Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation

Julian R. Homburger, Eric M. Green, Colleen Caleshu, Margaret S. Sunitha, Rebecca E. Taylor, Kathleen M. Rupple, Raghu Prasad Rao Metpally, Steven D. Colan, MichelleMichels, Sharlene M. Day, Iacopo Olivotto, Carlos D. Bustamante, Frederick E. Dewey, Carolyn Y. Ho, James A. Spudich, and Euan A. Ashley

*Molecular motors are the fundamental force-generating elements of muscle contraction. Variation in the human β-cardiac myosin heavy chain gene (MYH7) can lead to hypertrophic cardiomyopathy (HCM), a heritable disease characterized by cardiac hypertrophy, heart failure, and sudden cardiac death. How specific myosin variants alter motor function or clinical expression of disease remains incompletely understood. Here, we combine structural models of myosin from multiple stages of its chemo-mechanical cycle, exome sequencing data from two population cohorts of 60,706 and 42,930 individuals, and genetic and phenotypic data from 2,913 patients with HCM to identify regions of disease enrichment within β-cardiac myosin. We first developed computational models of the human β-cardiac myosin protein before and after the myosin power stroke. Then, using a spatial scan statistic model to analyze genetic variation in protein 3D space, we found significant enrichment of disease-associated variants in the converter, a kinetic domain that transduces force from the catalytic domain to the lever arm to accomplish the power stroke. Focusing our analysis on surface-exposed residues, we identified a larger region significantly enriched for disease-associated variants that contains both the converter domain and residues on a single flat surface on the myosin head described as the myosin mesa. Notably, patients with HCM with variants in the enriched regions have earlier disease onset than patients who have HCM with variants elsewhere. Our study provides a model for integrating protein structure, large-scale genetic sequencing, and detailed phenotypic data to reveal insight into time-shifted protein structures and genetic disease.

Hypertrophic cardiomyopathy | myosin | rare disease genetics | genetic burden

Myosin motors are molecular machines responsible for converting chemical energy into the mechanical force necessary for cell division, directed cell migration, vesicle trafficking, and muscle contraction [1]. Variants in the human β-cardiac myosin heavy chain gene (MYH7), cause hypertrophic cardiomyopathy (HCM), a genetic disease of the heart muscle characterized by an asymmetrical thickening of the ventricular walls and a decrease in the ventricular chamber size. HCM is the most common heritable heart disease, with a prevalence of one in 500 individuals [2]. Clinically, the course of HCM is variable, with some patients experiencing minimal symp-toms and others developing arrhythmia, heart failure, or sudden death [3]. Relationships between genotype and disease expression in HCM have been challenging to establish due to the absence of large-scale genetic population data and lack of multicenter sharing of patient genetic and clinical data [4, 5].

The MYH7 gene is highly constrained for genetic variation [6]. Few loss-of-function variants are observed in population cohorts, and identified pathogenic variants are mainly missense. However, although many missense genetic variants in MYH7 cause HCM, not all genetic changes within the gene lead to disease. There are competing hypotheses regarding the localization of pathogenic missense variants within MYH7, analysis of which can offer insight into the underlying mechanism of HCM. Investigators have suggested enrichment of pathogenic variants in many of the functional domains of β-cardiac myosin, including the converter domain, actin-binding site, and ATP-binding domain [7–9]. However, others have suggested there is no regional enrichment for HCM variation within MYH7 [4, 10]. These inconsistencies could be due to limited sample sizes or to a lack of reference cohorts for comparison. Without information about the natural distribution


Conflict of interest statement: J.A.S. is a founder of and owns shares in CytoKinetix, Inc. and Myokardia, Inc., biotechnology companies that are developing therapeutics that target the sarcomere. E.M.G. is an employee and owns shares in MyoKardia, Inc. E.A.A. is a founder of Personalis, Inc. C.D.B. is on the Scientific Advisory Boards of Ancestry.com, Personalis, Liberty Biosecurity, and Etalon DX, and is also a founder and chair of the Scientific Advisory Board of IdentifyGenomics.

Frees available online through the PNAS open access option.

Significance

Genetic variants in human β-cardiac myosin, which causes muscle contraction in the heart, can lead to hypertrophic cardiomyopathy (HCM), an inherited heart disease that can cause sudden death. New technologies have generated sequence data for large numbers of patients with HCM and unaffected individuals. In this study, we compare the protein structural locations of genetic variants of patients with HCM and the general population to identify spatial regions of the myosin that have a higher than expected proportion of genetic variants associated with HCM and earlier age at diagnosis. In addition, we develop new methods to interrogate the localization of genetic changes in protein structures. Our study demonstrates the power of combining clinical, genetic, and structural data to gain insight into Mendelian disease.

PNAS | June 14, 2016 | vol. 113 | no. 24 | 6701–6706

www.pnas.org/cgi/doi/10.1073/pnas.1606950113

© 2016 National Academy of Sciences

of rare variants within MYH7, it is impossible to distinguish regions of disease-variant enrichment from regions of increased genetic tolerance (11). In addition, some putatively pathogenic variants are later found at higher than expected allele frequencies in large, ethnically diverse population reference cohorts (12). The study suggesting enriched domains lacked reference cohorts; when a reference cohort was compared against genetic variants from a small sample of patients with HCM, the study failed to detect any significant enrichment for disease-associated variants (10). In addition, a focus on the linear sequence of the gene or previously discovered functional domains could overlook novel functional regions or enrichments spanning multiple domains. These discrepancies point to the need to test for regional HCM variant enrichment within MYH7 using both a large patient population and a large reference cohort while also accounting for the 3D structure of β-cardiac myosin.

Recent advances in next-generation sequencing technology have enabled the assembly of large datasets of human genetic variation in both unselected and disease-affected populations. Comparative analysis of these cohorts enables within-gene inference of disease burden and constraint, a measure of population tolerance to variation that can reveal insight into critical functional residues and disease etiology. The Exome Aggregation Consortium (ExAC) (13) and the DiscovEHR (14) cohort are exome sequencing cohorts of unrelated individuals, respectively, that provide valuable information regarding the rates and types of genetic variation seen within disease genes. In addition, the Sarcomeric Human Cardiomyopathy Registry (SHaRe) was established as an international consortium of HCM investigators and currently contains detailed longitudinal clinical data on 2,913 patients with HCM who have undergone genetic testing. As large clinical and population sequencing projects such as these become more prevalent, novel methods for statistical analysis of variant burden and constraint will be essential for gaining insight into disease and identifying intragenic regions enriched from disease-associated variation.

We hypothesized that assessing regional genetic tolerance in the context of time-shifted 3D structures would reveal insights into MYH7 and possible hot spots in the myosin structure of pathogenic variants within HCM. Here, we compare genetic data from the SHaRe, an international HCM registry (15), with variants identified in large-scale exome sequencing projects (13) to identify regions of enrichment for HCM-associated variation within pre- and poststroke structures of β-cardiac myosin. In addition, we develop a general statistical framework based on a modified version of a spatial scan statistic to search for regions of increased disease-associated variation in 3D protein structures and surfaces. Finally, we take advantage of the clinical and phenotypic data in the SHaRe to examine clinical differences between groups of patients based upon variant location within β-cardiac myosin. We demonstrate the power of combining clinical, genetic, and structural data to make inferences regarding disease etiology and 3D structural hotspots for HCM variants.

Results and Discussion

We first compared the linear distribution of missense variants in patients with HCM (SHaRe) with missense variants in a population reference cohort (ExAC). The ExAC cohort contains sequencing information from 60,706 individuals who were part of disease-specific (non-HCM) and population genetic studies, and the SHaRe database contains 2,913 individuals with HCM sequenced for MYH7. Although the ExAC reference contains some pathogenic sarcomere variants and likely some individuals with HCM, it is not enriched for individuals with the disease. We found 192 unique missense variants (in 474 patients with MYH7 variants) in the HCM cohort and 421 unique missense variants in the ExAC database (Dataset S1). In both cohorts, observed missense variants were very rare and the majority were observed only once (Fig. 1), consistent with previous reports of constraint within the MYH7 gene (6). Further, the vast majority of SHaRe patients with a rare variant in MYH7 carry only one such variant (457 of 474 patients), whereas 16 carry two rare missense variants and a single patient carries three different variants.

Both disease (SHaRe) and population/reference (ExAC) variants are nonuniformly distributed throughout the gene, and we find a significant difference in the linear distribution of rare variants between these cohorts [Kolmogorov–Smirnov (KS) test: $P = 5.0 	imes 10^{-11}$; Fig. 1]. Disease-associated missense variants are concentrated in the catalytic globular domain and the coiled-coil S2, consistent with some previous results (4) but in contrast to other recent comparisons (10). Even within these domains, however, distributions of disease and population variants are not the same (KS test: $P = 0.003$). In addition, missense variants in MYH7 in both the SHaRe and ExAC cohorts are extremely rare; the majority in both cohorts are observed only once (Fig. 1 B and C). These results suggest that the likelihood of MYH7 variants causing disease is due, in part, to their location within the gene.

Because molecular motors act in 3D space, we sought a method to investigate patterns of genetic tolerance in the folded structure of human β-cardiac myosin protein. We used multitemplate homology modeling of other myosin proteins in the pre- and poststroke states to build 3D models of human β-cardiac myosin containing the human ventricular light chains (Fig. 2 and Materials and Methods). These models represent two distinct phases of the actin-activated myosin chemomechanical cycle. Four fundamental regions of the myosin motor domain are included: the actin-binding site (Fig. 2, green residues), the ATP-binding pocket (red), the converter domain (blue), and the light chain-binding region or lever arm. In the prestroke state, the converter aligns with a relatively flat surface of the myosin head described as the myosin mesa. Based on its size (>20 nm$^2$), flat topology, and high degree of evolutionary conservation, this feature has been proposed as an interaction site for intra- or intermolecular binding (16). Following the force-producing lever arm stroke of a myosin head, the motor is in its poststroke state (Fig. 24) and the mesa falls out of alignment with the converter domain.

To identify 3D structural regions of interest, we applied a modified version of the spatial scan statistic (17, 18) to the prestroke and poststroke models of β-cardiac myosin S1. This statistic searches for spherical regions with an increased proportion of genetic variants in disease compared with reference cohorts. We defined any rare missense variants observed in patients with HCM in the MYH7 head as disease-associated (n = 110) and any variants seen only in the ExAC cohort as reference (or population) variants (n = 110). Twenty-two variants observed in both the SHaRe and ExAC cohorts were classified as disease-associated. In the myosin prestroke model, we find a striking increase in the proportion of disease-associated missense variation in a 15-A sphere centered
During systolic contraction of the heart, the converter domain serves the critical function of transducing force by swinging about 70° from its prestroke position (Fig. 3B, lever arm projecting outward). Variants in the converter domain have been shown to alter muscle power output and kinetics (19, 20) and have been associated with worse outcomes in HCM (21, 22). We find that the converter domain is the only spherical region significantly enriched for disease variation in contrast to previous reports, which hypothesized that there were many regions of enrichment throughout the myosin head (7–9) or others that found no regions of significant enrichment (10). Although location in an enriched region is not a necessary condition for pathogenicity, novel variants in regions strongly enriched for disease variation should be viewed with increased suspicion. In regions where nearby variants do not have similar effects, there will not be an enrichment of disease-associated variation. Our data provide a complementary line of evidence that variants in part of the β-cardiac myosin converter domain are poorly tolerated and individuals carrying variants in this region are prone to development of HCM.

To replicate these results, we sought an independent source of disease-associated and population genetic variation. We curated publications from medical centers not yet affiliated with the SHaRe registry (Dataset S3 and Supporting Information). We compared this set of 231 unique missense variants with 430 unique missense variants found in 42,930 exomes from unselected individuals in the Geisinger Health System sequenced by the Regeneron Genetics Center (DiscovEHR cohort) (14) (Dataset S4). The converter domain regions identified in both the pre- and poststroke states showed enrichment of disease-associated variants in the replication dataset (prestroke: \( P = 0.0079 \), poststroke: \( P = 4.1 \times 10^{-6} \)).

Surface regions of the β-cardiac myosin have been suggested to be important functional domains implicated in HCM (16). We extended our analysis to examine surface regions of β-cardiac myosin and search for regions enriched for HCM-associated variants. We first defined surface-exposed amino acids by their accessibility to a spherical probe with a radius of 2.5 Å (the approximate size of an amino acid side chain) and approximated the surface distance between any two residues (23) (Supporting Information). As expected, surface distances tended to be shorter between amino acids nearby in the primary sequence (Fig. S2B, diagonal). In addition, there were many regions where amino acids far apart in the primary
The enriched amino acid residues in the poststroke model move the mesa (Wilcoxon test: P = 0.003; Fig. S4). This proximal part of S2 has been shown to bind to the aminoterminal part of myosin-binding protein C (MyBP-C) (26), a sarcomere protein that is also frequently mutated in patients with HCM (27). The enrichment of disease-associated variants in this region suggests that binding between myosin S2 and MyBP-C (and potentially other partners) is important for development of HCM.

To investigate the contribution of the genetically constrained regions to disease further, we compared the clinical features of patients with variants in these regions with the clinical features of patients with variants elsewhere in MYH7. The clinical profile of HCM is highly variable, with some patients living a normal lifespan with minimal symptoms and others dying suddenly or requiring cardiac transplantation at a young age (28). Similarly, age at presentation of HCM varies widely among patients, and earlier onset is correlated with a more severe phenotype (29). We find that HCM patients with a variant inside the spherical enriched region are 11.2 y younger at diagnosis (24.9 (SE = 1.0) vs. 36.1 (SE = 1.2) y old; P = 5.6 \times 10^{-4}; Fig. 3 A and B) than patients harboring other variants in the myosin head. The presence of a variant in the HCM-enriched surface region is associated with a 10.0-y younger age at diagnosis (31.5 (SE = 2.4) vs. 41.5 (SE = 2.6) y old; P = 1.6 \times 10^{-3}) than in those patients with other surface variants (Fig. 3 C and D). In addition, we find an increased hazard for clinical outcomes in the surface enriched region after adjusting for differences in age at diagnosis (hazard ratio = 1.918, P = 0.023), although not in the spherical enriched region (Fig. S5 and Supporting Information). These findings demonstrate that analysis of genetic constraint in protein space can reveal domains with both increased disease burden and pathogenicity.

Our study demonstrates the power of integrating detailed structural information with large clinical and genetic databases to identify regions associated with functional importance and disease severity in Mendelian diseases. We find that variants associated with HCM are enriched in the \(\beta\)-cardiac myosin converter domain, where they lead to more severe outcomes. We provide the first evidence, to our knowledge, that similar clustering and pathogenicity are present in a surface spanning the converter domain and the relatively flat surface of the myosin catalytic domain called the myosin mesa (16). Because amino acid residues forming the mesa come from disparate locations in the nucleotide sequence, discovery of this region depends on integration of protein structural information. The pronounced shift of the converter/mesa surface during the power stroke raises the mechanistic hypothesis that these variants exert their deleterious effect selectively in the prestroke state, perhaps by disrupting dynamic binding interactions. In summary, these findings highlight the importance of considering data from human genetics in the context of the dynamic 3D protein structure, and illustrate an approach to structure/function analysis in genetic diseases.

Materials and Methods

\(\textit{SHaRe Database.}\) The \(\textit{SHaRe}\) is a multicenter database that pools deidentified patient-level data from established institutional datasets at participating sites. At the time of analysis, the registry contained clinical and genetic testing data on 2,913 patients with HCM, 514 with variants in \(\text{MYH7}\). This database contains individuals from nine inherited disease centers throughout the world and includes demographic data, medical history, echocardiogram data, genetic testing results, and many other data relating to cardiac health and clinical outcomes.

---

**Fig. 4.** Surface spatial scan analysis identifies a larger surface region enriched for HCM-associated missense variation. (A) Similar view of the prestroke model to the view in Fig. 2B, looking directly down onto the mesa. The residues are colored as follows: orange, surface region enriched for HCM variants; blue, missense variants seen only in the ExAC; red, missense variants seen in the \(\text{SHaRe}\); light gray, residues considered to be on the surface; dark gray, residues not considered to be on the surface. The HCM-enriched surface region identified covers the entire mesa plus the adjoining converter domain. (B) Same side view of the prestroke S1 motor domain as shown in Figs. 1 and 2A. (C) View of the prestroke model viewing the side opposite the mesa. This surface is not enriched for HCM variants. (D) Number of HCM variants (\(\text{SHaRe}\)) and non-disease-associated (ExAC) variants identified in the surface enriched region (Left) and in the sum of all other parts of the myosin surface (Right).
greater than the incidence in the general population. A large proportion of the HCM in the... variants for... cannot assume variants identified in the ExAC are not pathogenic. For example,... variants only in the ExAC population were considered to be disease-associated, regardless of their presence in the ExAC population. Rare missense variants identified only in the ExAC population were considered to be disease-associated. Within the MYH7 S1 head, we identified 103 disease-associated variants (seen in the SHARe and 110 reference (or population) variants seen only in the ExAC cohort. Twenty-two variants observed in both the SHARe and ExAC cohorts were classified as disease-associated. Enrichment testing was performed by searching for regions of the myosin protein that had a significantly higher proportion of disease-associated variants than non-disease-associated variants. The ExAC population likely contains some individuals with cardiomyopathy induced by changes in MYH7; therefore, we cannot assume variants identified in the ExAC are not pathogenic. For example, there are multiple individuals carrying a well-known pathogenic variant for HCM in the MYBPC3 gene, p.Arg502Trp, in the ExAC cohort. However, the incidence of cardiomyopathies in the ExAC population is not expected to be greater than the incidence in the general population. A large proportion of the rare variants identified in MYH7 in the ExAC cohort are likely not associated with cardiomyopathy. In contrast, rare variants identified in SHARe HCM patients are likely enriched for disease-causing variation.

We calculated the maximum population-specific allele frequency for each variant in the ExAC as the maximum frequency of the variant in any of the ExAC populations (excluding Finnish). To compare the burden of rare variants within each cohort, we removed from both sets (SHARe and ExAC) any allele with a population-specific allele frequency higher than 1:2,000. This filter removed 16 variants in MYH7 from our ExAC list (of 421 total variants), six of which were also reported in the SHARe list (of 192 total variants). Within the MYH7 head, only four ExAC variants were found at a higher frequency than 1,200 (of 136 total variants), two of which were also seen in the SHARe cohort (of 105 total variants).

For validation purposes, we also performed a subset of analyses including variants classified as “pathogenic” or “likely pathogenic” according to the American College of Medical Genetics and Genomics guidelines (30), although excluding variants of unknown significance, likely benign, and benign variants (Supporting Information). In addition, we performed a validation analysis using variants found in individuals of European ancestry from the ExAC and individuals with a reported race of white to ensure that global population structure was not confounding our analysis (Supporting Information).

We generated an independent validation dataset combining previously published analyses of HCM variants from other medical centers with 42,930 exomes from the DiscovEHR sequencing project (14). We included only missense variants and removed variants with an allele frequency greater than 1 in 2,000 in the DiscovEHR exomes (Supporting Information).

Development of Human β-Cardiac Myosin Protein Models. We developed human β-cardiac myosin S1 models based on the known motor domain structural data to represent the human form of the cardiac myosin. We retrieved the protein sequence of human β-cardiac myosin and the human cardiac light chains from the UniProt database (31): myosin heavy chain motor domain (MYH7), P12883; myosin essential light chain (MLC1), P08590; and myosin regulatory light chain (MLC2), P19016. We used a multitemplate homology modeling approach to build the structural coordinates of MYH7 (residues 1–840), MLC1 (residues 1–195), MLC2 (residues 1–166), and S2 (residues 841–1,280). We obtained the 3D structural model of S1 in the pre- and post-stroke states by integrating the known structural data from solved crystal structures (details are provided in Supporting Information). Homology modeling of the pre-stroke structure was performed with template structures of the smooth muscle myosin motor domain (32) [Protein Data Bank (PDB) ID code 1BR1; with MgADP.AIF4 bound at the active site, which is thought to mimic most closely the ADP.Pi-bound state, or pre-stroke state, of the myosin] and the scallop smooth muscle myosin light chain domain (33) (PDB ID code 3TSS). The templates used for the modeling of the post-stroke structure were obtained from the human β-cardiac myosin motor domain (25, 34) (PDB ID code 4P7H; no nucleotide in the active site) and the rigor structure from the squid myosin motor domain (35) (PDB ID code 3ISG; no nucleotide in the active site). Modeling was done using the MODELER package (36). Visualizations were performed using PyMOL version 1.7.4 (www.pymol.org).

Statistical Methods. Comparisons between the ExAC and SHARe variant locations in MYH7 were performed using the KS test statistic. All statistical analyses were performed in R version 3.1.2 (37), and many graphs were prepared using ggplot2 (38).

Spatial Scan Statistic. For the spatial scan analysis, we compared the locations of unique HCM-associated variants with the locations of reference variants. The spatial scan statistic exhaustively searches 3D windows of a predefined set of sizes and shapes throughout the human β-cardiac myosin molecule for regions with an increased proportion of HCM-associated variants. It identifies the 3D region of the protein with the largest binomial likelihood ratio test of enrichment of HCM-associated variants (17, 18) (Supporting Information). The maximum binomial ratio test statistic for the entire set of windows in the model was calculated, and significance was assessed through permutation of variant labels using 1,000 permutations. For validation, we performed the analysis above, removing all missense variants classified as variants of unknown significance using only missense variants observed in the European individuals (Supporting Information).

Surface Analysis. In addition to the spherical windows defined above, we define windows based upon the exposed surface of the human β-cardiac myosin molecule. For this analysis, we estimate the surface distance between any two amino acids based upon the solved crystal structure and surface calculations (Supporting Information). Based upon this set of pairwise calculated distances, we define planar surface regions of the β-cardiac myosin for analysis as all of the amino acids within a certain distance of any given “center” amino acid.
We exclude nonsurface amino acids, based upon their depth. We once again perform the spatial scan statistic as described above to identify surface regions of increased genetic burden.

S2 Fragment Analysis. We performed the spherical spatial scan statistic as described above to test for enrichment in the S2 fragment of myosin, including the region between amino acids 838 and 1,112 (Supporting Information).

Clinical Phenotype Analysis. For the analysis of age at diagnosis, only known probands were included (n = 260 myosin head, n = 201 myosin surface). Patients with multiple MYH7 variants were excluded from the age at diagnosis analysis, because the presence of multiple variants causes an earlier onset of HCM. We compared the primary diagnosis ages using a Wilcoxon test. The overall composite outcome combined the arrhythmic and heart failure outcomes, as well as the outcomes of atrial fibrillation, stroke, and all-cause death (Supporting Information). Individuals were considered to enter the study at their diagnosis age and were censored at their last known age. We compared hazard ratios for each region using the Cox proportional hazards model, adjusting for gender.

ACKNOWLEDGMENTS. We thank the investigators and participants in the SHaKe for their support of this project. We also thank Jonathan Fox for guidance during the early stages of the SHaKe. J.R.H. is supported by a Stanford Graduate Fellowship. R.E.T. is supported by NIH Grant F32 HL123247. This work supported by NIH Grants U01HG007436 (to C.D.B.), U01HG007708 (to E.A.A.), and U01HG006382 (to R.P.R.M.), and a grant from SAP (to C.D.B.), and a grant from the Autism Research Institute, to R.P.R.M., for the Autism Sequencing Consortium.

Spatial Scan Statistic

For the spatial scan analysis, we compared the locations of unique HCM-associated variants with the locations of reference variants in the myosin S1 subfragment. The spatial scan statistic exhaustively searches 3D windows of a predefined set of sizes and shapes throughout the human β-cardiac myosin molecule for regions that have an increased proportion of HCM-associated (SHaRe) variants. It identifies the 3D region of the protein with the largest binomial likelihood ratio test of enrichment of HCM-associated variants (17, 18).

Let $p_w$ be the proportion of variation within a window that is HCM-associated, and let $q_w$ be the proportion of variation outside the window, which is HCM-associated. For each window, we calculate the binomial likelihood ratio statistic comparing the null model, where $p_w$ and $q_w$ both equal the overall rate $r$ against the model where $p_w$ is not equal to $q_w$. The likelihood ratio statistic ($\Lambda_w$) for each window is as follows:

$$\ln(\Lambda_w) = \ln\left[\frac{p_w^r (1-p_w)^{y_w}}{(1-r)^{y_w}}\right] - \ln\left[\frac{p_r^r (1-p_r)^{y_r}}{(1-r)^{y_r}}\right],$$

where $y_w$ is the number of HCM-associated variants within the window, $y_r$ is the number of other variants within the window, $y_w$ is the number of HCM-associated variants outside the window, and $r$ is the overall proportion of HCM-associated variants. The test statistic is the maximum of the observed likelihood ratio statistic (maximum value of $\Lambda_w$) for all windows tested over the protein. Significance is assessed through permutation analysis by permuting the variant labels.

For the first analysis, we used spherical windows based upon the 3D locations of amino acids in the human β-cardiac myosin molecule. The myosin S1 models include residues 1–841. We excluded the loop regions between residues 205 and 211 and residues 627 and 640, because the positions of these amino acid residues are not well defined. We generated lists of unique missense variants associated with HCM and reference variants. Variants were assigned the 3D location in the human β-cardiac myosin model of the α-carbon atom of their corresponding residue. For acid residue in the myosin S1, we tested spherical windows with radii of 10, 12.5, 15, 17.5, 20, 22.5, and 25 Å centered on the α-carbon for enrichment of HCM-associated variants.

The maximum binomial ratio test statistic for the entire set of windows in the model was calculated, and significance was assessed through permutation of variant labels using 1,000 permutations.

Myosin S2 Spatial Scan Analysis

To assess enrichment in the S2 fragment of myosin, we performed a similar spatial analysis to the one performed on the S1 fragment. We used spherical window sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 Å centered on the α-carbon of each amino acid residue in the S2 model. For this analysis, we compared the region between amino acid residues 838 and 1,112. The analysis was truncated at residue 1,112 due to low sequencing coverage following residue 1,112 in the ExAC data.

Protein Modeling

We used homology modeling to determine the prestroke and poststroke structures of the β-cardiac myosin.

Homology modeling of the prestroke structure was performed with template structures of the smooth muscle myosin motor domain (32) (PDB ID code 1BR1, with MgADP.Aif4 bound at the active site, which is thought to mimic most closely the ADP.Pi bound state, or prestroke state, of the myosin) and the scallop smooth muscle myosin light chain domain (33) (PDB ID code 3TS5). The templates used for the modeling of the poststroke structure were obtained from the human β-cardiac myosin motor domain (34) (PDB ID code 4P7H, no nucleotide in the active site) and the rigor structure from the squid myosin motor domain (35) (PDB ID code 35G, no nucleotide in the active site). Missing regions in the myosin motor domain (loop1, loop2) were each separately built using the ModLoop program (39), and regions in the regulatory light chains that were not solved in the crystal structures were independently modeled using the I-TASSER prediction (40) method, and they were used as individual templates. Sequence alignment between MYH7, MLC1, and MLC2 with their respective structural templates was obtained. The models of pre- and poststroke structures were acquired using the MODELER package (36). We used a multitemplate modeling method: A total of 100 models were obtained, and the best model was selected based on the discrete optimized protein energy score. The structural models were energy-minimized using SYBYL7.2 (Tripos, Inc.) to remove potential short contacts. The final 3D models of the pre- and poststroke structures were validated using RAMPAGE (41), which provides a detailed check on the stereochemistry of the protein structure using the Ramachandran map.

The S2 region is a long coiled-coil structure; hence, we used the template from the Myosinome database (42). Modeling was done using the MODELER package, wherein selection of best model and validation of the model were done as described above. Visualizations were performed using PyMOL version 1.7.4 (www.pymol.org).

Estimation of Surface Distances

We developed the following procedure to estimate the surface distances between any two amino acids. First, we calculated the solvent-excluded surface for the human β-cardiac myosin models using the MSMS program with a 2.5-Å radius sphere as a probe, which approximates the size of amino acid side-chain interactions. For the calculated surface, the MSMS program returns a large net of vertices of the surface, each connected to multiple other points on the surface. We use these vertices and connections to build a weighted graph of the surface of the molecule, with the vertices as nodes and the edges as connections weighted by the Euclidean distance between the two connected vertices. Then, for each amino acid, we assign it to the point on the surface that is the closest point, on average, to all of the non-hydrogen atoms of the amino acid. This average distance is used as an estimate of the depth of the amino acid. Amino acids with average depths of greater than 4 Å for the surface model were considered not to be on the surface. We used the A-star algorithm to calculate the distance on the surface graph between two amino acid residues. In this analysis, we included only amino acids in the “head” of the human β-cardiac myosin molecule, defined as amino acids 1–784. The lever arm was excluded. We excluded the disordered loop regions between residues 205 and 211 and residues 627 and 640.

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 (43–52). We compared this set 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, we compared unique amino acid changes from each dataset for the validation analysis.

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 $\chi^2$ distribution with 1 df, because the alternative model has two parameters, whereas the null model has a single parameter. For the poststroke model, both regions are enriched in the validation set as well ($P = 0.0082$ for the surface region and $P = 0.00041$ for the spherical region).

**Age at Event Survival Analysis**

We performed an age at event survival analysis comparing individuals with variants in the identified regions vs. 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. For the age at event analyses, the composite outcomes were defined as follows. The arrhythmic heart failure outcome consisted of cardiac arrest, implantable cardioverter-defibrillator firing, and sudden cardiac death. The heart failure outcome combined the events of end-stage HCM (defined as the left ventricular ejection fraction falling below 55%), New York Heart Association class III or IV status, transplant operation, and left ventricular assistive device implantation. The overall composite outcome combined the arrhythmic and heart failure outcomes, as well as the outcomes of atrial fibrillation, stroke, and death (all causes).

We used the Kaplan–Meier method to estimate survival curves for two groups of patients: patients carrying variants in the prestroke spherical enriched region in the converter domain of β-cardiac myosin and those patients 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.

**Assessing the Effects of Variants of Unknown Significance 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 of unknown significance (VUS) according to the American College of Medical Genetics and Genomics guidelines and again performed the same spatial scan analysis. Within this framework, variants are classified as one of five categories: pathogenic, likely pathogenic, of unknown significance, likely benign, and benign. Within MYH7 in the SHaRe, there were 48 pathogenic, 82 likely pathogenic, 66 of unknown significance, 2 likely benign, and 0 benign variants. Within the MYH7 S1 head domain, there were 33 pathogenic, 48 likely pathogenic, 27 of unknown significance, 1 likely benign variant, and 0 benign variants. This assessment is to ensure that VUS are not driving the results of the analyses. Variant classifications were provided by each site and were made by the site’s clinical team or the clinical genetic testing laboratory that performed the test.

For the prestroke model, we once again identified the same 15-Å region centered at amino acid residue 736 ($P = 0.003$) as in the analysis that 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 reference 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 the SHaRe and ExAC. This analysis identified a similar enriched region centered at amino acid 749 with a 17.5-Å 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 ($P = 0.002$). Both the SHaRe and ExAC contain a large majority of individuals with European ancestry. As such, we currently do not have the power to perform this analysis in a subset of individuals from an additional population. As genetic datasets expand further in size, we hope to perform this analysis in other populations to identify additional regions of enrichment or to refine the enriched regions identified here.

Together, these analyses demonstrate that our analyses are not significantly affected by the inclusion of VUS or by ancestry differences between patients with HCM and population references.

**Assessing the Effects of Inclusion of Variants of Unknown Significance on the Clinical Phenotype Data**

To assess the effects of the exclusion of individuals with VUS on the clinical phenotype analysis, we performed the same clinical comparisons using only individuals with a pathogenic or likely pathogenic variant. Any variants with a frequency above 1/2,000 in the ExAC were excluded from the analysis. For the spherical enriched region, we find similar results as before, with a 9.79-y difference in diagnosis age ($P = 0.0036$) when comparing individuals with VUS in the surface enriched region vs. VUS in other surface regions of the myosin molecule.
Fig. S1. Poststroke model spherical spatial scan statistic identifies a region of the converter domain as enriched for HCM variants. (A) Side view of the poststroke S1 motor domain as shown in Fig. 2A, without the light chains attached to the S1. The orange residues define a sphere of residues in the motor domain, which 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 the ExAC; dark red, missense HCM variants seen in the SHaRe; light gray, all other residues. (B) Enriched region in the poststroke model of myosin, from a different perspective; coloring is as in A. (C) 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).
Fig. S2. Solvent-excluded surface mesh of myosin with the labeled mesa region and amino acid surface distances calculated for the myosin head. (A) 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. (B) Surface distance in angstroms between any two amino acid residues in the prestroke model of the myosin head, with darker shading indicating residues that are closer together. White lines indicate amino acids not located on the surface. Gray regions indicate distances of greater than 60 Å. 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 3D space (dark shading off the diagonal).
Fig. S3. Surface spatial scan statistic identifies a smaller surface region as enriched in the myosin poststroke model. (A) Spatial scan statistic identifies an enriched surface region in the poststroke myosin model that contains many of the residues in the enriched region in the prestroke myosin model. In this view, the orange residues define the surface of residues in the motor domain, which is the region significantly enriched for HCM variants. The S1 residues are colored as follows: orange, enriched region; blue, missense variants seen only in the ExAC; red, missense HCM variants seen in the SHaRe; light gray, all other surface residues; dark gray, nonsurface residues. (B) Side view of the enriched region; coloring is as in A. (C) Nonenriched side of the poststroke myosin; coloring is as in A. (D) 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).

Fig. S4. Enrichment of HCM residues in the proximal part of myosin S2. (A) Molecular structure of the coiled-coil S2 fragment, modeled as described in Materials and Methods. Only one α-helix of the coiled-coil is shown in detailed structure; the second α-helix is symmetrical and shown only in cartoon mode. HCM variants are labeled in red, whereas ExAC variants are labeled in blue. Orange shading indicates the identified enriched region. (B) 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.
Patients with HCN 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. (E) Hazard ratios and 95% confidence intervals for the spherical enriched region for each of the three outcome classes. (F) Kaplan–Meier survival curve for the composite outcome comparing the two regions. (G) Kaplan–Meier survival curve for the heart failure outcome comparing the two regions. (H) Kaplan–Meier survival curve for the arrhythmic outcome comparing the two regions.
Dataset S1. Missense variants found in SHaRe patients with HCM and from the ExAC

Missense variants in MYH7 identified in the SHaRe HCM patients and from the ExAC.

Dataset S2. Enriched regions identified in the myosin protein

Amino acids located in each of the regions enriched for HCM-associated variants. NA indicates regions that are in the disordered loops of the myosin or amino acids that were not on the surface in the surface analysis of the models.

Dataset S3. Missense variants identified through literature search

Missense variants were identified through an extensive literature search.

Dataset S4. Rare missense variants in β-cardiac myosin identified in DiscovEHR exomes

Unique missense variants were identified in the MYH7 gene in the Geisinger exome sequencing study.