bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data-driven modelling of mutational hotspots and *in-silico* predictors in hypertrophic cardiomyopathy

Waring A.J.[1], Harper A.R.[1,2], Salatino S.[1], Kramer C.M.[4], Neubauer S[2], 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
Wellcome Trust Centre for Human Genetics, Roosevelt Dr, Headington, Oxford OX3 7BN
+44 (0)1865 287601
martin.farrall@cardiov.ox.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Background

Although rare-missense variants in Mendelian disease-genes have been noted to cluster in specific regions of proteins, it is not clear how to consider this information when evaluating the pathogenicity of a gene or variant. Here we introduce methods for gene-association and variant-interpretation that utilise this powerful signal.

Methods

We present a case-control rare-variant association test, *ClusterBurden*, that combines information on both variant-burden and variant-clustering. We then introduce a data-driven modelling framework to estimate mutational hotspots in genes with missense variant-clustering and integrate further *in-silico* predictors into the models.

Results

We show that *ClusterBurden* can increase statistical power to scan for putative disease-genes, driven by missense variants, in simulated data and a 34-gene panel dataset of 5,338 cases of hypertrophic cardiomyopathy. We demonstrate that data-driven models can allow quantitative application of the ACMG criteria PM1 and PP3, to resolve a wide range of pathogenicity potential amongst variants of uncertain significance. A web application (*Pathogenicity_by_Position)* is accessible for missense variant risk prediction of six sarcomeric genes and an R package is available for association testing using *ClusterBurden.*

Conclusion

The inclusion of missense residue position enhances the power of disease-gene association and improves rare-variant pathogenicity interpretation.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

The clustering of pathogenic missense variants in specific regions or domains of proteins has been frequently reported [1-5]. A plausible mechanism underpinning this phenomenon is the presence of multiple loss or gain-of-function variants within functionally important domains [6]. Despite numerous examples of variant clustering, there have been few attempts to explicitly model variant residue position as a predictor of pathogenicity [7].

Pathogenic genes for Mendelian diseases were historically identified by linkage and candidate gene studies in multiple affected families [8]. Scanning of large exome patient cohorts offers 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 test to confirm the pathogenicity of candidate genes [9], as well as identify novel putative pathogenic genes [10]. However, for genes with non-uniform variant pathogenicity, including positional information, alongside burden, may provide an increase in power to detect undiscovered low-penetrance genes.

The American College of Medical Genetics and Genomics (ACMG) have produced guidelines to interpret variant pathogenicity [11]. These guidelines integrate diverse data and classify variants into five categories from benign to pathogenic. However, due to limited information available for many variants, they fall into the category; 'variant of uncertain significance' (VUS). Although positional information is covered by criteria PM1, there is a lack of evidence for mutational hotspots, resulting in underapplication of this criteria. Furthermore, although much work has gone into the development of *in-silico* prediction scores, alternative scores can be conflicting, leading to discordance amongst testing laboratories [12] and uncertainty in their application. However, wherever large patient cohorts are attainable, mutational hotspots and the uncertainty surrounding *in-silico* predictors can be directly estimated from the data.

Hypertrophic cardiomyopathy (HCM), a relatively common autosomal dominant disease (1 in 500 prevalence), is a major cause of heart disease in people of all ages [13] and a cause of sudden cardiac

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

death. In our cohort, eight sarcomeric genes collectively provide firm molecular diagnoses for ~27% of HCM patients, with a further ~13% of patients carry a VUS in the same genes. It has been suggested that disease and gene-specific approaches are needed to improve interpretation [14] and guidelines have been produced for specific genes and/or disease areas [15-18]. HCM is common enough to provide the large datasets needed for these gene-specific and data-driven approaches.

Here we propose new statistical approaches to explicitly include variant residue position in rare-missense variant association and interpretation; *ClusterBurden* for association testing and generalized additive models (*GAMs*) for hotspot estimation and application of *in-silico* prediction algorithms. We apply these methods to a large cohort of 5,338 HCM patients and up to 125,748 gnomAD [19] population controls. We demonstrate that utilising positional information increases power to detect disease-gene associations and elucidate the clustering signals present in 34 cardiomyopathy genes. We then use *GAMs* to model residue position and *in-silico* predictors for six core sarcomeric genes.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

Patient cohorts and simulated data

Next-generation sequence data for 34 cardiomyopathy genes (S1 Table) were available from two large HCM cohorts (S1A Methods); 2,757 probands referred to the Oxford Medical Genetics Laboratory (OMGL) for genetic testing and 2,636 probands recruited to the HCMR project [20]. High-coverage exonic sequences were captured by target enrichment and sequenced on the MiSeq platform (Illumina Inc.). Joint bioinformatic processing of both datasets followed the Genome Analysis ToolKit version 4 best practice guidelines (S1B Methods). OMGL variants were confirmed by Sanger sequencing and HCMR variants were manually checked by inspection of BAM files.

The gnomAD population reference database was used as a control group, which includes variant frequency data based on up to 125,748 individuals. For both cases and controls, only missense variants with a gnomAD population maximum allele frequency of less than 0.0001 [9-10] were included. This excludes potentially common ancestry-specific variants that are unlikely to be pathogenic for HCM.

To determine the theoretical performance of *ClusterBurden*, synthetic data were generated using a forward-time simulator (S2 Methods) designed to imitate rare-variants in genes with discrete exonic regions of increased pathogenic potential. Six different scenarios were considered, combining three clustering scenarios (uniform, a single pathogenic cluster and multiple pathogenic clusters) and two protein lengths (500 and 1,000 amino acid residues). For each scenario, 10,000 synthetic datasets were generated with 5,000 cases and 125,000 controls, variants were filtered at a simulated control MAF of <0.0001.

Detecting missense variant burden and clustering – ClusterBurden

Missense variants causing amino acid substitutions at specific positions in the linear protein sequence were numbered from 1 to *N*, where *N* is the length of the protein. In all positional analyses, we consider this number as the position of a missense variant. As the background distribution of variants

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

may be non-uniform, a cluster of variants observed in affected cases is insufficient in isolation to determine relevance to disease. Therefore to detect disease-relevant clustering, distributions were compared between affected cases and unaffected controls.

We propose *BIN-test* to evaluate these distributional differences. First, the protein's linear sequence of amino-acid residues is split into *k* bins of equal length for both cases and controls. A chi-squared two-sample test is applied to the resultant *k* x 2 contingency table of binned variant counts. In simple terms, the test assumes that for each bin, the relative frequency of observed variants in cases and controls is the same and is more significant depending on how many bins deviate from this expectation and by how much. We applied a $k \sim n^{2/5}$ heuristic [21] to select the optimal number of bins (*k*) dependent on *n*, the total number of observed variants. We compared the performance of the *BIN-test* with two other tests that compare distributions between two-samples; Anderson-Darling (AD) [22] and Kolmogorov-Smirnov (KS) [23]. 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 0.05 and ***n*** is the number of simulations [24].

Current methods to discover novel Mendelian disease genes focus on the burden of rare-variants in an affected cohort relative to controls. We propose an approach, *ClusterBurden,* which tests the joint hypothesis of an excess of rare missense variants *and* differential clustering, in case-control data. This was accomplished by combining the p-values from a burden test (Fisher's-exact test) with the *BIN-test*. As there are no known examples of a protective burden of rare exonic variants in cardiomyopathy, here we only consider an excess burden in the case group making it a one-sided test. Fisher's method [25] was then used to calculate the joint significance of the combined p-values. An important assumption of this method is that the contributing p-values are independent; this was assessed in simulated data by Spearman's rank correlation test [26]. The performance of *ClusterBurden* was compared, in simulated data, to two published position-informed association methods: DoEstRare [7] and CLUSTER [27], and three position-uninformed methods: C-alpha [28],

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SKAT [29], WST [30]. Then, both *BIN-test* and *ClusterBurden* were then applied to the 34 cardiomyopathy gene-panel in our HCM-gnomAD case-control cohort.

Hotspot estimation and *in-silico* predictor modelling using GAMs

To test the hypothesis that a variant's position can improve pathogenicity interpretation, we considered gene-specific models of variant clustering in cases and controls. By combining information on gene-level burden and variant positions, these data-driven models estimate the regional burden across the linear protein sequence to quantify mutational hotspots. The models were fitted in the GAM framework, [31] implemented in the R package "mgcv" [32]. The outcome variable was disease status, so each model was unsupervised with respect to previous classifications of pathogenicity. The training data included all rare-missense variants in cases and controls, including known pathogenic variants in the control set. Therefore, this approach implicitly models incomplete penetrance and benign background variation, leading to unbiased estimates of variant odds-ratios.

GAMs, as an extension of the linear modelling framework, are designed to deal with non-linear relationships of unknown complexity, between explanatory variables (e.g. residue position) and the response variable (e.g. case-control status). When a relationship is potentially non-linear, it is represented by a smooth curve instead of a straight line. These curves are inferred automatically using restricted maximum likelihood, which reduce over-fitting by penalising excessive 'wiggliness'.

Using this framework, we defined the structure of the *hotspot-model*, which models; carrier status (gene-level burden) and residue position (clustering). To incorporate gene-level burden, non-carriers must also be modelled. However, as variant-level features such as residue position are meaningless for non-carriers, so a nested model structure is required, whereby residue position is included *only* as an interaction with carrier status. Under these circumstances, the smoothed residue position term is multiplied by zero for non-carriers, excluding this undefined data from the model. The structure of the *hotspot-model* is as follows:

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$P = \beta_0 + \beta_1\, carrier_status + s_1(residue_position, \text{by} = carrier_status) + \boldsymbol{\varepsilon}$$

where **P** is the probability of being a case, $\boldsymbol{\beta_0}$ is the model intercept, $\boldsymbol{\beta_1}$ is a linear coefficient for *carrier_status*, $\mathbf{s_1}$ is a smoothed (i.e. non-linear) function for *residue_position*, **by** is used to generate factor-smooth interactions and **ε** is a binomial error term.

The feasibility of this approach is dependent on the number of observations, thereby limiting its application in our data to the six core sarcomeric genes; *MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *TNNI3* and *TNNT2,* each carrying at least 20 rare-missense variants. A *hotspot-model* was produced for each gene and raw model predictions for each residue position, in the form of logistic probabilities and standard errors, were transformed to ORs and 95% confidence intervals (CIs). There is currently no universal guidance on how to quantitatively apply ACMG criteria PM1. However, using the probability that a variant is a case variant as a proxy of pathogenicity, we can use predicted probabilities to attribute levels of evidence. Here we stratify variants based on the probability thresholds 0.9, 0.95 and 0.99 to represent supporting, moderate and strong evidence of pathogenicity. These correspond to ORs of approximately 10, 20 and 100.

As GAMs are additive in structure, it is straight-forward to include further predictors in the model. Here we experimented with the inclusion of variant prediction scores extracted from the dbNSFP4.0 database [33]. These *in-silico* prediction algorithms are covered by the ACMG criteria PP3, however, like criteria PM1, it is challenging to apply this criteria quantitatively. It is unclear which threshold determines a pathogenic variant with a given probability and whether these thresholds are consistent across genes. Both of these problems can be solved using gene-specific models as the relationship between the *in-silico* predictor and disease-status is automatically inferred. Furthermore, we can quantify uncertainty on the usage of these scores for each gene.

To avoid overfitting, we implemented a strict two-stage feature selection procedure, based on 24 *in-silico* scores from dbNSFP. 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, whereby

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

features are removed from the model one at a time and then remaining features are assessed again in a new model, until all features are significant (Bonferroni corrected for the number of features selected in stage 1). The resulting models, which are henceforth termed *in-silico-models*, assimilate evidence of gene-burden, variant clustering and pathogenicity prediction scores.

Model performance for these GAMs are best judged by the estimates of uncertainty accompanying predictions. However, to determine the relative ability of these models to predict pathogenicity, they were compared with models based on single *in-silico* predictors and expert variant classifications. Relative performance was assessed using the area under the curve (AUC) from the receiver operator characteristic (ROC) curves with ten cross-fold validations performed by dividing the data into training and test sets in the ratio of 80:20%.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

Testing the hypothesis of clustered missense variants

Under the null hypothesis of no excess burden or clustering, the false-positive rate of the *BIN-test* and AD test were adequately controlled in simulated data, whereas the KS test was overly conservative (S3 Table). The *BIN-test* had superior power than AD or KS under all clustering scenarios and protein lengths with on average 1.8-fold more power to detect clustering. As power covaries with the number of observed variants, power was higher for longer proteins as well as proteins with more pathogenic residues.

Correlations between p-values generated by the *BIN-test* and Fisher's exact test were compared for simulated data under 1) a null model of no association or 2) a disease model of over-burdened and clustered variants. For the disease model, there was a positive correlation (Spearman's rank correlation rho=0.40) between p-values, as anticipated as the power of these tests covaries with the number of observed variants. However, under the null model, the p-values were completely uncorrelated (i.e. rho = 0.00), satisfying the independence assumption of Fisher's method.

The false-positive rate for *ClusterBurden*, DoEstRare, CLUSTER and C-alpha were all well controlled in the simulated datasets (S3 Table). On the contrary, SKAT and WST showed markedly inflated false-positive rates under the null and were not examined further. *ClusterBurden* was the most powerful method when clustering was present with an average of 72% power, 3% higher than the second-best test DoEstRare. The best method under the uniform model (i.e. burden-only) was CLUSTER, which had ~5% more power than *ClusterBurden*. Amongst the position-informed tests, *ClusterBurden* was the most rapid to compute taking less than a second per gene whereas DoEstRare and CLUSTER took over 20 or 4 minutes respectively.

We then examined 34 cardiomyopathy genes for rare-missense variant associations with Fisher's-exact test (burden), *BIN-test* (cluster) and *ClusterBurden* (combined cluster and burden) in our cohorts

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of HCM cases and gnomAD controls (Figure 1). Significance thresholds were conservatively 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 then detected in 11 genes with Fisher's-exact test; *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 clustering 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 sarcomeric genes showed nominal evidence of clustering; *ACTC1* ($p<0.0412$) and *TPM1* ($p< 0.0494$). *ClusterBurden* confirmed the association for 11 genes that showed burden signals and calculated substantially lower p-values for the six core-sarcomeric genes with significant clustering, consistent with enhanced power for this approach in true disease-causing genes.

Hotspot and in-silico models

Figure 2 summarises the *GAM*-predictions for six sarcomeric genes in the *hotspot-models*. Visualising the predicted odds-ratios (OR) for each residue illuminates the local burden of rare-missense variants across each protein, identifying "mutational-hotspots" and highlighting areas of potential functional importance in HCM pathogenesis. Confidence in these predictions is tight for *MYH7* and *MYBP3,* conversely the genes with fewer observed variants have much broader confidence intervals. ORs from all models correlate strongly with expert manually assigned classifications, though there is substantial overlap between classes. Variants with a VUS classification show the highest spread in predictions.

Figure 3 displays *in-silico-model* predictions for individual variants in the same six genes. Due to strict feature selection, the number of predictors included in each model depends on the power to detect associations between features and disease status. This resulted in fewer features for genes with fewer observed variants, *MYL3* had no additional significant features. As residue position is included as a

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

predictor in each model, predictions generally follow the *hotspot-model*, however, due to additional *in-silico* predictors, ORs tend to vary from this pattern, stratifying risk for variants at the same position.

As with the *hotspot-models* there was strong correspondence between predictions and expert classifications (mean rho 0.41 across six models). In *MYH7*, mean predicted ORs for pathogenic, likely pathogenic and VUS variants observed in cases, were 74, 50 and 20 respectively. Again, the VUS class had highest heterogeneity, with predicted ORs ranging from 0.25 to 197 (*MYH7*). For half of these VUS's, limited information is available, as they are observed in a single case and absent in controls (singleton variants). The empirical ORs for these singletons, based only on case and control frequencies and adding 0.5 to zero-count cells (Haldane continuity correction [34]), had very wide 95% Cis: 44.9 [1.5, 1338.3]. However, *predicted* ORs for such variants can have greater precision with different point estimates depending on the precise amino-acid substitution. In MYH7, five singleton VUS's had predicted ORs greater than 100 and three had ORs less than 1.

The mean and standard deviation of AUC for ten cross-fold validations summarise overall model performance (Figure 4). The *in-silico-model* had a much higher mean AUC than any individual *in-silico* predictor in isolation. With the exception of *MYBPC3*, the *hotspot-model* out-performed any standalone score from dbNSFP. This suggests that residue position is more important in determining pathogenicity than the *in-silico* predictors in dbNSFP for these sarcomeric proteins. The AUC standard deviations for *MYH7* and *MYBPC3* were the smallest, suggesting they have the highest capacity to generalize to new variants.

The GAMs were then used to attribute evidence of pathogenicity based on the ACMG criteria PM1 and PP3. Using the ACMG OR thresholds described in the methods, Table 1 shows the proportion of variants in our cohort with evidence of pathogenicity predicted by the *hotspot* and *in-silico-models*. For the *hotspot-model*, the PM1 criteria was satisfied, with supporting (OR > 10) or moderate (OR > 20) evidence, for some variants in all genes except *MYL3.* Strong evidence of pathogenicity (OR > 100) was not predicted by the *hotspot-model* for any variant. Conversely, the *in-silico-model*, which

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

combined criteria PM1 and PP3, provided strong evidence of pathogenicity for many variants, including VUS's in *MYH7, MYBPC3, TNNT2* and *TNNI3*.

Table 1: Proportion of variants with evidence of pathogenicity in *hotspot* and *in-silico* models.

Each observed variant across six genes in our HCM cases was given supporting (OR > 10; Sup.), moderate (OR > 20; Mod.) or strong (OR > 100; Str.) evidence of pathogenicity based on model predictions from the *hotspot* and *in-silico* models. The proportion of variants with supporting, moderate or strong evidence are stratified by expert classifications made by Oxford Medical Genetics Laboratory.

	VUS				***Likely pathogenic***				***Pathogenic***			
	N	*Sup.*	*Mod.*	*Str.*	*N*	*Sup.*	*Mod.*	*Str.*	*N*	*Sup.*	*Mod.*	*Str.*
Hotspot-models												
MYH7	***123***	10%	28%	0%	***93***	33%	55%	0%	***36***	31%	61%	0%
MYBPC3	***97***	35%	12%	0%	***13***	38%	23%	0%	***6***	33%	67%	0%
TNNT2	***19***	21%	47%	0%	***4***	0%	50%	0%	***4***	0%	100%	0%
TNNI3	***18***	17%	67%	0%	***11***	9%	91%	0%	***4***	0%	100%	0%
MYL2	***12***	25%	0%	0%	***1***	100%	0%	0%	***2***	100%	0%	0%
MYL3	***10***	0%	0%	0%	***3***	0%	0%	0%	***1***	0%	0%	0%
In-silico models												
MYH7	***123***	18%	26%	4%	***93***	16%	58%	14%	***36***	8%	75%	17%
MYBPC3	***97***	16%	25%	7%	***13***	15%	54%	15%	***6***	17%	33%	50%
TNNT2	***19***	16%	26%	16%	***4***	0%	50%	25%	***4***	0%	50%	50%
TNNI3	***18***	22%	33%	22%	***11***	18%	45%	36%	***4***	0%	50%	50%
MYL2	***12***	0%	25%	0%	***1***	0%	100%	0%	***2***	50%	50%	0%
MYL3	***10***	0%	0%	0%	***3***	0%	0%	0%	***1***	0%	0%	0%

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A web application, *pathogenicity_by_postion*, is available to facilitate the exploration of the *hotspot* and *in-silico* models (R Shiny: https://adamwaring.shinyapps.io/Pathogenicity_by_position). Users can explore alternative models and submit their own missense variants to retrieve predicted ORs and support intervals. A further R package is available for cluster-detection and association testing using *BIN-test* and *ClusterBurden* (https://github.com/adamwaring/ClusterBurden).

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

We present here new statistical approaches to incorporate residue position in the analysis of rare-missense variants in Mendelian disease genes. Our association tests were well-calibrated in simulated data, the *BIN-test* detected significant clustering in almost all firmly-established HCM genes, and *ClusterBurden* gave superior power over a simple burden test. Data-driven models were applied to six core sarcomeric genes to estimate mutational hotspots and provides a flexible method for quantitative application of ACMG criteria PM1 and PP3. Our results demonstrate that residue position can be a powerful predictor of both gene and variant pathogenicity and can quantify the statistical uncertainty surrounding the application of *in-silico* algorithms.

BIN-test is a powerful method to test the hypothesis of variant clustering in known or putative disease genes. *ClusterBurden* is a gene association test with superior power over a standalone burden test in situations where pathogenic variants cluster in specific protein regions. Both tests keep false-positives below 5% and are rapid to compute, making them scalable for whole-exome scanning of very-large datasets like the UK Biobank. Although *ClusterBurden* has slightly reduced power whenever clustering is absent, we observed clustering for most well-established HCM genes where missense variants cause disease. Therefore, this method has the potential be more powerful to detect undiscovered low-penetrance genes than a burden-only test.

The most significant position signal was observed in the beta myosin heavy chain protein (*MYH7:* ENST00000355349), a finding that has been long recognised [35]. Most of the signal was observed in the myosin-head motor domain between residues 100-900, in two clusters peaking at 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).

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [36]. To explore whether the signal was overly driven by high-frequency founder mutations, seven variants with allele counts above 10; c.2429G>A, c.2308G>A, c.1624G>C, c.1504C>T, c.1484G>A, c.772G>A and c.655G>C (ENST00000545968), 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).

Eighty-nine percent of 27 case variants 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 [37]. Mutations between residues 92-110 have been previously noted to impair tropomyosin dependent functions in *TNNT2* [38] and six variants map to this region. In *TNNI3* (ENST00000344887), which encodes cardiac troponin I, 91% of 34 case 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 [39]. In *MYL2* (ENST00000228841), which encodes myosin regulatory light chain, 50% of 30 case 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, 79% of 14 case variants cluster between residues 125 and 175 (Figure 2) whereas control variants were more uniformly distributed.

GAMs were used to model variant pathogenicity based on mutational hotspots (*hotspot-model)* and *in-silico* predictors (*in-silico-model*). GAMs have attractive statistical properties, not necessarily shared by other machine-learning approaches, in that they can produce familiar interpretable results via

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variant-specific ORs and accompanying 95% confidence intervals. Unlike empirical ORs, based solely on observed frequencies for variants, GAM ORs draw upon a much larger pool of information, including features of other variants in the training set. This permits the estimation of variant-specific ORs whenever the empirical frequencies are uninformative. Furthermore, as the GAM response variable is case status, the models are unbiased by previous classifications, and account for both incomplete penetrance and benign background rare-variation.

Reassuringly, model predictions were positively correlated with expert manually-curated classifications. Using a probabilistic approach, we attributed different levels of evidence for the criteria PM1 and PP3. Currently for HCM, criteria PM1 is applied as moderate evidence to *MYH7* for variants that fall in the residue 181-937 motor-domain [15]. The *hotspot-model* extends this criteria to five more sarcomeric genes and stratifies evidence as either supporting or moderate. When *in-silico* predictors were included in the model, evidence was stratified as supporting, benign or occasionally strong. However, this relies on effectively collapsing two ACMG criteria into one, a relevant modification of the current additive guidelines [11].

CONCLUSIONS

Here we have demonstrated proof-of-concept data-driven modelling as an effective tool for variant interpretation in HCM. Collaboration for more cases could substantially improve model performance and reduce uncertainty in estimates, especially for genes with few observations such as *MYL2* and *MYL3*, but could also allow informative modelling of further HCM genes and other Mendelian diseases. Furthermore, the modelling approach developed could be modified into a Bayesian framework allowing inclusion of previously reported pathogenic and benign mutations, available from public resources such as clinVAR, as model priors, reducing uncertainty further. Our association methods, have proved theoretically powerful and could be used to scan for novel cardiomyopathy genes when large datasets such as 500K UK Biobank exomes become available.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

LIST OF ABBREVIATIONS

ACMG – American College of Medical Genetics

VUS – Variant of Uncertain Significance

HCM – Hypertrophic Cardiomyopathy

GAM – Generalized Additive Models

OMGL – Oxford Medical Genetics Laboratory

HCMR – Hypertrophic Cardiomyopathy Registry

BAM – Binary alignment map

AD – Anderson-Darling

KS – Kolmogorov-Smirnov

AUC – Area under the Curve

ROC – Receiver Operating characteristic

OR – Odds-ratio

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DECLARATIONS

Ethics

The research protocol was approved by the South Central - Oxford A Research Ethics Committee (REC reference: 14/SC/0190); written informed consent was obtained from all participants.

Availability of data and materials

Due to the confidential nature of some of the research materials supporting this publication not all of the data can be made accessible to other researchers. Please contact the corresponding author for more information.

Competing interests

None declared

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].

Author contributions

AW and MF led conception and design of work
AW contributed to data curation, conducted all analyses, developed statistical methods and associated software and wrote each draft of manuscript
MF supervised all analyses and interpretation of results
AH led the curation and quality control of the hypertrophic cardiomyopathy datasets
SS contributed to quality control of the data
CK and NS supported access to the HCMR dataset
KT and HW advised and guided the clinical aspects of the work
MF and KT reviewed and editing each draft of the manuscript

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.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 USA* 1994; 91:9622–9625
2. 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
3. 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
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, 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
6. 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
7. 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
8. Collins FS. Positional cloning moves from perditional to traditional. *Nat Genet* 1995; 9:347-50

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. Walsh R, Thompson K, Ware J, Funke B, Woodley J, McGuire K, Mazzarotto F, Blair E, Seller A, Taylor JC, Minikel EV, Exome Aggregation Consortium, MacArthur DG, Farrall M, Cook SA, Watkins H. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* 2017; 19:192-203.

10. 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

11. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. 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

12. Ghosh R, Oak N and Plon S. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* 2017; 225.

13. Watkins H, Ashrafian H, Redwood C. Inherited cardiomyopathies. N Engl J Med. 2011; 364:1643-56

14. 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

15. 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

16. 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

17. 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

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variant curation: recommendations from the ClinGen PTEN expert panel. *Hum Mutat* 2018; 39:1581-1592

18. 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
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291
20. Kramer CM, Appelbaum E, Desai MY, Desvigne-Nickens P, DiMarco JP, Friedrich MG, Geller N, Heckler S, Ho CY, Jerosch-Herold M, Ivey EA, Keleti J, Kim DY, Kolm P, Kwong RY, Maron MS, Schulz-Menger J, Piechnik S, Watkins H, Weintraub WS, Wu P, Neubauer S. Hypertrophic Cardiomyopathy Registry: The rationale and design of an international, observational study of hypertrophic cardiomyopathy. *Am Heart J* 2015; 170, 223–230
21. 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
22. Anderson TW, Darling DA. "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". *Ann Math Stat* 1952; 23:193–212
23. Kolmogorov AN. Sulla Determinazione Empirica di Una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari* 1933; 4:83-91

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24. Davison, A, & Hinkley, D. Bootstrap Methods and their Application (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. 1997

25. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd. 1925.

26. Spearman CE. The proof and measurement of association between two things. *Am J Psychol* 1904; 15:72–101

27. Lin W. Association Testing of Clustered Rare Causal Variants in Case-Control Studies. *PLoS One* 2014; 9: e94337

28. 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

29. 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

30. Madsen, BE, Browning, SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet* 2009; 5: e1000384

31. Hastie T, Tibshirani R. Generalized Additive Models. Chapman & Hall, London. 1990

32. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J Roy Stat Soc* 2011; 73:3-36

33. 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

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

35. 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

36. 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

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

38. 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

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

39. 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.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

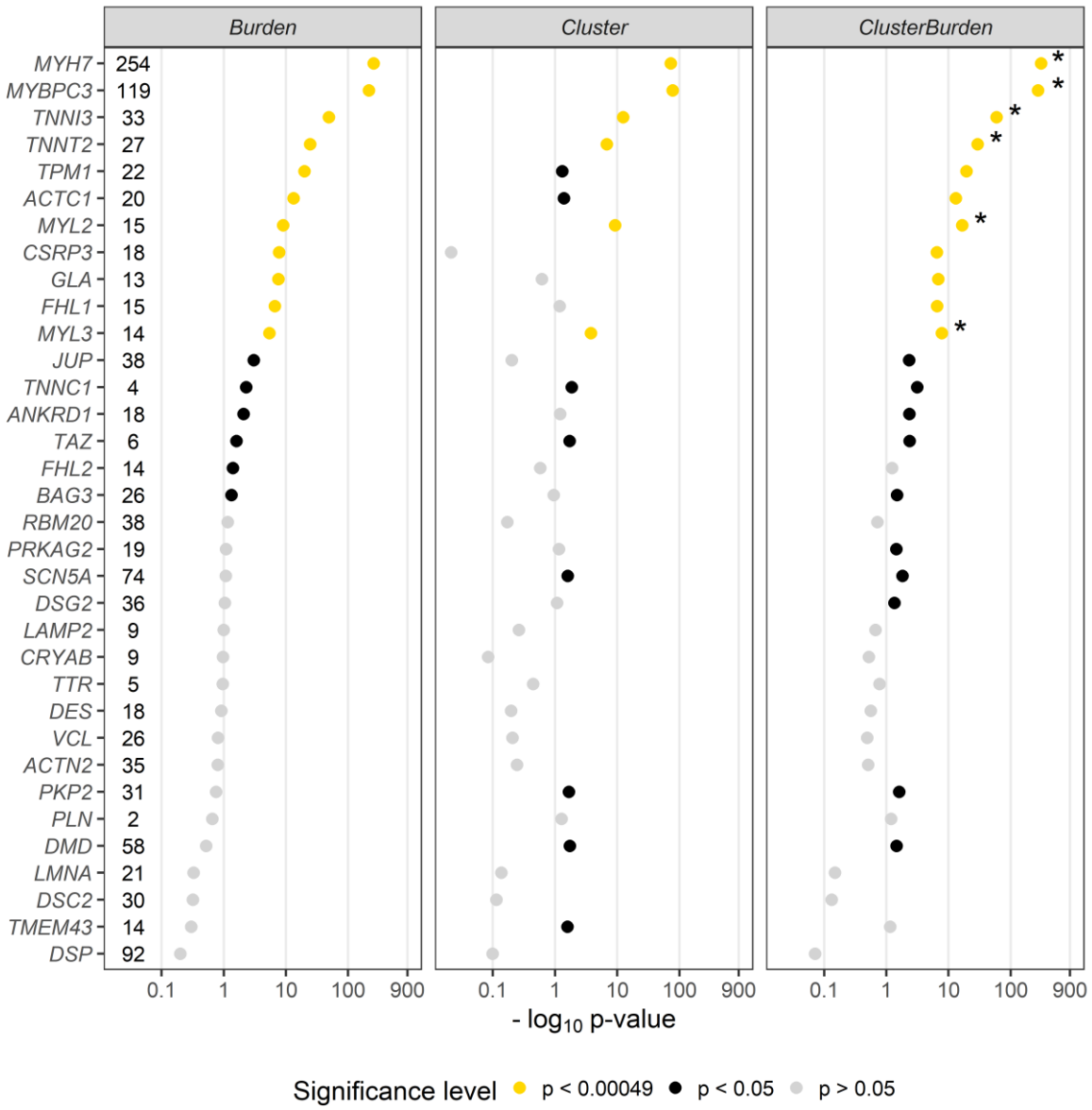

Figure 1: Association analysis using Fisher's-exact test (Burden), *BIN*-test (Cluster) and *ClusterBurden* of 34 genes commonly present on cardiomyopathy gene-panels.

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.

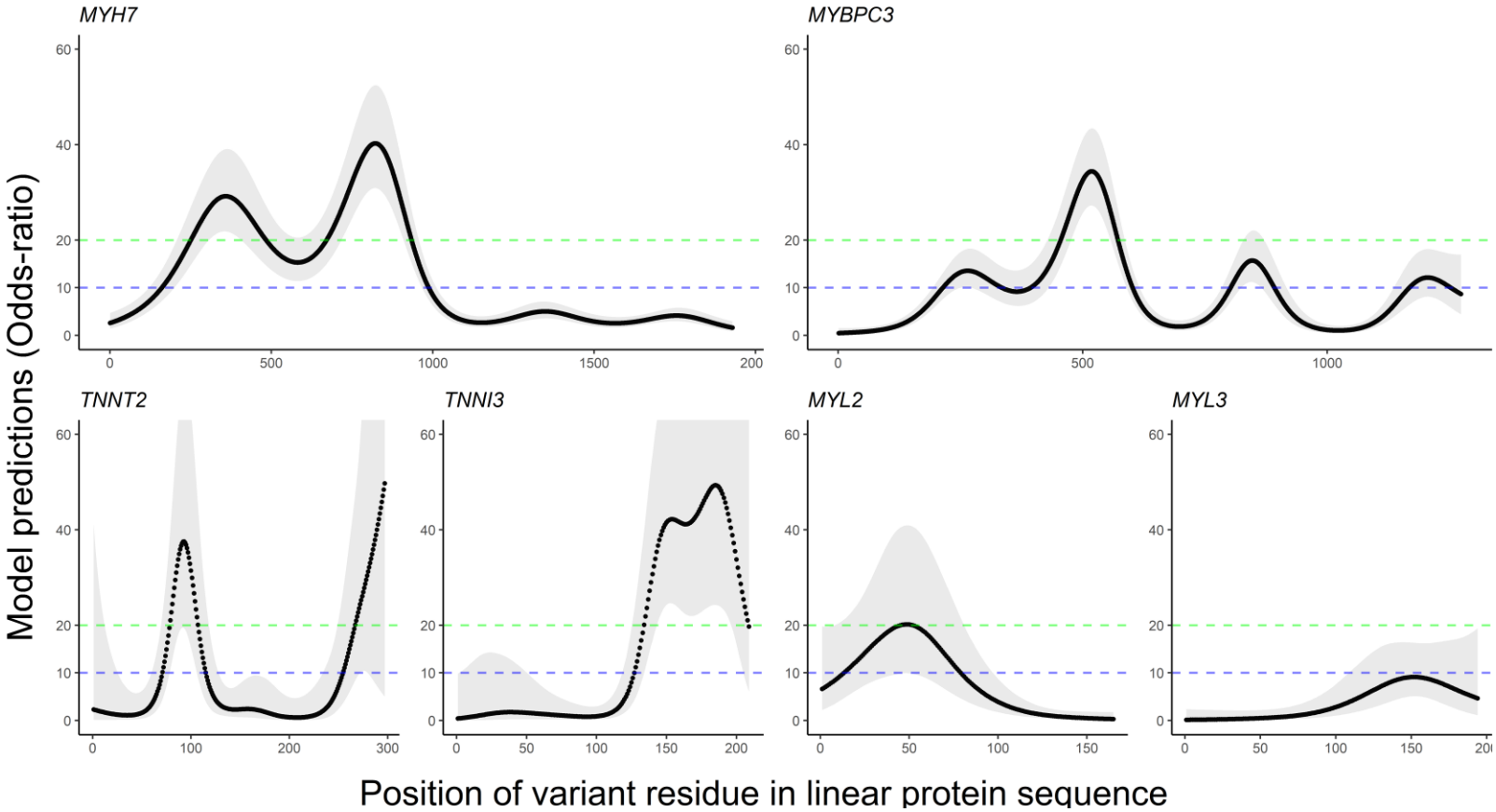

Figure 2: Odds-ratio (OR) predictions and 95% confidence intervals for *hotspot-models*.

Mutational hotspots were estimated for six firmly established HCM disease-genes. The 95% confidence intervals for model predictions are displayed as light grey shading. Regions above the blue (OR=10) and green (OR=20) dashed lines ascribe supporting and moderate evidence respectively, for criteria PM1.

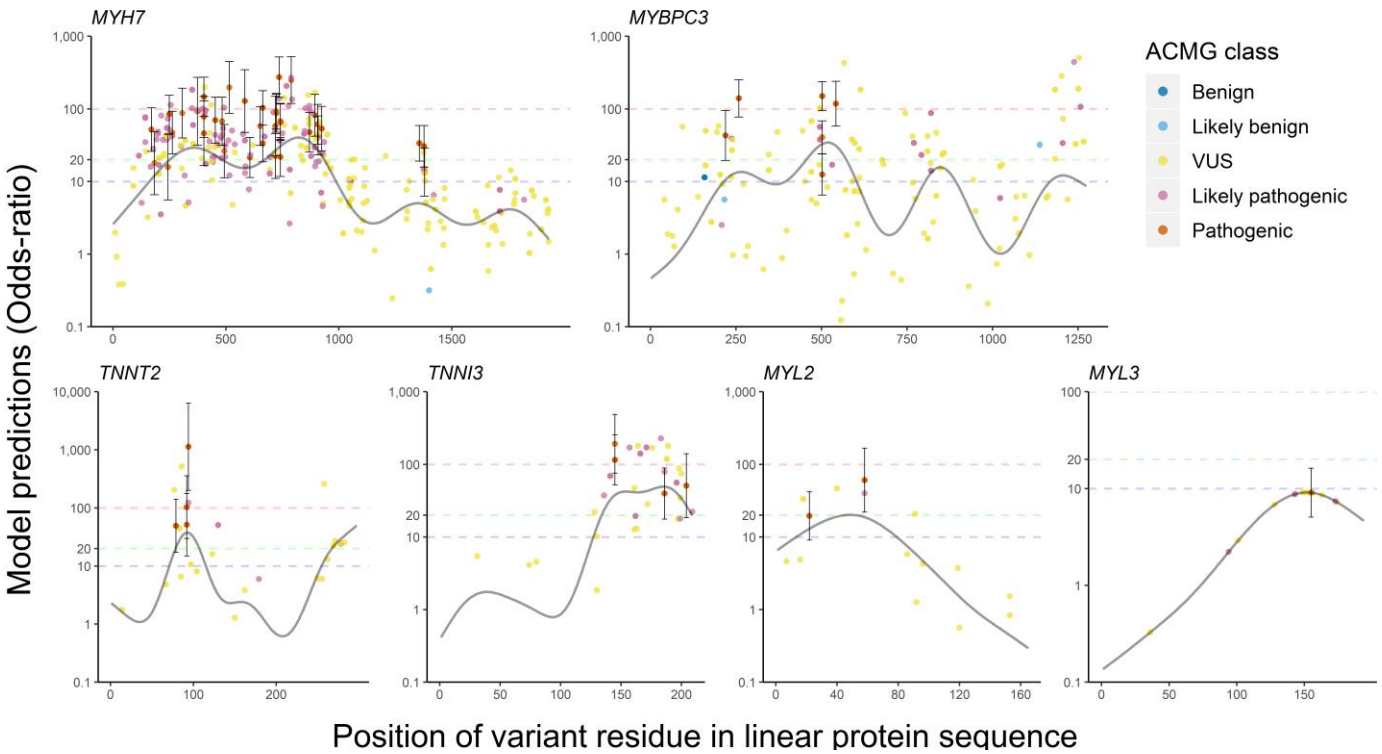

Figure 3: Odds-ratio (OR) predictions and 95% confidence intervals for *in-silico-models.*

Each point denotes a rare-variant in the HCM dataset and is coloured to indicate its expert classification. ORs on the y-axis are displayed on a $\log_{10}$ scale and were derived from the *in-silico-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 (*hotspot-model*). Dashed red lines indicate an OR of 1, dashed blue lines indicates the OR for the uniform burden model.

bioRxiv preprint doi: https://doi.org/10.1101/826164; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

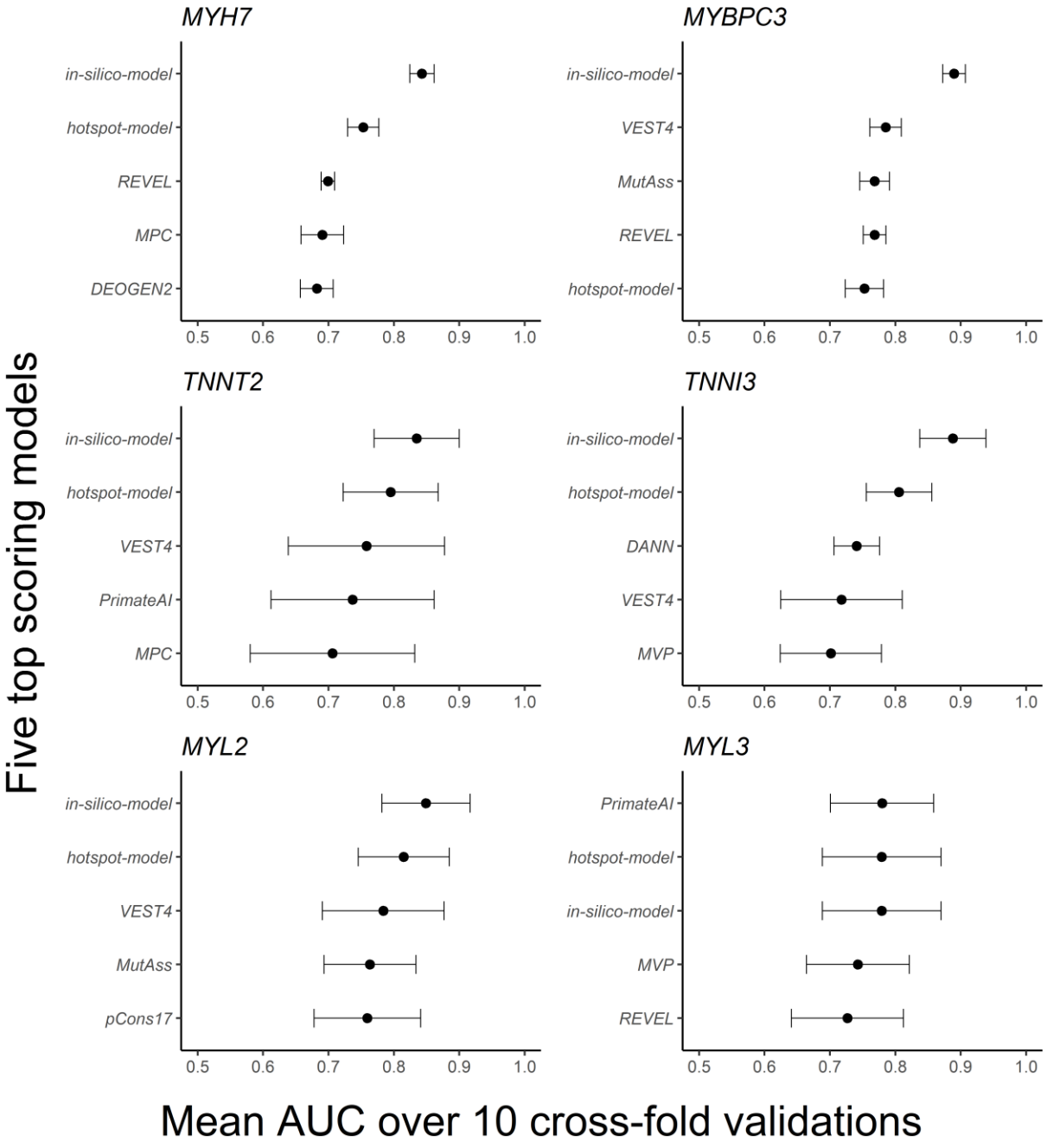

Figure 4: Means and standard deviations of area under the curve (AUC) across 10 cross-fold validations.

For each gene, the *in-silico-model* and *hotspot-model* are compared to each individual *in-silico* predictor from dbNSFP. Each model is trained on the same HCM-gnomAD case-control missense variants all filtered at a gnomAD population maximum frequency of 0.01%. Only the five highest mean AUC scoring models are displayed.