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, Michelle Michels, Sharlene M. Day, Iacopo Olivotto, Carlos D. Bustamante, Frederick E. Dewey, Carolyn Y. Ho, James A. Spudich, and Euan A. Ashley

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305; MyoKardia, Inc., South San Francisco, CA 94080; Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, CA 94305; Institute for Stem Cell Biology and Regenerative Medicine, Bangalore 560065, India; Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305; Department of Pediatrics (Cardiology), Stanford University School of Medicine, Stanford, CA 94305; Collaborative Bioinformatics Lab, Geisinger Clinic, Danville, PA 17822; Department of Cardiology, Boston Children's Hospital, Boston, MA 02115; Department of Cardiology, Erasmus Medical Center, 3015 CE Rotterdam, The Netherlands; Cardiovascular Division, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109; Referral Center for Cardiomyopathies, Careggi University Hospital, Florence, Italy 50134; Department of Biomedical Data Sciences, Stanford University, Stanford, CA 94305; Regeneron Genetics Center, Tarrytown, NY 10591; Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115.

Contributed by James A. Spudich, May 5, 2016 (sent for review March 31, 2016; reviewed by Elizabeth McNally and David D. Thomas)

Myosin 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 chemomechanical cycle, exome sequencing, and detailed phenotypic data to reveal insight into the relationship between structure, function, and disease expression in HCM.

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.


Conflict of interest statement: J.A.S. is a founder of and owns shares in Cytkinetics, 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 Board of Ancestry.com, Personalis, Liberty Biosecurity, and Etalon DX, and also a founder and chair of the Scientific Advisory Board of IdentifyGenomics.

Freely available online through the PNAS open access option.

1J.A.S. and E.A.A. contributed equally to this work.
2To whom correspondence may be addressed. Email: jspudich@stanford.edu or euan@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606950113/-/DCSupplemental.
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 the general population, respectively, that provide clinical 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 novel 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: \( P = 5.0 \times 10^{-11} \); 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: \( P = 0.005 \)). 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 nm²), flat topology, and high degree of evolutionary conservation, this feature has been proposed as an interaction site for intra- and 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 (\( n = 103 \)) 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-Å sphere centered...
on residue 736 ($P = 0.001$) (Fig. 34 and Dataset S2). This region, covering a subset of the converter domain, contains 17 missense variants observed in disease and no missense variants observed in reference data (Fig. 3C). Using the poststroke model of β-cardiac myosin, we again observed enrichment of disease-associated variants in a portion of the converter domain ($P < 0.001$) centered on residue 733 (Fig. S1).

Enrichment of disease-associated variants in both the pre- and poststroke states persists when including only variants formally classified as pathogenic or likely pathogenic (26 disease-associated of 103 total variants in the SHaRe; Supporting Information). Burden tests can also be sensitive to differences in the population prevalence of disease or differences in the population frequency of variants. To ensure that our results were robust to these potential biases, we performed the same analysis limited to individuals of European descent, the largest population in both the SHaRe and ExAC cohorts (2,519 of 2,913 patients in the SHaRe and 89 of 103 disease-associated variants; Supporting Information). We identified very similar enriched regions in both the pre- and poststroke models, indicating that our analysis is not being significantly affected by population stratification. However, large cohorts from other populations may help identify other regions of the myosin enriched for disease variation.

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 prestroke and poststroke states showed enrichment of disease-associated variants in the replication dataset (prestroke: $P = 0.0079$, poststroke: $P = 4.1 \times 10^{-4}$).

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
sequence were close together on the surface of the β-cardiac myosin (e.g., amino acids near residues 110 and 684).

The surface of β-cardiac myosin contains 568 of the 765 residues in the S1 domain (74%). Of these residues, 71 are associated with HCM (72% of all HCM variants) and 79 are found in the ExAC population (71% of all reference variants), suggesting that variants in both cohorts are evenly distributed between the surface and core of the protein (χ²: P = 0.51). We then applied our spatial scan statistic to the surface of β-cardiac myosin. Using the myosin pre-stroke model, we identified a region of the surface covering 277 of the 568 surface amino acid residues (P = 0.002; Fig. 4 A and B), including the converter domain and the mesa, which is highly enriched for disease-associated variation. Strikingly, the region contains 52 of the 71 surface HCM-associated missense variants and only 27 of the 79 surface non–disease-associated missense variants (Fig. 4D), whereas the remainder of the myosin surface (Fig. 4C) covers 291 residues and contains only 19 disease-associated variants compared with 52 non–disease-associated variants (Fig. 4D). The identified converter/mesa region was also enriched in the replication dataset (P = 4.3 × 10⁻⁵).

Using the same procedure to search the surface regions of the myosin poststroke structure, we detected a smaller enriched region of 122 amino acid residues again covering the converter but with a reduced portion of the myosin mesa (Fig. S3), a domain that has been proposed to be a binding site for another protein structural element (16). During the myosin power stroke, the converter moves away from the mesa (compare Figs. 2 A and B, 2B, and 4 and Fig. S3), so the enriched converter/mesa region is no longer contiguous and available for intra- or intermolecular interactions in the poststroke state (Fig. 2B and Fig. 4 A and B).

The enriched amino acid residues in the poststroke model move significantly more in 3D space between the pre- and poststroke models (Wilcoxon test: P < 2 × 10⁻¹⁶) than other amino acid residues on the surface of the protein. This result suggests that myosin conformational changes during the actin-activated chemomechanical cycle may be important not only for transducing force but also for modulating the size and shape of this surface region and altering its availability for binding to other protein structural elements. The functional importance of the converter and the mesa region is further supported by the presence of the binding site for omecamtiv mecarbil, a recently described small-molecule modulator of cardiac myosin currently in clinical trials for the treatment of heart failure (24, 25).

Next, we tested the myosin S2 fragment for regions enriched for disease-associated variation. The spatial scan analysis revealed that the first half of the S2 fragment is enriched for disease variants (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 × 10⁻⁷; 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 × 10⁻⁸] than in those patients with other surface variants (Fig. 5 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 β-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

**SHaRe Database.** The SHaRe is a multicenter database that pools unidentified 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 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.
Rare variants identified in the ExAC population is not expected to be among individuals carrying a well-known pathogenic variant for 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); P06590; 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 poststroke states by integrating the known structural data from solved crystal structures (details are provided in Supporting Information). Homology modeling of the prestroke 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 poststroke 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 poststroke 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 acid residues on the solvent-excluded surface calculated in a 10 Å sliding window (23) (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.

Homburger et al.
We exclude nonsurface amino acids, based upon their depth. We once again with multiple S2 Fragment Analysis. of increased genetic burden.

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 com- pose outcome combined the arrhythmic and heart failure outcomes, as well as the terminologies of atrial fibrillation, stroke, and all-cause death. 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 SHaRe for their support of this project. We also thank Jonathan Fox for guidance during the early stages of the SHaRe. 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.).