Population-Based Variation in Cardiomyopathy Genes

Jessica R. Golbus, BA; Megan J. Puckelwartz, PhD; John P. Fahrenbach, PhD;
Lisa M. Dellefave-Castillo, MS, CGC; Don Wolfgeher, BS; Elizabeth M. McNally, MD, PhD

Background—Hypertrophic cardiomyopathy and dilated cardiomyopathy arise from mutations in genes encoding sarcomere proteins including *MYH7*, *MYBPC3*, and *TTN*. Genetic diagnosis of cardiomyopathy relies on complete sequencing of the gene coding regions, and most pathogenic variation is rare. The 1000 Genomes Project is an ongoing consortium designed to deliver whole genome sequence information from an ethnically diverse population and, therefore, is a rich source to determine both common and rare genetic variants.

Methods and Results—We queried the 1000 Genomes Project database of 1092 individuals for exonic variants within 3 sarcomere genes *MHY7*, *MYBPC3*, and *TTN*. We focused our analysis on protein-altering variation, including nonsynonymous single nucleotide polymorphisms, insertion/deletion polymorphisms, or splice site altering variants. We identified known and predicted pathogenic variation in *MYBPC3* and *MYH7* at a higher frequency than what would be expected based on the known prevalence of cardiomyopathy. We also found substantial variation, including protein-disrupting sequences, in *TTN*.

Conclusions—Cardiomyopathy is a genetically heterogeneous disorder caused by mutations in multiple genes. The frequency of predicted pathogenic protein-altering variation in cardiomyopathy genes suggests that many of these variants may be insufficient to cause disease on their own but may modify phenotype in a genetically susceptible host. This is suggested by the high prevalence of *TTN* insertion/deletions in the 1000 Genomes Project cohort. Given the possibility of additional genetic variants that modify the phenotype of a primary driver mutation, broad-based genetic testing should be employed. **(*Circ Cardiovasc Genet*. 2012;5:391-399.)**

Key Words: cardiomyopathy ■ hypertrophic cardiomyopathy ■ myosin heavy chain ■ myosin-binding subunit ■ titin

Cardiomyopathy is a major cause of heart failure, and hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM) each have a significant heritable component.[1] Although debated, the estimated prevalence of HCM in the general population is 1:500.[2–5] DCM, classified as left ventricular dilatation and systolic dysfunction in the absence of other causes of cardiomyopathy, has an estimated prevalence of 1:2500.[5–7] Mutations in *MHY7*, the gene encoding β-myosin heavy chain, and *MYBPC3*, the gene encoding myosin-binding protein C, cause both HCM and DCM.[1,3,5] Mutations in the *TTN* gene, encoding the giant protein titin, have been associated with multiple forms of cardiomyopathy, including DCM and arrhythmogenic right ventricular cardiomyopathy.[8-10] The number and size of cardiomyopathy genes have limited the sensitivity of genetic testing since most testing has not included *TTN*. Improvements in sequencing technology now facilitate access to cardiomyopathy genes, including *TTN* with its 363 exons.[11]

Clinical Perspective on p 399

The genetic diagnosis of cardiomyopathy relies on completely sequencing the coding regions of cardiomyopathy genes since most pathogenic variation is rare or private. Common variation is atypical in inherited cardiomyopathy, although common variants may modify disease phenotype. Genomic variants are defined as rare if present at a minor allele frequency (MAF) of <0.5% or as low frequency if present at an MAF of 0.5% to 5%.[12] Since allele frequencies differ among ethnic and racial groups, interpretation of individual genetic variation is limited by the ethnic makeup of available genome databases. The 1000 Genomes Project is an ongoing consortium designed to deliver whole genome sequence from an ethnically varied population, with the goal of making publically available the genomes of 2500 individuals. In 2010, the 1000 Genomes pilot project was published, providing data from low-coverage (≈2-fold to 6-fold) whole-genome sequencing of 179 individuals, high-coverage (average 42-fold) whole-genome sequencing of 6 individuals in 2 trios, and exon-targeted sequencing (more than 50-fold coverage) of 8140 exons in 697 individuals.[12] Since that time, sequentially released data from each phase of the project has been made available. The most recent release, in February 2012, contained variant calls and phased genotypes across 1092 individuals, with an average coverage of >4-fold per

Received December 21, 2011; accepted June 20, 2012.

From the Department of Medicine and Department of Human Genetics, The University of Chicago, Chicago, IL.

The online-only Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.111. 962928/-/DC1.

Correspondence to Elizabeth McNally, The University of Chicago, 5841 S Maryland Ave, MC 6088, Room G611, Chicago, IL 60637. E-mail emcnally@uchicago.edu

DOI: 10.1161/CIRCGENETICS.112.962928

individual in those sequences released as part of the full project.[12]

Phenotypic data are unavailable from the subjects whose genomes were determined in the 1000 Genomes Project. Thus, interrogation of genetic variation in this cohort should be considered to include normal individuals and those with disease, with an expected prevalence of cardiomyopathy mirroring that of the population at large. Analysis of the 1000 Genomes Project dataset now allows for estimates of variant prevalence. Unlike the HapMap project, which provided information on common variation, the 1000 Genomes Project provides information on rare or private variation.[13] Herein, we queried the 1000 Genomes Project database to interrogate genetic variation in 3 sarcomeric genes (*MHY7*, *MYBPC3*, and *TTN*) to demonstrate how this database can be used to estimate genetic variation in the population. We focused our analysis on protein-altering variation (PAV), or those DNA deviations that predict a distinct protein from the referent genome sequence. PAVs are mainly single nucleotide polymorphisms (SNPs) that are nonsynonymous, or missense, in nature. PAV also includes insertion/deletions (indel) in coding regions, splice site altering variants, and nonsense polymorphisms. We found substantial variation, including protein-disrupting sequences, in *TTN*. We also identified previously reported and predicted pathogenic variation in *MYBPC3* and *MYH7* in the 1000 Genomes Project database at a frequency substantially higher than what would be predicted based on population-based studies of DCM and HCM prevalence. This data can be used to inform estimates of baseline genetic variation and, importantly, suggests that at-risk genotypes are more common in the population at large than previously expected.

Methods

Variant Discovery for Sarcomeric and Nonsarcomeric Genes

Exonic boundaries for *MYBPC*3 (ENST00000545968), *MYH7* (ENST00000355349), *CAMK2D* (ENST00000342666), *KCNQ1* (ENST00000155840), *KCNH2* (ENST00000262186), *SGCD* (ENST00000337851), *SGCG* (ENST00000218867), and *LMNA* (ENST00000368300) were downloaded from the Ensemble Genome Browser (www.ensemble.org). Exonic variants were extracted from the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release (No. ICHG2011) using the online data slicer available through the 1000 Genomes Project browser (http://browser.1000genomes.org). Variants were filtered based on their Phred quality score, discarding those below 29. Variants were subsequently submitted to the web site's Variant Effect Predictor (VEP), allowing for the analysis of potentially splice site altering variation within exons. The default settings were adjusted so as to return National Center for Biotechnology Information (NCBI) terms for variant consequences and to yield Sorts Intolerant From Tolerant (SIFT), Polyphen 2 (PP2), and Condel predictions and scores.[14–16] Variant predictions corresponding to the aforementioned transcripts were extracted. Variants were compared with those present in the National Institutes of Health National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/) using the Compare two Datasets tool available through the University of Pennsylvania Galaxy server (http://main.g2.bx.psu.edu/).

Variant Discovery for TTN

The 3 primary isoforms of titin (N2-BA, N2-A, and N2-B) result from differential splicing of the I-band encoding region of titin. The N2-BA isoform encodes the full-length protein and contains blocks of sequence that are specific to either N2-B or N2-A titin.[17] Both N2-BA and N2-B titin are expressed in the heart. The DNA sequence encoding the N2-BA isoform does not exist in the Ensemble or University of California Santa Cruz (UCSC) databases but instead must be deduced from the N2-B and N2-A isoforms.

The FASTA sequence for N2-A titin (NP_596869.4) was downloaded from the NCBI-Gene database and aligned with the Uniprot titin sequence Q8WZ42 using Uniprot's online sequence alignment tool (http://www.uniprot.org). The 927 amino acids absent from the N2-A isoform were extracted and run through Blat, a sequence alignment tool publically available through the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). The amino acids were mapped to positions 179 603 868 to 179 606 648 on chromosome 2 and were found to correspond to exon 45 in the N2-B isoform of titin (ENST00000460472) through cross-referencing of its exon boundaries on the Ensemble Genome Browser.

Exonic boundaries encoding the N2-A *TTN* isoform (ENST00000342992) and exon 45 of N2-B were downloaded from Ensemble. Exonic variants were extracted from the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release, using the online data slicer. Variants were filtered based on their Phred-scaled quality score, retaining only those variants with a score ≥29.[18,19] Exonic variants were then submitted to the Variant Effect Predictor (VEP), using the previously described settings. Variant consequences corresponding to ENST00000342992 and exon 45 of ENST00000460472 were extracted to yield all variation within N2-BA *TTN* as defined by the Uniprot consensus sequence Q8WZ42. Amino acid positions from the N2-A transcript were shifted from position 4380 onward to give their amino acid positions within N2-BA titin. *TTN* variants were compared with those present in the NHLBI ESP, as described above. Indels from the 1000 Genomes Project were identified in the low-coverage, whole-genome project. Indels are not available from the NHLBI ESP.

Variant Mapping to Protein Domains

Variants in *MYBPC3* and *TTN* were mapped to their respective locations within the protein using Uniprot sequence identifiers Q14896 and Q8WZ42. Variants in *MYH7* were mapped to their location within the protein using the boundaries outlined in Blair et al.[20,21] Pathogenic variants were identified through comparison with those listed on the online Human Gene Mutation Database (HGMD) Professional release 2011.3 (http://www.hgmd.org/).

Ethnic Distribution of TTN Variants

Subjects were matched to their corresponding ethnic groups (African, American, Asian, European) using the supplement provided with the integrated phase 1 release (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel). Variants were eliminated from this analysis in cases where the minor allele frequency exceeded that of the reference allele. The number of total PAVs per individual was assessed within each ethnic group for *MYH7*, *MYBPC3*, and *TTN*.

Statistical Analysis

A Kruskal-Wallis one-way analysis of variance was performed in PRISM to test for differences in variant load between ethnic groups. Posttest analysis was performed using Dunn's multiple comparison test.

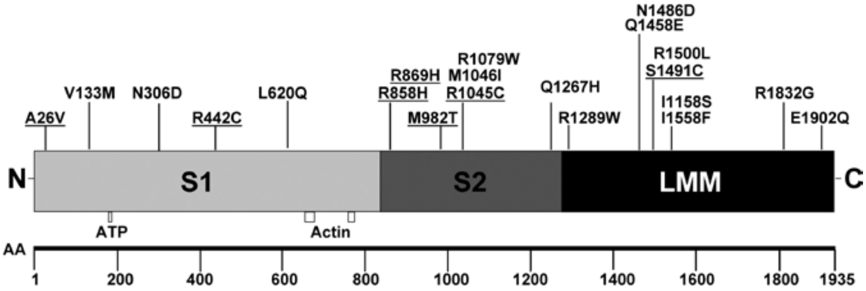

Figure 1. Missense variants identified in the *MYH7* gene in 1000 Genomes Project. The position of 21 rare missense variants from the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release. Variants also reported as pathogenic in the Human Genome Mutation Database (HGMD) are underlined. Adenosine trisphosphate (ATP) and actin refer to the ATP and actin-binding sites respectively.

Results

Extracting Protein-Coding Variants From Cardiomyopathy-Associated Genes From the 1000 Genomes Project Database

The variants used for this analysis were extracted from the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release. This call set derives from low-coverage, whole-genome, and exome sequencing data and contains phased genotype calls across 1092 individuals for SNPs, indels, and larger deletions. We queried the dataset for SNPs in 3 genes associated with cardiomyopathy: *MYH7*, encoding the major myosin heavy chain in the heart; *MYBPC3*, encoding myosin-binding protein C3; and *TTN*, encoding the giant protein titin.[21] *TTN* and *MYBPC3* SNPs were identified in data derived from low-coverage genome sequencing, as they were not included in the targeted exome analysis performed as part of the pilot project. *MYH7* SNPs were identified in both the low-coverage, whole-genome sequence data and the higher-coverage, targeted exome sequence. Of the SNPs in *MYH7,* 48.8% were found in both whole-genome and exome sequencing, 45.2% in exome sequencing alone, and 6.0% in whole-genome sequencing alone.

Variant Discovery in 3 Sarcomere Genes

Of the 11153 total variants in *MYH7*, *MYBPC3*, and *TTN*, 1136 were SNPs and 17 were indels, the later of which were all found in *TTN*. Indels ranged in size from 1 to 3bps, with the addition of 1 larger 18-bp deletion. Of the 1153 coding SNPs, 386 were synonymous, 15 of which were exonic SNPs predicted to alter a splice site, and 726 resulted in missense changes. Three nonsense variants were detected, all of which were in *TTN* (online-only Data Supplement Table I). Of these protein-altering variants (PAVs), 79.0% were defined as rare (MAF <0.5%), and 14.6% were low-frequency (defined as a MAF of 0.5% to 5.0%). Thus, 93.6% of variants with the potential to alter protein structure or function are found in a small fraction of the population, underscoring the contribution of infrequent variation to genotype diversity.

Distribution of MYH7 and MYBPC3 Variants

Figure 1 depicts the protein structure encoded by *MYH7* and displays the position of PAV detected in 1092 individuals from the 1000 Genomes Project database. Twenty-one different PAVs were found in the 1000 Genomes Project cohort (Table 1). Eleven fall within the S1 and S2 domains of the *β*-myosin heavy chain protein, whereas the remaining 10 are within the rod domain. Of the 21 PAVs in *MYH7*, 8 were reported by both PP2 and SIFT to be damaging (online-only Data Supplement Table II).[14,16] All 21 variants were present in the 1000 Genomes Project database as rare or low-frequency variants.

Figure 2 shows those PAVs detected in *MYBPC3* in the 1000 Genomes Project database. Twenty-two different PAVs were found (Table 2). Eight PAVs were reported by both PP2 and SIFT to be damaging (Table 2). All PAVs were present at a rare or low frequency.

Spatial Distribution of TTN Variants

Within *TTN*, 711 PAVs were detected in the 1000 Genomes Project database from 1092 individuals. Because of the large number of nonsynonymous SNPs (nsSNPs), we focused our analysis on indels. All indels had a Phred-scaled quality score of ≥29, indicating a base-call accuracy of almost 99.9%. Figure 3 shows the 17 indels detected, including 1 large 18-bp deletion, I11443-E11449del, found in 62 individuals with a MAF of 0.03 (Table 3). This 18-bp deletion falls within the PEVK region that regulates extensibility of titin.[22] Two individuals had multiple *TTN* indels, the implications of which depend on both phase and effect on protein frame. One had a frameshift combined with an in-frame deletion, and the second had a frameshift and 2 in-frame deletions. One individual was homozygous for the 18-bp deletion, and no individuals were homozygous for a frameshifting variant in TTN. Of the 17 indels in *TTN*, 7 were found in the Ig-like domains and none within the fibronectin type III domains. Of these 17 indels, 6 are in-frame deletions, and 11 are frameshifts that disrupt titin's carboxyl-terminus. Considering only the 11 indels that disrupt *TTN*, 3.2% of the population (35/1092) has a frameshift in *TTN*.

Variation in Sarcomere Genes Compared With Nonsarcomere Genes

The complexity of and requirements for sarcomere assembly are highly intricate. Because of the multiple protein-protein interactions required for proper assembly of the sarcomere, it has been suggested that sarcomere proteins, especially myosin, may be less tolerant of protein-coding variation.[23] To assess this, we queried both common and rare variation in the 3 sarcomere genes (*MYH7*, *MYBPC3*, and *TTN*) and compared this to several genes encoding proteins important to cardiac function. We evaluated *CAMK2D*, encoding calcium/calmodulin-dependent protein kinase 2 subunit δ; *KCNQ1* and *KCNH2*, encoding voltage-gated potassium channels linked

Table 1. ***MYH7*** **Protein-Altering Variants Found in the 1000 Genomes Project**

AA Change	Domain	SIFT Prediction (Score)	PP2 Prediction (Score)	MAF (No. of Occurrences From n=2184 Alleles)
A26V*	S1domain	T (0.25)	Benign (0.017)	0.0032 (7)
V133 M	S1 domain	D (0.04)	Probably damaging (0.998)	0.0005 (1)
N306D	S1 domain	T (0.52)	Possibly damaging (0.235)	0.0005 (1)
R442C*	S1 domain	D (0)	Probably damaging (1)	0.0005 (1)
L620Q	S1 domain	D (0.03)	Probably damaging (0.992)	0.0005 (1)
R858H*	S2 domain	D (0.02)	Benign (0.002)	0.0005 (1)
R869H*	S2 domain	T (0.07)	Possibly damaging (0.496)	0.0005 (1)
M982T	S2 domain	D (0)	Benign (0.06)	0.0005 (1)
R1045C*	S2 domain	D (0)	Benign (0.012)	0.0005 (1)
M1046I	S2 domain	T (0.2)	Benign (0.002)	0.0005 (1)
R1079W	S2 domain	D (0)	Possibly damaging (0.682)	0.0005 (1)
Q1267H	S2 domain	T (0.06)	Benign (0.001)	0.0009 (2)
R1289W	LMM	D (0)	Probably damaging (0.997)	0.0005 (1)
Q1458E	LMM	T (0.24)	Benign (0.037)	0.0005 (1)
N1486D	LMM	T (0.07)	Possibly damaging (0.539)	0 (0)
S1491C*	LMM	T (0.1)	Benign (0.001)	0.01 (17)
R1500L	LMM	D (0)	Probably damaging (0.982)	0 (0)
I1558F	LMM	D (0.02)	Probably damaging (0.99)	0.0005 (1)
I1558S	LMM	T (0.05)	Probably damaging (0.99)	0.0009 (2)
R1832G	LMM	D (0)	Possibly damaging (0.831)	0.005 (1)
E1902Q	LMM	T (0.16)	Benign (0.166)	0.0009 (2)

SIFT indicates Sorts Intolerant From Tolerant; PP2, polyphen 2; MAF, minor allele frequency; T, tolerated in SIFT; and D, deleterious in SIFT.

*Reported in the Human Gene Mutation Database (HGMD).

to hereditary long QT syndrome; *SGCD* and *SGCG*, encoding the dystrophin-associated proteins δ-sarcoglycan and γ-sarcoglycan; and *LMNA*, encoding the nuclear membrane protein Lamin A (online-only Data Supplement Table 1). The latter 3 genes have been linked to cardiomyopathy.[24] When considering PAVs/1 kb of coding sequencing, sarcomeric genes averaged 5.77 variants/kb compared with 6.35 variants/kb for nonsarcomeric genes (online-only Data Supplement Table 3). The range of PAV/1 kb of coding region varied considerably, from 2.29 to 14.84, suggesting that sarcomeric genes fall within this range. Among the 3 sarcomere genes, *MYH7* had the greatest amount of total variation but the least amount of PAV, with only 26.6% of the total exonic SNPs altering protein in *MYH7* compared with 73.3% and 68.2% PAV in *MYBPC3* and *TTN*.

Known or Reported Disease-Associated Variants in 1000 Genomes Project

Twenty-two SNPs detected in the 1000 Genomes Project database were also found in the Human Genome Mutation Database (HGMD), a comprehensive online database of inherited disease mutations (www.hgmd.org) (online-only Data Supplement Table 4). Of the 22 SNPs in the 1000 Genomes Project also identified in the HGMD, 11 were in *MYBPC3*, 7 in *MYH7,* and 4 in *TTN*. These SNPs represent a small portion of pathogenic variants reported in the HGMD, 3.6%, 3.2%, and 17.4% of published disease-causing mutations in *MYBPC3*, *MYH7*, and *TTN*, respectively. Within the 1000 Genomes Project population, 18 of the 22 mutations also present in the HGMD appeared as rare variants while 3 others were present at a low frequency. One variant, S236G

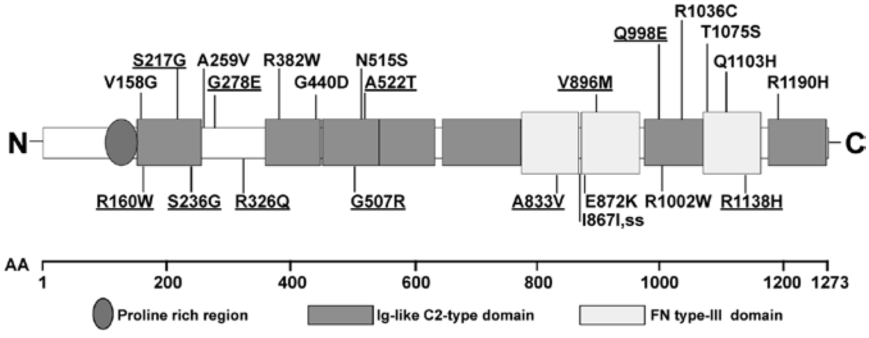

Figure 2. Missense variants identified in the *MYBPC3* gene in the 1000 Genomes Project. Cardiac myosin-binding protein C, with depiction of 22 missense variants, one of which alters a splice site from the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release. Variants present in the Human Genome Mutation Database (HGMD) are underlined.

Table 2. ***MYBPC3* Protein-Altering Variants in the 1000 Genomes Project**

AA Change	SIFT Prediction (Score)	PP2 Prediction (Score)	MAF (No. of Occurrences From n=2184 alleles)
V158 M	D (0.01)	Possibly damaging (0.269)	0.05 (101)
R160W*	D (0)	Probably damaging (1)	0.0014 (3)
S217G*	D (0.01)	Benign (0.154)	0.0005 (1)
S236G*	T (1)	Benign (0)	0.09 (189)
A259V	T (0.3)	Benign (0.002)	0.0023 (5)
G278E*	T (0.14)	Possibly damaging (0.637)	0.0046 (10)
R326Q*	D (0.01)	Benign (0.02)	0.0018 (4)
R382W	D (0.02)	Probably damaging (0.936)	0.01 (31)
G440D	T (1)	Benign (0.021)	0.0005 (1)
G507R*	D (0)	Probably damaging (1)	0.0027 (6)
N515S	T (0.34)	Probably damaging (0.994)	0.0018 (4)
A522T*	T (0.2)	Benign (0.143)	0.0014 (3)
A833V*	T (0.52)	Probably damaging (0.888)	0.01 (14)
I867I g->a synSS	n/a	n/a	0.02 (34)
E872K	T (0.05)	Probably damaging (0.994)	0.0005 (1)
V896 M*	T (0.05)	Benign (0.005)	0.0032 (7)
Q998E*	T (0.11)	Possibly damaging (0.799)	0.01 (17)
R1002W	D (0)	Probably damaging (1)	0.0027 (6)
R1036C	D (0)	Probably damaging (0.931)	0.0009 (2)
T1075S	T (0.05)	Benign (0.01)	0.0005 (1)
Q1103H	D (0)	Probably damaging (0.997)	0.0005 (1)
R1138H*	D (0)	Probably damaging (1)	0.0009 (2)
R1190H	T (0.05)	Probably damaging (1)	0.0005 (1)

SIFT indicates Sorts Intolerant From Tolerant; PP2, polyphen 2; MAF, minor allele frequency; D, deleterious in SIFT; T, tolerated in SIFT; and synSS, synonymous splice site.

*Reported in the Human Gene Mutation Database (HGMD). The exon-encoded predicted splice variant cannot be computed by SIFT or PP2.

in *MYBPC3*, was present at a higher frequency (MAF=0.09; n=189) and therefore is not likely an HCM-associated mutation, given its high prevalence.

To address the possibility of multiple variants within individuals, we evaluated the sarcomeric mutations found in both the HGMD and the 1000 Genomes Project (online-only Data Supplement Table 4). *MYBPC3* S236G was excluded from this analysis, given its frequency (MAF=0.09, n=189). Ninety-nine individuals had at least 1 pathogenic variant, and, of these, 93 individuals were heterozygous for a single variant. Three individuals from European and American cohorts were homozygous for *MYBPC3* mutations (V896 M, Q998E). One European individual was homozygous for the mutation S1491C in *MYH7*. Two individuals had 2 pathogenic variants each; 1 in *MYH7* and 1 in *MYBPC3 (*R1138H *MYBPC3* plus M982T *MYH7* and S217G *MYBPC3* plus

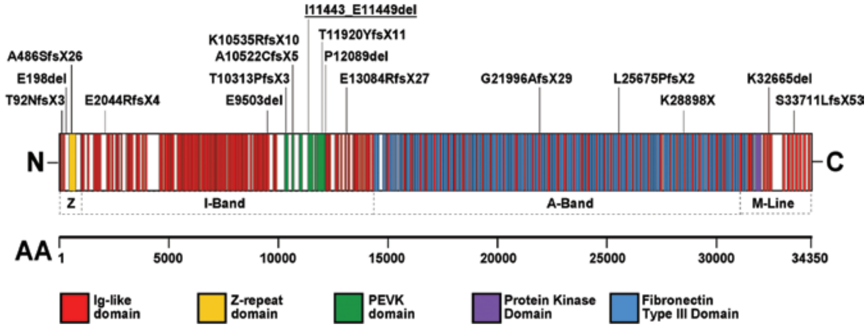

Figure 3. Frameshifting mutations identified in *TTN* in the 1000 Genomes Project. The titin isoform N2-BA is shown. Titin is linearly depicted with its 152 Ig-like domains in red and 132 fibronectin type III domains in blue. Depicted are the 17 indels identified in the 1000 Genomes Project February 2012 updated genotypes for the integrated phase 1 release. Six of 17 indels do not disrupt reading frame. One (underlined) that deletes 6 amino acids in the PEVK region is found in 5% of the population. Variants are shown relative to the titin Uniprot Sequence identifier Q8WZ42. The dashed lines below the protein schematic indicate the location of variants within the sarcomere.

Table 3. ***TTN* Insertion/Deletion Polymorphisms**

Location	Reference/ Alternate Allele	AA Change	Domain	MAF (No. of Occurrences in n=2184 Alleles)
179666885–179666886	G/GT	Thr92fs	Ig-like 1	0.0037 (8)
179664626–179664628	ACTT/A	Glu198del		0.0009 (2)
179658211–179658212	C/CT	Ala486fs	Z-repeat 2	0.0027 (6)
179640461–179640462	C/CT	Glu2044fs		0.0009 (2)
179570049–179570051	CTCT/C	Glu9503del	Poly-Glu	0.0009 (2)
179554281	GT/G	Thr10313fs	PEVK	0.0005 (1)
179549673–179549674	A/AG	Ala10522fs	PEVK	0.0018 (4)
179548796	TC/T	Lys10535fs	PEVK	0.0005 (1)
179514942–179514959	TTTTCCTCTTCAGGAGCAA/T	Ile11443_Glu11449del	PEVK	0.03 (62)
179505311–179505312	A/AG	Thr11920fs	PEVK	0.0005 (1)
179501447–179501449	CAGG/C	Pro12089del	Ig-like 80	0.0018 (4)
179495075–179495076	TTC/T	Glu13084fs		0.0014 (3)
179428912–179428913	A/AG	Leu25675fs	Ig-like 125	0.0005 (1)
179439949	GC/G	Gly21996fs	Ig-like 115	0.0005 (1)
179414950–179414952	TTAA/T	Lys28898del	Ig-like 133	0.0005 (1)
179395288	GA/G	Ser33711fs	Ig-like 148	0.0032 (7)
179398424–1179398426	ACTT/A	Lys32665del	Ig-like 144	0.0005 (1)

MAF indicates minor allele frequency.

S1491C *MYH7*). Thus, 5 of 1092 individuals in the 1000 Genomes Project carry multiple reported pathogenic mutations in sarcomeric genes. This number also exceeds population estimates of cardiomyopathy prevalence, suggesting that these variants may be insufficient to cause disease but may modify phenotype in the context of a primary driver mutation.

We employed 3 protein prediction algorithms as part of our analysis: SIFT, PP2, and Condel, a normalized composite of the prior 2 programs.[12] We used Condel to analyze previously reported mutations now identified in the 1000 Genomes Project dataset and found that it predicted 15 of the 17 possible variants to be pathogenic. Using this same approach, Condel predicted that 83.7% (36/43) of the missense variants detected in *MYH7* and *MYBPC3* would be pathogenic.

Comparison to the NHLBI ESP

We queried the NHLBI ESP[25] for total SNPs in *MYH7*, *MYBPC3*, and *TTN*. This is a database of high-coverage exonic sequence that includes 5379 samples drawn from African American and European American individuals. Average coverage in the NHLBI ESP across the 3 sarcomeric genes *MYH7*, *MYBPC3*, and *TTN* is higher (114.32-, 55.42-, and 108.73-fold, respectively) than the lower coverage 1000 Genomes Project, but the ESP lacks indel calls at this time; 51.2%, 60.6%, and 63.5% of all SNPs identified in the 1000 Genomes Project in *MYH7, MYBPC3,* and *TTN* respectively were also identified in the NHLBI ESP. Those SNPs in the 1000 Genomes Project but not in the NHLBI ESP were rare (present at an average frequency of 0.0020, 0.0037, and 0.0010 for *MHY7*, *MYBPC3*, and *TTN*, respectively.) Those SNPs found in the NHLBI ESP but not the 1000 Genomes Project were similarly rare (present at an average frequency of 0.00018, 0.00020, and 0.0049, respectively). The similar frequency of rare alleles in the low-coverage 1000 Genomes Project and the high-coverage NHLBI ESP argues that these findings are not false-positives related to low-coverage sequence. The different ethnic groups that constitute each sequenced cohort likely account for the difference in observed rare alleles.

Distribution of Variants by Race and Ethnicity

A Kruskal-Wallis one-way analysis of variance was used to test for differences in PAV in *MYH7, MYBPC3*, and *TTN* among ethnic groups (African, American, Asian, European) in the 1000 Genomes Project cohort. Five variants were removed before this analysis, as their MAF exceeded that of the reference allele. The frequency of variants differed significantly across the 4 groups in *MYH7* [H(3)=148.1, P<0.0001], *MYBPC3* [H(3)=10.05, P=0.0181], and *TTN* [H(3)=262.9, P<0.0001], (Figure 4). On posttest comparisons, Africans were shown to have significantly more variation in *MYH7* than all other ethnic groups at 0.68 variants/individual while Asians were shown to have significantly more variation in *TTN* at 28.76 variants/individual. The latter differed from the results of the 1000 Genomes pilot project in which African individuals were reported to have the greatest variation across the genome, as measured by the number of total and novel variants within each ethnic group.[12] Thus, variant frequencies differ between ethnic groups at the level of individual genes, and, notably, the extent and direction of these ethnic differences changes between genes.

Discussion

Pathogenic Variation in the 1000 Genomes Project Is Higher Than Expected; Implications for Testing

The 1000 Genomes Project launched in 2008 with the goal of creating an ethnically diverse public reference database for DNA polymorphisms. Unprecedented in its scope, the project is the first to approximate the breadth of human genetic

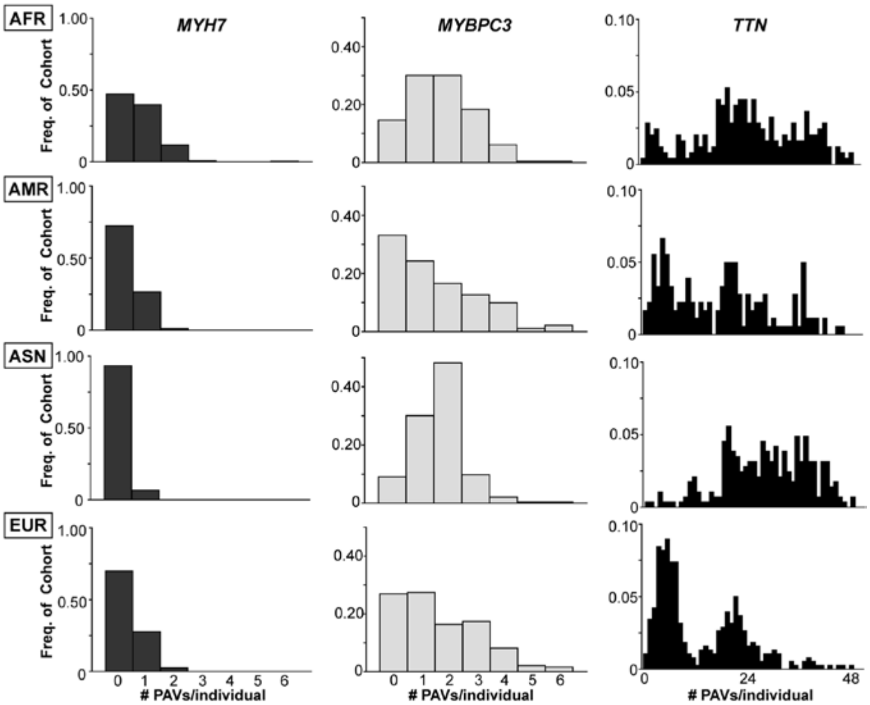

Figure 4. Protein-altering variation (PAV) within the 1000 Genomes Project cohort differs by cohort. The histograms above show the percentage of individuals within each cohort (AFR=African; AMR=American; ASN=Asian; EUR=European) having a specified number of PAVs per gene. This analysis included all PAV, except those 5 instances in which the MAF exceeded that of the reference allele. The amount of PAV in specified genes differed by racial groups, and this difference was significant (*MYH7*, [H(3)=148.1, *P*<0.0001]; *MYBPC3*, [H(3)=10.05, *P*=0.0181]; and *TTN*, [H(3)=262.9, *P*<0.0001]; *P* values are from nonparametric analysis of variance [ANOVA]). On posttest comparisons, African individuals had more variation in *MYH7* than all other groups, whereas Asians had significantly more variation in *TTN*.

variation and provide a quantitative framework in which to predict potentially relevant variants. We queried the database of 1092 individuals for 3 sarcomeric genes, as they were thought to contain little sequence variation.[21,23,26] This cohort, in the absence of phenotypic data, should be viewed as a representative sample of the population at large, containing both healthy and diseased individuals. Within the 1000 Genomes Project database, we identified variants that were previously reported as pathogenic in the HGMD. Although reported as mutations, some of these polymorphisms are not sufficient to cause cardiomyopathy, as several have been reported in normal populations. We also found variants not reported in the HGMD that are predicted to be pathogenic using computational algorithms. Summing these with the observed *TTN* frameshifts yields a frequency that substantially exceeds population estimates of the combined HCM and DCM prevalence. It is likely that some of these variants are positioned to modify disease outcome and favors a comprehensive sequence analysis. Supporting this idea, compound and double mutations have been noted in younger, more affected individuals.[27] Therefore, we favor comprehensive sequencing analysis especially in younger individuals.

Currently, assessing gene variants involves the use of prediction tools, segregation studies within families, and animal/cellular models to predict pathogenicity. As cardiomyopathy genetic testing panels become more expansive and the number of variants increases, costly and labor-intensive tools for analysis will prove challenging. Improved prediction tools will be necessary to better anticipate the effect of a given PAV. Better algorithms, along with comprehensive testing, should improve the information obtained from genetic analysis.

Indels in TTN

We also uncovered a high frequency of indels in *TTN*, all of which have a base call accuracy of almost 99.9%. In light of the high-cumulative frequency of *TTN* indels in the general population (9%), *TTN* may be positioned to modify phenotype. Just over 5% of the general population has an 18-bp in frame deletion in the proline rich region of titin. This region of titin has been closely linked to the elastic properties of titin and its ability to sense and respond to sarcomere length.[22] This allele will be important to track in studies of cardiac modifiers given its prevalence and potential functional role. Notably, 3% of the population has a protein-truncating *TTN* allele. This estimate is similar to the recent findings from Herman et al where protein-truncating *TTN* mutations were found in 25% of DCM subjects and in 3% of a normal control population.[28] Our analysis, unlike Herman et al, focused only on indels as protein-truncating mutations in *TTN* and did not include predicted *TTN* intronic splice-site mutations. Thus, our analysis is expected to underestimate the population prevalence of protein-truncating mutations in *TTN*.

Likely Underestimates of Pathogenic and Predicted Pathogenic Variation

With increasing sequencing coverage, the ability to detect private variation increases.[29] High-coverage sequence of *MYH7* yielded additional variants that were not detected in the low-coverage sequencing. The variation described in *TTN* and *MYBPC3* derives only from low-coverage whole-genome sequencing, and, therefore, our study probably underestimates variation. Overall, it is unlikely that these variants represent false-positives SNP. As of last year's reporting, the pilot project low-coverage genome sequence is estimated to have a <5% false discovery rate for SNPs and a 1% to 3% genotype error rate.[12] Our analysis also did not query variation in intronic sequence, including potentially pathogenic splice-site altering sequence, thus providing an additional source for pathogenic variation that is not represented in our baseline estimates of genetic variation. At present, the current 1000 Genomes Project database contains data on 1092 individuals and will extend sequence coverage to 2500 individuals to provide further information on rare and low-frequency variation.

Interpreting Rare Variation in Cardiomyopathy Genes

Given the extent of rare PAV in both sarcomeric and nonsarcomeric genes, the mere presence of rare variation does not imply pathogenicity. Segregation of a rare variant within a family or demonstration of impaired protein function within a biological context is often used to interpret the findings of genetic testing; however, these in vitro studies are often impractical, and not all family structures support this analysis, especially in families with diseases that limit lifespan. The continued development of analyses that better predict variant pathogenicity will be required. This analysis highlights the challenges that accompany analyzing population-based genetic information and applying findings to the individual patient with cardiomyopathy. The 1000 Genomes Project dataset will aid in interpreting genetic testing information, although its use is limited by lack of phenotypic correlates. In case-control studies of genotype and phenotype, control subjects should be carefully considered if the subjects have not been clinically evaluated or especially if they are too young to yet manifest disease

Sources of Funding

Supported by the Doris Duke Charitable Foundation, the Sarnoff Foundation, and NIH R01 HL092443 and NIH HL61322.

Disclosures

None.

References

1. Watkins H, Ashrafian H, Redwood C. Inherited cardiomyopathies. *N Engl J Med*. 2011;364:1643–1656.
2. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. 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*. 1995;92:785–789.
3. Gersh BJ, Maron BJ, Bonow RO, Dearani JA, Fifer MA, Link MS, Naidu SS, Nishimura RA, Ommen SR, Rakowski H, Seidman CE, Towbin JA, Udelson JE, Yancy CW; American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines; American Association for Thoracic Surgery; American Society of Echocardiography; American Society of Nuclear Cardiology; Heart Failure Society of America; Heart Rhythm Society; Society for Cardiovascular Angiography and Interventions; Society of Thoracic Surgeons. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2011;124:e783–e831.
4. Otsuka H, Arimura T, Abe T, Kawai H, Aizawa Y, Kubo T, Kitaoka H, Nakamura H, Nakamura K, Okamoto H, Ichida F, Ayusawa M, Nunoda S, Isobe M, Matsuzaki M, Doi YL, Fukuda K, Sasaoka T, Izumi T, Ashizawa N, Kimura A. Prevalence and distribution of sarcomeric gene mutations in Japanese patients with familial hypertrophic cardiomyopathy. *Circ J*. 2012;76:453–461.
5. Xu Q, Dewey S, Nguyen S, Gomes AV. Malignant and benign mutations in familial cardiomyopathies: insights into mutations linked to complex cardiovascular phenotypes. *J Mol Cell Cardiol*. 2010;48:899–909.
6. Petretta M, Pirozzi F, Sasso L, Paglia A, Bonaduce D. Review and metaanalysis of the frequency of familial dilated cardiomyopathy. *Am J Cardiol*. 2011;108:1171–1176.
7. Millat G, Bouvagnet P, Chevalier P, Sebbag L, Dulac A, Dauphin C, Jouk PS, Delrue MA, Thambo JB, Le Metayer P, Seronde MF, Faivre L, Eicher JC, Rousson R. Clinical and mutational spectrum in a cohort of 105 unrelated patients with dilated cardiomyopathy. *Eur J Med Genet*. 2011;54:e570–e575.
8. Gerull B, Gramlich M, Atherton J, McNabb M, Trombitas K, Sasse-Klaassen S, Seidman JG, Seidman C, Granzier H, Labeit S, Frenneaux M, Thierfelder L. Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat Genet*. 2002;30:201–204.
9. Gerull B, Atherton J, Geupel A, Sasse-Klaassen S, Heuser A, Frenneaux M, McNabb M, Granzier H, Labeit S, Thierfelder L. Identification of a novel frameshift mutation in the giant muscle filament titin in a large Australian family with dilated cardiomyopathy. *J Mol Med (Berl)*. 2006;84:478–483.
10. Taylor M, Graw S, Sinagra G, Barnes C, Slavov D, Brun F, Pinamonti B, Salcedo EE, Sauer W, Pyxaras S, Anderson B, Simon B, Bogomolovas J, Labeit S, Granzier H, Mestroni L. Genetic variation in titin in arrhythmogenic right ventricular cardiomyopathy-overlap syndromes. *Circulation*. 2011;124:876–885.
11. Linke WA. Sense and stretchability: the role of titin and titin-associated proteins in myocardial stress-sensing and mechanical dysfunction. *Cardiovasc Res*. 2008;77:637–648.
12. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–1073.
13. International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437:1299–1320.
14. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11:863–874.
15. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*. 2011;88:440–449.
16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–249.
17. LeWinter MM, Granzier H. Cardiac titin: a multifunctional giant. *Circulation*. 2010;121:2137–2145.
18. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998;8:186–194.
19. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8:175–185.
20. Blair E, Redwood C, de Jesus Oliveira M, Moolman-Smook JC, Brink P, Corfield VA, Ostman-Smith I, Watkins H. Mutations of the light meromyosin domain of the beta-myosin heavy chain rod in hypertrophic cardiomyopathy. *Circ Res*. 2002;90:263–269.
21. Buvoli M, Hamady M, Leinwand LA, Knight R. Bioinformatics assessment of beta-myosin mutations reveals myosin's high sensitivity to mutations. *Trends Cardiovasc Med*. 2008;18:141–149.
22. Miller MK, Granzier H, Ehler E, Gregorio CC. The sensitive giant: the role of titin-based stretch sensing complexes in the heart. *Trends Cell Biol*. 2004;14:119–126.
23. Freeman K, Nakao K, Leinwand LA. Low sequence variation in the gene encoding the human beta-myosin heavy chain. *Genomics*. 2001;76:73-80.
24. Dellefave L, McNally EM. The genetics of dilated cardiomyopathy. *Curr Opin Cardiol*. 2010;25:198–204.
25. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337:64–69.
26. Hamady M, Buvoli M, Leinwand LA, Knight R. Estimate of the abundance of cardiomyopathic mutations in the beta-myosin gene. *Int J Cardiol*. 2010;144:124–126.

27. Ingles J, Doolan A, Chiu C, Seidman J, Seidman C, Semsarian C. Compound and double mutations in patients with hypertrophic cardiomyopathy: implications for genetic testing and counselling. *J Med Genet*. 2005;42:e59.
28. Herman DS, Lam L, Taylor MR, Wang L, Teekakirikul P, Christodoulou D, Conner L, DePalma SR, McDonough B, Sparks E, Teodorescu DL, Cirino AL, Banner NR, Pennell DJ, Graw S, Merlo M, Di Lenarda A, Sinagra G, Bos JM, Ackerman MJ, Mitchell RN, Murry CE, Lakdawala NK, Ho CY, Barton PJ, Cook SA, Mestroni L, Seidman JG, Seidman CE. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med*. 2012;366:619–628.
29. Clark MJ, Chen R, Lam HY, Karczewski KJ, Euskirchen G, Butte AJ, Snyder M. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol*. 2011;29:908–914.

CLINICAL PERSPECTIVE

The 1000 Genomes Project is an international consortium designed to provide full genomic sequence information from an ethnically varied population. The clinical history of 1000 Genomes Project participants is unknown, so information from the project is best used to determine population estimates of sequence variation. We queried the 1000 Genomes Project database for genetic variation in three cardiomyopathy genes, MYH7 (β myosin heavy chain), MYBPC3 (myosin binding protein C), and TTN (titin). We focused our analysis on rare variation that was either reported previously as pathogenic or was predicted to be pathogenic by computational algorithms. These 3 genes encode sarcomere proteins, and mutations in these genes cause familial dilated and hypertrophic cardiomyopathy. We identified predicted and previously reported pathogenic variation at a much higher frequency than the incidence of both dilated and hypertrophic cardiomyopathy. Specifically, we found 9% of genomes had insertion/deletion polymorphisms in TTN. The functionality of this genetic variation is not known given the absence of clinical information from 1000 Genomes Project participants; however, these variants may be important genetic modifiers that increase the risk for developing heart failure in the general population. For patients with a history of familial cardiomyopathy, these findings favor comprehensive genetic testing to identify both primary mutations and secondary disease modifiers.