Genetics inMedicine | ARTICLE

A genome-first approach to aggregating rare genetic variants in *LMNA* for association with electronic health record phenotypes

Joseph Park, AB[1,2], Michael G. Levin, MD[2], Christopher M. Haggerty, PhD[3,4], Dustin N. Hartzel, BS[4], Renae Judy, MS[5], Rachel L. Kember, PhD[1], Nosheen Reza, MD[2,6] Regeneron Genetics Center[7], Marylyn D. Ritchie, PhD[1,8], Anjali T. Owens, MD[2,6], Scott M. Damrauer, MD[5] and Daniel J. Rader, MD[1,2,9]

Purpose: "Genome-first" approaches, in which genetic sequencing is agnostically linked to associated phenotypes, can enhance our understanding of rare variants' contributions to disease. Loss-of-function variants in *LMNA* cause a range of rare diseases, including cardiomyopathy.

Methods: We leveraged exome sequencing from 11,451 unselected individuals in the Penn Medicine Biobank to associate rare variants in *LMNA* with diverse electronic health record (EHR)–derived phenotypes. We used Rare Exome Variant Ensemble Learner (REVEL) to annotate rare missense variants, clustered predicted deleterious and loss-of-function variants into a "gene burden" ($N =$ 72 individuals), and performed a phenome-wide association study (PheWAS). Major findings were replicated in DiscovEHR.

Results: The *LMNA* gene burden was significantly associated with primary cardiomyopathy ($p = 1.78E-11$) and cardiac conduction disorders ($p = 5.27E-07$). Most patients had not been clinically diagnosed with *LMNA* cardiomyopathy. We also noted an association with chronic kidney disease ($p = 1.13E-06$). Regression analyses on echocardiography and serum labs revealed that *LMNA* variant carriers had dilated cardiomyopathy and primary renal disease.

Conclusion: Pathogenic *LMNA* variants are an underdiagnosed cause of cardiomyopathy. We also find that *LMNA* loss of function may be a primary cause of renal disease. Finally, we show the value of aggregating rare, annotated variants into a gene burden and using PheWAS to identify novel ontologies for pleiotropic human genes.

Genetics in Medicine (2020) 22:102–111; https://doi.org/10.1038/s41436-019-0625-8

Keywords: genome-first; rare variants; phenome-wide association studies (PheWAS); *LMNA*; electronic health records (EHRs)

INTRODUCTION

The study of the genetic basis of human disease has traditionally utilized a "phenotype-first" approach in which persons with phenotypic disease traits are genotyped or sequenced to identify gene variants that may be associated with or causal for disease.[1,2] A "genome-first" approach in which sequencing is applied to large heterogeneous populations with subsequent determination of the associated phenotypes is of interest.[3,4] This approach can be applied to health-care populations with extensive electronic health record (EHR) phenotype data, thus permitting an unbiased approach to phenome-wide association studies (PheWAS) to determine the clinical impact of specific genetic variants.[5,6] In addition to identifying previously unsuspected gene ontologies, this approach may also reveal that many patients with single-gene Mendelian disorders are not clinically diagnosed.[7]

Large-scale exome sequencing allows for the identification of rare exonic variants. Statistical aggregation tests that interrogate the cumulative effects of multiple rare variants in a gene (i.e., "gene burden") increase the statistical power of regression analyses and enable gene-based association studies to describe the implications of mutated genes in human disease. Gene burden PheWAS in large health-care populations could increase the potential to uncover novel

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [3]Department of Imaging Science and Innovation and The Heart Institute, Geisinger, Danville, PA, USA; [4]Biomedical and Translational Informatics Institute, Geisinger, Danville, PA, USA; [5]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [6]Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA, USA; [7]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA; [8]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [9]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. Correspondence: Daniel J. Rader (rader@pennmedicine.upenn.edu)

Submitted 29 April 2019; accepted: 18 July 2019
Published online: 6 August 2019

consequences of gene variants in the human disease phenome. One approach to gene burden PheWAS is to focus only on predicted loss-of-function (pLOF) variants,[6] but could lead to lack of power due to their infrequency. To address this issue, private and very rare missense variants could be added to substantially increase the number of genotypic cases. However, a major challenge is deciding which missense variants to include in gene burden tests of association.

The unbiased genome-first approach is an ideal system for studying the effects of rare variants in genes with known pleiotropy. Pathogenic variants in *LMNA* are highly pleiotropic and cause several rare diseases including dilated cardiomyopathy, familial partial lipodystrophy type 2, and Emery–Dreifuss muscular dystrophy, among others.[8–11] We leveraged the Penn Medicine Biobank (PMBB, University of Pennsylvania), a large academic biobank with exome sequencing linked to EHR data, to evaluate in detail the phenotypes associated with rare pLOF and annotated deleterious missense variants in *LMNA*. In addition to mining qualitative ICD-based diagnosis codes, we interrogated EHR data for quantitative phenotypic traits via analyses of clinical imaging and laboratory measurements. Our findings represent the first report of a genome-first approach to examining the clinical effects of pLOF and predicted deleterious missense variants in *LMNA*.

MATERIALS AND METHODS

Setting and study participants

All individuals recruited for the Penn Medicine Biobank (PMBB) are patients of clinical practice sites of the University of Pennsylvania Health System. Appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing, and access to all available EHR data. The study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki.

The DiscovEHR cohort was used to replicate major findings. DiscovEHR is a collaboration between the Geisinger Health System and Regeneron Genetics Center in which exome sequencing was performed on biospecimens collected and linked to EHR data through Geisinger's MyCode Community Health Initiative.[12]

Exome sequencing

This study included a subset of 11,451 individuals in the PMBB who had exome sequencing. We extracted DNA from stored buffy coats and then obtained exome sequences as generated by the Regeneron Genetics Center (Tarrytown, NY). These sequences were mapped to GRCh37 as previously described.[13] For subsequent phenotypic analyses, we removed samples with low exome sequencing coverage (i.e., less than 75% of targeted bases achieving 20× coverage; $N = 46$), high missingness (i.e., greater than 5% of targeted bases; $N = 14$), high heterozygosity ($N = 97$), dissimilar reported and genetically determined sex ($N = 104$), genetic evidence of sample duplication ($N = 89$), and cryptic relatedness (i.e., closer than third-degree relatives; $N = 145$) with overlap among categories, leading to a total of 455 removed from our database. Of note, among the 72 individuals identified as carrying one of pLOF variants or missense variants with Rare Exome Variant Ensemble Learner (REVEL)[14] scores of at least 0.65 who were used for the primary analyses of this work, 4 individuals were removed from subsequent analyses due to low coverage ($N = 2$), sex discordance ($N = 1$), and being part of a parent–child pair ($N = 1$).

Exome sequencing in the DiscovEHR cohort was also performed by the Regeneron Genetics Center, as previously described.[6,15] In addition to exclusions for sequence quality, sample duplicates, and sex discordance, we excluded 31,399 individuals with closer than third-degree relatedness, yielding a study set of 61,056 individuals.

Variant annotation and selection for gene burden association testing

For both PMBB and DiscovEHR, variants were annotated using ANNOVAR[16] as pLOF or missense variants. pLOFs were defined as frameshift insertions or deletions, gain or loss of stop codon, and disruption of canonical splice site dinucleotides. Only variants with minor allele frequencies (MAF) ≤0.1% per the Genome Aggregation Database (gnomAD) were considered for inclusion in the gene burden association testing. Several approaches to inclusion of rare variants in the gene burden were applied, including pLOFs only, additional ClinVar pathogenic variants, and inclusion of missense variants that were scored deleterious by 5/5 algorithms (SIFT[17], PolyPhen2 HumDiv, PolyPhen2 HumVar[18], LRT[19], MutationTaster[20]). To capture additional individuals with potentially pathogenic missense variants, we utilized REVEL, an ensemble method for predicting the pathogenicity of missense variants,[14] to score rare missense variants in *LMNA*.

Clinical data collection

International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes and procedural billing codes, medications, and clinical imaging and laboratory measurements were extracted from the patients' EHR. All laboratory values measured in the outpatient setting were extracted for participants from the time of enrollment in the Biobank until 3 March 2018; all units were converted to their respective clinical traditional units. Minimum, median, and maximum measurements of each measurement were recorded per individual. Glomerular filtration rate (GFR) estimates were calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation, given its superiority to the Modification of Diet in Renal Disease Study (MDRD) equation in patient populations with normal or mildly reduced eGFR. Inpatient and outpatient echocardiography measurements were extracted if available for participants from 1 January 2010 until 9 September 2016; outliers for each echocardiographic parameter (less than Q1 – 1.5*IQR or

greater than Q3 + 1.5*IQR) were removed. Similarly, minimum, median, and maximum values for each parameter were recorded per patient.

For DiscovEHR, phenotypes were retrieved from Geisinger's Phenomic Initiative database, which incorporates numerous sources (including the EHR) into a common data model. Patient demographics and ICD-10 codes from inpatient and outpatient encounters were retrieved as of 28 November 2018. ICD-9 codes were mapped to equivalent ICD-10 codes using underlying diagnosis codes.

Phenome-wide association studies

A PheWAS approach was used to determine the phenotypes associated with predicted deleterious variants in *LMNA* carried by individuals in PMBB.[21] ICD-10 encounter diagnoses were mapped to ICD-9 via the Center for Medicare and Medicaid Services 2017 General Equivalency Mappings (https://www.cms.gov/Medicare/Coding/ICD10/2017-ICD-10-CM-and-GEMs.html) and manual curation. Phenotypes for each individual were then determined by mapping ICD-9 codes to distinct disease entities (i.e., PheCodes) using the R package "PheWAS."[22] Patients were determined to have a certain disease phenotype if they had the corresponding ICD diagnosis on two or more dates, while phenotypic controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses.

Each disease phenotype was tested for association with the *LMNA* gene burden using a logistic regression model adjusted for age, age^2, gender, and the first ten principal components of genetic ancestry. We used an additive genetic model to collapse predictably deleterious *LMNA* variants via an extension of the fixed threshold approach.[23] Given the relatively high percentage of individuals of African ancestry present in PMBB, PheWAS analyses were performed separately by European and African genetic ancestry and combined with inverse variance weighted meta-analysis. Our association analyses considered only disease phenotypes with at least 200 cases (≥~1.75% prevalence in the cohort), based on a prior simulation study for power analysis of PheWAS.[24] This led to the interrogation of 333 total phenotypes, and we used a Bonferroni correction to adjust for multiple testing ($p = 0.05/333 \approx 1.5E\text{-}04$).

Replication of major PheWAS findings in DiscovEHR was performed using a logistic regression model adjusted for age, age^2, sex, and the first four principal components of ancestry. Dilated cardiomyopathy was defined as two or more encounter diagnoses of I42.0 ("Dilated cardiomyopathy"), or two or more instances of I42.8 ("Other cardiomyopathies")/I42.9 ("Cardiomyopathy, unspecified") diagnoses and mention of "dilated" in the underlying diagnosis code. Chronic kidney disease was defined as two or more encounter diagnoses of N18.3 ("Chronic kidney disease, stage 3 [moderate]"). For both phenotypes, patients with only one encounter diagnosis were excluded from analysis.

Statistical analyses

To compare available echocardiographic and serum laboratory measurements between carriers of predicted deleterious *LMNA* variants and genotypic controls, we used a nonparametric statistical model to compare each clinical measurement between the two groups using the Wilcoxon rank-sum test (i.e., Mann–Whitney *U* test). Additionally, comparisons were made using robust linear regression, adjusted for age, age^2, gender, and the first ten principal components of genetic ancestry, in both the overall population and individuals of European ancestry alone. Furthermore, 95% confidence intervals (CIs) and *p* values were corrected by bootstrapping with 1000 replicates via the adjusted percentile method. All statistical analyses, including PheWAS, were completed using R version 3.3.1 or version 3.5 (Vienna, Austria).

RESULTS

Phenome-wide association studies for gene burden of deleterious variants in *LMNA*

Among the 11,451 individuals in PMBB with exome sequencing, we identified a total of 11 individuals carrying one of nine different pLOF variants (including five frameshift insertions/deletions, one gain of stop codon, and three variants disrupting canonical splice site dinucleotides) in *LMNA* (Table S1). All 11 individuals carrying pLOF variants had a diagnosis of either "primary/intrinsic cardiomyopathy," "cardiac conduction disorders," or both, confirming that heterozygous pLOF variants in *LMNA* have a high penetrance for cardiomyopathy. Interestingly, only 4 of these 11 individuals had received clinical genetic testing to confirm their laminopathies.

A PheWAS on the 11 carriers with pLOFs alone showed a signal for cardiomyopathy (Fig. S1) but had insufficient power; furthermore, most known pathogenic *LMNA* variants are missense variants. Therefore, we identified 167 individuals with one of 88 rare (MAF ≤ 0.1% in gnomAD) missense variants in *LMNA* (Table S1). We aggregated pLOF variants and missense variants annotated as pathogenic in ClinVar ($N = 9$ different variants, 20 carriers) and performed PheWAS ($N = 33$ carriers), resulting in a stronger signal for cardiomyopathy that was significant (Fig. S2). Given that many of the rare *LMNA* variants were of unknown pathogenicity, we combined missense variants predicted to be deleterious by a consensus of 5/5 algorithms (SIFT[17], PolyPhen2 HumDiv, PolyPhen2 HumVar[18], LRT[19], MutationTaster[20]), one of the standard approaches for combining pLOF variants with computationally predicted pathogenic missense variants[6] ($N = 14$ different variants, 24 carriers; Table S1), in a gene burden PheWAS ($N = 35$ carriers; Fig. **1a**). The signal for cardiomyopathy diagnoses was even stronger, additionally identifying related diagnoses such as "first-degree atrioventricular (AV) block," "sinoatrial node dysfunction," and "congestive heart failure."

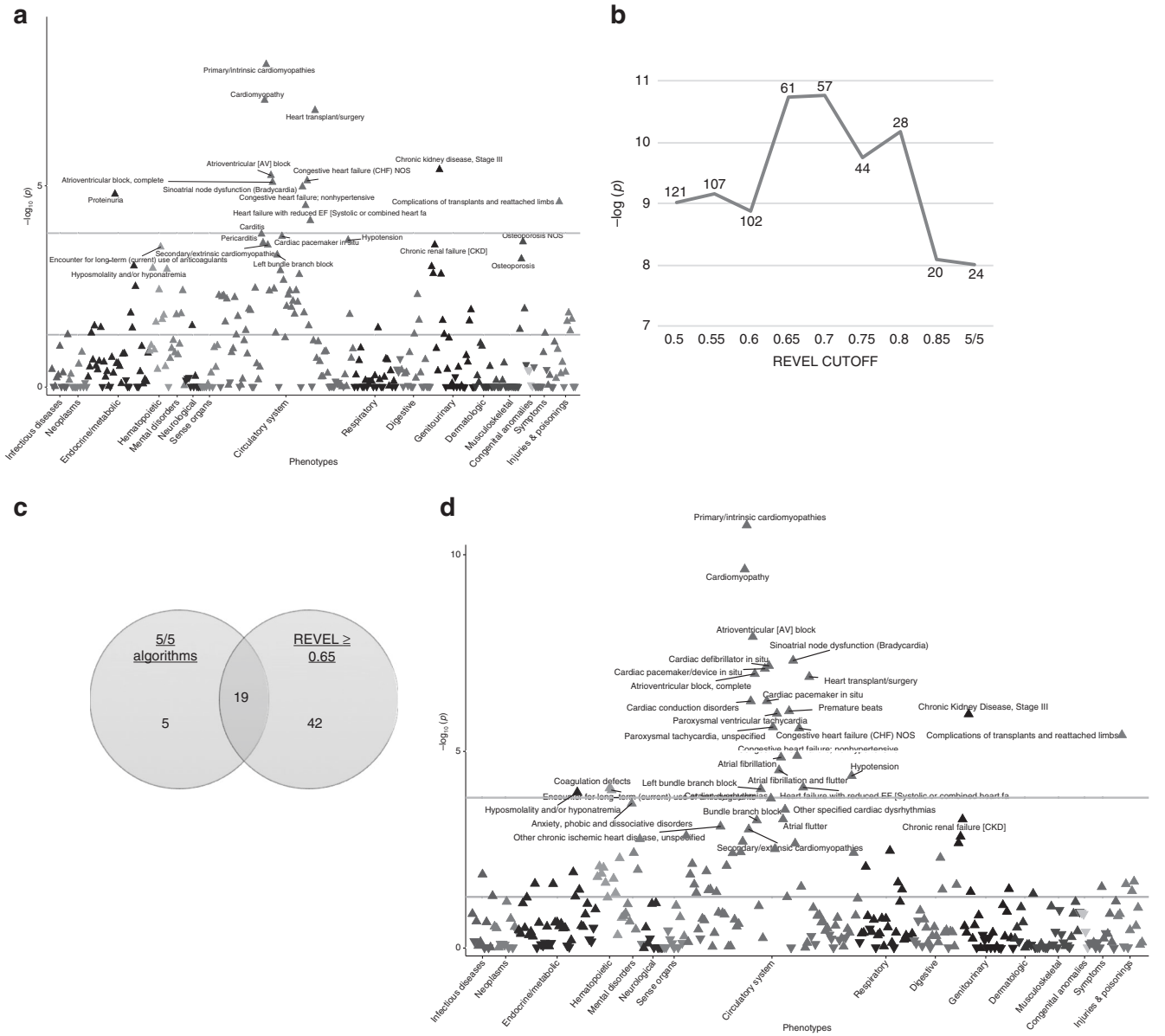

Fig. 1 Phenome-wide association studies (PheWAS) of predicted deleterious *LMNA* variants. Gene burden tests of association for predicted loss-of-function (pLOF) variants and predicted deleterious missense variants in *LMNA*. (**a**) Gene burden PheWAS of pLOF variants ($N = 11$ carriers) and missense variants predicted to be deleterious by 5/5 algorithms (SIFT, PolyPhen2 HumDiv, PolyPhen2 HumVar, MutationTaster, and LRT; $N = 24$). The blue line represents a p value of 0.05, and the red line represents the Bonferroni corrected significance threshold to adjust for multiple testing ($p = 0.05/333$). (**b**) Plot of p value for gene burden association with "primary/intrinsic cardiomyopathy" using pLOF variants and missense variants predicted to be deleterious per various REVEL cutoff scores as well as 5/5 algorithms. Each point is labeled with the number of exome-sequenced individuals who are carriers for missense variants in each threshold category without using a minor allele frequency threshold. (**c**) Venn diagram of number of exome-sequenced carriers for missense variants predicted to be deleterious by 5/5 algorithms and/or with a REVEL score ≥0.65. (**d**) Gene burden PheWAS of pLOF variants ($N = 11$) and missense variants with REVEL scores of at least 0.65 ($N = 61$). The blue line represents a p value of 0.05, and the red line represents the Bonferroni corrected significance threshold to adjust for multiple testing ($p = 0.05/333$).

However, we noted that there were a substantial number of carriers for rare missense variants in *LMNA* that did not meet the 5/5 criteria who were diagnosed with "primary/intrinsic cardiomyopathy" (Table S1), suggesting that this algorithmic filter was too stringent. To capture more individuals with pathogenic missense variants, we utilized REVEL, which has been reported to more accurately distinguish pathogenic from neutral missense variants, particularly those with MAFs less than 0.5%, compared with other predictive methods.[14] Analysis of variance on ClinVar-annotated variants showed that REVEL scores correlate with clinical pathogenicity (Table S2). While a threshold of 0.5 has been suggested,[14] we experimented with REVEL score thresholds in bins of 0.05 to evaluate the optimal score cutoff for capturing the most robust association with cardiomyopathy as a positive control (Fig. 1b). Of note, all REVEL cutoff scores of at least 0.5

Table 1 Demographics, clinical characteristics, and significant cardiovascular PheWAS associations for individuals in Penn Medicine Biobank (PMBB) carrying a predicted deleterious *LMNA* variant

Basic demographics	*LMNA*$^{+/-}$	*LMNA*$^{+/+}$	OR	*p* value
N	68	10,928	-	-
Male, *N* (%)	38 (55.9)	6489 (59.4)	-	0.625
Median age (at biobank entry), years	63.4	67.9	-	0.021
Race				
AFR, *N* (%)	12 (17.6)	2191 (20.0)	-	-
AMR, *N* (%)	4 (5.9)	303 (2.8)	-	-
EAS, *N* (%)	0 (0)	79 (0.7)	-	-
EUR, *N* (%)	51 (75.0)	8208 (75.1)	-	-
SAS, *N* (%)	1 (1.5)	114 (1.0)	-	-
Clinical cardiometabolic diagnoses				
Diabetes mellitus, *N* (%)	26 (38.2)	3508 (32.1)	1.31	0.298
Hypertension, *N* (%)	51 (75.0)	7957 (72.8)	1.12	0.785
Coronary artery disease, *N* (%)	32 (47.1)	4765 (43.6)	1.15	0.624
Myocardial infarction, *N* (%)	14 (20.6)	2214 (20.3)	0.98	0.881
Heart failure, *N* (%)	41 (60.3)	4159 (38.1)	0.40	2.40E-04
Dilated cardiomyopathy, *N* (%)	19 (27.9)	610 (5.6)	8.57	4.48E-09
Heart transplant, *N* (%)	14 (20.6)	379 (3.5)	7.21	1.00E-07
PheCodes				
Primary/intrinsic cardiomyopathy, *N* (%)	35 (58.3)	1608 (18.3)	6.37	1.78E-11
Cardiac conduction disorders, *N* (%)	42 (82.4)	2594 (44.4)	7.13	5.27E-07
Atrial fibrillation, *N* (%)	39 (81.3)	3352 (50.8)	5.64	1.42E-05
Atrioventricular (AV) block, *N* (%)	15 (62.5)	565 (14.8)	14.02	1.22E-08
Sinoatrial node dysfunction (bradycardia), *N* (%)	15 (62.5)	544 (14.4)	13.67	4.89E-08
Paroxysmal ventricular tachycardia, *N* (%)	27 (75.0)	1318 (28.9)	7.59	1.09E-06
Cardiac pacemaker/device in situ, *N* (%)	36 (80.0)	1849 (36.3)	8.53	7.90E-08
Cardiac defibrillator in situ, *N* (%)	28 (75.7)	1263 (28.0)	9.20	6.65E-08
Congestive heart failure, nonhypertensive, *N* (%)	40 (64.5)	3504 (42.1)	3.38	1.29E-05
Heart failure with reduced EF, *N* (%)	20 (47.6)	1415 (22.7)	3.82	8.23E-05
Heart transplant/surgery, *N* (%)	15 (40.5)	472 (8.9)	6.67	1.27E-07
Chronic kidney disease, stage III, *N* (%)	15 (30.6)	746 (10.3)	4.91	1.13E-06

Top and middle: Basic demographic characteristics (top) and cardiometabolic diagnoses (middle) for 68 of 72 heterozygous carriers of predicted loss-of-function (pLOF) variants ($N = 11$) and missense variants with REVEL scores of at least 0.65 ($N = 61$) (represented as *LMNA*$^{+/-}$) compared with noncarriers in the overall PMBB population (represented as *LMNA*$^{+/+}$). Each characteristic is labeled with count data in the *LMNA* carrier population and the rest of PMBB, as well as *p* values for two-tailed Fisher's exact tests. Of note, 4 of 72 carriers were not included due to additional genotypic quality check measures (see "Materials and Methods"). Bottom: Representative cardiovascular and renal PheCodes identified by gene burden phenome-wide association studies (PheWAS) for predicted deleterious exonic variants in *LMNA* (predicted loss-of-function variants and missense variants with a REVEL score of at least 0.65, $N = 72$). Patients were determined to have a certain PheCode if they had the corresponding International Classification of Diseases (ICD) diagnosis on two or more dates, while phenotypic controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses. Each phenotype is labeled with count and proportion data in the *LMNA* carrier population and the rest of PMBB, as well as odds ratios and *p* values attributable to *LMNA* carrier status via logistic regression adjusted for age, age^2, gender, and the first ten principal components of genetic ancestry.
AFR African, *AMR* mixed American, *EAS* East Asian, *EF* ejection fraction, *EUR* European, *OR* odds ratio, *SAS* South Asian.

performed better in identifying association with "primary/intrinsic cardiomyopathy" compared with the usage of 5/5 algorithms.

We chose a REVEL cutoff score of 0.65 given its optimal *p* value for association with "primary/intrinsic cardiomyopathy" (Fig. **1b**) while maintaining relatively high numbers of carriers for predictably deleterious *LMNA* variants. This cutpoint included 19 of the 24 carriers (11 of the 14 variants) that met the 5/5 criteria, but also included 42 additional carriers (21 variants) that did not meet the 5/5 criteria (Fig. **1c**). PheWAS of the *LMNA* gene burden of pLOF variants plus missense variants with REVEL scores of at least 0.65 ($N = 72$ carriers) revealed a much more robust signal for cardiomyopathy and related phenotypes (Fig. **1d**, Table **1**). Of note, the signal was more statistically robust compared with other recently developed ensemble methods for predicting pathogenicity such as VEST3[25,26] (Fig. S3), M-CAP[27] (Fig. S4), and CADD[28] (Fig. S5). Furthermore, we addressed potential issues of small sample sizes by using Firth's penalized likelihood approach, and found that beta and *p* value estimates were consistent with exact logistic regression (Table S3). Importantly, only 6 of the 35 individuals with a rare deleterious variant in *LMNA* and a diagnosis of "primary/intrinsic cardiomyopathy" had been molecularly diagnosed

with a *LMNA* variant (Table 1), indicating that *LMNA* cardiomyopathy is substantially underdiagnosed. Furthermore, 15 missense variants with REVEL scores >0.5 that are annotated as variants of uncertain significance or having conflicting interpretations of pathogenicity had at least one carrier with a diagnosis of "primary/intrinsic cardiomyopathy" and/or "cardiac conduction disorder" (Table S1).

Given the variety of cardiovascular traits that were highly significant in the REVEL-informed gene burden PheWAS for *LMNA*, we addressed whether these are independent signals. After running association analyses among all individuals with a phenotype of "primary/intrinsic cardiomyopathy," we found that the entire spectrum of cardiovascular PheWAS signals disappeared, suggesting that the other cardiac phenotypes were secondary to primary cardiomyopathy in carriers of the deleterious *LMNA* variants (Fig. S6).

In addition to cardiac disease phenotypes, our REVEL-informed *LMNA* gene burden PheWAS also identified phenome-wide significant disease phenotypes that are not typically defined as laminopathies, including "chronic kidney disease, stage III" ($p = 1.13E-06$; Fig. **1d**, Table 1). The relative persistence of the association signal for "chronic kidney disease, stage III" ($p = 1.33E-03$) when controlling for primary cardiomyopathy suggests an independent pathophysiological mechanism for renal failure in the context of loss of function in *LMNA* (Fig. S6).

We replicated these observations in the DiscovEHR cohort using the same approach (pLOFs plus REVEL score ≥0.65; Table S4a). There was a significant association between *LMNA* gene burden and dilated cardiomyopathy (odds ratio [OR]: 4.2 [95% CI: 1.3–10.0], $p = 0.005$; Table S4b). Furthermore, the association of *LMNA* gene burden with chronic kidney disease was also replicated (OR: 1.6 [95% CI: 1.1–2.5], $p = 0.02$; Table S4b).

Association of *LMNA* gene burden with cardiovascular imaging and clinical laboratory data

To build upon the PheWAS findings, we took a deeper dive into the cardiovascular imaging and laboratory EHR data (Table 1). First, we analyzed the cardiac structures of these individuals by interrogating available echocardiography data. By doing so, we also aimed to better define the PheCode "primary/intrinsic cardiomyopathy," which does not differentiate between the different types of primary cardiomyopathy. Carriers of rare deleterious *LMNA* variants had heart morphology consistent with dilated cardiomyopathy when compared with the rest of the PMBB population with echo data available (Table 2, Table S5a, b). More specifically, carriers had significantly increased left atrial volume indices, decreased left ventricular ejection fractions, decreased left ventricular outflow tract velocity time integrals, and increased mitral E/A ratios as an indication for weak atrial contraction.

We also conducted similar quantitative analyses for select clinical laboratory measurements. Carriers of predicted deleterious *LMNA* variants had significantly elevated alanine transaminase (ALT) and aspartate transaminase (AST) levels when compared with individuals not carrying a predicted deleterious *LMNA* variant (Table 3, Table S6a). In the overall population, carrier status was significantly associated with increased total cholesterol levels (Table 3, Table S6a,b). Furthermore, maximum blood triglyceride levels trended to be elevated among carriers ($p = 0.0559$; Table 3). These laboratory features are consistent with subclinical features of partial lipodystrophy, such as fatty liver and dyslipidemia.

Table 2 Cardiac architecture for carriers of presumed deleterious variants in *LMNA* is consistent with dilated cardiomyopathy

Echo parameter	$LMNA^{+/-}$ median (IQR) *N*	$LMNA^{+/+}$ median (IQR) *N*	Beta	*p*
Left atrial volume index, maximum	52.592 (37.491, 60.381) 20	36.982 (27.329, 49.802) 2648	11.582	0.00649
Left ventricular end systolic diameter PLAX, maximum (cm)	4.010 (3.563, 4.637) 31	3.490 (2.970, 4.290) 4643	0.512	0.0159
Left ventricular diastolic diameter PLAX, maximum (cm)	5.282 (4.792, 5.792) 32	4.980 (4.420, 5.591) 4696	0.188	0.235
Left ventricular ejection fraction (LVEF), minimum	45.00 (40.00, 55.00) 33	55.00 (40.00, 65.00) 5506	−7.501	0.0162
Left ventricular outflow tract (LVOT) velocity time integral, minimum (cm)	17.070 (14.200, 21.200) 28	19.100 (15.300, 23.175) 3846	−2.801	0.0114
Mitral E/A ratio, maximum	1.753 (1.413, 2.541) 26	1.312 (0.942, 1.901) 4529	0.517	0.0124

Comparison of representative echocardiography parameters for cardiac size and functionality between heterozygous carriers of predicted loss-of-function variants and missense variants with REVEL scores of at least 0.65 (represented as $LMNA^{+/-}$), and individuals in the Penn Medicine Biobank (PMBB) not carrying one of presumed deleterious variants with echocardiographic data available (represented as $LMNA^{+/+}$). Data is represented as median, respective first and third quartiles, the number of individuals from each population with available measurement data, and corresponding beta and *p* value attributable to *LMNA* carrier status via robust linear regression adjusted for age, age^2, gender, and the first ten principal components of genetic ancestry. 95% confidence intervals and *p* values were corrected by bootstrapping with 1000 samples.

IQR interquartile range, *PLAX* parasternal long-axis view.

Table 3 Clinical laboratory measurements for carriers of presumed deleterious variants in *LMNA* is consistent with subclinical features of partial lipodystrophy and renal disease

Lab parameter	$LMNA^{+/-}$ Median (IQR) N	$LMNA^{+/+}$ Median (IQR) N	p
ALT, maximum (U/L)	58.50 (32.25, 143.50) 50	36.00 (23.00, 62.00) 8459	1.13E-04
AST, maximum (U/L)	53.50 (35.50, 109.50) 50	35.00 (25.00, 63.00) 8392	1.48E-04
Total cholesterol, maximum (mg/dL)	208.00 (180.00, 248.00) 43	196.00 (162.00, 231.00) 6037	0.0259
LDL, maximum (mg/dL)	116.00 (90.50, 143.50) 43	114.00 (88.00, 145.00) 5982	0.998
HDL, minimum (mg/dL)	41.00 (29.00, 50.75) 42	39.00 (31.00, 50.00) 5978	0.693
Triglycerides, maximum (mg/dL)	185.00 (100.50, 319.00) 43	149.00 (102.00, 224.00) 6189	0.0559
Creatine kinase, maximum	133.50 (85.00, 196.50) 6	113.00 (71.00, 183.00) 1512	0.541
eGFR, minimum (mL/min/1.73 m^2)	38.26 (18.26, 54.64) 53	56.94 (32.52, 79.05) 8238	5.20E-05
Albumin (serum), minimum (g/dL)	3.00 (2.40, 3.70) 50	3.50 (2.90, 3.90) 8049	4.34E-03
Urine protein, maximum (mg/dL)	41.00 (20.00, 262.50) 8	22.00 (9.00, 85.00) 801	0.162

Unadjusted comparison via Wilcoxon rank-sum test of representative clinical laboratory parameters between heterozygous carriers of predicted loss-of-function variants and missense variants with REVEL scores of at least 0.65 (represented as $LMNA^{+/-}$), and individuals in the Penn Medicine Biobank (PMBB) not carrying one of presumed deleterious variants with serum laboratory data available (represented as $LMNA^{+/+}$). Data are represented as median, respective first and third quartiles, the number of individuals from each population with available measurement data, and corresponding *p* value for Wilcoxon rank-sum test.
ALT alanine aminotransferase, *AST* aspartate aminotransferase, *eGFR* estimated glomerular filtration rate, *HDL* high-density lipoprotein, *IQR* interquartile range, *LDL* low-density lipoprotein.

While only 2 of the 72 carriers of predicted deleterious variants had an ICD diagnosis of "lipodystrophy," there were 44 carriers with a phenotype of "hyperlipidemia," 20 carriers with a diagnosis of "type 2 diabetes," and eight with "secondary diabetes mellitus." Comprehensive investigation of physical exam notes written by health-care providers for individuals with these related metabolic phenotypes showed no mention of loss of subcutaneous fat from the extremities, trunk, or gluteal region, which is the classic presentation specific to partial lipodystrophy type 2.

Finally, regarding the identification of "chronic kidney disease, stage III" from our REVEL-informed gene burden PheWAS, we compared quantitative markers of renal disease between carriers of predicted deleterious *LMNA* variants and noncarriers in PMBB. We found that carrier status was associated with significantly decreased eGFR and serum albumin levels (Table 3, Table S6a, b). Furthermore, eGFR was still significantly decreased among carriers of predicted deleterious *LMNA* variants after adjusting for lifetime diagnosis of both congestive heart failure and diabetes mellitus, as well as adjusting for each diagnosis separately (Table 4). Additionally, serum albumin was also significantly decreased for carriers of predicted deleterious *LMNA* variants after adjusting for both heart failure and diabetes mellitus lifetime diagnoses (Table 4).

Table 4 Renal clinical laboratory measurements for carriers of presumed deleterious variants in *LMNA* are consistent with primary renal disease

Lab parameter	Beta	p
Adjusted for heart failure		
eGFR, minimum (mL/min/1.73 m^2)	−9.633	0.0149
Albumin (serum), minimum (g/dL)	−0.234	0.0842
Adjusted for diabetes mellitus		
eGFR, minimum (mL/min/1.73 m^2)	−16.121	4.59E-05
Albumin (serum), minimum (g/dL)	−0.399	5.65E-04
Adjusted for HF + DM		
eGFR, minimum (mL/min/1.73 m^2)	−10.648	0.00554
Albumin (serum), minimum (g/dL)	−0.264	0.0283

Comparison of estimated glomerular filtration rate (eGFR) and serum albumin between heterozygous carriers of predicted loss-of-function variants and missense variants with REVEL scores of at least 0.65, and individuals in the Penn Medicine Biobank (PMBB) not carrying one of presumed deleterious variants with serum laboratory data available, adjusted for lifetime congestive heart failure diagnosis (top), diabetes mellitus diagnosis (middle), and lifetime diagnoses of both heart failure (HF) and diabetes mellitus (DM) (bottom). Data are represented as beta and *p* value attributable to *LMNA* carrier status via robust linear regression adjusted for lifetime diagnosis of heart failure and/or diabetes mellitus as well as the first ten principal components of genetic ancestry. 95% confidence intervals and *p* values were corrected by bootstrapping with 1000 samples. eGFR not adjusted for age, age^2, and gender given the dependence of eGFR on age and gender per the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation. Serum albumin additionally adjusted for age and age^2.
eGFR estimated glomerular filtration rate.

DISCUSSION

While exome-wide interrogation of patients with shared phenotypic traits has been successful in identifying many new genetic variants associated with rare human disease, proving causality of disease due to pathogenic genetic variants in humans in vivo remains enigmatic.[29,30] We attempt to address the limitations of traditional phenotype-first approaches through this study, which represents a genome-first approach to analyzing the clinical manifestations of predicted deleterious variants in *LMNA* by fully utilizing available EHR data. Our study serves as an example of a genome-first approach for studying the medical consequences of rare pLOF and deleterious missense genetic variants in specific genes within the context of large health-care biobanks linked to extensive EHR phenotypic data.

An important area of research in precision medicine initiatives is to create a platform by which health-care providers can make accurate diagnoses based on a wide variety of personalized health data, including individuals' genetic information. However, current genetic panels offered at most health-care institutions cover only a small portion of genetic variants implicated in rare human diseases.[31] We suggest that the pipeline for interpretation of variants in *LMNA* identified via clinical genetic testing should be updated, as indicated by the number of variants of uncertain significance (VUS) identified in PMBB that we suggest may be pathogenic given the combination of their association with cardiomyopathy and/or arrhythmia and their predicted deleteriousness. Additionally, we found that important molecular diagnoses were missed, as many carriers for predicted deleterious variants in *LMNA* with dilated cardiomyopathy had not been sequenced for *LMNA*. In our analysis of PMBB, 35 individuals with a diagnosis of "primary/intrinsic cardiomyopathy" had a rare deleterious variant in *LMNA* and only six had been previously tested and molecularly diagnosed with a *LMNA* variant, suggesting that there is a lack of genetic testing for laminopathies in patients with cardiomyopathy of unknown etiology. Currently, *LMNA* genetic testing is not routinely offered to all patients with dilated cardiomyopathy unless a genetic cause is suspected to underlie dilated cardiomyopathy as a primary condition.[32–34] Furthermore, all six individuals who received testing were identified as carriers for known pathogenic variants, suggesting that some carriers of potentially pathogenic variants annotated as VUS as well as novel variants would not have been identified even if offered genetic testing in the clinic. Similarly, familial partial lipodystrophy due to a pathogenic *LMNA* variant is also likely underdiagnosed.

Although there are no current therapies specific to *LMNA* cardiomyopathy, there is benefit to making the molecular diagnosis with regard to providing an etiology for the cardiomyopathy, predicting clinical course and complications, and testing other family members at risk. More effective molecular diagnoses can lead to change in medical management for these individuals who are at high risk for arrhythmic sudden cardiac death.[35,36] In the clinical setting, dilated cardiomyopathy patients with confirmed pathogenic *LMNA* variants are often referred for electrophysiologic risk stratification earlier than other patients with nongenetic dilated cardiomyopathy. Thus, while evaluation of the contribution of individual variants remains clinically challenging and a definitive classification of pathogenicity for each presumed deleterious variant is hard to predict, our analyses suggest that earlier identification of laminopathies through an improved framework promoting genetic testing in the clinical setting using a comprehensive and updated variant panel is warranted to provide earlier, preventive treatments.

Additionally, the increased number of specific pathogenic variants in *LMNA* identified through this genome-first approach will provide greater insight into LMNA structure–function. Interestingly, 19 of 29 known ClinVar-annotated pathogenic missense variants cause a deviation from arginine in various locations of the *LMNA* protein product, highlighting a potential importance of the positively charged arginine in the LMNA protein structure, consistent with previous studies identifying arginine in many splicing binding sites for generating prelamin A and lamin C.[37] Notably, among novel missense variants discovered in this study, 8 of 18 variants with REVEL scores of at least 0.65 cause deviations from arginine, consistent with the prevalence of these changes in known clinically pathogenic missense variants.

This approach to inclusion of REVEL-annotated likely deleterious missense variants in a gene burden has the advantage of increasing the power for gene burden PheWAS analyses that can identify novel gene ontologies, as seen by the identification of advanced renal disease in the context of loss of function in *LMNA*. While renal abnormalities are possible direct clinical sequelae related to heart failure and diabetes mellitus, pathophysiological mechanisms for renal failure due to pathogenic *LMNA* variants through primary, noncardiorenal processes have recently been suggested.[38,39] We report impaired renal function and hypoalbuminemia in the context of loss of function in *LMNA*, even after adjusting for both a lifetime diagnosis of congestive heart failure and diabetes mellitus, suggesting a pathophysiology for renal failure due to a proteinuric, primary nephrotic clinical picture that may be confounded by, yet independent of, the pathophysiology of heart failure in dilated cardiomyopathy and the overlap with diabetes in partial lipodystrophy. Our results suggest a clinical or subclinical nephrotic phenotype due to loss-of-function variants in *LMNA* that may have been further masked by comorbid cardiac and metabolic disease traits, calling for follow-up studies interrogating primary renal disease as a potential novel laminopathy.

In conclusion, we used an approach to include pLOFs and REVEL-annotated deleterious missense variants in *LMNA* in a gene burden to show by PheWAS, using a relatively small number of carriers, significant associations with primary dilated cardiomyopathy, laboratory values consistent with partial lipodystrophy, and a novel finding of chronic kidney disease. We demonstrate the importance of deeply

interrogating quantitative data in the EHR to uncover important clinical and subclinical information relevant to other rare laminopathies implicated by deleterious *LMNA* variants. Our approach suggests an expanded role for clinical genetic testing for patients who present with primary dilated cardiomyopathy or early pathophysiologic signs like conduction defects. Importantly, our study also lays a methodological framework by which future studies can uncover novel gene–disease relationships and identify novel pathogenic loss-of-function variants across the human genome through genome-first analyses of large, heterogeneous health care–based populations.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-019-0625-8) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank JoEllen Weaver, David Birtwell, Heather Williams, Paul Baumann, and Marjorie Risman.

DISCLOSURE

Research reported in this paper was supported by the National Human Genome Research Institute of the National Institutes of Health under award number F30HG010442. A.T.O. receives funding support from the Winkelman Family Fund for Cardiac Innovation, not related to this work. S.M.D. receives research support to the University of Pennsylvania from RenalytixAI and CytoVas, not related to this work. The other authors declare no conflicts of interest.

REFERENCES

1. Chong JX, Buckingham KJ, Jhangiani SN, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet. 2015;97:199–215.
2. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014;312:1870–1879.
3. Stessman HA, Bernier R, Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. Cell. 2014;156:872–877.
4. Mefford HC. Genotype to phenotype-discovery and characterization of novel genomic disorders in a "genotype-first" era. Genet Med. 2009;11:836–842.
5. Lim ET, Wurtz P, Havulinna AS, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. 2014;10:e1004494.
6. Dewey FE, Murray MF, Overton JD, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016;354:aaf6814.
7. Abul-Husn NS, Manickam K, Jones LK, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science. 2016;354:aaf7000.
8. Worman HJ, Bonne G. "Laminopathies": a wide spectrum of human diseases. Exp Cell Res. 2007;313:2121–2133.
9. Benedetti S, Menditto I, Degano M, et al. Phenotypic clustering of lamin A/C mutations in neuromuscular patients. Neurology. 2007;69:1285–1292.
10. Capell BC, Collins FS. Human laminopathies: nuclei gone genetically awry. Nat Rev Genet. 2006;7:940–952.
11. Genschel J, Schmidt HH. Mutations in the LMNA gene encoding lamin A/C. Hum Mutat. 2000;16:451–459.
12. Carey DJ, Fetterolf SN, Davis FD, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med. 2016;18:906–913.
13. Dewey FE, Gusarova V, Dunbar RL, et al. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. N Engl J Med. 2017;377:211–221.
14. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–885.
15. Staples J, Maxwell EK, Gosalia N, et al. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. Am J Hum Genet. 2018;102:874–889.
16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
17. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–1081.
18. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–249.
19. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–1561.
20. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11:361–362.
21. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31:1102–1110.
22. Carroll RJ, Bastarache L, Denny JCR. PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014;30:2375–2376.
23. Price AL, Kryukov GV, de Bakker PI, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86:832–838.
24. Verma A, Bradford Y, Dudek S, et al. A simulation study investigating power estimates in phenome-wide association studies. BMC Bioinformatics. 2018;19:120.
25. Carter H, Douville C, Stenson PD, et al. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 suppl 3:S3.
26. Douville C, Masica DL, Stenson PD, et al. Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016;37:28–35.
27. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48:1581–1586.
28. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–315.
29. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508:469–476.
30. Marian AJ. Causality in genetics: the gradient of genetic effects and back to Koch's postulates of causality. Circ Res. 2014;114:e18–21.
31. Burkett EL, Hershberger RE. Clinical and genetic issues in familial dilated cardiomyopathy. J Am Coll Cardiol. 2005;45:969–981.
32. Newton-Cheh C. Should identifying a titin truncating variant change the management of patients with dilated cardiomyopathy? J Am Coll Cardiol. 2017;70:2275–2277.
33. Hasselberg NE, Haland TF, Saberniak J, et al. Lamin A/C cardiomyopathy: young onset, high penetrance, and frequent need for heart transplantation. Eur Heart J. 2017;39:853–860.
34. Ellepola CD, Knight LM, Fischbach P, Deshpande SR. Genetic testing in pediatric cardiomyopathy. Pediatr Cardiol. 2017;39:491–500.
35. Anselme F, Moubarak G, Savoure A, et al. Implantable cardioverter-defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Heart Rhythm. 2013;10:1492–1498.
36. Taylor MR, Fain PR, Sinagra G, et al. Natural history of dilated cardiomyopathy due to lamin A/C gene mutations. J Am Coll Cardiol. 2003;41:771–780.
37. Lee JM, Nobumori C, Tu Y, et al. Modulation of LMNA splicing as a strategy to treat prelamin A diseases. J Clin Invest. 2016;126:1592–1602.

38. Thong KM, Xu Y, Cook J, et al. Cosegregation of focal segmental glomerulosclerosis in a family with familial partial lipodystrophy due to a mutation in LMNA. Nephron Clin Pract. 2013;124:31–37.

39. Imachi H, Murao K, Ohtsuka S, et al. A case of Dunnigan-type familial partial lipodystrophy (FPLD) due to lamin A/C (LMNA) mutations complicated by end-stage renal disease. Endocrine. 2009;35:18–21.