Enhanced prediction of gene and missense rare-variant pathogenicity by joint analysis of gene burden and amino-acid residue position

Waring A.J.[1], Harper A.R.[1,2], Salatino S.[1], Kramer C.M.[4], Neubauer S[2], HCMR Investigators, Thomson K.L.[2,3], Watkins H.[1,2], Farrall M.[1,2]

1. Wellcome Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

2. Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Division of Cardiovascular Medicine, John Radcliffe Hospital, Oxford, OX3 9DU

3. Oxford Medical Genetics Laboratories, Churchill Hospital, Oxford, OX3 7LE

4. University of Virginia Health System, Charlottesville, VA, USA

Correspondence to: Martin Farrall (martin.farrall@cardiov.ox.ac.uk)

Author contributions

Waring: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation

Harper, Salatino, Kramer, Neubauer, HCMR Investigators: Data Curation

Thomson: Conceptualization, Writing – Review & Editing

Watkins: Conceptualization, Writing – Review & Editing

Farrall: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Abstract

Although rare missense variants underlying a number of Mendelian diseases have been noted to cluster in specific regions of proteins, this information may be underutilized when evaluating the pathogenicity of a gene or variant. We introduce *ClusterBurden* and *GAMs*, two methods for rapid association testing and predictive modelling, respectively, that combine variant burden and amino-acid residue clustering, in case-control studies. We show that *ClusterBurden* increases statistical power to identify disease genes driven by missense variants, in simulated and experimental 34-gene panel for hypertrophic cardiomyopathy. We then demonstrate that *GAMs* can be used to apply the ACMG criteria PM1 and PP3 quantitatively, and resolve a wide range of pathogenicity potential amongst variants of uncertain significance. An R package is available for association testing using *ClusterBurden,* and a web application (*Pathogenicity_by_Position)* is available for missense variant risk prediction using GAMs for six sarcomeric genes. In conclusion, the inclusion of amino-acid residue positional information enhances the accuracy of gene and rare variant pathogenicity interpretation.

Author Summary

Two statistical methods have been developed that utilize signal in the residue position of missense variants. The first is a rapid association method that tests the joint hypothesis of an excess of rare-variants and rare-variant clustering. The method, *ClusterBurden*, is powerful when rare-missense variants cluster in discrete pathogenic regions of the protein. It can be applied to exome-scans to discover novel Mendelian disease-genes, that may not be identified by classic burden testing. The second method is a statistical model for rare-missense variant interpretation. It provides superior predictive performance compared to generic *in silico* predictors by training on our large case-control dataset. The method represents a data-driven quantitative approach to apply hotspot and *in-silico* prediction criteria from the ACMG variant interpretation guidelines.

1. Introduction

It has been frequently reported that pathogenic missense variants, tend to cluster in specific regions or domains of proteins [1-7]. A plausible mechanism underpinning this phenomenon is the presence of multiple loss or gain-of-function variants within functionally important domains, which disrupt critical aspects of protein function [8]. Despite numerous examples of variant clustering, there have been few attempts to explicitly model variant residue position as a predictor of pathogenicity [9].

Pathogenic genes for Mendelian diseases were historically identified by linkage and candidate gene studies in multiple affected families [10]. Advances in high-throughput DNA sequencing technology allow scanning of whole-exome or whole-genome sequencing of patient cohorts to offer an alternative strategy to identify novel pathogenic genes and variants. The aggregated burden of variants in affected cases compared to healthy controls has proved to be a useful statistical test to confirm the pathogenicity of candidate genes [11] as well as identify novel putative pathogenic genes [12]. However, for genes where variant pathogenicity is not uniform, including positional information alongside burden may improve power to detect associated genes.

The American College of Medical Genetics and Genomics (ACMG) have produced general guidelines to interpret variant pathogenicity [13]. These guidelines integrate a variety of diverse data, including population frequency, functional and segregation data, and classify variants into five categories from benign to pathogenic. However, due to low counts of observed variants in case datasets, a lack of segregation data and functional evidence, or presence in control datasets, many variants fall into the category 'variant of uncertain significance' (VUS).

Hypertrophic cardiomyopathy (HCM), a relatively common disease (1 in 500 prevalence), is an exemplar for this. It is a major cause of heart disease in people of all ages [14-15] and a cause of sudden cardiac death. Eight sarcomeric genes collectively provide firm molecular diagnoses for ~27% of HCM patients, with a further ~13% of patients carrying VUS in the same genes. It has been suggested that disease and gene-specific approaches are needed to improve interpretation [16] and guidelines have been produced for specific genes and/or disease areas [17-21]. Missense variant clustering is a gene-specific metric that falls under the ACMG evidence category PM1('mutational hotspot'). However, there has previously been limited data to define these 'hotspots' quantitatively. Therefore this category is only used subjectively for a limited number of genes.

Here we introduce two new statistical methods to aid in the identification of novel causal genes and reassess the pathogenicity of variants in well-established disease genes. We show that information on a variants' amino-acid position can usefully augment power to detect novel pathogenic genes compared to simpler, mutation burden-based tests. We apply the methods to an extensive dataset of 5,338 HCM cases and use 125,748 gnomAD population controls [22], to visualise the landscape of burden and position signals across 34 cardiomyopathy genes. Finally, we develop and apply a flexible statistical modelling framework that can integrate variant burden with residue annotation data to predict pathogenicity potential in six well-established pathogenic HCM genes.

2. Methods

A computationally rapid, rare variant association test (*ClusterBurden*) was developed to test the joint hypothesis of an excess of rare missense variants, clustered with respect to amino-acid residue position in case-control data. This was accomplished by combining p-values from a rare variant burden test with a second binning test that detects variant clustering.

2.1 ClusterBurden – A combined rare variant burden and cluster test

A 2 x 2 contingency table was constructed to summarise variant carrier status in cases and controls. The tables, which sometimes include low (fewer than 5) or zero cell counts, are conveniently analysed by Fisher's exact test (FE). While we recognize that, as the number of observed variants is not fixed, a singly-conditioned exact test of odds-ratios (OR) is a more appropriate statistical test for case-control studies [23], FE was pragmatically used for its speed of computation and ease of implementation. As there are no known examples of a protective burden of rare exonic variants in cardiomyopathy, we consider only one-sided significance tests, where the p-value is the proportion of contingency tables with ORs *higher* than the observed one.

To test for distributional differences, along the length of a protein, the protein's linear sequence of amino-acid residues was split into k bins of equal length. To test the hypothesis that variants cluster in specific bins, in affected cases compared to controls, we applied a chi-squared two-sample test (hereafter *BIN-test*) to a k x 2 contingency table of binned variant counts in cases and controls. 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 R is the frequency in bin K_i for the cases and S is the frequency in bin K_i for controls. The test statistic B is asymptotically chi-squared distributed with k degrees of freedom. We used the $k \sim n^{2/5}$ heuristic [24] to select the number of bins (k) dependent on n, the number of observations. We compared the performance of the *BIN-test* with Anderson-Darling (AD) [25] and Kolmogorov-Smirnov (KS) [26] two-sample distributional tests.

Fisher's method [27] was used to calculate the joint significance of the contributing burden and cluster tests by summing the natural logarithms of the two p-values, multiplied by minus two, to produce a chi-squared statistic with 4 degrees of freedom. An important assumption of this method is that the two p-values are uncorrelated; this was assessed in simulated data using Spearman's rank correlation test [28].

2.2 A combined rare variant burden and annotation prediction model

We fitted gene-specific generalized additive models (GAM) [29] implemented in the R package "mgcv" [30], to model rare missense variants in firmly-established pathogenic genes. The model was trained on disease status in our HCM-gnomAD case-control dataset and is unsupervised with respect to variant pathogenicity. The model is therefore not reliant on previous classifications and includes all rare-variants in both cases and controls to make estimates of odds-ratios. This allows us to quantify pathogenicity and uncertainty for variants, taking background variation and incomplete penetrance into account.

GAMs adaptively model linear and non-linear relationships of varying complexity, between explanatory variables (e.g. burden, residue position) and the response variable (case-control status). Non-linear terms specified as smooth functions are built from underlying basis functions whose linear combination sum to smooth the predictor fully. Automatic optimisation by penalized maximum likelihood reduces over-fitting as increased 'wiggliness' comes at a user-specified cost. GAMs therefore facilitate a flexible and parsimonious non-linear model-selection strategy to integrate rare variant burden and amino-acid annotation data.

The primary predictive features were carrier status (to model rare-variant gene-level burden) and residue position (to model clustering). Secondary predictive features included several variant prediction scores (e.g. SIFT [31]) extracted from the dbNSFP4.0 database [32]. To include the gene 'burden' in the model, non-carriers (i.e. samples without a variant in the gene) were included in the model. However, variant-level features such as amino-acid position are undefined (i.e. meaningless) for non-carriers, so a nested hierarchical model structure is required, whereby features are included in the model only as an interaction

with carrier status. Carrier status as a binary indicator variable is then multiplied with the model matrix for smoothed terms, such as amino-acid position. For non-carriers, the indicator variable is zero to exclude this undefined data from the analysis.

The structure of an example GAM with three independent predictors, a [0,1] indicator variable for carrier status (*car*), a continuous variable for amino-acid position (*aapos*) and the continuous variable for SIFT score (*sift_score*) is as follows:

$$Ln(P/1\text{-}P) = \beta_0 + \beta_1\, car + s_1\,(aapos,\ \text{by} = car) + s_2\,(sift\_score,\ \text{by}=car) + \varepsilon$$

where ***Ln(P/1-P)*** is a logistic function specifying the probability of being a case (***P***), $\boldsymbol{\beta_0}$ the model intercept, $\boldsymbol{\beta_1}$ a linear coefficient for *car*, $\mathbf{s_1}$ a smoothed (i.e. non-linear) function for *aapos*, $\mathbf{s_2}$ a smoothed function for *sift_score*, **by** is used to generate factor smooth interactions and $\boldsymbol{\varepsilon}$ is a binomial random residual error term. GAMs were fitted using thin-plate regressions and parameters were estimated by restricted maximum likelihood (REML). A strict two-stage feature selection procedure was implemented to avoid overfitting. In stage 1, only features with a marginal p-value < 0.002 (0.05/24 for Bonferroni correction) were selected. In stage 2, backwards elimination was implemented using p-values Bonferroni corrected for the number of features selected in stage 1.

Six genes (*MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *TNNI3* and *TNNT2)* each carrying at least 20 rare missense variants in our HCM cases and with a significant positional signal, were informative for GAM analysis. Carrier status (gene burden) and amino-acid residue position (local burden) were included as primary predictive features (*posGAM*). A two-stage selection of secondary features (e.g. SIFT scores) was then performed to model the relationship between HCM status with the primary features and any retained secondary features (*fullGAM*). ORs with standard errors were computed for each variant to predict their respective strengths of association. Two other sarcomeric genes that showed evidence of excessive burden, *ACTC1* and *TPM1,* but had insufficient evidence of clustering, so positional GAM modelling was less informative for these genes.

Our GAM models are capable of integrating pathogenicity data underpinning two ACMG criteria; PM1 (mutational hotspot) and PP3 (*in silico* prediction algorithms). There is currently no quantitative way to activate PM1 and PP3. For criteria PP1 and PS4, [17] propose OR thresholds for *MYH7* to quantify evidence of pathogenicity as 10-30 for supporting, 30-100 for moderate and >100 for strong. Adopting the same thresholds, GAMs can compute variant-specific ORs that can quantitatively apply the PM1 and PP3 criteria, as supporting, moderate or strong evidence.

The relative performance of alternative GAM models was assessed by receiver operator characteristic (ROC) curves across ten cross-fold validations using 80:20 splits. Model predictions were stratified by manual variant classifications, obtained from an in-house database to determine correlation.

2.3 Simulated and observed data

A computationally rapid forward-time rare variant simulator was coded in R to model missense variants in proteins, clustered in pathogenic regions under high selection (S1 Methods). Briefly, demography was based on European population history, mutations followed an infinite sites model and mating was dependent on selection. Three clustering scenarios were considered: 1) a uniform pathogenicity model 2) a single cluster model equivalent to a protein with one discrete pathogenic region and 3) a multiple cluster model where the protein contains several pathogenic regions. Simulation parameters such as mutation rate were tuned to generate a simulated case-control cohort with properties comparable to weakly associated genes in our HCM-gnomAD dataset. Simulated ORs were kept intentionally low for two reasons; to ensure power was <95% for each test to facilitate comparisons of power, and to highlight that the power to detect associations for relatively low penetrance genes requires highly efficient methods.

Synthetic data were generated for 5,000 cases and 125,000 controls and variants were filtered at a minor allele frequency of <0.0001. The type 1 error and power of *ClusterBurden* was compared to two published

position-informed association methods: DoEstRare [9] and CLUSTER [33], and three position-uninformed methods: C-alpha [34], SKAT [35], WST [36], using 10,000 replicate datasets, under the three different clustering scenarios and two different protein lengths (500 and 1,000). Power and type 1 error were calculated using the *(r+1)/(n+1)* estimator where ***r*** represents the number of simulated datasets with p-values less than the 0.05 and ***n*** is the number of simulations [37]. Implementation details for all tests considered are found in S1 Table.

Next-generation sequence (NGS) data for 34 cardiomyopathy genes were available for two large HCM cohorts; 2,757 probands referred to the Oxford Medical Genetics Laboratory (OMGL) for genetic testing and 2,636 HCM probands recruited to the HCMR project [38]. High coverage exonic sequences were captured by target enrichment and sequenced on the MiSeq platform (Illumina Inc.). Bioinformatic processing of NGS data followed the Genome Analysis ToolKit version 4 best practice guidelines (https://software.broadinstitute.org/gatk/best-practices/). OMGL variants were confirmed by Sanger sequencing, HCMR variants were manually checked by inspection of BAM files.

In all analyses, only missense variants were considered, defined as single nucleotide polymorphisms that cause a single amino acid residue substitution in the protein sequence. The gnomAD population reference database was used as a control group, which includes variant frequency data based on up to 125,748 individuals. Individual sample-level information is unavailable for gnomAD, so we assumed that no sample carried more than one variant in each gene. Although individual sample-level information was available for the HCM cases, the same simplification was applied. This was justified by the low proportion of case samples with multiple variants across the gene. For 5,338 samples in the combined OMGL-HCMR dataset, over 34 genes, there were an average of 0.77 samples with multiple missense variants post-filtering per gene. We define rare variants by reference to their allele frequencies in unaffected controls, which for rare Mendelian diseases is approximated by population reference databases such as gnomAD [11-12]. Rare variants were selected with a gnomAD population maximum (*popmax*) allele frequency less than 0.0001 to exclude potentially common, and thus unlikely to be pathogenic for HCM, ancestry-specific polymorphic variants.

3. Results

3.1 ClusterBurden – a clustered rare variant burden test

Correlation between p-values from the *BIN-test* and FE tests were compared under the null and disease (cluster/burden) models in simulated data. For the disease model, there was an expected positive correlation (Spearman's rank correlation rho=0.398) between the p-values, as the power of both tests covary with the number of observed data points (i.e. number of rare variant carriers). However, under the null model, the p-values were completely uncorrelated, satisfying the independence assumption of the Fisher p-value combination method.

Table 1 shows summary statistics for each simulation model as well as type 1 error and power estimates. For the disease models mean ORs across 10,000 replicate populations range from 1.71 to 2.49 across the different simulation models. Under the null hypothesis of no excessive burden or differential clustering, the type 1 errors for the BIN and Anderson-Darling (AD) tests were adequately controlled, and the Kolmogorov-Smirnov (KS) test was slightly conservative. The *BIN-test* had greater power than AD or KS under both clustering scenarios.

Table 1: Type 1 error and power comparisons and properties of simulated datasets. Mean estimates from 10,000 simulations are shown for the null or disease hypothesis for 6 scenarios, consisting of three clustering models (uniform, one cluster and multiple clusters) and two protein lengths (500 and 1000); estimates for the disease model are followed by the null model value in parentheses. The upper table reports the number of unique variants in cases and controls, as well as the burden odds-ratio. The lower table shows the proportion of significant results (alpha < 0.05) under each scenario for the disease model (i.e. power) and the null model (i.e. type 1 error). The 'Uniform' model describes a gene where mutated amino-acid residues uniformly increase the risk of disease. The 'One cluster' model represents a gene with a single causal region and the 'Multi-cluster' model represents a gene with multiple causal regions. The 'Null' model represents a gene with neither burden nor position signal. For the clustered models, *ClusterBurden* was the most powerful test and *BIN-test* was the most powerful position test (highlighted in green). Power was not computed for WST and SKAT as the type 1 error for these tests was excessively inflated.

	Clustering model:	**Uniform**		**One cluster**		**Multiple clusters**	
	Protein length:	**500**	**1000**	**500**	**1000**	**500**	**1000**
Cases	**Unique variants**	13.1 (5.8)	26.1 (11.6)	9.7 (6.2)	19.4 (12.6)	12.4 (6.3)	24.8 (12.7)
Controls	**Unique variants**	81.9 (82.2)	163.9 (164.5)	85.7 (85.9)	171.1 (172.2)	86.0 (86.3)	172.3 (172.9)
	Burden odds-ratio	2.49 (1.02)	2.49 (1.03)	1.71 (1.02)	1.71 (1.03)	2.28 (1.01)	2.26 (1.02)

Burden	**Fisher-exact test**	0.8 (0.042)	0.97 (0.05)	0.39 (0.044)	0.58 (0.05)	0.69 (0.042)	0.87 (0.053)
Position	***BIN-test***	**0.058 (0.051)**	**0.058 (0.047)**	**0.31 (0.051)**	**0.52 (0.05)**	**0.45 (0.05)**	**0.73 (0.047)**
	Kolmorgorov-Smirnov	0.036 (0.023)	0.039 (0.03)	0.16 (0.024)	0.31 (0.032)	0.26 (0.026)	0.49 (0.033)
	Anderson-Darling	0.056 (0.051)	0.054 (0.049)	0.18 (0.05)	0.31 (0.05)	0.27 (0.05)	0.48 (0.049)
Position-informed RVATs	***ClusterBurden***	**0.72 (0.05)**	**0.94 (0.05)**	**0.48 (0.049)**	**0.7 (0.053)**	**0.76 (0.047)**	**0.93 (0.052)**
	DoEstRare	0.8 (0.058)	0.96 (0.062)	0.46 (0.06)	0.64 (0.06)	0.74 (0.058)	0.9 (0.063)
	CLUSTER	0.82 (0.054)	0.97 (0.056)	0.42 (0.055)	0.61 (0.059)	0.71 (0.053)	0.88 (0.057)
Generic rare-variant association tests	**C-alpha**	0.75 (0.051)	0.94 (0.056)	0.42 (0.053)	0.59 (0.058)	0.7 (0.051)	0.87 (0.057)
	WST	*NA* (0.17)	*NA* (0.089)	*NA* (0.16)	*NA* (0.086)	*NA* (0.16)	*NA* (0.08)
	SKAT	*NA* (0.18)	*NA* (0.2)	*NA* (0.18)	*NA* (0.21)	*NA* (0.18)	*NA* (0.2)

Type 1 errors for *ClusterBurden*, DoEstRare, CLUSTER and C-alpha were all well controlled. Conversely, SKAT and WST showed markedly inflated false-positives under the null and were not examined further. *ClusterBurden* was the most powerful method when clustering was present, whereas CLUSTER was most powerful under the uniform (i.e. burden-only) association model. Amongst the position-informed tests, *ClusterBurden* was the most rapid to compute per gene (<1 second) whereas DoEstRare took >20 minutes and CLUSTER took >4 minutes.

3.2 ClusterBurden analysis of cardiomyopathy gene panel

We examined 34 cardiomyopathy genes for rare missense variant associations with the FE (burden), *BIN-test* (cluster) and *ClusterBurden* (combined cluster/burden) tests in HCM cases and gnomAD controls (**Fig. 1**). Significance thresholds were Bonferroni adjusted to allow for 34 genes x 3 methods (i.e. p-values adjusted for 102 tests to $p < 0.00049$). Significant burden signals were detected in 11 genes with FE; *MYH7* ($p < 5.44 \times 10^{-252}$), *MYBPC3* ($p < 1.74 \times 10^{-229}$), *TNNI3* ($p < 1.46 \times 10^{-50}$), *TNNT2* ($p < 1.11 \times 10^{-24}$), *TPM1* ($p < 6.56 \times 10^{-21}$), *ACTC1* ($9.61 \times 10^{-14}$), *GLA* ($1.61 \times 10^{-10}$), *FLH1* ($1.02 \times 10^{-9}$), *MYL2* ($1.87 \times 10^{-9}$), *CSRP3* ($3.56 \times 10^{-8}$) and *MYL3* ($6.53 \times 10^{-6}$). The *BIN-test* detected significant cluster signals for 6 core sarcomeric genes; *MYH7* ($p < 1.36 \times 10^{-73}$), *MYBPC3* ($p < 1.55 \times 10^{-78}$), *TNNI3* ($p < 3.34 \times 10^{-13}$), *MYL2* ($p < 5.83 \times 10^{-10}$), *TNNT2* ($p < 1.69 \times 10^{-7}$) and *MYL3* ($p < 1.7 \times 10^{-4}$). Two additional core sarcomeric genes did not show Bonferroni significant associations; *ACTC1* ($p<0.0412$) and *TPM1* ($p< 0.0494$). *ClusterBurden* confirmed the association for 12 genes that showed burden signals and calculated substantially lower p-values for all eight core-sarcomeric genes, consistent with enhanced power for this approach.

3.3 Combining rare variant burden, amino-acid position and annotation scores to predict pathogenicity

Figure 2 summarises the results of GAM analyses of six sarcomeric proteins. The relationship between a variant's predicted OR and its location within the linear amino-acid sequence illuminates the architecture of HCM association across each pathogenic gene. Predicted ORs vary substantially across the linear sequence of the proteins. For all models, except for *MYL3* which has no secondary features, features from dbNSFP further partitioned variant risk. Risk predictions in the *posGAM* (i.e. burden and position only) and *fullGAM* (i.e. including primary and secondary features) are accompanied by 95% confidence intervals (S1-12 Figures). For each gene, features that passed selection and their marginal p-values are displayed in Table 2. Due to the covariance of power to detect an association and the number of observations, genes with more variants supported the use of more features.

Table 2: Features used to generate gene-specific generalized-additive models and their marginal p-values for six sarcomeric genes.

Gene	Features	P-values
MYH7	Residue position, CADD, MPC, MVP, MetaLR, MutationAssessor, PrimateAI, REVEL and SiPhy 29way logodds	9.9×10^{-65}, 5.5×10^{-25}, 9.2×10^{-57}, 1.3×10^{-25}, 2×10^{-33}, 1.5×10^{-7}, 9.6×10^{-15}, 2.3×10^{-39} and 1.3×10^{-8}
MYBPC3	Residue position, CADD, Deogen2, MetaLR, MutationAssessor, PROVEAN, REVEL, and VEST4	1.7×10^{-36}, 5.6×10^{-37}, 6.7×10^{-52}, 1.7×10^{-38}, 3.4×10^{-55}, 7.3×10^{-34}, 1.3×10^{-55} and 7.6×10^{-68}
TNNT2	Residue position, MPC, SiPhy 29way logodds and Phylop 100way vertebrate	3.4×10^{-7}, 8×10^{-6}, 2.8×10^{-4} and 3.9×10^{-6}

TNNI3	Residue position and MPC	$6.8x10^{-11}$ and $1.2x10^{-7}$
MYL2	Residue position, MutationAssessor and primateAI	$2.1x10^{-7}$, $1.6x10^{-5}$ and $2.7 x 10^{-4}$
MYL3	Residue position	$3.1x10^{-4}$

The receiver operator characteristic (ROC) area under the curve (AUC) metric was calculated for ten cross-fold validations with 80:20 splits. The two GAM models, *fullGAM* and *posGAM*, were compared to models that stratified samples into cases or controls based on individual in silico variant prediction scores. Linear thresholds for the individual score models were determined to give the maximum possible AUC for each score, conditional on the observed data. Figure 3 displays the mean and standard deviation AUC across the ten cross-fold splits for the five highest-scoring models in each gene. For all six HCM genes, the *fullGAM* had a higher mean AUC than any individual in silico predictor. With the exception of *MYBPC3*, the *posGAM* performed better than any *in silico* predictor from dbNSFP. The AUC standard deviations for *MYH7* and *MYBPC3* were considerably lower than the remaining genes.

There was a strong correspondence between *fullGAM* predictions and expert classifications made by OMGL for variants in each gene (Fig. 4). In *MYH7* cases, mean predicted ORs for pathogenic, likely pathogenic and VUS variants were 74, 50 and 20 respectively. Notably, the VUS class had high heterogeneity, with predicted ORs ranging from 0.25 to 197. For 53% of these VUS's, limited information is available, as they are observed in a single case and were not present in gnomAD. Based on these observed frequencies, and after Haldane continuity correction [39], the empirical OR for each of these variants is 44.9 with wide uninformative 95% CIs [1.5, 1338.3.] Conversely, ORs generated by the *fullGAM* for each of these variants have much higher confidence and provide a range of different point estimates. Five of these VUS's (p.Glu894Lys, p.Met435Thr, p.Lys865Glu, p.Phe758Cys, p.Gly407Val) have a predicted ORs greater than 100 with lower 95% CIs of at least 57. Three (p.Glu45Asp, p.Gln27Arg, p.Gln1237His) have ORs less than 1 with upper 95% CIs at least below 1.6.

Predictions from the GAMs allow quantitative application of the ACMG PM1 and PP3 criteria. For *MYH7*, utilising the PM1 rule in the *posGAM* model gives; none, supporting or moderate evidence for 25.9%, 53.7% and 20.4% of variants respectively. With the inclusion of PP3 in the fullGAM model; 13%, 18%, 58.4% and 10.6% of variants were assigned none, supporting, moderate or strong evidence of pathogenicity.

A web application, *Pathogeniicty by postion*, was developed to facilitate the exploration of the GAM modelling approach (R Shiny: https://adamwaring.shinyapps.io/Pathogenicity_by_position/). Users can explore models and submit their own missense variants to retrieve predicted ORs and support intervals. An R package is available for association testing using *ClusterBurden* (https://github.com/adamwaring/ClusterBurden).

4. Discussion

We have developed two new analytic methods; *ClusterBurden* and *GAMs,* which incorporate information on amino-acid residue position to examine the pathogenic potential of rare coding variants in Mendelian disease genes. We apply the methods to gene panel data from HCM patients to illustrate the applications of this approach for rare variant interpretation, and for pathogenic gene discovery.

4.1 ClusterBurden

ClusterBurden is a gene association test with superior power over a standalone burden test in situations where rare pathogenic variants cluster in specific protein regions. *ClusterBurden* was devised to be suitable for scanning large-scale whole-exome sequencing projects designed to identify novel pathogenic genes for rare Mendelian diseases. The combination of *FE* and *BIN-test* to model clustering and burden minimizes the computation overhead required to calculate p-values making analyses of >20,000 genes [40] in hundreds of thousands of cases and controls, practical in terms of execution time and computer memory requirements. For genes with a significant *BIN-test* 'cluster' p-value, *ClusterBurden* calculated considerably lower p-values that the traditional Fisher's exact 'burden' test, implying an increase in statistical power (Table 1). Although *ClusterBurden* has reduced power whenever clustering is absent, we observe clustering for the majority of well-established HCM genes where missense variants cause disease, so expect that the method will often be more powerful to detect novel Mendelian genes than a burden-only test.

4.2 GAMs

We show that generalized additive models can be informative to assess pathogenicity of rare coding variants based on our study of several well-established HCM genes. We report strikingly different predicted ORs depending on where in the linear protein sequence a variant falls. For 5 out of 6 core sarcomeric genes; *MYH7*, *TNNT2*, *TNNI3*, *MYL2* and *MYL3*, variant residue position relative to gnomAD population controls was the best predictor of case or control status. For *MYBPC3,* several bioinformatics annotation features in dbNSFP had improved predictive performance over residue position.

GAMs have attractive statistical properties that are not necessarily shared by other machine-learning approaches, in that they produce familiar interpretable results via variant-specific ORs and quantify uncertainty in estimates by 95% confidence intervals. Unlike empirical ORs that are based on the observed case-control frequencies of the given variant in isolation, GAM ORs draw upon a much larger pool of information, including *in silico* prediction scores and features of other variants in close proximity. These estimates implicitly combine information on both pathogenicity and penetrance, have much tighter confidence intervals (even when observed counts are very low), and could be calculated for novel singleton or hypothetical variants.

Model predictions were positively correlated with classifications made by experts at OMGL. Discordant predictions highlight variants with potential for reclassification. GAM predictions therefore offer a useful metric for stratification of rare missense variants and suggest a natural ranking to variants that can be considered for reclassification as (likely) pathogenic or (likely) benign. The GAMs developed here overlap with two ACMG criteria; PM1 and PP3. The *posGAM* model represents a quantitative data-driven approach to applying criteria PM1. Under the ACMG guidelines, classifications are made when a specific number of criteria have been met at different levels (e.g. one strong and two moderate for a pathogenic classification). Therefore there is potential to lose information when multiple rules are captured in a single model i.e. *fullGAM*. However, in most cases, the addition of supporting evidence from PP3 will not impact classification. It is therefore possible in these circumstances, to combine signals PM1 and PP3 in the *fullGAM* to use available information as efficiently as possible.

For genes burdened by clustered variants, GAM improved performance over other commonly used *in silico* predictors that are not optimized on *gene-specific* data. Unlike most variant prediction algorithms, including all those available in dbNSFP, where models are trained using variant pathogenicity as the response label, the GAM response variable is case status and not variant pathogenicity. This is necessary to ascribe ORs that simultaneously represent pathogenicity potential *and* penetrance. However, this does impose a limit on the maximum predictive accuracy for each model, as this is heavily influenced by incomplete penetrance. However, the relative increase in AUC when using this gene-specific approach, compared to generic *in silico* predictors was substantial.

4.3 Significant position signals in the HCM gene panel

After *popmax* filtering at 0.0001, we detected Bonferroni significant clustering of variants in six genes in HCM cases compared to the gnomAD reference controls; *MYH7, MYBPC3*, *TNNT2*, *TNNI3*, *MYL2* and *MYL3* (Figure 1). The strongest position signal was observed in the beta myosin heavy chain protein (*MYH7:* ENST00000355349), a finding that has been long recognised [41-42]. The highest variant density is observed in residues 100-900 that overlaps with the myosin-head motor domain, two peak densities in this region centre on residues 370 and 830 (Figure 2). The relatively low variant density in cases and high density in controls in the carboxy-terminus of this protein might lead an observer to hypothesise a regional protective effect on HCM risk (S13 Figure). In sharp contrast, the GAM model predicts a modestly excessive burden (OR ~3) across this entire region discounting the likelihood of a localised protective effect (Figure 2).

A strong position signal, driven by four potential clusters, was observed in the *MYBPC3* gene (ENST00000545968), which encodes cardiac myosin-binding protein C C10 (Figure 2). Clusters peaked at residues 260, 518, 864 and 1274, which respectively fall in domains C1, C3, C7 and C10. Multiple functional roles are suspected for the region containing the C1 domain, including binding to myosin S2 and actin. The C10 domain is also a possible titin binding site [43]. To explore whether the signal was overly driven by founder mutations found at high frequencies in our cardiomyopathy cases, seven variants with allele counts above 10; p.Arg810His, p.Asp770Asn, p.Glu542Gln, p.Arg502Trp, p.Arg495Gln, p.Glu258Lys and p.Val219Leu were masked in a sensitivity analysis. In their absence, there is still strong evidence of a position signal ($p < 3 \times 10^{-9}$) and the remaining peak densities overlap with the locations of the (masked) founder mutations (S14 Figure).

The majority of variants, 24 (89%) in the *TNNT2* gene (ENST00000509001), which encodes cardiac troponin T, map to clusters between residues 67-179 and residues 250-282 (Figure 2). The first peak at residue 90 overlies a previously reported region at residues 79-179 that binds tropomyosin [44-45]. Mutations between residues 92-110 have been previously noted to impair tropomyosin dependent functions in *TNNT2* [46] and 6 of 27 variants map to this region (p.Ala104Val, p.Lys97Asn, p.Arg94His, p.Arg94Leu, p.Arg92Gln and p.Arg92Trp). In *TNNI3* (ENST00000344887), which encodes cardiac troponin I, 31 (91%) of variants mapped to a cluster spanning residues 128-209. This accords with previous studies documenting disease-causing variant clustering in the carboxy-terminus of this sarcomeric protein [47]. In *MYL2* (ENST00000228841), which encodes myosin regulatory light chain, 15 variants cluster between residues 25 and 100, whereas control variants tended to cluster towards the C-terminus (Figure 2). In *MYL3* (ENST00000395869), which encodes myosin essential light chain, 11 out of 14 variants cluster between residues 125 and 175 whereas control variants were more uniformly distributed.

5. Conclusion

As the GAM modelling framework is a data-dependent approach to pathogenicity interpretation, increasing the size of the training dataset should increase the accuracy and confidence of model predictions for HCM. This is especially relevant if this model is to be integrated within the ACMG classification pipeline. With large datasets to drive these models, criteria PM1 and PP3 can be applied to confidently to variants observed in the clinic. Furthermore, this modelling approach has general application to other Mendelian diseases with sufficiently large case cohorts for a data-driven modelling approach.

Our present analytic methods assume an autosomal dominant genetic model. With sample-level information to distinguish homozygotes and compound heterozygotes, it is conceivable to extend *ClusterBurden* and the GAM methods to analyse a recessively inherited disease by judicious choices of indicator-variable coding. The GAM approach could also be extended from the 1-dimensional linear protein sequence to 3-dimensional protein structures, by including smoothed linear variables to model x, y and z protein coordinates. This is potentially a more informative way to model variant clustering; however it is limited by the availability of complete high-resolution 3-dimensional structures. For the HCM genes we

examined, no suitably complete structures were available; as more structures are solved and possibly for other Mendelian diseases, a 3-D GAM analysis might offer further improvements in variant interpretation.

In conclusion, with the assembly of large patient and population control datasets to quantify mutation clustering, missense residue position is an important feature to consider in analyses of rare-variants in Mendelian diseases. The *ClusterBurden* and *GAM* methods have the potential to improve power to detect novel pathogenic genes and probe in detail the genetic architecture of risk variants, analyses that could improve interpretation of genetic testing to provide more reliable information to families and patients.

Funding

Wellcome Trust doctoral studentship (203834/Z/16/Z) to AJW, MRC doctoral studentship to ARH, Welcome Trust core award (203141/Z/16/Z, MF, HW), the Oxford BHF Centre of Research Excellence (RE/13/1/30181, MF, HW), HW has received support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. CK, SN and HW received support from a National Heart, Lung, and Blood Institute [grant U01HL117006-01A1].

Acknowledgements

We would like to acknowledge Anuj Goel for his bioinformatics support in data curation and Michael Bowman for his support in accessing data from the clinical genetics laboratory.

Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflict of Interest – none declared

REFERENCES

1. 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. 1994; 91:9622–9625
2. Robertson SP, Twigg SR, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, Kenwrick SJ et al. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. 2003; 33:487–491
3. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics. 2008; 40:395–402
4. 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
5. Fine DM, Wasser WG, Estrella MM, Atta MG, Kuperman M, Shemer R et al. APOL1 Risk Variants Predict Histopathology and Progression to ESRD in HIV-Related Kidney Disease. J Am Soc Nephrol. 2012; 23:343–350
6. Nicolas G, Charbonnier C2, Wallon D, Quenez O, Bellenguez C, Grenier-Boley B et al. SORL1 rare variants : a major risk factor for familial early-onset Alzheimer's disease. Mol Psychiatry. 2016; 21:831–836
7. Schneppenheim R, Michiels JJ, Obser T, Oyen F, Pieconka A, Schneppenheim S, Will K, Zieger B, Budde U. A cluster of mutations in the D3 domain of von Willebrand factor correlates with a distinct subgroup of von Willebrand disease : type 2A / IIE. Blood. 2018; 115:4894–4902.
8. Lelieveld SH, Wiel L, Venselaar H, Pfundt R, Vriend G, Veltman JA, Brunner HG, Vissers LELM, Gilissen C. Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. Am J Hum Genet. 2017; 101:478–484
9. 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: e0179364
10. Collins FS. Positional cloning moves from perditional to traditional. Nat Genet. 1995; 9:347-50
11. Walsh R, Thompson K, Ware J, Funke B, Woodley J, McGuire K et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genetic in Medicine. 2017; 19:192-203.
12. Guo MH, Plummer L, Chan Y-M, Hirschorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data Am J Hum Genet. 2018; 103:522-34
13. 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
14. Watkins H, Ashrafian H, Redwood C. Inherited cardiomyopathies. N Engl J Med. 2011; 364:1643-56
15. Maron BJ, Olivotto I, Spirito P, Casey SA, Bellone P, Gohman TE, Graham KJ, Burton DA, Cecchi F. Epidemiology of hypertrophic cardiomyopathy-related death: revisited in a large non-referral-based patient population. Circulation. 2000; 102:858–64
16. 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
17. Kelly MA, Caleshu C, Morales A, Buchan J, Wolf Z, Harrison SM et al. Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. Genet Med. 2018; 20:351-359

18. 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
19. 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
20. 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
21. Romanet P, Odou M, North M, Saveanu A, Coppin L, Pasmant E, Mohamed A, Goudet P, Borson-Chazot F, Calender A, Béroud C, Lévy N, Giraud S, Barlier A. Proposition of adjustments to the ACMG-AMP framework for the interpretation of MEN1 missense variants. Hum. Mutat. 2019; 40:661-674
22. Lek et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291
23. Ludbrook J. Analysis of 2 × 2 tables of frequencies: matching test to experimental design. Int J Epidemiol. 2008; 37:1430-1435
24. Mann H, Wald A. On the choice of the number and width of classes for the chi-square test of goodness of fit. Ann Math Stat. 1942; 13:306–317
25. Anderson TW, Darling DA. "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics. 1952; 23:193–212
26. Kolmogorov AN. Sulla Determinazione Empirica di Una Legge di Distribuzione. Giornale dell'Istituto Italiano degli Attuari. 1933; 4:83-91
27. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd. 1925.
28. Spearman CE. The proof and measurement of association between two things. Am J Psychol. 1904; 15:72–101
29. Hastie T, Tibshirani R. Generalized Additive Models. Chapman & Hall, London. 1990
30. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J Roy Stat Soc. 2011; 73:3-36
31. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protocols. 2016; 11: 1-9
32. Liu X, Jian X, and Boerwinkle E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Hum Mut. 2011; 32:894-9
33. Lin W. Association Testing of Clustered Rare Causal Variants in Case-Control Studies. PLoS One. 2014; 9: e94337
34. Neale B, Rivas M, Voight B, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell S, Roeder K, Daly K. Testing for an Unusual Distribution of Rare Variants. PLoS Genet. 2011; 7: e1001322
35. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011; 89:82–93
36. Madsen, BE, Browning, SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet. 2009; 5: e1000384
37. Davison, A, & Hinkley, D. Bootstrap Methods and their Application (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. 1997
38. Kramer, C. M. et al. Hypertrophic Cardiomyopathy Registry: The rationale and design of an international, observational study of hypertrophic cardiomyopathy. Am. Heart J. 2015; 170, 223–230
39. Haldane JB. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet. 1956; 20:309-11

40. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey A, Salzberg SL. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018; 19:208
41. 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
42. Blair E, Redwood C, Oliveria M, Moolman-Smook JC, Brink P, Corfield VA, Ostman-Smith I, Watkins H. Mutations of the Light Meromyosin Domain of the β-Myosin Heavy Chain Rod in Hypertrophic Cardiomyopathy. Circ Res. 2002; 90:263–9
43. 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
44. Pearlstone JR, Smillie LB. The binding site of skeletal alpha-tropomyosin on troponin-T. Can. J. Biochem. 1977; 55:1032–8
45. Heeley DH, Golosinska K, Smillie LB. The effects of troponin T fragments T1 and T2 on the binding of nonpolymerizable tropomyosin to F-actin in the presence and absence of troponin I and troponin C. J. Biol. Chem. 1987; 262:9971–8
46. 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
47. 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.

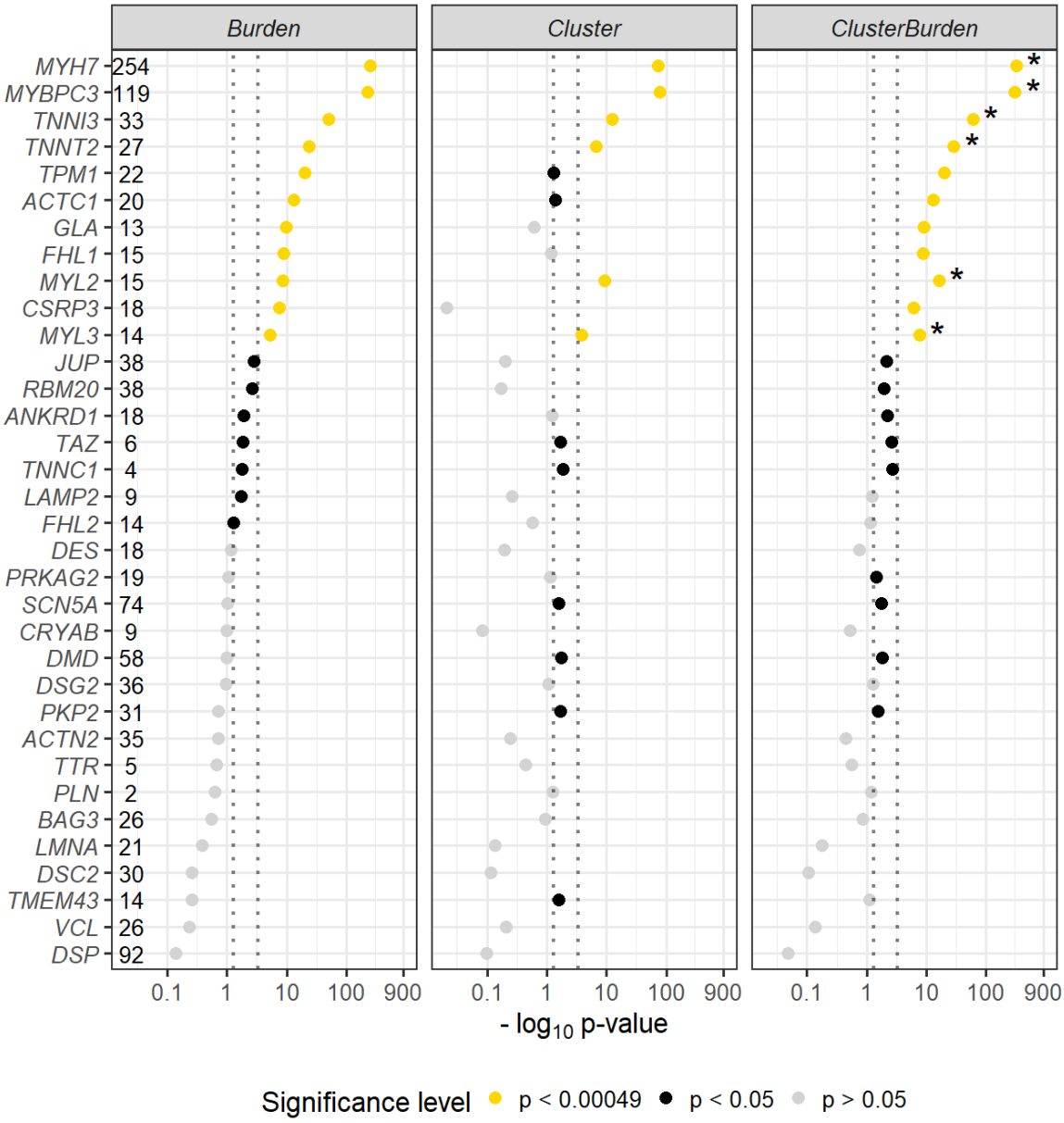

Figure 1: Association analysis with Fisher's-exact test (Burden), *BIN*-test (Cluster) and *ClusterBurden* of 34 cardiomyopathy genes.

Our case-control dataset contains 5,338 hypertrophic cardiomyopathy cases and 125,748 gnomAD controls. For all tests only missense variants with a *popmax* MAF less than 0.01% were considered. P-values are presented on a -$\log_{10}$ scale. The number of observed case variants in each gene is displayed next to the gene symbol. P-values displayed in yellow are significant after Bonferroni correction for 34 genes x 3 tests ($p < 0.00049$), p-values in black are nominally significant ($p < 0.05$) and p-values in grey are insignificant ($p > 0.05$). Asterix's denote genes where the *ClusterBurden* p-value is lower than the Burden p-value. Two vertical dotted lines at 0.05 and 0.00049 indicate the nominal and Bonferroni significance thresholds.

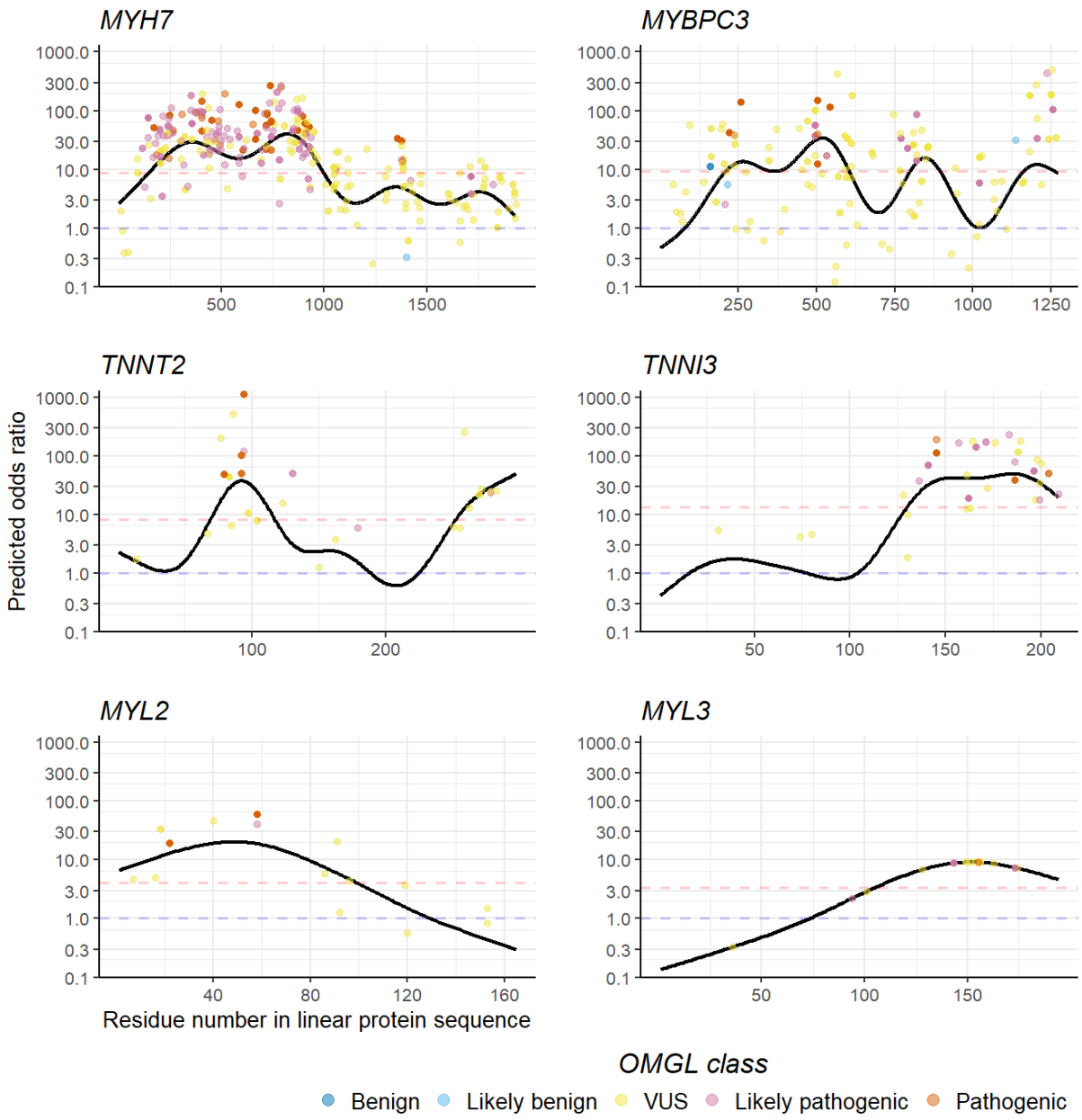

Figure 2: Risk predictions generated by GAM for rare-missense variants in *MYBPC3, MYH7, TNNI3, TNNT2, MYL2* and *MYL3*.

Each point denotes a rare-variant in the HCM dataset and is coloured to indicate its expert Oxford Medical Genetics Laboratory classification. ORs on the y-axis are displayed on a $\log_{10}$ scale and were derived from *full*GAM models including gene-burden, residue position and gene-specific significant secondary features from dbNSFP. The solid black curvy lines represent the predictions for each residue in the protein for a gene-burden and position model (*posGAM*). Dashed red lines indicate an OR of 1, dashed blue lines indicates the OR for the uniform burden model.

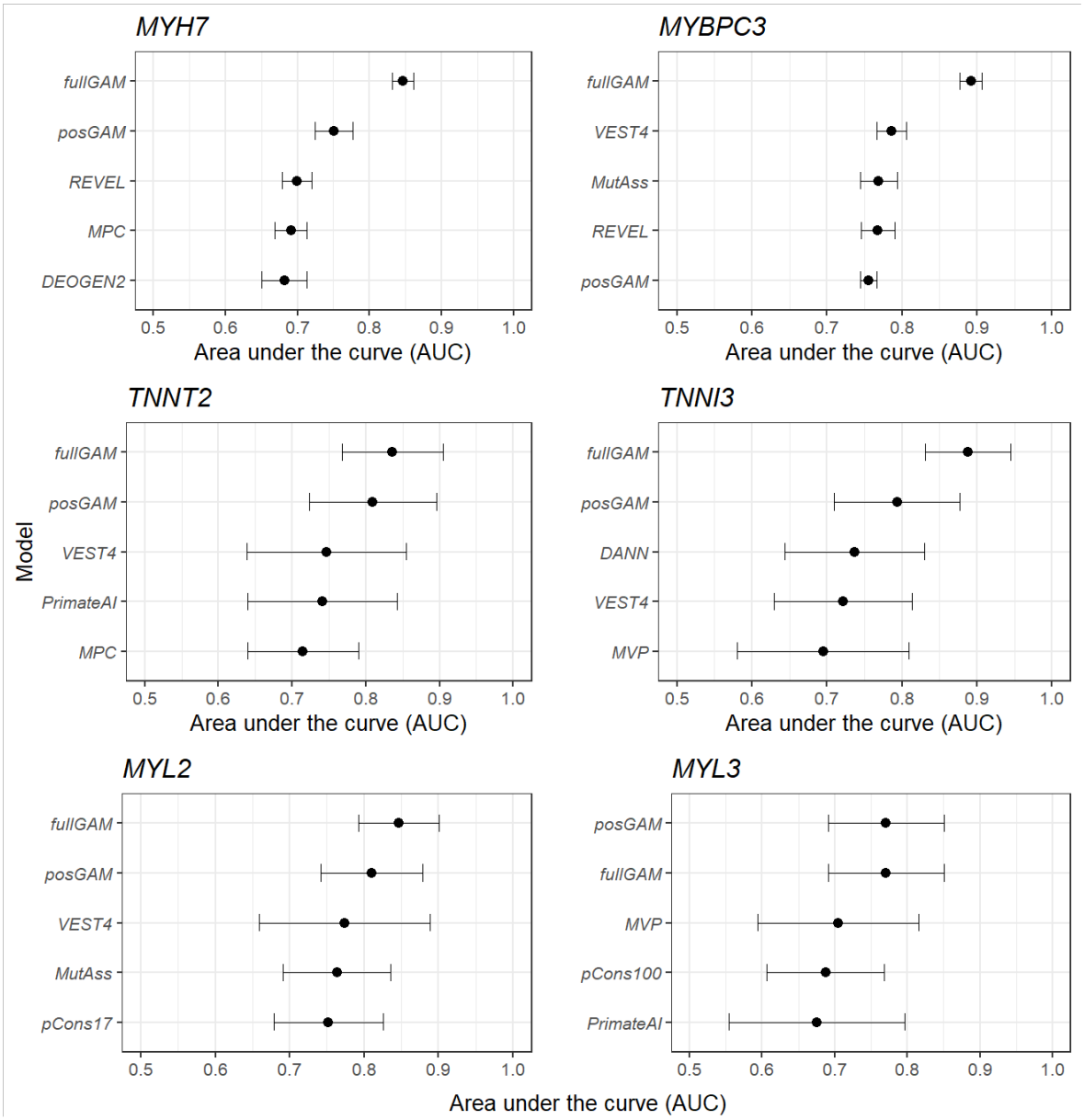

Figure 3: Mean and standard deviations of area under the curve (AUC) metrics across different models after 10 cross-fold validations in our HCM-gnomAD dataset.

For each gene, the *full*GAM and *pos*GAM models are compared to each individual *in silico* predictor from dbNSFP. For each gene, only the five highest mean AUC scoring models are displayed.

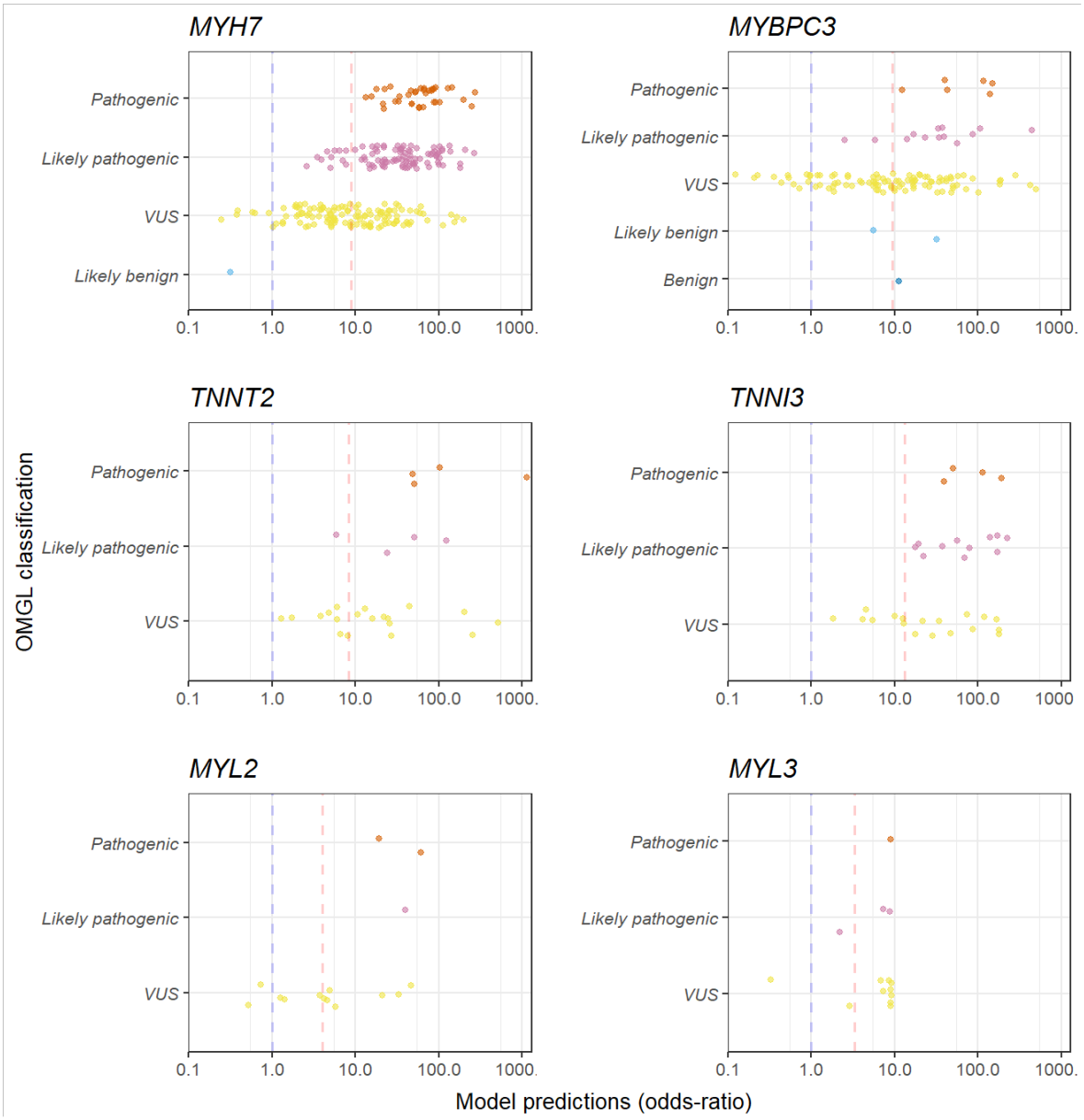

Figure 4: Model predictions stratified by their expert classifications for HCM variants in *MYH7*, *MYBPC3*, *TNNT2*, *TNNI3*, *MYL2* and *MYL3*.

Classifications for each variant were manually curated by the Oxford Medical Genetics Laboratory. Risk predictions generated by *full*GAM models are displayed on the x-axis on a $\log_{10}$ scale. Only variants with a *popmax* frequency of less than 0.01% were considered, excluding most variants with a benign or likely benign classification.

Supporting information captions

S1 Table: Implementation details for all statistical tests used for type 1 error and power calculations in this study.

S1 Methods: Description of the forward-time simulation algorithm used to simulate rare-clustered variants.

S1-6 Figures: ORs from the *posGAM* model for all amimo acid residues in *MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3.*

Each plot shows predictions (odds-ratios) and 95% confidence intervals, for each possible residue in the protein, where residue position and carrier status (gene burden) are the only model predictors. The GAM is trained on cardiomyopathy cases (n=5,338) and gnomAD controls (n=125,748). The dashed red line indicates an odds-ratio of 1 and the blue dashed line indicates the odds-ratio for the uniform burden model (i.e. gene odds-ratio).

S7-12 Figures: ORs from the *fullGAM* model for our observed HCM variants in *MYH7, MYBPC3, TNNT2, TNNI3, MYL2, MYL3.*

Each plot shows predictions (odds-ratios) generated by GAM, trained on cardiomyopathy cases (n=5,338) and gnomAD controls (n=125,748). Each point is a variant in our hypertrophic cardiomyopathy dataset, is coloured by its expert classification (made by the Oxford Medical Genetics Laboratory), and is accompanied by a 95% confidence interval bar. The dashed red line indicates an odds-ratio of 1, the blue dashed line indicates the odds-ratio for uniform burden, and the solid black line is the marginal odds-ratio for a model of only amino-acid residue number.

S13 Figure: Distribution and risk predictions for rare-missense *MYH7* variants in our case-control cohort.

The variant positions and training data for the GAM are a case cohort of 5,338 hypertrophic cardiomyopathy cases and 125,748 gnomAD controls. The density plot (lower panel) may give the impression that there is an excess of control variants in the C-terminus of the *MYH7* protein; however the GAM model (upper panel) resolves this potential misinterpretation and clearly shows an odds-ratio greater than 1 for the entire protein.

S14 Figure: Rare-missense variant clustering in *MYBPC3* with and without potential founder mutations.

Variant clustering model (*posGAM*) are generated for three different frequency filtering strategies. The model identifies four discrete regions with high pathogenic potential regardless of whether the founder mutations are included in the analysis. However, the magnitude of the predicted ORs are higher under normal filtering conditions (e.g. *popmax* < 0.01%).