	0.323
LQTS	*SCN5A*	20	1190	0.0182	0.65	0.65	0.65	0.35	0.35
DCM	*RBM20*	14	487	0.0299	0.643	0.643	0.643	0.357	0.357
HCM	*GLA*	19	87	0.192	0.632	0.632	0.632	0.368	0.368
DCM	*LMNA*	14	397	0.0371	0.579	0.571	0.579	0.421	0.429
DCM	*MYH7*	39	1140	0.033	0.573	0.538	0.574	0.426	0.462
DCM	*TNNI3*	7	124	0.0534	0.573	0.571	0.573	0.427	0.429

HCM	*TPM1*	30	80	0.271	0.518	0.467	0.538	0.462	0.533
DCM	*TPM1*	5	80	0.0514	0.435	0.8	0.412	0.588	0.2

The best predictive performance was from the ARVC gene *DSC2,* which had an accuracy of 83.3%. This represents a correct classification for 6/7 cases (TPR=85.7%) and 531/638 controls (TNR=83.2%). Although a supporting N-terminal hotspot was identified in this gene, there were so few case variants to train the model, confidence intervals are wide, and it is unclear whether this would still hold true with the addition of more ARVC samples. Four likely pathogenic variants have been reported in *clinvar* for this gene (p.Arg375Gly, p.Thr275Met, p.Phe250Ser and p.Arg203His), and only p.Thr275Met occurs in this ARVC training set. Three of these variants overlap with the identified supporting hotspot region (100-302). For both *gnomad_e2* and *gnomad_g2*, less than 25% of rare-missense variants map to this region, while 4/5 unique ARVC variants map to this region. However, in a previous aggregation of clinical reports (Walsh *et al.* 2017), only 2 of the 7 variants observed in 351 ARVC patients mapped to this region. To reconcile these opposing pieces of evidence and to confirm this hotspot, additional validation data is required.

The next best classifiers were the *hostpot models* for the troponin genes; *TNNT2* (accuracy=82.6%) and *TNNI3* (accuracy=76.2%) in HCM. Both had sufficient numbers of samples to reduce confidence intervals to reasonable levels and may indicate realistic estimates of the accuracy of residue position as a predictor. Other established HCM-causing genes; *MYL3*, *FHL1*, *MYL2*, *MYH7* and *MYBPC3* are also amongst the most accurate models. The ARVC gene *PKP2* also stands out again but with the same caveat of the frequently occurring p.His733Asp variant. Overall, ten models achieved an accuracy greater than 70%, demonstrating useful predictive value for this predictor. The worst performing models were from the *TPM1* gene in both HCM and DCM. This implies that the residue position is not a good predictor of affection status and therefore pathogenicity in this gene.

3.2.6 Stratification against expert classifications

Most variants in the inherited heart disease datasets were classified internally by clinical scientists at the OMGL using an ACMG based framework. *Hotspot model* predicted ORs were stratified by their classification (Figure 3.8). For each variant, where no OMGL classification was available, *clinvar* classifications were used instead, if available. This was the case for all LQTS variants. Correlation between expert classifications and model predictions were assessed using Spearman's rank-order correlation test. To achieve this, the categorical classifications were treated as an ordinal variable, ranked from 1 (benign) to 5 (pathogenic). Only variants passing the strict *popmax* filter were included in this comparison, so very few likely benign and benign variants are observed.

Seven disease-gene pairs had a significant correlation (alpha<0.05) between model predictions and ranked classifications; DCM-MYH7 (rho=0.53), DCM-TNNI3 (rho=0.79), HCM-MYBPC3 (rho=0.73), HCM-MYH7 (rho=0.33), HCM-MYL2 (rho=0.83), HCM-TNNI3 (rho=0.29) and LQTS-KCNQ1 (rho=0.26). Generally, models demonstrated an expected positive correlation. A mean Rho of 0.21 was observed across all genes. However, for some genes, there was a non-significant negative correlation, such as *TPM1* (Rho=-0.29) and *DSC2* (Rho=-0.49). As these classifications may be inaccurate, and because the VUS are all either pathogenic or benign, uncorrelated results from this analysis do not inevitably mean the models are wrong. For the case of *DSC2*, the negative correlation is due to the 5 VUSs having higher mean predictions than the two observed LP variants. Perhaps indicating these variants should be upgraded to LP.

VUSs had a very wide spread of model predictions with some very low and very high ORs. This is especially the case for the genes *MYH7* and *MYBPC3,* which harbour the most VUSs. In *MYH7*, mean predicted ORs for pathogenic, likely pathogenic and VUS variants observed in cases, were 74, 50 and 20 respectively. However, VUSs had predicted ORs ranging from 0.25 to 197. Half of these VUSs were observed in a single case and absent in controls (private singletons). The empirical ORs

for these variants, based on the case and control frequencies and adding 0.5 to zero-count cells, Haldane's continuity correction (Haldane 1956), have wide 95% CIs: 44.9 [1.5, 1338.3]. However, *hotspot model* ORs for such variants had greater precision with varying point estimates depending on the precise amino-acid substitution. In *MYH7*, five of these singleton VUSs had predicted ORs greater than 100 and three had ORs less than 1. This high variance in the VUS group lowered the correlation estimates for many genes, especially *MYH7* and *TNNI3*.

3.2.7 Comparison with previously discovered HCM hotspots

Hotspot estimation in HCM has been tackled previously by Walsh *et al.* (2019). They developed an unsupervised 1-dimensional clustering algorithm (henceforth abbreviated as 1dC), to detect clusters of missense variants in HCM proteins using clinical reports for several thousand patients. The approach assesses local enrichment of variant counts via a binomial test, within and outside sliding windows across the protein. Optimal window size is determined by an iterative process and a boundary trimming procedure is implemented to refine clusters. The significance of trimmed clusters is then assessed using a final binomial test, and the etiological fraction is assessed using ExAC controls. Their approach identified clusters in six genes (Figure 3.8).

There was some overlap between identified clusters and also some mismatch. In 1dC the gene *CSPR3* is found to have an N-terminal cluster whereas no clusters are detected in the *hotspot-model*. Although an N-terminal cluster of variants was identified in our cases, this matched with an N-terminal cluster of variants with controls. This highlights the importance of case-control hotspot detection. In the case-only approach, local regions of increased background variation could be misinterpreted as hotspots. If the gene was uniformly overburdened, then this region would then be confirmed as a local region of excess variation in cases whereas in actual fact the whole gene is overburdened.

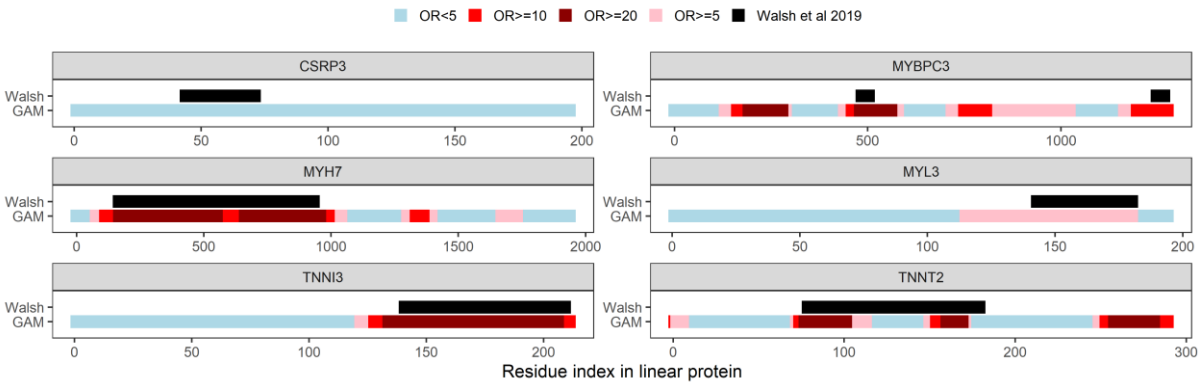

Figure 3.8: Comparison of predicted mutational hotspots with those predicted by 1dC. Only the six genes with hotspots detected by 1dC are considered. GAM hotspots are stratified into four classes by predicted OR. 1dC hotspots were not stratifiable.

For *TNNT2*, predictions differed between the two approaches, two central clusters detected by the *hotspot model* were collapsed into one cluster in 1dC and start and end clusters were not detected using 1dC. Inspection of the Walsh *et al.* (2019) data revealed an almost identical pattern of observed variants throughout the gene. For the 1dC approach, each assessment of clustering in windows across the gene, compared within window counts with the rest of the gene. For the case of multiple clusters, when the window does focus on a real cluster, outside-window counts are skewed by the excess of variants in the additional clusters, diminishing the difference between the true cluster and the rest of the gene. It is possible that the clusters not detected by 1dC fell victim to this bias. For the central region of excess case variation, the *hotspot model* identifies two clusters broken up by region not predicted to be a mutational hotspot. It is clear from both datasets that case variation is sparser in this region. For the data used here, the region with predicted ORs below ten from residues 110 to 150 has four observed case variants and 28 observed control variants. Although this does make it overburdened (OR~3), it falls far short of our conservative level of supporting evidence (OR>=10), and it is difficult to justify allocating strong evidence, as 1dC does.

For the remaining four genes, there was at least some overlap between the two approaches. For *MYH7*, the two large clusters in the N-terminal largely mapped to a large cluster identified by 1dC,

however the *hotspot model* refined this cluster into regions of supporting (OR>=10) and moderate (OR>=20) evidence, whereas 1dC assigns strong evidence for the whole region. In *MYBPC3,* two of the four hotspots identified by GAM were also observed in 1dC. For *MYL3*, the weak hotspot identified in GAM was also observed in 1dC. However, their estimated etiological fraction for this cluster did not reach 0.95 and therefore also did not have strong evidence under their framework. A strong cluster in the C-terminus of *TNNI3* was identified in both approaches.

In the GAM approach, an N-terminal cluster was found in *MYL2* whereas none were found in 1dC. Due to the small size of this gene (166 residues) and the relatively low proportion of the genetic heterogeneity of HCM it explains (~1-2%), few variants were observed in this gene (n=24). As Walsh *et al.* 2019 had even fewer cases, it is possible that their power was too low to detect this signal. Similarly, while GAM noted further hotspots in *FHL1*, *ACTC1* and *FLNC*, these were not detected by 1dC.

3.2.8 Constrained-coding regions (CCRs), common variants and pathogenic ClinVAR variants

In the last few years, much work has gone into identifying regions of high intra-species constraint by searching for contiguous regions with markedly reduced variation in large population control cohorts. In Table 3.5 and Figure 3.9, GAM hotspots are compared with constrained-coding regions (CCRs) identified by Havrilla *et al.* (2019). CCRs were specified as coding regions depleted of missense variants controlled for coverage and CpG density (mutability). CCRs here are displayed by their percentile across the exome. For example, a CCR greater than 99% denotes a region of constraint in the top 99% of the exome. Part of the dataset used to estimate these CCRs was GnomAD exomes and genomes, resulting in some overlap with the controls used to estimate hotspots. Notably, no CCR data was available for *GLA* and *FHL1* and was incomplete for *KCNH2* and *KCNQ1*.

Table 3.5: Mean constrained-coding region percentiles for protein residues stratified by their *hotspot model* predictions. The Rho correlation coefficient is calculated using Spearman's rank correlation test. Empty cells indicate no residues had predicted ORs within the given band.

		Mean CCR percentile					
Disease	**Gene**	*OR<5*	*OR>=5*	*OR>=10*	*OR>=20*	*OR>=100*	*Rho*
ARVC	*DSC2*	36.2	38.4	36.6			-0.021
ARVC	*PKP2*	29.4	40.5	30.1	31.6		0.075
BrS	*SCN5A*	17.7	43	45.2	57		0.178
DCM	*LMNA*	46.9	62.6				0.286
DCM	*MYH7*	41.3	61.5				0.276
DCM	*RBM20*	32.8	25.5				0.00828
DCM	*TNNI3*		48.5				0.153
DCM	*TNNT2*	35.9	62.7	52.4	48.5		0.06
DCM	*TPM1*	70.5	73.4	73.5			-0.145
HCM	*ACTC1*	61.2	81.3	82.1	85.1		0.27
HCM	*MYBPC3*	47.4	40.8	44.3	30.9		-0.0743
HCM	*MYH7*	41.8	46.1	53.7	62.4		0.31
HCM	*MYL2*	29.5	34.7	44.2	37.3		0.236
HCM	*MYL3*	34.5	43.8				0.229
HCM	*TNNC1*	51.5	76.7	88.3	94		0.512
HCM	*TNNI3*	39.6	36.2	57.7	62.9		0.445
HCM	*TNNT2*	45.2	38.2	43.2	68		0.104
HCM	*TPM1*		73				0.145
LQTS	*KCNH2*	42.7	48.5	53	62.5	78.6	0.338
LQTS	*KCNQ1*	28.3	38.6	39.7	51.5	51.1	0.188
LQTS	*SCN5A*	44.5	45.3				-0.0227

Visual inspection of hotspots and CCRs (Figure 3.9) revealed some cases of overlap but few exact and clear matches. For HCM, partial overlaps were found in *ACTC1*, *MYH7, TNNC1, TNNI3* and the last hotspot of *TNNT2.* For DCM, partial overlaps were found with *MYH7* and *LMNA*. In LQTS partial overlap is found in all four genes but is most notable in *KCNQ1* and *KCNH2*. Similarly, numerical summaries (Table 3.6) reveals both weak and strong relationships, between *hotspot model* ORs and CCR percentiles, for different genes. Generally mean CCR percentiles were higher for higher OR groups, for example; *SCN5A* (Brs), *LMNA* (DCM), *MYH7* (DCM), *ACTC1* (HCM), *MYH7* (HCM), *MYL2* (HCM), *TNNC1* (HCM), *KCNQ1* (LQTS) and *KCNH2* (LQTS). The N-terminal *TNNC1* hotspot strongly overlapped with regions of high CCR percentiles. Mean CCR for the OR<5 region was 51.5, whereas the OR>=20 region was 94. Spearman rank correlation coefficient for the correlation between predicted OR and CCR was 0.512 for this gene, the highest across the whole set. However, for some

genes, the relationship was not as strong leading to Rho estimates close to zero, such as; *DSC2* (ARVC), *PKP2* (ARVC), *RBM20* (DCM), *MYBPC3* (HCM) and *SCN5A* (LQTS).

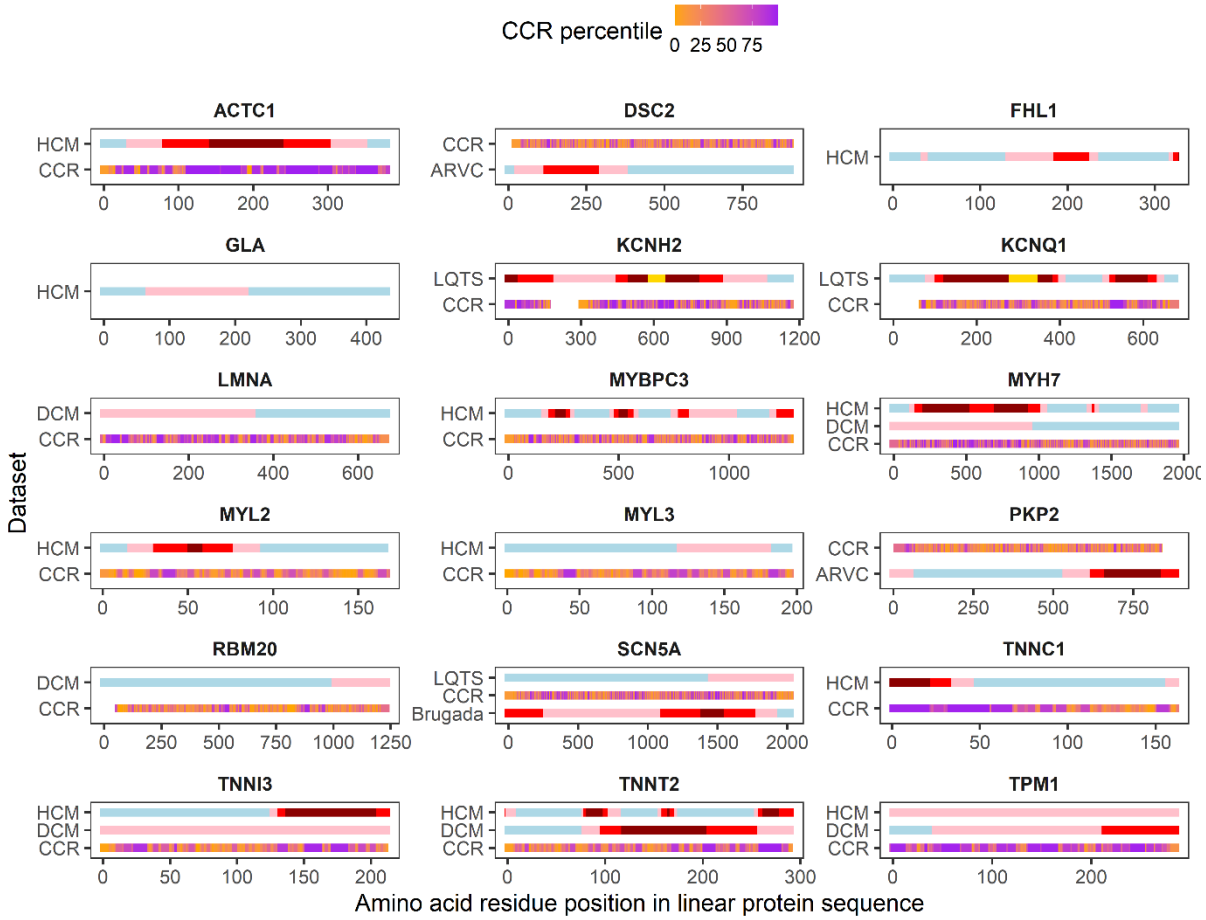

Figure 3.9: Comparison of estimated hotspots with constrained coding regions (CCRs) identified by Havrilla *et al.* 2019. CCRs are coloured by their percentile of all CCRs throughout the exome with darker purple corresponding to highly constrained regions. *Hotspot model* predictions are again stratified by their predicted ORs.

3.2.9 Overlap with known protein features

To determine whether the GAM hotspots in *n-significantly* clustered genes overlapped with known genomic features of interest such as; domains, experimentally defined regions, secondary structure and post-translational modifications were downloaded from UniProt (Figure 3.10). For most genes,

so many annotations were available that hotspots in almost any location would have had some overlap. However, overlapping features may indicate the disease mechanism for some hotspots.

The p.His733Asp-driven hotspot in the C-terminus of *PKP2* overlapped with the third ARM-repeat motif. This motif, named armadillo repeat, is found in many proteins, including plakophilins such as *PKP2*, the only plakophillin expressed in the heart (Bass-Zubek *et al.* 2009). Overlap of ARVC variants and these conserved ARM-repeats has been reported previously (Syrris *et al.* 2006). A small mutational hotspot was reported in residues 611-615 (Al-Jassar *et al.* 2013), and a founder variant p.C796R near this location was confirmed to cause ARVC by LoF effects and HI (Kirchner *et al.* 2012).

In *FHL1*, a mildly enriched region at the N-terminus (OR>=5) overlapped with a C4-type Zinc-Finger domain and the central mild and supporting (OR>=10) regions partially overlapped with the second LIM zinc-binding domains. However, the strongest hotspot in the C-terminal (OR>=20) did not overlap with any known protein features. Although no region achieved an OR above 5 in *FLNC* the C-tail region, where the largest density of case variants was found, it partially overlapped with multiple regions of interest. In HCM, for *MYH7*, as discussed previously, the main hotspot overlaps with the myosin motor domain but also the IQ domain in the head region of the protein. The smaller hotspot in the coiled-coil tail region of the protein overlaps with a structurally helical region but no known functional important motifs. In *MYBPC3*, the first strongest hotspot, as well as the final hotspot, overlap with Ig-like C2-type domains whereas the third hotspot overlaps both this domain and a Fibronectin type-III domain. The *MYL2* hotspot was located in the first EF-hand domain, whereas the *MYL3* hotspot was located in the second EF-hand domain. The C-terminal hotspot in *TNNI3* partially overlaps with a region involved in binding TNC and actin and is mostly helical in structure. Few genomic annotations were available for *TNNT2,* but the C-tail hotspot overlapped the only helical region of the protein. The strong hotspot in the *TNNC1* N-terminal domain partially overlapped with the first EF-hand domain in a structurally helical region.

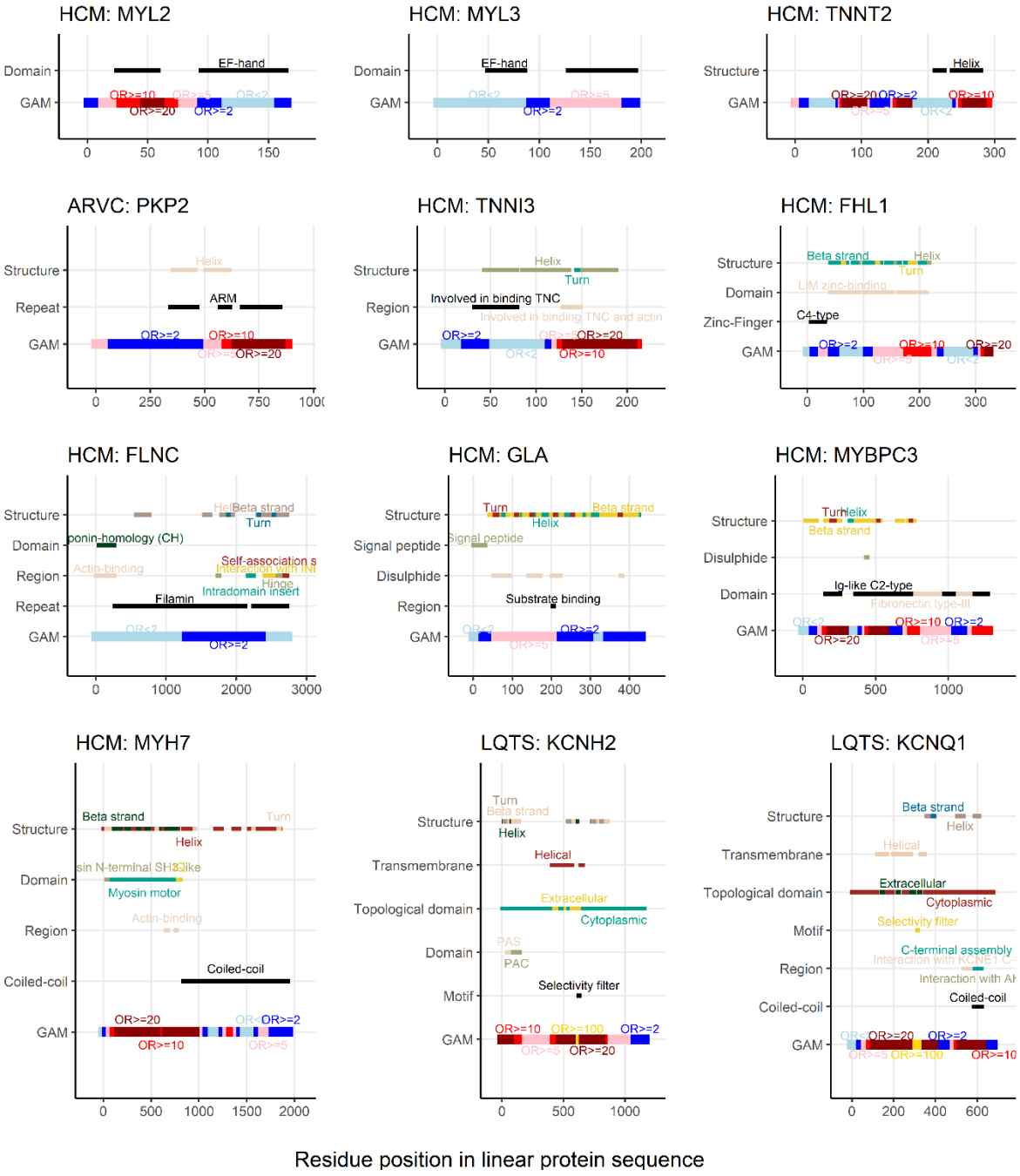

Figure 3.10: Overlap of estimated hotspots with genomic features downloaded from UniProt. UniProt feature type is labelled on the y-axis and includes; structure, domain, region and others. Features are labelled with their specific feature names within the plot e.g. for 'Structure' this may be 'Turn', 'Beta Strand' and 'Helix'.

In LQTS, for *KCNH2* the first hotspot overlapped with the PAS and PAC domains whereas the second hotspot overlapped many features including a transmembrane domain, intramembrane domain

and selectivity filter. For *KCNQ1,* the first hotspot was located in the helical transmembrane domain and overlapped the selectivity filter motif. The second hotspot overlapped a cytoplasmic region with multiple features of interest, including; the C-terminal assembly domain, and regions that interact with *KCNE1* and *AKAP*. Interestingly both strong hotspots in the potassium channel genes overlapped with the single selectivity filter motif in each gene.

It is unclear whether these overlaps are driven by chance or contain any generalizable or useful information. Several hotspots collocated with structurally helical regions or known domains. Correlation was assessed between the *hotspot model* predictions and overlap or non-overlap of these annotations. For these genes at least, there was a significant correlation between regions secondary structure and predicted ORs. Mean ORs for regions with no annotation (OR=6.24) and beta-strand (OR=6.81) were much lower than that for helical regions (OR=15.71). Similarly, mean ORs for regions overlapping (OR=12.03) and not-overlapping domains (OR=5.83) were significantly different. However, with a small sample size of genes, it is unclear if these results are generalizable or simply a summary of the investigated gene set.

3.2.10 Common and ClinVAR inside and outside of clusters

The counts of GnomAD variants with a *popmax* greater than 0.1% (*common_e2*) and likely pathogenic or pathogenic *clinvar* variants (*cln_pat*h), were compared in regions of differing predicted ORs (Table 3.6). Only gene-disease pairs with at least one predicted OR greater than ten across all residues were included. First genes were split into residues with an OR >= 10 (*or+10*) and residues with an OR < 10 (*or-10*). This was followed by further splitting the latter group into those with an OR between 10 and 20 (*or10-20*) and those with an OR >= 20 (*or+20*). In the first instance, when splitting regions at the OR-equals-ten threshold, there was an *n-significant* difference in the counts of *common_e2* variants in *MYH7*, *KCNH2* and *KCNQ1*, and a significant difference in the counts of *cln_path* variants in *PKP2*, *MYH7*, *TNNI3*, *TNNT2*, *KCNH2*, and *KCNQ1*. Overall there were

over 5.3 times as many *cln_path* variants in *or+10* regions and roughly 1.8 times as many *cln_path* variants in *or+20* regions compared to in regions with *or10-20* regions. Less *common_e2* variants were present in *or+10* regions compared to *or-10* regions (OR=0.7) and *or+20* regions relative to *or10-20* regions (OR=0.86).

Table 3.6: Counts of common and previously reported pathogenic variants inside and outside of clusters. The OR type indicates regions of the gene with predicted ORs between 10 and 20 (OR10) and above 20 (OR20), ignoring regions with ORs below 10. *N* and *Size* columns indicate the number of observations and number of residues inside and outside of clusters. *P-value* indicates a Fisher's-exact test p-value. '*Odds*' indicates odds-ratio and '*Lwr'* and '*Upr'* indicate 95% confidence intervals surrounding these.

				Inside cluster		Outside cluster					
Disease	**Gene**	**Dataset**	**OR type**	**N**	**Size**	**N**	**Size**	**P-value**	**Odds**	**Lwr**	**Upr**
ARVC	PKP2	ClinVAR	OR10	7	280	3	600	0.0157	4.99	1.13	30.1
ARVC	PKP2	ClinVAR	OR20	4	202	4	78	0.23	0.388	0.07	2.13
ARVC	PKP2	Common	OR10	46	280	126	600	0.207	0.782	0.53	1.14
ARVC	PKP2	Common	OR20	36	202	10	78	0.475	1.39	0.636	3.29
HCM	FHL1	ClinVAR	OR10	3	57	10	266	0.71	1.4	0.24	5.66
HCM	FHL1	ClinVAR	OR20	1	1	3	56	0.128	16.3	0.179	1470
HCM	FHL1	Common	OR10	5	57	16	266	0.556	1.46	0.401	4.38
HCM	FHL1	Common	OR20	1	1	5	56	0.183	10.3	0.119	886
HCM	MYBPC3	ClinVAR	OR10	17	431	25	844	0.409	1.33	0.667	2.6
HCM	MYBPC3	ClinVAR	OR20	3	166	17	265	0.0348	0.282	0.052	0.998
HCM	MYBPC3	Common	OR10	74	431	156	844	0.65	0.929	0.678	1.26
HCM	MYBPC3	Common	OR20	39	166	36	265	0.0307	1.73	1.02	2.92
HCM	MYH7	ClinVAR	OR10	112	1000	51	933	2.83E-05	2.04	1.44	2.94
HCM	MYH7	ClinVAR	OR20	96	677	16	326	3.46E-05	2.89	1.66	5.34
HCM	MYH7	Common	OR10	39	1000	86	933	7.73E-06	0.422	0.278	0.63
HCM	MYH7	Common	OR20	23	677	16	326	0.298	0.692	0.345	1.42
HCM	MYL2	ClinVAR	OR10	2	51	5	114	1	0.895	0.083	5.69
HCM	MYL2	ClinVAR	OR20	1	13	2	38	1	1.45	0.023	30.1
HCM	MYL2	Common	OR10	3	51	13	114	0.4	0.518	0.091	2
HCM	MYL2	Common	OR20	1	13	2	38	1	1.45	0.023	30.1
HCM	TNNI3	ClinVAR	OR10	21	84	6	127	0.0003	5.26	1.95	16.6
HCM	TNNI3	ClinVAR	OR20	21	73	2	11	0.73	1.58	0.306	15.7
HCM	TNNI3	Common	OR10	5	84	20	127	0.079	0.379	0.107	1.1
HCM	TNNI3	Common	OR20	5	73	2	11	0.261	0.382	0.054	4.48
HCM	TNNT2	ClinVAR	OR10	18	92	15	197	0.0118	2.56	1.16	5.73
HCM	TNNT2	ClinVAR	OR20	17	59	7	33	0.634	1.35	0.473	4.28
HCM	TNNT2	Common	OR10	11	92	26	197	0.853	0.906	0.387	2
HCM	TNNT2	Common	OR20	5	59	9	33	0.0753	0.314	0.076	1.15

LQTS	KCNH2	ClinVAR	OR10	145	676	11	484	2.30E-20	9.43	5.04	19.5
LQTS	KCNH2	ClinVAR	OR20	98	383	47	293	0.0158	1.59	1.08	2.39
LQTS	KCNH2	Common	OR10	59	676	90	484	2.22E-05	0.47	0.325	0.674
LQTS	KCNH2	Common	OR20	22	383	37	293	0.00594	0.455	0.25	0.812
LQTS	KCNQ1	ClinVAR	OR10	162	450	4	224	4.17E-20	20.1	7.56	75.7
LQTS	KCNQ1	ClinVAR	OR20	157	379	6	71	2.40E-05	4.89	2.08	14.1
LQTS	KCNQ1	Common	OR10	39	450	38	224	0.00765	0.511	0.309	0.847
LQTS	KCNQ1	Common	OR20	29	379	11	71	0.0744	0.495	0.227	1.15

OR	* ***OR10*** = (0-10)(10-Inf), ***OR20*** = (10-20,20-Inf)
Dataset	* ***ClinVAR*** = ClinVAR likely pathogenic/pathogenic, ***Common*** = GnomAD exomes popmax > 0.1%

In *MYH7*, only 39 *common_e2* variants were observed across the 1000 *or+10* residues compared to 86 observed in the remaining 933 residues ($p<7.7x10^{-6}$, OR=0.42). When splitting *or+10* regions, although not significant (p<0.298) counts of *common_e2* variants were again less in the *or+20* regions (OR=0.692). In the potassium channels *KCNH2* and *KCNQ1*, roughly half as many *common_e2* variants were found in *or+10* regions and after splitting again roughly half in *or+20* regions. The same pattern was also identified in the HCM genes *TNNT2* and *TNNI3* but was not as clear in *MYBPC3*, *MYL2* and *FHL1* and the ARVC gene *PKP2*. In *MYBPC3*, although not statistically significant, slightly fewer common variants were observed in *or+10* regions (OR=0.93) but when stratifying *or+10* group further, more common variants were found in *or+20* residues (OR=1.73). The same pattern is true of *MYL2* but only 16 *common_e2* variants were available for these comparisons. *FHL1* was the biggest outlier as the only gene with more *common_e2* variants in *or+10* regions. An analysis of the distribution of *common_e2* variants and rare case variants using the *BIN-test* found at least *n-significant* signals across all genes with patterns of clustering consistent with earlier comparisons to *gnomad_e2* rare variants.

In *PKP2*, seven *cln_path* variants were observed in the 280 residues in *or+10* regions and only three were observed in the 600 *or-10* regions (p<0.016, OR=5). There were over double as many *cln_path* variants in the *or+10* region of *MYH7* ($p<2.8x10^{-5}$, OR=2). Subdividing this region further revealed

an excess in *or+20* regions ($p<3.5x10^{-5}$, OR=2.89). Over 5 times as many *cln_path* variants were observed in *or+10* regions in *TNNI3* (p<0.0003, OR=5.3), with 21 observed in 84 *or+10* residues and only 6 observed in 127 *or-10* residues. In *TNNT2*, *or+10* regions had more than twice as many *cln_path* variants (p<0.01, OR=2.56) and *or+20* regions were moderately enriched (OR=1.35). The LQTS potassium channel genes were massively enriched for *cln_path* variants with ORs of 9.4 and 20.1, respectively. For both genes, further enrichment was found in *or+20* regions.

3.3 Integrating *in-silico* algorithms

3.3.1 Overview

The ACMG criterion PP3 covers all *in-silico* evidence available for the assessment of variant pathogenicity (Table 3.7). It is applied as supporting evidence and only when all algorithms support a pathogenic classification. As there is no guidance on which tools to consider or which thresholds determine pathogenicity, there is discordance between testing laboratories when applying this criterion (Ghosh *et al.* 2017). However, with large datasets of case variants these scores can be tuned to the disease and gene of interest, and uncertainty boundaries can be estimated. This allows you to answer the questions; which scores are meaningful for my disease/gene? How does the prediction score correlate with the pathogenicity of variants? How confident are we that these scores demonstrate pathogenicity?

As GAM models are additive in nature, they are designed to cope with an arbitrary number of additional covariates, providing the number of covariates is not greater than the number of observed data points. Therefore, the predictive utility of including scores from variant prediction algorithms was investigated. For each observed variant in our cases and *gnomad_e2* controls, features were annotated with twenty-nine variant prediction scores, freely available from the dbNSFP v4.0 database (Liu et al. 2016), using Ensembl's Variant Effect Predictor tool (VEP; McLaren

et al. 2016). These *in-silico* predictive features were included alongside carrier status and residue position to produce our *hotspot+* models. The structure of the *hotspot+* models is as follows;

$$ln(P/1-P) = \beta 0 + \beta 1\ carrier + s1\ (pos,\ by=carrier) + s2\ (dbnsfp^{1},\ by=carrier) + \ldots + s3(dbnsfp^{n},\ by=carrier) + \varepsilon$$

This structure is an extension of the *hotspot* models with additional smooth functions for *n* dbNSFP features included in the model.

Table 3.7: Predictors from dbNSFP used in the final *hotspot+* models. Predictor classes are approximate and not mutually exclusive as some predictors draw from more than one. Meta = meta-predictor (combined predictions from other methods using a machine learning model), Sequence-based = sequence model (in this case hidden-Markov model) where predictions are influenced by adjacent positions, Conservation = conservation of nucleotide or residue position between species, Constraint = intra-species constraint assessed from human reference cohorts, Structure = considers protein structure. Classes are

Predictor	**Class**	**Citation**
CADD	Meta	Kircher *et al*. 2014
Deogen2	Meta	Raimondi *et al*. 2017
FATHMM	Sequence-based	Shihab *et al*. 2013
GERP	Conservation	Davydov *et al.* 2010
MetaLR or MetaSVM	Meta	Dong *et al.* 2015
MPC	Constraint	Samocha *et al*. 2017
MutationTaster	Meta	Scwartz *et al.* 2014
MutationAssessor	Conservation	Reva *et al*. 2007
MVP	Meta	Hongjian *et al*. 2018
PhastCons (vertebrate, mammalian, primate)	Conservation	Siepel *et al*. 2005
Polyphen	Structure	Adzhubei *et al.* 2013
PrimateAI	Meta	Sundaram *et al*. 2018
Provean	Conservation	Choi *et al*. 2012
REVEL	Meta	Ioannidis *et al*. 2016
SiPhy 29way	Conservation	Garber *et al.* 2009
VEST4	Meta	Carter *et al*. 2013

3.3.2 Feature selection

Although maximum predictive performance would be attained by including all dbNSFP scores into each model, this also increases the risk of overfitting, and for many genes with low observed counts, prevents models from fitting as there are fewer unique covariate combinations than observations. Therefore, a strict two-stage feature selection process was implemented. In step one, the marginal association of each feature was calculated in a GAM model containing only that feature as a predictor. Only features with a model p-value below alpha (0.05), Bonferroni corrected for the total number of dbNSFP scores, was assessed in stage 2 ($p<0.00167$). In the second stage, a backwards elimination routine was implemented, whereby in each iteration, if any predictors had a conditional p-value above the alpha, Bonferroni corrected for the number of features that passed stage 1, the feature with the highest p-value was removed from the model. This process was repeated until all remaining features had p-values lower than the significance threshold. Because of the stringency of this approach, for some genes no additional features, beyond protein position, were included in the model (Table 3.8).

Table 3.8: Additional predictors passing feature selection for genes that were *hotspot+* modelled. Genes without additional features (beyond carrier status and residue position) are not displayed.

Disease	Gene	Additional predictors for *hotspot+* model
HCM	*FLNC*	REVEL
HCM	*GLA*	Polyphen2
HCM	*MYBPC3*	MutationTaster, MutationAssessor, FATHMM, Provean, MetaSVM, MetaLR, REVEL, PrimateAI
HCM	*MYH7*	Polyphen2, MutationAssessor, REVEL, MVP, Deogen2, CADD, GERP NR
HCM	*MYL2*	MutationAssessor, vest4
HCM	*TNNI3*	MPC, GERP NR
HCM	*TNNT2*	PrimateAI
DCM	*LMNA*	MetaSVM
DCM	*MYH7*	REVEL
DCM	*RBM20*	REVEL
ARVC	*PKP2*	MutationTaster, FATHMM, PrimateAI
LQTS	*KCNH2*	MutationAssessor
LQTS	*KCNQ1*	REVEL
LQTS	*SCN5A*	PrimateAI

For some algorithms, not only do disease predictions differ between cases and controls, but disease predictions differ between different case groups (Figure 3.11). For example, in PrimateAI, the distributions of HCM and *gnomad_e2* are fairly similar, although HCM has fewer low values leading to a higher mean score (0.72) compared to *gnomad_e2* (0.63). For ARVC, DCM and LQTS, the distribution of scores peak at different locations, with the highest scores resulting from LQTS variants. In fact, for many algorithms (e.g. FATHMM, MetaLR, REVEL, Deogen2), distributions of LQTS variant scores are visually separable from the remaining datasets. Four genes have a *b-significant* burden in LQTS, but most variants are in the two strongly associated genes; *KCNQ1* and *KCNH2*. It is possible that there is a bias in these algorithms that better detects the type of pathogenic variants in these genes, especially as they cluster in specific structural and functional regions, or alternatively, there could be a higher ratio of pathogenic to benign variants in these four genes. Given the substantially high ORs in *KCNQ1* (OR=45) and *KCNH2* (OR=23), this is a viable possibility. Often ARVC is visually separable from the other datasets and is based on only one gene; *PKP2*. This could be driven by the few variants in this gene, including the high-frequency p.His733Asp variant.

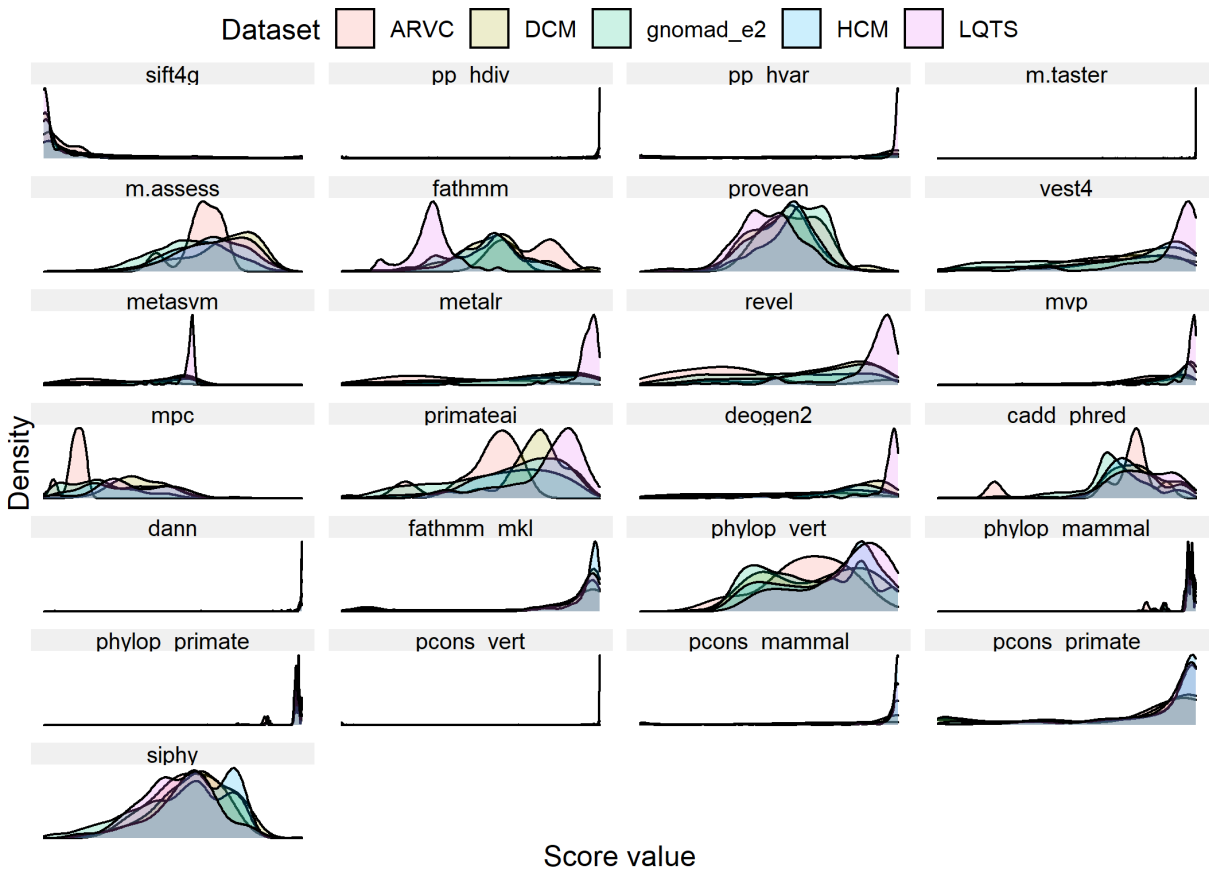

Figure 3.11: Distribution of *in-silico* algorithm predictor values for each disease dataset (ARVC, DCM, HCM, LQTS) and *gnomad_e2* controls. For each dbNSFP predictor (each subplot), and each cohort, the Gaussian density of variant scores are displayed. Only variant-scores in firmly associated genes were considered for each dataset. For *gnomad_e2*, all variants in all genes considered in the disease cohorts were included.

The extent of correlation between predictor scores and case (1) control (0) status of observed rare-missense variants was investigated. In HCM (Figure 3.12), there was a clear pattern of increased correlation between meta-predictors and cohort status. The best in terms of mean correlation across genes were all meta-predictors; VEST4, REVEL and MVP. The algorithms with the next highest correlation were Polyphen2 (hvar) and Polyphen2 (hdiv). Although not specifically meta-predictors, as they do not use the results from other prediction algorithms, they do draw from a very large and diverse pool of features. The scores with the least correlation were mostly conservation scores but also FATHMM which has been demonstrated to have poor performance when distinguishing pathogenic and non-pathogenic variants in the same gene. All findings were

consistent with previous investigations on the performance of *in-silico* prediction scores across a wide-set of pathogenic and benign variants (Ghosh *et al.* 2017).

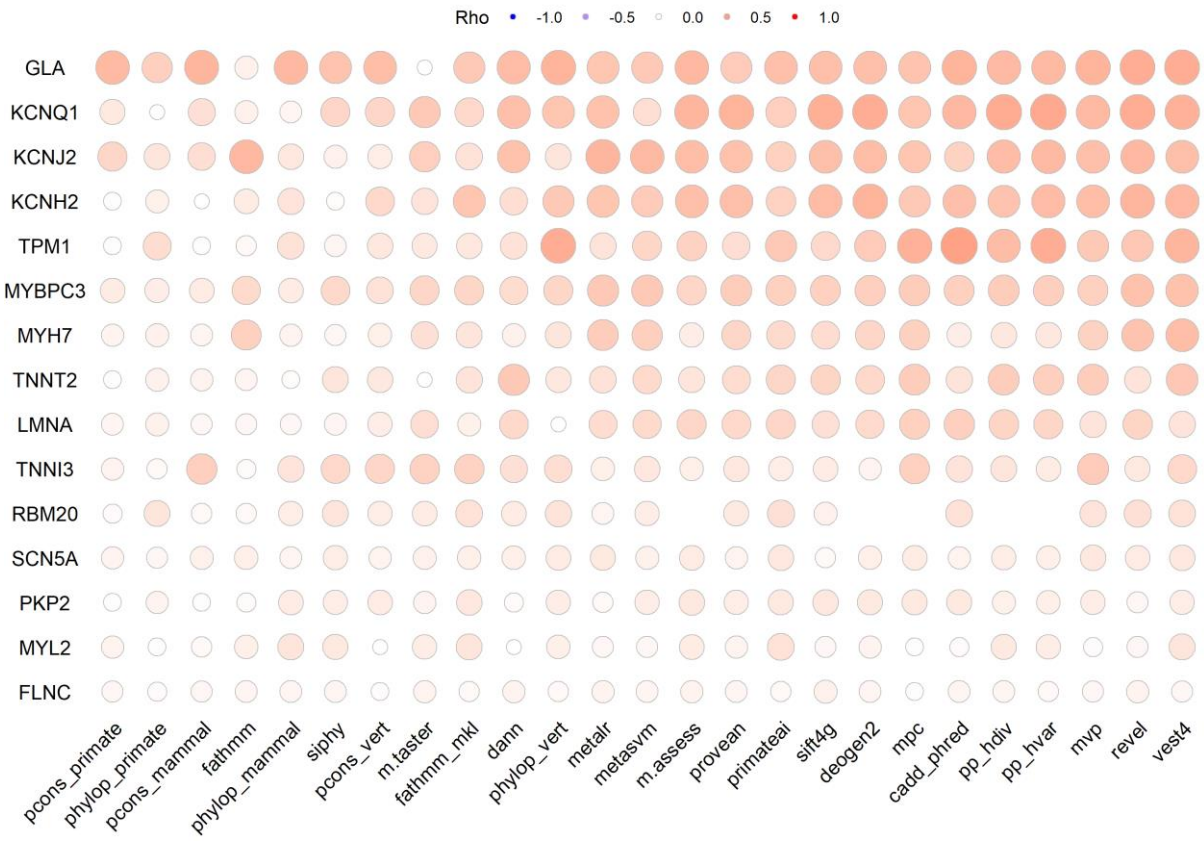

Figure 3.12: Correlation between missense-variant pathogenicity-score and case-control status. Cases included all variants in significantly overburdened heart-disease genes, controls included all *gnomad_e2* variants in the same genes. dbNSFP predictors and genes are ordered by their total correlation (Rho) across each column or row. VEST4 scores have the highest correlation with case-control status in this experiment and *GLA* variants have the highest correlation with case-control status across all scores.

The extent of collinearity between different predictor scores was assessed by their Spearman rank correlation coefficients (Figure 3.13). Unsurprisingly, the meta-predictors had the highest correlation with other algorithms. Conversely, FATHMM had the least correlation with other algorithms. One notable finding was that the algorithm MPC, which relies partially on the regional constraint, was not highly correlated with other predictor scores but was reasonably highly

correlated with cohort status of observed variants. However, as it uses *gnomad_e2* data in its training set, there is also type-1 circularity here (Subsection 1.2.5). Although collinearity indicates redundancy in the feature set, it is not necessarily a problem for the model fitting process, provided there is sufficient data. However, for collinear features, or in the case of GAM smooth functions, concurved features, coefficient weights are split arbitrarily between the correlated features. This may make it more complicated to interpret which features are the most important, as coefficients are conditional on other features in the model.

Figure 3.13: Correlation between *in-silico* prediction algorithms based on their scores for variants in the inherited heart disease genes. Scores are ordered by their total correlation across all gen genes. REVEL was the most correlated with other pathogenicity scores (reflecting its meta-predictor status).

3.3.3 Hotspot+ models

Hotspot+ models were generated for all disease-gene pairs with at least *n-significant* burden signals. Predictors for these models included both burden and residue position, as with the *hotspot models*, as well as additional features passing the strict feature selection procedure. The resulting models are displayed in Figure 3.14 for each gene. Plotting against the protein position on the x-axis reveals predictions still largely follow the trend of the positional signal. However, the increased information from pathogenicity prediction algorithms allows the predictors to fall above or below this line. Predictions can now vary even at the same residue for different alternative amino acids, giving higher resolution.

For *MYBPC3*, odds-ratio predictions range from 0.00591 to 566, substantially wider than for the original *hotspot model* which ranged from 0.588-34.7. The same was seen in *MYH7*, where the range of predictions now vary from 0.0674 to 830, as opposed to 0.683-59.1. The result of this is that more variants now reach the threshold for strong evidence (OR>=100). The model is now drawing relevant information from two ACMG criteria PM1 and PP3. If we collapse these into one criterion then many more variants are given; supporting, moderate or strong evidence. For each model, manual variant classifications were stratified by evidence levels (Table 3.9).

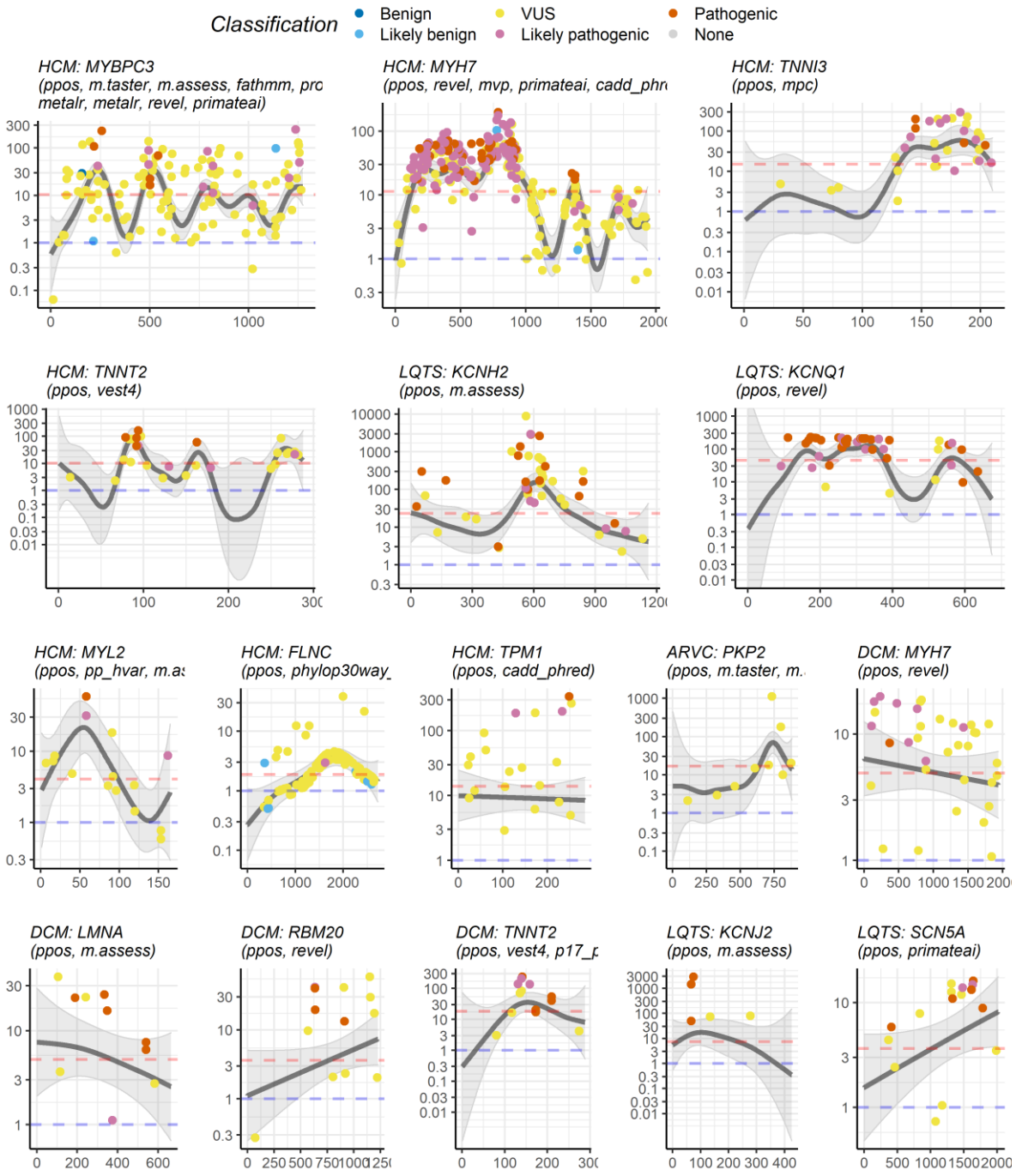

Figure 3.14: Predicted ORs for all case-variants in the *hotspot+ modelled* genes. Red lines indicate the gene-level odds-ratio (the case excess of rare-missense variants relative to *gnomad_e2*), blue lines intersect the y-axis at one and grey ribbons display the 95% confidence intervals. Training data from the case-cohort are plotted and coloured by their pathogenicity classification.

For many VUS variants, the *hotspot+ models* provided evidence of pathogenicity. In *PKP2*, of the 17 samples carrying a VUS, twelve were provided with strong evidence that it was the causal variant. In DCM, 5/7 *LMNA* VUS carriers were given moderate evidence, 11/29 *MYH7* VUS carriers were given supporting evidence and 5/10 *RBM20* VUS carriers were given supporting or moderate evidence. Overall, out of 21 LP or P variants in these DCM genes, four were given moderate evidence, and nine were given supporting evidence. With the remaining 8 given no evidence in these models (i.e. predicted OR<10). In *MYH7* and *MYBPC3* VUS carriers in the HCM cohort, many samples are now provided with strong evidence that it is the causal variant. In *MYBPC3*, 177 samples received a VUS classification, 13 of these now can be provided with strong evidence and 82 with moderate evidence. For the 191 *MYH7* VUS carriers, one was provided with strong evidence and 54 with moderate evidence. There were 523 VUS carriers across the HCM genes; of these, 181 were provided with either strong or moderate evidence of pathogenicity. Of the 624 LP or P variants, 523 were provided with moderate or strong evidence.

In LQTS, genes were not provided with classifications by OMGL, so classifications here are from *clinvar* and are not necessarily disease-specific. However, 93/109 LP and P variants were given moderate or strong evidence, and 21/41 VUS variants were given moderate or strong evidence. To take into account confidence in model predictions, variants were stratified by their lower 90% (one-sided) confidence intervals instead of point estimates. These correspond to a 90% chance that the true OR is greater than this value. Although fewer samples were given evidence using this conservative approach, many were still provided with supporting (n=228), moderate (n=510) and strong evidence of pathogenicity (n=138).

Table 3.9: Evidence categories for each observed case variant. Only cells where at least one variant has supporting evidence are displayed. Evidence levels in a:b:c:d format where a = none, b = supporting, c = moderate and d = strong. For example, 1:0:0:1 corresponds to one variant with no evidence and one with strong.

A) Point	**Classification**		
Disease: Gene	*VUS*	*LP*	*P*
ARVC: *PKP2*	3: 3: 1: 12		
DCM: *LMNA*	2: 0: 5: 0		3: 1: 2: 0
DCM: *MYH7*	18: 11: 0: 0	2: 6: 0: 0	2: 0: 0: 0
DCM: *RBM20*	5: 1: 4: 0	0: 0: 1: 0	0: 2: 1: 0
HCM: *FLNC*	105: 3: 5: 0		
HCM: *GLA*	4: 9: 0: 0	2: 1: 0: 0	0: 1: 0: 0
HCM: *MYBPC3*	51: 31: 82: 13	5: 16: 52: 1	0: 1: 35: 74
HCM: *MYH7*	71: 36: 54: 1	27: 20: 113: 5	0: 20: 152: 1
HCM: *MYL2*	12: 1: 0: 0	1: 0: 1: 0	0: 0: 9: 0
HCM: *TNNI3*	4: 4: 6: 5	0: 3: 26: 9	0: 0: 6: 15
HCM: *TNNT2*	9: 2: 13: 2	3: 0: 3: 0	0: 0: 17: 4
LQTS: *KCNH2*	5: 2: 7: 11	2: 0: 2: 2	1: 1: 4: 14
LQTS: *KCNQ1*	3: 1: 2: 1	0: 0: 7: 12	1: 0: 7: 45
LQTS: *SCN5A*	6: 3: 0: 0	0: 2: 0: 0	6: 3: 0: 0

B) Lower 90%			
Disease: Gene	*VUS*	*LP*	*P*
ARVC: *PKP2*	7: 0: 1: 11		
DCM: *LMNA*	2: 5: 0: 0		4: 2: 0: 0
DCM: *MYH7*	26: 3: 0: 0	5: 3: 0: 0	
DCM: *RBM20*	6: 3: 1: 0	0: 1: 0: 0	2: 1: 0: 0
HCM: *FLNC*	111: 2: 0: 0		
HCM: *MYBPC3*	89: 53: 35: 0	21: 22: 31: 0	1: 7: 43: 59
HCM: *MYH7*	96: 30: 36: 0	30: 37: 96: 2	0: 35: 137: 1
HCM: *MYL2*		1: 1: 0: 0	0: 0: 9: 0
HCM: *TNNI3*	9: 2: 8: 0	3: 10: 24: 1	0: 2: 19: 0
HCM: *TNNT2*	17: 1: 8: 0	5: 0: 1: 0	0: 0: 21: 0
LQTS: *KCNH2*	7: 2: 8: 8	2: 1: 2: 1	2: 1: 11: 6
LQTS: *KCNQ1*	4: 0: 3: 0	1: 3: 5: 10	2: 1: 11: 39

Evidence levels in a:b:c:d format (a = none, b = supporting, c = moderate, d = strong)

3.3.4 *Stratification by pathogenicity classifications*

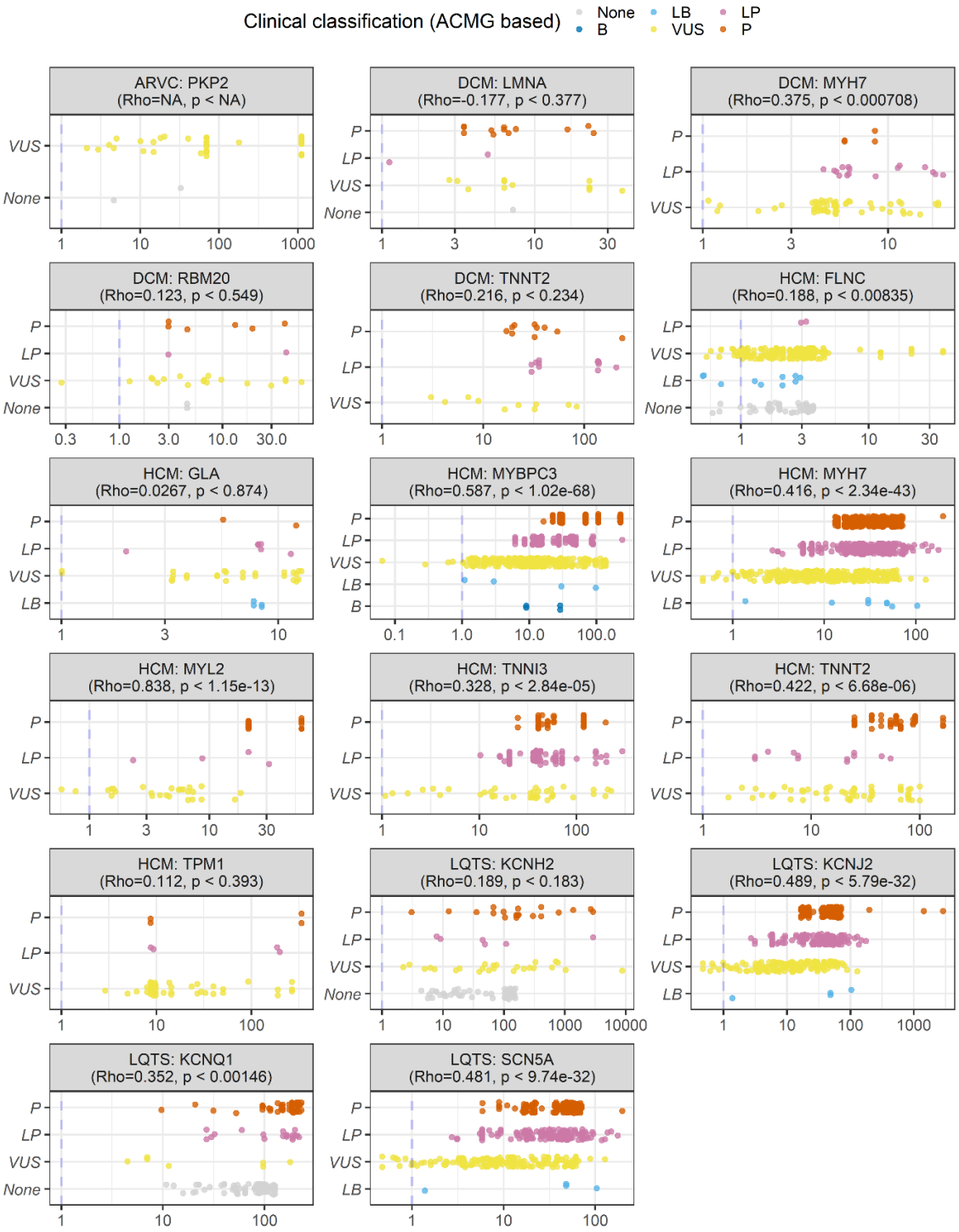

Figure 3.15: *Hotspot+ model* predictions for each case-variant, stratified by their clinical classifications. P = Pathogenic, LP = likely pathogenic, VUS = variant of uncertain significance. Where 'None' is indicated, this means the specific variants have neither been classified by OMGL nor submitted to ClinVAR with a classification.

Hotspot+ predicted ORs were stratified by expert classifications (Figure 3.15) made using the ACMG framework by clinical scientists at OMGL (or if unavailable – *clinvar*). A positive correlation was observed in all genes, excluding *PKP2,* which had only VUSs. The best performance was observed in *MYL2* (Rho=0.838, $p<1.2x10^{-13}$), which showed a clear distinction between VUS, LP and P variants. A reasonably strong correlation was also observed in the HCM for genes; *MYBPC3* (Rho=0.59), *MYH7* (Rho=0.42), *TNNT2* (Rho=0.42) and *TNNI3* (Rho=0.33), as well as in DCM for *MYH7* (Rho=0.38), and LQTS for *KCNJ2* (Rho=0.49), *SCN5A* (Rho=0.48) and *KCNQ1* (Rho=0.35).

3.3.5 Prediction metrics

Prediction metrics were recalculated with the addition of these new features (Table 3.10). Fifteen out of sixteen models had improved accuracy compared to the *hotspot models*. The HCM model for *TNNT2* had reduced accuracy. Most likely, this was due to overfitting. The models with the best performance boost were those with the weakest *hotspot models* including; DCM *LMNA*, DCM *RBM20*, HCM *TPM1* and LQTS *KCNJ2*. For *MYBPC3*, accuracy improved by 82.3%, a twelve percent increase on the accuracy of the *hotspot model*. In *MYH7*, there was a 9.1% increase in the accuracy of the model.

Table 3.10: Model accuracy metrics for *hotspot+ models* and the relative different between those scores and those reported earlier for *hotspot models*. TPR = true-positive rate, TNR = true-negative rate, FPR = false-positive rate, FNR = false-negative rate.

Disease	Gene	Accuracy		TPR		TNR		FPR		FNR	
ARVC	*PKP2*	0.79	(+0.049)	0.789	(+0.052)	0.79	(+0.049)	0.21	(-0.049)	0.211	(-0.052)
DCM	*LMNA*	0.786	(+0.207)	0.786	(+0.215)	0.786	(+0.207)	0.214	(-0.207)	0.214	(-0.215)
DCM	*MYH7*	0.666	(+0.093)	0.667	(+0.129)	0.666	(+0.092)	0.334	(-0.092)	0.333	(-0.129)
DCM	*RBM20*	0.766	(+0.123)	0.786	(+0.143)	0.766	(+0.123)	0.234	(-0.123)	0.214	(-0.143)
DCM	*TNNT2*	0.752	(+0.064)	0.75	(+0.062)	0.752	(+0.064)	0.248	(-0.064)	0.25	(-0.062)

HCM	*FLNC*	0.637	(+0.033)	0.636	(+0.033)	0.637	(+0.033)	0.363	(-0.033)	0.364	(-0.033)
HCM	*GLA*	0.698	(+0.066)	0.684	(+0.052)	0.701	(+0.069)	0.299	(-0.069)	0.316	(-0.052)
HCM	*MYBPC3*	0.817	(+0.114)	0.816	(+0.112)	0.817	(+0.115)	0.183	(-0.115)	0.184	(-0.112)
HCM	*MYH7*	0.777	(+0.073)	0.776	(+0.07)	0.777	(+0.075)	0.223	(-0.075)	0.224	(-0.07)
HCM	*MYL2*	0.744	(+0.012)	0.792	(+0.042)	0.736	(+0.007)	0.264	(-0.007)	0.208	(-0.042)
HCM	*TNNI3*	0.807	(+0.045)	0.859	(+0.026)	0.774	(+0.056)	0.226	(-0.056)	0.141	(-0.026)
HCM	*TNNT2*	0.792	(-0.034)	0.792	(-0.019)	0.792	(-0.04)	0.208	(+0.04)	0.208	(+0.019)
HCM	*TPM1*	0.791	(+0.273)	0.8	(+0.333)	0.788	(+0.25)	0.212	(-0.25)	0.2	(-0.333)
LQTS	*KCNH2*	0.823	(+0.097)	0.824	(+0.099)	0.823	(+0.097)	0.177	(-0.097)	0.176	(-0.099)
LQTS	*KCNJ2*	0.98	(+0.376)	1	(+0.397)	0.979	(+0.375)	0.021	(-0.375)	0	(-0.397)
LQTS	*KCNQ1*	0.811	(+0.132)	0.797	(+0.113)	0.814	(+0.136)	0.186	(-0.136)	0.203	(-0.113)
LQTS	*SCN5A*	0.75	(+0.146)	0.75	(+0.147)	0.75	(+0.146)	0.25	(-0.146)	0.25	(-0.147)

3.3.6 Comparison to dbNSFP predictors

The next goal was to determine whether our data-driven position-based modelling approach gave superior predictive performance than general pathogenicity or functional predictors. This was achieved by comparing the *hotspot+* models to the models generated by single *in-silico* predictors available from dbNSFP. The individual models were also generated by GAMS. For each gene, the mean and standard deviation of the area under the curve was calculated across 10-cross fold validations (CV-folds). Cross-fold validation is a method to prevent overfitting without splitting data into a training set and a testing set. Multiple models are generated with different samples withheld for testing each time. Here 80:20 splits were used, which keeps 80% of the data for training and tests on the remaining 20%. This was done for ten different subsets of training and testing data (Figure 3.16).

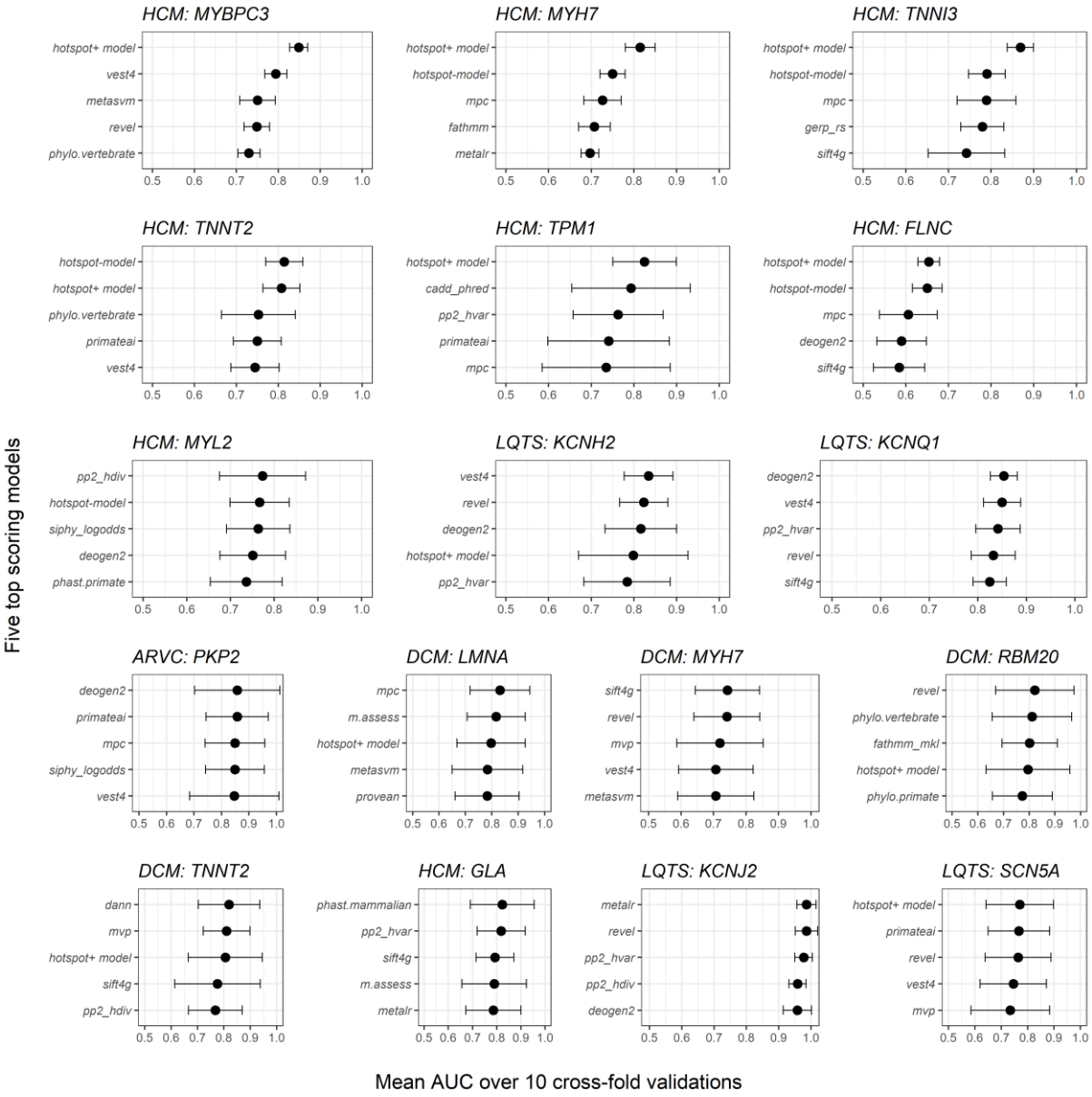

Figure 3.16: Mean and standard deviation of AUROC scores across 10 cross-fold validations. The predictive performance of *hotspot* and *hotspot+ models* were compared against models trained with a single-feature from dbNSFP.

The metric used for comparison was the receiver-operating characteristic (ROC) area under the curve (AUC). It calculates specificity and sensitivity at different thresholds for converting the logistic model probability output into a case-control binary outcome. For binary outcomes, random guesses, proportional to the number of cases and controls, should give an average AUC of 50%. Therefore, accuracy above this minimum suggests a predictive performance greater than expected

by chance. The highest ROC AUC (AUROC), is 1, which denotes 100% accuracy. The CV-fold approach provides insight into the generalizability of a model. Here, the mean AUROC across CV-folds indicates a model performance adjusted for model variance (i.e. overfitting).

For many genes, the *hotspot+ model* did give a clear performance boost compared to other *in-silico* prediction scores. This was most notable in the HCM genes, especially *MYBPC3* and *MYH7*, but also *TNNI3*, *TNNT2*, *TPM1* and *FLNC*. For the remaining genes, AUROC standard deviations are generally wide and differences in mean AUROC between approaches are small. This indicates more data is needed to train these models. Interestingly, for many genes, the *hotspot model* gave superior performance than any generic *in-silico* prediction algorithms. This provides evidence of the importance of this gene-specific metric for variant interpretation.

3.4 Discussion

3.4.1 Conceptual underpinnings of the hotspot models

Hotspot-models were created to identify local mutational hotspots in inherited heart disease genes. These hotspots are of interest as they represent an increased prior probability for pathogenicity of variants observed in clinical cases. They therefore have implications for the interpretation of rare-missense variants in a clinical context but are also relevant to our understanding of disease pathogenesis.

Many previous approaches to *in-silico* pathogenicity prediction were developed based on complex supervised or semi-supervised machine learning approaches such as support vector machines (Kircher *et al*. 2014, Dong *et al.* 2015), random forests (Carter *et al.* 2013), ensemble models (Ioannidis *et al.* 2016) or deep learning models (Hongjian *et al.* 2018, Sundaram *et al.* 2018). These approaches rely on huge amounts of training data of putatively benign or pathogenic variants. For the purposes of hotspot estimation, as it is gene-specific, the same magnitude of training data is

not available. Furthermore, model interpretability, not only is predictive performance of interest here, but also inference. For this reason, generalized-additive models were used as a parsimonious and interpretable modelling strategy to both infer mutational hotspots and predict variant pathogenicity.

Coverage was identified as a potential confounder for the identification of mutational hotspots. Although the ultimate solution would be to perfect sequencing technology to attain 100% reliable calls. In lieu of this, regions were weighted by local coverage, with the intention of up-weighting regions of low coverage. As with the *BIN-test* inverse-coverage weighting procedure, there were some limitations to this approach. Firstly, a region was only up-weighted by increasing the contribution to the log-likelihood of variants in that region. Therefore, if no variants were observed in a region, it was not up-weighted. For this reason, all regions with less than 50% of samples with at least 10X coverage were completely excluded from the analysis. However, where rare-variant data is already very sparse, this approach is suboptimal and may fail to up-weight some regions effectively. Furthermore, by up-weighting only observed variants, the true location of missing variants are not captured, leading to increased variance in the model.

The outcome variable used in this analysis was affection status i.e. whether each sample is a case or control. However, in practice, the affection status is known, and interest lies in determining whether the variant is pathogenic or benign, so predictions can be made before clinical onset of the disease. An alternative approach would be to model an outcome based on classifications of pathogenic and benign variants. However, there are limitations to this latter strategy. Clinical classifications of variant pathogenicity are imperfect, and misclassifications add systematic-reinforcing bias to the models. Furthermore, many variants are VUSs and there is not a clear way to incorporate these into pathogenic-benign training data. Finally, it becomes difficult to integrate burden information. For strongly associated genes, the high burden of case variants suggests that most variants are associated with the disease. It is therefore useful to capture this information.

Another consequence of modelling a sample-level outcome variable was that to reduce noise, only rare-variants which have a high prior probability of being pathogenic were included in the training data. This does mean that common control variants are excluded from the analysis. These variants are very likely to be benign and therefore contain useful information regarding background variation of neutral variation and may help detection of mutational hotspots. In fact, in subsection 3.2.10, common GnomAD variants were observed less frequently in identified mutational hotspots. Despite this, it is unclear how to include this information while retaining model interpretability with respect to the regional burden. For a model where the desired output is not an estimate of burden, the model could potentially include all putatively benign variants including *gnomad_e2* control variants of any frequency. In this scenario, unless each variant is included in the model a single time regardless of frequency, a much more sophisticated weighting strategy would have to be implemented to prevent common variants from dominating the model.

To model residue position using GAM smooth functions, the discrete residue position must be considered as a continuous covariate. This approach makes the assumption that the relationship between pathogenicity and affection status varies continuously and smoothly across the protein-coding region. Therefore, the pathogenicity of neighbouring residues are expected to be similar, but also the penetrance of neighbouring residues are expected to be similar. This is in contrast to the discrete model usually applied to genetic sequences of interest such as genes, promoters and enhancers; however, it is unclear whether the same discrete model applies to pathogenicity potential within a protein. The discrete model in this scenario models the hypothesis that there is a specific disease-causing functional domain with variants equally likely to cause disease at any point within this domain. For example, in previous studies of mutational hotspots in HCM (Walsh *et al.* 2019), specific genetic ranges are identified in the genes and characterised by a uniform measure of pathogenicity. The correct approach is dependent on how neighbouring residues in a protein affect protein function. A discrete model would suggest that specific groups of contiguous residues all affect protein function in exactly the same way. Alternatively, the continuous approach

posits that the effect of variants is dependent on how close they are to a functionally or structurally important region. Although the latter is a plausible hypothesis, the correct approach could only be validated with a huge amount of data or detailed experimental functional analysis.

3.4.2 GAM hotspots versus 1dC hotspots

Hotspot identification was a problem simultaneously tackled by Walsh *et al.* (2019). Their approach (1dC), based on a partially overlapping dataset, had some conceptual differences between *hotspot-models*. These include case-only scanning, boundary trimming and uniform allocation of pathogenicity. 1dC compares inside-window case-counts to outside-window case-counts within sliding windows across the protein. Multiple problems may arise with this approach. Even when the window does focus on a pathogenic region, the outside-window regions may not be neutral due to other pathogenic regions in the protein. In this case, the outside-window counts are skewed, decreasing the relative difference and reducing power to detect hotspots. This may be the case for *TNNT2*, where multiple clusters are detected in the *hotspot models* but only one large central cluster was detected by 1dC (Figure 3.8).

The 1dC approach may also miss less distinct hotspots. For example, in the more uniformly burdened *TNNC1* gene, GAM *hotspot modelling* detects an N-terminal moderate hotspot, and in the weakly clustered *ACTC1* gene, GAM detects a moderate central hotspot (Figure 3.7). Due to the relatively modest case-variant clustering, it would have been difficult to detect those using 1-dimensional scans. However, regions of substantial burden relative to controls exist. For both these genes, hotspots are identified when considering both case variants and the relative depletion of rare-missense variants in controls. Both examples corroborated by high-levels of intra-species constraint (Table 3.5). Similarly, for low penetrance HI genes like *FLNC*, where the GAM approach suggests a C-terminal cluster that may inspire more analysis into this novel gene, a 1-dimensional clustering approach is unlikely to reveal this relationship.

Another key difference between the approaches is the strength of evidence proposed for identified hotspots. Walsh *et al.* (2019) suggest an etiological fraction of 0.95 to ascribe strong evidence of pathogenicity, whereas moderate evidence is recommended for an equivalent probability, here (OR>=20). Due to the non-trivial amount of background variation observed, a conservative threshold is recommended. 1dC also makes the assumption that all variants within an identified cluster share the same etiological fraction. As previously discussed for *TNNT2*, the 1dC central cluster is split into two clusters using GAMs (Figure 3.8). These two clusters reach ORs over 20, but the small region between them has an empirical OR of only 3. The 1dC approach estimates an etiological fraction of 0.95 and provides strong evidence for this entire region, whereas the GAM approach correctly identifies only low enrichment in this region.

3.4.3 Meaningful hotspot overlaps

Identified hotspots were compared with constrained-coding regions (CCRs) identified by Havrilla *et al.* (2019). As the *gnomad_e2* rare-variants formed part of the training set to identify these regions of intra-species constraint, there was inevitably some overlap based on this alone. However, the estimates were also based on additional exomes and genomes as well as common variants. The estimates were also controlled for both coverage and mutation rate. In this sense, these intra-species constraint measures provided some augmenting information. Correspondence between *hotspot models* and CCRs were strong in some genes and weak in others (Table 3.5; Figure 3.9). It is easy to see why they would be correlated, especially as many hotspots were accompanied by depleted regions in controls. However, it is also somewhat intuitive why they may not overlap. Firstly, while both approaches are gene-specific, only the *hotspot models* are disease-specific. Many of the genes investigated are associated with multiple disease phenotypes, including many genes (e.g. *MYH7*, *TNNT2*, *TNNI3*, *and TPM1*) that are associated with both HCM and DCM. While the CCRs aggregate gene regions associated with each of these diseases, the *hotspot models* focus in on the disease-relevant region. The highest CCRs, regions of extreme constraint, are most likely

associated with embryonic lethality or other high-mortality phenotypes. The inherited heart diseases investigated are not as lethal, and often patients survive long enough to reproduce. This lowers the level of constraint and obscures the delineation of high-constraint regions.

Despite some relationship between secondary structures, presence of protein domains and some protein and sequence properties, there were no clear hypotheses resulting from this analysis, primarily due to the small gene-set investigated (Subsection 3.2.10). It is certainly not possible to identify or characterise mutational hotspots based purely on the boundaries of known protein features. This is important to recognize as a naïve approach may place too much weight on known functional domains. There is almost certainly connections between hotspots and structural and functional features. These are likely to be complex involving a large number of interactions between features and will require large training sets to model effectively.

Comparison of hotspots with previously reported pathogenic *clinvar* variants and common GnomAD variants revealed a strong overlap (Table 3.6). It is likely that incorporation of these data would improve the resolution of the positional component of the *hotspot models*. However, this would complicate inclusion of burden into the model. An alternative approach would be to integrate these data as Bayesian priors for hotspot locations. By constraining models based on prior data, where the training data and prior data agree on hotspot locations, standard errors could be reduced, reducing uncertainty to a level fit for clinical utility.

3.4.4 Evaluation of hotspot models

Specific and detailed explorations of how well these models generalize to new data were not conducted due to difficulties in aggregating a suitable validation case-series. Therefore, evaluation metrics were calculated by holding back training data to test on. Although of academic interest, the evaluation metrics for these genes are useful only in the sense they give an indication of the importance of protein position as a predictive feature. In practice, only variants in a confirmed case

are subject to interpretation. The only metrics of interest in this scenario are the true-positive and false-negative rate. The recommended guidelines state that a variant with 90% probability of pathogenicity can be considered a likely pathogenic variant (Harrison and Rehm 2019). In line with this threshold, no model had a sufficiently high true-positive rate to classify any variants as likely pathogenic. However, as multiple lines of evidence are considered when evaluating the pathogenicity of a variant, the contribution of these models may corroborate evidence from additional sources leading to a pathogenic classification.

As a data-driven approach, the models will continue to improve as more data becomes available. Although most of these genes that were successfully modelled were in HCM, DCM has more reported causal genes, many of which cause disease via missense variants. As our HCM cohort was over five times as large, it is likely that with an increased sample size for this cohort, more genes may be highlighted with mutational hotspots, including; *LMNA*, *MYH7* and *RBM20*, which yielded specific regions with ORs greater than 5. The same may be true for additional genes in the other diseases investigated.

3.4.5 Hotspot+ models

Hotspot-models were extended to include additional *in-silico* scores from pathogenicity prediction algorithms (*hotspot+ models*). With the addition of these features, variant predictions varied extensively. The model now incorporated evidence relevant to two ACMG criteria; PM1 (mutational hotspots) and PP3 (*in-silico* evidence). Separation of case and control variants generally improved in cross-fold validations and model predictions were correlated with clinical classifications. AUC comparisons revealed that data-driven gene-specific models often outperformed generic prediction scores trained on much larger datasets.

The *hotspot+ models* were specified using a consistent strict automatic feature selection procedure. The rationale underlying this strategy was that even when holding out unseen test data,

the datasets were so small that analysis of multiple complex models could have yielded some that overfit to the training but still performed well on the test data by chance. However, in the effort to avoid overfitting, some genuinely generalizable features may have been ignored. In fact, as many of the algorithms have already been calibrated on Mendelian disease variants, therefore they may be informative even if they are not strongly correlated in the specific variants observed in this dataset. In fact, for some low penetrance genes that explain only a small proportion of the genetic heterogeneity of the diseases, it is possible no observed case variants are the primary disease-causing variants. In this case, any correlation between variant scores and cohort status is by chance. For genes with no additional predictors after feature selection, models may have benefitted with the addition of an algorithm that has previously been reported as reliable (Ghosh *et al.* 2017) or was highly correlated across all inherited heart disease genes (Figure 3.13), such as REVEL or VEST4. Hidden overfitting would then need to be assessed in additional validation data.

The *hotspot+ modelling* approach analysis may suffer from type 1 circularity (Subsection 1.2.5). Some of the variants observed in the inherited heart disease cohort may have been uploaded to public repositories of disease variation and consequently used in the training data for the pathogenicity prediction algorithms. Implications of this include reinforcement of previous classifications leading to bias. In future iterations of these models, variants specific to the investigated diseases should be excluded from the training sets of these algorithms. Furthermore, scores using intra-species constraint should be excluded, or controls should be changed to those not-overlapping with the training set for these algorithms.

Integration of *in-silico* evidence in *hotspot+ models* increased both model accuracy (Table 3.10) and correlation with clinical classifications (Figure 3.15) compared to *hotspot models*. However, applying the model in line with the ACMG variant interpretation guidelines becomes more complicated as the model draws from information relevant to two criteria PM1 and PP3. As an additive framework, rules are currently assessed only one at a time and combinations of different

evidence levels map to specific final classifications. The simplest solution would be to collapse the two criteria into a single one. An objective of this data-driven approach was to quantify the uncertainty in *in-silico* prediction algorithms. This was done by providing evidence-based on the lower confidence estimates from the model. For many genes including *MYH7*, *MYBPC3*, *KCNH2* and *KCNQ1,* many carriers had lower confidence estimates high enough to confidently ascribe evidence. Comparing model predictions with clinical classifications made using ACMG guidelines revealed moderate correlation and a very heterogeneous group VUS variants. High ORs were predicted for some VUSs leading to moderate or strong evidence of pathogenicity, but there were also some very low ORs suggesting some of these are possibly benign. In this sense, the GAM predictions may be useful to stratify this group into variants both more likely to be pathogenic *and* more likely to be benign.

Chapter 4

Missense-variant clustering

4.1 The prevalence of missense-variant clustering

In this thesis, missense-variant clustering has been proposed as a common phenomenon in Mendelian disease-genes. This was alluded to briefly in the introduction in the context of deafness-associated genes (Figure 1.2). Here, widespread evidence of variant clustering is explored further, using genes present in the publicly accessible database, ClinVAR (*clinvar*; Landrum *et al.* 2018). *Clinvar* accepts submissions of variants with clinical classifications derived from clinical testing laboratories, research or the literature. A VCF containing all variants available in the September 2019 release was downloaded from the *clinvar* FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/). This VCF contained the genome coordinates of each variant as well as uploader-derived information such as their assessment of clinical significance and the patient's condition. The VCF was annotated

with VEP to obtain features such as overlapping gene and variant consequence (i.e. missense, synonymous or LoF) and population frequencies from *gnomad_e2* and *gnomad_g2*.

The dataset contains data for 351,710 protein-coding variants including; 13,474 benign, 58,771 likely benign, 197,853 VUS, 25,172 likely pathogenic and 55,960 pathogenic. For *clinvar* variants with an annotation of benign or likely benign, 72.4% were synonymous, and the *gnomad_e2* allele count for the remaining 27.6% non-synonymous variants was high (median=197). As population allele frequency is the key piece of evidence used to provide a benign classification (Richards *et al.* 2015) this fitted with expectations. In total, 6015 genes were represented in *clinvar*. However, this includes 1499 genes with only one uploaded variant and a further 1092 genes with two to four variants. There were 209,993 missense variants (59.7%), making this variant type the most commonly reported. Synonymous variants were the next most common group (74,027; 21.0%) followed by frameshift (34002; 9.7%) and stop gained (21,323; 6.1%).

Gnomad_e2 and *gnomad_g2* were used as controls. VCFs for these datasets (version 2) were downloaded from the GnomAD website. Each dataset was re-annotated using VEP to be consistent with the *clinvar* dataset. In *gnomad_e2*, there were 8,624,738 variants including 1,993,507 variants with a population maximum allele frequency above 0.1%. In total there were 19,745 genes represented, including only 168 genes with less than five missense variants. Although *clinvar* variants have a bias towards variant types with a higher probability of being deleterious, *gnomad_e2* variant types follow the same order as *clinvar*; 47.6% missense, 22.7% synonymous, 2% frameshift and 1.4% stop gained.

In all analyses, genes were only included if they had at least five unique missense variants in both the cases (*clinvar)* and controls (*gnomad*), comprising up to 3045 genes, but dependent on the additional filtering criteria. For each analysis, *gnomad* variants were filtered by their allele frequency and *clinvar* variants were filtered by their clinical significance (provided with their submission). *Gnomad* variants were defined as common if they had a population maximum allele

frequency above 0.1% and uncommon otherwise. The clinical significance of ClinVAR variants was assigned to a five-class system, including; benign, likely benign, VUS, likely pathogenic or pathogenic. Variants with more than one classification, including conflicting interpretations of pathogenicity, were assigned a single class by taking the average of the classifications. This was achieved by ranking each class from 1 (benign) to 5 (pathogenic) and taking the average. A non-observed VUS, weighted ¼ of the real classifications, was included in each calculation to force a conservative classifications in the case of ties. The result was rounded to the nearest integer and converted to the corresponding classification. Variants with classifications from miscellaneous classes such as; 'other', 'risk factor', 'affects', 'association', 'drug response' or 'not provided', were excluded analyses unless specified otherwise.

Multiple filtering routines generated case-control datasets for analysis. Clustering of missense variants was then assessed using the *BIN-test* for the different filtering scenarios. Missense variants were supplemented by inframe insertions or deletions (indels) that affect three or fewer amino acid residues, which included 84.6% of the inframe indels across the cohorts (132,353/156,420).

In this chapter, clustering signals are assessed in a *gnomad_e2* and *gnomad_g2* putative null models to assess effectiveness of the *BIN-test* reciprocal mean 10X coverage weighting scheme. This is followed by an analysis of *gnomad_e2* and *clinvar* to detect clustered genes. Clusters in these genes were then defined using a novel algorithm. The properties of these clusters were then compared with known protein features downloaded UniProt (The UniProt Consortium 2019), as well as occurrence of variants in *clinvar* and *gnomad*.

4.2 GnomAD exomes versus GnomAD genomes null model

To investigate the empirical false-positive rate of the *BIN-test*, clustering was assessed between *gnomad_e2* and *gnomad_g2*. Both datasets contain samples generally from large common-disease projects, such as cardiovascular disease or diabetes, and have severe paediatric diseases filtered out. The distribution of diseases between the groups is unlikely to be exactly the same, so it is

theoretically possible that if a common disease group was overrepresented in one cohort and underrepresented in the other, then a true association could occur. However, the substantial background noise of other disease groups in each cohort make it arguably unlikely that true signals will rise above this noise. Both common (*popmax* > 0.1%) and rare (*popmax* < 0.1%) were investigated.

To explore potential confounding, genes were flagged based on coverage metrics and protein lengths. Proteins with over 3000, 2000 and 1000 residues were flagged as very high, high and moderate concern (pl_flag). Proteins with over 10%, 30% and 50% of *BIN-test* bins with lower than 80% coverage were flagged as very high, high and moderate concern (cov_flag1). Proteins with mean coverage less than 60%, 70% and 80% were flagged as very high, high or moderate concern (cov_flag2). Where the gene did not fall into any of these categories, it was labelled as "pass".

Common variants in both cohorts were analysed first. As these are substantially fewer and should be shared between both cohorts, significant results here would flag problems with the experimental design. Reassuringly, no *BIN-test* p-value was significant. In the next stage, rare-variants were examined. Overall, 6.3% of p-values were less than 0.05, six of which were *b-significant* (Table 4.1; Figure 4.1). The lambda genomic inflation factor (lambda) was used to assess the overall deviation from the expected null model. This metric is the ratio of the median test statistic observed and the expected median, and should be 1 for a true null model. The genomic inflation factor lambda was 1.024 for this analysis. After removing variants flagged for coverage and very large proteins (>3000 residues), false-positives were reduced to 5.8%, only four *b-significant* results remained and lambda dropped to 0.99.

Table 4.1: Bonferroni significant *BIN-test* p-values in the analysis of GnomAD exomes versus GnomAD genomes. pl_flag = protein length flag, cov_flag1 = coverage flag type 1, cov_flag2 = coverage flag type 2.

Gene name	*BIN-test* p-value	pl_flag	cov_flag1	cov_flag2
MUC4	1.78e-15	vhigh	moderate	vhigh
FGF2	4.47e-07	pass	pass	moderate
ZNF763	8.95e-07	pass	pass	pass
GTF2IRD2	1.58e-06	pass	pass	moderate
PPP1R13L	1.77e-06	pass	moderate	high
CABP1	1.85e-06	pass	pass	moderate

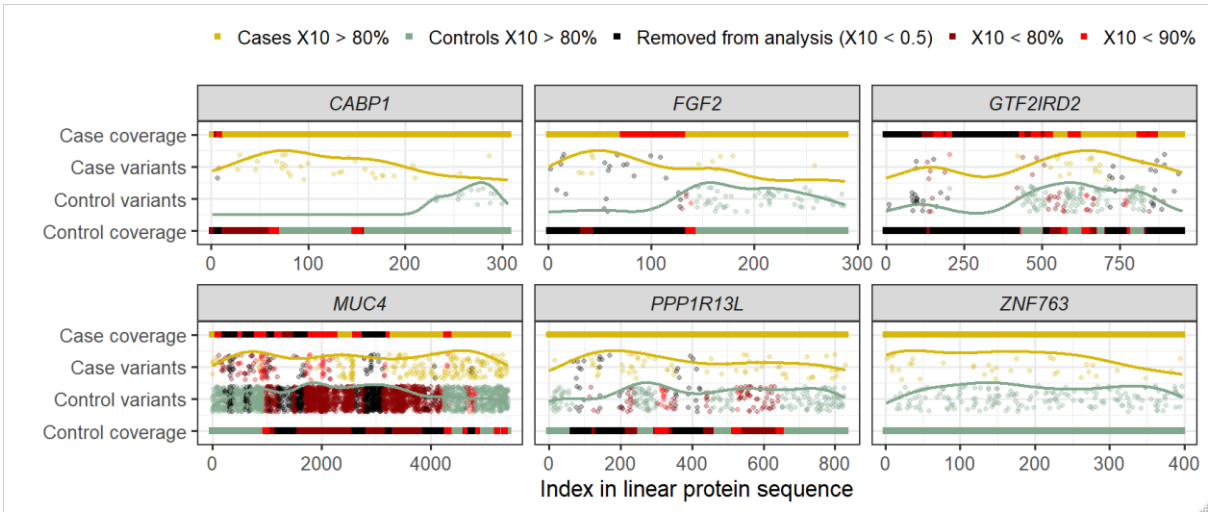

Figure 4.1: Distribution of variants and coverage in Bonferroni significant *BIN-test* genes in the analysis of GnomAD exomes (controls) versus GnomAD genomes (cases). Observed variants displayed as points in the graph are coloured by the coverage at that site, just as the two strips denoting case and control coverage. For example, any variant coloured as black, will have been excluded from the analysis (10X < 0.5).

Of the six *b-significant* genes, *MUC4* had the most-significant p-value ($p < 1.8 \times 10^{-15}$). This gene also was flagged as a gene that had global coverage issues (cov_flag1 = moderate), a substantial number of *BIN-test* bins with low coverage (cov_flag2 = very high) and was a very large protein (pl_flag = very high). It is therefore plausible that imperfect coverage control drove this signal. For the remaining genes, only *ZNF763* passed all the filters and visual inspection indicated no clear coverage-related reason why it was significant ($p < 8.9 \times 10^{-7}$). *ZNF763* is a zinc-finger protein, not predicted as particularly constrained for either missense of loss-of-function variant (Lek *et al.* 2016). The signal seems to be driven by a reduction in variation in the C-terminus of *gnomad_g2*. Inspection of the coverage files for this gene revealed no region had fewer than 99.7% of samples

with at least 10X coverage in *gnomad_g2* and 98.3% in *gnomad_e2*. P-values across all genes generally indicated a uniform distribution with a modest peak near zero driving the lambda inflation (Figure 4.2).

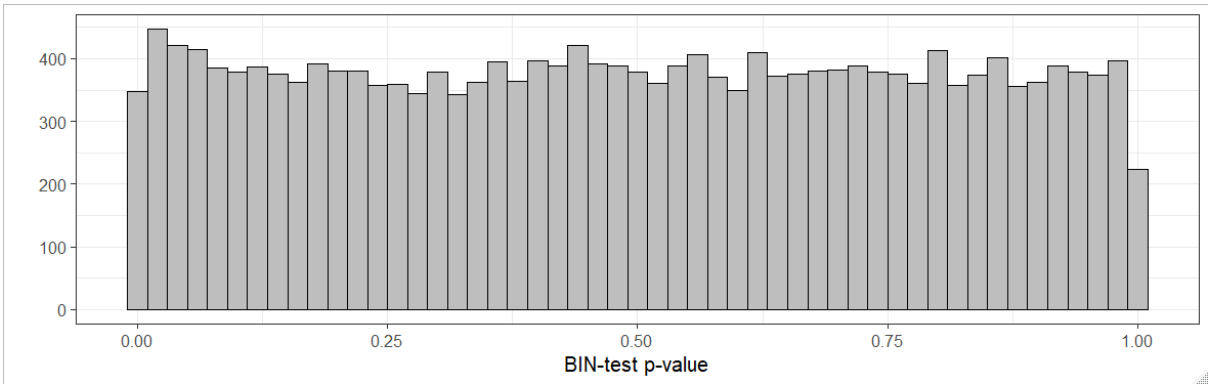

Figure 4.2: Distribution of *BIN-test* p-values in the putative null analysis between GnomAD exomes versus GnomAD genomes. In this analysis, rare-variants (*popmax* < 0.1%) were included as many times as they were observed, and uneven coverage was controlled for using reciprocal weighting.

The same null model analysis was repeated without controlling for coverage i.e. without inverse coverage weighting for *BIN-test* bins. In this analysis many more significant signals arose. Overall, 16.2% of genes yielded *n-significant* results (alpha<0.05), 3.2% of genes yielded *b-significant* results (alpha<2.58e-6) and the genomic inflation metric lambda was 1.61. To explore whether genes flagged for issues overly contributed to these significant results, lambda inflations statistics were calculated for each combination of these flags (Table 4.2). Wherever genes were grouped by a concern-level moderate or greater, lambda was above 1. Protein length was the most impactful confounder based on the lambda metric of 2.273 for all proteins longer than 3000 residues (n=148) and 2.061 for all proteins longer than 2000 residues (n=299). For the 12,491 genes with pass for all three filters the lambda statistic was 0.86. For this gene group the method is overly conservative, potentially as a result of the reduced power associated with small genes with few observed variants. To determine this, the statistic was recalculated with only genes over 200 residues in

length. The lambda then rose to 1.02 for all proteins with length greater than 200 (for proteins smaller than this the lambda was 0.32).

Table 4.2: Lambda metrics stratified by potential confounder flags for *BIN-test* p-values in a *gnomad_e2-gnomad_g2* putative null model. Genes are stratified by each level of each flag to form overlapping subsets of genes. Each cell displays lambda in the format 'all gene' / 'all genes greater than 200 residues'.

	Very	High	Moderate	Pass
Coverage flag 1	1.05/1.89	1.56/1.76	1.29/1.93	0.99/1.12
Coverage flag 2	1.13/1.81	1.35/1.73	1.12/1.17	0.96/1.09
Protein length flag	2.27/2.27	2.06/2.03	1.56/1.52	0.93/1.08

It is evident that the controlled results are a substantial improvement to the uncontrolled analysis, supporting the application of the novel coverage weighting procedure. However, false-positives still breach the nominal 5% level indicating either coverage is not being controlled for completely, further factors are confounding the analysis or true signals exist in this putatively null model. Nevertheless, these imperfections were accepted and subsequent analyses were performed with *gnomad_e2* controls using the *BIN-test* with reciprocal coverage weighting in the controls and regions with less than 50% of samples with 10X coverage in *gnomad_e2* were excluded. Coverage information was not available for *clinvar*, an important point that will be covered in the discussion.

4.3 GnomAD exomes vs ClinVAR

As a liberal baseline, the distribution of protein positions for all *clinvar* and *gnomad_e2* variants, regardless of their clinical significance or MAF, were compared; *all-all* analysis. Clustering was assessed in 3045 genes that had at least five unique variants in each cohort. The coverage-adjusted *BIN*-test detected *n-significant* clustering in 525 genes and *b-significant* (alpha < 0.05/3045) clustering in a further 143 genes.

Each cohort was then refined to include only common (putatively benign) variants and variants with a classification of VUS, likely pathogenic (LP) or pathogenic (P) in *clinvar* (*common-vus+*). After

filtering, 2837 of these genes had at least five unique variants in each cohort. In the *clinvar* dataset, there was an average of 58 VUSs, seven LP and eight P variants across these genes. In *gnomad_e2,* there were 90 common variants on average. The *BIN-test* yielded 509 *n-significant* genes and 78 *b-significant* genes (alpha < 0.05/2837). Increasing stringency further by excluding all VUSs, leaving only LP and P variants (*common-lp+*), reduced the filtered set to 1340 genes. On average, there were nine LP, 11 P and 86 *gnomad_e2* common variants in each gene. The *BIN-test* yielded 583 *n-significant* signals, 178 n-*significant* (0.05/1340) and 160 *b-significant* (0.05/2837). As a visual example of the landscape of significant results, the p-values from the *common-lp+* analysis are displayed in a Manhattan-style plot in Figure 4.3.

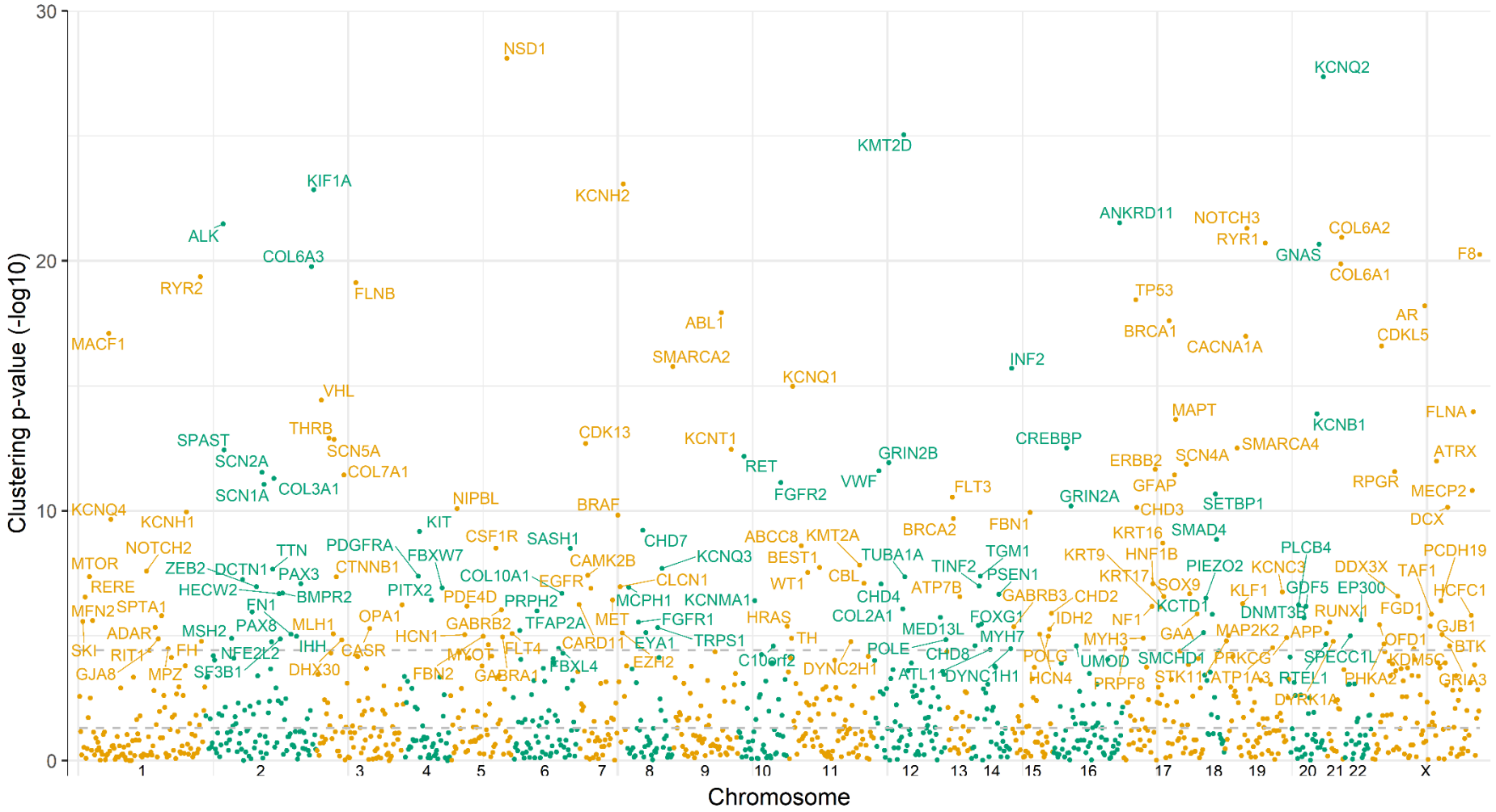

Figure 4.3: P-values (-log10) for differential distribution of variant protein positions between ClinVAR variants with an annotation likely pathogenic or pathogenic and GnomAD exomes variants with a population maximum frequency greater than 0.1%. A total of 1299 genes each with at least five variants in each cohort was included. Bonferroni significant genes are labelled with their HGNC gene symbols.

None of the *b-significant* genes arising in *gnomad_e2* versus *gnomad_g2* null model, reached *b-significance* in the *all-all*, *common-vus+* or *common-lp+* analysis. However, there was overlap between the three *gnomad-clinvar* analyses. For *b-significant* genes (Figure 4.4A), 53 genes were observed in all three filtering scenarios, 17 additional genes were found by including VUS's on top of likely pathogenic and pathogenic variants and 51 additional genes arose when including all variants in *gnomad_e2* and *clinvar*. Ninety genes were detected only under the most stringent filtering conditions (*common-lp+*) where the highest number of genes reached significance. A similar pattern is observed for n-significant variants (Figure 4.4B) and many genes were shared between the filtering scenarios with the most overlap between the common-VUS+ and the common-LP+ approaches.

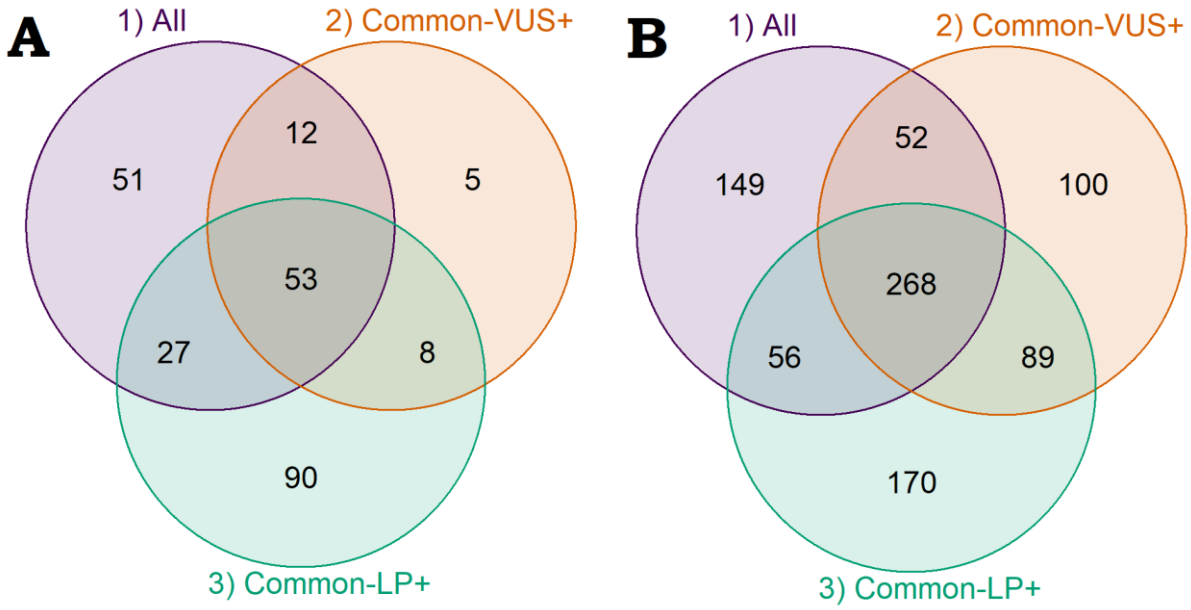

Figure 4.4: Overlap of genes with significant clustering difference between GnomAD and ClinVAR for three experimental scenarios. A) overlap with Bonferroni-significant genes and B) overlap with nominally significant genes. 1) All = all variants in GnomAD and ClinVAR, 2) Common-VUS+ = common GnomAD variants and ClinVAR variants with a classification of VUS, LP or P, 3) Common-LP+ = common GnomAD variants and ClinVAR variants with a classification of LP or P.

Overall, reducing *clinvar* to only LP/P variants yields the most results. However, a substantial number of genes were only detected when including all *gnomad* variants as opposed to only the common. Therefore, the analysis was repeated with all *gnomad* variants and *vus+* and *lp+ clinvar* variants. There were 2872 and 1363 genes post-filtering for *all-vus+* and *all-lp+* respectively. For *all-vus+*, there were 576 *n-significant* genes and 162 *b-significant* genes. For *all-lp+*, there were 654 *n-significant* genes and 292 *b-significant* genes. Using all *gnomad_e2* variants, regardless of frequency, yielded substantially more signals than limiting to common variants only. Visualising the overlaps between gene sets of the common-only versus the all-*gnomad_e2* analyses (Figure 4.5), reveals that while a few signals were lost, a substantial number of *b-significant* signals were not found when using only common *gnomad_e2* variants.

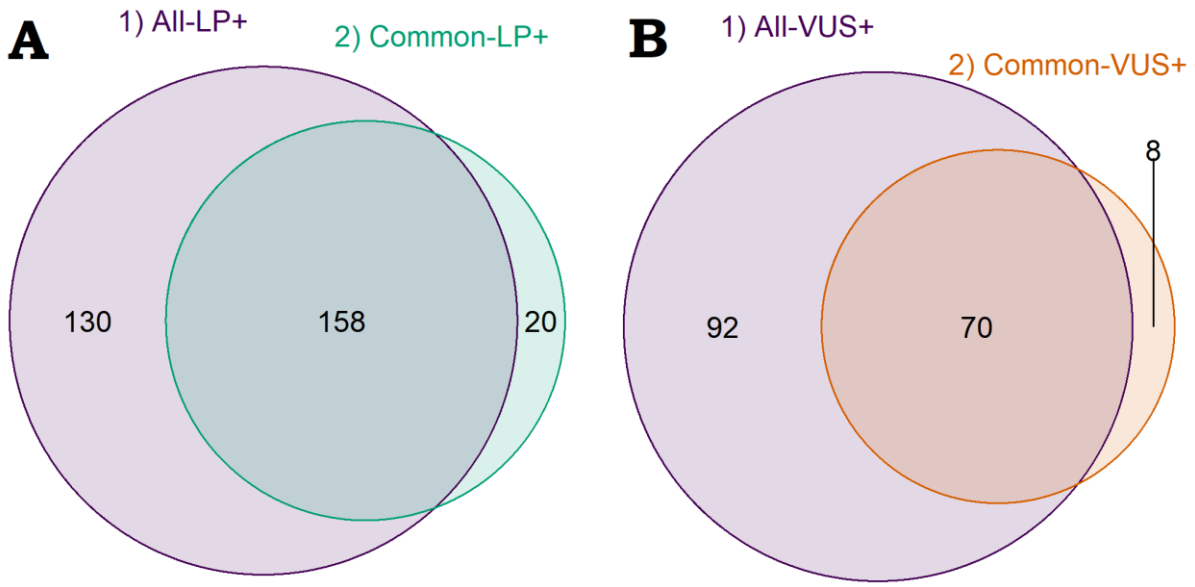

Figure 4.5: Overlap of Bonferroni significant clustering signals when using only common GnomAD variants (Common-X) versus using all GnomAD variants (All-X). A) overlap when including only LP or P ClinVAR variants, B) overlap when including VUS, LP or P ClinVAR variants.

4.3.1 Comparison of gene groups present

The gene groups present in GnomAD, *clinvar* and *b-significant* gene lists were analysed for overrepresentation (Fisher's-exact) using PANTHER (Thomas *et al.* 2003). Firstly, PANTHER protein classes in all genes with a submission in *clinvar* were compared to all human protein-coding genes. The top hits revealed overrepresentation of three main protein classes in *clinvar*. Enzymes; metabolite interconversion enzymes (p<3.06e-21, OR=1.61), oxidoreductase (p<4.31e-11, OR=1.85) and dehydrogenase (p<8.77e-8, OR=2.41). Transporters; transporter (p<3.55e-13, OR=1.68), ion channel (p<2.39e-08, OR=2.25) and primary active transporter (p<9.31e-6, OR=1.82). Extracellular matrix proteins; extracellular matrix protein (p<1.51e-8, OR=2.17) and extracellular matrix structural protein (p<2.02e-7, OR=2.7). Interestingly, stark underrepresentation of immune-related proteins was also observed in *clinvar*; immunity protein (p<1.19e-18, OR=0.21) and immunoglobulin (p<8.2e-18, OR=0.02).

Next, genes with a *b-significant* clustering signal in *all-vus+* (n=162) were compared against all *clinvar* genes (n=5421). The top hits were; ion channel (p<4.85e-5, OR=4.83) and DNA-binding transcription factor (p<5.39e-5, OR=3.28). Other significantly overrepresented classes included; G-protein, ligand-gated ion channel, voltage-gated ion channel, gap junction intermediate filament binding protein and P53-like transcription factor. Clustered ion channel genes driving this signal included five voltage-gated sodium ion channels; *SCN1A*, *SCN2A*, *SCN4A*, *SCN5A*, *SCN8A*, two ryanodine voltage-gated ion channel receptors; *RYR1* and *RYR2*, as well as; *GABRB2*, *GABRB3* and *BEST1*. Clustered transcription factors included; *MEF2C, FOXL2, ZBTB18, SMAD3, THRB, HNF1A, ERF, CAMTA1, EGR2, TP63, CTCF, ZEB2, NFE2L2, MYC, HNF1B, YY1, ZBTB20, SOX11, FOXC1, PBX1* and *MEIS2*. Extending this analysis to all *n-significant* genes (n=556) increases this list to 63 transcription factors significantly overrepresented by a factor of 2.6. These results suggest protein types more prone to missense-variant clustering. Interestingly, an underrepresentation of metabolite interconversion enzymes was

also observed (p<1.77e-3, OR=0.34), perhaps suggesting enzymes are less likely to be associated with clustered missense variants.

Next, genes *b-significant* in the *all-lp+* analysis (n=292) were compared for over-representation compared to the remaining *clinvar* genes not in this set (n= 5147). The top hits were ion channel (p<1.8e-5, OR=4.54), G-protein (p<1.97e-5, OR=6.56), and DNA-binding transcription factor (p<8.01e-4, OR=1.93). Ion channels here included all ion channel genes referenced in the previous paragraph and a further six genes; *CHRNB2*, *GABRA1*, *ITPR1*, *TRPV4*, *GLRA1* and *GABRG2*. Amongst which 2 additional gamma-aminobutyric acid receptor subunits bringing the total of these to four. The clustered G-proteins included; *GNAS*, *RHOBTB2*, *GNAQ*, *NRAS*, *KRAS*, *ATL1*, *RAC1*, *RIT1*, *GNAO1* and *HRAS.* Again, an under-representation of metabolite interconversion enzyme was discovered (p<4.5e-3, OR=0.43).

All *clinvar* genes versus *all-lp+* were then analysed against *clinvar* for other gene groups. For PANTHER pathways, the top results were Gonadotropin-releasing hormone receptor pathway (p<2.37E-04, OR=3.37) and CCKR signaling map (p<5.67E-04, OR=4.12) but also included p53 pathway feedback loops, Angiogenesis and Integrin signaling pathway. For PANTHER GO-slim molecular functions, 18/20 top hits were overlapping lists of transporters or channels including the top hit passive transmembrane transporter activity (p<1.38e-10, OR=3.84). DNA binding and transcription factor related functions also appeared multiple times in the *b-significant* signals. PANTHER GO-slim biological processes painted the same picture with transporters dominating the output. The top hits from Reactome pathways was Developmental Biology (p<3.62e-11, OR=2.86), Signal Transduction (p<3.80e-7, OR=1.89) and Generic Transcription Pathway (9.18e-7, OR=2.58). These results may suggest gene groups and protein types where missense variant clustering is more likely to be observed.

4.3.2 Properties of b-significant genes

Missense constraint, protein length and domain or motif annotation were assessed for gene lists defined by their clinical significance in *clinvar* and their missense-variant clustering (Table 4.6). Mean missense constraint is significantly higher (p<2.2e-16) in genes with at least one submission of a missense variant in *clinvar* (n=5421; Z=0.991) compared to *gnomad_e2* genes without a submission in *clinvar* (n=14,277; Z=0.670). The mean missense constraint is higher again when limiting to genes with a VUS variant (n=5186; Z=1.02) and higher again with only genes with a LP/P variant (n=3440; Z=1.11). The highest mean constraint scores was observed in *b-significant all-vus+* genes (Z=2.55). This group also has the highest mean protein length (n=162; pl=1425). Although, mean missense constraint was higher in the *all-vus+* analysis than the *all-lp+* analysis, it was not significantly different for *b-significant* (p < 0.22) or *n-significant* (p < 0.37) genes. Of the 288 *b-significant all-lp+* genes, 64.7% had a known functional domain and 27.6% had a known functional motif. This was the highest of all groups, followed closely by *b-significant all-vus+* genes.

Table 4.6: The mean missense constraint (observed/expected and Z-score), the mean protein length and the percentage of genes with a known protein domain or motif, in different gene lists. *Gnomad_e2* = all *gnomad_e2* genes, *Clinvar* = all genes with at least one submission to ClinVAR, *Clinvar-vus+ and Clinvar-lp+* = all genes with at least one submission with a classification of VUS/LP/P, or LP/P respectively. The remaining rows display results for significantly clustered genes for different case-control cohorts. *Bsig* = Bonferroni significant, *nsig* = nominally significant and *all* = all gnomAD variants (common and rare).

Gene lists	N	Missense OE	Missense Z	Protein Length	Domains	Motif
Gnomad_e2	19,697	0.874	0.759	583	0.429	0.117
Clinvar	5,421	0.853	0.991	799	0.469	0.155
Clinvar-vus+	5,186	0.848	1.020	798	0.470	0.156
Clinvar-lp+	3,440	0.834	1.110	799	0.453	0.163
Bsig all-vus+	162	0.682	2.550	1425	0.627	0.242

Bsig all-lp+	288	0.707	2.330	1286	0.647	0.276
Nsig all-vus+	576	0.715	2.010	996	0.575	0.234
Nsig all-lp+	654	0.736	1.940	1041	0.568	0.219

4.3.3 Properties of missense variant clusters

The nature of missense-variant distributional differences driving *BIN-test* significance were assessed in the top *b-significant* genes. For the most significant genes in the *all-lp+* analysis, LP/P *clinvar* variants tended to cluster in tight regions of the protein in which a corresponding depletion in *gnomad_e2* was also observed (Figure 4.6). For the genes, *NSD1*, *GNAS*, *CREBBP*, *COL6A3*, *KIF1A, GRIN2B*, *ALK* and *ZBTB20*, almost all variants classified as LP/P in *clinvar* are observed in a single (relatively small) cluster. In most, a relative depletion in controls is also evident. This is especially apparent in the C-terminal of *GNAS* and *ZBTB20* for which the control depletion is severe suggesting strong purifying selection acting on this region. Also notable from Figure 4.6, is the overlap of some clusters with functional domains. This overlap is especially discernable in *GNAS* (G-alpha), *KIF1A* (Kinesin motor) and *ALK* (Protein kinase).

In other genes, more complex patterns emerge. For *KCNQ2,* associated with early infantile epileptic encephalopathy, two clusters are observed. For *RYR1* (a myopathy associated gene), a strong C-terminal cluster is observed, accompanied by the characteristic control depletion, but with a further uniform background of LP/P variants. For *SCN1A* (associated with severe myoclonic epilepsy in infancy) and *RYR2* (associated with catecholaminergic polymorphic ventricular tachycardia), multiple clusters regions of increased pathogenicity potential exist. It is worth noting the potential clinical significance of the two highly significant Ryanodine receptors genes *RYR1* (8.8×10^{-21}) and *RYR2* (1.1×10^{-19}). Both are present on ACMG's list of genes to be reported as incidental findings. However, this analysis corroborates prior isolated evidence of mutation clustering in these genes (Medeiros-Domingo *et al.*

2009), potentially highlighting a route to determining the clinical relevance of these genes with more confidence.

When including the high number of variants with uncertain classifications in *clinvar* (Figure 4.7), some of the top 12 genes overlap with the previous analysis (*ZBTB20*, *NSD1*, *KCNQ2*, *GNAS* and *SCN1A*), but in general VUSs added noise to the neat clusters observed in Figure 4.6. The top gene in the first analysis *NSD1*, associated with Sotos Syndrome, now sees a much more uniformly spread set of rare-missense variation and a less significant p-value. The same is true for *SCN1A* and *KCNQ2*. It is quite possible, that the classification of either LP/P or VUS is somewhat predetermined by the clusters detected here, boosting the chance of a pathogenic classification when variants fall in the tight regions from 4.6.

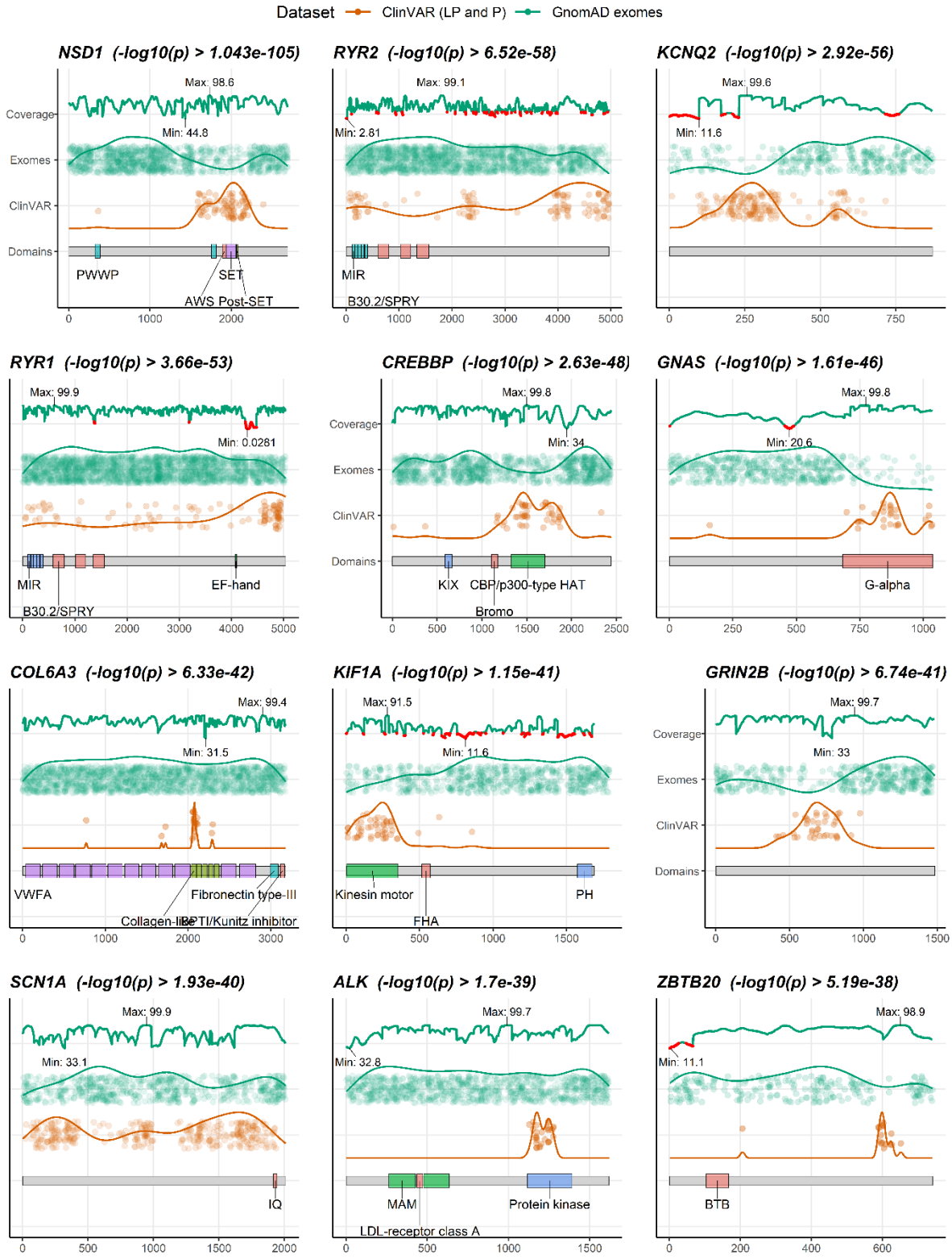

Figure 4.6: Top 12 most significant genes in an analysis of variant positions between all *gnomad*_e2 variants and *clinvar* likely pathogenic and pathogenic variants. Mean coverage information is displayed for *gnomad_e2*, regions below 30X coverage are coloured in red. Protein domains, annotated from UniProt, are displayed in the lower row of each plot.

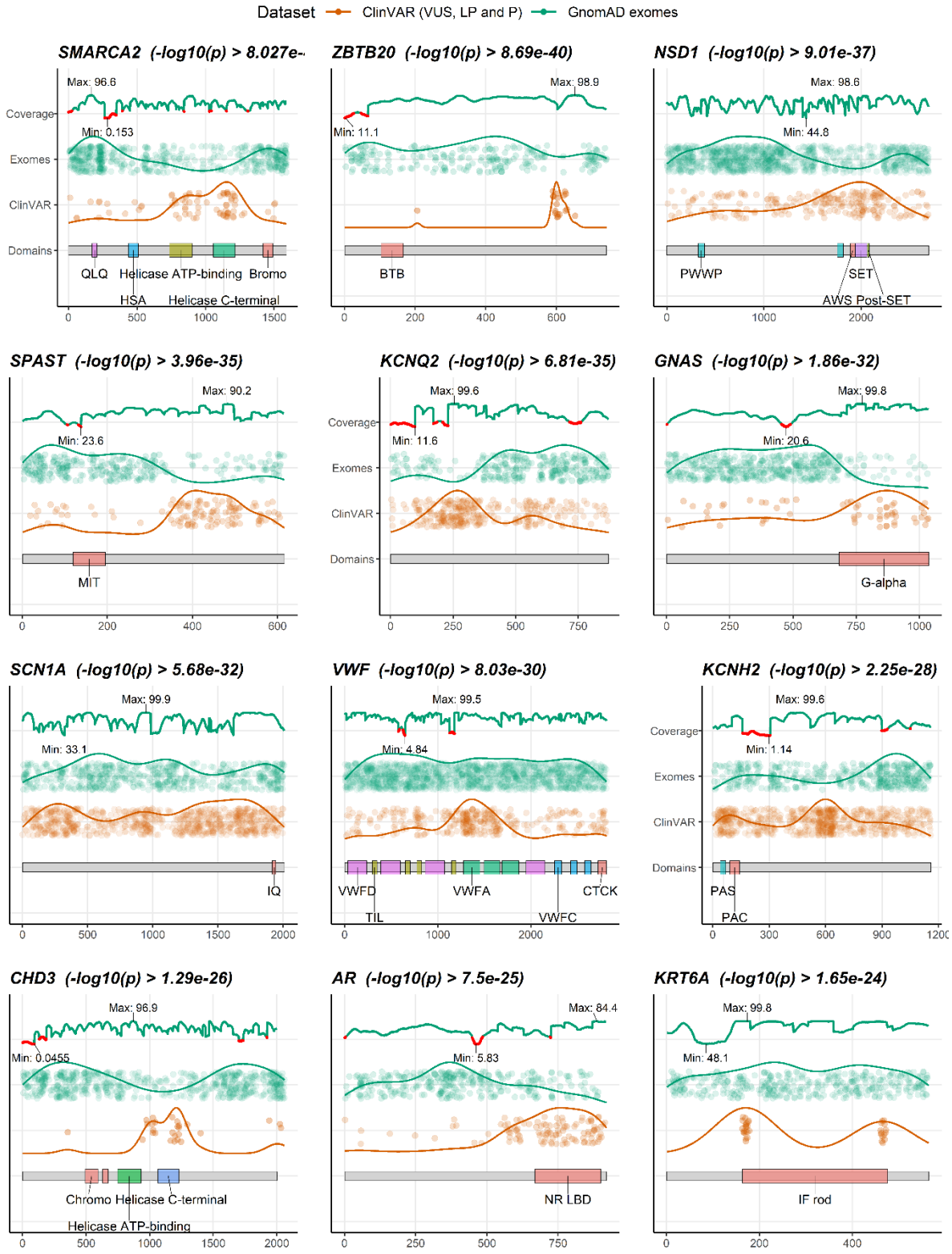

Figure 4.7: Top 12 most significant genes in an analysis of variant positions between all *gnomad_e2* variants and *clinvar* likely pathogenic and pathogenic variants. Mean coverage information is displayed for *gnomad_e2*, regions below 30X coverage are coloured in red. Protein domains, annotated from UniProt, are displayed in the lower row of each plot.

4.3.4 Determining "clusters"

To examine the properties of these putative pathogenic clusters numerically, an algorithm was written to identify their locations. Although previously GAMs were used to achieve this, here as no sample sizes were available for *clinvar*, burden could not be assessed. Furthermore, as many regions within *clinvar* had zero variation. When attempting to fit a GAM in such a scenario, this produces large regions of infinite confidence intervals and disrupts smoothing in adjacent regions. Finally, as predicted ORs may vary widely between genes, it was unclear which OR-threshold to use to determine a "cluster" under this framework, making automation of this approach more complicated.

Intuitively, for tight clusters, a simple 1-dimensional clustering algorithm is sufficient to estimate start and end locations of the clusters. However, for the more complex clustering patterns, this may not capture more subtle differences in variant density between the cohorts. Therefore, variant counts were assessed at each residue plus ten residue flanks on either side. This was done at every residue for both cohorts. Then the proportion of total observations in each cohort that were present in the local window surrounding each residue was compared between the cohorts using a chi-squared proportion test. A one-sided test was used to identify only regions with excess variation in cases. All significant residues ($p<0.05$) were considered part of a cluster. Clusters that were within 10 residues of each other were then merged. The estimated cluster locations for the top 20 most significant genes in the *all-lp+* and *all-vus+* analyses are displayed in Figure 4.8. On visual inspection, estimated clusters seemed to correctly map to regions enriched for variation. Tight clusters of increased variant density were identified as specific clusters and more evenly distributed variants were not included in these clusters. For the overlapping genes such as *SCN1A* and *KCNQ2*, the addition of VUS variants make little difference to the cluster locations. This approach was accepted as sufficient to allow the subsequent analysis of the properties of clusters in significant genes.

Figure 4.8: Clusters delimited by the clustering algorithm, in the top 20 most significant genes in the *all-lp+* and *all-vus+* analyses. The distribution of *gnomad_e2* variants (green) and *clinvar* variants (orange) are plotted alongside the inferred clusters (black).

A total of 1213 separate clusters were identified by the algorithm for the 290 genes with a *b-significant BIN-test* p-value in the *all-lp+* analysis. Cluster sizes ranged from 1 to 171 residues, had a mean of 24.7 residues and a skew towards smaller sizes. The number of clusters per gene ranged from 1 to 29 with a mean of 4.2. The gene with 29 clusters was *FBN1* (Figure 4.9). *FBN1* is a large protein (residues=2871)

associated with Marfan syndrome. Clustering of variants in cysteine residues at EGF-like domains was reported very early in investigations of this gene (Dietz *et al.* 1992). Subsequent identification of hotspots associated with neonatal Marfan syndrome, severe Marfan syndrome and mild Marfan syndrome were observed in almost all 65 exons (Palz *et al.* 2000). In a more recent investigation, it was reported that most variants cluster in exons 24-32. Four clusters occurred between these exons starting from residue 1045 and ending at residue 1219 and highlighted by a red circle in Figure 4.9. Assuming a uniform distribution, 14 control and 9 case variants are expected but only eight control variants and seventeen case variants are observed.

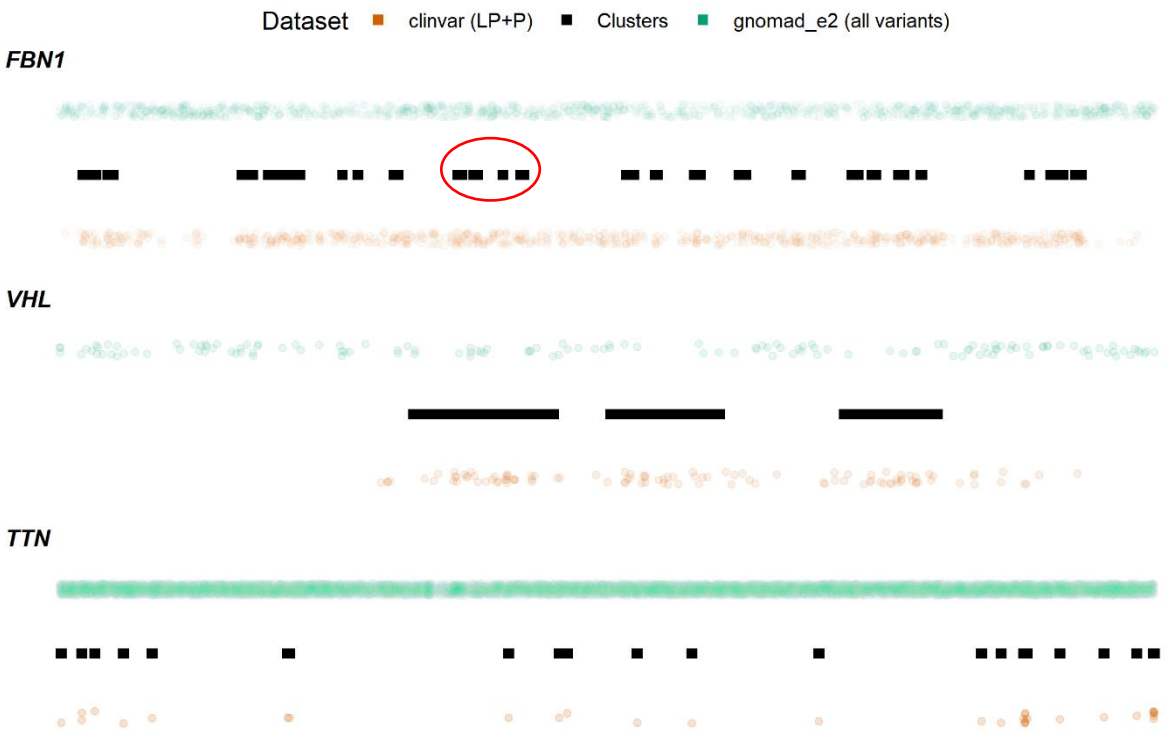

Figure 4.9: Clusters in three genes identified as extreme cases in the analysis of cluster properties. *FBN1* had the most clusters of any gene, *VHL* had the most clusters per residue and *TTN* had the fewest number of clusters per residue. The four clusters in *FBN1* highlighted by the red circle cover exons 24-32, a region reported to have the highest case variants.

As the number of clusters observed correlated with protein length (Rho=0.67), the number of clusters per residue was calculated to identify genes with a high number of clusters relative to their size. The

proteins ranges from 0.06 clusters per 100 residues to 1.4 clusters per 100 residues. The maximum was in the small protein *VHL* (n=213), which had three clusters (Figure 4.8) and the minimum was in massive *TTN,* which had 22 clusters (Figure 4.8). *VHL* is a tumour suppressor gene associated with Von Hippel-Lindau Syndrome and pathogenic variants have been reported to cluster in two functionally important regions between residues 91-113 and 157-171 (Arunachal *et al.* 2016). The first two clusters mostly overlap the first reported cluster and the third cluster here almost exactly matches the second reported cluster. The clusters in *VHL* seem to overlap with groups of LP/P variants, whereas the *TTN* "clusters" seem to result from single variants against a vast background of control variants. For this reason, in subsequent analyses clusters comprised of single variants were excluded.

When restricting clusters to those with at least 2 case variants, 707 clusters remain across 283 genes. Analysis of the relative excess of case variants and relative depletion of control variants in these clusters was compared. Again expected variant numbers were calculated based on uniformly distributed variants across the entire protein. The mean excess of case-variants was 8.3 with a maximum of 137 seen in *TTN*. This was derived from a cluster of only five variants spanning residues 31702 to 31742 at exons 341 and 342. All five variants were associated with hereditary myopathy with early respiratory failure in *clinvar* and are rare in *gnomad_e2*. The second biggest excess was also in *TTN* and the third was in the large *MACF1* gene (nresidues= 5430), indicating a bias towards larger proteins. In fact, in the top 30 clusters with the highest case excess, all but one had a protein length over 1000. The exception was a small cluster in *GRIA4* (nresidues=901) of which 4/5 pathogenic *clinvar* variants fall within.

In the corresponding control regions, the mean relative depletion was 0.55, indicating that on average roughly half of the expected variation was seen in these regions. Some of the clusters (n=43), had zero observed control variants. The mean size of these clusters was only 23 residues. The largest cluster with no associated background variation was observed in *GRIN1,* which has a majority classification of 'Neurodevelopmental disorder with or without hyperkinetic movements and seizures' in *clinvar*.

Between residues 628 and 693, 15 of 37 case variants are observed but 0 of 168 *gnomad_e2* variants are observed. This cluster is in the centre of the transmembrane domains that form the ion channel pore of this receptor and variant clustering here has been previously reported (Lemke *et al.* 2016).

The simplified uniform background assumption has multiple weaknesses. For the controls, uneven mutation rates and coverage affect this background distribution. Although coverage was controlled for, mutation rates were not. Fortunately, population constraint estimates stratified by both coverage and mutation rates have been calculated by Havrilla *et al.* 2019. Cluster regions were overlapped with constrained-coding region percentiles. For each cluster, the mean CCR was calculated, the mean of all CCR scores for each cluster was 77, whereas the mean across the non-cluster positions in the same genes was only 42. This significant difference in CCRs ($p<2.2e-16$) provides more evidence that there is genuine depletion of control variants in the identified clusters.

Of the *b-significant* genes in the *all-lp+* and *all-vus+* analysis, 18 further genes were identified only with the inclusion of VUS variants and 157 were identified with VUS variants excluded. The mean –log10 p-values for each gene group was largely similar (*all-vus+* = 8.94; *all-lp+* = 9.01). Although slightly higher in the all-*vus+* exclusive group (2.14 vs 2.09), the mean missense Z score was not significantly different ($p < 0.87$). One difference between the gene groups, was that a higher proportion of genes in the LP/P exclusive approach were annotated with a functional domain (0.33, 0.65, $p < 0.01$). Clusters were assessed in genes only identified when including VUS variants using the clustering algorithm (Figure 4.10). Some interesting patterns appear. In *ATN1* and *YY1*, small C-terminal clusters of pathogenic variants are enhanced by additional VUS variants occurring at the same locations In *DST*, *RNF213*, *PALLD* and *CAMTA1*, only VUS variants are available and map to specific clusters. In other genes, the clusters either enhance that of the pathogenic variants or define their own clusters. For example, in the far C-terminal of *UGT1A4*, a VUS cluster is located at a control depleted site. In *SMAD3*, an important tumour suppressor gene, broad differences in variant densities in the C and N-terminal between VUS variants and control variation predict two clusters. These results potentially highlight

genes where variant clustering may be a useful and an underutilized signal in variant classification. These may represent targets where the GAM *hotspot modelling* approach in case-control cohorts may be effective.

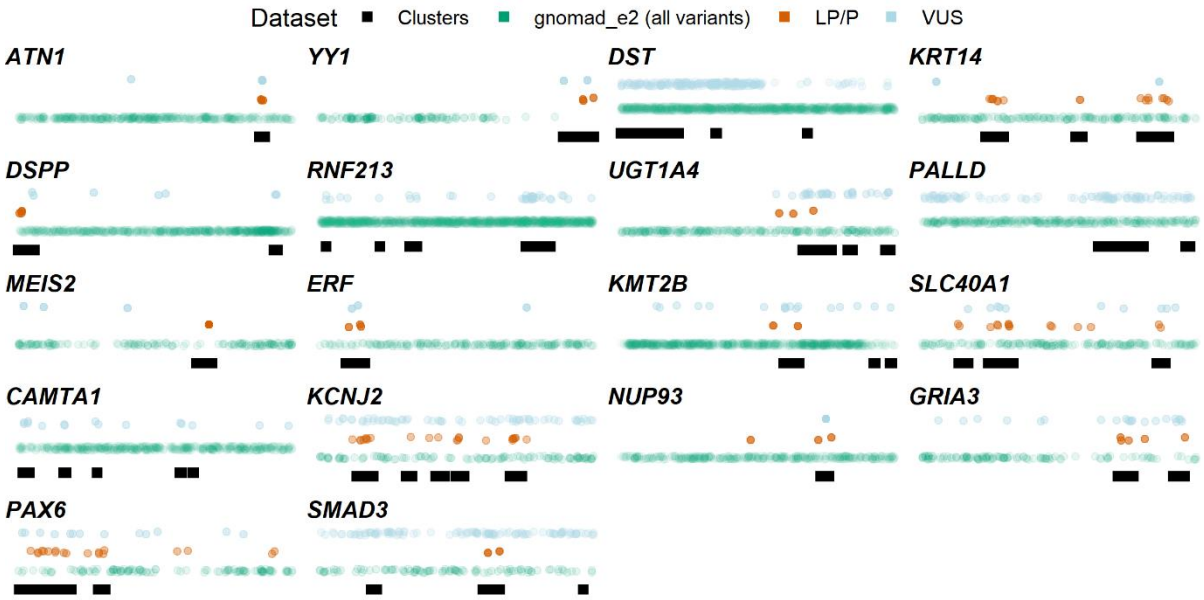

Figure 4.11: Clusters in genes only *b-significant* when VUSs were included alongside LP and P variants. Blue points are c*linvar* VUSs, orange points are *clinvar* LP/Ps, green points are *gnomad_e2* variants and black rectangles are the inferred clusters, presumably driven by both VUS and LP/P variants or only VUSs in the genes without variants annotated as pathogenic (*CAMTA1, DST, RNF213* and *PALLD*).

4.3.5 Overlap with protein features

Clusters were annotated with any UniProt protein features that they overlapped with including all features noted in Table 4.7. Ninety percent (n=1093) of the clusters overlapped with at least one feature, 39% overlapped with a functional domain and 26% overlapped with a functional region. The most overlaps were observed in a 121 residue cluster in *MAPT*, which overlapped with twelve cross-link annotations, thirteen structure annotations, five glycosylations, one disulphide, three repeats and 1 functional region: Microtubule-binding domain.

Table 4.7: UniProt protein feature types and some examples of subtypes within each feature. Datasets were annotated with features from the groups below and overlap with clusters was assessed.

Feature	N overlaps	Example feature names
Topological domain	419	Cytoplasmic, Extracellular, Lumenal
Transmembrane	219	Helical
Domain	518	ABC transmembrane type-1 2, ABC transporter 2
Region	419	Interaction with PEX19, Required for homodimerisation
Structure	1486	Beta strand, Turn, Helix
Motif	33	Kinase activation loop, Basolateral sorting signal
Repeat	131	ANK 2, ANK 3, ARM 2
Compositional-bias	41	Pro-rich, Leu-rich, Ser-rich
Glycosylation	44	O-linked (HexNAc...) tyrosine, N-linked (GlcNAc...) asparagine
Peptide	2	P3(40), P3(42)
Zinc-Finger	54	NR C4-type, GATA-type, PHD-type
Cross-link	46	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin
Coiled-coil	12	Coiled-coil
Intramembrane	27	ECO:0000269|PubMed:29809153
Disulphide	200	Interchain, Or C-1586 with C-1678
Signal peptide	2	Signal peptide
Lipidation	2	S-palmitoyl cysteine

For the 518 clusters that overlapped with a domain, on average 89% of the cluster residues fell within the domain boundaries including 388 clusters where every residue was within the domain. On the other hand, the mean proportion of each domain that the clusters covered was only 17% indicating that clusters were smaller than domains. Some clusters overlapped strongly with specified domains (Figure 4.10).

In the *DCTN1* gene associated with neurodegeneration, the single cluster overlapped strongly with a CAP-Gly domain. All confident pathogenic variants in this protein have been identified in this domain which binds microtubules (Vilariño-Güell *et al.* 2009). In the *DEAF1* gene, associated with neurodevelopmental delay, pathogenic variants strongly overlapped with the SAND domain, a finding that is well-known (Nabais *et al.* 2020). In *FBN2*, associated with Contractural Arachnodactyly and Macular Degeneration, there is a strong overlap between the central EGF-like 17 domain and the identified cluster. Clustering in this gene has been reported to occur between exon 24-34 (Ratnapriya *et al.* 2014) covering this domain and the tail of pathogenic variants extending towards the C-terminus. In *FLNB*, associated with Larsen syndrome, variants have been reported to cluster in the Calponin-

homology domain as observed here and also repeat 14 and 15 (Zhang *et al.* 2006). *MEF2C* is a MADS-box transcription factor in which variants cluster in the MADS-box domain. For *NOTCH3*, although pathogenic variants are observed across the N-terminal region of the protein, a strong cluster was observed in the F-like 1 domain. For *NSD1*, a cluster overlapping with the SET domain was identified however more pathogenic variants flank this region. For *OFD1*, overlap with the *OFD1* domain is observed. For *SHANK3*, pathogenic variants overlap with the C-terminal SAM domain. For the top 20 genes with the best overlap between domains and clusters, 15 were annotated by GO as being involved in development. Five were involved in extracellular matrices (*UMOD*, *VWF*, *COL6A3*, *FBN2* and *FBN1*) and six were involved in calcium ion binding (*UMOD*, *STIM1*, *FBN1*, *FBN2*, *NOTCH2* and *NOTCH3*).

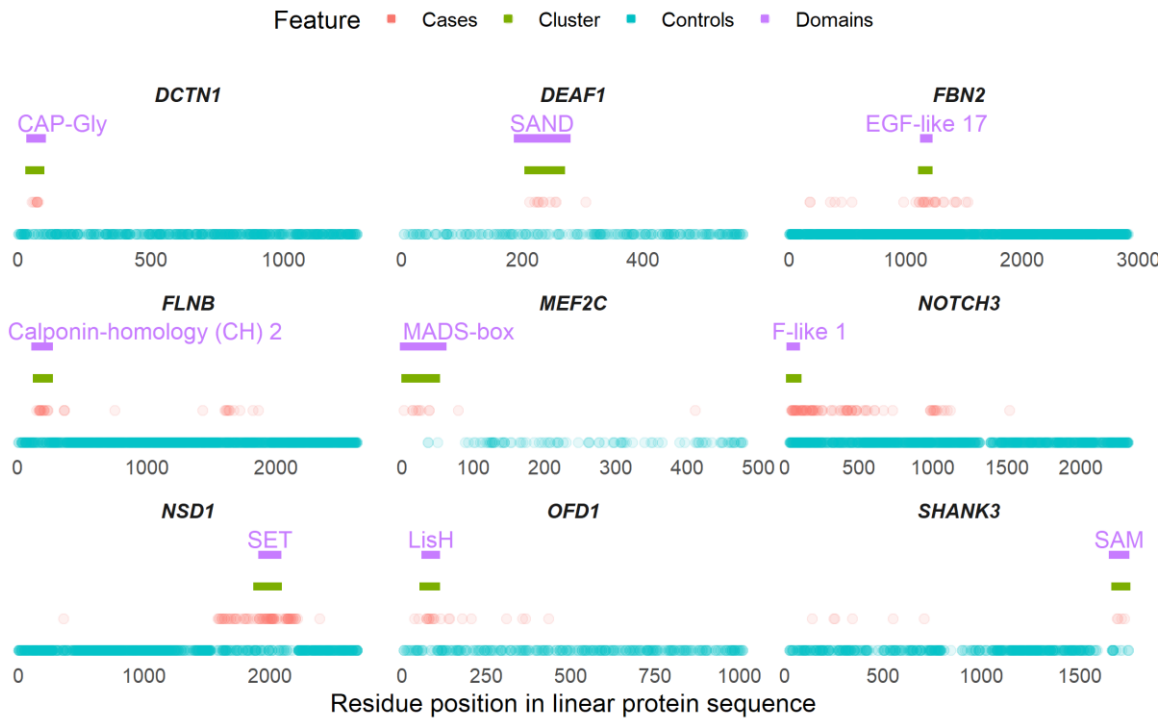

Figure 4.10: Identified clusters and functional protein domains with the highest proportion of residues overlapping in each group. Only the overlapping cluster and domain are displayed. Case (red) and control (blue) variants are plotted below the domain (purple) and predicted cluster (green). Domains are labelled by their name within each subplot.

Secondary structural information comprising either; beta-strand, alpha-helix or turn, was available from UniProt. For each gene, the total number of clustered and non-clustered residues overlapping with each of these secondary structure categories was determined. For each gene, a Fisher's exact test was used to determine if there was a significant difference in the proportion of clustered residues that overlapped with these structural features. Overall 293 genes (*b-significant all-lp+*) were assessed for overlap with structural annotation. Of these genes, 210 have at least one region annotated as structurally helical, 140 of these genes had a p-value below 0.05 and 96 had a p-value below 0.00024 (Bonferroni corrected for 210 genes). Slightly fewer genes (n=187) had a beta strand annotation, of these 122 had a p-value below 0.05 and 69 had a p-value below 0.00027 (Bonferroni corrected for 187 genes). The final group "turn", was annotated in 174 genes, 52 of which had p-values below 0.05 and 13 with p-values below 0.00096 (Bonferroni corrected for 52 genes). The same test was then applied to all genes combined for each secondary structure category. For all categories, the p-value was calculated as 2.2×10^{-16} reaching the lower limit of the software. Helical regions had the greatest effect size OR=2.4 [2.32-2.48], followed by "turn" OR=2.16 [1.96-2.39] and lastly beta-strand OR=1.89 [1.81-1.97]. These results suggest that perhaps these regions could be weighted more heavily in a positional analysis. As secondary structural information is readily available for all genes, this would be relatively easy to accomplish.

4.4 Inherited heart disease genes

Eight of the inherited heart disease gene-panel genes (n=35) were observed amongst the *b-significant* genes in the analysis of distribution between *clinvar* pathogenic and all *gnomad_e2* variants. When including VUS variants, only six genes are observed missing *KCND3*, *DMD* and *LMNA* and gaining *KCNJ2*. All genes except *KCND3* and *RYR2* were picked up in the previous analyses of rare-missense regional burden (chapter 3.2). Here, as the majority classification for *KCND3* this gene is spinocerebellar ataxia, it is likely that the clustering is specific to that disease. In fact, it was not shown to be associated with LQTS or BrS in our data. The same is true of *RYR2* associated with

catecholaminergic polymorphic ventricular tachycardia. Although this is an inherited heart disease sequenced at OMGL, so few samples were available it was not included in the analysis. The LQTS genes *KCNH2* and *KCNQ1*, as well as the HCM gene *MYH7*, demonstrated the same pattern of clustering as observed by the *hotspot models* in chapter 3.2. The association of *DMD* here is driven by Becker Muscular dystrophy associated pathogenic variants in the tails of the protein with corresponding control depletion in the same regions. Interestingly, the presumed artefactual association with HCM found an excess of in the centre of this protein (Figure 3.5). Here pathogenic *LMNA* variants are seen in excess in the N-terminus whereas control variants are modestly depleted here. Interestingly, the weak cluster observed in the DCM samples corresponds with this (Figure 3.5). For *SCN5A*, a strong BrS associated cluster is observed in the C-terminus of the protein, again this strongly corresponds to what was observed in the *hotspot models* (Figure 3.5).

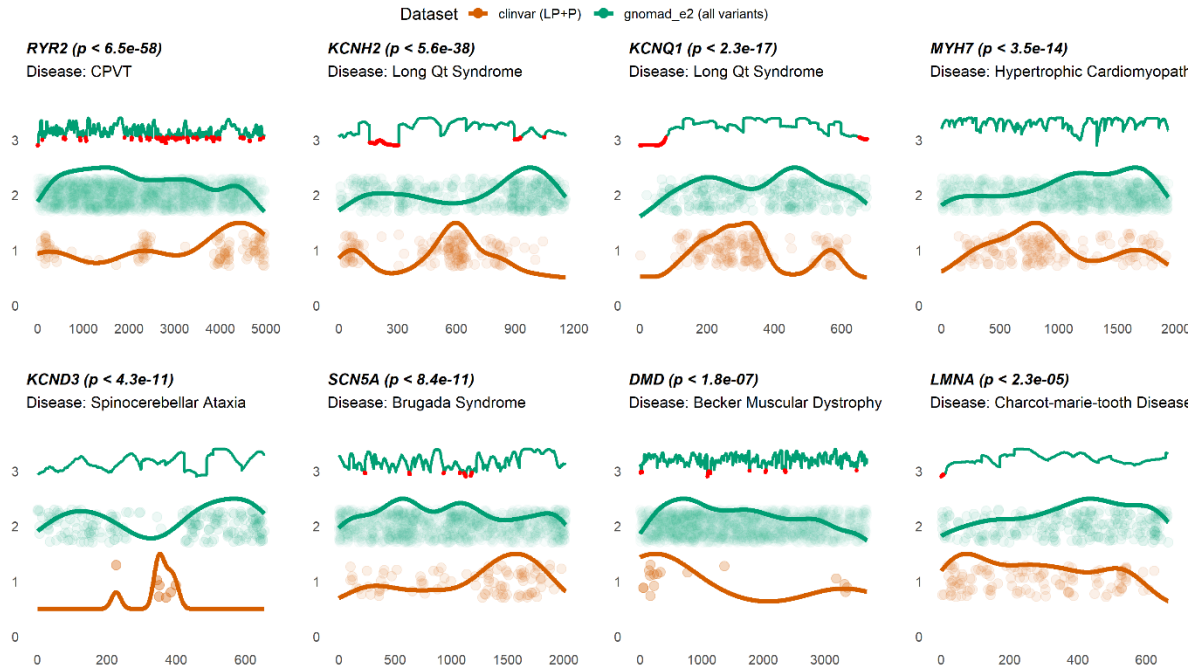

Figure 4.12: Significant clustering of pathogenic *clinvar* variants and all *gnomad_e2* variants in inherited heart disease genes investigated in chapters 2 and 3. Coverage for *gnomad_e2* is displayed, however coverage was unavailable for *clinvar*. The majority disease-association submitted with *clinvar* variants is displayed under the gene name above each plot. The clustering p-value is displayed alongside each gene name.

4.5 Discussion

In this chapter, the prevalence of missense-variant clustering in Mendelian disease-genes was explored by drawing case-variants from the *clinvar* database of clinically relevant variants and control-variants from GnomAD. The *BIN-test* scan was applied for different case-control datasets defined by specific filtering strategies and many significant results emerged from each analysis. The properties of significant clustered genes and missense-variant clusters themselves were then investigated. As well as, clearly capturing the ubiquity of missense-variant clustering, the results here show promising strides into the prior identification of clustered-genes and cluster locations, both of which could be used to improve *BIN-test* and *ClusterBurden* exome-scans in the future. These preliminary analyses suggest, specific protein classes are more likely to harbour variant clusters (e.g. ion channels). These clusters are constrained for both rare and common variants, they overlap strongly with previously classified pathogenic variants, they tend to overlap with protein domains and regions with defined secondary structure. Continuation of this project to incorporate more data and more features may yield great insight into cluster detection strategies and mutational hotspot inference.

The putative null model (*gnomad_e2* versus *gnomad_g2*) revealed that the proposed coverage control approach was effective in reducing the number of false-positives. However, significant signals still remained and uneven coverage, as well as high protein length, are characteristics of significantly clustered genes in this analysis. This may indicate that coverage control is still imperfect and that clustering in large proteins should be interpreted with caution. At the very least, these factors need to be considered as competing hypotheses for significant results. Of course, it does not make sense that protein length could be a confounder in its own right. Results from chapter 2.4.3 indicate no relationship between type-1 error and protein length. The likely explanation for its appearance as such, is its correlation with power. This causes any underlying biases in the experimental setup to be amplified in large genes. For example, if the call-rate was very-marginally higher in one cohort due to the sequencing technology or bioinformatics processing, then this baseline difference would probably

be undetected in small genes with few data. However, in large genes, although the effect size is still small, the large number of observations gives high power to significantly detect this small effect.

The filtering strategy that yielded the most signals, in both relative and absolute terms, was the analysis using every *gnomad_e2* variant regardless of frequency and limiting *clinvar* variants to only those with a likely pathogenic and pathogenic classification. Although this reduced the gene set to only 1363 genes, 654 of these were *n-significant* and 288 of these were *b-significant*. Analysis of the clusters driving the *b-significant* signals revealed many genes where strong grouping of pathogenic variants was observed in specific protein regions. When analysis was broadened to include VUSs, although many of the same genes were detected, variants were more uniformly distributed often obscuring neat clusters detected with only pathogenic variants. This may partly explain their VUS status and indicates the importance of protein position in variant interpretation.

Although some identified clusters overlapped strongly with protein domains (Figure 4.10), domain boundaries were not definitive delineators of cluster locations. This further supports data-driven approaches to cluster identification, as opposed to a naïve functional domain specification. This mismatch may be due to imperfect resolution in experimentally derived domains or the possibility that functional domains can be influenced by variants falling outside their linear boundaries. This is potentially a problem that better defined three-dimensional structures can resolve. Overall, genes with significant clustering were more likely to harbour a known functional domain compared to other *clinvar* genes, and clusters overlapped with these domains more than expected by chance. This indicates that functional domains (or motifs) may be useful for prior weighting of clustered genes or gene regions.

Global missense constraint Z-scores were higher in clusters detected by VUS/LP/P variants than just LP/P variants in both *n-significant* and *b-significant* genes lists (Table 4.6). Although not-significant which may be a result of relatively low samples sizes, this may be attributable to different disease-mechanisms for each group. Many LP/P variants clusters are small and defined and may indicate DN

or GoF effects. Conversely VUS/LP/P variant clusters are sometimes more subtle and widespread and may indicate HI mechanisms. For the DN and GoF, as only small regions are strongly associated with disease outcomes, the overall constraint is reduced at a gene-wide level. Therefore, ranking by global missense constraint may not be slightly less effective in genes whose disease mechanism is non-HI. Furthermore, as the global burden differences may also be reduced in these genes, due to regions of background variation in both cohorts, these genes may be good examples of when including both burden and positional information is useful.

4.5.1 Limitations

Although variant distribution in *gnomad_e2* controls were adjusted by coverage, there was no way to do this for the *clinvar* case variants. This is because each submission to this database is not associated with coverage data. There may be cases where historic clinical sequencing focused on specific domains hypothesised to be involved with the disease, especially in large proteins, causing artefactual clusters. Additionally, and perhaps even more important, is the self-perpetuating bias from classifications. Even when the whole gene is well-covered, variants in previously identified hotspots, may be disproportionately likely to receive a pathogenic classification. Therefore, when limiting *clinvar* variants to only LP and P, resulting clusters may need unbiased validation in complete case-series as with the *hotspot models* in chapter 2. However, as many clusters fall in regions with corresponding control depletion (relatively unbiased), this corroborates their status as genuine clusters.

Chapter 5

Frequency filtering and ancestry

5.1 Subtle confounders of using population controls

The use of large publically available population controls is highly convenient for population-genetic analyses but there usage introduces numerous potential confounders. One of these already discussed is that of uneven and unmatched coverage. Another, is the probable batch-effects arising from the consideration heterogeneous unmatched datasets, such as those originating from differences in exome-capture technologies used. These caveats cast uncertainty on the calibration of parameters estimated by statistical inference on these data and may explain the inflation in synonymous burden observed in subsection 2.4.1. In this chapter, the impact of the method of frequency filtering, broad-scale ancestry differences and the interaction between the two, are considered.

5.1.1 Determining "popmax"

The population reference database GnomAD (Karczewski *et al.* 2020) has the aggregated exome sequences of 125,748 samples. The samples within each dataset have been stratified into different ancestry groups, summarized in Table 5.1. For each dataset, although sample-level information is not provided, variants are annotated with the count and the effective allele number of each PCA-derived ancestry group. This permits ancestry-specific frequency estimation.

Table 5.1: PCA-derived ancestry groups and the maximum sample-size of the GnomAD exomes (version 2) cohort. As sequencing coverage varied across samples and genes, the sample size is effectively different for different regions of the exome.

Ancestry	**Description**	**Exomes version 2**
AFR	African/African American	7,652
AMR	Admixed American	16,791
ASJ	Ashkenazi Jewish	4,925
EAS	East Asian	8,624
FIN	Finnish	11,150
NFE	Non-Finnish European	55,860
SAS	South Asian	15,391
OTH	Other (population not assigned)	2,743

In chapters 2 and 3, a "*strict*" *popmax* filter was used, which removed any variant observed over 0.01% in any subpopulation or *gnomad_e2 gnomad_g2*. The *strict* approach included frequencies in ASJ, FIN, OTH and *gnomad_g2* in the popmax calculation*,* whereas the normal *popmax* filter, provided in the GnomAD VCF, used only the subpopulations; AFR, AMR, NFE, SAS and EAS (Lek *et al.* 2016). *Gnomad_g2* was included as an additional and separate filtering group as the data generally had fewer regions of low coverage, and may contain genuinely non-rare variants that were not observed due to exome-capture-technology-related low coverage in *gnomad_e2*. Although the Finnish (FIN) and Ashkenazi Jewish (ASJ) populations have a high number of founders variants, so could contain high-frequency pathogenic variants (Kerminen *et al.* 2017; Zlotogora 2018), it was hypothesized that this

would add more noise than signal. Finally, it was unclear why variants frequent in the heterogeneous unmatched ancestry group "other" (OTH) should not be excluded. Due to the potential sensitivity of the positional-based methods (*BIN-test* and *hotspot models*) to high-count variants, the strict approach seemed to be a harmonious strategy. However, this filter was later identified to give overly conservative estimates of allele frequency, especially in subpopulations with few samples. This caused some variants classified as pathogenic or likely pathogenic to be excluded from the analyses in chapters 2 and 3 (Table 5.2).

Table 5.2: A list of variants classified as pathogenic or likely pathogenic (by OMGL) with *strict popmax* frequencies above 0.0001 (putatively common). HCM = hypertrophic cardiomyopathy, DCM = dilated cardiomyopathy, ARVC = arrhythmogenic right-ventricular cardiomyopathy, BrS = Brugada syndrome. *Strict* here refers to the *popmax* frequency across eight *gnomad_e2* subpopulations and *gnomad_g2*.

Dataset	Gene	AC	Class	Protein	AC GnomAD	Strict	Group
ARVC	*MYBPC3*	1	P	p.Arg502Trp	10	0.000165235	OTH
DCM	*MYBPC3*	1	P	p.Arg502Trp	10	0.000165235	OTH
ARVC	*MYBPC3*	1	LP	p.Gln208His	55	0.00442923	ASJ
ARVC	*DSG2*	1	LP	p.Gly812Ser	6	0.000159276	G2
BrS	*DSG2*	1	LP	p.Gly812Ser	6	0.000159276	G2
DCM	*TTR*	1	P	p.Val142Ile	283	0.0156865	AFR
BrS	*SCN5A*	1	LP	p.Gly1319Val	10	0.000139024	AFR
DCM	*SCN5A*	1	LP	p.Gly1319Val	10	0.000139024	AFR
HCM	*MYBPC3*	4	LP	p.Arg597Gln	5	0.000119048	AMR
HCM	*MYBPC3*	2	LP	p.Gly531Arg	8	0.000166889	OTH
HCM	*MYBPC3*	96	P	p.Arg502Trp	10	0.000165235	OTH
HCM	*MYBPC3*	1	LP	p.Gln208His	55	0.00442923	ASJ
HCM	*PKP2*	3	LP	p.His877Gln	42	0.00133917	SAS
HCM	*MYL2*	7	P	p.Glu22Lys	5	0.000162866	OTH
HCM	*MYH7*	2	LP	p.Arg1053Gln	16	0.000739098	FIN
HCM	*MYH7*	9	LP	p.Arg869His	6	0.000123047	AFR
HCM	*MYH7*	26	P	p.Ala797Thr	6	0.000123031	AFR
HCM	*MYH7*	1	LP	p.Asp750His	1	0.000163079	OTH
HCM	*MYH7*	1	LP	p.Glu497Asp	2	0.000198413	ASJ
HCM	*MYH7*	6	LP	p.Arg143Trp	7	0.000108731	EAS
HCM	*TPM1*	3	P	p.Asp175Asn	4	0.000138876	FIN
HCM	*DSG2*	1	LP	p.Gly812Ser	6	0.000159276	G2
HCM	*TTR*	1	P	p.Val50Met	26	0.000488599	OTH
HCM	*TTR*	3	P	p.Val142Ile	283	0.0156865	AFR
HCM	*TNNI3*	7	LP	p.Arg162Trp	10	0.000166871	EAS
HCM	*TNNI3*	4	LP	p.Arg145Gln	4	0.000111272	EAS
HCM	*SCN5A*	1	LP	p.Arg340Gln	18	0.000258465	AFR

The list includes 20 different variants across nine genes. The well-known *MYBPC3* founder variant p.Arg502Trp (Saltzman *et al.* 2010), present once in ARVC and DCM, and present 96 times in HCM, was excluded due to its frequency in the OTH group. The pathogenic *MYH7* variant p.Ala797Thr is a known founder variant (Ross *et al.* 2017) associated with a mild/moderate form of HCM (e Mattos *et al.* 2016). This was present 26 times in HCM, but excluded due to its frequency in the *gnomad_e2* African sub-population (AFR). Other high allele count variants excluded, included *MYH7*: p.Arg869His and p.Arg143Trp, and *MYL2*: p.Glu22Lys. The exclusion of these relatively common pathogenic variants may have influenced the results of the models in chapter 3. Interestingly, for *MYBPC3*, *MYH7* and *MYL2*, pathogenic clusters were still determined in the location of the excluded variants (Figure 3.3). The largest allele count variant in the list, *TTR*: p.Val142Ile, is observed 283 times in *gnomad_e2* and is a recognized high-frequency African pathogenic variant known to cause Transthyterin cardiac amyloidosis and cardiomyopathy (Perfetto *et al.* 2015).

Although this filter was conceptually justified, the small size of some subpopulations caused inaccurate estimates of allele frequency. For example, the OTH population is only 2,743 samples. Any variant present in this dataset was filtered from the analysis, as even singletons had an estimated frequency above 0.01% (~0.018%). Further to this, in both the AFR and EAS subpopulations, all doubletons were also filtered from the analysis. This is an unfortunate and unforeseen consequence of the filtering approach. Over 50% of variants in *gnomad_e2* (53.2%) are singletons, and the mean frequency for these variants is 8.4×10^{-6}. If we take this as representative of accurate variant frequencies, it is probable that many OTH variants were in fact rare (i.e. lower than 0.00018).

5.1.2 Poisson confidence interval frequency estimates

A superior filtering approach, to avoid over-filtering, would be to take subpopulation sample-sizes into account when calculating allele frequencies. An approach to achieve this has been proposed by

Whiffin *et al.* (2017) and was integrated into the GnomAD datasets as the 'filtering allele frequency' (Karczewski *et al.* 2020). The approach is approximately equivalent to taking a lower Poisson confidence interval (pCI) of the empirical frequency estimate. For example, the Poisson lower 95% confidence interval of the MAF for a singleton observed in the OTH group (n=2,743), is 9.3×10^{-6}, passing the filter. Similarly, a variant observed twice passes this filter (6.4×10^{-5}), and a variant observed three times is excluded (1.4×10^{-4}).

The 95% pCI approach retains many variants that would have been filtered based on their frequency point estimates. By taking the lower confidence interval, it is intrinsically liberal, favouring the retention of possibly rare (and potentially pathogenic) variants, over the exclusion of non-rare (and potentially benign). This may be intuitive for variant prioritisation, particularly when limited number of genes are investigated yielding few candidate variants. However, for aggregation analysis, such as burden or clustering testing, this may preserve more noise than signal, potentially reducing power. Here, different lower confidence intervals are explored to determine a suitable balance between signal and noise.

For each subpopulation, the maximum observed count before a variant exceeds the filtering frequency threshold, was explored for different lower pCI thresholds, in Table 5.3. When using the unadjusted point estimate, simply calculated by dividing the allele count by the allele number, all variants in OTH or ASJ, all doubletons and above in AFR or EAS, all tripletons and above in FIN and all variants observed four times or more in AMR or SAS, were excluded. When using any pCI approach investigated, no singleton was excluded in any population. The least liberal threshold (pCI80), excludes any doubletons and above in OTH, but all thresholds excluded tripletons and above. For ASJ, a variant can occur twice without being filtered by any threshold but a variant observed three times or more is filtered by both pCI90 and pCI80.

Table 5.3: Point estimates (upper table) and multiple Poisson lower confidence intervals (lower table) for allele frequencies of variants occurring twice, thrice or four times in each *gnomad_e2*

ancestry group. PCI95, pCI90 and pCI80 correspond to Poisson lower confidence intervals 95%, 90% and 80% representing 2.5%, 5% and 10% lower probability tails respectively. Singleton variants were below 0.01% in all populations for all lower confidence thresholds and were therefore not displayed. Light red cells highlight frequencies above 0.01% and would therefore fail the filter at this threshold.

Pop	AN	Point estimate AC=1	Point estimate AC=2	Point estimate AC=3	Point estimate AC=4
AFR	15304	6.54E-5	1.31e-4	1.96e-4	2.61e-4
AMR	33582	2.98E-5	5.96e-5	8.93e-5	1.19e-4
ASJ	9850	1.02E-4	2.03e-4	3.05e-4	4.06e-4
EAS	17248	5.81E-5	1.16e-4	1.74e-4	2.32e-4
FIN	22300	4.48E-5	8.97e-5	1.35e-4	1.79e-4
NFE	111720	8.93E-6	1.79e-5	2.69e-5	3.58e-5
SAS	30782	3.25E-5	6.5e-5	9.75e-5	1.3e-4
OTH	5486	1.82E-4	3.65e-4	5.47e-4	7.29e-4

Pop	AN	pCI95 AC=2	pCI90 AC=2	pCI80 AC=2	pCI95 AC=3	pCI90 AC=3	pCI80 AC=3	pCI95 AC=4	pCI90 AC=4	pCI80 AC=4
AFR	15304	2.32e-5	3.47e-5	5.39e-5	5.34e-5	7.2e-5	1e-4	8.93e-5	1.14e-4	1.5e-4
AMR	33582	1.06e-5	1.58e-5	2.45e-5	2.43e-5	3.28e-5	4.57e-5	4.07e-5	5.2e-5	6.84e-5
ASJ	9850	3.61e-5	5.4e-5	8.37e-5	8.3e-5	1.12e-4	1.56e-4	1.39e-4	1.77e-4	2.33e-4
EAS	17248	2.06e-5	3.08e-5	4.78e-5	4.74e-5	6.39e-5	8.9e-5	7.92e-5	1.01e-4	1.33e-4
FIN	22300	1.59e-5	2.38e-5	3.7e-5	3.67e-5	4.94e-5	6.88e-5	6.13e-5	7.82e-5	1.03e-4
NFE	111720	3.18e-6	4.76e-6	7.38e-6	7.32e-6	9.86e-6	1.37e-5	1.22e-5	1.56e-5	2.06e-5
SAS	30782	1.15e-5	1.73e-5	2.68e-5	2.66e-5	3.58e-5	4.99e-5	4.44e-5	5.67e-5	7.46e-5
OTH	5486	6.48e-5	9.69e-5	1.5e-4	1.49e-4	2.01e-4	2.8e-4	2.49e-4	3.18e-4	4.19e-4

It is evident that the different filtering approaches would yield different datasets. The goal here is to detect the threshold which produces the optimal signal and noise ratio in the resulting dataset. To estimate this, prior classifications of variant pathogenicity could be used as a proxy for signal (pathogenic) and noise (benign). However, variant misclassifications causes bias and the usage of frequency to determine classifications (especially for benign variants) causes circularity. Therefore, the missense burden OR between HCM and *gnomad_g2* for firmly established pathogenic genes with sufficient variation to provide confident estimates, was assessed at different pCI thresholds. This

included only the core sarcomeric genes in this preliminary analysis. As a control, the synonymous burden OR was also calculated. For missense variants, the pCI 80% approach yielded the highest or equal highest OR for all eight genes. For synonymous variants, pCI 80% yielded the highest or equal highest OR for all genes excluding *ACTC1*. Overall ORs were similar between the approaches. Although conducted with limited data, these inconclusive results highlight the challenge in determining the optimal solution. Future analyses could utilise multiple firmly established genes in matched case-control datasets across UK Biobank or similar datasets.

5.1.3 Potential founder variants in Ashkenazi Jews and Finnish subpopulations

While the previous section examined effects at the aggregated-level, the following investigates individual high-frequency variants that may influence results. For our inherited heart gene-panel data, using the standard *popmax* filter (AFR, AMR, EAS, NFE, and SAS) at 0.01%, 3017 variants pass. However, 28 variants from this group have a count above 20 in ASJ and 8 have a count above 10 in FIN (Table 5.4). An LP variant in 16 FIN samples (*MYH7*: p.Arg1053Gln) is recognised as a pathogenic HCM-associated Finnish founder variant (Jääskeläinen *et al.* 2019). In ASJ, the variant *MYBPC3:* p.Gln208His is observed only once in 5,338 HCM cases but 44 times in ASJ and a further 11 times in the remainder of *gnomad_e2*. The remainder are either VUS variants or have no classification, from OMGL or *clinvar*. For *MYBPC3:* p.Gln208His and the remaining 34 variants here, no published evidence corroborating their status as a high-frequency founder could be identified. Further to this, some are already classified as benign. In conclusion, these variants may be likely to add noise to analyses. For burden, their high counts would inflate the aggregate value for that gene, for clustering, each variant may drive a cluster at its protein position.

Table 5.4: GnomAD exomes variants passing the standard *popmax* filter (<0.01%) but present in high frequencies in ASJ or FIN (putative founder variants). BrS = Brugada syndrome, ARVC = arrhythmic

right-ventricular cardiomyopathy, HCM = hypertrophic cardiomyopathy, DCM = dilated cardiomyopathy.

Disease	Gene	Type	HGVSp	Class	AC	AC (e2)	Pop	AC(pop)
BrS	*RYR2*	Synonymous	p.Ala1240=		1	40	FIN	24
ARVC	*MYBPC3*	Missense	p.Gln208His	LP	1	55	ASJ	44
BrS	*HCN4*	Synonymous	p.Arg82=		1	37	FIN	36
BrS	*TMEM43*	Missense	p.Ala57Thr		1	23	ASJ	13
BrS	*SCN5A*	Missense	p.Asp1243Asn	VUS	1	36	ASJ	25
DCM	*DSP*	Missense	p.Asp230Asn	LB	1	135	ASJ	123
DCM	*DSP*	Missense	p.Val495Met	VUS	1	47	ASJ	35
HCM	*LMNA*	Missense	p.Asn660Asp	LB	1	29	FIN	21
HCM	*ACTN2*	Missense	p.Arg852Gln		2	19	ASJ	13
HCM	*ACTN2*	Missense	p.Ala887Thr	VUS	1	36	ASJ	27
HCM	*VCL*	Synonymous	p.Phe126=		2	64	ASJ	45
HCM	*ANKRD1*	Missense	p.Tyr274His		1	25	FIN	16
HCM	*RBM20*	Synonymous	p.Ser485=		1	18	FIN	12
HCM	*RBM20*	Synonymous	p.Pro715=		3	53	ASJ	47
HCM	*RBM20*	Synonymous	p.Ser748=		1	35	ASJ	27
HCM	*MYBPC3*	Missense	p.Asp605Asn	VUS	1	42	ASJ	33
HCM	*MYBPC3*	Missense	p.Gln208His	LP	1	55	ASJ	44
HCM	*PKP2*	Missense	p.Pro276Ser		1	42	ASJ	34
HCM	*MYH7*	Missense	p.Arg1053Gln	LP	2	16	FIN	16
HCM	*MYH7*	Missense	p.Gly10Arg	VUS	1	17	ASJ	12
HCM	*FHL2*	Missense	p.Gln134His		2	46	ASJ	39
HCM	*SCN5A*	Missense	p.Ala185Thr		1	98	FIN	92
HCM	*DSP*	Missense	p.Asp230Asn	LB	3	135	ASJ	123
HCM	*DSP*	Missense	p.Val495Met	VUS	1	47	ASJ	35
HCM	*DSP*	Synonymous	p.Ala1505=		1	50	FIN	47
HCM	*DSP*	Missense	p.Ser2821Leu		1	13	ASJ	11
HCM	*FLNC*	Synonymous	p.Pro856=		2	117	ASJ	103
HCM	*FLNC*	Missense	p.Gly1674Ser	VUS	3	37	ASJ	32

Only high-frequency founders present at least once in the heart data are present in Table 5.4. Across *gnomad_e2* in the same genes, 131 high-frequency variants are observed, including 67 in ASJ and 64 in FIN. The mean allele count across is 34 and ranges from 11 to 139. No high-frequency variants (ac>10) were observed in OTH, supporting the heterogeneous nature of this group. No further LP/P

variants were observed here in OMGL or *clinvar*. These variants, which are probable founder variants due to the discrepancy in their frequencies between ASJ or FIN and other populations, could add bias in burden or position testing. However, the bias would be proportional to how mismatched the ancestry of cases and controls were. For the case of using *gnomad_e2* as controls, as these founder populations are well studied, a disproportionate number of FIN and ASJ samples entered this dataset. Therefore, for most associated case-datasets, including the heart-data used here, these variants would induce bias into the analysis. In light of this, a *strict popmax* is recommended, providing that over-filtering is avoided by using lower pCI estimates of variant frequencies.

5.1.4 Revisiting hotspot models for core HCM genes

The core sarcomeric genes; *MYH7*, *MYBPC3*, *TNNI3*, *TNNT2*, *MYL2*, *MYL3*, *TPM1* and *ACTC1* were remodeled using the *strict popmax* of lower 90% pCI estimates of allele frequency (Figure 5.1). On visual inspection, model predictions remain stable and follow the same patterns observed in chapter 3.2. However, when predictions were stratified by supporting (OR>=10) and moderate (OR>=20) evidence of pathogenicity, small but important differences are observed. Some protein regions such as the centre of the N-terminal motor domain in *MYH7*, the first N-terminal cluster of *MYBPC3* and the only clusters in *MYL2* and *ACTC1*, lose moderate evidence of pathogenicity for some residues. For *MYL3*, the excess in the C-terminal is now high enough to reach supporting evidence (OR>=10).

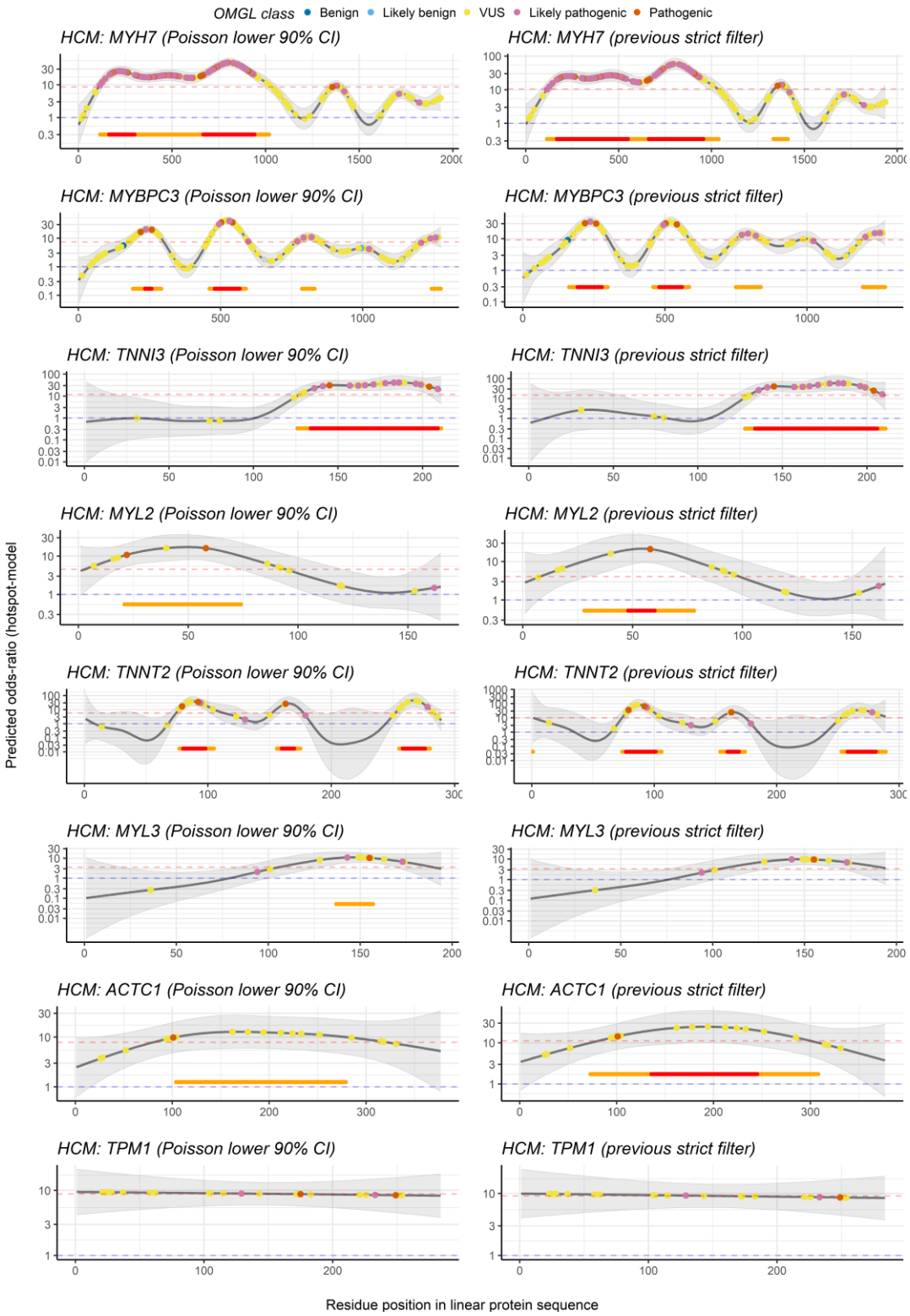

Figure 5.1: *Hotspot models* for core sarcomeric HCM-associated genes with Poisson 90% lower confidence interval frequency estimates for variants (left) versus point estimates of frequency (right). On the lower section of each subplot, red and orange lines represent regions of moderate and supporting evidence of pathogenicity.

5.2 Burden by ancestry

5.2.1 Mean exonic burden

For each variant in *gnomad_e2*, allele counts for each PCA-derived ancestry group is provided with the publicly available VCF. Here, the burden of missense and synonymous variation per ancestry group is explored. Ancestry-specific allele frequencies were first re-estimated using the lower pCI 90% approach outlined in section 5.1. Only sites with at least 50% of samples covered to 10X in all ancestry groups were considered. This left 5,068,797 missense variants and 2,419,076 synonymous variants. The *Strict popmax* variant-frequency (*pci90_strict*) was then employed to stratify variants into the non-overlapping frequency bands; <0.01% (*very-rare*), 0.01% - 0.1% (*rare*) and 0.1% - 1% (*uncommon*). For each band and ancestry group, the number of missense and synonymous alternative alleles were counted. This was then divided by the sample size of each group to estimate the average number of exonic variants per exome (Figure 5.2).

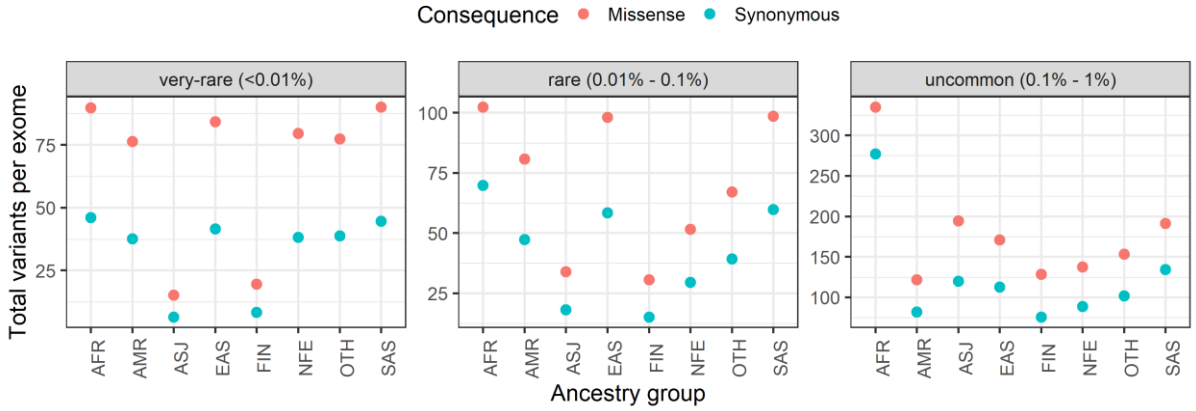

Figure 5.2: Mean counts of missense and synonymous variants per exome for eight ancestry groups in *gnomad_e2*. The frequency filtering method was conducted by calculating the maximum of all Poisson lower 90% confidence allele frequency estimates in all eight GnomAD exomes subpopulations.

For *very-rare* variants, the mean exonic burden was substantially lower in the founder populations FIN and ASJ. These ancestry groups are not the smallest in *gnomad_e2,* eliminating sample-size related artefacts as the driver of this signal. Both these populations were excluded from the provided *popmax*

frequency provided in the *gnomad_e2* dataset due to their strong founder effects (Lek *et al.* 2016). FIN is a genetically isolated founder population (Kerminen *et al.* 2017; Kääriäinen et al. 2017) widely utilised in population genetics studies as disease-causing variation may reach relatively high frequencies (Lim *et al.* 2014). It is previously noted that rare neutral variation is reduced in this population, as observed here. ASJ is another well-studied founder population, similarly characterised by high frequency deleterious alleles and a reduction of neutral rare-variation (Slatkin 2004; Simons and Sella 2016; Zlotogora 2018). Outside of these two well-known founder populations, the mean burden of missense and synonymous variation between ancestry groups (AFR, AMR, EAS, NFE, OTH and SAS) are similar in the *very-rare* band. Statistical comparison of these means could not be undertaken, as without sample-level information, variance could not be estimated.

Global inter-ancestry differences in *very-rare* missense burden may bias burden tests in unmatched cohorts. In an extreme scenario, ASJ controls and AFR cases, the baseline burden differences would give an odds-ratio of 6.26, surely resulting in false-positives. Ignoring, ASJ and FIN, the next most extreme comparison in this data is of AMR and AFR giving a baseline OR difference of 1.18. In a more realistic scenario of NFE-dominated mixed ancestry cohorts, this OR may fall closer to 1. Nevertheless, it is still likely to be an important bias in burden testing.

For the *rare* band (*pci90_strict* 0.01% - 0.1%), although FIN and ASJ display the least variation per exome again, they are closely followed by NFE. In this frequency band, more inter-ancestry variance is seen in the non-founder populations. For *uncommon* variants (*pci90_strict* 0.1% - 1%), the AFR population shows a drastic increase compared to other populations. The increase in *uncommon* variation in AFR, compared to other populations, is usually attributed to the high degree of within continent genetic diversity, which is higher than any other ancestry groups (Rotimi *et al.* 2017) supporting the 'out-of-Africa' hypothesis (Shriner *et al.* 2015). Although FIN and ASJ demonstrated similar patterns in the *very-rare* and *rare* bands, they deviate for *uncommon* variants, with ASJ having the second highest mean burden in this group. This undoubtedly reflects interesting differences in the

demographic histories of these founder populations, though identifying the nature of this is beyond the scope of this chapter.

The burden of missense variants is consistently higher than synonymous variants; most likely attributable to the increased empirical probability of a random mutation event causing a missense rather than a synonymous variant. As the frequency band increases, the relative difference between the two variant types decreases. This is probably due to purifying selection over many generations. Noteworthy across different ancestry groups, is the consistency in the proportional difference between missense and synonymous variant burden. Table 5.5 displays the mean synonymous variant burden divided by the mean missense burden. For all MAF groups, ASJ and FIN are again the outliers, however this could not be statistically evaluated without sample-level information. For *very-rare* variants, the relative difference is increased indicating more missense variants relative to synonymous variants were observed. The same is observed in the *rare* band and to a lesser extent in the *uncommon* band and is possibly a consequence of decreased purifying selection in this group. Also consistent outliers across MAF groups, was AFR, which had fewer missense variants relative to synonymous variants, perhaps indicating increased purifying selection in this population.

Table 5.5: Ratio of missense to synonymous variants in each population at different MAF groups. Heat coding assists in pinpointing potential outliers.

Population	Very-rare (<0.01%)	Rare (0.01% - 0.1%)	Uncommon (0.1% - 1%)
AFR	1.95	1.47	1.21
AMR	2.03	1.71	1.48
ASJ	2.36	1.88	1.62
EAS	2.03	1.68	1.52
FIN	2.34	2.03	1.7
NFE	2.09	1.75	1.55
OTH	2	1.71	1.5
SAS	2.02	1.65	1.42

5.2.2 Different frequency filtering methods

To determine the impact of ancestry-specific exonic burden using different frequency filtering approaches, the analysis in subsection 5.2.1 was repeated with; the pre-annotated *gnomad_e2 popmax* filter, and the lower 90% pCI confidence adjustments for the global (pci_global), *popmax* (*pci_popmax*) and *strict* (*pci_strict*) filters (Figure 5.3). Perhaps the most notable difference is between the *pci_strict* and *pci_popmax* filters. In the *pci_strict* approach, which includes ASJ and FIN in the *popmax* calculation, there is a clear depletion of *very-rare* variants in these populations (as in subsection 5.21). Conversely, if they are excluded from the calculation, the mean exonic burden in these groups is closer to the other ancestry groups, albeit with FIN still exhibiting the lowest mean across ancestries. Despite this change, it is important to recognise that the FIN and ASJ *very-rare* variants (<0.01% frequency band) under the *pci_popmax* filter, is a heterogeneous group of both rare variants and common founder variants not excluded by their frequency in AMR, AFR, EAS, NFE or SAS. Therefore, the decreased variance in inter-ancestry mean-exonic-burden is due to chance, as high-frequency founder variants inflate the mean burden.

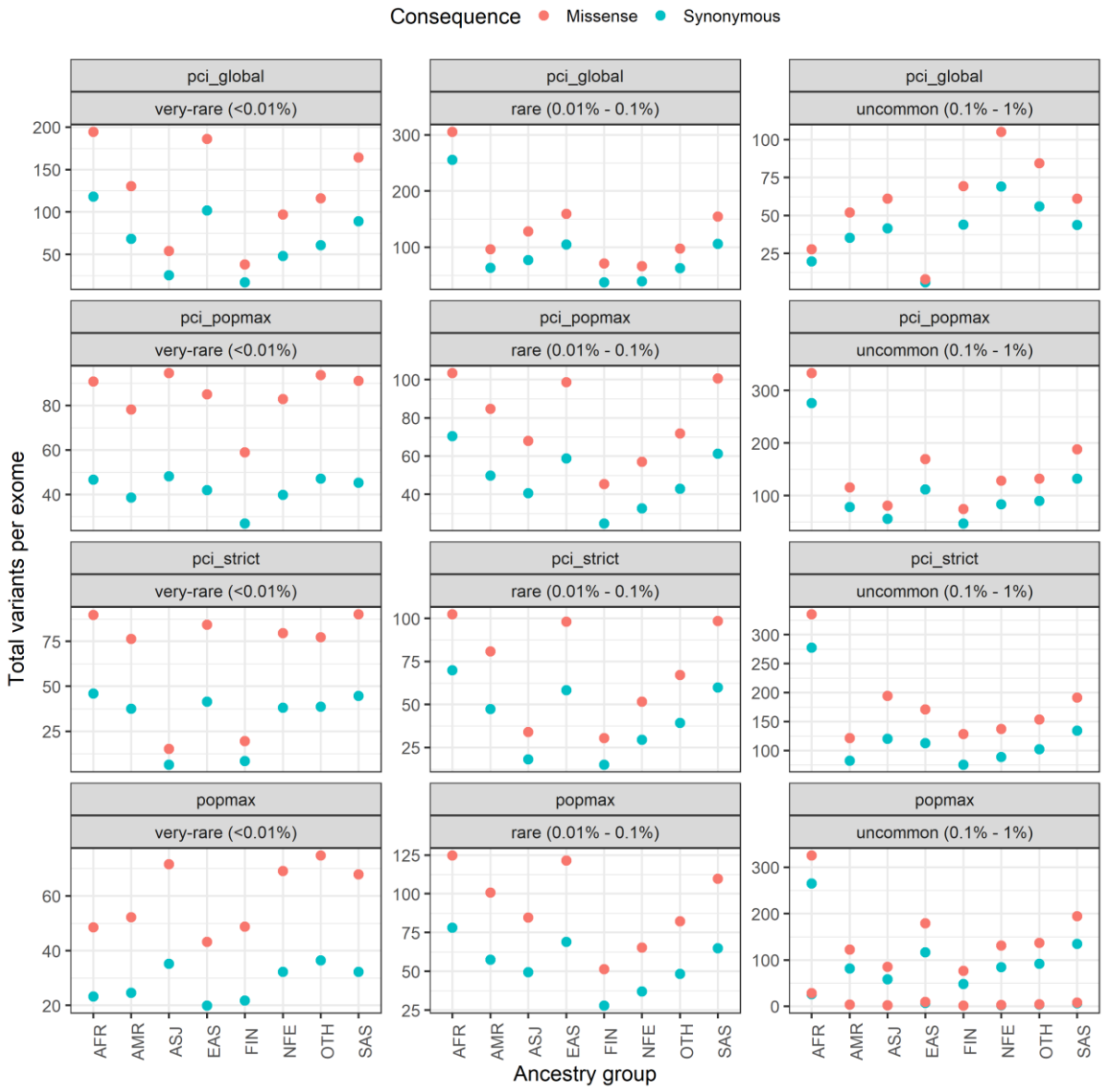

Figure 5.3: Mean counts of missense and synonymous variants per exome for eight ancestry groups in *gnomad_e2*. The frequency filtering method (pci_strict) was conducted by calculating the maximum of all Poisson lower 90% confidence allele frequency estimates in all eight GnomAD exomes subpopulations. For pci_popmax, ASJ, FIN and OTH were excluded from this calculation and for pci_global the Poisson estimate was of the total frequency across all samples. The popmax filter, refers to the unadjusted pre-annotated filter excluding ASJ, FIN and OTH.

For *very-rare* variants, there are roughly twice as many variants per exome using the *pci_global* filter compared to the *popmax* filtering approaches (*pci_popmax, pci_strict* and *popmax*). While FIN and ASJ harbour the least variation again, when using the *pci_global* filter, there is also increased inter-

ancestry variability in other ancestry groups, with patterns comparable to the *rare* band in the *popmax*-based approaches. The standard *popmax* filter, without pCI adjustment, results in fewer *very-rare* variants per exome and a different pattern of ancestry specific burden. When excluded from the *popmax* filter, ASJ now has one of the highest mean exonic burden of *very-rare* variants. However, as this population was not used for filtering, variants frequent in this population but not in other populations (founders) were included amongst these *very-rare* counts. The AFR and EAS are now among the lowest mean-burden groups, probably due to their small sizes and the consequent over-filtering when using point estimates of variant frequencies (Table 5.3).

For *rare* and *uncommon* variants, the patterns of variation for all *popmax*-based approaches was broadly similar but with a noticeable decrease in variation for ASJ and FIN using the *strict* approach. The global filter again produces patterns of ancestry-specific burden similar to the frequency band above (0.1% - 1%) in *popmax*-based approaches. One hypothesis to explain this is that the global filter fails to exclude non-rare-population specific variants, hence generating a pattern of mean-burden more similar to variants in the frequency band above.

5.2.3 Variance in ancestry-specific burden across genes

In the previous subsection, it was noted that excluding ASJ from the *popmax* calculation results in a mean exonic burden of *very-rare* variants comparable with other populations (Figure 4.13). As excluding this population from the *popmax* calculation results in a heterogeneous variant variant-set, containing both rare and founder variants, this was almost certainly driven by high-frequency founder variants overinflating the mean exonic burden estimate. If this was the case, then founder variants would not equally impact all genes. To explore this, missense burden adjusted for protein length, was calculated at every gene. Then, the mean and standard deviation of these gene-level burdens were calculated for each population under both *pci_popmax* and *pci_strict* filters (Figure 4.15).

As expected, although comparable in mean exonic burden (Figure 5.4) and mean burden per gene (Figure 4.15), the standard deviation in burden across genes in ASJ and FIN were the highest of all subpopulations by a wide margin. Whereas, when *strict* filtering was employed, although the mean burden was markedly different in these populations, the standard deviation was visible lower. This was the case for both *very-rare* (<0.01%) and *rare* (0.01% - 0.1%) variants. Consequently, these populations may confound burden testing under both the *pci_popmax* or *pci_strict* filter. For *pci_strict*, reduced rare variation causes systematic bias in baseline burden across all genes. For *pci_popmax*, fluctuations in the presence of founder-variants between genes may cause stochastic bias, whereby founder variants in some genes can inflate burden ORs or cause artefactual "clusters".

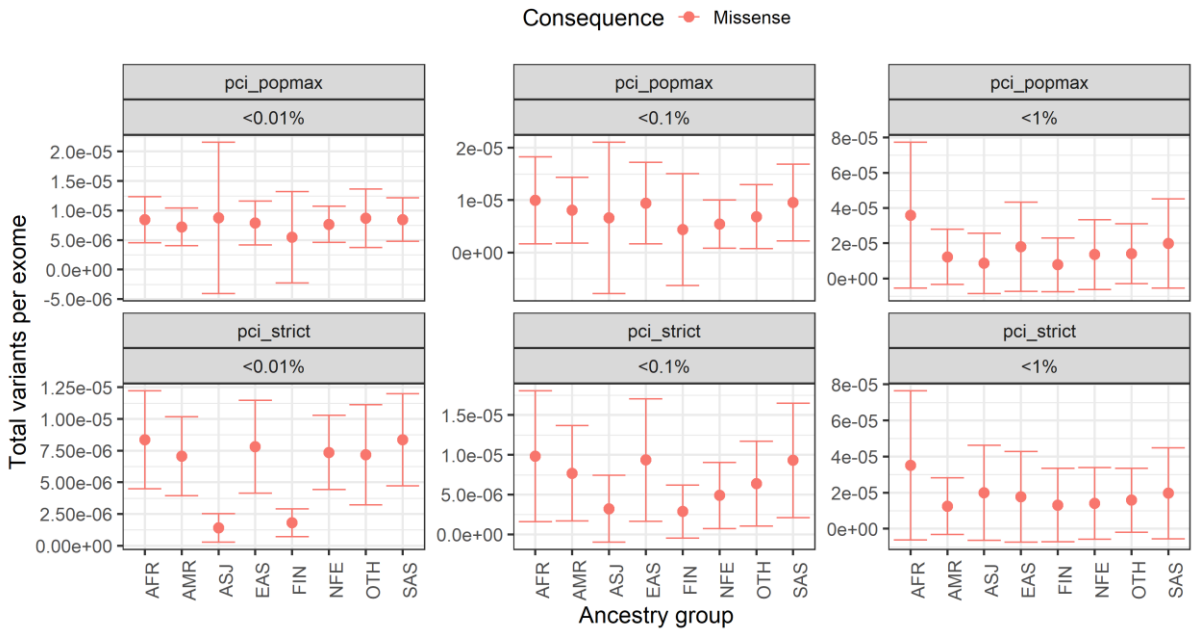

Figure 5.4: Summary of missense burden across all genes in the exome for different ancestry groups, frequency bands and filtering methods. The mean (point) and standard deviation (error bar) summarise 18,736 ancestry-specific gene burdens.

The coefficient of variation (*coefvar*: standard deviation divided by the mean) was calculated for each ancestry group and filtering mode (Table 5.6). Using the *pci_popmax* filter, ASJ had the highest *coefvar*

across gene burdens, with a standard deviation 1.5 times bigger than its mean. This was closely followed by FIN, which has a *coefvar* of 1.4. The OTH group, also excluded from the *popmax* calculation, also had a higher *coefvar* than AFR, AMR, EAS, NFE and SAS. This pattern persisted into the *rare* frequency band. For FIN and ASJ, the increased variation in burden between genes even extended into the uncommon frequency band, though to a lesser extent. When considering the *pci_strict* filter, although variation is reduced compared to the *pci_popmax* approach, with low standard deviation on visual inspection (Figure 4.15), *coefvars* are still larger for *very-rare* variants in FIN, ASJ and OTH and for *rare* variants in ASJ and FIN. Therefore, for these populations, inter-gene variance in burden was higher in both filters, but substantially so when using the *popmax* filter.

Table 5.6: Coefficient of variation for ancestry-specific burden of missense variants across all genes in the exome. Coefficient of variation is calculated across eight ancestry groups, three frequency bands and both *popmax* and *strict* filtering. Heatmap colour coding highlights outliers across populations.

	Popmax			*Strict*		
Population	**Very-rare (<0.01%)**	**Rare (0.01% - 0.1%)**	**Uncommon (0.1% - 1%)**	**Very-rare (<0.01%)**	**Rare (0.01% - 0.1%)**	**Uncommon (0.1% - 1%)**
AFR	0.46	0.83	1.2	0.46	0.84	1.2
AMR	0.44	0.78	1.3	0.44	0.78	1.2
ASJ	1.5	2.2	2	0.8	1.3	1.3
EAS	0.47	0.82	1.4	0.47	0.82	1.4
FIN	1.4	2.4	2	0.61	1.2	1.6
NFE	0.4	0.85	1.5	0.4	0.85	1.4
OTH	0.57	0.89	1.2	0.55	0.84	1.1
SAS	0.44	0.77	1.3	0.43	0.77	1.3

5.2.4 Ancestry variant burden in high and low-constraint genes

Patterns of missense-variant burden were contrasted in genes with high and low predicted missense-constraint (Figure 5.5). To achieve this, constraint scores were downloaded for all genes from GnomAD

(https://gnomad.broadinstitute.org/downloads/). Genes in the top 10% and lowest 10% of missense constraint Z-scores were extracted from the dataset to represent extreme groups of constrained and unconstrained genes. The most striking difference between the gene-groups, was that highly constrained genes had fewer variants on average than the low constraint genes. This was expected as sparseness of missense variation in these genes was used to construct the constraint scores. Interestingly, in the *uncommon* band, the vast increase in missense burden seen in AFR in Figure 4.13, is markedly reduced for high constraint genes, bringing the burden to a level more comparable with the other populations. This supports the hypothesis that the excess seen in this population (Figure 4.13), relative to other groups, is driven by genetic drift over their long evolutionary history. Corroborating this, for synonymous variants and low constrained genes, AFR still have substantially higher burden than other groups.

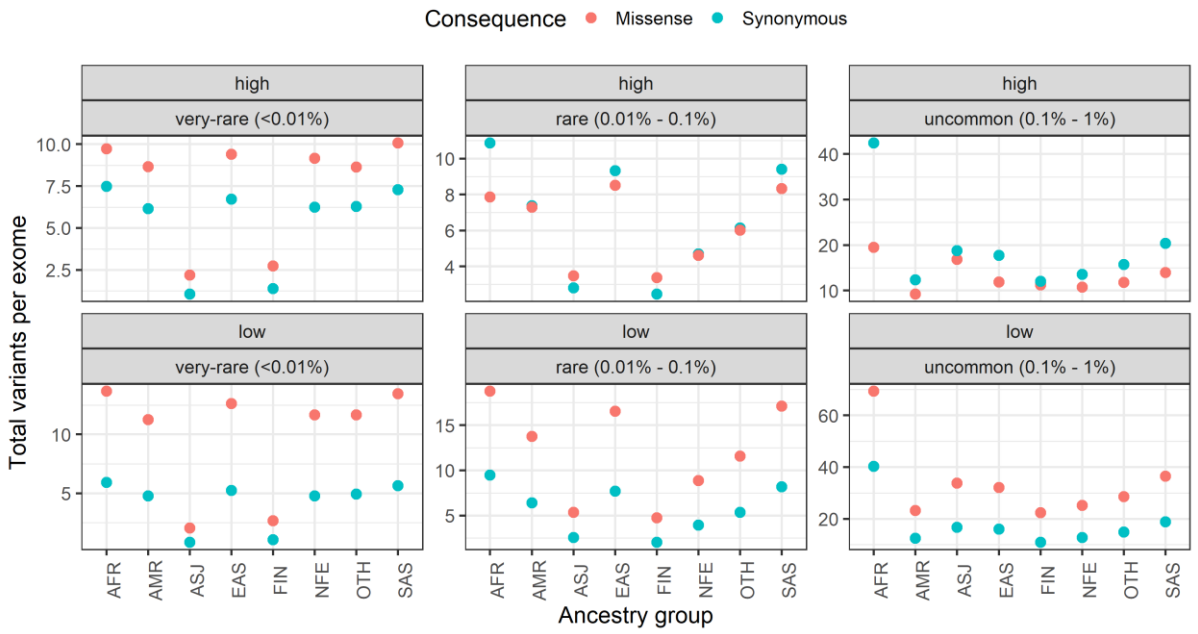

Figure 5.5: Mean number of variants per exome for three variant-rarity-levels and eight *gnomad_e2* ancestry groups, in both high and low missense-constrained genes. High and low constraint genes were defined as the top and bottom 10% of missense constraint Z-scores calculated by Lek *et al.* (2016).

5.3 Discussion

In this chapter, two supplementing considerations relevant to burden and positional testing (especially using GnomAD controls) were considered, including choice of frequency filtering strategy and ancestry-specific variant burden. The filtering strategy has the potential to strongly bias association results, as the accuracy of empirical frequency estimates is reduced in small subgroups and the relationship between variant rarity and potential pathogenicity differs between different subpopulations. For ancestry, the different genealogical histories of ancestry groups produce unique site frequency spectrums of variants. This cause differences in the baseline burden of variants between different subpopulations.

5.3.1 Frequency filtering

In previous chapters, the point estimates of subpopulation allele frequencies were entered into the *popmax* calculation. Due to the small sample sizes of some ancestry groups, this resulted in the exclusion of many variants that were probably rare. To overcome this, the point estimates were replaced with Poisson lower confidence interval estimates. This prevented over-filtering of smaller populations, however the optimum lower confidence threshold for aggregation analyses could not be determined in this study. However, an opposing bias may also occur. Whereby an inflated number of qualifying rare-variants are observed in ancestry groups underrepresented or absent in reference cohort used for frequency filtering (Petrovski *et al.* 2016). This is because their true frequency cannot be estimated. The clear solution to this problem is larger and more diverse reference cohorts. Given the huge number of projects ongoing such as UK Biobank exomes, US national initiative, GenomeAsia 100K, H3African, this will soon be the reality.

Throughout this thesis, the frequency filtering approach included all *gnomad_e2* subpopulations in the *popmax* calculation. This removed all ancestry-specific common variants from the analysis, as common penetrant pathogenic variants are inconsistent with the prevalence of the diseases under

investigation. However, founder variants in ASJ and FIN that may be penetrant and pathogenic, are thought to exist. In subsection 5.1.3, variants observed at high frequencies in ASJ or FIN but which passed the standard *popmax* filter, were identified. One potential pathogenic founder was corroborated by published reports confirming its likely pathogenic status (*MYH7*: p.Arg1053Gln) but it was not possible to confirm any pathogenic association for any of the other 35 high-frequency ASJ or FIN founders in our heart-data or the 131 in *gnomad_e2* across the same genes. For this reason, it is logical to conclude that including these variants in aggregation analyses may decrease the signal-noise ratio.

5.3.2 Ancestry burden

Using ancestry-specific variant counts in the *gnomad_e2* dataset, inter-ancestry differences in missense and synonymous variant burden were investigated. Interesting patterns of variant burden emerged between ancestry groups and allele frequency bands and variant consequence (missense or synonymous). These patterns were contrasted for different frequency filtering methods, in high and low constraint genes and inter-gene variance in burden was explored. Results were consistent with previous observations of the site-frequency spectrum in the investigated populations and highlighted the importance of ancestry-control in association studies.

The burden of missense variants varies across different ancestry groups. This is a different from fine-scale population stratification known to be a potential confounder of rare-variant association studies (Mathieson and McVean 2012). Instead, differences in the ancestral histories of broad-scale ancestry groups result in more or fewer rare-variants, on average, across samples. The consequence of this is that for unmatched ancestry case-control cohorts, the baseline expected odds-ratio, averaged across genes, is not equal to one, and in the most extreme case in our data, can be as high as 6.26. Clearly, this would lead to false-positives in association studies in unmatched cohorts. Due to inter-gene variance in this baseline difference, a scaling factor would be insufficient to control for this.

In subsection 5.13, it was noted that founder variants in ASJ and FIN are possible confounders of association studies when these populations are not used for filtering. In subsection 5.2.1, it is clear that when including them in the *strict popmax* filter, a different bias emerges, one of reduced rare-variant burden in these populations. This has previously been noted by Povysil *et al.* (2019), whom indicated lower rates of exome-wide rare-variation in consanguineous and bottlenecked populations. In alternative *popmax* approaches where these populations are not included in the *popmax* calculation, the number of "rare" ASJ variants is actually more than any other population and the number of rare FIN variants is equivalent to EAS, AFR and AMR. However, this is due to inflation of the mean by non-rare founder variants. This was supported by gene-level variance in counts, which was very high in FIN and ASJ using the *pci_popmax* filter (Figure 4.15). Without the option of ancestry adjustment, the best solution would therefore be to exclude these populations from the analysis if cohorts were not matched.

Overall in terms of ancestry-specific confounding for rare-variant association studies, for broadly matched ancestry groups, which are well-represented in reference controls, it has been proposed that little population structure influences rare-variants and zero population stratification underlies very rare, potentially private *de-novo* variants (Zhu *et al.* 2017). For this reason, ancestry–adjustment may not be essential to incorporate at the statistical method level, but instead relies on only broadly matched cohorts and large-diverse reference populations. As well studied founder populations, FIN and ASJ comprise a disproportionate number of samples in the GnomAD exomes dataset, together they make up 16,075 samples, which is 12.9% of the total. It is unlikely that case datasets derived in an ancestry-agnostic framework, such as aggregation of clinical sequence data in the UK, would have such a high proportion of these rare-variant sparse ancestry groups. This may partly explain the results in chapter 2.5, where more genes than expected harboured *n-significant* missense variant burden p-values and the lambda inflation factor for synonymous burden was inflated.

Chapter 6

Conclusion

This thesis was undertaken to investigate the genetic phenomenon of rare-missense variant clustering in Mendelian disease-genes. At the outset, it was posited that this phenomenon is ubiquitous across protein-coding genes underlying susceptibility to rare-diseases. It was hypothesised that this characteristic signal would enhance the power of gene-association studies. And finally, it was conjectured that quantitative data-driven approaches to variant interpretation would improve on the current qualitative framework. Here, in this final chapter, I argue that I have been successful in this endeavour, I have contributed to the field of clinical genetics and I have discovered hopeful avenues for the continuation of this research. I discuss the results presented in chapters 2 to 5, and conclude with the limitations, contributions and exciting future directions of this work.

6.1 ClusterBurden

In chapter 2, a gene-based association test *ClusterBurden* was developed to test for association between genes and a disease outcome using both burden and positional information. The framework underlying *ClusterBurden* was the combination of two separate tests, one to detect burden, a Fisher's exact test, and another to detect clustering. Multiple methods were explored to determine the most powerful strategy to identify clustering. The victor, in terms of power, was the chi-squared two-sample test, which gave vastly superior power than commonly used KS-test and AD-test. After settling on the appropriate procedure for binning variant counts ($k = n^{2/5}$ where k = number of evenly sized bins and n = number of observed variants). The approach, named *BIN-test*, was then combined with a burden test to form the overall RVAT: *ClusterBurden*. In simulated data, *ClusterBurden*, and *BIN-test*, had well-calibrated type-1 error rates and were more powerful than alternative methods when disease-causing variants clustered in cases. A final test, LRT-dbNSFP, was developed as an additional complementary and uncorrelated approach. Here, a model of pathogenicity prediction scores was generated to predict case-control status based on the specific missense variants observed in each cohort. A likelihood ratio test was then used to assess significance of this model.

A common strategy to handle uneven coverage in case-control studies is to exclude variant sites below a specified coverage threshold in either cases or controls. Here, to retain as much data as possible, a novel strategy to control for coverage was introduced for the *BIN-test*. The counts in each bin were weighted by the reciprocal of coverage in the spanned region for cases and controls. Although regions with less than 50% of samples covered to at least 10X in either cohort were still excluded, this approach allowed retention of the majority of the clinical data available. In a putatively null analysis of synonymous burden, the approach reduced the number of *n-significant* signals relative to an unadjusted analysis, giving some indication of its effectiveness.

All methods were applied to a large set of clinical sequences from inherited heart disease patients. BAM files were aggregated over 5 years of clinical sequencing at Oxford Medical Genetics Laboratory, and combined with the large HCMR dataset. This produced one of largest known dataset of its kind. Using the GnomAD exomes population cohort as a control group, burden (Fisher's-exact test), clustering (*BIN-test*), combined (*ClusterBurden*) and pathogenicity-score signals (LRT-dbNSFP), were assessed in each gene-disease pair. Significant burden signals were identified in almost all firmly established disease-genes reported to cause HCM, DCM, ARVC, LQTS and BrS. Significant clustering was detected in many genes, including eight HCM genes, one DCM gene (*MYH7*), one ARVC gene (*PKP2*) and two LQTS genes (*KCNH2* and *KCNQ1*). Due to the relatively small samples sizes in all diseases except HCM, it is easy to envisage that with increasing data, this significantly clustered list of inherited heart disease genes would also increase. The combination of both methods into a unified test; *ClusterBurden*, rediscovered all firmly established genes identified via burden testing and boosted significance in all with clustering signals. This included the recently identified *FLNC* which had only a modest burden effect size but strong clustering signal, emphasising the value of this approach. The final test, LRT-dbNSFP, discovered Bonferroni-significant differences in pathogenicity prediction scores in a total of 18 genes across the analysis, all of which with published evidence supporting their true association.

Overall, it is clear that positional information and pathogenicity prediction scores contain useful information for the assessment of gene pathogenicity. Where rare-missense variants are thought to cause the disease, the addition of these signals, may provide uncorrelated complementary information that augments association studies. Although inherited heart diseases were used as a test case for these methods, there is no reason why they could not be applied to other diseases with sufficient data, although modifications to data pre-processing would be necessary for different genetic models such as recessive. Incorporating these methods in sufficiently-powered exome-wide scans, may elevate power to discover novel low penetrance genes.

6.2 GAMs

In chapter 3, the narrative progresses from gene-level association to variant-level association. Here, the position of a variant is used to inform the probability of its pathogenicity. Using generalised-additive models, regional burden estimates were made in significantly associated inherited heart disease genes; *hotspot models*. These models were used to infer mutational hotspots where observed variants had an increased prior probability of disease association. Incorporated into the widely followed ACMG guidelines as criterion PM1, mutational hotspots represent useful information regarding a variant's potential pathogenicity. Using the *hotspot models*, a quantitative data-driven strategy to apply criterion PM1 was proposed. This allowed stratification of regions of supporting, moderate and strong evidence of pathogenicity for fifteen genes, in line with PM1.

Models were then extended to include further *in-silico* predictors from the variant prediction score database dbNSFP. The models, named *hotspot+ models*, integrated both burden, clustering and pathogenicity prediction data that could include functional consequences of variants, inter-species conservation and protein structural information. The predictive performance of these models, judged by their ability to separate case and control variants in our heart disease and GnomAD exomes dataset, was substantial higher relative to generic prediction scores in almost all genes.

Few quantitative statistical approaches to variant interpretation exist in the current guidelines. This is mainly attributable to the limited data available to make data-driven inference on the specific criteria used for interpretation. Here, much hard work has gone into aggregating sufficient data to achieve sample sizes capable of reliable inference. The result is an unbiased modelling procedure for mutational hotspots and *in-silico* predictors that can boost confidence in application of criteria PM1 and PP3. This has great scope for translatability and can be used by clinical scientists to make more accurate assessments of variant pathogenicity in real patients. This may reduce the number of incorrect classifications as well as the number of patients who may now have their pathogenic variant

confidently identified. This will not only benefit the patients but also their relatives via cascade screening.

6.3 Prevalence of missense-variant clustering

In chapter 4, an analysis was conducted to support the claim that missense-variant clustering is a common phenomenon in Mendelian disease genes. Variants in one of the most comprehensive resources of clinically-relevant variants (*clinvar*), were assessed for clustering relative to the GnomAD exomes population reference cohort. Hundreds of significantly clustered genes were detected affirming the importance of this signal for disease association. Properties of clustered genes and missense-variant clusters were explored signifying the potential of determining prior likelihood of clustered genes and cluster locations in future iterations.

Before comparing variant locations between GnomAD and *clinvar*, a putative null model was examined by comparing variant positions between GnomAD exomes and GnomAD genomes. Due to different sequencing technologies involved for whole-exome and whole-genome sequencing, coverage between the datasets substantially differed with GnomAD genomes having a lower average but more even coverage. Reassuringly, the inverse-coverage weighting procedure reduced lambda from 1.61 to 1.02, supporting the effectiveness of this coverage-control technique.

When scanning for missense-variant clusters in *clinvar* relative to GnomAD controls, the most rewarding experimental setup involves the inclusion of only likely pathogenic and pathogenic *clinvar* variants and all GnomAD variants regardless of frequency. Of the 1363 genes viable for this analysis, 654 were at least nominally significant and 293 of these were significant after multiple testing correction. This comes with the important caveat that “clusters” can be self-perpetuating by increasing the probability of a pathogenic classification for variant within them, as well as historic partial sequencing of genes, predetermining clusters at these regions when the whole gene is analysed. Both the results and this caveat demonstrate the importance for positional-based analysis

of missense-variants, to identify the many clustered regions but also to assess the validity of previously assumed clusters that may have been identified with lower statistical certainty.

An analysis of significantly clustered genes revealed over-representation of specific protein groups (ion channels and transporters), higher than average missense constraint and increased likelihood of having an annotated functional domains or motif. An algorithm was developed to automatically define cluster locations in these significantly associated genes. Predicted clusters overlapped with functional domains more than expected by chance and were significantly associated with regions characterised with secondary structure annotations of helical, beta-strand or turn. These results are important to inform the direction of future analyses that aim to determine a weighting scheme for the burden and position component in *ClusterBurden* for exome-wide scans. Cluster location properties may also provide an opportunity to weight *BIN-test* bins to give higher weight to regions more likely to harbour missense-variant clustering. Finally, hotspot inference, made only using empirical data in chapter 3, could be extended to include other features predictive of mutational hotspots.

Although leaving much room for continuation, this analysis represents the foundation of an entirely new project aimed at identifying useful priors for exome-scans. By identifying which genes would be useful to apply the *BIN-test* to, and which regions of the gene should be weighted more in this analysis. The eventual result of this, would be a more-consistently well-powered RVAT that could avoid power-loss in non-clustered genes and improve power to detect association in clustered genes. As a follow up, it may be fruitful to conduct a comprehensive literature and database search, to generate a meaningful and reliable dataset of missense-variant clusters. With a sufficient amount of training data, it may be possible to generate a machine learning models to predict clustered genes and clusters. While this may be useful in its own right it may also inform priors for *ClusterBurden* exome-scans enhancing the method.

6.4 Frequency filtering and ancestry-specific baseline burden

In chapter 5, the variant frequency filtering strategy was reconsidered. The original method for filtering was addressed to tackle two potential biases; the inaccuracy of frequency estimates with low sample sizes and the inclusion, or exclusion, of the Finnish (FIN), Ashkenazi Jewish (ASJ) and other (OTH) ancestry groups in the *popmax* calculation. A filtering strategy adopting Poisson lower confidence intervals effectively prevented over-filtering in small subpopulations. However, the optimal subpopulations to include in the popmax calculation still remained an issue. In the original ExAC paper (Lek *et al.* 2016), the authors recommended a popmax style approach by taking the maximum of the five major populations in the dataset but this approach fails to exclude high-frequency variants specific to FIN, ASJ or OTH. For variant-centric interpretation, it would be important not to exclude these, as they may be high-frequency pathogenic variants, however, few of these high-frequency variants in our data had published evidence of pathogenicity. This suggests that these variants may add noise to aggregation analyses, with the added problem that their high carrier count would give them too much weight in the analysis.

To further complicate matters, in an analysis of ancestry-specific burden, it was clear that the founder populations (FIN and ASJ), had reduced rare-variation relative to other populations, highlighting another source of potential bias. This is especially problematic when using *gnomad_e2* as a control group as the ancestry is not only unlikely to be matched to the cases, but these rare-variant sparse populations are over-represented in the ancestry composition. Due to the high variance in burden between genes in these populations, there is no simple and effective measure to control for this in unmatched ancestry groups.

This analysis highlighted specific modes of confoundment by ancestry and frequency filtering, and exposed the unreliability of GnomAD as a control group. Although, the heterogeneous and unmatched nature of this control group makes it unsuitability intuitive, the extent and nature of the problems were characterised in greater detail here. However, solutions to some of these problems, including choice of filtering populations and broad-scale ancestry matching, are relevant for future analyses of

this type. These lessons will remain useful as GnomAD is superceded by new and larger reference datasets.

6.5 Contributions, future directions and limitations

Like almost all data-driven statistical genetics studies, the investigations reported here may be improved as increasingly large datasets become available, such as the UK Biobank exomes. Specifically, there will be opportunities for well-powered exome-wide scans using *ClusterBurden*, or future iterations of this method. As the power of this approach may be superior to traditional methods, applying this framework may help to discover low penetrance inherited heart disease genes, or other Mendelian disease-genes.

For *ClusterBurden*, and the other proposed methods in chapter 2 (*BIN-test*, LRT-dbNSFP), some design improvements have been considered. An important finding in this thesis and elsewhere (Samocha *et al.* 2017; Evans *et al.* 2019; Havrilla *et al.* 2019; Silk *et al.* 2019), is that regions of increased pathogenicity potential in Mendelian disease-genes are often associated with regions of missense constraint in population controls. This signal, presumably driven by purifying selection in these regions, is automatically considered in the case-control methods developed here, as purifying selection is implicitly modelled by the numerous reference controls. However, it may also promise further improvements. In this thesis GnomAD was used as a control group. In future analyses, better matched case-control cohorts should be used. This provides the opportunity to use GnomAD as an auxiliary dataset to inform regional missense constraint at the population level and indicate genes and regions more likely to harbour missense-variant clustering.

For *ClusterBurden*, the combination of p-values is achieved using Fisher's method, which gives equal weight to both the position and burden signal. If genes were characterised by a measure of variance in regional missense constraint, then these scores could be used to weight the contribution of each test when combined using Stouffer's method (Stouffer *et al.* 1949). This is based on the hypothesis

that prior likelihood of missense variant clustering can be predicted by genic patterns of regional missense constraint. This hypothesis would be investigated first to determine the optimal metric to relate the two. In addition to this, regional missense constraint may inform particular protein regions that should be considered with more weight in a *BIN-test* analysis. As well as regional missense constraint, other predictors of clustered genes and locations such as functional domains and secondary structure could be included as priors to *ClusterBurden*.

Although 3-dimensional structural information is unavailable or limited for the majority of inherited heart disease-genes, recent advances in deep learning prediction algorithms (Senior *et al.* 2020) and experimental technologies (Yip *et al.* 2020) may lead to imminent availability of high-resolution protein structures. Access to this information would theoretically improve power for both *BIN-test* and the *hotspot models*. This is because functional regions may be distal in the linear sequence but proximal in the tertiary structure (Tokheim *et al.* 2016). Future iterations of the methods proposed here could attempt to utilise this additional information. For the *BIN-test*, linear binning could be replaced with 3D binning, which may increase the probability of binning together functionally related variants. For *hotspot models,* GAMs are commonly used for geographical modelling, and are capable of generating 3-dimensional smooth functions to potentially model 3D clusters.

Data-driven GAMs have been demonstrated as an effective way to model burden and position to produce models of regional burden (*hotspot models*). Their utility does not stop at inherited heart diseases and the field would benefit from this style of analysis to many more genes and disease areas. This would extend the consistent and quantitative application of the ACMG criterion PM1. Future advances of this approach coincide with that of *ClusterBurden* and *BIN-test.* Most crucially, for accurate inference of mutational hotspots, improvements to the data are necessary. This follows from the conclusions of chapter 5 and would involve larger and more diverse reference cohorts for frequency filtering, better matched cases and controls (at least in terms of ancestry), and separation of control and reference cohort.

Finding a balance between idealised experiments with perfectly matched samples and the pragmatic need to advance knowledge on the basis of already available, albeit imperfect, data, led to the adoption of GnomAD exomes as controls for many analyses. However, as ancestry, sequencing technology and bioinformatics processing were not identical between GnomAD and the clinically sequenced cases, this represented an important potential source of bias and confounding. Furthermore, frequencies in the GnomAD population cohort were used for filtering of both case and control variants. This may induce bias, as the filtering of non-rare variants in GnomAD may be more assured than the removal of non-rare variants in the cases, especially for ancestry-specific variants from ancestries underrepresented in GnomAD. Despite these caveats, the huge size of this population cohort makes it a valuable resource to demonstrate the key hypotheses of this thesis. In the future, new datasets may be available to overcome these problems. For example, the UK Biobank exomes will provide larger sample sizes, the opportunity for better ancestry matching via the associated SNP-array data and will allow separation of the filtering population and control population.

The methods developed here in chapters 2 and 3 have been made available to the statistical and clinical genetics communities. In appendix 1, an R package for association testing alongside many helper functions is detailed. In appendix 2, a web application for *hotspot modelling*, written in the R:Shiny framework is described. Both methods are published in a peer-reviewed journal (Waring *et al.* 2020) to share results with researchers in the realm of medical genetics.

Above all else, this thesis promotes the statistical integration of missense variant protein position in the assessment of disease genes or variants. It is demonstrated that variant clustering is a common feature of Mendelian disease-genes. It is clear that in the case of a clustered-gene, including this information in association testing is more powerful that the traditional burden approach. Finally, robust statistical inference of mutational hotspots improves interpretation of rare-variants in inherited heart disease-genes.

References

1. Ader F, De Groote P, Réant P, Rooryck-Thambo C, Dupin-Deguine D, Rambaud C, Khraiche D, Perret C, Pruny JF, Mathieu-Dramard M, et al.. FLNC pathogenic variants in patients with cardiomyopathies: prevalence and genotype-phenotype correlations. Clin Genet. 2019; 96:317–329. doi: 10.1111/cge.13594
2. Adler A, Novelli V, Amin AS, et al. An International, Multicentered, Evidence-Based Reappraisal of Genes Reported to Cause Congenital Long QT Syndrome. Circulation. 2020;141(6):418-428. doi:10.1161/CIRCULATIONAHA.119.043132
3. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013; 7 (7).
4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. Nat Methods. 2010, 7: 248-249.
5. Agresti A. Categorical data analysis: New York: John Wiley & Sons; 1996
6. Alamo L, Ware J, Pinto A, Gillilan R, Seidman J, Seidman C, Padron R. Effects of myosin variants on interacting-heads motif explain distinct hypertrophic and dilated cardiomyopathy phenotypes. eLife 2017;6:e24634 DOI: 10.7554/eLife.24634.
7. Alfares A, Kelly M, McDermott G, Funke B et al.. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. Genetics in Medicine 2015; 17: 880-888. https://www.nature.com/articles/gim2014205
8. Al-Jassar, Caezar & Bikker, Hennie & Overduin, Michael & Chidgey, Martyn. (2013). Mechanistic Basis of Desmosome-Targeted Diseases. Journal of molecular biology. 1778. 10.1016/j.jmb.2013.07.035.
9. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019 Jan 8;47(D1):D1038-D1043. doi:10.1093/nar/gky1151. PMID: 30445645.
10. Amendola LM. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. The American Journal of Human Genetics. Volume 98, Issue 6, 2 June 2016, Pages 1067-1076

11. Arunachal G, Pachat D, Doss CG, Danda S, Pai R, Ebenazer A. Molecular Characterization of a Novel Germline VHL Mutation by Extensive In Silico Analysis in an Indian Family with Von Hippel-Lindau Disease. Genet Res Int. 2016;2016:9872594. doi:10.1155/2016/9872594
12. Auton, A., Abecasis, G., Altshuler, D. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393
13. Azaiez H, Booth KT, Ephraim S, Crone B, Black-Ziegelbein EA, Marini RJ, Shearer AE, Sloan-Heggen CM, Kolbe D, Casavant T, Schnieders MJ, Nishimura C, Braun T, Smith RJ. "Genomic Landscape and Mutational Signatures of Deafness-Associated Genes" American Journal of Human Genetics. September 20, 2018, DOI:https://doi.org/10.1016/j.ajhg.2018.08.006
14. Bamshad, M., Ng, S., Bigham, A. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12,** 745–755 (2011). https://doi.org/10.1038/nrg3031
15. Bass-Zubek AE, Godsel LM, Delmar M, Green KJ. Plakophilins: multifunctional scaffolds for adhesion and signaling. Curr Opin Cell Biol. 2009;21(5):708-716. doi:10.1016/j.ceb.2009.07.002
16. Bauce B, Basso C, Rampazzo A, Beffagna G, Daliento L, Frigo G, Malacrida A, Settimo L, Danieli G, Thiene G, Nava A. Clinical profile of four families with arrhythmogenic right ventricular cardiomyopathy caused by dominant desmoplakin mutations, European Heart Journal, Volume 26, Issue 16, August 2005, Pages 1666–1675, https://doi.org/10.1093/eurheartj/ehi341
17. Baumgartner W, Weiß P and Schindler S. A Nonparametric Test for the General Two-Sample Problem. Biometrics. Vol. 54, No. 3 (Sep., 1998), pp. 1129-1135
18. Bean LJH, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Med 9, 111 (2017).
19. Bissler JJ, Cicardi M, Donaldson VH, Gatenby PA, Rosenii FS, Sheffer AL, Davis AE. A cluster of mutations within a short triplet repeat in the C1 inhibitor gene. *Proc Nati Acad Sci USA* 1994; 91:9622–9625
20. Bollen IAE, van der Velden J. The contribution of mutations in MYH7 to the onset of cardiomyopathy. Neth Heart J. 2017;25(12):653-654. doi:10.1007/s12471-017-1045-5
21. Brugada et al. 2018. Present Status of Brugada Syndrome. Journal of the American College of Cardiology. Volume 72, Issue 9, August 2018. http://www.onlinejacc.org/content/72/9/1046
22. Brugada J, Campuzano O, Arbelo E, Sarquella-Brugada G, Brugada R. Present Status of Brugada Syndrome. J Am Coll Cardiol. 2018 Aug, 72 (9) 1046-1059.

23. Bycroft, C., Freeman, C., Petkova, D. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018). https://doi.org/10.1038/s41586-018-0579-z
24. Campuzano et al. 2019. Genetic interpretation and clinical translation of minor genes related to Brugada syndrome. Volume40, Issue6. Pages 749-764.
25. Campuzano O, Fernández-Falgueras A, Iglesias, Brugada R. Brugada Syndrome and PKP2: Evidences and uncertainties. Int J Cardiol. 2016;214:403-5
26. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 (Suppl 3): S3 10.1186/1471-2164-14-S3-S3
27. Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., Hubbard, T., Jostins, L., Maltby, N., Mahon-Pearson, J., McVean, G., Nevin-Ridley, K., Parker, M., Parry, V., Rendon, A., Riley, L., Turnbull, C., & Woods, K.. (2017). *The National Genomics Research and Healthcare Knowledgebase* (Version 5). figshare. https://doi.org/10.6084/m9.figshare.4530893.v5 ([])
28. Chaves-Markman ÂV, Markman M, Calado EB, et al. GLA Gene Mutation in Hypertrophic Cardiomyopathy with a New Variant Description: Is it Fabry's Disease?. Arq Bras Cardiol. 2019;113(1):77-84. Published 2019 Jul 10. doi:10.5935/abc.20190112
29. Chen et al. 1998. Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. Nature. 1998 Mar 19;392(6673):293-6.
30. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013;9(1):e1003224. pmid:23358228
31. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet*. 2015;97(2):199-215. doi:10.1016/j.ajhg.2015.06.009
32. Cochran WG. The χ2 test of goodness of fit. Annals of Mathematical Statistics. 1952;23:315–345.
33. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005, 15: 901-913.
34. Corrado D, Link MS, Calkins H. Arrhythmogenic right ventricular cardiomyopathy. N Engl J Med. 2017;376(1):61–72.
35. Curtis D, Coelewij L, Liu S. et al. Weighted Burden Analysis of Exome-Sequenced Case-Control Sample Implicates Synaptic Genes in Schizophrenia Aetiology. Behav Genet 48, 198–208 (2018). https://doi.org/10.1007/s10519-018-9893-3

36. Dietz HC, Saraiva JM, Pyeritz RE, Cutting GR, Francomano CA. Clustering of fibrillin (FBN1) missense mutations in Marfan syndrome patients at cysteine residues in EGF-like domains. Hum Mutat. 1992;1(5):366-374. doi:10.1002/humu.1380010504
37. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–2137. pmid:25552646
38. Dowle M and Srinivasan A (2019). data.table: Extension of `data.frame`. R package version 1.12.2. https://CRAN.R-project.org/package=data.table.
39. e Mattos B, Luís Scolari F, Torres M, Simon L, de Freitas V, Giugliani R, Matte U. Prevalence and Phenotypic Expression of Mutations in the MYH7, MYBPC3 and TNNT2 Genes in Families with Hypertrophic Cardiomyopathy in the South of Brazil: A Cross-Sectional Study. 3.10.1 Arq. Bras. Cardiol. 2016 vol.107 no.3.
40. Ehlermann, P., Weichenhan, D., Zehelein, J. et al. Adverse events in families with hypertrophic or dilated cardiomyopathy and mutations in the MYBPC3gene. BMC Med Genet 9, 95 (2008). https://doi.org/10.1186/1471-2350-9-95
41. Evans P, Wu C, Lindy A, et al. Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets. Genome Res. 2019;29(7):1144–1151. doi:10.1101/gr.240994.118
42. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Silverman EK, Pagano M, Mangold E, Lange C. 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinformatics. 2012; 28 (23);3027–3033
43. Fine DM, Wasser WG, Estrella MM, Atta MG, Kuperman M, Shemer R, Rajasekaran A, Tzur S, Racusen LC, Skorecki K. APOL1 Risk Variants Predict Histopathology and Progression to ESRD in HIV-Related Kidney Disease. *J Am Soc Nephrol* 2012; 23:343–350
44. Flashman E, Redwood C, Moolman-Smook J, Watkins H. Cardiac Myosin Binding Protein C: Its role in physiology and disease. Circulation 2004; 94:1279-1289
45. Friedrich FW, Wilding BR, Reischmann S *et al.* Evidence for FHL1 as a novel disease gene for isolated hypertrophic cardiomyopathy, Human Molecular Genetics, Volume 21, Issue 14, 15 July 2012, Pages 3237–3254, https://doi.org/10.1093/hmg/dds157

46. Friedrich, F.W., Reischmann, S., Schwalm, A. et al. FHL2 expression and variants in hypertrophic cardiomyopathy. Basic Res Cardiol 109, 451 (2014). https://doi.org/10.1007/s00395-014-0451-8

47. Fritz Scholz and Angie Zhu (2019). kSamples: K-Sample Rank Tests and their Combinations. R package. version 1.2-9. https://CRAN.R-project.org/package=kSamples

48. Fukuyama M, Ohno S, Ichikawa M, Takayama K, Fukumoto D, Horie M. P5860: Novel RYR2 mutations causative for long QT syndromes. 2017. European Heart Journal. 38. 10.1093/eurheartj/ehx493.P5860.

49. Furqan et al. 2017. Care in Specialized Centers and Data Sharing Increase Agreement in Hypertrophic Cardiomyopathy Genetic Test Interpretation. Circulation: Genomic and Precision Medicine 10 (5). https://www.ahajournals.org/doi/10.1161/CIRCGENETICS.116.001700

50. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54–62.

51. Garmany R, Giudicessi JR, Ye D, Zhou W, Tester D, Ackerman M. Clinical and functional reappraisal of alleged type 5 long QT syndrome: Causative genetic variants in the KCNE1-encoded minK β-subunit. Heart Rhythm 2020.

52. Geisterfer-Lowrance AT, Kass S, Tanigawa G, Vosberg H, McKenna W, Seidman C and Seidman J. A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation. Cell 1990; 62 (5): 999-1006. https://www.sciencedirect.com/science/article/pii/009286749090274I?via%3Dihub

53. Gelb B, Cavé H, Dillon MW, Gripp KW, Lee J, Mason-Suares H, Rauen K, Williams B, Zenker M, Vincent L; ClinGen RASopathy Working Group. ClinGen's rasopathy expert panel consensus methods for variant interpretation. Genet. Med. 2018; 20:1334-1345

54. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18(1):225. Published 2017 Nov 28. doi:10.1186/s13059-017-1353-5

55. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet*. 2012;13(2):135-145. Published 2012 Jan 18. doi:10.1038/nrg3118

56. Girolami F, Ho C, Semsarian C, Baldi M, Will M, Baldini K, Torricelli F, Yeates L, Cecchi F, Ackerman M and Olivotto L. Clinical Features and Outcome of Hypertrophic Cardiomyopathy Associated With Triple Sarcomere Protein Gene Mutations. JACC 2010; 55 (14). http://www.onlinejacc.org/content/55/14/1444.abstract

57. Goerg, Sebastian & Kaiser, Johannes. (2009). Nonparametric Testing of Distributions—the Epps–Singleton Two-Sample Test using the Empirical Characteristic Function. Stata Journal. 9. 454-465. 10.1177/1536867X0900900307.

58. Gómez, J. , Lorca, R. , Reguero, J. R. , Morís, C. , Martín, M. , Tranche, S. , … Cote, E. (2017). Screening of the filamin C gene in a large cohort of hypertrophic cardiomyopathy patients. Circulation Cardiovascular Genetics, 10 10.1161/CIRCGENETICS.116.001584

59. Grimm DG, Azencott CA, Aicheler F, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513-523. doi:10.1002/humu.22768

60. Guo MH, Dauber A, Lippincott MF, Chan YM, Salem RM, Hirschhorn JN. Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. Am J Hum Genet. 2016;99(3):527-539. doi:10.1016/j.ajhg.2016.06.031

61. Guo, M. H., Plummer, L., Chan, Y.-M., Hirschhorn, J. N. & Lippincott, M. F. Burden testing of rare variants identified through exome sequencing via publicly available control data. Am. J. Hum. Genet. 103, 522–534 (2018).

62. Haldane JB. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet 1956; 20:309-11

63. Hall, C.L., Sutanto, H., Dalageorgou, C. et al. Frequency of genetic variants associated with arrhythmogenic right ventricular cardiomyopathy in the genome aggregation database. Eur J Hum Genet 26, 1312–1318 (2018). https://doi.org/10.1038/s41431-018-0169-4

64. Harrison, S.M., Rehm, H.L. Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. Genome Med 11, 72 (2019). https://doi.org/10.1186/s13073-019-0688-9.

65. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. 2019. A map of constrained coding regions in the human genome. Nat Genet 51: 88–95.

66. Henderson DM, Lee A, Ervasti JM. Disease-causing missense mutations in actin binding domain 1 of dystrophin induce thermodynamic instability and protein aggregation. *Proc Natl Acad Sci USA* 2010; 107:9632–7

67. Hershberger RE, Morales A. LMNA-Related Dilated Cardiomyopathy. 2008 Jun 12 [Updated 2016 Jul 7]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1674/

68. Hershberger, R., Hedges, D. & Morales, A. Dilated cardiomyopathy: the complexity of a diverse genetic architecture. Nat Rev Cardiol 10, 531–547 (2013). https://doi.org/10.1038/nrcardio.2013.105

69. Ho et al. 2018. Genotype and Lifetime Burden of Disease in Hypertrophic Cardiomyopathy. Circulation. 138 (14): 1387-1398.

70. Hoischen A, van Bon BW, Gilissen C, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010;42(6):483-485. doi:10.1038/ng.581

71. Hongjian Qi, Chen Chen, Haicang Zhang, John J. Long, Wendy K. Chung, Yongtao Guan, Yufeng Shen. MVP: predicting pathogenicity of missense variants by deep learning. 2018; BioRxiv: doi: https://doi.org/10.1101/259390.

72. Horvat, C., Johnson, R., Lam, L. et al. A gene-centric strategy for identifying disease-causing rare variants in dilated cardiomyopathy. Genet Med 21, 133–143 (2019). https://doi.org/10.1038/s41436-018-0036-2

73. Hosseini et al. 2018. Reappraisal of Reported Genes for Sudden Arrhythmic Death: Evidence-Based Evaluation of Gene Validity for Brugada Syndrome. Circulation. 2018;138:1195–1205. https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.035070

74. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics. 2008; 40:395–402

75. Ingles et al. 2017. Nonfamilial Hypertrophic Cardiomyopathy. Circulation 10 (2). ahajournals.org/doi/full/10.1161/CIRCGENETICS.116.001620

76. Ingles J, Goldstein J, Thaxton C, Caleshu C, Corty EW, Crowley SB, Dougherty K, Harrison SM, McGlaughon J, Milko LV, et al.. Evaluating the clinical validity of hypertrophic cardiomyopathy genes.Circ Genom Precis Med. 2019; 12:e002460. doi: 10.1161/CIRCGEN.119.002460

77. Ioannidis NH, Rothstein JH, Pejaver V, et al.. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics, Volume 99, Issue 4, 2016, Pages 877-885.

78. Ionita-Laza I, Makarov V, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Amer J Human Genet. 2012;90(6):1002–13.

79. Iqbal et al. (2020). Burden analysis of missense variants in 1,330 disease-associated genes on 3D provides insights into the mutation effects. bioRxiv
80. Ito K, Patel PN, Gorham JM, et al. Identification of pathogenic gene mutations in LMNA and MYBPC3 that alter RNA splicing. Proc Natl Acad Sci U S A. 2017;114(29):7689-7694. doi:10.1073/pnas.1707741114
81. Jääskeläinen P , Vangipurapu J , Raivo J , et al . Genetic basis and outcome in a nationwide study of Finnish patients with hypertrophic cardiomyopathy. ESC Heart Failure 2019;6:436–45.doi:10.1002/ehf2.12420
82. Jacoby D and McKenna W. Genetics of inherited cardiomyopathy. European Heart Journal 2012; 33 (3): 296-304.
83. Jarcho JA, McKenna W, Pare P, Solomon S, Holcombe R, Dickie S, Levi T, Donis-Keller H, Seidman J and Seidman C. Mapping a Gene for Familial Hypertrophic Cardiomyopathy to Chromosome 14q1. The New England Journal of Medicine 1989; 321: 1372-1378. https://www.nejm.org/doi/full/10.1056/NEJM198911163212005?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed
84. Kääriäinen H, Muilu J, Perola M, Kristiansson K. Genetics in an isolated population like Finland: a different basis for genomic medicine?. J Community Genet. 2017;8(4):319-326. doi:10.1007/s12687-017-0318-4
85. Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7
86. Kelly MA, Caleshu C, Morales A, Buchan J, Wolf Z, Harrison SM, Cook S, Dillon MW, Garcia J, Haverfield E, Jongbloed JDH, Macaya D, Manrai A, Orland K, Richard G, Spoonamore K, Thomas M, Thomson K, Vincent LM, Walsh R, Watkins H, Whiffin N, Ingles J, van Tintelen JP, Semsarian C, Ware JS, Hershberger R, Funke B. Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. Genet Med 2018; 20:351-359
87. Kerminen S, Havulinna AS, Hellenthal G, et al. Fine-Scale Genetic Structure in Finland. G3 (Bethesda). 2017;7(10):3459-3468. Published 2017 Oct 5. doi:10.1534/g3.117.300217
88. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–315. pmid:24487276

89. Kirchner, Florian & Schuetz, Anja & Boldt, Leif-Hendrik & Martens, Kristina & Dittmar, Gunnar & Haverkamp, Wilhelm & Thierfelder, Ludwig & Heinemann, Udo & Gerull, Brenda. (2012). Molecular Insights into Arrhythmogenic Right Ventricular Cardiomyopathy Caused by Plakophilin-2 Missense Mutations. Circulation. Cardiovascular genetics. 5. 400-11. 10.1161/CIRCGENETICS.111.961854.
90. Koehler K, Larntz K. An empirical investigation of goodness-of-fit statistics for sparse multinomials. Journal of the American Statistical Association. 1980;75:336–1278
91. Kryukov G., Shpunt A., Stamatoyannopoulos J., Sunyaev S. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. PNAS. 106 (10): 3871-3876.
92. Kryukov GV, Pennacchio LA and Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007 Apr;80(4):727-39.
93. Kulldorff M, Nagarwalla N: Spatial disease clusters: detextion and inference. Statistics in Medicine. 1995, 14: 799-810.
94. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009, 4: 1073-1081
95. Kutyifa et al. 2018. Clinical aspects of the three major genetic forms of long QT syndrome (LQT1, LQT2, LQT3). https://onlinelibrary.wiley.com/doi/full/10.1111/anec.12537
96. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062-D1067.
97. Lee et al. 2018. Incident Atrial Fibrillation Is Associated With MYH7 Sarcomeric Gene Variation in Hypertrophic Cardiomyopathy. Circulation 11 (9). https://www.ahajournals.org/doi/full/10.1161/CIRCHEARTFAILURE.118.005191
98. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5-23. doi:10.1016/j.ajhg.2014.06.009
99. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13:762–775.
100. Lek, M., Karczewski, K., Minikel, E. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016). https://doi.org/10.1038/nature19057
101. Lelieveld SH, Wiel L, Venselaar H, Pfundt R, Vriend G, Veltman JA, Brunner HG, Vissers L, Gilissen C. Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes. Am. J. Hum. Genet., 101 (2017), pp. 478-484

102. Lemke JR, Geider K, Helbig KL, et al. Delineating the GRIN1 phenotypic spectrum: A distinct genetic NMDA receptor encephalopathy. Neurology. 2016;86(23):2171-2178. doi:10.1212/WNL.0000000000002740

103. Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

104. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, et al. (2014) Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. PLoS Genet 10(7)

105. Lin WY, Lou XY, Gao G, Liu N (2014) Rare Variant Association Testing by Adaptive Combination of P-values. PLoS One 9: e85728.

106. Lin W-Y. Association testing of clustered rare causal variants in case-control studies. PloS One. 2014;9: e94337. pmid:24736372

107. Liu BH. Statistical genomics: linkage, mapping, and QTL analysis: CRC press; 1997

108. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010;6:e1001156

109. Liu X, Wu C, Li C and Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Human Mutation. 37:235-241.

110. Ludbrook J. Analysing 2 × 2 contingency tables: Which test is best? 2013. CEPP. https://doi.org/10.1111/1440-1681.12052

111. Lynch TL 4th, Ismahil MA, Jegga AG, et al. Cardiac inflammation in genetic dilated cardiomyopathy caused by MYBPC3 mutation. J Mol Cell Cardiol. 2017;102:83-93. doi:10.1016/j.yjmcc.2016.12.002

112. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5:e1000384.

113. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-2873.

114. Mann H, Wald A. On the choice of the number and width of classes for the chi-square test of goodness of fit. Annals of Mathematical Statistics. 1942;13:306–317.

115. Manolio TA, Chisholm RL, Ozenberger B, et al. Implementing genomic medicine in the clinic: the future is here. *Genet Med* 2013;**15**:258–267.

116. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT and Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. Circulation 1995; 92 (4): 785-9. https://www.ncbi.nlm.nih.gov/pubmed/7641357

117. Maron BJ, Roberts WC, Arad M, et al. Clinical outcome and phenotypic expression in LAMP2 cardiomyopathy. JAMA. 2009;301(12):1253-1259. doi:10.1001/jama.2009.371

118. Maron, Rowin and Maron 2019. Paradigm of Sudden Death Prevention in Hypertrophic Cardiomyopathy. ahajournals.org/doi/full/10.1161/CIRCRESAHA.119.315159

119. Marstrand et al. 2019. Abstract 15763: Predictors of End-Stage Hypertrophic Cardiomyopathy. Circulation. 140 (1). ahajournals.org/doi/abs/10.1161/circ.140.suppl_1.15763

120. Mathieson I and McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44,** 243–246 (2012). https://doi.org/10.1038/ng.1074

121. McCallum KJ and Ionita-Laza L. Empirical Bayes scan statistics for detecting clusters of disease risk variants in genetic studies. Biometrics. 2015. https://doi.org/10.1111/biom.12331 biometrics

122. McKoy G, Protonotarious N, Crosby A, Tsatsopoulou A, Anastasakis A, Coonar A, Norman M, Baboonian C, Jeffrey S and McKenna W. Identification of a deletion in plakoglobin in arrhythmogenic right ventricular cardiomyopathy with palmoplantar keratoderma and woolly hair (Naxos disease). The Lancet 2000; 335 (9221): 2119-2124. https://www.sciencedirect.com/science/article/pii/S0140673600023795?via%3Dihub#bib10

123. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology Jun 6;17(1):122. (2016) doi:10.1186/s13059-016-0974-4

124. McNally E and Mestroni L. Dilated Cardiomyopathy: Genetic Determinants and Mechanisms. Cirulation 2017; 121: 731-748. https://www.ahajournals.org/doi/full/10.1161/circresaha.116.309396

125. McNally E, MacLeod H, Dellefave-Castillo L. Arrhythmogenic Right Ventricular Cardiomyopathy. 2005 Apr 18 [Updated 2017 May 25]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1131/

126. Medeiros-Domingo A, Bhuiyan ZA, Tester DJ, et al. The RYR2-encoded ryanodine receptor/calcium release channel in patients diagnosed previously with either catecholaminergic polymorphic ventricular tachycardia or genotype negative, exercise-induced long QT syndrome: a comprehensive open reading frame mutational analysis. *J Am Coll Cardiol*. 2009;54(22):2065-2074. doi:10.1016/j.jacc.2009.08.022

127. Merner N, Hodgkinson K, Haywood A, Connors S, French V, Drenckhahn J, Kupprion C, Ramadanova K, Thierfelder L, McKenna W, Gallagher B, Morris-Larkin L, Bassett A, Parfrey P and Young T. Arrhythmogenic right ventricular cardiomyopathy type 5 is a fully penetrant, lethal arrhythmic disorder caused by a missense mutation in the TMEM43 gene. American Journal of Human Genetics 2008; 82 (4): 809-821. https://www.ncbi.nlm.nih.gov/pubmed/18313022

128. Mester J, Ghosh R, Pesaran T, Huether R, Karam R, Hruska K, Costa H, Lachlan K, Ngeow J, Barnholtz-Sloan J, Sesock K, Hernandez F, Zhang L, Milko L, Plon S, Hegde M, Eng C. Gene-specific criteria for PTEN variant curation: recommendations from the ClinGen PTEN expert panel. Hum Mutat 2018; 39:1581-1592

129. Mestroni L, Brun F, Spezzacatene A, Sinagra G, Taylor MR. GENETIC CAUSES OF DILATED CARDIOMYOPATHY. Prog Pediatr Cardiol. 2014;37(1-2):13-18. doi:10.1016/j.ppedcard.2014.10.003

130. Mogensen J, Murphy RT, Kubo T, Bahl A, Moon JC, Klausen IC, Elliott PM, McKenna WJ. Frequency and clinical expression of cardiac troponin I mutations in 748 consecutive families with hypertrophic cardiomyopathy. J Am Coll Cardiol 2004; 44: 2315-25.

131. Moore RK, Abdullah S, Tardiff JC. Allosteric effects of cardiac troponin TNT1 mutations on actomyosin binding: a novel pathogenic mechanism for hypertrophic cardiomyopathy. Arch Biochem Biophys. 2014;552-553:21-28. doi:10.1016/j.abb.2014.01.016

132. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007;615(1-2):28-56. doi:10.1016/j.mrfmmm.2006.09.003

133. Moss AJ, Shimizu W, Wilde AA, et al. Clinical aspects of type-1 long-QT syndrome by location, coding type, and biophysical function of mutations involving the KCNQ1 gene. Circulation. 2007;115(19):2481–2489. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3332528/#!po=50.0000

134. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. (2015) The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test

Hypotheses About Complex Disease. PLoS Genet 11(4): e1005165. https://doi.org/10.1371/journal.pgen.1005165

135. Nabais Sá, M.J., Jensik, P.J., McGee, S.R. et al. De novo and biallelic DEAF1 variants cause a phenotypic spectrum. Genet Med 21, 2059–2069 (2019). https://doi.org/10.1038/s41436-019-0473-6

136. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7:e1001322

137. Ng, S., Buckingham, K., Lee, C. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42,** 30–35 (2010). https://doi.org/10.1038/ng.499

138. Nielsen MW, Holst AG, Olesen SP, Olesen MS. The genetic component of Brugada syndrome. Front Physiol. 2013 Jul 15;4:179.

139. Norgett E, Hatsell S, Carvajal-Huerta L, Cabezas J, Common J, Purkis P, Whittock N, Leigh I, Stevens H and Kelsell D. Recessive mutation in desmoplakin disrupts desmoplakin-intermediate filament interactions and causes dilated cardiomyopathy, woolly hair and keratoderma. Human Molecular Genetics 2000; 9 (18): 2761-2766. https://www.ncbi.nlm.nih.gov/pubmed/11063735

140. Novelli V, Malkani K, Cerrone M. Pleiotropic Phenotypes Associated With PKP2 Variants. Front Cardiovasc Med. 2018;5:184. Published 2018 Dec 18. doi:10.3389/fcvm.2018.00184

141. Oren M, Rotter V. Mutant p53 gain-of-function in cancer. Cold Spring Harb Perspect Biol 2010;2

142. Oza A, DiStefano M, Hemphill S, Cushman B, Grant A, Siegert R, Shen J, Chapin A, Boczek N, Schimmenti L, Murry J, Hasadsri L, Nara K, Kenna M, Booth K, Azaiez H, Griffith A, Avraham K, Kremer H, Rehm H, Amr S, Abou Tayoun A; ClinGen Hearing Loss Clinical Domain Working Group. Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. Hum Mutat 2018; 39: 1593-1613

143. Palm T, Graboski S, Hitchcock-DeGregori SE, Greenfield NJ. Disease-causing mutations in cardiac troponin T: identification of a critical tropomyosin-binding region. Biophys J 2001; 81:2827-37

144. Palz M, Tiecke F, Booms P, et al. Clustering of mutations associated with mild Marfan-like phenotypes in the 3' region of FBN1 suggests a potential genotype-phenotype correlation. Am J Med Genet. 2000;91(3):212-221. doi:10.1002/(sici)1096-8628

145. Papadakis et al. 2018. The Diagnostic Yield of Brugada Syndrome After Sudden Death With Normal Autopsy. Journal of the American College of Cardiology. Volume 71, Issue 11, March 2018. http://www.onlinejacc.org/content/71/11/1204?sso=1&sso_redirect_count=2&access_token=&intcmp=trendmd

146. Passantino et al. 2018. Cardiomyopathies in children – inherited heart muscle disease: Overview of hypertrophic, dilated, restrictive and non-compaction phenotypes. Progress in Pediatric cardiology. 52: 8-15. https://www.sciencedirect.com/science/article/pii/S1058981318300134

147. Pav SE. 2018. BWStest: Baumgartner Weiss Schindler Test of Equal Distributions. R package version 0.2.2.

148. Pearlstone JR, Smillie LB. The binding site of skeletal alpha-tropomyosin on troponin-T. Can J Biochem 1977; 55:1032–8

149. Peng B., Liu X. (2011). Simulating Sequences of the Human Genome with Rare Variants. Human Hereditary. 70; 287-291.

150. Pérez-Palma et al. (2019) Identification of pathogenic variant enriched regions across genes and gene families.

151. Perfetto F, Frusconi S, Bergesio F, et al. The Val142Ile transthyretin cardiac amyloidosis: more than an Afro-American pathogenic variant. *J Community Hosp Intern Med Perspect*. 2015;5(2):26931. Published 2015 Apr 1. doi:10.3402/jchimp.v5.26931

152. Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, French Exome Consortium, Schott J, Redon R, Bellanger L, Dina C. DoEstRare : A statistical test to identify local enrichments in rare genomic variants associated with disease. PLoS One 2017; 12.

153. Peters S, Trummel M and Meyners W. Prevalence of right ventricular dysplasia-cardiomyopathy in a non-referral hospital. International Journal of Cardiology 2004; 97 (3): 499-501. https://www.sciencedirect.com/science/article/pii/S016752730400049X?via%3Dihub

154. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol*. 2016;17(1):157. Published 2016 Jul 14. doi:10.1186/s13059-016-1016-y

155. Petukh M, Kucukkal TG and Alexov E. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum Mutat. 2015 May;36(5):524-534.

156. Platzer K, Yuan H, Schütz H, et alGRIN2B encephalopathy: novel findings on phenotype, variant clustering, functional consequences and treatment aspectsJournal of Medical Genetics 2017;54:460-470.

157. Poetter K, Jiang H, Hassanzadeh S, Master S, Chang A, Dalakas M, Rayment I, Sellers J, Fananapazir L and Epstein N. Mutations in either the essential or regulatory light chains of myosin are associated with a rare myopathy in human heart and skeletal muscle. Nature Genetics 1996; 13: 63-69. https://www.nature.com/articles/ng0596-63

158. Povysil, G., Petrovski, S., Hostyk, J. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat Rev Genet 20, 747–759 (2019). https://doi.org/10.1038/s41576-019-0177-4

159. Price AL., Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 2010;86:832–838.

160. Quarta G, Muir A, Pantazis A, Syrris P, Gehmlich K, Garcia-Pavia P, Ward D, Sen-Chowdhry S, Elliott P and McKenna W. Familial evaluation in arrhythmogenic right ventricular cardiomyopathy: impact of genetics and revised task force criteria. Circulation 2011; 123 (23): 2701-2709. https://www.ncbi.nlm.nih.gov/pubmed/21606390

161. Rahmani G, Kraushaar G, Dehghani P. Diagnosing burned-out hypertrophic cardiomyopathy: Daughter's phenotype solidifies father's diagnosis. J Cardiol Cases, 11 (2014), pp. 78-80

162. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins, Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W201–W206, https://doi.org/10.1093/nar/gkx390

163. Rangaraju A, Lova SM, Calambur N, Nallaria P. Significance of intronic and synonymous MYBPC3 gene variations in hypertrophic cardiomyopathy. Gene reports 2017; 8: 94-99.

164. Ratnapriya R, Zhan X, Fariss RN, et al. Rare and common variants in extracellular matrix gene Fibrillin 2 (FBN2) are associated with macular degeneration. Hum Mol Genet. 2014;23(21):5827-5837.

165. Razali M, Yap B. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. 2011. J. Stat. Model. Analytics. 2.

166. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM. ClinGen — The Clinical Genome Resource. N Engl J Med. 2015; 372:2235-42

167. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8(11):R232. pmid:17976239

168. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J et al. 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. Genet. Med. 2015; 17:405-24.

169. Ritchie GR, Flicek P. Computational approaches to interpreting genomic sequence variation. Genome Med 6, 87 (2014).

170. Robertson SP, Twigg SR, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, Kenwrick SJ, Kim CA, Morava E, Newbury-Ecob R, Orstavik KH, Quarrell OW, Schwartz CE, Shears DJ, Suri M, Kendrick-Jones J, Wilkie AO; OPD-spectrum Disorders Clinical Collaborative Group. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. *Nat Genet* 2003; 33:487–491

171. Ross, Samantha & Bagnall, Richard & Ingles, Jodie & Tintelen, J. Peter & Semsarian, Christopher. (2017). Burden of Recurrent and Ancestral Mutations in Families With Hypertrophic CardiomyopathyCLINICAL PERSPECTIVE. Circulation: Cardiovascular Genetics. 10. e001671. 10.1161/CIRCGENETICS.116.001671.

172. Rotimi CN, Bentley AR, Doumatey AP, Chen G, Shriner D, Adeyemo A. The genomic landscape of African populations in health and disease, Human Molecular Genetics, Volume 26, Issue R2, 01 October 2017, Pages R225–R236,

173. Routledge RD. Resolving the Conflict over Fisher's Exact Test. The Canadian Journal of Statistics / La Revue Canadienne de Statistique. Vol. 20, No. 2 (Jun., 1992), pp. 201-209.

174. Rucco R, Liparoti M, Jacini F, et al. Mutations in the SPAST gene causing hereditary spastic paraplegia are related to global topological alterations in brain functional networks. Neurol Sci. 2019;40(5):979-984. doi:10.1007/s10072-019-3725-y

175. Ruwald MH, Parks XX, Moss AJ, Zareba W, Baman J, McNitt S, et al. Stop-codon and C-terminal nonsense mutations are associated with a lower risk of cardiac events in patients with long QT syndrome type 1. Heart Rhythm, 13 (2016), pp. 122-131

176. Saadi I, Semina EV, Amendt BA. Identification of a dominant negative homeodomain mutation in Rieger syndrome. J Biol Chem. 2001;276:23034–23041.

177. Saltzman AJ, Mancini-DiNardo D, Li C, et al. Short communication: the cardiac myosin binding protein C Arg502Trp mutation: a common cause of hypertrophic cardiomyopathy. Circ Res. 2010;106(9):1549-1552. doi:10.1161/CIRCRESAHA.109.216291

178. Samocha K, Robinson E, Sanders S. et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet 46, 944–950 (2014).

179. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. 2017. Regional missense constraint improves variant deleteriousness prediction. bioRxiv 10.1101/148353

180. Sarquella-Brugada, G., Campuzano, O., Arbelo, E. et al. Brugada syndrome: clinical and genetic findings. Genet Med 18, 3–12 (2016). https://doi.org/10.1038/gim.2015.35

181. Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN. Detecting genomic clustering of risk variants from sequence data: cases versus controls. Hum Genet. 2013;132(11):1301-1309. doi:10.1007/s00439-013-1335-y

182. Schultheiss, H., Fairweather, D., Caforio, A.L.P. et al. Dilated cardiomyopathy. Nat Rev Dis Primers 5, 32 (2019). https://doi.org/10.1038/s41572-019-0084-1

183. Schwartz et al. 2019. Prevalence of the Congenital Long-QT Syndrome. Circulation. 2009;120:1761–1767. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.109.863209

184. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014 Apr;11(4):361-2

185. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010, 7: 575-576

186. Senior AW., Evans R., Jumper J. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020). https://doi.org/10.1038/s41586-019-1923-7

187. Shendure J, Findlay G and Synder M. Genomic medicine-progress, pitfalls, and promise. Cell, 177 (2019), pp. 45-57.

188. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013.Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34: 57–65. 10.1002/humu.22225

189. Shriner, D., Tekola-Ayele, F., Adeyemo, A. et al. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. Sci Rep 4, 6055 (2015)

190. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.

191. Simons and Sella, 2016. Yuval B. Simons, Guy Sella. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr. Opin. Genet. Dev., 41 (December) (2016), pp. 150-158

192. Sims, D., Sudbery, I., Ilott, N. et al. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–132 (2014).
193. Slatkin M. A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. Am J Hum Genet, 75 (2004), pp. 282-293
194. Strande, N.T., Brnich, S.E., Roman, T.S. et al. Navigating the nuances of clinical sequence variant interpretation in Mendelian disease. Genet Med 20, 918–926 (2018). https://doi.org/10.1038/s41436-018-0100-y
195. Sundaram L, Gao H, Padigepati, SR et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 50, 1161–1170 (2018). https://doi.org/10.1038/s41588-018-0167-z
196. Syrris P, Ward D, Asimaki A, Sen-Chowdhry S, Ebrahim HY, Evans A, Hitomi N, Norman M, Pantazis A, Shaw AL, Elliott PM, McKenna WJ. Clinical Expression of Plakophilin-2 Mutations in Familial Arrhythmogenic Right Ventricular Cardiomyopathy. Originally published16 Jan 2006https://doi.org/10.1161/CIRCULATIONAHA.105.561654Circulation. 2006;113:356–364
197. Tango T (1984) The detection of clustering of disease in time. Biometrics 40:15–26
198. Tanjore R, Rangaraju A, Vadapalli S, Remersu S, Narsimhan C, Nallari P. Genetic variations of β-MYH7 in hypertrophic cardiomyopathy and dilated cardiomyopathy. Indian J Hum Genet. 2010;16(2):67-71. doi:10.4103/0971-6866.69348
199. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* **20,** 1054–1060 (2018). https://doi.org/10.1038/gim.2017.210
200. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet*. 2015;47(7):717-726. doi:10.1038/ng.3304
201. Taylor MR, Slavov D, Ku L, Di Lenarda A, Sinagra G, Carniel E, Haubold K, Boucek MM, Ferguson D, Graw SL, Zhu X, Cavanaugh J, Sucharov CC, Long CS, Bristow MR, Lavori P, Mestroni L; Familial Cardiomyopathy Registry; BEST (Beta-Blocker Evaluation of Survival Trial) DNA Bank. Prevalence of desmin mutations in dilated cardiomyopathy.Circulation. 2007; 115:1244–1251. doi: 10.1161/CIRCULATIONAHA.106.646778
202. Tester DJ, Ackerman MJ. Genetics of long QT syndrome. Methodist Debakey Cardiovasc J. 2014;10(1):29-33. doi:10.14797/mdcj-10-1-29
203. The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, , it
204. Thierfelder L, Watkins H, MacRae C, Lamas R, McKenna W, Vosberg H, Seidman J and Seidman C. α-tropomyosin and cardiac troponin T mutations cause familial hypertrophic

cardiomyopathy: A disease of the sarcomere. Cell 1994; 77 (5); 701-712. https://www.sciencedirect.com/science/article/pii/009286749490054X?via%3Dihub#!

205. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res., 13: 2129-2141 .

206. Thomson KL, Ormondroyd E, Harper AR, et al. Analysis of 51 proposed hypertrophic cardiomyopathy genes from genome sequencing data in sarcomere negative cases has negligible diagnostic yield. *Genet Med*. 2019;21(7):1576-1584. doi:10.1038/s41436-018-0375-z

207. Tiso N, Stephen D, Nava A, Bagattin A, Devaney J, Stanchi F, Larderet G, Brahmbhatt B, Brown K, Bauce B, Muriago M, Basso C, Thiene G, Danieli G and Rampazzo A. Human Molecular Genetics 2001; 10 (3): 189-194. https://www.ncbi.nlm.nih.gov/pubmed/11159936

208. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL and Karchin R. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016 Jul 1;76(13):3719-31.https://www.sciencedirect.com/science/article/pii/S0002929715002451#bib14

209. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S. 2017. Optimising genomic medicine in epilepsy through a gene-customised approach to missense variant interpretation. Genome Res 27: 1715–1729.

210. Valdés-Mas, R., Gutiérrez-Fernández, A., Gómez, J. et al. Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. Nat Commun 5, 5326 (2014). https://doi.org/10.1038/ncomms6326

211. van den Hoogenhof MMG, Beqqali A, Amin AS, van der Made I, Aufiero S, Khan MAF, Schumacher CA, Jansweijer JA, van Spaendonck-Zwarts KY, Remme CA, Backs J, Verkerk AO, Baartscheer A, Pinto YM, Creemers EE. RBM20 mutations induce an arrhythmogenic dilated cardiomyopathy related to disturbed calcium handling.Circulation. 2018; 138:1330–1342. doi: 10.1161/CIRCULATIONAHA.117.031947

212. Vergult, S., Dheedene, A., Meurs, A. et al. Genomic aberrations of the CACNA2D1 gene in three patients with epilepsy and intellectual disability. Eur J Hum Genet 23, 628–632 (2015). https://doi.org/10.1038/ejhg.2014.141

213. Vilariño-Güell C, Wider C, Soto-Ortolaza AI, et al. Characterization of DCTN1 genetic variability in neurodegeneration. Neurology. 2009;72(23):2024-2028. doi:10.1212/WNL.0b013e3181a92c4c

214. Waddell-Smith et al. 2020. Pre-Test Probability and Genes and Variants of Uncertain Significance in Familial Long QT Syndrome. Heart, Lung and Circulation. https://www.sciencedirect.com/science/article/pii/S1443950619315811

215. Wallace et al. 2019. Long QT Syndrome: Genetics and Future Perspective. Pediatric Cardiology volume 40, pages1419–1430(2019). https://link.springer.com/article/10.1007/s00246-019-02151-x

216. Walsh R, Mazzarotto F, Whiffin N, et al. Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: the case of hypertrophic cardiomyopathy. *Genome Med*. 2019;11(1):5. Published 2019 Jan 29. doi:10.1186/s13073-019-0616-z

217. Walsh R, Thomson K, Ware J. et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet Med 19, 192–203 (2017). https://doi.org/10.1038/gim.2016.90

218. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164, 2010

219. Wang, Q., Shashikant, C.S., Jensen, M. et al. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Sci Rep 7, 885 (2017).

220. Waring A, Harper A, Salatino S, et al. Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathyJournal of Medical Genetics Published Online First: 30 July 2020. doi: 10.1136/jmedgenet-2020-106922

221. Watkins H, Conner D, Thierfelder L, Jarcho J, MacRae C, McKenna W, Maron B, Seidman M and Seidman C. Mutations in the cardiac myosin binding protein–C gene on chromosome 11 cause familial hypertrophic cardiomyopathy. Nature Genetics 1995; 11: 434-437. https://www.nature.com/articles/ng1295-434

222. Watkins H, Rosenzweig A, Hwang DS, Levi T, McKenna W, Seidman CE, Seidman JG. Characteristics and prognostic implications of myosin missense mutations in familial hypertrophic cardiomyopathy. N Engl J Med 1992; 326:1108-14

223. Whiffin, N., Minikel, E., Walsh, R. et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med 19, 1151–1158 (2017). https://doi.org/10.1038/gim.2017.26

224. White LF, Bonetti M, Pagano M. The Choice of the Number of Bins for the M Statistic. *Comput Stat Data Anal*. 2009;53(10):3640-3649. doi:10.1016/j.csda.2009.03.005

225. Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73, 3–36.

226. Yip KM, Fischer N, Paknia E, Chari A, Stark H. Breaking the next Cryo-EM resolution barrier – Atomic resolution determination of proteins! BioRxiv. doi: https://doi.org/10.1101/2020.05.21.106740

227. Zhang D, Herring JA, Swaney SS, et al. Mutations responsible for Larsen syndrome cluster in the FLNB protein. J Med Genet. 2006;43(5):e24. doi:10.1136/jmg.2005.038695

228. Zhu X, Padmanabhan R, Copeland B, et al. A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLoS Genet*. 2017;13(11):e1007104. Published 2017 Nov 29. doi:10.1371/journal.pgen.1007104

229. Zlotogora, J., Patrinos, G.P. & Meiner, V. Ashkenazi Jewish genomic variants: integrating data from the Israeli National Genetic Database and gnomAD. Genet Med 20, 867–871 (2018).

Appendix 1

R package – *ClusterBurden*

The statistical methods developed here; *BIN-test* and *ClusterBurden* were written into an R package (ClusterBurden) which has been made available for download from the GitHub repository: https://github.com/adamwaring/ClusterBurden. Alongside these statistical methods, the package allows the automatic usage of GnomAD population controls, has functions optimised for quick exome-scans and multiple helper and visualisation functions.

Six functions in the package relate to the statistical methods presented in this chapter; *BIN_test*, *burden_test*, *combine_ps*, *BIN_test_WES*, *burden_test_WES* and *ClusterBurden_WES*. The former three functions are designed for a single use case (i.e. for one gene/protein), whereas the latter three functions are optimised for whole-exome scans, by vectorising many of the calculations. For the *BIN_test*, the minimum a user can provide is the case and control residue positions in the gene of interest, with no further arguments this will calculate a p-value for *BIN-test* with k = n^(2/5) bins, where n is the total number of observed variants. The *burden_test* function simply carries out a FE-

test, but with simple inputs which at a minimum must be the variant counts for cases and controls and the sample sizes. The *combine_ps* functions allows the combination of p-values using either Fisher's method or a weighted Stouffer's method. If either contributing p-value is zero, sometimes the case due to rounding of extremely small values, then the returned result is also zero. Both the *BIN_test* and *burden_test* provide the option of applying Yate's continuity correction to zero count cells and coverage control is implemented when coverage files are supplied by the optional arguments. GnomAD coverage files are available for automatic coverage control of *gnomad_e2* and *gnomad_g2* datasets. For the **_WES* functions, vectorised implementations speed up the calculation of thousands of p-values to allow whole-exome scans to be conducted in seconds. The underling statistical methods are exactly the same.

For the single-gene tests, warning messages indicating potential founders are output to the terminal. For example, variants with unexpectedly high allele counts can have a large influence over the results. Therefore, if variant counts are extreme outliers to the background site-frequency spectrum of each cohort, then these are flagged as potential confounders and may benefit from a sensitivity analysis to further explore their impact. For all functions, basic coverage statistics are reported; for example, the number of residue positions and observed variants excluded by low coverage in either cohort. More advanced coverage statistics are output using the option 'covstats=T'; for example, the specific ranges in the linear protein sequence that were excluded from the analysis.

The resulting output of the WES functions also contain columns that flag potential confounders. The first of these is "pl_flag" which flags large proteins. As huge proteins are more likely to produce false-positive results, protein length is checked and the concern level is reported as; moderate, high and very high for proteins over; 1000, 2000 and 3000 residues respectively. The next flag "cov_flag1", indicates the mean proportion of samples with at least 10X coverage, after excluding residue positions with low coverage. Concern level is reported as; moderate, high and very high for mean 10X coverage of; <80%, <70% and <60% respectively. The final flag "cov_flag2", indicates the proportion of *BIN-test*

bins with less than 80% coverage and is reported as; moderate, high and very high for; 10%, 30% and 50% of bins respectively.

Many functions exist to make whole-exome scans easier with these methods including the automatic use of GnomAD controls. Compressed data files containing the *gnomad_e2* and *gnomad_g2* variant data, as well as their associated coverage files, are available as data for the package. The function *collect_gnomad_controls* is provided to allow automatic retrieval and filtering of GnomAD controls, with the option to switch dataset (exomes, genomes) dependent on the minimum coverage threshold desired. For example, if the minimum coverage threshold is set at 90% and the desired control set is *gnomad_e2*, then for genes with less than 90% coverage in *gnomad_e2*, *gnomad_g2* data will be returned instead along with the correct sample sizes and coverage files.

To support users formatting their own data for use with the package, multiple functions were created. These include a function for annotating user data with GnomAD frequencies, for filtering purposes. The variant-set used for filtering includes all GnomAD variants available from their website, regardless of whether they passed the filters, this is in accordance with most annotation software available. For users to format their coverage files in line for use in the statistical analysis, *format_coverage* was provided to convert the common genome position based coverage files (i.e. chromosome, position) into a transcript and protein position based coverage file (i.e. gene, protein position). This allows users to automatically map their coverage files to canonical ensembl transcripts and the amino acid residue positions of the protein. Coverage statistics for the three bases that make up a amino acid residue, are averaged to produce per-residue coverage statistics. Further functions include extracting protein positions from HGVSp protein consequence format using regular expressions (*get_residue_position*) and excluding variants from a dataset where coverage for either cases or controls falls below a specific threshold (*remove_uncovered_residues*).

To visualise results, functions using the popular ggplot package were made available to plot data and results. To follow up a significant *BIN-test* analysis, the *plot_residuals* function allows visualisation of

the standardised residuals from the analysis to identify regions contributing heavily to the significance. Also available are functions to plot the raw variant positions and coverage data for one (*plot_distribs*) or multiple significant genes from a WES analysis (*plot_signif_distribs*). For exome-wide scans a manhattan-style plotting function is available to identify significant genes and the general distribution of p-values across the genome. Finally the function *plot_features* automatically annotates gene data with a wide range of sequence features from Uniprot, such as structural information and protein domains, and plots the variant distributions alongside chosen feature types.

Each function is fully documented and four R markdown vignettes are supplied alongside the R package which demonstrate some example cases of using the package. Vignettes include a guide on using the statistical tests provided, a guide to whole-exome scans using automatic GnomAD controls, a guide to formatting user data for the functions provided and exploration of the behaviour of *BIN-test* under a putative *gnomad_e2* versus *gnomad_g2* null model.

Appendix 2

Shiny application: Pathogenicity by position

To facilitate access to these variant interpretation models, a web application was developed in the RShiny framework for HCM genes (https://adamwaring.shinyapps.io/Pathogenicity_by_position/). The website comprises four pages; home/about, input, results and analysis. The first page (Home) includes some background to the approach, a link to the paper and instructions on how to use the app, including a youtube video tutorial. The input page (User input) allows researchers to input any observed, novel or hypothetical variants possible from single nucleotide missense variants in the HCM genes. Predicted ORs are returned for these variants from both the *hotspot* and *hotspot+ models*. In the results page (Visualise model), predictions from the training data are displayed alongside predictions for any input entered for users to view predictions in the context of previously observed variants. The final page (Analyse features) allows users to compare which features are the most

important for prediction. The option to specify new models is available alongside access to the pre-trained models.

Appendix 3

More on the burden results…

MYL2 and *MYL3* are both thick filament sarcomeric genes with very strong evidence of association with HCM. A potential explanation for their relatively low rank amongst the *b-significant* genes is their small sizes (166 and 195 amino acid residues respectively). This results in fewer mutations leading to fewer pathogenic variants and lower power to detect an excess of rare-missense variants. *FHL1* ($p < 1.8 \times 10^{-14}$) and *CSRP3* (1.1×10^{-5}), also amongst the *b-significant* genes, are both members of the LIM domain proteins. *FHL1*, four and a half LIM domains 1, is an X-linked gene, associated with *TTN*, and predominantly expressed in striatic muscle. It is also linked to other muscle syndromes but can cause isolated HCM. It was first discovered to be associated with HCM in 2012 (Friedrich *et al.*) when it came to attention as a candidate gene after being shown to cause distinct myopathies that were sometimes associated with cardiomyopathy. The role of *FHL1* in the pathogenesis of HCM was subsequently demonstrated by molecular studies (Friedrich *et al.* 2014). *CSRP3*, ranked twelfth in significance, is involved in structure related to the sarcomere and also has strong published evidence of association. Variants in this gene have been associated with a reduced penetrance with mild expressivity (Gómez *et al.* 2017). This is likely to result in more sub-clinical cases and explains its relatively lower odds-ratio; 3.11.

FLNC is an interesting addition to the significant subset of genes as it was discovered relatively recently by whole-exome sequencing (Valdés-Mas *et al.* 2014). It encodes a protein involved in both striated and cardiac muscle function, however, variants have been demonstrated to cause isolated HCM. Although highly significant ($p < 8.1 \times 10^{-10}$), it has a relatively low odds-ratio of 1.87 with significance in part driven by its large size (2,725 residues). This low odds-ratio suggests a substantial amount of

incomplete penetrance for variants in this gene. Furthermore, it may also explain why it was only discovered relatively recently.

The gene *GLA* ($p < 5.6 \times 10^{-8}$), ranked ninth in significance, is thought to cause the X-linked disorder Fabry disease. However, specific variants are associated with non-syndromic FD, accompanied by myocardical hypertrophy, clinically comparable to HCM (Chaves-Markman *et al.* 2019). *DMD* ($p < 8.6 \times 10^{-5}$) is an unexpected member of this elite group of HCM-associated genes, and not present on many HCM-specific gene-panels (Ingles *et al.* 2019, Walsh *et al.* 2017). *LAMP2* ($p < 0.0014$), narrowly *b-significant*, is a phenocopy gene, definitely associated with Danon disease, but can also present as isolated left ventricular hypertrophy (Maron *et al.* 2009). Here an odds-ratio of 2.71 does directly relate to the proportion of pathogenic variants in the gene and penetrance, as it is also affected by the proportion of individuals with Danon disease that exhibit symptoms mistaken for hypertrophic cardiomyopathy.

A further eleven genes were nominally significant (*n-significant*). Amongst them, another troponin *TNNC1* ($p < 0.0043$), which has moderate evidence of association in a recent assessment of the clinical validity of HCM genes (Ingles *et al.* 2019). However, the remaining *n-significant* genes have little evidence of association. *JUP* ($p < 0.018$), although a known causative gene for ARVC (Li *et al.* 2011), has no prior evidence of association with HCM. *ANKRD1* ($p < 0.021$), although previously connected to HCM, has been reported to have limited true evidence of association (Ingles *et al.* 2019). Similarly, the remaining *n-significant* genes; *SCN5A, BAG3, PRKAG2, DES, PKP2, FHL2, DSG2* and *RBM20,* also have limited evidence of association. For the remaining 18 genes with p-values above 0.05, there were no genes that were expected to harbour an excess of rare-variants.

In DCM, eight genes passed multiple testing correction, and a further two were *n-significant*. Variants in many sarcomeric genes cause diverse cardiomyopathy phenotypes (Bollen and van der Velden 2017). Based on this burden analysis, *MYH7* is clearly an important gene for both HCM and DCM. Although the highest-ranking gene based on p-value, with an OR of 4.9, it is actually ranked eighth in

effect size. Multiple hypotheses have been put forth to explain the role of *MYH7* variants in both HCM and DCM including; different impact on energy compromise, does-effect of the mutant protein or even additional environmental and genetic modifiers (Tanjore *et al.* 2010). Recent studies have revealed that *MYH7* variants cause either HCM or DCM by disrupting contraction and relaxation of the sarcomere. While DCM is caused by the contractility effects of variants, HCM is caused by disruption to the interaction of myosin-head pairs (Alamo *et al.* 2017).

TTN, the most frequently involved gene in dilated cardiomyopathy (Mestroni et al. 2014), was excluded from this section of the analysis due to the complexity resulting from its size. *LMNA* ($p < 2.2 \times 10^{-6}$), ranked third in significance and a known cause of idiopathic DCM, is often associated with conduction system disease that occurs before heart failure symptoms (Hershberger and Morales 2016). *RBM20* ($p < 6.6 \times 10^{-5}$), is a splicing factor that is involves in key cardiac genes and its variants cause aggressive DCM that carry an increased risk of ventricular arrhythmias (van den Hoogenhof *et al.* 2018). *DSP* ($p < 8.5 \times 10^{-5}$), the cell-to-cell adhesion protein desmoplankin, is a known cause of ARVC (Bauce *et al.* 2005) but its association with DCM here is unexpected. The association of *MYBPC3* ($p < 0.00035$) is more complex, as although there have been associations between *MYBPC3* variants and DCM (Ehlermann *et al.* 2008), this may be an end-stage phenotype of HCM described as "burn-out phase" (Lynch *et al.* 2017; Rahmani *et al.* 2015). Variants in *PKP2* ($p < 0.00133$) have been associated with DCM (Novelli *et al.* 2018) and *DES* ($p < 0.0015$) is a phenocopy gene that caused Desmin-related myofibrillar myopathy but can also cause isolated DCM (Taylor *et al.* 2007). For the two *n-significant* results; variants in *FLH1* have been previously associated with DCM (Friedrich *et al.* 2012) and there is some evidence for an association with *FLNC*.

PKP2, the gene encoding Plakophilin-2, is the strongest associated gene, consistent with previous data that suggest that 45-73% of ARVC cases are attributed to pathogenic variants in this gene (Hall *et al.* 2018). With an odds-ratio of 16.7, *PKP2* also has one of the largest excess of rare-missense variants observed across our data. This may be in part due to the reduced genic heterogeneity for this disease.

The gene *JUP* has been previously associated with ARVC in a recessive fashion. It was not possible to identify compound heterozygotes in *gnomad_e2* without sample-level information, so it was not possible to test for a burden in a recessive manner. Causal variants in the genes; *DSP*, *DSC2* and *DSG2*, can cause both dominant and recessive ARVC (Hall *et al.* 2018). This may be one factor in their reduced penetrance compared to *PKP2* variants, as gene-level odds-ratios are at least three times smaller than *PKP2* for these genes. Although TMEM43 has been previously associated with ARVC (McNally *et al.* 2017), the proportion of cases attributed to variants in this gene is low (0-2%), which alongside low sample size (n=266) may explain its lack of association here.

Pathogenic variants in the gene *RYR2* was more commonly associated with catecholaminergic polymorphic ventricular tachycardia (CPVT), however evidence for its association with LQTS is also published (Fukuyama *et al.* 2017). *KCNQ1, KCNH2* and *SCN5A* are defined as the three major non-syndromic LQTS genes while *CALM2* and *CACNA1C* (Nielsen *et al.* 2013) are minor genes. *KCNJ2* is associated with Andersen-Tawil syndrome (Tester and Ackerman 2014). Genes in this group has the highest observed odds-ratio across all diseases. *KCNQ1* (OR=45.1), *CALM2* (OR=26.6) and *KCNH2* (OR=22.90) has the highest odds-ratios across all disease-gene pairs. With only a marginally significant p-value, it was surprising to see such a high odds-ratio for *CALM2*. This was driven by a single variant in 584 LQTS patients and eight variants in 122, 893 *gnomad_e2* samples. The confidence intervals surrounding this estimate were therefore very wide (95% Cis = 0.6-197.7).