S1 Methods: A – Hypertrophic cardiomyopathy cohorts used in this analysis. B – Bioinformatics pipelines used for processing of raw data.

A) A total of 5,393 cases of hypertrophic cardiomyopathy (HCM) were available by combing two datasets; the Oxford Medical Genetics Laboratory and the Hypertrophic Cardiomyopathy Registry (HCMR). There were 2,757 probands sequenced at OMGL, all referred by cardiac specialists between 2013 and 2018. The HCMR cohort consisted of 2,636 probands with a clinical diagnosis of HCM, also sequenced between 2013 and 2018.

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

S2 Methods: Implementation details and properties of the forward-time simulator.

Overview

A forward-time simulation program was developed in the R programming language using object orientated programming with R's built-in reference classes. The data.table package (Dowle and Srinivasan 2019) was used to efficiently simulate populations. The simulator is capable of evolving protein sequences with differential selection across their linear form, imitating variant clustering observed in our hypertrophic cardiomyopathy genes. The simulator is sufficient to imitate real world datasets for the purpose of comparing power and type 1 error for different methods.

The demographic model, followed Kryukov *et al.* (2009) maximum likelihood model, and includes an emulation of recent European population history. The three-stage model begins with a burn-in stage of 10,000 generations with an initial diploid population of 8,100 individuals. A brief "bottleneck", acknowledging the out of Africa hypothesis, followed for 100 generations with a reduced population size of 7,900. Most importantly, there was a final stage of exponential growth allowing the population to reach a size of 900,000 individuals in 370 generations. Each generation evolved via a three-stage process of mutation, selection and mating.

Protein length

To explore the relationship between protein length (number of amino acid residues) and statistical power, protein sequences were simulated at two different lengths; 500 and 1,000 residues. The chosen values for protein length fall within the range of observed protein lengths for both SWISSPROT (n=20,233, mean=560) and the HCM gene panel (n=35, mean=688; **Fig. 1**).

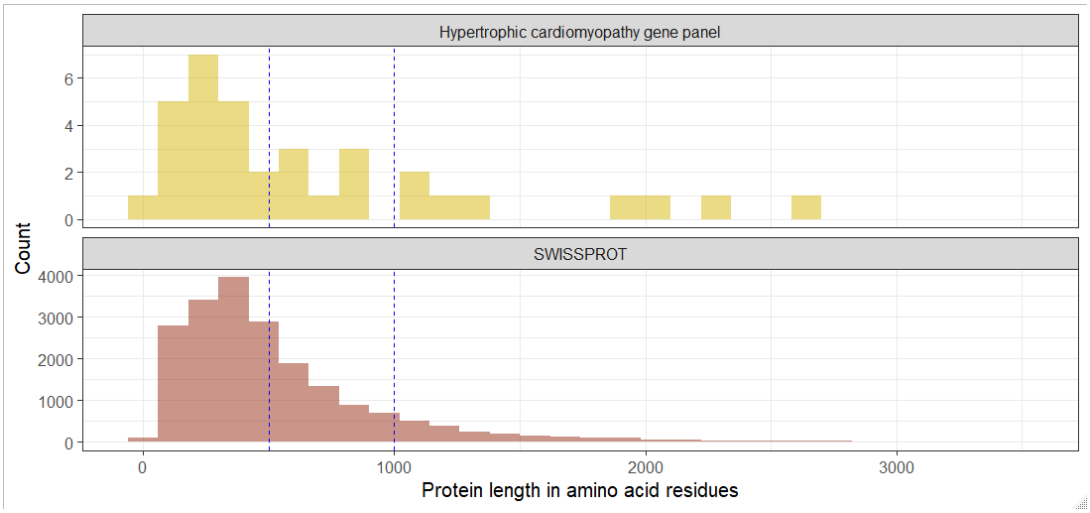

Figure 1: Protein lengths in 20,233 SWISSPROT proteins and 35 genes associated with hypertrophic cardiomyopathy. *The x-axis is truncated at 3,500 residues, excluding very large proteins in SWISSPROT, such as TTN, that otherwise obscure the distribution. Blue dashed vertical lines at 500 and 1,000 denote the selected protein lengths for our simulations.*

Mutations

In each generation the number of new mutations was randomly drawn from a binomial distribution X~B(n,p) where *n* is the total number of mutable sites in the population and *p* is the mutation rate. The mutation rate was calculated by multiplying 1.8 x 10^{-8}, the per-nucleotide base mutation rate in the human genome (Peng and Liu 2011), by the 3 bases in a codon and the empirical probability of a missense mutation 0.619. Thus, we made the simplifying assumption that all base changes are equally likely and all amino acids are present in equal proportions. This approximates an average missense mutation rate of 3.34 x 10^{-8}.

To determine whether the simulated data was broadly similar to our observed data we compared the total number of segregating sites (SS) and the site frequency spectrum (SFS) to our HCM-gnomAD case-control dataset (**Fig. 2**). When SS were adjusted for sequence length and sample size, the number of SS were similar to the real data but had fewer on average. The SFS were comparable between the real and simulated data.

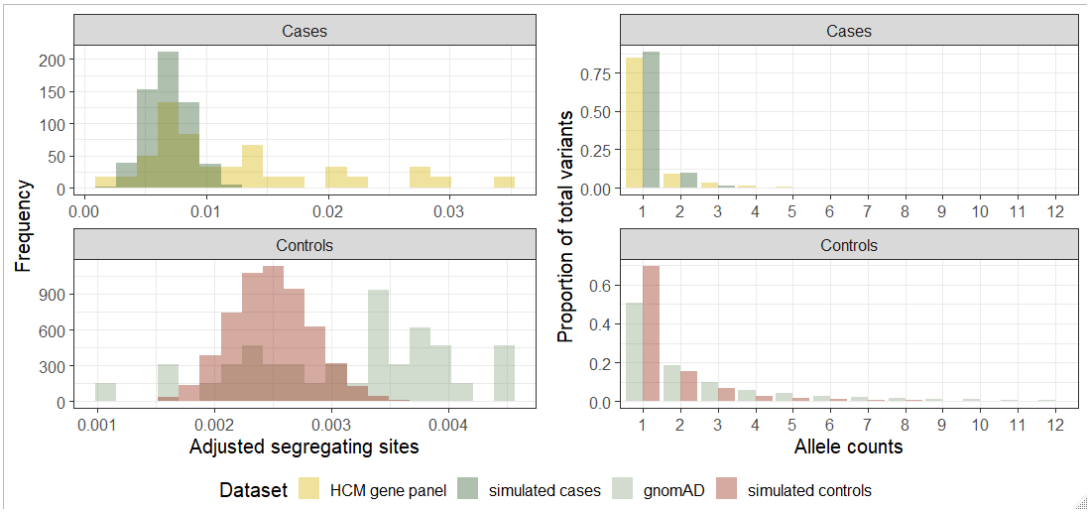

Figure 2: The number of segregating sites and the site-frequency spectrum for simulated data and observed data. Segregating sites are adjusted for sample size and total number of sites in the sequence.

Clustering models

Three different amino-acid residue clustering models were explored; uniform, single-cluster and multiple-cluster. In the uniform model all variants in a protein are equally likely to be pathogenic. In the clustered models, only specific regions of the protein harbour pathogenic variants. Both clustering scenarios were apparent in the HCM gene panel data.

To implement the clustering scenarios, discrete pathogenic regions were specified before evolution. Cluster sizes were randomly drawn to span one-twentieth to one-quarter of the total protein length, were randomly located and could overlap. In the multiple-cluster model there was 70:30% chance of two or three pathogenic regions respectively. Each residue in the sequence was assigned a static selection coefficient (SSC). For the uniform model the SSC was 1 for every residue. For the clustered

models, residues that fell within a pathogenic region (PR) were given SSCs of 5 and residues outside PRs were given SSCs of 1. Normalized SSCs were calculated by dividing by the mean SSC across the protein so the resultant mean SSC of the sequence was 1. After evolution begins each new mutation was assigned a random selection coefficient (RCC) drawn from an exponential distribution (**Fig. 3**). Selection coefficients varied between 0.0001 and 0.1 where the log transformation of the distribution is uniform in this range. The SSC and the RCC were multiplied to give the final selection coefficient (FSC).

A non-random mating process was implemented whereby the chance of success was additively dependent on the FSC of each variant carried by an individual. As most variants were detrimental to survival, most carriers had reduced fitness and a reduced probability to mate. As the primary interest was in rare variants, recombination was ignored.

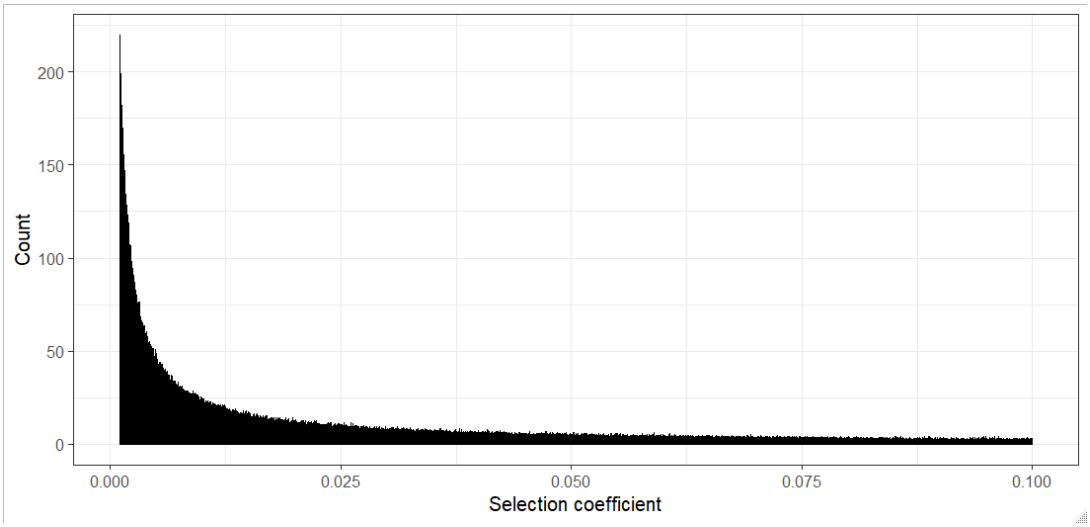

Figure 3: The distribution of randomly allocated selection coefficients assigned to new mutations in the simulated populations. *The minimum coefficient was 0.0001 and the largest coefficient was 0.1.*

To monitor the selection procedure, FSCs were compared to the final allele counts of simulated mutations. There is an inverse relationship between allele count and mean selection FSC where high frequency variants had weaker FSCs (**Fig. 5; upper panel**). Variants within pathogenic regions, defined by the clustered models, had higher FSCs than variants outside these regions (**Fig. 5; lower panel**).

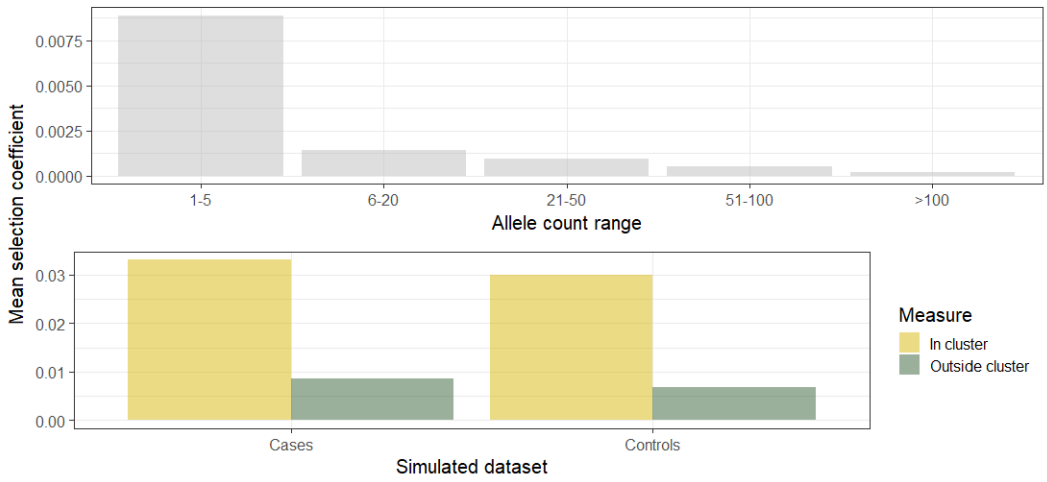

Figure 5: Selection coefficients of simulated variants. *The upper panel shows the relationship between the final allele count of simulated variants and their static selection coefficients. The lower panel shows the average selection coefficient of variants that fall inside and outside the causal cluster(s).*

Simulated datasets

In order to generate simulated cohorts with differential distributions of variants, simulated phenotype was based on both the FSC and variant position. This is necessary as selection coefficient alone is insufficient to define a variant as disease-causing. This is analogous to an incompletely penetrant variant or a disease that does not lower reproductive fitness. Odds-ratios were drawn from an L-shaped distribution with mean 2.60 and median 2.35. All variants in the population were then ranked by FSC and assigned the equivalent ranked odds-ratio, therefore variants with the highest FSC had the highest odds-ratios. These odds-ratios were then multiplied by 3 or 1/3 depending on whether they fall within or outside a pathogenic region. Affection status was then simulated based on these odds-ratios according to the formula;

$$P(Affected|X) = \frac{e^{\alpha+BX}}{1 + e^{\alpha+BX}}$$

Here X is the vector of genotypes, B is their respective odds ratios and alpha was defined by log(prevalence / 1 –prevalence). Prevalence was specified as 1/500 to correspond to the population prevalence of hypertrophic cardiomyopathy (Maron et al 1995). After each sample was assigned a phenotype, 2,500 cases and 123,000 controls were randomly selected from the population. For each gene length we simulated 10,000 populations in order to produce precise estimates of type 1 error with a binomial exact 95% confidence interval (500 successes in 10,000 trials) of 0.0458 to 0.0545.

References

Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.2. https://CRAN.R-project.org/package=data.table.

Kryukov G., Shpunt A., Stamatoyannopoulos J., Sunyaev S. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. PNAS. 106 (10): 3871-3876.

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

Maron B., Gardin J., Flack J., Gidding S., Kurosaki T., Bild D. (1995). Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. Circulation. 92 (4): 785-9.

S3 Methods: Controlling for uneven coverage between cases and controls

GAMs

Low coverage NGS regions were present in the GnomAD exomes or the HCM data for several of the genes under investigation for mutational hotspots (Figure 6). Although clinical laboratories routinely survey poorly covered NGS regions by Sanger sequencing, this data could not be accessed due to patient confidentiality reasons. The GnomAD exomes data constitutes an aggregation of multiple exome sequencing projects using a variety of capture and sequencing technologies and it also suffers from uneven coverage – most notably in *MYBPC3*. The consequence of this was that the effective sample size varied across each protein as not all samples were covered to sufficient depth for fully inclusive missense variant-calling at every residue.

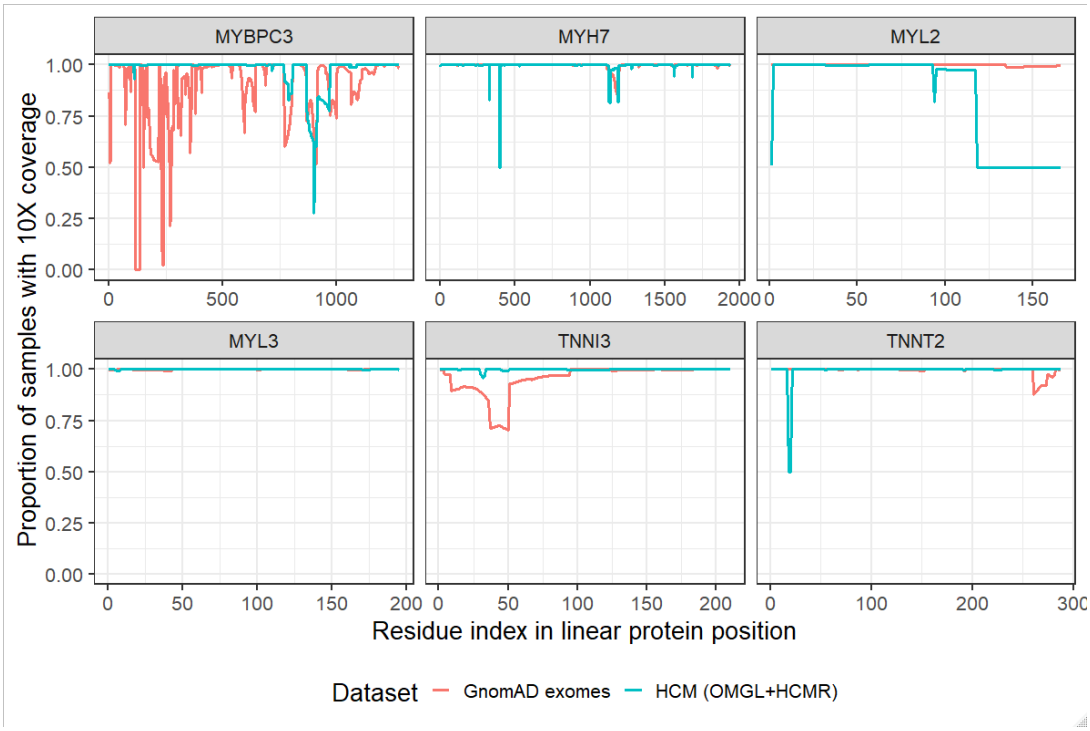

Figure 6: Proportion of samples with at least 10X coverage for all residues in the proteins investigated for mutational hotspots.

As uneven coverage at the nucleotide level could skew distributions of observed amino-acid variants within each cohort, giving rise to artificial hotspots, this was adjusted for in the model by assigning prior weights to the contribution of each variant residue to the log likelihood. Each variant was weighted by the reciprocal of the mean 10X coverage in the local region for its cohort (i.e. cases or controls). The local region was specified as the variant residue plus five residue flanks towards the N and C termini (i.e. a total of 11 residues). The metric used to summarise coverage (the proportion of samples with at least 10X coverage) is widely used to specify low coverage regions where DNA variant calls may be missed (Sims *et al.* 2014, Wang *et al.* 2017, Guo *et al.* 2018, Povysil *et al.* 2020). The consequence of this weighting procedure is to up-weight low coverage regions in the model. Note that

the routine NGS QC criteria applied to the HCM case and gnomAD control data mean that missense variants in low coverage regions are effectively masked (i.e. variant counts at these residues are set to zero).

We explored the impact of the re-weighting approach for coverage adjustment on the GAM hot-spot analysis (Figure 7). Variants were assigned to idealised case-control cohorts carrying 100 or 200 total variants for 1) a disease model containing one C-terminal hotspot and 2) a null model of uniform distribution of variants. For each of these four scenarios; null 100, null 200, disease 100 and disease 200, three models were trained. The first was with the original 100% coverage data (Unmodified model). In the second, the region between 100 and 200 residues was designated as a low coverage region (10X = 0.5) and 50% of observations were randomly removed (LowCov model). This low coverage region was then adjusted in model 3 by the weighting procedure (Adjusted model). For all four scenarios, the adjustment procedure successfully compensated for low coverage to produce very similar models to the unmodified data.

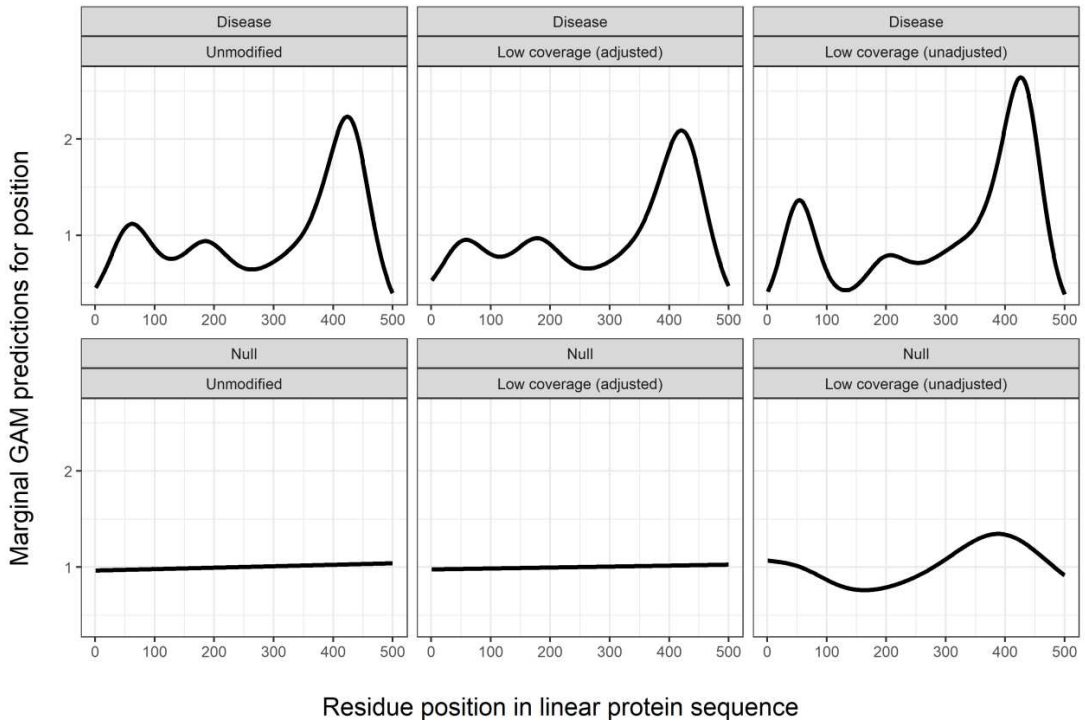

Figure 7: Effect of reciprocal coverage weighting on GAM models for simulated sequences with low coverage regions.

BIN-test

The coverage was evaluated across the 34 genes examined in this paper (Figure 8). Regions of low coverage were most notable in the genes; *CRYAB, DMD*, *FHL1, GLA, LAMP2, LMNA, MYBPC3, MYL2, SCN5A* and *TAZ*. The approach used to adjust for uneven coverage in the *BIN-test* was similar to that used for GAMs. After binning the data, the missense variant counts in each cohort was divided by the coverage in that region, thereby imputing missing data where the coverage was below 1. The coverage for each bin was calculated as the mean proportion of samples with at least 10X coverage over all nucelotides encompassed by the bin.

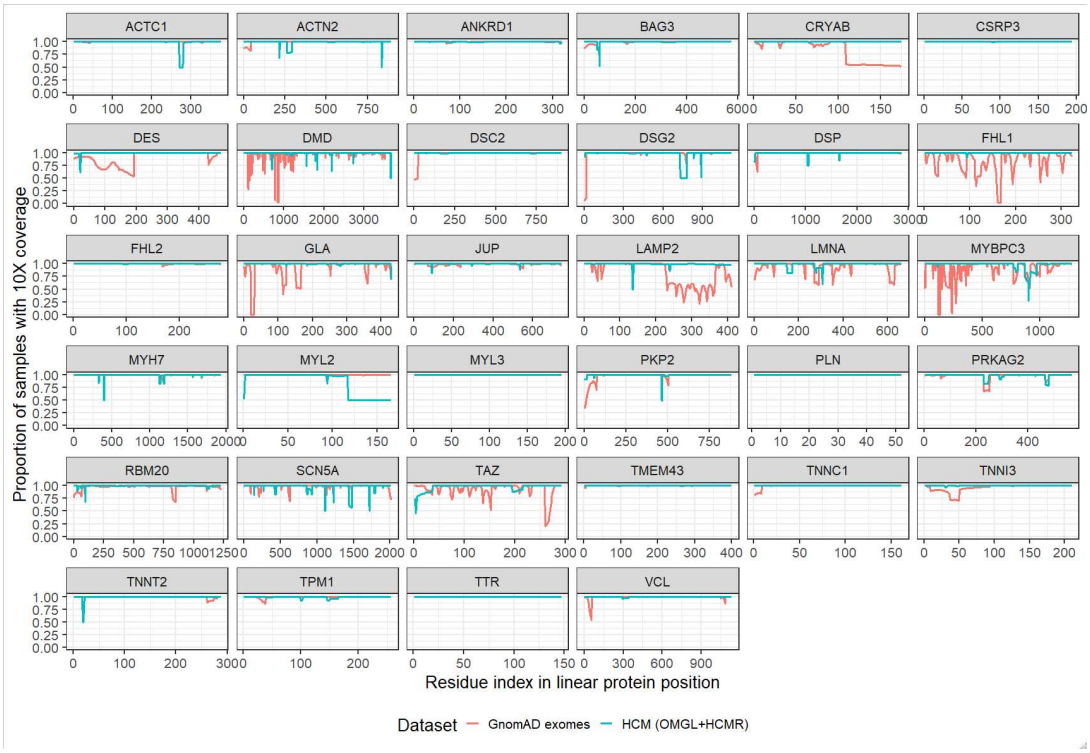

Figure 8: Proportion of samples with at least 10X coverage for all residues across 34-cardiomyopathy gene panel genes in HCM (OMGL+HCMR) and GnomAD exomes.

To assess the performance of this adjustment procedure under a scenario assumed to represent the null, *BIN-test* p-values were calculated for synonymous variants in HCM cases and GnomAD exomes controls adjusted and unadjusted for coverage (Figure 9). Reassuringly, no results were Bonferroni significant with or without adjustment for coverage. Furthermore, coverage adjustment reduced the significance of 5/6 nominally significant unadjusted p-values leaving only three nominally significant results, all of which were above 0.01.

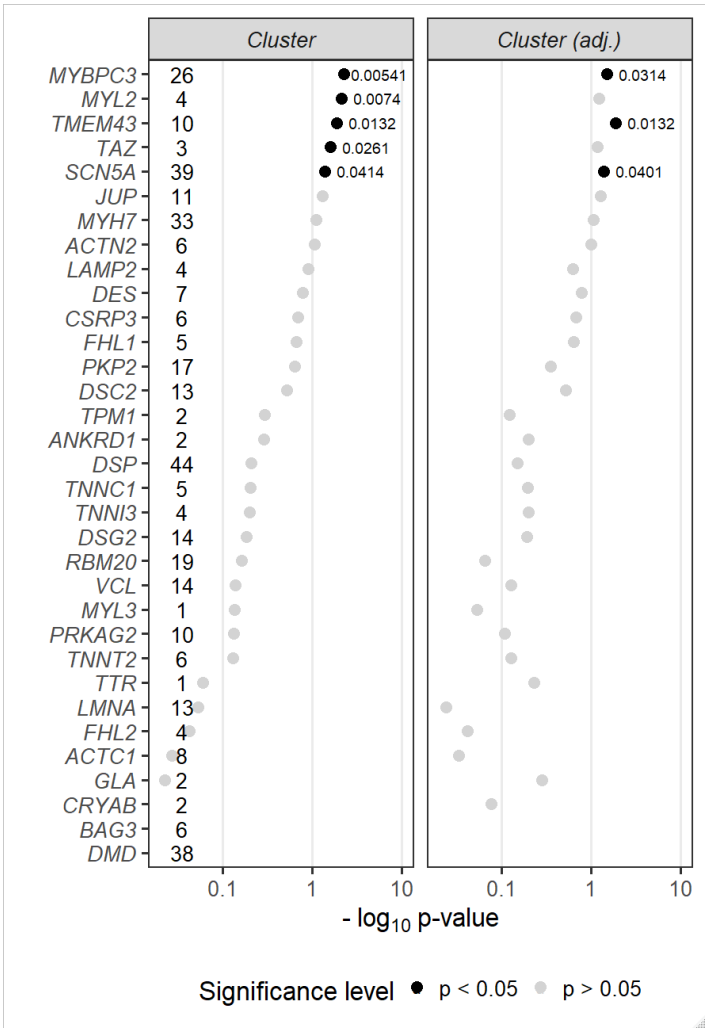

Figure 9: Burden, clustering and combined p-values for synonymous variants in HCM cases versus GnomAD exomes controls.

References

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

Sims, D., Sudbery, I., Ilott, N. et al. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–132 (2014).

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

Povysil, G., Petrovski, S., Hostyk, J. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat Rev Genet 20, 747–759 (2019).