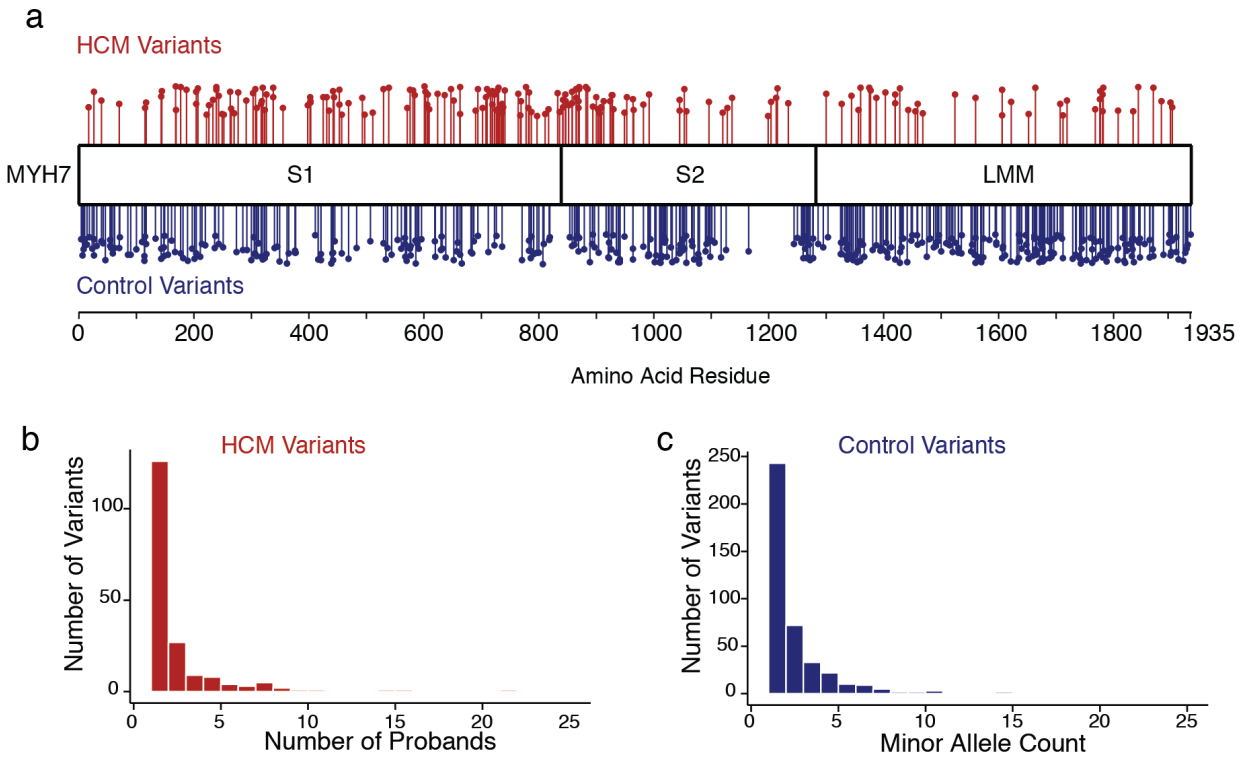
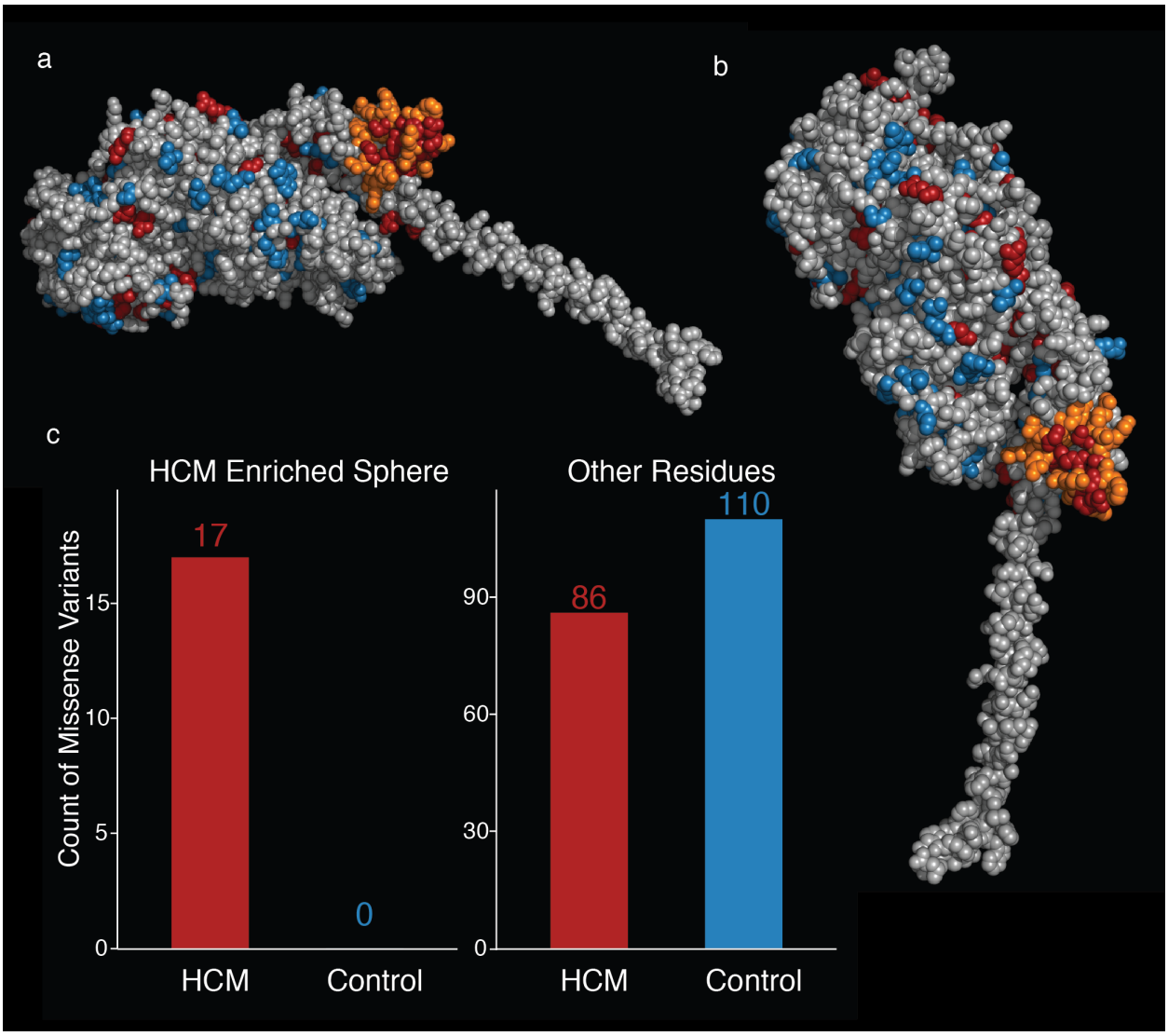
**Supplement for: Multi-dimensional structure function relationships in human β-cardiac myosin from population scale genetic variation**

Julian R. Homburger[1], Eric M. Green[2], Colleen Caleshu[3,4], Margaret Sunitha[5], Rebecca Taylor[6], Kathleen M. Ruppel[6,7], SHaRe Investigators, Steven D. Colan[8], Michelle Michels[9], Sharlene Day[10], Iacopo Olivotto[11], Carlos D. Bustamante[1,12], Frederick Dewey[13], Carolyn Y. Ho[14], James A. Spudich[5,6*†], Euan A. Ashley[1,3*†]

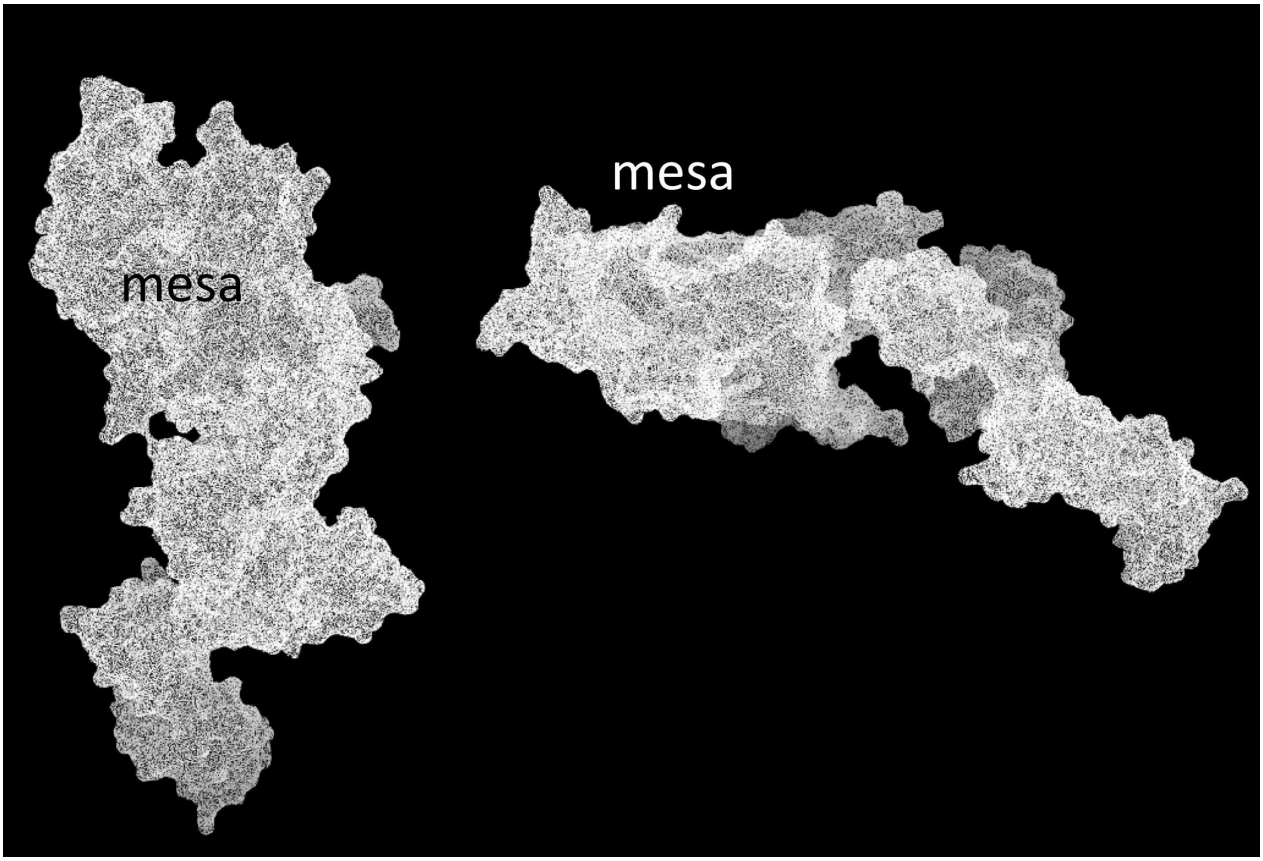
## Supplementary Figures



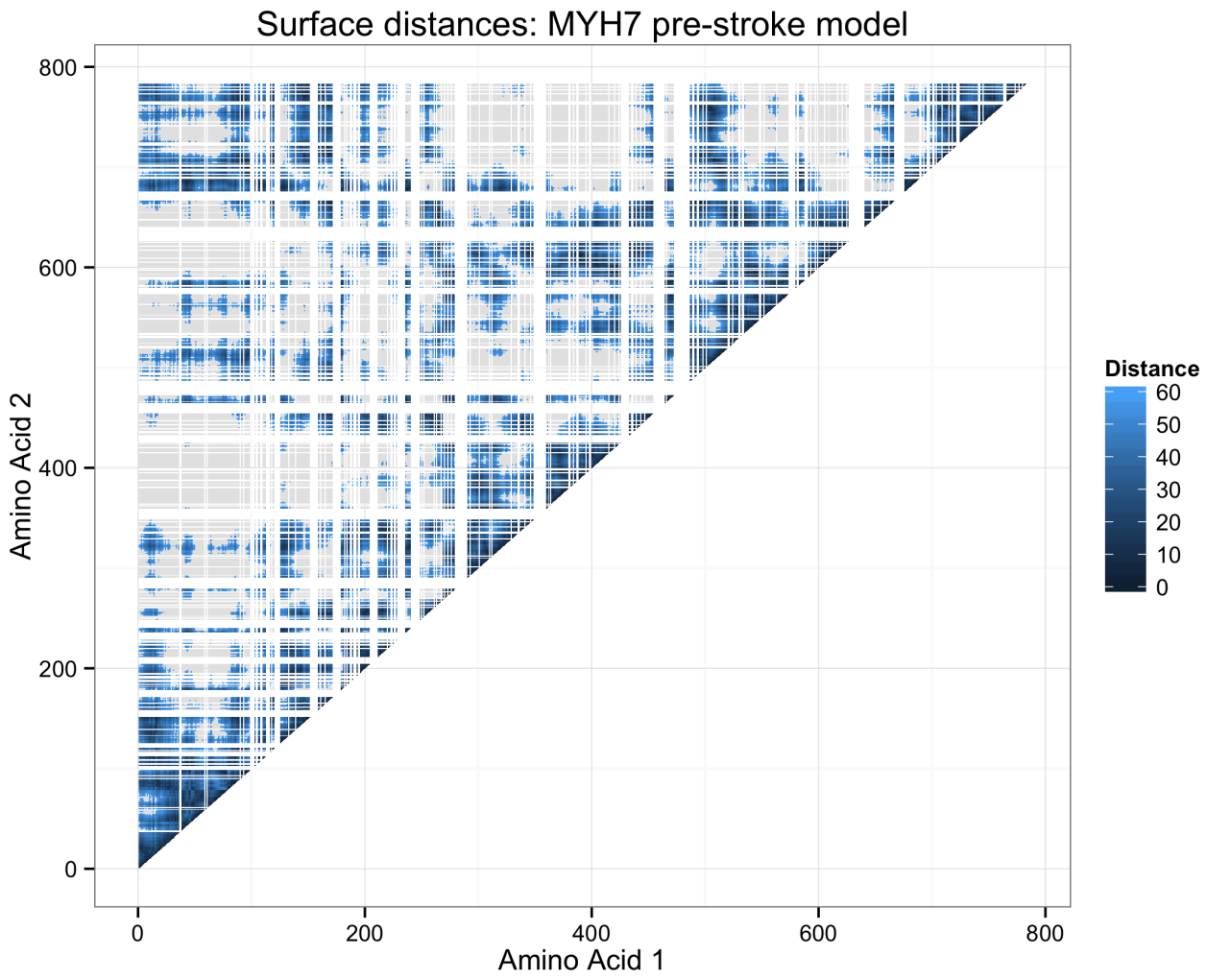

**Supplementary Figure S1. Differences in the position of missense variants between HCM and control cohorts in human β-cardiac myosin.** Missense variants identified in SHaRe HCM patients are in red and those identified in ExAC individuals are in blue. (b) Histogram of frequency of observation of *MYH7* missense variants in SHaRe HCM patients. (c) Histogram of the number of observed alleles of *MYH7* missense variants in the ExAC. 15 missense variants with a frequency above 0.005 are not shown.



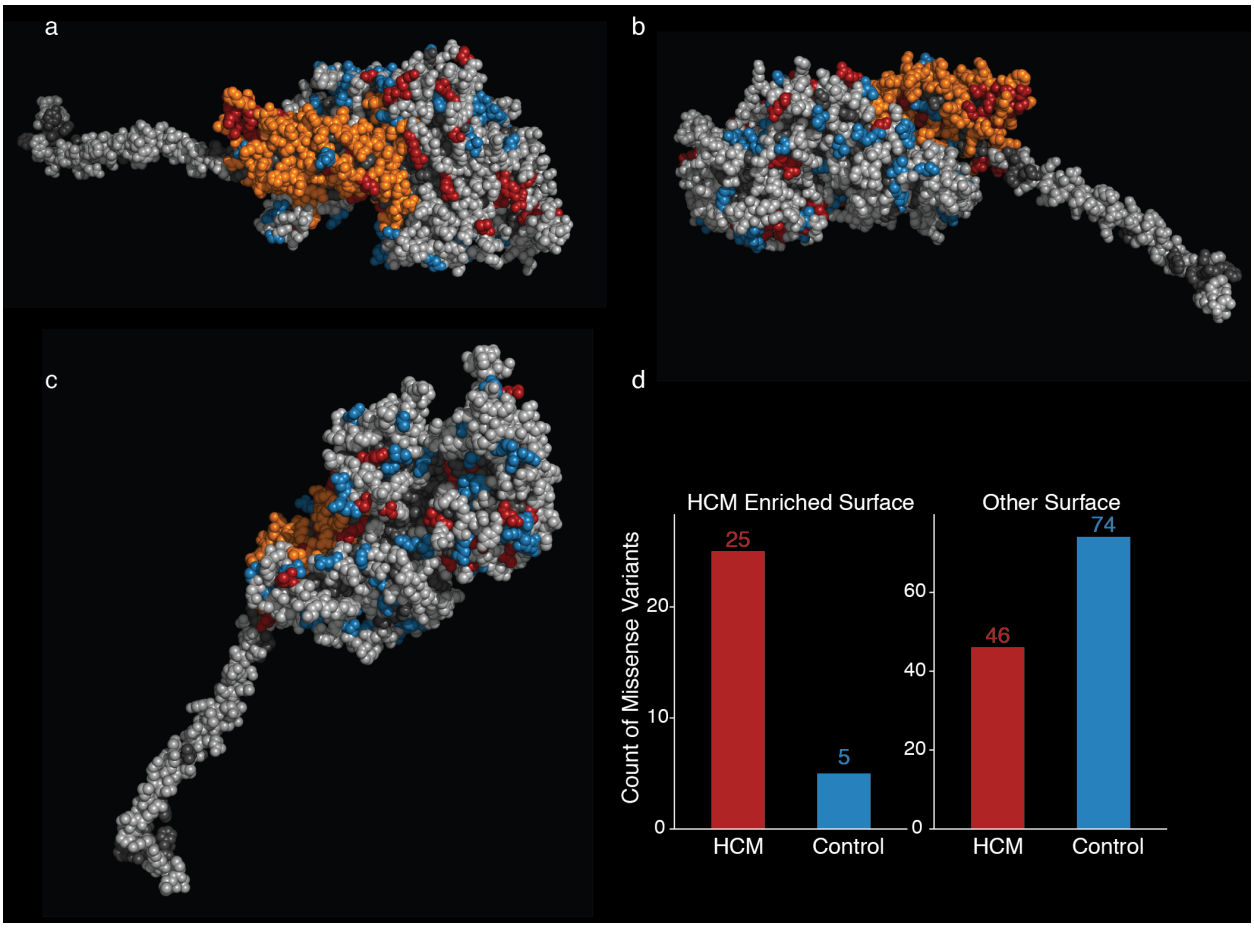

**Supplementary Figure S2. The post-stroke model spherical spatial scan statistic identifies a region of the converter domain as enriched for HCM variants.** (a) The side view of the post-stroke S1 motor domain as shown Figure 2a, without the light chains attached to the S1. The orange residues define a sphere of residues in the motor domain that is the only region significantly-enriched for HCM variants. The S1 residues are colored as follows: orange – enriched region, blue – missense variants seen only in Exac, red – missense HCM variants seen in SHaRe, light grey – all other residues. (b). The enriched region in the post-stroke model of myosin, from a different perspective. Coloring is as in (a). (c). The number of HCM-variants (SHaRe) and non-disease associated (ExAC) variants identified in the spherical enriched region (left) and in the sum of all other parts of the myosin (right).



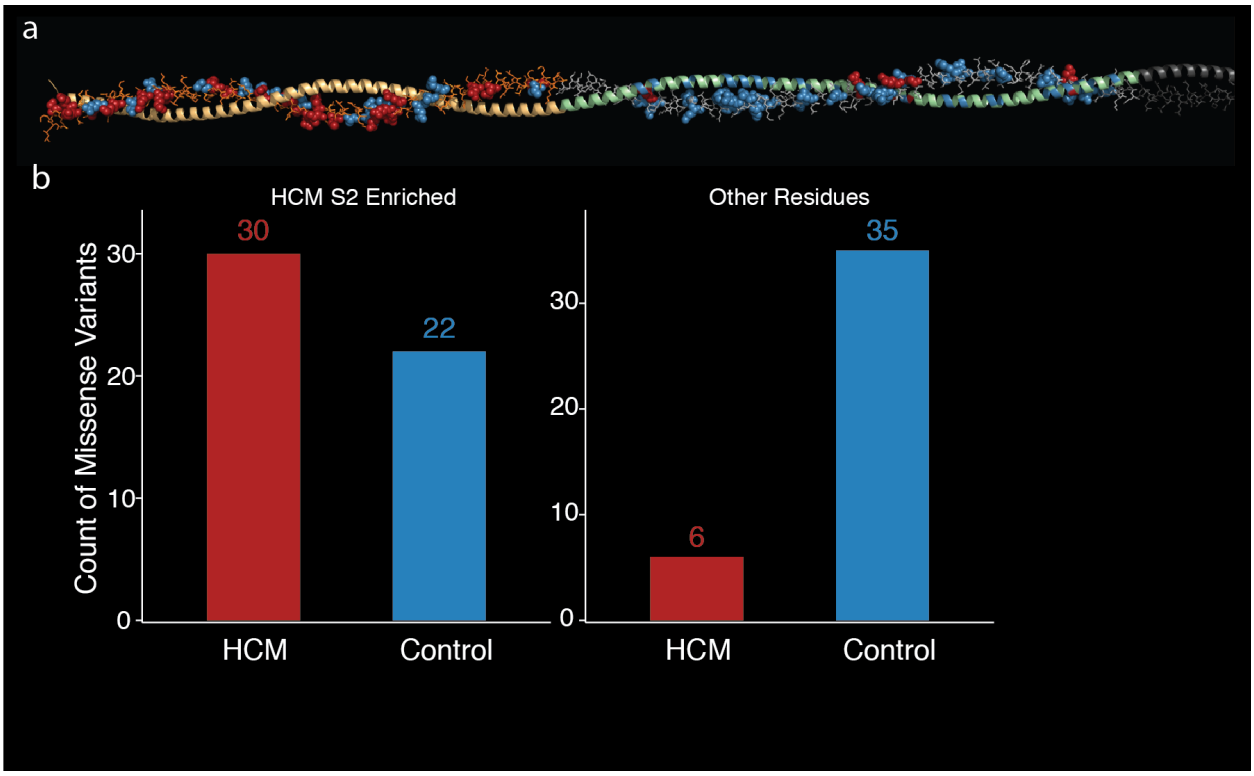

**Supplementary Figure S3. The solvent excluded surface mesh of myosin with the labeled mesa region.** The solvent excluded surface is represented by a network of points in space. This weighted network is used to calculate the surface distance between any two points. The mesh was calculated using the MSMS program with a 2.5 Å surface probe. The mesa region is labeled.



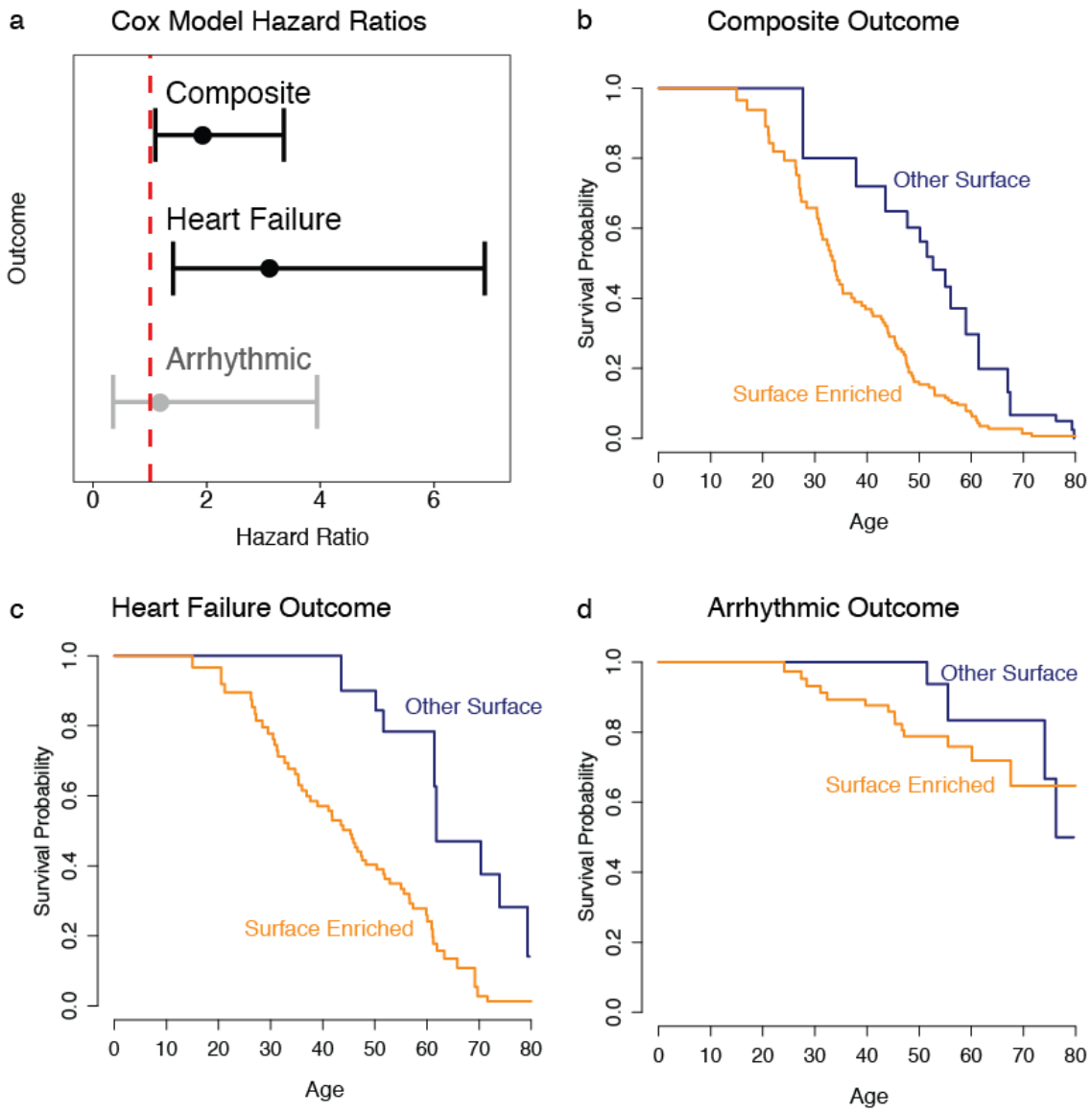

**Supplementary Figure S4. Amino acid surface distances calculated for the myosin head from the solvent excluded surface.** The surface distance in Angstroms between any two amino acid residues in the pre-stroke model of the myosin head, with darker shading indicating residues that are closer together. White lines indicate amino acids not located on the surface. Grey regions indicate distances of greater than 60 Angstroms. Residues near each other in the linear sequence of the protein are often located near each other (dark shading along diagonal). In addition, there are many instances where residues that are far apart in the linear sequence are close together in three-dimensional space (dark shading off the diagonal).



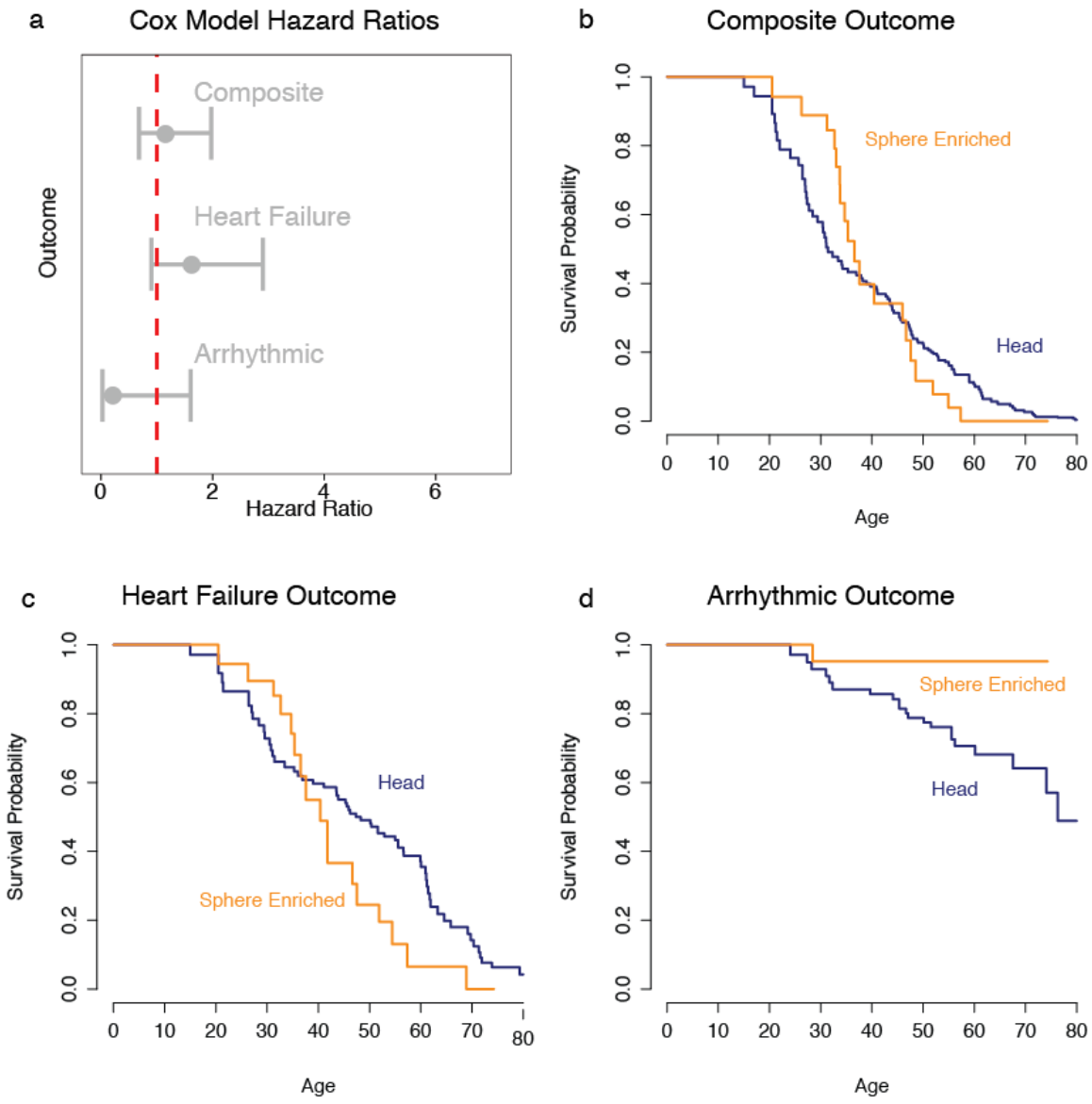

**Supplementary Figure S5. The surface spatial scan statistic identifies a smaller surface region as enriched in the myosin post-stroke model.** (a) The spatial scan statistic identifies an enriched surface region in the post-stroke myosin model that contains many of the residues in the enriched region in the pre-stroke myosin model. In this view, the orange residues define the surface of residues in the motor domain that is the region significantly-enriched for HCM variants. The S1 residues are colored as follows: orange – enriched region, blue – missense variants seen only in Exac, red – missense HCM variants seen in SHaRe, light grey – all other surface residues, dark grey - non-surface residues. (b) A side view of the enriched region. Coloring as in (a). (c) The non-enriched side of the post-stroke myosin, coloring as in (a). (d) The number of HCM-variants (SHaRe) and non-disease associated (ExAC) variants identified in the surface enriched region (left) and in the remaining regions of the myosin surface (right).




**Supplementary Figure S6. Enrichment of HCM residues in the proximal part of myosin S2.** (a) Molecular structure of the coiled-coil S2 fragment, modeled as described in Methods. Only one α-helix of the coiled-coil is shown in detailed structure, the second α-helix is symmetric and shown only in cartoon mode. HCM variants are labeled in red, while ExAC variants are in blue. Orange shading indicates the identified enriched region. (b) The number of HCM-variants (SHaRe) and non-disease associated (ExAC) variants identified in the enriched region of the S2 compared with the remainder of the S2 model.




**Supplementary Figure S7. HCM patients who carry a missense variant in the surface enriched region of *MYH7* have an increased hazard for clinical events.** (a) Hazard Ratios and 95% confidence intervals for the surface enriched region for each of the three outcome classes. (b) Kaplan-Meier survival curve for the composite outcome comparing the two regions. (c) Kaplan-Meier survival curve for the heart failure outcome comparing the two regions. (d) Kaplan-Meier survival curve for the arrhythmic outcome comparing the two regions.




**Supplementary Figure S8. No significant differences in hazard for clinical events between HCM patients who carry a variant in the spherical enriched region and the remainder of the *MYH7* gene.** (a) Hazard Ratios and 95% confidence intervals for the spherical enriched region for each of the three outcome classes. (b) Kaplan-Meier survival curve for the composite outcome comparing the two regions. (c) Kaplan-Meier survival curve for the heart failure outcome comparing the two regions. (d) Kaplan-Meier survival curve for the arrhythmic outcome comparing the two regions.

**Supplementary Information:**

**Supplementary Text 1. Replication Dataset for Genetic Analyses**

To replicate the associations above, we searched the literature for HCM studies that reported the observed *MYH7* variants and that occurred at medical centers not part of the SHaRe consortium. We combined the reported *MYH7* missense variants previously reported in the literature into a single set of unique missense variants [41–51]. We compared this against exome sequencing data of *MYH7* from the DiscovEHR sequencing project involving the Regeneron Genetics Center and Geisinger Health System using the spatial scan statistic. *MYH7* missense variants were extracted from exome sequencing data generated on 42,930 individuals as described. We used the observed set of *MYH7* missense variants identified through this exome sequencing set as a general population cohort. Due to the fact that some literature studies only reported amino acid differences, for the validation analysis we compared unique amino acid changes from each data set.

We tested the enriched regions identified from the SHaRe vs. ExAC analysis for HCM-associated variant enrichment in the validation data. We performed a likelihood ratio test comparing the enrichment of literature HCM-associated variants with the variants found in the DiscovEHR exome population cohort. For each of the enriched regions, we calculated the likelihood ratio test statistic for enrichment for HCM-associated variation. Two times the logarithm of the likelihood ratio statistic follows a chi-square distribution with one degree of freedom, since the alternative model has two parameters while the null model has a single parameter. For the pre-stroke model, both the spherical enriched region ($p = 0.0019$) and the surface enriched region ($p = 2.5 \text{ X } 10\text{^-}5$) validate. For the post-stroke model, both regions are enriched in the validation set as well ($p = 0.00017$ for the surface region and $p = 0.00011$ for the spherical region).

**Supplementary Text 2. Age to event survival analysis.**

We performed an age at event survival analysis comparing individuals with variants in the identified regions versus individuals with other variants in the head of the human β-cardiac myosin molecule. We focused on multiple types of composite clinical outcomes in SHaRe subjects, including combined measures of heart-failure outcomes and arrhythmic events. We used the Kaplan-Meier method to estimate survival curves for two groups of patients: patients carrying variants in the pre-stroke spherical variant enriched region in the converter domain of β-cardiac myosin and those carrying variants located elsewhere in the β-cardiac myosin head. Patients are censored at their last known age and enter the study at their diagnosis age. We also tested for differences in the survival between patients with variants in the enriched region identified on the surface of β-cardiac myosin and patients with variants in other surface amino acids. Using a Cox proportional hazards model, we find no significant differences between individuals with variants in the spherical hit and individuals with variants in other regions of the β-cardiac myosin head. We do find significant differences in the hazard of the heart failure and

composite outcomes between individuals with variants in the surface enriched region and individuals with variants in other regions of the myosin surface.

**Supplementary Text 3. Assessing the effects of VUS and ancestry on the spatial scan results.** To assess whether the observed enrichments of HCM-associated variants are affected by the inclusion of missense variants with less compelling evidence for pathogenicity, we excluded any sites with only variants classified as unknown significance (VUS) and again performed the same spatial scan analysis. For the pre-stroke model, we once again identified the same 15 Angstrom region centered at amino acid residue 736 ($p = 0.003$) as the analysis which included all missense variants. For the surface analysis, we also find the same surface region is identified when excluding all VUS variants ($p < 0.001$). For the rigor surface analysis, we also identify the same surface region as in the full analysis ($p < 0.001$).

Some burden analyses can be strongly confounded by ancestry differences between case and control groups. To test if our results were confounded by population structure, we performed the spatial scan analysis only including variants seen in individuals with European ancestry in both SHaRe and ExAC. This analysis identified a similar enriched region centered at amino acid 749 with a 17.5 Angstrom radius ($p = 0.003$), which included 75% of the residues identified in the previous analyses. For the surface analysis, the same surface region is the most significant hit when including only European variants ($p = 0.007$). For the rigor surface analysis, we also identify the same surface region as the analysis including all of the variants ($p=0.002$). Together, these analyses demonstrate that our analyses are not significantly affected by the inclusion of VUS or by ancestry differences between cases and controls.

**Supplementary Text 4. Assessing the effects of VUS on the clinical phenotype data.**

To assess the effects of the inclusion of individuals with Variants of Unknown significance (VUS) on the clinical phenotype analysis, we performed the same clinical comparisons using only individuals with a pathogenic or likely pathogenic variant. For the spherical enriched region, we find similar results as before, with a 9.79 year difference in diagnosis age (Wilcoxon $p = 3.2 \times 10^{-4}$). For the surface enriched region, there is a 7.96 decrease in age at diagnosis ($p = 0.0062$) when comparing individuals in the surface enriched region vs. those in other surface regions of the myosin molecule.