2021

A Genome-First Approach To Investigating The Biological And Clinical Relevance Of Exome-Wide Rare Coding Variation Using Electronic Health Record Phenotypes

Joseph Park
University of Pennsylvania

A Genome-First Approach To Investigating The Biological And Clinical Relevance Of Exome-Wide Rare Coding Variation Using Electronic Health Record Phenotypes

Abstract

Genome-wide association studies (GWAS) have successfully described the roles of common genetic variation on human diseases by analyzing large populations recruited based on a shared phenotype, but the biological and clinical relevance of numerous genes remain incompletely described through these 'phenotype-first' methodologies. Much of the unexplained genetic contribution to disease risk and variability in complex traits may belong to the very rare and private spectrum of alleles, a range traditionally ignored by GWAS. Furthermore, the phenotype-first approach is likely to miss unexpected phenotypic consequences of genetic variants, such as those that may not be feasible to study in a phenotype-first approach due to rarity of the condition. The Penn Medicine BioBank, a healthcare system-based database of genotype, whole-exome sequencing, and electronic health record data, allows for an unbiased, 'genome-first' approach to describing the relationships between genetic variants and human disease traits captured in the clinical setting. Through 'gene burden' tests that interrogate the cumulative effects of multiple rare and private variants in a gene that are predicted to affect gene function, this dissertation aims to characterize the clinical manifestations of diseases and traits caused by rare, predicted loss-of-function and predicted deleterious missense variants on an exome-wide and/or phenome-wide scale. These analyses uncover previously unsuspected medical and biological consequences of loss-of-function variants in multiple genes. In summary, this dissertation will investigate the biological and clinical relevance of disease-associated genes by investigating the association of rare coding variation found in whole-exome sequencing with phenotypes derived from the EHR.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Daniel J. Rader

Second Advisor

Marylyn D. Ritchie

Subject Categories

Bioinformatics | Genetics | Medicine and Health Sciences

A GENOME-FIRST APPROACH TO INVESTIGATING THE BIOLOGICAL AND CLINICAL RELEVANCE OF EXOME-WIDE RARE CODING VARIATION USING ELECTRONIC HEALTH RECORD PHENOTYPES

Joseph Park

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation	Co-Supervisor of Dissertation
________________________	________________________
Daniel J. Rader, M.D.	Marylyn D. Ritchie, Ph.D.
Seymour Gray Professor of Molecular Medicine	Professor of Genetics

Graduate Group Chairperson

____

Benjamin F. Voight, Ph.D.
Associate Professor of Systems Pharmacology and Translational Therapeutics

Dissertation Committee
Katherine L. Nathanson, M.D., Pearl Basser Professor for BRCA-Related Research (Chair)
Hakon Hakonarson, M.D., Ph.D., Professor of Pediatrics
Jinbo Chen, Ph.D., Professor of Biostatistics
Iain Mathieson, Ph.D., Assistant Professor of Genetics
Ron Do, Ph.D., Associate Professor of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

A GENOME-FIRST APPROACH TO INVESTIGATING THE BIOLOGICAL AND CLINICAL RELEVANCE OF EXOME-WIDE RARE CODING VARIATION USING ELECTRONIC HEALTH RECORD PHENOTYPES

Dedicated to Hi Choon Park, Sang Hoon Lee, Eun Hee Jung, Soon Hwa Choi,

Seungho Park, and Sookhyung Park

ACKNOWLEDGMENT

The past four years have been a life-changing experience for me. Since I was a child, I have many memories of working hard to constantly accomplish goals and pushing myself to my limits, often being relentless to put it nicely, or perhaps being stubborn. But why I did that since such a young age is unclear to me. Perhaps it was due to an intrinsic quality of my personality. I remember always competing to stand out, whether it be auditioning to be admitted into the Juilliard School's Pre-College Division, then trying to surpass thousands of peers at Stuyvesant High School to get into a good college, then comparing myself to my peers at Harvard College to get into a good medical school—the struggle never seemed to stop. One could say that these struggles worked out on paper, but once I came to the Perelman School of Medicine at the University of Pennsylvania to start my medical training as an MD-PhD student, I started to feel the effects of burnout, losing motivation and questioning the purpose of having constantly pushed myself. My performance also started to wane, so I even questioned my enthusiasm for becoming a physician-scientist and became unsure of what I wanted to do.

This is when I met my mentors Daniel J. Rader and Marylyn D. Ritchie. I want to thank them foremost for sparking a passion within me to study human genomics and for rekindling my original motivations to become a physician-scientist. All the work I have done during my dissertation would surely not have been possible without their mentorship. I truly felt that they always kept me in their thoughts no matter how busy they were and wished the best for me in both academics and life in general. They taught me how to truly do science and perhaps more importantly how to effectively convey the significance of our scientific work to others. I will never forget the American Society of Human Genetics annual conference in 2019: the largest platform talk I have ever given, the ecstatic moment that Marylyn and I shared as I was announced as winner of the Charles J. Epstein Trainee Award for Excellence in Human Genetics Research, and the phone calls I had with Dan before and after the conference as he prepared me and

congratulated me from afar despite not being able to make it to Houston, Texas. That was the moment when many things clicked for me—that I was doing the right kind of work, that I was being mentored by the best, and that I had a future in this field. And suddenly I became determined again to push myself and make the best out of my dissertation work, and importantly to become passionate about a future career as a physician-scientist striving to better characterize the genomics of human disease to ultimately make tangible improvements to human society. I sincerely thank them for trailblazing a career path for me and for helping me to become comfortable in my own skin as a scientist. I now clearly understand the reason for having pushed myself, and am eager to continue with purpose.

I also want to thank my family, especially my grandparents (Hi Choon Park, Sang Hoon Lee, Eun Hee Jung, and Soon Hwa Choi) and my parents (Seungho Park and Sookhyung Park), to whom I dedicate this dissertation. When my parents and I left our home in South Korea when I was 3 years old, a major reason was for me to have the potential to develop into who I want to be, unaffected by the confinements of Korea's strict societal hierarchies. Both my grandparents and parents had to sacrifice the ability to be with each other, and my family came to the United States with virtually nothing. I dedicate this dissertation to my grandparents, who all passed away during my college and early medical school years and unfortunately would not be able to see this, for their love and sacrifices for my family to have a better future and for always supporting me from afar. I'm sure they would be absolutely overjoyed to know that I've made it up to here. And I dedicate this dissertation to my parents, for whom I wish that everything I do leads to their happiness and well-being. They brought me to the States starting from ground zero so that I might have a brighter future, and all they have ever done has only been for my benefit. This dissertation serves as a testament that they have clearly succeeded at being the best possible parents and fulfilled their life goals.

Finally, I thank God for His love, and for always guiding me and molding me into a better citizen of the world.

ABSTRACT

A GENOME-FIRST APPROACH TO INVESTIGATING THE BIOLOGICAL AND CLINICAL RELEVANCE OF EXOME-WIDE RARE CODING VARIATION USING ELECTRONIC HEALTH RECORD PHENOTYPES

Joseph Park

Daniel J. Rader and Marylyn D. Ritchie

Genome-wide association studies (GWAS) have successfully described the roles of common genetic variation on human diseases by analyzing large populations recruited based on a shared phenotype, but the biological and clinical relevance of numerous genes remain incompletely described through these ‘phenotype-first’ methodologies. Much of the unexplained genetic contribution to disease risk and variability in complex traits may belong to the very rare and private spectrum of alleles, a range traditionally ignored by GWAS. Furthermore, the phenotype-first approach is likely to miss unexpected phenotypic consequences of genetic variants, such as those that may not be feasible to study in a phenotype-first approach due to rarity of the condition. The Penn Medicine BioBank, a healthcare system-based database of genotype, whole-exome sequencing, and electronic health record data, allows for an unbiased, ‘genome-first’ approach to describing the relationships between genetic variants and human disease traits captured in the clinical setting. Through ‘gene burden’ tests that interrogate the cumulative effects of multiple rare and private variants in a gene that are predicted to affect gene function, this dissertation aims to characterize the clinical manifestations of diseases and traits caused by rare, predicted loss-of-function and predicted deleterious missense variants on an exome-wide and/or phenome-wide scale. These analyses uncover previously unsuspected medical and biological consequences of loss-of-function variants in multiple genes. In summary, this dissertation will investigate the biological and clinical relevance of disease-associated genes by investigating the association of rare coding variation found in whole-exome sequencing with phenotypes derived from the EHR.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF APPENDICES

CHAPTER 1. Introduction: Transitioning from a phenotype-first to a genome-first era of genetic association studies

The study of the genetic basis of human disease has traditionally utilized a ‘phenotype-first’ approach in which individuals who share a phenotypic disease trait are recruited for genotyping or sequencing to identify gene variants that may be associated with or causal for the phenotype of interest. Within this realm, how we have historically studied human genetic variation can be broadly organized by minor allele frequency (MAF) into studies of common variants versus rare variants. On one end of the allele frequency spectrum, we have studied common (MAF > 1%) genetic variants in populations recruited based on a shared phenotype using methods like genome-wide association studies (GWAS) to describe how these common variants of small effect can contribute to common complex diseases, such as coronary artery disease, type 2 diabetes, and cancer.[1-3] However, GWAS have not historically interrogated low-frequency (0.1% < MAF ≤ 1%) and rare (MAF ≤ 0.1%) coding variation given its reliance on genotype arrays that include common single nucleotide polymorphisms (SNPs), and more extensive exome or genome sequencing to capture rare genetic variants had been limited to relatively small sample sizes. On the other end of the allele frequency spectrum, rare variants, which are predicted to have larger effect sizes, have been studied mostly in the context of Mendelian disorders such as cystic fibrosis, familial hypercholesterolemia, and muscular dystrophy.[4,5] Identification of rare loss of gene function variants through high-throughput DNA sequencing in such human cohorts has allowed for the discovery of novel gene associations with various Mendelian diseases. However, because these variants are rare, it is difficult to study their clinical implications using traditional GWAS analysis approaches due to lack of statistical power especially when using a phenotype-first approach, and thus the clinical implications of most rare variants are unknown.

To address these limitations in traditional phenotype-first approaches to genetic association studies, this dissertation tackles the issue of how to systematically search for rare genetic variants that confer risk for or resilience to human disease. A ‘genome-first’ approach in which sequencing is first applied to large heterogeneous populations with subsequent determination of the associated phenotypes is of recent interest, and transitioning to this approach may help fill in additional puzzle pieces in the missing heritability of many human diseases and also provide new insights into the role of many genes in human biology and phenotypes.[6,7] To overcome the issue of statistical power while maintaining a high prevalence of human diseases, this approach can be applied to healthcare populations with extensive electronic health record (EHR) phenotype data.[8] Furthermore, this would permit an unbiased approach to ‘phenome-wide association studies’ (PheWAS), in which a single DNA variant is associated with various phenotypes to determine the clinical impact of specific genetic variants on the human disease phenome.[9] Importantly, a genome-first approach to PheWAS has the advantage of testing for pleiotropy in addition to discovering new genotype-phenotype relationships.[10] Thus, if we could combine whole-exome sequencing (WES) with EHR phenotype data, it would allow for a genome-first approach to associating rare human genetic variants with various phenotypic data captured natively in the documentation of clinical care.

But because rare variants are often too rare to study in a univariate association test due to lack of statistical power,[11] genome-first approaches to genetic association tests like those utilizing PheWAS have traditionally focused on single common variants.[12] However, gene-based association studies that aggregate rare and private variants per gene into a ‘gene burden’ could allow us to overcome the issue of statistical power and interrogate the association of a single gene with various phenotypes.[13] There are various ways to collapse genetic variants into gene burdens to conduct genetic association studies for rare variants. One apparent place to start is to test the most damaging variants in a gene by collapsing rare predicted loss-of-function (pLOF) variants, which include frameshift insertions/deletions, gain of stop codon (nonsense), and splice site dinucleotides, into a gene burden such that each individual is scored based on the number of rare pLOF alleles carried to ultimately compare rare pLOF carriers versus non-carriers. Importantly, rare pLOF variants are predicted to have the largest effect on phenotypes through a mechanism of haploinsufficiency for heterozygous carriers, whether it be via nonsense-mediated decay or truncated proteins with altered function. Another approach is to assess predicted deleterious missense (pDM) variants, although they may have smaller effects on phenotypes and their directionality of effect may not always be predictable when compared to pLOF variants. For example, one may select pDM variants for inclusion into gene burdens according to clinical classifications of pathogenicity (*e.g.* ClinVar, InterVar)[14,15] and/or *in silico* algorithmic predictions of probability for being damaging to the gene product (*e.g.* SIFT, PolyPhen2 HumDiv/HumVar, LRT, MutationTaster, REVEL).[16-20] Yet another method for variant selection for gene burden testing could be to select variants defined by a specific maximum minor allele frequency threshold. Clearly, the list is not limited to these examples, and there are various additional ways to select rare coding variants for inclusion in gene burden tests of association.[8,10,21]

This dissertation entails extensive genome-first analyses that detail various ways to collapse genetic variants into gene burdens to conduct gene-based association studies, whether it be unbiased using a PheWAS approach or targeted to selected quantitative and/or qualitative phenotypes (Figure 1.1). These analyses leveraged the Penn Medicine BioBank (PMBB, University of Pennsylvania), a large academic medical biobank enriched for various human diseases with WES data linked to EHR phenotype data, to test the performance of variant selection methods and to conduct gene burden association tests to uncover new gene biology and function. Importantly, because the PMBB is a healthcare population of ~60,000 individuals with a higher prevalence of disease compared to other population-based biobanks like the UK Biobank, there is more potential for discovery of novel gene-phenotype relationships.[22] PMBB is also an ancestrally diverse population, with about a quarter of the population being of African ancestry. While previous genetic association studies have typically focused on European ancestry individuals, PMBB's ancestral diversity allows us to characterize previously undescribed genetic variants that are enriched among individuals of African ancestry.

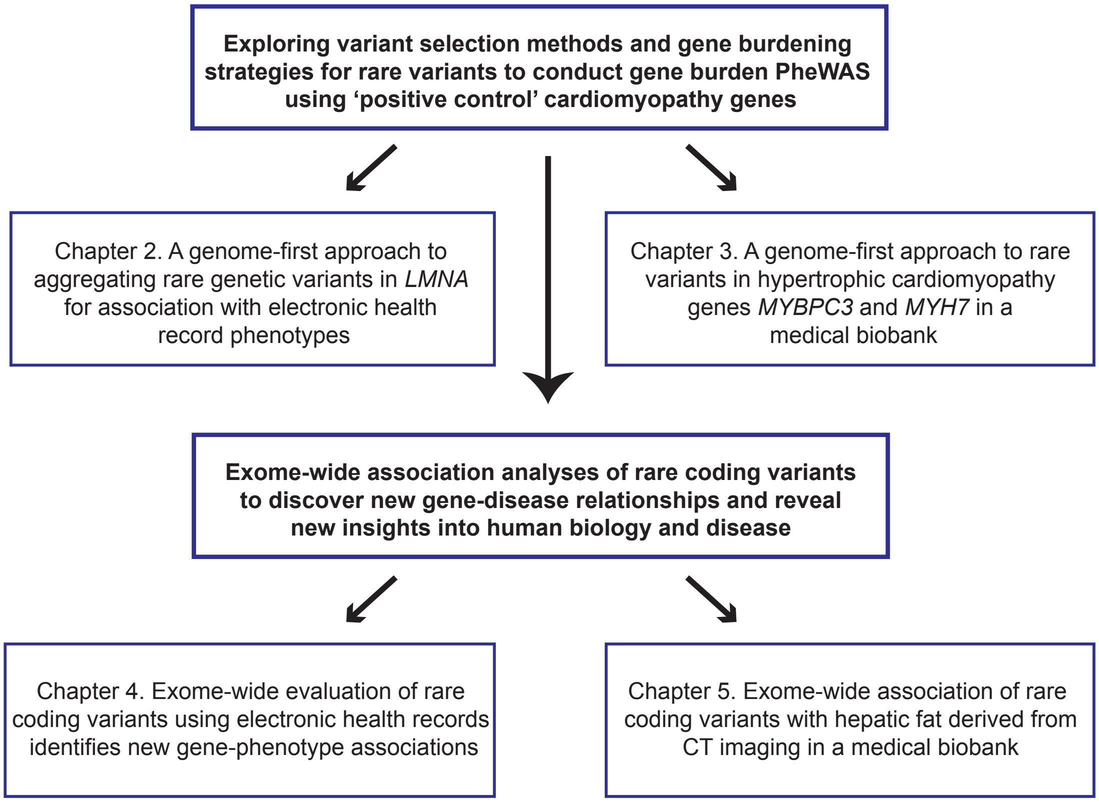
Exploring variant selection methods and gene burdening strategies for rare variants to conduct gene burden PheWAS using 'positive control' cardiomyopathy genes
Chapter 2. A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes
Chapter 3. A genome-first approach to rare variants in hypertrophic cardiomyopathy genes MYBPC3 and MYH7 in a medical biobank
Exome-wide association analyses of rare coding variants to discover new gene-disease relationships and reveal new insights into human biology and disease
Chapter 4. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations
Chapter 5. Exome-wide association of rare coding variants with hepatic fat derived from CT imaging in a medical biobank

Figure 1.1. Flow chart summarizing each experimental dissertation chapter

This dissertation explores genome-first approaches to variant selection methods and gene burdening strategies for rare coding variants to conduct gene burden PheWAS in PMBB. Importantly, Chapters 2 and 3 entail application of the unbiased genome-first approach to 'positive control' genes with known phenotype associations which represent valuable systems for comparison of variant selection methods and gene burdening strategies. More specifically, Chapter 2 focuses on the gene *LMNA*, in which pathogenic variants are known to be highly pleiotropic and cause several rare diseases including dilated cardiomyopathy. This chapter shows how we found the ensemble tool REVEL[20] outperformed other *in silico* prediction algorithms in selecting pDM variants for inclusion in the *LMNA* gene burden to achieve the strongest association with the expected dilated cardiomyopathy phenotype. Subsequent follow-up PheWAS analyses on the ideal *LMNA* gene burden as well as review of clinical charts revealed that pathogenic *LMNA* variants are an underdiagnosed cause of cardiomyopathy and that *LMNA* loss-of-function may be a primary cause of renal disease. Similarly, Chapter 3 focuses on the two genes for which pathogenic variants account for up to 50% of all clinically recognized cases for hypertrophic cardiomyopathy, namely *MYBPC3* and *MYH7*. This chapter shows how we found that the approach to aggregating rare variants for these two genes produced drastically different results: pLOFs but not pDM variants in *MYBPC3* were strongly associated with HCM, whereas pDM but not pLOF variants in *MYH7* were strongly associated with HCM. Importantly, this chapter shows the importance of evaluating both pLOF and pDM variants for gene burden testing for discovery of new gene-disease relationships and identification of new pathogenic loss-of-function variants across the human genome.

Then, in Chapters 4 and 5, the variant selection and gene burdening strategies interrogated in Chapters 2 and 3 are applied on an exome-wide scale to conduct exome-wide association studies for a variety of EHR-derived phenotypes in PMBB. Chapter 4 evaluates rare pLOF variants for each individual gene on an exome-wide scale by applying gene burden PheWAS to every adequately powered gene, essentially representing the first ‘exome-by-phenome-wide association study’ in a medical biobank. In addition to verifying several ‘positive control’ gene-phenotype associations, Chapter 4 details how this approach led to discovery of 21 novel and robustly replicated gene-phenotype relationships through interrogation of pLOF and pDM variants in PMBB, several other medical biobanks, and the population-based UK Biobank. Finally, Chapter 5 evaluates rare variant gene burdens as well as single low-frequency and common coding variants on an exome-wide scale for association with hepatic fat quantifications derived from abdominal CT scans in PMBB, representing the first exome-wide association study for hepatic fat in a medical biobank. Importantly, Chapter 5 confirms that the approaches taken in

Chapters 2-4 for qualitative phenotypes can also be applied to quantitative phenotypes, verifies ‘positive control’ single variants which have previously been associated with differences in hepatic fat or risk for non-alcoholic fatty liver disease (NAFLD), and discovers new single variants and rare variant gene burdens associated with differences in hepatic fat.

In summary, as the genome-first approach is becoming increasingly utilized in parallel to the traditional phenotype-first approach, and the prevalence of high-throughput sequencing in healthcare-based biobank populations rises and subsequently identifies more rare coding variants, this dissertation explores the value and importance of genome-first approaches to investigating the biological and clinical relevance of exome-wide rare coding variation using EHR phenotypes in medical biobanks.

CHAPTER 2. A genome-first approach to aggregating rare genetic variants in *LMNA* for association with electronic health record phenotypes

This chapter was adapted from:

Park J, Levin MG, Haggerty CM, Hartzel DN, Judy R, Kember RL, Reza N; Regeneron Genetics Center, Ritchie MD, Owens AT, Damrauer SM, Rader DJ. A genome-first approach to aggregating rare genetic variants in *LMNA* for association with electronic health record phenotypes. *Genetics in Medicine*. 2020 Jan;22(1):102-111. Doi: 10.1038/s41436-019-0625-8.

2.1. Abstract

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 10,996 unselected individuals in the Penn Medicine BioBank to associate rare variants in *LMNA* with diverse EHR-derived phenotypes. We used 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.

2.2. 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.[4,5] A 'genome-first' approach in which sequencing is applied to large heterogeneous populations with subsequent determination of the associated phenotypes is of interest.[6,7] This approach can be applied to healthcare 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.[8,23] 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.[24]

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 healthcare populations could increase the potential to uncover novel 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,[8] 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.[25-28] 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 (pDM) variants in *LMNA*.

2.3. Materials and Methods

2.3.1. 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.[29]

2.3.2. Exome sequencing

This study included a subset of 11,451 individuals in 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.[30] For subsequent phenotypic analyses, we removed samples with low exome sequencing coverage (*i.e.* less than 75% of targeted bases achieving 20x 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 3rd degree relatives; N=145) with overlap among categories, leading to a total of 455 removed from our database for a total study set of 10,996 individuals. Of note, among the 72 individuals identified as carrying one of pLOF variants or missense variants with Rare Exome Variant Ensemble Learner (REVEL)[20] scores of at least 0.65 who were used for the primary analyses of this work, four 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.[8,31] In addition to exclusions for sequence quality, sample duplicates, and sex discordance, we excluded 31,399 individuals with closer than 3rd degree relatedness, yielding a study set of 61,056 individuals.

2.3.3. Variant annotation and selection for gene burden association testing

For both PMBB and DiscovEHR, variants were annotated using ANNOVAR[32] as predicted loss-of-function (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,[16] PolyPhen2 HumDiv, Polyphen2 HumVar,[17] LRT,[18] MutationTaster[19]). To capture additional individuals with potentially pathogenic missense variants, we utilized an ensemble method for predicting the pathogenicity of missense variants called REVEL to score rare missense variants in *LMNA*.[20]

2.3.4. 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 March 3, 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 CKD-EPI Creatinine equation, given its superiority to the MDRD equation in patient populations with normal or mildly reduced eGFR. Inpatient and outpatient echocardiography measurements were extracted if available for participants from January 1, 2010 until September 9, 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 November 28, 2018. ICD-9 codes were mapped to equivalent ICD-10 codes using underlying diagnosis codes.

2.3.5. 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.[33] 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".[34] Patients were determined to have a certain Phecode if they had the corresponding ICD diagnosis on 2 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.[35] 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.[36] This led to the interrogation of 333 total phenotypes, and we used a Bonferroni correction to adjust for multiple testing (p=0.05/333≈1.5E-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.

2.3.6. 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 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).

2.4. Results

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

Among 10,996 individuals in PMBB with exome sequencing included in this study, 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* (Appendix A, Table S1). All 11 individuals carrying pLOF variants were cases for Phecodes “primary/intrinsic cardiomyopathy,” “cardiac conduction disorders,” or both, confirming that heterozygous pLOF variants in *LMNA* have a high penetrance for cardiomyopathy. Interestingly, only four of these 11 individuals had received clinical genetic testing to confirm their laminopathies.

A PheWAS on the 11 carriers for pLOFs alone showed a signal for cardiomyopathy (Appendix A, Figure 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* (Appendix A, 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 (Appendix A, Figure 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,[16] PolyPhen2 HumDiv, Polyphen2 HumVar,[17] LRT,[18] MutationTaster[19]), one of the standard approaches for combining pLOF variants with computationally predicted pathogenic missense variants[8] (N=14 different variants, 24 carriers; Appendix A, Table S1), in a gene burden PheWAS (N=35 carriers; Figure 2.1.A). The signal for cardiomyopathy diagnoses was even stronger, additionally identifying related Phecodes such as “first degree AV block”, “sinoatrial node dysfunction”, and “congestive heart failure”.

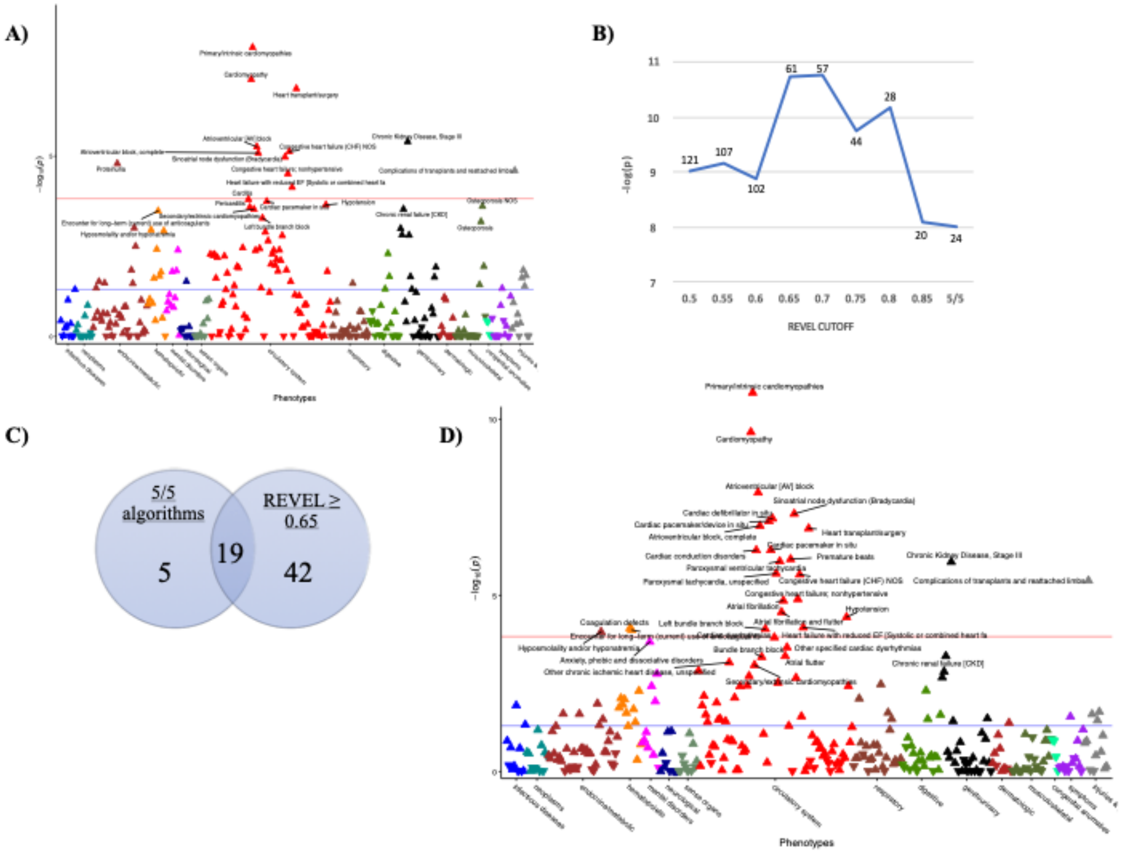

Figure 2.1. Phenome-wide association studies (PheWAS) of predicted deleterious *LMNA* variants. Gene burden tests of association for predicted loss-of-function (pLOF) variants and predicted deleterious missense (pDM) variants in *LMNA*. (A) Gene burden PheWAS of pLOF variants (N=11 carriers) and missense 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” (Appendix A, 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 to other predictive methods.[20] Analysis of variance on ClinVar-annotated variants showed that REVEL scores correlate with clinical pathogenicity (Appendix A, Table S2). While a threshold of 0.5 has been suggested,[20] 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 (Figure 2.1.B). Of note, all REVEL cutoff scores of at least 0.5 performed better in identifying association with “primary/intrinsic cardiomyopathy” compared to 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” (Figure 2.1.B) 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 (Figure 2.1.C). 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 (Figure 2.1.D, Table 2.1). Of note, the signal was more statistically robust compared to other recently developed ensemble methods for predicting pathogenicity such as VEST3[37,38] (Appendix A, Figure S3), M-CAP[39] (Appendix A, Figure S4), and CADD[40] (Appendix A, Figure 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 (Appendix A, Table S3). Importantly, only six of the 35 individuals with a rare deleterious variant in *LMNA* and a Phecode diagnosis of “primary/intrinsic cardiomyopathy” had been molecularly diagnosed with a

LMNA variant (Table 2.1), 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 Phecode diagnosis of “primary/intrinsic cardiomyopathy” and/or “cardiac conduction disorder” (Appendix A, Table S1).

Basic demographics	***LMNA***$^{+/-}$	***LMNA***$^{+/+}$	**OR**	**p-value**
N	68	10928	-	-
Male, N (%)	38 (55.9)	6489 (59.4)	-	0.625
Median Age (at biobank entry), yr	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)	2.47	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

Table 2.1. Demographics, clinical characteristics, and significant cardiovascular Phecode associations for individuals in PMBB carrying a predicted deleterious *LMNA* variant.
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 to non-carriers 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, four of 72 carriers were not included due to additional genotypic quality-check measures (see Methods). Bottom: representative cardiovascular and renal Phecodes identified by gene burden 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 ICD diagnosis on 2 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, EUR-European, SAS-South Asian

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 (Appendix A, Figure 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; Figure 2.1.D, Table 2.1). 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* (Appendix A, Figure S6).

We replicated these observations in the DiscovEHR cohort using the same approach (pLOFs plus REVEL score ≥ 0.65; Appendix A, Table S4A). There was a significant association between *LMNA* gene burden and dilated cardiomyopathy (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; Appendix A, Table S4B).

2.4.2. 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. 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 to the rest of the PMBB population with echo data available (Table 2.2; Appendix A, 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 to individuals not carrying a predicted deleterious *LMNA* variant (Table 2.3; Appendix A, Table S6A). In the overall population, carrier status was significantly associated with increased total cholesterol levels (Table 2.3; Appendix A, Table S6A-B). Furthermore, maximum blood triglyceride levels trended to be elevated among carriers (p=0.0559; Table 2.3). These laboratory features are consistent with subclinical features of partial lipodystrophy, such as fatty liver and dyslipidemia. While only two 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 healthcare 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.

Echo parameter	***LMNA*$^{+/-}$ Median (IQR)** **N**	***LMNA*$^{+/+}$ Median (IQR)** **N**	**ß**	**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

Table 2.2. Cardiac architecture for carriers of presumed deleterious variants in *LMNA* is consistent with dilated cardiomyopathy. 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 PMBB not carrying one of presumed deleterious variants with echocardiographic data available (represented as *LMNA*$^{+/+}$). Data is represented as median, respective 1st and 3rd 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.

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

Table 2.3. Clinical laboratory measurements for carriers of presumed deleterious variants in *LMNA* is consistent with subclinical features of partial lipodystrophy and renal disease. 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 PMBB not carrying one of presumed deleterious variants with serum laboratory data available (represented as *LMNA*$^{+/+}$). Data is represented as median, respective 1st and 3rd quartiles, the number of individuals from each population with available measurement data, and corresponding p-value for Wilcoxon rank sum test.

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 non-carriers in PMBB. We found that carrier status was associated with significantly decreased eGFR and serum albumin levels (Table 2.3;

Appendix A, 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 2.4). 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 2.4).

Lab parameter	**ß**	**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

Table 2.4. Renal clinical laboratory measurements for carriers of presumed deleterious variants in *LMNA* are consistent with primary renal disease. 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 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 and diabetes mellitus (bottom). Data is 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 CKD-EPI equation. Serum albumin additionally adjusted for age and age^2.

2.5. 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.[41,42] 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 healthcare biobanks linked to extensive EHR phenotypic data.

An important area of research in precision medicine initiatives is to create a platform by which healthcare 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 healthcare institutions cover only a small portion of genetic variants implicated in rare human diseases.[43] 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.[44-

[46] 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.[47,48] In the clinical setting, dilated cardiomyopathy patients with confirmed pathogenic *LMNA* variants are often referred for electrophysiologic risk stratification earlier than other patients with non-genetic 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.[49] 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, non-cardiorenal processes have recently been suggested.[50,51] 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 healthcare-based populations.

CHAPTER 3. A genome-first approach to rare variants in hypertrophic cardiomyopathy genes *MYBPC3* and *MYH7* in a medical biobank

This thesis chapter was adapted from:

Park J, Packard EA, Levin MG, Judy RL, Regeneron Genetics Center, Damrauer SM, Day SM, Ritchie MD, Rader DJ. A genome-first approach to rare variants in hypertrophic cardiomyopathy genes *MYBPC3* and *MYH7* in a medical biobank. *medRxiv.* 2021.05.26.21257880. doi: 10.1101/2021.05.26.21257880

3.1. Abstract

‘Genome-first’ approaches to analyzing rare variants can reveal new insights into human biology and disease. Because pathogenic variants are often rare, new discovery requires aggregating rare coding variants into ‘gene burdens’ for sufficient power. However, a major challenge is deciding which variants to include in gene burden tests. Pathogenic variants in *MYBPC3* and *MYH7* are well-known causes of hypertrophic cardiomyopathy (HCM), and focusing on these ‘positive control’ genes in a genome-first approach could help inform variant selection methods and gene burdening strategies for other genes and diseases. Integrating exome sequences with electronic health records among 41,759 participants in the Penn Medicine BioBank, we evaluated the performance of aggregating predicted loss-of-function (pLOF) and/or predicted deleterious missense (pDM) variants in *MYBPC3* and *MYH7* for gene burden phenome-wide association studies (PheWAS). The approach to grouping rare variants for these two genes produced very different results: pLOFs but not pDM variants in *MYBPC3* were strongly associated with HCM, whereas the opposite was true for *MYH7*. Detailed review of clinical charts revealed that only

38.5% of patients with HCM diagnoses carrying an HCM-associated variant in *MYBPC3* or *MYH7* had a clinical genetic test result. Additionally, 26.7% of *MYBPC3* pLOF carriers without HCM diagnoses had clear evidence of left atrial enlargement and/or septal/LV hypertrophy on echocardiography. Our study shows the importance of evaluating both pLOF and pDM variants for gene burden testing in future studies to uncover novel gene-disease relationships and identify new pathogenic loss-of-function variants across the human genome through genome-first analyses of healthcare-based populations.

3.2. Introduction

'Genome-first' approaches, in which genetic variants of interest are first identified and then analyzed for association with phenotypes, can be used to inform the genetic basis of human disease and reveal new insights into gene function and human biology.[6] Particularly when applied to medical biobanks consisting of healthcare populations with DNA sequencing data linked to extensive electronic health record (EHR) phenotype data, genome-first approaches allow for agnostic phenome-wide association studies (PheWAS) to determine the clinical impact of specific genetic variants.[8,9] With the rising use of large-scale whole-exome sequencing (WES) and identification of many rare coding variants predicted to impact protein structure or function, studies are increasingly interrogating the cumulative effect of multiple rare variants in a gene (*i.e.* 'gene burden') to increase the statistical power of regression analyses and enable gene-based association studies to describe the implications of mutated genes in human disease.[13]

We have previously shown that gene burden PheWAS applied to large healthcare populations has the potential to uncover novel consequences of rare coding variants in the human disease phenome.[22] One approach to gene burden PheWAS is to focus only on predicted loss-of-function (pLOF) variants, but this could lead to lack of power due to their infrequency. Additional coding variation could be added to substantially increase the number of effect alleles, but a major challenge is deciding which variants to include in gene burden tests of association. Furthermore, there are many *in silico* algorithms that can predict the probability that a variant may have a deleterious effect on its gene product as well as various filters that can be applied for variant selection based on variant type and frequency, but very little large-scale functional data that can be used to annotate missense variants.

Application of the unbiased genome-first approach to ‘positive control’ genes with known phenotype associations represents a valuable system for comparison of variant selection methods and gene burdening strategies, as we previously showed for the gene *LMNA* and its association with dilated cardiomyopathy.[10] Pathogenic variants in *MYBPC3* and *MYH7* are known to cause hypertrophic cardiomyopathy (HCM) and together account for up to 50% of all clinically recognized HCM cases and at least 75% of HCM cases for which a pathogenic variant is identified.[52] We leveraged the Penn Medicine BioBank (PMBB, University of Pennsylvania), a large academic medical biobank with WES linked to EHR data, to evaluate in detail the performance of methods for aggregating pLOF and/or annotated predicted deleterious missense (pDM) variants in *MYBPC3* and *MYH7* for gene burden association studies. Additionally, we followed up on our genome-first approach with review of EHR charts to describe the clinical characteristics of variant carriers identified through this gene burden study.

3.3. Materials and Methods

3.3.1. 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. This 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.

3.3.2. Whole-exome sequencing

This study included a subset of 43,731 individuals in the PMBB who had undergone whole-exome sequencing. We extracted DNA from stored buffy coats and then mapped exome sequences as generated by the Regeneron Genetics Center (Tarrytown, NY) to GRCh38 as previously described.[10] Samples with low exome sequencing coverage, high missingness (*i.e.* greater than 5% of targeted bases), dissimilar reported and genetically determined sex, and genetic evidence of sample duplication were not included in this subset.[10,22] For subsequent phenotypic association analyses, we removed samples with evidence of 1st and 2nd-degree relatedness, leading to a total of sample size of 41,759 for analysis.

3.3.3. Variant annotation and selection for gene burden association testing

For PMBB, variants were annotated using ANNOVAR[32] as pLOF or missense variants. pLOFs were defined as frameshift insertions or deletions, gain of stop codon, and disruption of canonical splice site dinucleotides. For splicing variants, we removed those with SpliceAI scores < 0.2 for loss or gain of acceptor or donor site.[53] Several approaches to inclusion of rare variants in the gene burden were applied, including pLOFs only, additional ClinVar pathogenic variants, and inclusion of predicted deleterious missense (pDM) variants that were scored deleterious by 4/4 algorithms (SIFT[16], PolyPhen2 HumDiv, Polyphen2 HumVar[17], MutationTaster[19]). To capture additional individuals with potentially pathogenic missense variants, we utilized an ensemble method for predicting the pathogenicity of missense variants called REVEL[20] to score rare missense variants in *MYBPC3* and *MYH7*. Finally, we overlapped a list of all 19 expert-adjudicated pathogenic missense variants in *MYBPC3* from the SHaRe Database to identify high-confidence pathogenic missense variants in PMBB.[54]

3.3.4. Clinical data collection

All International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes, clinical imaging and laboratory measurements were extracted from the patients' EHR. All ICD diagnosis codes and outpatient laboratory measurements available up to July 2020 were extracted for PMBB participants. Inpatient and outpatient echocardiography measurements were extracted if available for participants from September 2005 until November 2018. Outliers for each echocardiographic parameter (values >10 median absolute deviations from the median) were removed. Minimum, median, and maximum measurements of each quantitative trait were recorded per individual.

3.3.5. Phenome-wide association studies

A PheWAS approach was used to determine the phenotypes associated with predicted deleterious variants in *MYBPC3* or *MYH7* carried by individuals in PMBB[33]. 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"[34]. Patients were determined to be a case for a certain Phecode if they had the corresponding ICD diagnosis on 2 or more dates, while controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as individuals meeting control exclusion criteria based on default PheWAS phenotype mapping protocols were not considered in statistical analyses.

Each disease phenotype was tested for association with a gene burden of pLOF and/or pDM variants using a logistic regression model adjusted for age, sex, and the first ten principal components of genetic ancestry. We used an additive genetic model to aggregate variants into gene burdens as previously described.[22] PheWAS analyses were performed separately by African and European genetic ancestry and then combined with inverse variance weighted meta-analysis. Our association analyses considered only disease phenotypes with at least 20 cases based on a prior simulation study for power analysis of rare variant gene burden PheWAS.[22] This led to the interrogation of 1396 total Phecodes, and we used a Bonferroni correction to adjust for multiple testing (p=0.05/1396=3.58E-05).

3.3.6. Statistical analyses

We also created a single HCM phenotype by combining Phecodes "Hypertrophic obstructive cardiomyopathy" and "Other hypertrophic cardiomyopathy" for association with various gene burdens in conjunction with PheWAS, using a logistic regression model adjusted for age, sex, and the first ten principal components of genetic ancestry. These HCM-specific analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. Additionally, to compare available echocardiographic and serum laboratory measurements between carriers of predicted deleterious variants and genotypic controls, we used linear regression adjusted for age, sex, and the first ten principal components of genetic ancestry. These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. For echocardiographic comparison of *MYBPC3* pLOF and pathogenic missense variant carriers versus controls, *MYH7* pLOF and REVEL-informed pDM variant carriers were removed from controls. Likewise, for echocardiographic comparison of carriers of pLOF and pDM with REVEL≥0.5 variants in *MYH7* versus controls, carriers of *MYBPC3* pLOF and pathogenic missense variants were removed from controls. All statistical analyses were completed using R version 3.5 (Vienna, Austria).

3.3.7. Review of clinical charts

We reviewed the clinical charts of patients with HCM who also carry a pLOF or pDM variant in *MYBPC3* or *MYH7* to assess the prevalence of clinical genetic testing for a molecular diagnosis of carrying a pathogenic *MYBPC3* or *MYH7* variant. We also reviewed the clinical charts of *MYH7* pLOF carriers to characterize a cardiac phenotype among these individuals without an HCM diagnosis. Finally, we interrogated the clinical charts of cases for “muscular wasting and disuse atrophy” who also carried a pLOF or pDM variant in *MYH7* to assess the prevalence of clinical genetic testing for a molecular diagnosis of *MYH7*-related myopathy.

3.4. Results

3.4.1. pLOF variants in MYBPC3 were strongly associated with HCM

Among 41,759 unrelated individuals with WES in PMBB (Table 3.1), we identified 45 individuals carrying one of 33 predicted loss-of-function (pLOF) variants in *MYBPC3*, including 13 frameshift insertions/deletions, 9 gain of stop codon, and 11 splicing variants disrupting canonical splice site dinucleotides (Figure 3.1.A; Appendix B, Table S1). PheWAS of the gene burden of pLOF variants in *MYBPC3* showed phenome-wide significant associations with HCM and related cardiac phenotypes such as cardiac conduction disorders, heart failure, heart transplant/surgery, and use of cardiac defibrillator (Table 3.2; Appendix B, Figure S1). 15 of the 45 individuals with a rare pLOF had a clinical diagnosis of HCM (Phecodes “Hypertrophic obstructive cardiomyopathy” or “Other hypertrophic cardiomyopathy”) (Appendix B, Table S1). 12 of 33 pLOFs were annotated in ClinVar as pathogenic or likely pathogenic (P/LP), and of the 18 carriers of these P/LP variants, 6 had a clinical diagnosis of HCM. Of the 21 pLOF variants without a P/LP classification in ClinVar, 9 were carried by a total of 9 individuals with diagnoses of HCM (Appendix B, Table S1).

Basic demographics	
Total population, N	41759
Female, N (%)	20731 (49.6)
Median age, years	63
Genetically informed ancestry	
AFR, N (%)	10217 (24.5)
AMR, N (%)	572 (1.4)
EAS, N (%)	672 (1.6)
EUR, N (%)	29362 (70.3)
SAS, N (%)	564 (1.4)
Phecodes	
Primary/intrinsic cardiomyopathies, N (%)	2912 (7.6)
Hypertrophic obstructive cardiomyopathy, N (%)	199 (0.6)
Other hypertrophic cardiomyopathy, N (%)	184 (0.5)
Cardiac dysrhythmias, N (%)	12130 (35.5)
Atrial fibrillation, N (%)	5885 (21.0)
Atrial flutter, N (%)	2324 (9.5)
Paroxysmal ventricular tachycardia, N (%)	1960 (8.2)
Ventricular fibrillation and flutter, N (%)	467 (2.1)
Cardiac conduction disorders, N (%)	5688 (20.5)
Cardiac pacemaker/device in situ, N (%)	3177 (12.6)
Cardiac defibrillator in situ, N (%)	1945 (8.1)
Congestive heart failure; nonhypertensive, N (%)	6224 (16.9)
Heart transplant/surgery, N (%)	792 (2.5)
Cardiac shunt/heart septal defect, N (%)	547 (1.4)
Cardiogenic shock, N (%)	188 (0.5)
Muscular wasting and disuse atrophy, N (%)	52 (0.1)

Table 3.1. Penn Medicine BioBank whole exome-sequenced cohort characteristics.
Basic demographic characteristics and representative Phecodes identified by gene burden PheWAS for *MYBPC3* and *MYH7*. Each characteristic is labeled with count data and percent prevalence where appropriate. Individuals were determined to be a case for a Phecode if they had the corresponding ICD diagnosis on two or more dates, while controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as those under control exclusion criteria based on Phecode mapping protocols were not considered. AFR-African, AMR-Mixed American, EAS-East Asian, EUR-European, SAS-South Asian

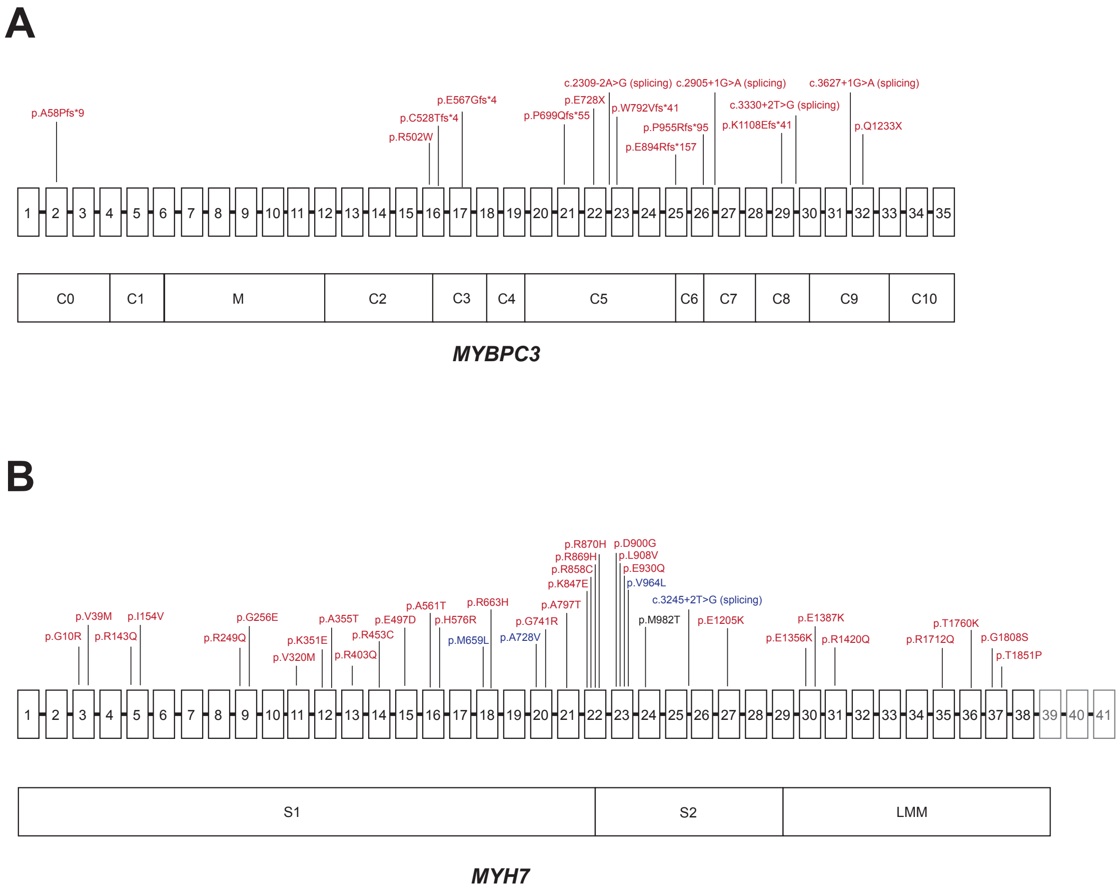

Figure 3.1. Distribution of disease-associated variants in *MYH7 and MYBPC3.* A) Schematic of *MYBPC3* gene with exons 1-35 (exons not to scale) above and domains below, with variants labeled in red denoting amino acid change (or location of splice variant) for pLOF or adjudicated missense variants that were associated with HCM in PMBB. B) Schematic of *MYH7* gene with exons 1-41 (exons not to scale) above and domains below. Variants are labeled in red denoting amino acid change for pDM variants with REVEL≥0.5 that were associated with HCM in PMBB, and variants are labeled in blue denoting amino acid change (or location of splice variant) for pLOF or pDM variants with REVEL≥0.5 that were associated with "Muscular wasting and disuse atrophy" in PMBB. Note: exons 39-41, which do not encode for protein, are grayed out. Note 2: p.M982T in *MYH7* was associated with both HCM and muscular wasting and disuse atrophy, but via different individuals, and is thus labeled in black.

Review of clinical charts of the 15 pLOF carriers with a clinical diagnosis of HCM revealed that only 5 had a clinical genetic test report, all of which were concordant with the WES pLOF results. While all 5 individuals with clinical genetic testing had a family history of HCM, only

3 of 10 untested individuals had a family history of HCM noted in their charts. Chart review of the 30 *MYBPC3* pLOF carriers without a clinical HCM diagnosis revealed that 17 had received at least one transthoracic echocardiogram, and 8 of these individuals had left atrial enlargement and/or hypertrophy of the septum or another segment within the left ventricle. Additionally, 10 of 30 pLOF carriers without a diagnosis of HCM had a history of atrial fibrillation.

MYBPC3						
Gene Burden	**Beta**	**SE**	**OR**	**P**	**Carrier N**	**HCM N**
ClinVar P/LP (pLOF + missense)	4.177	0.453	65.161	3.15E-20	45	8
pLOF only	5.112	0.411	166.008	1.52E-35	45	15
pDM (REVEL≥0.6) only	0.908	0.316	2.48	4.05E-03	628	10
Adjudicated missense (SHaRe)	2.785	0.76	16.204	2.46E-04	22	2
pLOF + pDM	1.949	0.214	7.023	7.99E-20	673	25
pLOF + adjudicated missense	4.402	0.321	81.638	1.03E-42	67	17

MYH7						
Gene Burden	**Beta**	**SE**	**OR**	**P**	**Carrier N**	**HCM N**
ClinVar P/LP (pLOF + missense)	4.737	0.301	11.404	1.29E-55	77	22
pLOF only	-10.658	315.401	2.35E-05	9.73E-01	27	0
pDM (REVEL≥0.5) only	2.059	0.183	7.839	2.01E-29	858	37
pLOF + pDM	2.033	0.183	7.635	9.23E-29	885	37

Table 3.2. Association of *MYBPC3* and *MYH7* gene burdens with HCM in PMBB. Summary statistics for gene burden associations with HCM (combining Phecodes “Hypertrophic obstructive cardiomyopathy” and “Other hypertrophic cardiomyopathy”) in PMBB using ClinVar P/LP, pLOF only, missense only (pDM and/or adjudicated), and pLOF + missense variants in *MYBPC3* (top) and *MYH7* (bottom). Each gene burden association is reported as beta, standard error (SE), odds-ratio (OR), p value, the number of carriers for variants included in the gene burden, and the number of carriers having the HCM phenotype.

3.4.2. Predicted deleterious missense (pDM) variants in MYBPC3 were not associated with HCM

Based on ‘phenotype-first’ presentations of HCM, most pathogenic variants in *MYBPC3* are pLOF variants.[55] ClinVar shows that 482 frameshift, nonsense, or splicing variants in *MYBPC3* are classified as P/LP, whereas only 54 missense variants are annotated as P/LP.[14] In PMBB, the gene burden of all nonsynonymous coding variants in *MYBPC3* classified by ClinVar as P/LP (N=45 heterozygous carriers) was strongly associated with HCM as expected (Table 3.2; Appendix B, Figure S2). However, while a gene burden of just the 12 pLOF variants classified as P/LP (N=18 heterozygous carriers) also showed phenome-wide significant associations with HCM Phecodes (Appendix B, Figure S3A), a gene burden of the 17 missense variants classified as P/LP (N=27 heterozygous carriers) showed a much weaker association with HCM that was not phenome-wide significant (p=9.50E-05) (Appendix B, Figure S3B). We previously have shown that prediction of deleteriousness for missense variants using the ensemble tool REVEL[20] is highly correlated with clinical annotations for missense variants in *LMNA*.[10] We similarly applied REVEL to missense variants in *MYBPC3* and found through an analysis of variance on ClinVar-annotated variants that REVEL scores showed essentially no correlation with annotations of clinical pathogenicity for *MYBPC3* (Appendix B, Table S2). We experimented with REVEL score thresholds in bins of 0.05 to evaluate the optimal score cutoff for capturing the most robust association of missense variants in *MYBPC3* associated with HCM. We found that all REVEL cutoff thresholds showed very weak association with HCM, although a REVEL threshold score of 0.6 was relatively optimal (Table 3.2; Appendix B, Figure S4). Additionally, application of the aggregation of predicted deleterious missense (pDM) variants chosen by a consensus of algorithms (SIFT, PolyPhen2 HumDiv, PolyPhen2 HumVar, and MutationTaster)—one of the standard approaches for predicting the deleteriousness of missense variants—also showed weak association with HCM (Appendix B, Figure S4).

Given the weak or insignificant association of missense variants in *MYBPC3* with HCM based on algorithms predicting deleteriousness of missense variants alone, we overlapped a list of 19 expert-adjudicated pathogenic missense variants in *MYBPC3* from the SHaRe Database to identify high-confidence pathogenic missense variants in PMBB.[54] We identified 6 of these high-confidence pathogenic missense variants in *MYBPC3* carried by a total of 22 individuals among the PMBB WES cohort (Appendix B, Table S3). A gene burden including just these 6 missense variants showed a similar degree of association with HCM compared to a gene burden of ClinVar P/LP missense variants, and was more strongly associated with HCM compared to a burden of pDM variants based on all REVEL score thresholds as well as 4/4 algorithms (Table 3.2; Appendix B, Figure S4). 2 of the 22 carriers were diagnosed with HCM, and chart review of the 20 carriers without an HCM diagnosis revealed 5 individuals with mild concentric hypertrophy or dilated atria noted on transthoracic echocardiography.

Including pDMs together with pLOFs in a gene burden has the potential to increase power for finding phenotype associations. When we combined pLOFs with pDMs having REVEL≥0.6, we noted no improvement in the association of the gene burden with HCM (Table 3.2). Only when the adjudicated pathogenic missense variants were included with pLOFs did we see a stronger association with HCM compared with pLOFs alone (Figure 3.2, Table 3.2). We further interrogated this *MYBPC3* gene burden combining pLOFs and adjudicated pathogenic missense variants by analyzing available echocardiography data in PMBB. Carriers of pLOF and adjudicated pathogenic missense variants in *MYBPC3* on average had increased left ventricular posterior wall (LVPW) diastolic thickness, interventricular septum (IVS) diastolic thickness, left atrial (LA) volume index, and left ventricular outflow tract (LVOT) peak gradient (Table 3.3) compared to non-carriers, although not always at clinically relevant thresholds.

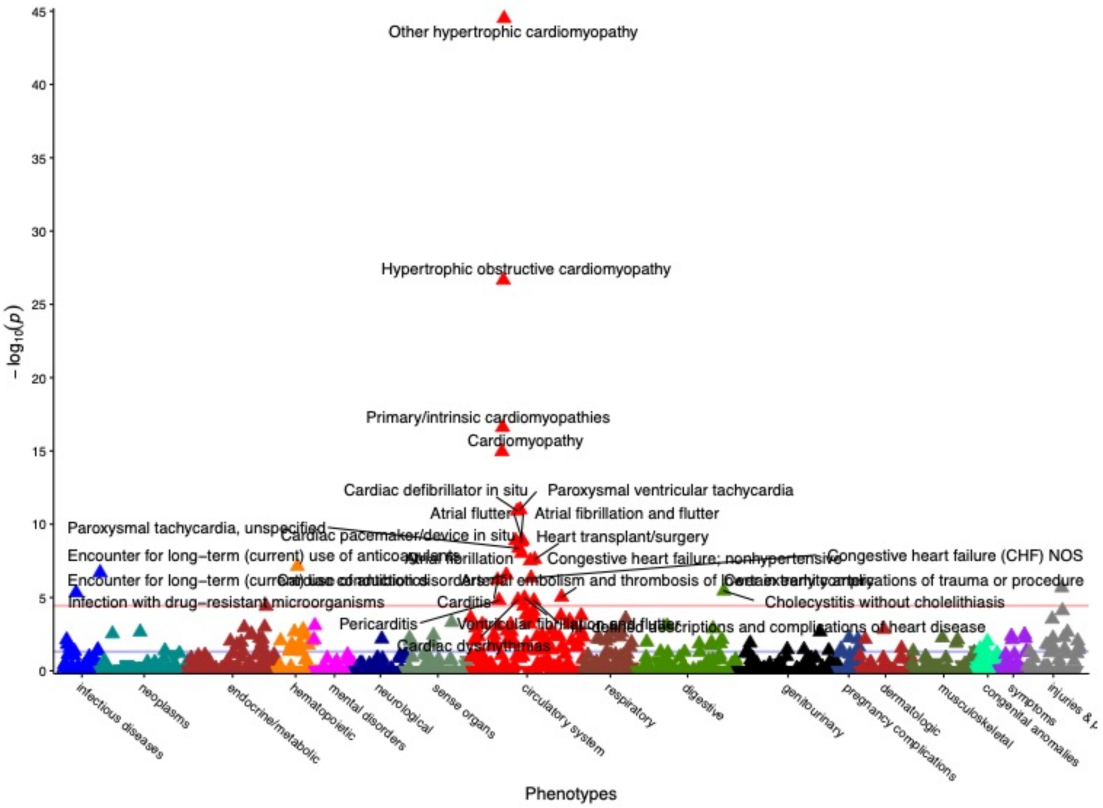

Figure 3.2. Gene burden PheWAS of pLOF and adjudicated pathogenic missense variants in *MYBPC3.* Gene burden PheWAS of pLOF variants (N=45, Appendix B, Table S1) and adjudicated pathogenic missense variants from SHaRE (N=22, Appendix B, Table S3) in *MYBPC3*. Phecodes are plotted along the x axis to represent the phenome, and the association of the gene burden with each Phecode is plotted along the y axis representing –log$_{10}$(p value). The red line represents the Bonferroni-corrected significance threshold to adjust for multiple testing (p=3.58E-05), and the blue line represents a nominal significance threshold (p=0.05).

3.4.3. pLOF variants in MYH7 were not significantly associated with HCM but had suggestive associations with other cardiac phenotypes

We identified 27 individuals carrying one of 23 pLOF variants in *MYH7*, including 5 frameshift insertion/deletions, 13 gain of stop codon, and 5 splicing variants (Appendix B, Table S4). Of note, none of these pLOF variants were in the last exon of *MYH7.* In contrast to *MYBPC3*, PheWAS of the *MYH7* pLOF only gene burden showed no association with HCM (p=0.973) (Table 3.2; Appendix B, Figure S5). We confirmed through chart review that none of the 27 pLOF carriers had an HCM diagnosis (Appendix B, Table S4). Interestingly, however, this pLOF gene burden had weak evidence for association with "Cardiac shunt/heart septal defect" (p=1.12E-04, N=3), suggesting the possibility of a cardiac phenotype among heterozygous carriers for pLOF variants in *MYH7*. Detailed review of the clinical charts of the 27 *MYH7* pLOF carriers confirmed a history of patent foramen ovale or atrial septal defect in 3 pLOF carriers, and also revealed that 13 of 27 had received echocardiograms, of which 5 individuals had mild concentric left ventricular hypertrophy.

Echo Parameter	***MYBPC3* Median (IQR)**	**Control Median (IQR)**	**Beta**	**SE**	**P**	***MYBPC3 N***	**Control N**
LVPW diastolic thickness, max (cm)	1.320 (1.160, 1.525)	1.100 (0.951, 1.284)	0.257	0.043	2.86E-09	31	11928
IVS diastolic thickness, max (cm)	1.510 (1.203, 1.871)	1.150 (0.989, 1.340)	0.414	0.050	7.68E-17	30	11854
LA volume index, max (mL/m^2)	48.593 (40.224, 52.374)	33.725 (24.929, 44.948)	12.567	3.744	7.89E-04	22	8128
LVOT peak gradient, max (mmHg)	6.554 (4.162, 9.000)	4.410 (3.254, 6.052)	2.605	0.500	1.91E-07	29	8807
Echo Parameter	***MYH7* Median (IQR)**	**Control Median (IQR)**	**Beta**	**SE**	**P**	***MYH7 N***	**Control N**
LVPW diastolic thickness, max (cm)	1.130 (0.934, 1.285)	1.100 (0.951, 1.284)	0.045	0.014	1.63E-03	303	11928
IVS diastolic thickness, max (cm)	1.190 (1.001, 1.390)	1.150 (0.989, 1.340)	0.071	0.016	1.09E-05	300	11854
LA volume index, max (mL/m^2)	36.796 (27.763, 50.182)	33.725 (24.929, 44.948)	5.235	1.242	2.51E-05	214	8128
LVOT peak gradient, max (mmHg)	4.801 (3.467, 6.554)	4.410 (3.254, 6.052)	0.458	0.187	1.41E-02	222	8807

Table 3.3. Echocardiographic analyses for *MYBPC3* and *MYH7* in PMBB. Top: comparison of echocardiographic parameters representative of HCM with gene burden of pLOF variants and adjudicated pathogenic missense variants from SHaRE in *MYBPC3* versus non-carriers. Carriers of pLOF variants and

pDM variants with REVEL≥0.5 in *MYH7* were removed from this analysis. Median and interquartile ranges (IQR) for each echo parameter are listed for each test group, and association results are listed as beta, standard error (SE), and p value. The number of individuals with each echo parameter available are also listed per test group. Bottom: comparison of echocardiographic parameters representative of HCM with gene burden of pLOF variants and pDM variants with REVEL≥0.5 in *MYH7* versus non-carriers. Carriers of pLOF variants and adjudicated pathogenic missense variants from SHaRE in *MYBPC3* were removed from this analysis. Median and interquartile ranges (IQR) for each echo parameter are listed for each test group, and association results are listed as beta, standard error (SE), and p value. The number of individuals with each echo parameter available are also listed per test group. LVPW – left ventricule posterior wall, IVS – interventricular septum, LA – left atrial, LVOT – left ventricular outflow tract

3.4.4. Predicted deleterious missense (pDM) variants in MYH7 were strongly associated with HCM

Most known pathogenic variants in *MYH7* are missense[56], as exemplified by the ClinVar database which has classified 272 missense variants in *MYH7* as P/LP, whereas there are only 20 frameshift, nonsense, or splicing variants annotated as P/LP.[14] In PMBB, we found a total of 43 missense variants annotated by ClinVar as P/LP (N=75 heterozygous carriers) as well as just two pLOFs annotated as P/LP (N=2 carriers). As expected, a gene burden comprised only of these ClinVar P/LP nonsynonymous variants had a very strong association with HCM (Table 3.2; Appendix B, Figure S6). Of the 75 carriers with ClinVar P/LP missense variants, 22 had a diagnosis of HCM, while neither of the two carriers for pLOFs annotated in ClinVar as P/LP had an HCM diagnosis.

We then asked how a computational approach to selecting *MYH7* missense variants would perform and explored in detail the association of pDM variants in *MYH7* with HCM and other phenotypes. We applied REVEL to missense variants in *MYH7* and found through an analysis of variance on ClinVar-annotated variants that REVEL scores were highly correlated with annotations of clinical pathogenicity for *MYH7* (Appendix B, Table S5). We then experimented with REVEL score thresholds in bins of 0.05 to evaluate the optimal score cutoff for capturing the most robust association of missense variants in *MYH7* with HCM, with the goal of potentially capturing deleterious missense variants in *MYH7* that lack a pathogenic ClinVar classification. We found that a REVEL cutoff score of 0.5 had the optimal p value for association of *MYH7* pDM variants with HCM (Table 3.2; Appendix B, Figure S7). Of note, this REVEL-based association was more strongly associated with HCM compared to the aggregation of pDM variants predicted deleterious by a consensus of algorithms (SIFT, PolyPhen2 HumDiv, PolyPhen2 HumVar, and MutationTaster), even while identifying more variants (303 variants with REVEL≥0.5 vs. 166 variants passing consensus of algorithms) (Appendix B, Figure S7). The HCM-prevalent missense variants were concentrated among the globular S1 head and coiled-coil S2 domains of *MYH7* (Figure 3.1.B; Appendix B, Table S6). Among 858 total carriers for pDM variants in *MYH7* with REVEL≥0.5, 37 heterozygous carriers for 33 different pDM variants had a diagnosis of HCM (Appendix B, Table S6). 15 *MYH7* pDM variants with REVEL≥0.5 lacking a classification of P/LP in ClinVar were carried by individuals with a diagnosis of HCM (Appendix B, Table S6). Chart review of these 37 carriers confirmed the diagnosis of HCM and indicated that only 15 of the 37 carriers had clinical genetic testing for HCM in their chart (all were concordant with the WES *MYH7* variant). While 13 of 15 individuals with a clinical genetic test in the chart noted a family history of HCM, only half of untested individuals had a documented family history.

3.4.5. Gene burden of pLOF and pDM variants in MYH7 was also associated with a skeletal muscle phenotype

In order to increase power for finding associations, we aggregated the 23 pLOF variants in *MYH7* (N=27 carriers) with the 303 pDM variants with REVEL≥0.5 in *MYH7* (N=858 carriers) into a gene burden. PheWAS of this combined *MYH7* gene burden showed, as expected, phenome-wide significant associations with HCM (Figure 3.3, Table 3.2), but the strength of this association was not greater than with a pDM gene burden alone. This expanded gene burden was also significantly associated with related cardiac phenotypes “Heart transplant/surgery” and “Cardiac defibrillator in situ” (Figure 3.3). We also linked the expanded gene burden with available echocardiography data in PMBB, and found that carriers of pLOF and pDM variants in *MYH7* on average had increased left ventricular posterior wall (LVPW) diastolic thickness, interventricular septum (IVS) diastolic thickness, left atrial (LA) volume index, and left ventricular outflow tract (LVOT) peak gradient (Table 3.3) compared to non-carriers, although not always at clinically relevant thresholds.

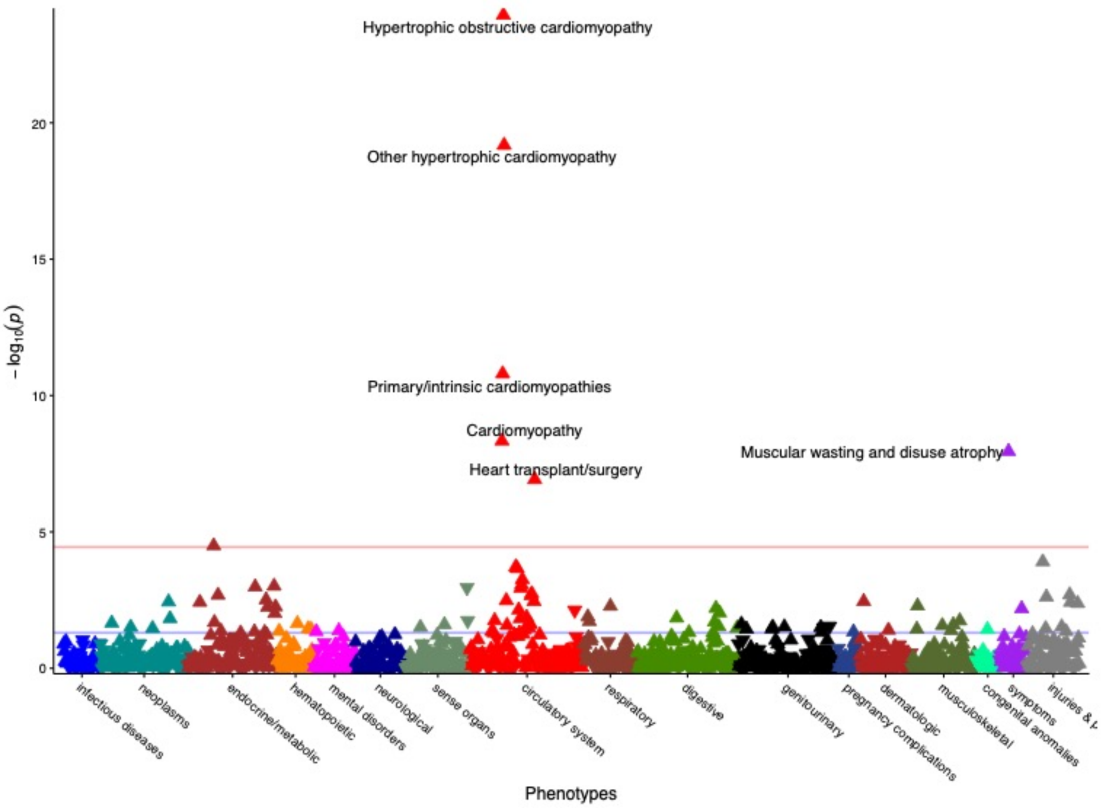

Figure 3.3. Gene burden PheWAS of pLOF + pDM variants in *MYH7*. Gene burden PheWAS of pLOF variants (N=27, Appendix B, Table S4) and pDM variants with REVEL≥0.5 (N=858, Appendix B, Table S6) in *MYH7*. Phecodes are plotted along the x axis to represent the phenome, and the association of the gene burden with each Phecode is plotted along the y axis representing $-\log_{10}$(p value). The red line represents the Bonferroni-corrected significance threshold to adjust for multiple testing (p=3.58E-05), and the blue line represents a nominal significance threshold (p=0.05).

Additionally, we found a phenome-wide significant association with “Muscular wasting and disuse atrophy” with this expanded gene burden (Figure 3.3) that was not seen with the gene burden limited to ClinVar P/LP variants alone (Appendix B, Figure S6). We identified 1 pLOF variant and 4 missense variants with REVEL≥0.5 carried by a total of 7 individuals who had the Phecode of “Muscular wasting and disuse atrophy” (Figure 3.1.B; Appendix B, Table S4, Table S6). Of note, all 5 variants lacked a P/LP classification in ClinVar, and none of the 7 individuals had a diagnosis for HCM. Chart review of these individuals revealed secondary myopathy and generalized muscular wasting rather than a primary myopathy diagnosis. Of those individuals with electromyography performed, none revealed a primary neuromuscular diagnosis. We also compared available serum creatine kinase (CK) measurements between carriers for pLOF variants or pDM with REVEL≥0.5 in *MYH7* (total of 73 carriers with CK available) versus all non-carriers, and there was no significant association between the *MYH7* gene burden and CK concentrations (p=0.549).

3.5. Discussion

Gene burden PheWAS applied to large healthcare-based biobanks has the potential to elucidate the medical consequences of gene variants on the human disease phenome.[22] A logical first step is to perform PheWAS focused only on predicted loss-of-function (pLOF) variants, which has the advantage of interrogating the largest effect sizes that a gene burden may have on associated phenotypes, but could lead to lack of power due to their infrequency. In aggregating pLOF variants in *MYBPC3* and *MYH7* for gene burden PheWAS in PMBB, only the *MYBPC3* pLOF gene burden was associated with HCM, while the *MYH7* pLOF gene burden was associated with other cardiac phenotypes and not HCM. This is an expected finding given that truncating variants account for >90% of cases of *MYBPC3* HCM,[57] while only missense *MYH7* variants have a demonstrated association with HCM.[58] Furthermore, detailed review of clinical charts confirmed a diagnosis of HCM among a subset of *MYBPC3* pLOF carriers while revealing mild concentric left ventricular hypertrophy and patent foramen ovale or atrial septal defect among a subset of *MYH7* pLOF carriers. Functional studies have shown that pLOF variants in *MYBPC3* promote nonsense-mediated decay pathways to contribute to the pathogenesis of HCM through haploinsufficiency.[55,59] pLOF variants in *MYH7*, on the other hand, are not associated with HCM; however, recently pLOF variants in *MYH7* have been associated with left ventricular noncompaction cardiomyopathy, which has low diagnostic accuracy on echocardiography and thus could have been missed in our review of clinical charts.[60]

In principle, missense variants could be combined with pLOFs to substantially increase the number of effect alleles and power for novel discovery, but a major challenge is deciding which variants to include in gene burden tests of association. Including predicted deleterious missense (pDM) variants based on *in silico* prediction algorithms in addition to pLOFs in gene burdens increases power for disease associations in some genes, but the performance of such algorithms for missense variants may not be consistent across all genes. Our interrogation of missense variants based on prediction algorithms like REVEL serves as a testament to this notion; while pDM variants in *MYH7* showed strong correlation with clinical annotations of pathogenicity as well as strong associations with the expected HCM phenotype in PMBB, predictions of deleteriousness for missense variants in *MYBPC3* did not correlate with clinical annotations of pathogenicity and were not associated with HCM. The difference in performance may be due to different mechisms of HCM pathogenesis. For example, *MYH7* missense variants exert dominant negative effects on sarcomeric function, while haploinsufficiency and allelic imbalance in *MYBPC3* is the major mechanism leading to HCM.[61,62]

Importantly, these differences show that aggregating only pLOFs for gene burden association testing may be more appropriate for some genes like *MYBPC3*, while inclusion of additional pDM variants chosen based on *in silico* predictions may be more beneficial for increasing the number of effect alleles for others like *MYH7*, as we have similarly shown for

LMNA.[10] Additionally for *MYBPC3*, in which *in silico* prediction of missense variants performed poorly, we show that expert-adjudication of annotations of pathogenicity for missense variants based on *in vitro* and *in silico* analyses of missense variants in SHaRe, the largest comprehensive HCM cohort assembled to date, is superior in selecting missense variants for addition to gene burden testing.[54,57] For other genes in which *in silico* predictions perform poorly for missense variants, we suggest that high-confidence pathogenic missense variants be added after expert adjudication to a degree that essentially parallels a saturated mutagenesis experiment.

A key goal in precision medicine initiatives is to promote a platform by which healthcare providers can make accurate diagnoses based on a wide variety of personalized health data, including individuals' genetic information. Genetic testing of patients with HCM can identify the precise genetic cause of disease, improve diagnostic accuracy in a patient with an ambivalent diagnosis, and allow for cascade screening in family members.[63] However, pathogenic HCM-causing variants found on more comprehensive sequencing can also be missed in clinical practice, which makes the genome-first approach applied to WES for review of clinical charts for carriers of pathogenic and potentially deleterious variants crucial.[64] Our review of clinical charts revealed that a substantial number of patients clinically diagnosed with HCM did not appear to have had genetic testing, including those who were found in WES to have a pathogenic HCM-associated variant in *MYBPC3* or *MYH7*. We found that about a third of these HCM individuals without a genetic diagnosis were diagnosed with HCM before age 30. We also noted that patients with HCM without a genetic test in their chart had a significantly lower rate of a family history of HCM versus those who received genetic testing. While genetic testing is now routinely offered in specialized cardiomyopathy clinics, it is still broadly underutilized and patients may benefit from referral to specialized HCM clinics for genetic diagnosis and family screening.[65] Additionally, we found that on average carriers for P/LP or predicted deleterious variants without a diagnosis of HCM had increased LA size and LV wall thickness compared to non-carriers, subclinical features that may warrant clinical follow-up when noted on echocardiography. Importantly, many carriers for P/LP or predicted deleterious variants without a diagnosis of HCM had no clinical or echocardiographic evidence of cardiac disease, showing the value of a genome-first approach for estimating penetrance of single-gene Mendelian disorders like HCM.

A major advantage of the genome-first approach to conducting gene burden PheWAS is the potential to capture pleiotropy. We found that the pLOF + pDM (REVEL ≥ 0.5) gene burden for *MYH7* had phenome-wide significant associations with both HCM and "Muscular wasting and disuse atrophy." Importantly, while *MYBPC3* is specifically expressed in the heart, *MYH7* is also expressed in skeletal muscle,[66] and *MYH7*-related myopathies are an emerging and underdiagnosed group of muscle diseases of childhood and adulthood.[67] It has been reported that variants in *MYH7* which cause skeletal muscle disorders may cluster in the distal regions of the rod domain (light meromyosin domain, LMM with or without cardiac involvement).[67] We found that myopathy-associated variants in *MYH7* were located in the neck and hinge (distal S1 and S2 domains) of *MYH7* from exons 18-24, suggesting that variants which affect skeletal muscle function may not be limited to the LMM domain. Of note, patients diagnosed with "muscular wasting and disuse atrophy" who carried a predicted deleterious *MYH7* variant did not have a primary myopathy diagnosis but rather had secondary muscular wasting, indicating that *MYH7*-related myopathy may be a 'second hit' phenomenon that can be unmasked by comorbidities.

In conclusion, we used a genome-first approach to include pLOFs and selected missense variants in *MYBPC3* and *MYH7* in gene burdens to show their significant associations with HCM, as well as the *MYH7*-specific association with myopathy. We demonstrate gene-specific differences in appropriate coding variant selection for gene burden testing to uncover important clinical and subclinical features relevant to associated diseases implicated by predicted deleterious variants. Our approach also suggests an expanded opportunity for clinical genetic evaluation and referral to multidisciplinary HCM centers. Importantly, our study demonstrates the value of assessing both pLOF and pDM variants for gene burden testing in future studies to uncover novel gene-disease relationships and identify novel pathogenic loss-of-function variants across the human genome through genome-first analyses of healthcare-based populations.

CHAPTER 4. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations

This thesis chapter was adapted from:

Park J, Lucas AM, Zhang X, Chaudhary K, Cho JH, Nadkarni G, Dobbyn A, Chittoor G, Josyula NS, Katz N, Breeyear JH, Ahmadmehrabi S, Drivas TG, Chavali VRM, Fasolino M, Sawada H, Daugherty A, Li Y, Zhang C, Bradford Y, Weaver J, Verma A, Judy RL, Kember RL, Overton JD, Reid JG, Ferreira MAR, Li AH, Baras A, LeMaire SA, Shen YH, Naji A, Kaestner KH, Vahedi G, Edwards TL, Chen J, Damrauer SM, Justice AE, Do R, Ritchie MD, Rader DJ. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. *Nature Medicine*. 2021 Jan;27(1):66-72. doi: 10.1038/s41591-020-1133-8.

4.1. Abstract

The clinical impact of rare loss-of-function variants has yet to be determined for most genes. Integration of DNA sequencing data with electronic health records (EHR) could enhance our understanding of the contribution of rare genetic variation to human disease.[8] By leveraging 10,900 whole exome sequences linked to EHR data in the Penn Medicine BioBank (PMBB), we addressed the association of the cumulative effects of rare predicted loss-of-function (pLOF) variants per individual gene on human disease on an exome-wide scale, as assessed using a set of diverse EHR phenotypes. After discovering 97 genes with exome-by-phenome-wide significant phenotype associations ($p < 10^{-6}$), we replicated 26 of these in PMBB, as well as in three other medical biobanks and the population-based UK Biobank (UKB). Of these 26 genes, five had associations that have been previously reported and represented positive controls, whereas 21

had phenotype associations not previously reported, among which were genes implicated in glaucoma, aortic ectasia, diabetes mellitus, muscular dystrophy, and hearing loss. These findings show the value of aggregating rare pLOF variants into "gene burdens" for identifying new gene-disease associations using EHR phenotypes in a medical biobank. We suggest that application of this approach to even larger numbers of individuals will provide the statistical power required to uncover unexplored relationships between rare genetic variation and disease phenotypes.

4.2. Introduction

A "genome-first" approach, in which genetic variants of interest are identified and then subsequently associated with phenotypes, has the potential to inform the genetic basis of human disease and reveal new insights into gene function and human biology.[6] This approach can be applied to "medical" biobanks consisting of healthcare populations with DNA sequence data linked to extensive EHR phenotype data, thus permitting "phenome-wide association studies" (PheWAS) as an agnostic approach to determining the clinical impact of specific genetic variants.[9] Genome-first approaches utilizing PheWAS have primarily focused on individual common variants of modest effect.[12] Very rare and private coding variants are more likely to have larger effect sizes and are of great interest, but are generally too rare to study in a univariate fashion.[11] Aggregation of multiple rare variants in a gene (*i.e.* "gene burden") not only increases the statistical power of regression analyses but also enables gene-based association studies to describe the clinical implications of loss of gene function in human disease.[13]

Previously, we leveraged the Penn Medicine BioBank (PMBB, University of Pennsylvania), a large academic medical biobank with whole-exome sequencing (WES) data linked to EHR data, to show that aggregating rare, loss-of-function variants in a single gene or targeted sets of genes to conduct gene burden PheWAS has the potential to uncover novel pleiotropic relationships between the gene and human disease.[10,68] We applied rare pLOF-based gene burden PheWAS on an exome-wide scale, utilizing WES data to conduct exome-by-phenome-wide association studies (ExoPheWAS) to evaluate in detail the clinical phenotypes (*i.e.* phecodes) associated with rare pLOF variants on a gene-by-gene basis across the human exome, and replicated our top results in several other medical biobanks.

4.3. Materials and Methods

4.3.1. Setting and study participants

All individuals who were 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, access to all available electronic health record (EHR) data, and permission to recontact for future studies. 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.

In addition to our robustness validation analyses within PMBB, replication analyses were conducted using the WES dataset from an additional set of independent African-American individuals in PMBB (PMBB2), BioMe, DiscovEHR, UK Biobank (UKB), as well as imputed genotype data in BioVU, for evaluation of the robustness of gene-phenotype associations identified in PMBB. For replication analyses in BioMe, DiscovEHR, and BioVU, each study was approved by the Institutional Review Board of each respective biobank’s institution. Access to the UK Biobank for this project was from Application 32133.

4.3.2. Genetic sequencing

This PMBB study dataset included a subset of 11,451 individuals in the PMBB who have undergone whole-exome sequencing (WES). For each individual, we extracted DNA from stored buffy coats and then obtained exome sequences generated by the Regeneron Genetics Center (Tarrytown, NY). These sequences were mapped to GRCh37 as previously described.[10] Furthermore, for subsequent phenotypic analyses, we removed samples with low exome sequencing coverage (*i.e.* less than 75% of targeted bases achieving 20x coverage), high missingness (*i.e.* greater than 5% of targeted bases), high heterozygosity, dissimilar reported and genetically determined sex, genetic evidence of sample duplication, and cryptic relatedness (*i.e.* closer than 3rd degree relatives), leading to a total of 10,900 individuals.

For replication studies in PMBB2, we interrogated an additional 6,935 individuals of African American ancestry in PMBB who were exome-sequenced by the Regeneron Genetics Center. We focused our replication efforts on 6,432 individuals after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, and those with dissimilar reported and genetically determined sex. These sequences were mapped to GRCh38.

For replication studies in BioMe, we interrogated 6,470 individuals of African ancestry, 8,735 individuals of European ancestry, and 8,784 individuals of Hispanic ancestry with WES data linked to EHR diagnosis phenotypes after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, and those with dissimilar reported and genetically determined sex. These sequences were mapped to GRCh38.

For replication studies in DiscovEHR, we interrogated 70,734 individuals of European ancestry exome-sequenced on the IDT platform and a separate set of 59,133 individuals of European ancestry exome-sequenced on the VCRome platform. We focused our replication efforts on 85,450 individuals (N=48,413 for IDT, N=37,037 for VCRome) after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, those with dissimilar reported and genetically determined sex, and those that self-identified as Hispanic/Latino. These sequences were mapped to GRCh38.

For replication studies in UKB, we interrogated the 34,629 individuals of European ancestry (based on UKB's reported genetic ancestry grouping) with ICD-10 diagnosis codes available among the 49,960 individuals who had WES data as generated by the Functional Equivalence (FE) pipeline. We focused our replication efforts on 32,268 individuals after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, and those with dissimilar reported and genetically determined sex. The PLINK files for exome sequencing provided by UKB were based on mappings to GRCh38.

For replication studies in BioVU, which has genotype but not large-scale WES data, we focused on a select group of single variants that showed replication in PMBB, PMBB2, and/or UKB. We interrogated these variants for association with specific phecodes in 10,456 individuals of African American ancestry and 55,944 individuals of European ancestry after removing samples with poor genotype quality, individuals closer than 3rd degree relatives, and those with dissimilar reported and genetically determined sex. These sequences were mapped to GRCh37.

Additional information regarding population characteristics, recruitment, and ethical oversight can be found in the Life Sciences Reporting Summary of this study.

4.3.3. Variant annotation and selection for association testing

For all cohorts analyzed, genetic variants were annotated using ANNOVAR (version 2018Apr16)[32] as predicted loss-of-function (pLOF) or missense variants according to the NCBI Reference Sequence (RefSeq) database. pLOF variants were defined as frameshift insertions/deletions, gain/loss of stop codon, or disruption of canonical splice site dinucleotides. Predicted deleterious missense (pDM) variants were defined as those with Rare Exonic Variant Ensemble Learner (REVEL)[20] scores ≥ 0.5. Minor allele frequencies for each variant were determined per Non-Finnish European, African, and Latino minor allele frequencies reported by the Genome Aggregation Database (gnomAD) v2.[69] pLOF and REVEL-informed missense variants were selected for gene burden testing or univariate association analyses per ancestry group in each cohort according to each ancestry's corresponding ancestry-specific minor allele frequency thresholds (rare variants with MAF ≤ 0.1% for gene burden testing, single variants with MAF > 0.1%).

4.3.4. Clinical data collection

International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) disease diagnosis codes and procedural billing codes, medications, and clinical imaging and laboratory measurements were extracted from the patients' EHR for PMBB. 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) via Phecode Map 1.2 using the R package "PheWAS".[34] 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.

All laboratory values measured in the outpatient setting were extracted for participants from the time of enrollment in PMBB until March 20, 2019; all units were converted to their respective clinical Traditional Units. Minimum, median, and maximum measurements of each laboratory measurement were recorded for each individual and used for all association analyses. Inpatient and outpatient echocardiography measurements were extracted if available for participants from January 1, 2010 until September 9, 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 for each patient and used for association analyses.

ICD-9 and ICD-10 codes were similarly mapped to phecodes in PMBB2, BioMe, DiscovEHR, and BioVU for replication studies. For UKB, we used the provided ICD-10 disease diagnosis codes for replication studies, and individuals were determined to have a certain disease phenotype if they had one or more encounters for the corresponding ICD diagnosis given the lack of individuals with more than two encounters per diagnosis, while phenotypic controls consisted of individuals who never had the ICD code. Individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses.

4.3.5. Association studies

A phenome-wide association study (PheWAS) approach was used to determine the phenotypes associated with rare (MAF ≤ 0.1% in gnomAD) pLOF variants carried by individuals in PMBB for the discovery experiment.[33] Each disease phenotype was tested for association with each gene burden or single variant using a logistic regression model adjusted for age, age^2, sex, and the first ten principal components (PCs) of genetic ancestry. We used an additive genetic model to collapse variants per gene via the fixed threshold approach.[35] Given the high percentage of individuals of African ancestry present in the discovery PMBB cohort, association analyses were performed separately in European (N=8,198) and African (N=2,172) genetic ancestries and combined with inverse variance weighted meta-analysis. Only genes with at least 25 carriers of pLOFs were analyzed in the discovery analysis (N=1,518). Our association analyses considered only disease phenotypes with at least 20 cases, leading to the interrogation of 1,000 total phecodes. All association analyses were completed using R version 3.3.1 (Vienna, Austria). Power analyses were conducted using QUANTO version 1.2.4.[70]

We further evaluated the robustness of our gene-phenotype associations in the same PMBB discovery cohort by 1) associating the aggregation of rare (MAF ≤ 0.1%) pDM variants in gene burden association tests and 2) testing pLOFs and pDM variants with MAF > 0.1 in univariate association tests. We ensured that individuals were non-overlapping across rare pLOFs, rare deleterious missense, and single variant groups. Rare deleterious missense gene burdens and single variants were analyzed for association with the specific phenotype identified in the pLOF-based gene burden discovery, as well as with related phenotypes in their corresponding phecode families (integer part of phecode). For example, to replicate an association of a gene burden with hypothetical phecode 123.45, we associated other variants in the same gene with phecode 123.45 as well as other related phenotypes under the phecode family 123 (*e.g.* 123.6). Notably, we checked for the presence of mutual carriers between each gene's pLOF-based gene burdens and subsequently interrogated missense-based gene burdens or single variants due to linkage disequilibrium and/or rare chance, and only reported replications for which the significant phenotypes' associations were not being driven by mutual carriers. All association studies in PMBB were based on a logistic regression model adjusted for age, age^2, sex, and the first 10 PCs of genetic ancestry.

Additionally, we replicated our findings in PMBB2, BioMe, DiscovEHR, and UKB for genes of interest using pLOF-based gene burden, REVEL-informed missense-based gene burden, and/or univariate association analyses from discovery in PMBB. A specific set of single variants were further replicated in BioVU. Association statistics were calculated similarly to PMBB, such that each disease phenotype was tested for association with each gene burden or single variant using a logistic regression model adjusted for age, age^2, sex, and the first 10 PCs of genetic ancestry. In BioMe, the summary statistics obtained from running the logistic regression model separately in individuals of European, African, and Hispanic ancestry were meta-analyzed. In DiscovEHR, the summary statistics obtained from running the logistic regression model separately in individuals of European ancestry on the IDT versus VCRome platforms were meta-analyzed. In BioVU, the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry were meta-analyzed. All association analyses for PMBB, PMBB2, BioMe, DiscovEHR, UK Biobank, and BioVU were completed using R version 3.3.1 or later (Vienna, Austria). Further information about association studies in each cohort can be found in the Life Sciences Reporting Summary of this study.

4.3.6. Undercalling of variants in UK Biobank

Given the undercalling of variants largely limited to ~3.25% of the exome target regions in the FE pipeline data, we found that 3 of the 97 genes having associations with p<E-06 from the discovery phase overlap with the undercalled exonic regions, namely *CES5A*, *CYP2D6*, and *ZC3H3*. While all other analyses in this study included variants with less than 5% missingness, we included variants with at least 65% call rate for these three genes, understanding that undercalling per variant is random per individual.

4.3.7. Statistical analyses of clinical measurements

In order to compare available measurements for echocardiographic parameters and serum laboratory values between carriers of predicted deleterious variants and genotypic controls in PMBB, we utilized linear regression adjusted for age, age^2, sex, and the first 10 PCs of genetic ancestry in individuals of European ancestry only. These analyses were conducted with the minimum, median, and maximum value as the dependent variable for each echocardiographic parameter and clinical lab measure. All statistical analyses, including PheWAS, were completed using R version 3.3.1 or later (Vienna, Austria).

4.3.8. Chart review to validate robust gene-phenotype associations

To confirm our curated list of robust exome-by-phenome-wide significant associations, we manually chart reviewed the EHR for each carrier of rare pLOF variants in genes that showed at least one mode of replication in any cohort. Importantly, for each gene, we aimed to adjudicate the diagnoses of carriers who were flagged as cases for the relevant associated phenotype. We removed associations for which chart review reduced the prevalence of the diagnosis among carriers and thus changed the association to p > E-06. Furthermore, we removed associations for which chart review could not identify a common underlying etiology among all cases for the diagnosis, paying special attention to phecodes that group "other" diagnoses that do not fit into disease-specific ICD codes (*i.e.* "other diseases of blood and blood-forming organs").

We discovered on chart review that individuals who were cases for phecodes "hypertrophic obstructive cardiomyopathy" or "other hypertrophic cardiomyopathy" in PMBB were patients with hypertrophic cardiomyopathy who were being assigned one of the codes due to the lack of a single ICD diagnosis code for hypertrophic cardiomyopathy. Thus, we defined a new phenotype for hypertrophic cardiomyopathy encompassing cases for either phecode, and repeated the association with the pLOF gene burdens of *MYBPC3* (positive control) and *BBS10* (novel), and confirmed their associations as exome-by-phenome-wide significant (Appendix C, Table S22).

4.3.9. Analysis of publicly available expression datasets from NCBI GEO

We interrogated microarray and RNA-seq data publicly available on the NCBI Gene Expression Omnibus (GEO) platform (https://www.ncbi.nlm.nih.gov/geo/).[71] To investigate the novel association between *CILP* and aortic ectasia, we interrogated 11 different microarray and RNA-seq datasets of human fibroblasts from various tissues treated with TGF-ß (GSE1724, GSE65069, GSE64192, GSE39394, GSE79621, GSE68164, GSE97833, GSE97823, GSE135065, GSE125519, GSE40266). Differential expression for each dataset was interrogated using the GEO2R software via a moderated t-statistic. Meta-analysis of differential expression across the datasets was achieved using the Fisher's combined probability test. Visualization of the meta-analyzed differential expression was achieved using the R package "MetaVolcanoR 1.0.1". Identification of the top 1% of differentially expressed genes across all datasets was achieved using the Topconfects method.[72]

We also analyzed microarray data from muscle biopsies in tibial muscular dystrophy patients versus control (GSE42806) to validate the novel association between *MYCBP2* and muscle spasms. Differential expression was interrogated using the GEO2R software via a moderated t-statistic.

4.3.10. In silico analyses for PPP1R13L expression in ocular tissues

To understand the functional relevance of *PPP1R13L* in the eye, we evaluated its expression in human ocular tissues using the publicly available Ocular Tissue Database (OTDB; https://genome.uiowa.edu/otdb/).[73] The OTDB consists of gene expression data for eye tissues from 20 normal human donors, generated using Affymetrix Human Exon 1.0 ST arrays and described as Probe Logarithmic Intensity Error (PLIER) values, where individual gene expression values are normalized with its expression in other tissues.

4.3.11. Gene expression in DBA/2J mouse ocular tissues

We assessed the gene expression of *Ppp1r13l* in mouse ocular tissues using the publicly available Glaucoma Discovery Platform (http://glaucomadb.jax.org/glaucoma). This platform provides an interactive way to analyze RNA sequencing data obtained from retinal ganglion cells (RGCs) isolated from retina and optic nerve head of a 9-month-old female D2 mouse, which is an age-dependent model of ocular hypertension/glaucoma, and D2-Gpnmb$^+$ mouse that do not develop high IOP/glaucoma.[74] For transcriptomic studies, four distinct groups were compared based on axonal degeneration and gene expression patterns. The transcriptome of D2 group 1 is identical to the control strain (D2-Gpnmb$^+$), while D2 groups 2–4 exhibit increasing levels of molecular changes relevant to axonal degeneration when compared to control group. We used the Datgan software to assess the differential expression of *Ppp1r13l* in the retina.[75]

4.3.12. Immunolocalization of PPP1R13L in human retina

To study the localization of PPP1R13L protein in different retinal layers of the human eye, we performed immunofluorescence on formalin-fixed paraffin-embedded section (N=3) obtained from normal 68-year old donor's cadaver eyes with a commercially available antibody, anti-PPP1R13L (Cat# 51141-1-AP, Proteintech, IL, USA). Antigen retrieval was performed in 1X citrate buffer (Life Technologies) warmed to 95°C for 30 minutes. Sections were allowed to cool to room temperature and subsequently blocked in 10% normal goat serum with 1% bovine serum albumin in 1X TBS buffer for one hour. The retinal distribution of PPP1R13L protein was visualized by incubating the retinal section with rabbit polyclonal anti-PPP1R13L antibody at 1:300 dilution overnight at 4°C, followed by chicken anti-rabbit IgG conjugated with Alexa Fluor 594 (Cat# A21442, Life Technologies, Carlsbad, CA) at 1:3000 dilution. Nuclei were stained with the use of Vectashield DAPI in the mounting media. The images were captured using a Zeiss Imager Z1 fluorescence microscope equipped with AxioVS40 software version 4.8.1.0.

4.3.13. Human iPSC-RGC cultures

The human iPSCs were generated from keratinocytes or blood cells via polycistronic lentiviral transduction (Human STEMCCA Cre-Excisable constitutive polycistronic [OKS/L-Myc]

Lentivirus Reprogramming Kit, Millipore) and characterized with a hES/iPS cell pluripotency RT-PCR kit.[76] The induced pluripotent stem cell-derived retinal ganglion cells (iPSC-RGCs) for our studies were derived using small molecules to inhibit BMP, TGF-beta (SMAD) and Wnt signaling to differentiate retinal ganglion cells (RGCs) from iPSCs. The iPSCs were differentiated into pure iPSC-RGCs with structural and functional features characteristic of native RGC cells based on a previous protocol.[77]

4.3.14. Evaluating oxidative stress in iPSC-RGCs

Induced pluripotent stem cell-derived retinal ganglion cells (iPSC-RGCs) were incubated with increasing amounts of H_2O_2 overnight before replacing the cultures with complete media. The cells were collected 24 hours after the H_2O_2 treatment, and levels of *PPP1R13L* transcripts were assessed using quantitative RT-PCR and gene expression primers, Fwd-5'-TGCCCCAATTCTGGAGTAGG-3' and Rev-5'- CGGCACGTGGACACAGATT-3' following previously established protocols.[78] Mean expression levels (±standard error of mean) were calculated by analyzing at least three independent samples with replica reactions and presented on an arbitrary scale that represents the expression over the housekeeping gene *ACTB*. Relative gene expression was quantified using the comparative Ct method. The relative gene expression was compared against no treatment control to obtain normalized gene expression. A two-tailed unpaired Student's *t* test was used for statistical analysis.

4.3.15. Single-cell RNA-seq of human pancreatic islets in type 1 diabetes and control subjects

Pancreatic islets were procured from the HPAP consortium under Human Islet Research Network (https://hirnetwork.org/) with approval from the University of Florida Institutional Review Board (IRB # 201600029) and the United Network for Organ Sharing (UNOS). A legal representative for each donor provided informed consent prior to organ retrieval. For type 1 diabetes (T1D) diagnosis, medical charts were reviewed and C-peptide was measured in accordance with the American Diabetes Association guidelines, leading to five individuals with T1D and six control individuals. T1D individuals were 50% female, and had a median age of 29.5 and median BMI of 21.25. Control individuals were 60% female, and had a median age of 13 and median BMI of 17.3. All individuals were of Caucasian race. Organs were recovered and processed as previously described.[79] Pancreatic islets were cultured and dissociated into single cells as previously described.[80] Total dissociated cells were used for single-cell capture for each of the donors.

The Single Cell 3' Reagent Kit v2 or v3 was used for generating scRNA-seq data. 3,000 cells were targeted for recovery per donor. All libraries were validated for quality and size distribution using a BioAnalyzer 2100 (Agilent) and quantified using Kapa (Illumina). For samples prepared using The Single Cell 3' Reagent Kit v2, the following chemistry was performed on an Illumina HiSeq4000: Read 1: 26 cycles, i7 Index: 8 cycles, i5 index: 0 cycles, and Read 2: 98 cycles. For samples prepared using The Single Cell 3' Reagent Kit v3, the following chemistry was performed on an Illumina HiSeq 4000: Read 1: 28 cycles, i7 Index: 8 cycles, i5 index: 0 cycles, and Read 2: 91 cycles. Cell Ranger 2.1.0 (10x Genomics) was used for bcl2fastq conversion using the command "cellranger mkfastq --id= --run= --csv= --localmem=64 --localcores=30". Cell Ranger 2.1.0 was used for aligning, filtering, counting, and cell calling with the command "cellranger count --id= --transcriptome= --fastqs= --localmem=64 --localcores=35". Samples were aggregated using Cell Ranger 2.1.0 using the command "cellranger aggr --id= --csv=".

Seurat 3.0.2 (http://satijalab.org/seurat/)[81] was used for filtering, UMAP generation, and initial clustering. Genes were kept that were in 0.01% of cells (3 cells), resulting in 74% of genes remaining for analysis (24,986 of 33,694 genes). Cells with at least 200 genes were kept; however, all cells had at least 200 genes, so this filtering didn't eliminate any of the 35,134 cells.

nFeature, nCount, percent.mt, nFeature vs nCount, and percent.mt vs nCount plots were generated to ascertain the lenient filtering criteria of 200 > nFeature < 7,500, percent.mt < 30, and nCount <100,000, which led to the filtering out of 66 cells (35,066 cells remaining). Data was then log-normalized, and the top 2,000 variable genes were detected using the “vst” selection method. The data was then linearly transformed, and PCA was carried out on the scaled data, using the 2,000 variable genes as input. To determine the dimensionality of the data (*i.e.* how many principal components to choose when clustering), we employed two approaches: (1) a Jackstraw-inspired resampling test that compares the distribution of p values of each PC against a null distribution and (2) an elbow plot that displays the standard deviation explained by each principal component. Based on these two approaches, 14 PCs with a resolution of 2 was used to cluster the cells, and non-linear dimensionality reduction (UMAP) was used with 14 PCs to visualize the dataset.

DoubletFinder 2.0[82] was used to demarcate and remove potential doublets in the data as previously described, with the following details: paramSweep_v3 was used, doubletFinder_v3 was used, 14 PCs were used for pK identification (no ground-truth), and the following parameters were used when running doubletFinder_v3: PCs = 14, pN = 0.25, pK =0.005, nExp = nEx_poi.adj, sct = FALSE. The doublets had higher nCount than the singlets identified using this method, and the 807 doublets were removed from further analyses.

Following doublet removal, the raw data for the remaining 34,259 cells was log normalized, the top 2,000 variable genes were detected, the data underwent linear transformation, and PCA was carried out, as described above. Both the Jackstraw-inspired resampling test and an elbow plot of standard deviation explained by each principal component were used to determine the optimal dimensionality of the data, as described above. Based on these two approaches, 11 PCs with a resolution of 1.2 was used to cluster the cells, and UMAP was used with 11 PCs to visualize the 28 clusters detected.

Garnett was used for initial cell classification as previously described.[83] In brief, a cell type marker file with 17 different cell types was compiled using various resources,[80,81,84] and this marker file was checked for specificity using the “check_markers” function in Garnett by checking the ambiguity score and the relative number of cells for each cell type. A classifier was then trained using the marker file, with “num_unknown” set to 150, and this classifier was then used to classify cells and cell type assignments were extended to nearby cells, “clustering-extended type” (Louvain clustering).

TooManyCells 2.0.0.0 was then used to cluster and visualize the 34,259 single cells, as previously described.[85] Briefly, the raw data from the 34,259 cells were not filtered and were normalized by total count and gene normalization by median count followed by frequency-inverse document frequency (tf-idf) using the flags --normalization "BothNorm and --no-filter. The “clustering-extended type” cell labels from Garnett, as well as the demarcation of canonical cell markers, were used to identify broad classes of cell types found within the pancreas, of which we focused on four: Beta, Stellate, Endothelial, and Immune cells.

Differential genes were found using edgeR 3.24.3 through TooManyCells with the normalization “NoneNorm” to invoke edgeR single cell preprocessing, including normalization and filtering. Briefly, edgeR fits normalized expression data to a negative binomial model and uses an exact test with false discovery rate (FDR) control to determine differential expressed genes.[86]

4.3.16. Single-cell RNA-seq of mouse aorta

All animal experiments were performed following protocols approved by the Institutional Animal Care and Use Committee at Baylor College of Medicine in accordance with the guidelines of the National Institutes of Health. The Center for Comparative Medicine at Baylor College of Medicine monitors the environmental conditions in the animal husbandry rooms. All mice housed in standard ventilated cages, floor area 65 in^2, maximum 4 mice per cage. Room temperatures are maintained at 70˚F ± 2˚. Normal humidity for animal holding rooms ranges from 30% to 70%. The standard light timer is set on a 14-hour light cycle with the lights coming on at 6 AM and off at 8 PM.

Ascending aortic samples were harvested from Mef2c-Cre ROSA26RmT/mG male mice (N=5) and were pooled in Hanks' Balanced Salt Solution (HBSS, #14175095, Thermo Fisher Scientific) with 10% fetal bovine serum. Extra aortic tissues were removed and the aortic tissues were cut into small pieces. To digest the aortas, samples were subsequently incubated with an enzyme cocktail (3 mg/ml collagenase type II (LS004176, Worthington); 0.15 mg/ml collagenase type XI (C7657, Sigma-Aldrich); 0.24 mg/ml hyaluronidase type I (H3506, Sigma-Aldrich); 0.1875 mg/ml elastase (LS002290, Worthington); 2.38 mg/ml HEPES (H4034, Sigma-Aldrich)) in Ca/Mg contained-HBSS (#14025092, Thermo Fisher Scientific) for 60 minutes at 37 °C. The cell suspension was filtered through a 40 μm cell strainer (CLS431750-50EA, Sigma-Aldrich), centrifuged at 300 g for 10 minutes, and resuspended using cold HBSS (#14175095) with 5% fetal bovine serum. Cells were stained with DIPI and were sorted to select viable cells (≥ 95% viability) by flow cytometry (FACS Aria III, BD Biosciences).

The cells were dispensed onto the Chromium Controller (10x Genomics) and indexed single cell libraries were constructed by a Chromium Single Cell 3' v2 Reagent Kit (10x Genomics). cDNA libraries were then sequenced in a pair-end fashion on an Illumina NovaSeq 6000. Raw FASTQ data was aligned by Cell Ranger 3.0 with GRCh38. Mapped unique molecular identifier (UMI) counts were imported into Seurat 3.1.4 and built into Seurat objects using the "CreateSeuratObject" function. Cells expressing less than 200 or more than 5000 genes were filtered out for exclusion of non-cell or cell aggregates. Cells with more than 10% mitochondrial genes were also excluded. Data was then normalized and processed into scaled data. Principal component analysis (PCA) and non-linear dimensional reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE) were performed to create clusters and those visualization. The

“FindAllMarkers” function in Seurat was used to identify the conserved marker genes in each cluster.

4.3.17. Single-cell RNA-seq of human aorta

The protocol for collecting human aortic tissue samples for scRNA-seq study was approved by the Institutional Review Board at Baylor College of Medicine. Written informed consent was provided by all participants before enrollment. All experiments conducted with human tissue samples were performed in accordance with the relevant guidelines and regulations. Ascending aortic samples were acquired from 3 controls (2 female and 1 male, heart transplant recipient or lung transplant donor) and 8 individuals with ascending thoracic aortic aneurysm (4 female and 4 male). Additional information can be found in the Life Sciences Reporting Summary of this study. For each sample, a piece of aortic tissue (1-2 cm^2) was torn into thin layers and cut into small pieces in Hanks’ balanced salt solution (HBSS, without Ca^{2+} and Mg^{2+}) (Gibco, Waltham, MA, USA) with 10% fetal bovine serum (FBS). Small pieces of tissue were then moved to enzyme cocktail prepared with 3 mg/ml collagenase type II (LS004176, Worthington Biochemical Corp., Lakewood, NJ, USA), 0.15 mg/ml collagenase type XI (H3506, Sigma Corp., Kanagawa, Japan), 0.25 mg/ml soybean trypsin inhibitor (LS003571, Worthington), 0.1875 mg/ml elastase lyophilized (LS002292, Worthington), 0.24 mg/ml hyaluronidase type I (H3506, Sigma), and 2.38 mg/ml 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, H4034, Sigma) in HBSS (with Ca^{2+} and Mg^{2+}) (14025092, Thermo Fisher Scientific, Waltham, MA, USA) and were digested in a 37˚C water bath for 1 to 2 hours. Tissue dissociation was examined under a microscope. Cell suspensions were collected by using a 40-μm cell strainer (CLS431750-50EA, Corning, Inc., Corning, NY, USA), centrifuged at 300 g for 10 minutes, and resuspended in HBSS (without Ca^{2+} and Mg^{2+}) (14175095, Thermo Fisher) with 5% FBS, followed with incubation on ice for 30 minutes. Cells were then stained by using a live and dead cell kit (L3224, Thermo Fisher) and were submitted for flow cytometry (BD) for the collection of live singlet cells. The living cell rate was further examined under a microscope by using trypan blue (T8154, Sigma Corp., Kanagawa, Japan) staining.

Single-cell suspensions were submitted to the 10X Genomics Chromium System (10x Genomics, Pleasanton, CA, USA), followed by the construction of 3' gene expression v3 libraries and sequencing on an Illumina NovaSeq 6000. Raw FASTQ data alignment was processed by using Cell Ranger 3.0, with GRCh38 as a reference. Mapped unique molecular identifier (UMI) counts were loaded to R for further analysis. The single-cell sequencing data were filtered by using Seurat 3.0 with the following criteria: gene count per cell >200 and <4000 (or 5000), percentage of mitochondrial genes <10%, and no HBB gene detected in the cell. Data were then normalized and processed into scale data, linear dimensional reduction, cluster finding, and nonlinear dimensional reduction for visualization according to the Seurat manual. To identify clusters in multiple combined datasets, we performed additional integration after normalization and before scale. The conserved (marker) genes for each cluster were identified by using the function "FindAllMarkers" in Seurat. For reclustering, the UMI count of cells of interest were extracted and analyzed similarly to clusters identified in multiple combined datasets.

4.4. Results and Discussion

4.4.1. Discovery: ExoPheWAS in PMBB

We interrogated a dataset of 10,900 individuals with WES data in PMBB (Table 4.1) for carriers of rare (MAF ≤ 0.1% in gnomAD) pLOF variants, which include frameshift insertions or deletions, gain or loss of stop codon, and disruption of canonical splice site dinucleotides. The distribution of the number of carriers for rare pLOF variants per gene was on a negative exponential distribution (Appendix C, Figure S1). We chose to interrogate genes with at least 25 heterozygous carriers for rare pLOFs (N=1,518 genes), for which we show that statistical power to detect association is sufficient as a function of effect size and the associated phenotype's number of cases (Appendix C, Figure S2). We collapsed rare pLOF variants into gene burdens across these 1518 genes for ExoPheWAS analyses with 1000 binary phecodes with at least 20 cases (Figure 4.1). Given that p values for gene burden association studies interrogating rare loss-of-function variants may be inflated due to their higher likelihood of increasing disease risk compared to other variants,[87] we found that our associations roughly deviated from the fitted expected distribution at an observed p<E-06 (Appendix C, Figure S3). We identified 97 gene burdens with phenotype associations at p<E-06 (Figure 4.2; Appendix C, Table S1). We addressed potential inflation issues regarding small sample sizes by using Firth's penalized likelihood approach, and found that beta and significance estimates were consistent with exact logistic regression (Appendix C, Table S1).

Basic demographics	
N	10900
Female, N (%)	4432 (40.7)
Median Age (at biobank entry), yr	67.0
Genetically informed ancestry	
AFR, N (%)	2172 (19.9)
AMR, N (%)	304 (2.8)
EAS, N (%)	79 (0.7)
EUR, N (%)	8198 (75.2)
SAS, N (%)	114 (1.0)
Cardiovascular phenotypes	
Essential hypertension, N (%)	6441 (59.1)
Ischemic Heart Disease, N (%)	5008 (45.9)
Myocardial infarction, N (%)	1640 (15.0)
Cardiomyopathy, N (%)	1976 (18.1)
Congestive heart failure; nonhypertensive, N (%)	3695 (33.9)
Heart transplant/surgery, N (%)	518 (4.8)
Cardiac dysrhythmias, N (%)	5784 (53.1)
Atrial fibrillation and flutter, N (%)	3782 (34.7)
Cerebrovascular disease, N (%)	1706 (15.7)
Peripheral vascular disease, N (%)	954 (8.8)
Aortic aneurysm, N (%)	836 (7.7)
Atherosclerosis, N (%)	539 (4.9)

Endocrine/metabolic phenotypes	
Type 2 diabetes, N (%)	2799 (25.7)
Overweight, obesity and other hyperalimentation, N (%)	2275 (20.9)
Hyperlipidemia, N (%)	6231 (57.2)
Hypercholesterolemia, N (%)	2034 (18.7)
Hypothyroidism, N (%)	1314 (12.1)
Gout and other crystal arthropathies, N (%)	811 (7.4)
Gastrointenstinal phenotypes	
Esophagitis, GERD and related diseases, N (%)	2526 (23.2)
Gastrointestinal hemorrhage, N (%)	660 (6.1)
Diverticulosis and diverticulitis, N (%)	610 (5.6)
Chronic liver disease and cirrhosis, N (%)	449 (4.1)
Renal phenotypes	
Chronic renal failure (CKD), N (%)	2135 (19.6)
End stage renal disease, N (%)	510 (4.7)
Kidney replaced by transplant, N (%)	283 (2.6)
Neuropsychiatric phenotypes	
Mood disorders, N (%)	1353 (12.4)
Anxiety, phobic and dissociative disorders, N (%)	1322 (12.1)
Delirium dementia and amnestic and other cognitive disorders, N (%)	123 (1.1)
Respiratory phenotypes	
Chronic airway obstruction, N (%)	1314 (12.1)
Asthma, N (%)	920 (8.4)
Obstructive sleep apnea, N (%)	1623 (14.9)
Respiratory failure, insufficiency, arrest, N (%)	697 (6.4)
Sensory phenotypes	
Cataract, N (%)	796 (7.3)
Hearing loss, N (%)	579 (5.3)
Glaucoma, N (%)	449 (4.1)
Congenital phenotypes	
Cardiac and circulatory congenital anomalies, N (%)	780 (7.2)
Genitourinary congenital anomalies, N (%)	151 (1.4)
Cystic kidney disease, N (%)	108 (1.0)
Congenital anomalies of great vessels, N (%)	77 (0.7)

Table 4.1. Demographics and disease prevalence of the PMBB discovery cohort. Demographic information and clinical phenotypic counts for all individuals with whole-exome sequencing linked to electronic health records in the Penn Medicine BioBank (PMBB). Clinical phenotypes were defined by phecodes (see Materials and Methods). Data is represented as count data with percent prevalence in the population in parentheses, where appropriate.

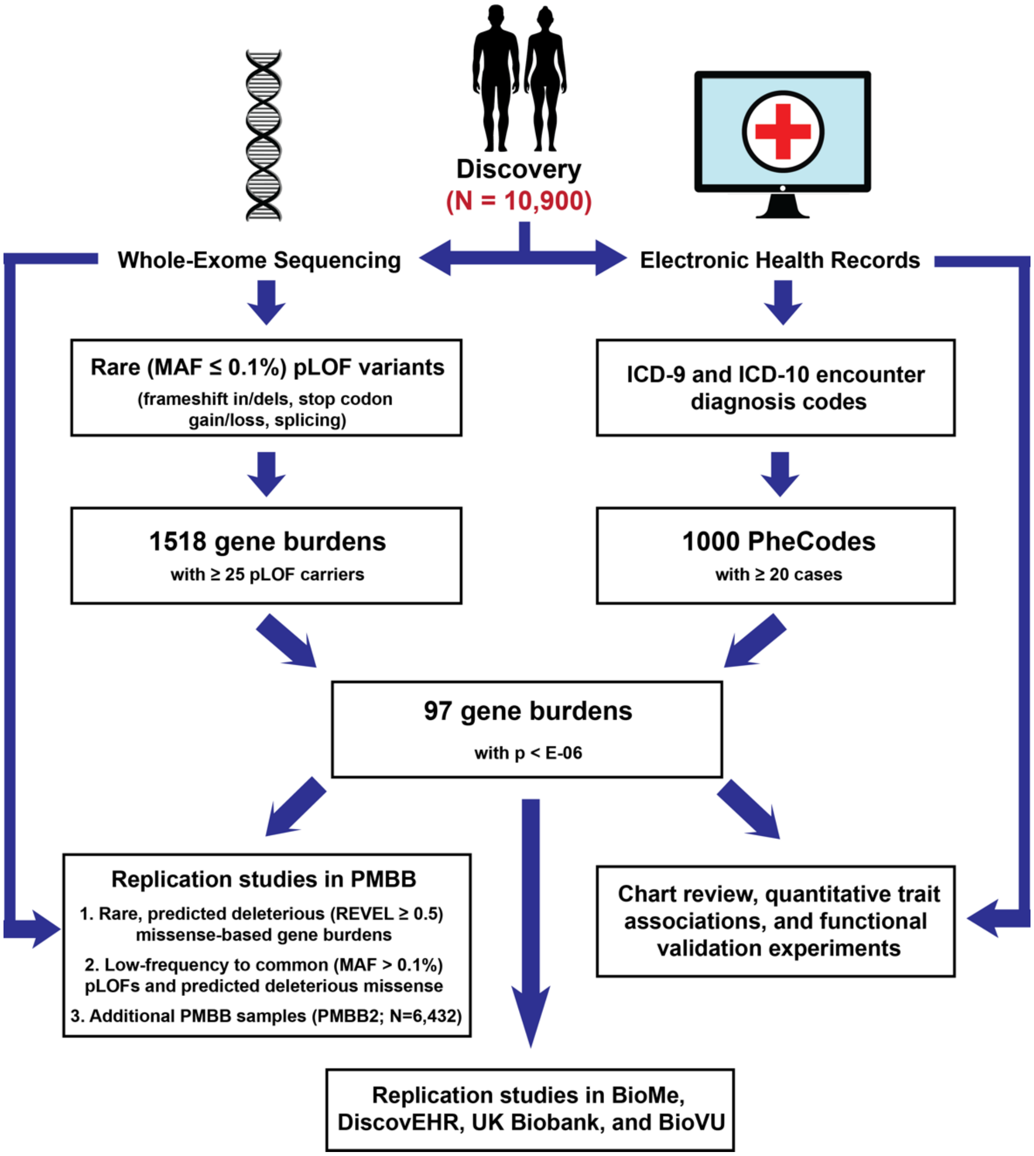
Discovery
(N = 10,900)
Whole-Exome Sequencing
Electronic Health Records
Rare (MAF ≤ 0.1%) pLOF variants
(frameshift in/dels, stop codon gain/loss, splicing)
ICD-9 and ICD-10 encounter diagnosis codes
1518 gene burdens
with ≥ 25 pLOF carriers
1000 PheCodes
with ≥ 20 cases
97 gene burdens
with p < E-06
Replication studies in PMBB
1. Rare, predicted deleterious (REVEL ≥ 0.5) missense-based gene burdens
2. Low-frequency to common (MAF > 0.1%) pLOFs and predicted deleterious missense
3. Additional PMBB samples (PMBB2; N=6,432)
Chart review, quantitative trait associations, and functional validation experiments
Replication studies in BioMe, DiscovEHR, UK Biobank, and BioVU

Figure 4.1. Flow chart for exome-by-phenome-wide association analysis using electronic health record phenotypes. Flow chart diagram outlining the primary methodologies used for conducting the exome-by-phenome-wide association study and for evaluation of the robustness of the associations, indicating that 97 genes had associations at a significance level of p<E-06 via logistic regression. The pathways starting with short descending arrows represent the *discovery phase*, in which predicted loss-of-function (pLOF)-based gene burdens were studied on an exome-by-phenome-wide scale in 10,900 individuals from the Penn Medicine BioBank (PMBB). “Replication studies in PMBB” refers to analyses of gene-phenotype associations using REVEL-informed missense-based gene burdens and univariate analyses within the discovery PMBB cohort, as well as in an independent cohort of African Americans in the PMBB (the PMBB2 cohort; N=6,432). Additional replication studies included analyses of gene-phenotype

associations using pLOF-based gene burdens, REVEL-informed missense-based gene burdens, and univariate analyses in BioMe (N=23,989), DiscovEHR (N=85,450), and the UK Biobank (N=32,268), as well as univariate analyses in BioVU (N=66,400).

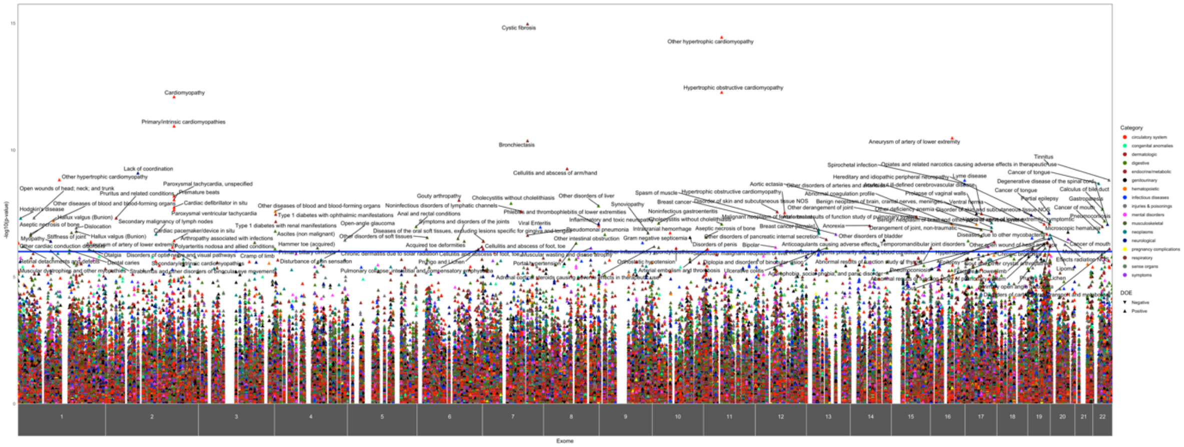

Figure 4.2. ExoPheWAS plot exhibits the landscape of gene-phenotype associations across the exome and phenome in the Penn Medicine BioBank. Plot of the results of the exome-by-phenome-wide association study (ExoPheWAS) in the Penn Medicine BioBank for 1518 gene burdens of rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants. The x-axis represents the exome and is organized by chromosomal location. The location of each gene along the x-axis is according to the gene's genomic location per Genome Reference Consortium Human Build 37 (GRCh37). The association of each gene burden with a set of 1,000 phecodes is plotted vertically above each gene, with the height of each point representing the $–\log_{10}$(p value) of the association between the gene burden and phecode using a logistic regression model. Each phecode point is color-coded according to the phecode group, and the directionality of each triangular point represents the direction of effect (DOE). The blue line represents the significance threshold at p=E-06 to account for multiple hypothesis testing.

4.4.2. Replication in PMBB, other medical biobanks, and UK Biobank

We evaluated the robustness of the significant gene-phenotype associations identified via our pLOF-based ExoPheWAS analyses by testing the associations in the same PMBB cohort between a separate group of rare *likely deleterious* exonic missense variants in the 97 significant genes with the same disease phenotypes that were identified in discovery (Figure 4.1). We utilized REVEL, an ensemble method for predicting the pathogenicity of missense variants,[20] to define predicted deleterious missense (pDM) variants (REVEL score ≥ 0.5) given the tool's success in identifying likely pathogenic variants for gene burden association studies.[10] First, we separately collapsed rare (MAF ≤ 0.1%), REVEL-informed pDM variants to test discovery-driven associations with their corresponding phenotypes (Appendix C, Table S2). We also interrogated single variants, including both pLOF variants and pDM (REVEL ≥ 0.5) variants, in the 97 genes identified in discovery that were of sufficient frequency (MAF > 0.1%) and therefore were not included in either of the gene burden analyses (Appendix C, Table S3).

We also endeavored to replicate our significant ExoPheWAS discovery analysis associations (Figure 4.1) using a separate cohort of 6,432 African Americans in PMBB who were exome-sequenced (PMBB2; Appendix C, Table S4-6), as well as two additional medical biobanks with WES linked to EHR phenotypes, namely BioMe (Mount Sinai; Appendix C, Table S7-9) and DiscovEHR (Geisinger Health System; Appendix C, Table S10-12), and the population-based UK Biobank (UKB) (Appendix C, Table S13-15). For each of the 97 significant genes, we interrogated: 1) gene burdens after collapsing rare (MAF ≤ 0.1%) pLOF variants, 2) gene burdens after collapsing non-overlapping rare (MAF ≤ 0.1%) REVEL-informed pDM variants, and 3) single pLOF or REVEL-informed pDM variants with MAF > 0.1% for association with their discovery phenotypes. Finally, we further interrogated a targeted list of univariate replications in BioVU (Vanderbilt; Appendix C, Table S16).

4.4.3. Positive control gene-phenotype associations

We identified a total of 26 robust genes using a Diverse Convergent Evidence (DiCE) approach[88] for ranking associations using a combination of the number of significant replications and functional validation (Table 4.2; Appendix C, Table S17). Five of these genes can be considered positive control gene-disease associations. A gene burden of rare pLOFs in *CFTR* was significantly associated with cystic fibrosis (CF), a recessive condition caused by biallelic variants in *CFTR*. This was driven by individuals with a rare pLOF who had a second deleterious *CFTR* variant—predominantly ΔF508—that was not included in the pLOF gene burden. This association of the *CFTR* pLOF gene burden with CF was not replicated in other biobanks due to the extremely low case prevalence of CF (Appendix C, Table S18). The *CFTR* pLOF gene burden was also significantly associated with bronchiectasis independent of a CF diagnosis and occurred in individuals without a second *CFTR* variant; this finding replicated in all interrogated cohorts. While a predisposition to bronchiectasis due to haploinsufficiency of *CFTR* has been suggested,[89] our finding strengthens this observation. *TTN* is a known dilated cardiomyopathy gene that replicated convincingly across other cohorts. *MYBPC3* is a known hypertrophic cardiomyopathy (HCM) gene that replicated in BioMe and DiscovEHR, but not in UKB, where HCM had a case-control ratio of an order of magnitude lower than the medical biobanks (Appendix C, Table S18). These results indicate that medical biobanks have a different—and sicker—population that enables discovery of associations of human diseases driven by rare genetic variants. A pLOF gene burden in *BRCA2* was associated with breast cancer and replicated in all biobanks. *BRCA1* was associated with breast cancer in discovery (p=1.29E-04) but due to power did not meet our significance threshold. Finally, *CYP2D6* is a P450 enzyme known to metabolize opioids;[90] we found that *CYP2D6* was significantly associated with adverse effects of therapeutic opiate use.

Gene	Phecode Description	Discovery P	Replications (N)	Clinical and/or Experimental Evidence
BRCA2	Breast cancer	1.72E-07	4	✓
CFTR	Bronchiectasis	2.27E-07	10	✓
	Pseudomonal pneumonia	4.21E-11	5	✓
	Cystic fibrosis	1.05E-15	1	✓
CYP2D6	Opiates and related narcotics causing adverse effects in therapeutic use	1.50E-09	3	✓
MYBPC3	Hypertrophic cardiomyopathy	3.49E-15	5	✓
TTN	Cardiomyopathy	7.83E-13	10	✓
	Cardiac conduction disorders	6.45E-09	10	✓
	Cardiac dysrhythmias	1.77E-08	12	✓
ABCA10	Benign neoplasm of brain, cranial nerves, meninges	7.26E-08	2	
	Abnormal results of function study of pulmonary system	1.54E-07	3	
BBS10	Hypertrophic cardiomyopathy	2.89E-08	1	✓
CES5A	Abnormal coagulation profile	8.10E-08	5	
CILP	Aortic ectasia	4.29E-08	3	✓
CTC1	Temporomandibular joint disorders	3.76E-07	3	
DNAH6	Lack of coordination	7.93E-10	2	
DNHD1	Aseptic necrosis of bone	2.67E-07	4	
EFCAB5	Prolapse of vaginal walls	3.19E-08	3	
EPPK1	Phlebitis and thrombophlebitis of lower extremities	9.19E-08	3	
FER1L6	Muscular wasting and disuse atrophy	7.18E-07	3	✓
FLG2	Stiffness of joint	1.76E-07	2	
MYCBP2	Spasm of muscle	2.08E-07	2	✓
PPP1R13L	Primary open angle glaucoma	7.29E-07	2	✓
RGS12	Type 1 diabetes	6.48E-08	5	✓
RTKN2	Orthostatic hypotension	7.24E-07	5	
SCNN1D	Cardiac conduction disorders	4.52E-07	5	
TGM6	Lipoma	2.77E-07	4	
TRDN	Acquired toe deformities	3.90E-07	3	
WDR87	Ventral hernia	1.70E-07	4	
ZNF175	Tinnitus	3.24E-10	3	✓
ZNF334	Microscopic hematuria	1.69E-07	3	

Table 4.2. List of robust exome-by-phenome-wide significant gene-phenotype associations. List of genes among 97 pLOF-based gene burdens with phenotype associations at p<E-06 in the Penn Medicine BioBank (PMBB) discovery cohort that were most robust according to a Diverse Convergent Evidence (DiCE) approach, which integrates successful replication of the association with clinical and experimental evidence. For replication studies, gene-phenotype associations were evaluated for their robustness by interrogating REVEL-informed missense-based gene burdens and single variants in the same discovery PMBB cohort, and pLOF-based gene burdens, REVEL-informed missense-based gene burdens, and single variants in an independent cohort of African Americans in PMBB (the PMBB2 cohort) as well as in BioMe, DiscovEHR, and the UK Biobank (UKB). Targeted single variants that showed successful replication in PMBB, PMBB2, and UKB were additionally analyzed in BioVU. Each gene-phecode association is labeled with the corresponding p value from logistic regression analyses in the discovery phase in PMBB as well as the number of total replications and existence of clinical/experimental evidence, as fully detailed in Appendix C, Table S17. Only associations with at least two total checkmarks in Appendix C, Table S17, where each successful mode of replication in a particular biobank (*e.g.* pLOF burden in BioMe) or the existence of clinical/experimental evidence is labeled with a checkmark, were deemed robust and were included in this table. Previously known associations were considered to represent positive controls and are listed at the top of the table, and are separated from novel associations by a double line. Positive control and novel associations are each ranked alphabetically by gene name.

4.4.4. Novel gene-phenotype associations

We identified 20 robust genes with novel disease associations that had at least two additional replications beyond the discovery experiment, and one strongly supported by the DiCE analysis (Table 4.2; Appendix C, Tables S2-S17). Some have prior biological plausibility, and for others we generated additional functional data supporting a biological basis to these associations. For example, a *BBS10* gene burden was significantly associated with HCM. *BBS10* is one of at least 19 genes implicated in autosomal recessive Bardet-Biedl Syndrome and accounts for ~20% of all cases.[91] *BBS10* is expressed in the heart[66] and cardiac abnormalities have been reported in Bardet-Biedl Syndrome, including hypertrophy of the interventricular septum,[92] but cardiac abnormalities due to haploinsufficiency of *BBS10* have not been described. We interrogated echocardiography data in carriers of rare pLOF variants in *BBS10* in PMBB compared with non-carriers and found increased left ventricular outflow tract (LVOT) stroke volume, consistent with cardiac hypertrophy (Appendix C, Table S19). Rare pLOF variants in *SCNN1D*, which encodes the delta subunit of the epithelial sodium channel (δENaC), were associated with cardiac conduction disorders and replicated robustly across medical biobanks. *SCNN1D* is expressed in the heart (unlike epithelial tissue-specific expression for *SCNN1A* and *SCNN1B*),[93] there is an association between 1p36 deletions (which contains *SCNN1D*) and congenital heart defects,[94] and decreased expression of δENaC may contribute to disrupted Na+ and K+ homeostasis in ischemic heart diseases.[95] The association between rare pLOFs in *ZNF175* and tinnitus (additionally, hearing loss barely missed our significance threshold), which replicated in BioMe, DiscovEHR, and UKB, is supported by the finding that mice with loss-of-function in *Zfp719* (the mouse ortholog) are profoundly deaf and have abnormal Preyer reflex (auditory startle response)[96] as well as raised auditory brainstem response thresholds.[97] *Zfp719* is expressed in inner and outer hair cells of the mouse ear,[98] and human *ZNF175* has a suggested role in neurotrophin production and neuronal survival.[99]

Rare pLOFs in *FER1L6* were robustly associated with muscular wasting and disuse atrophy. *FER1L6* is a member of the ferlin family of genes, and mutations in *FER1L1* (dysferlin) are known to cause recessive forms of muscular dystrophy.[100] Importantly, loss of the zebrafish ortholog *Fer1l6* has been shown to lead to deformation of striated muscle and delayed cardiac development.[101] Similarly, pLOFs in *MYCBP2*, an E3 ubiquitin-protein ligase critical in neuromuscular development in mice,[102] *Drosophila*,[103] and *C. elegans*,[104] were associated with muscular spasms and dystrophy. Mice lacking the mouse ortholog *Phr1* are lethal at birth without taking a breath due to incomplete innervation of the diaphragm by markedly narrower phrenic nerves that contain fewer axons than controls.[102] We found that *MYCBP2* showed significantly decreased expression in various lower extremity muscle tissues in tibial muscular dystrophy in humans (Appendix C, Figure S4). Our findings suggest that haploinsufficiency in *FER1L6* or *MYCBP2* increases the risk of developing dystrophic skeletal muscle.

Rare pLOFs in *CES5A* were robustly associated with abnormal coagulation. Upon further investigation of EHR lab data in PMBB, we found that carriers of rare pLOF variants in *CES5A* had increased international normalized ratios (INR; ß=8.2, p=2.13E-02, N=5,275) and partial thromboplastin times (PTT; ß=13.9, p=2.07E-02, N=3,786) compared to non-carriers. Through chart review, we found an enrichment of gastrointestinal bleeding episodes following use of anti-platelet medications among carriers for rare pLOF variants in *CES5A*. *CES5A* is part of the family of carboxylesterases, which are known metabolizers of various orally bioavailable drugs, including the anti-platelet medications aspirin and clopidogrel.[105] Given its predominant expression in the liver,[66] it is thus plausible that haploinsufficiency of *CES5A* predisposes to adverse effects of anti-platelet medications.

Another novel finding was that rare pLOF variants in *PPP1R13L*, one of the most evolutionarily conserved inhibitors of p53,[106] were associated with primary open angle glaucoma—a disease of the optic nerve head (ONH) that causes progressive vision loss. We interrogated the expression of *PPP1R13L in silico* using the Ocular Tissue Database (OTDB) and found that it is highly expressed in ocular tissues, with optic nerve and the ONH among the highest (Appendix C, Table S20). Retinal ganglion cells (RGCs) are the primary cells affected by glaucoma, and cells in the ONH such as astroglia, microglia, and endothelial cells mediate RGC degeneration in response to stress such as increased intraocular pressure. We investigated whether *Ppp1r13l* is differentially expressed in the mouse ONH in glaucoma by comparing microarray gene expression datasets of the ONH.[107] We found *Ppp1r13l* expression to be highest during late-early to moderate stages of glaucoma (Appendix C, Figure S5A). Additionally, inhibition of *PPP1R13L* has been shown to exacerbate retinal ganglion cell (RGC) death following axonal injury.[108] We found that the PPP1R13L protein is predominantly localized to the ganglion cell layer in the adult human retina with some expression in the outer and inner plexiform layers, confirming its role in RGC function (Appendix C, Figure S5B). Using human induced pluripotent stem cell-derived RGCs (iPSC-RGCs), we found that oxidative stress markedly upregulated *PPP1R13L* expression (Appendix C, Figure S5C) to a much greater extent than even superoxide dismutase 1 (*SOD1*), which is known to be transcriptionally upregulated in response to oxidative stress. Thus, *PPP1R13L* is expressed in RGCs, is significantly upregulated by oxidative stress, and may help to prevent RGC death from p53 activation and p53-mediated apoptosis in primary open angle glaucoma.[109] Our results are consistent with the concept that haploinsufficiency of *PPP1R13L* in RGCs increases the visual consequences of primary open angle glaucoma.

Another interesting novel finding was that rare pLOF variants in *RGS12* were associated with type 1 diabetes mellitus and its complications. In PMBB, carriers of rare pLOFs in *RGS12* had higher median values for random serum glucose than non-carriers (ß=16.9, p=2.91E-02, N=5,389). *RGS12*, an inhibitor of signal transduction in G protein signaling, contains an N terminus PDZ domain which selectively binds to and represses the macrophage IL-8 receptor CXCR2.[110] Activation of macrophage CXCR2 by IL-8 is pro-inflammatory, and its antagonism leads to attenuation of immune cell infiltration and cytokine release as well as a shift of macrophages to the anti-inflammatory M2 state, thereby counteracting inflammatory signal pathways in diabetes.[111] To further investigate *RGS12* in type 1 diabetes, we generated single-cell RNA-seq data in human pancreatic islets from type 1 diabetes and control subjects collected by the Human Pancreas Analysis Program (HPAP; https://hpap.pmacs.upenn.edu) and interrogated *RGS12* expression in distinct functional cells. We found that while *RGS12* showed no significant differential expression in pancreatic endocrine or exocrine cells in type 1 diabetes versus control, there was a substantial reduction of expression of *RGS12* in peri-islet CD45+ macrophages in type 1 diabetes (Appendix C, Figure S6). These results are consistent with a model that RGS12 dampens islet macrophage inflammatory responses and that haploinsufficiency of *RGS12* predisposes to greater islet inflammation and higher risk of type 1 diabetes.

Additionally, rare pLOF variants in *CILP* were associated with aortic ectasia, or dilatation of the aorta often associated with connective tissue disorders. Chart review of *CILP* pLOF carriers showed an enrichment for ascending thoracic aortic aneurysms. *CILP* encodes an extracellular matrix protein and is best known for its expression in chondrocytes.[112] However, *CILP* is also expressed in the cardiovascular system,[66] and has been shown to be involved in cardiac remodeling in response to pressure overload.[113] We performed single-cell RNA-seq of normal mouse aorta and found that *Cilp* expression was localized mainly to adventitial fibroblasts in the aorta, but showed no significant expression in aortic smooth muscle cells (Appendix C, Figure S7A-B). Single-cell RNA-seq of human aorta confirmed that *CILP* is localized to aortic fibroblasts (Appendix C, Figure S7C-D). Importantly, *CILP* has been reported to modulate *TGFB1* signaling and *IGF1*-induced proliferation,[114] and dysregulated TGF-ß signaling has been shown to contribute to the pathogenesis of thoracic aortic aneurysm formation.[115] To further interrogate the relationship between *CILP* and *TGFB1* in human fibroblasts, we conducted a meta-analysis of 11 independent microarray and RNA-seq datasets for human fibroblasts from various tissues treated with TGF-ß from the Gene Expression Omnibus (GEO). We found that *CILP* was in the top 1% of significantly upregulated genes in human fibroblasts when treated with TGF-ß ($\log_2$ fold change = 1.964, p = 3.60E-29; Appendix C, Figure S7E), confirming its role in a functional feedback loop with TGF-ß as similarly seen in the context of chondrocyte metabolism.[112] Furthermore, *CILP* was differentially co-expressed with *IGF1* as well as genes implicated in aortic ectasia including *SMAD3*, *ACTA2*, *MYH11,* and *ELN* (Appendix C, Figure S7E).[115] Our findings suggest that haploinsufficiency of *CILP* predisposes to the risk of developing thoracic aortic dilatation, perhaps through compromising the structural integrity of the aortic wall and contributing to dysregulation of TGF-ß signaling.

4.4.5. Conclusions

There has been a significant gap of knowledge regarding the clinical implications of genetic variants overrepresented among Africans due to the lack of ancestral diversity in the populations that have been studied in previous genetic association studies.[116] Our discovery study included 19.9% African ancestry individuals, and three of our replication cohorts included substantial numbers of African-Americans (6,432 in PMBB2, 6,470 in BioMe, and 10,456 in

BioVU). Interestingly, we identified 16 rare predicted deleterious single variants which are African ancestry-specific and that replicated associations with the same disease in which a pLOF gene burden was associated in discovery (Appendix C, Table S21). None of these rare variants exist in the GWAS catalog or have been previously mentioned in the published literature. Our findings suggest that larger experiments of this type in ethnically diverse cohorts are imperative for improving our understanding of the contribution of ancestry-specific rare genetic variants to human disease.

A significant challenge in rare variant association studies is the difficulty of performing replication studies. Here we show the value of evaluating the robustness of gene burden associations by interrogating other deleterious variants in the same genes (but in different individuals) in the same biobank cohort. We also performed replication studies in another cohort in PMBB as well as in two other medical biobanks with WES data. These provided more replication than the UKB, which is a population-based biobank and is widely recognized to have a "healthy volunteer selection bias"[117] and has lower prevalence of the specific diseases than the medical biobanks (Appendix C, Table S18). This may be one factor explaining the relative lack of novel findings in gene burden studies using UKB for discovery.[21,118] Finally, we show that one should not expect a uniform fit for p values when interrogating the cumulative effect of rare pLOF variants, and that the validity of the results is due as much to robust replication in other cohorts as to the determination of a particular significance threshold. To this end, our study emphasizes the value of medical biobanks for discovery of novel gene-disease associations based on rare variants.

In conclusion, we demonstrate the feasibility and value of aggregating rare pLOF variants into gene burdens on an exome-wide scale for association with EHR-derived phenotypes in a medical biobank for discovery of novel gene-disease relationships. Our compelling novel findings based on initial discovery in < 11,000 whole exomes suggest that much larger experiments of this type are likely to be highly informative and will lead to many new insights into the biology of human phenotypes and diseases.

CHAPTER 5. Exome-wide association of rare coding variants with hepatic fat derived from CT imaging in a medical biobank

5.1. Abstract

Background: Non-alcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease in Western countries and can lead to metabolic dysfunction, liver inflammation, end-stage liver disease. While hepatic fat is a highly heritable trait, previous genetic studies of hepatic fat quantifications derived from abdominal imaging scans have traditionally focused on common variants and have been largely underpowered for interrogation of additional rarer genetic variation.

Methods: We linked whole-exome sequences (WES) to hepatic fat quantifications derived from clinical CT scans using machine learning in a subset of 10,283 individuals in the Penn Medicine BioBank (PMBB). We conducted exome-wide discovery analyses for single variants (MAF>0.1%) as well as for gene burdens of rare predicted loss-of-function (pLOF) variants (MAF≤0.1%) in PMBB. We also tested a burden of rare pLOF ± predicted deleterious missense (pDM) variants (REVEL≥0.5) for genes nominated by the single variant discovery. We performed replication in the UK Biobank (UKB) by linking WES to MRI-based liver proton density fat fractions (PDFF).

Results: Exome-wide significant single variants confirmed previously described associations (*e.g.* variants in *PNPLA3*, *TM6SF2*, *SAMM50*) and revealed new variants in *FGD5* and *CITED2* associated with hepatic fat. We also found that a gene burden of rare predicted deleterious coding variants in certain genes nominated by single variants (*e.g. PNPLA3*) were associated with hepatic fat. Additionally, a burden of rare pLOFs in *LMF2* were exome-wide significant in their association with increased hepatic fat, a finding that replicated in UKB.

Conclusion: We show the value and feasibility of conducting an exome-wide evaluation of common and rare variants for association with hepatic fat as quantitated from abdominal CT

scans via machine learning. Importantly, this work represents the first exome-wide association studies for hepatic fat quantifications in a medical biobank. In addition to confirming signals for single variants previously associated with hepatic fat, we also found new common and rare variants associated with hepatic fat which replicated in UKB. We suggest that this approach applied to larger ancestrally diverse populations will uncover new genetic modulators of intrahepatic fat.

5.2. Introduction

Hepatic steatosis, or excess accumulation of intrahepatic fat, is a major risk factor for metabolic dysfunction, liver inflammation, and end-stage liver disease accompanied by high morbidity and mortality.[119] In particular, non-alcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease in Western countries, and there is growing evidence that the clinical burden of NAFLD extends beyond liver-related morbidity and mortality such as increasing risk for type 2 diabetes mellitus, cardiovascular disease, and chronic kidney disease.[120] While NAFLD has high heritability and our understanding of the genetic underpinnings of NAFLD has advanced, known genetic risk variants still explain only a small fraction of heritability, suggesting the existence of additional genetic variation that may confer risk for or protection from NAFLD which have yet to be uncovered.[121]

Large-scale systematic quantification of hepatic fat derived from clinical imaging to measure the extent of hepatic steatosis has proven to be important for discovery of genetic variants that contribute to risk of developing NAFLD.[122-124] However, genetic studies in this realm have traditionally focused on analysis of common variants and have been largely underpowered for interrogation of additional rare genetic variation. The Penn Medicine BioBank (PMBB) is a large academic medical biobank enriched for disease with genetic sequencing linked to electronic health record (EHR) phenotypes, in which a substantial number of participants have received abdominal computed tomography (CT) scans in the course of routine clinical care. We automated hepatic fat quantification derived from abdominal CT scans using a machine learning approach. In this study, we present an exome-wide analysis of the association of rare coding variants with hepatic fat in a subset of ~10K individuals in the PMBB for whom CT-derived hepatic fat quantitation and whole exome sequence data were available.

5.3. Materials and Methods

5.3.1. 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. These analyses focused on the subset of PMBB participants (N=10,283) who had both CT-derived hepatic fat quantitation and whole exome sequence data available (Table 5.1). This 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.

5.3.2. Clinical data collection

For PMBB, all International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes, clinical imaging and laboratory measurements were extracted from the patients' EHR. All ICD diagnosis codes and outpatient laboratory measurements available up to July 2020 were extracted for PMBB participants. Non-contract abdominal CT images available up to March 2019 were extracted for PMBB participants if available (N=14,249). All laboratory values measured in the outpatient setting were extracted for participants from the time of enrollment in PMBB until July 2020; all units were converted to their respective clinical

Traditional Units. Minimum, median, and maximum measurements of each laboratory measurement were recorded per individual for association analyses. Minimum, median, and maximum values for hemoglobin A1C, alkaline phosphatase, ALT, AST, and triglycerides were log-transformed to normalize their distributions.

5.3.3. Quantification of hepatic fat

In PMBB participants, hepatic fat was quantitated from abdominal CT scans using a fully automated image curation and organ labeling technique using deep learning as previously described.[125] Hepatic fat was quantitated as the difference in mean Hounsfield Units (HU) between the spleen and liver (spleen HU – liver HU) to create a measure that is directly proportional to intrahepatic fat. Minimum, median, and maximum measurements of hepatic fat were recorded per individual given the multiple independent CT scans available per patient.

5.3.4. Phenome-wide association of hepatic fat with EHR diagnoses and traits

A phenome-wide association study (PheWAS) approach was used to determine the phenotypes associated with the quantitative trait of median hepatic fat in PMBB for the 10,283 unrelated individuals in PMBB with both exome sequences and quantitated hepatic fat available.[33] 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".[34] Patients were determined to have a certain disease phenotype if they had the corresponding ICD diagnosis on 2 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 Phecode was tested for association with quantitated hepatic fat using a logistic regression model adjusted for age, sex, and PC1-10 of genetic ancestry. PheWAS analyses were performed as trans-ancestral cosmopolitan analyses. Our association analyses considered only disease phenotypes with at least 20 cases based on a prior simulation study for power analysis.[22] This led to the interrogation of 1396 total Phecodes, and we used a Bonferroni correction to adjust for multiple testing (p=0.05/1396=3.58E-05).

5.3.5. Whole-exome sequencing, variant annotation, and selection for association testing

This study included a subset of 43,731 individuals in the PMBB who had undergone whole-exome sequencing (WES). We extracted DNA from stored buffy coats and then mapped exome sequences as generated by the Regeneron Genetics Center (Tarrytown, NY) to GRCh38 as previously described. Samples with low exome sequencing coverage, high missingness (*i.e.* greater than 5% of targeted bases), dissimilar reported and genetically determined sex, and genetic evidence of sample duplication were not included in this subset. For subsequent phenotypic association analyses, we removed samples with evidence of 1st and 2nd-degree relatedness, leading to a total of sample size of 41,759 for analysis.

Genetic variants were annotated using ANNOVAR (version 2019Oct24)[32] for information regarding variant effect as determined by the NCBI Reference Sequence (RefSeq) database,[126] Rare Exonic Variant Ensemble Learner (REVEL) scores for missense variants,[20] and allele frequencies reported by the Genome Aggregation (gnomAD) v2.[69] Predicted loss-of-function (pLOF) variants were defined as frameshift insertions or deletions, gain of stop codon, and disruption of canonical splice site dinucleotides. For splicing variants, we removed those with SpliceAI scores < 0.2 for loss or gain of acceptor or donor site.[53] For single variant association tests in the PMBB discovery, all nonsynonymous coding variants and splicing variants with minor allele frequency (MAF) > 0.1% in Africans or non-Finnish Europeans in gnomAD were selected for association testing. For gene burden association tests, rare (MAF ≤ 0.1% in gnomAD) pLOF variants were aggregated per gene with or without rare missense variants with REVEL score ≥ 0.5.

5.3.6. Exome-wide association studies for hepatic fat

This study focused on a subset of 10,283 unrelated individuals in PMBB with both exome sequences and quantitated hepatic fat available. For exome-wide association studies with hepatic fat, individuals with ICD9/10 diagnosis codes indicating chronic hepatitis B or C (B18.0-B18.2, 070.32, 070.21, 070.22, 070.23, 070.31, 070.33, 070.54) or alcohol-related conditions or dependence, such as alcoholic liver disease (571.0, K70.0), alcoholic hepatitis (571.1, K70.1), alcoholic fibrosis and sclerosis of the liver (571.2, K70.3), alcoholic cirrhosis of liver and/or ascites (571.2, K70.2), alcoholic hepatic failure, coma, and unspecified alcoholic liver disease (571.3, K70.4, K70.40, K70.41, K70.9), and alcohol dependence (303.0, 303.9, F10.229, F10.20), were excluded (N=689), leading to total sample size of 9,594 for analyses.

Exome-wide association studies for hepatic fat were conducted in two stages, namely single variant discovery and gene burden discovery. For the discovery analyses, single variants and gene burdens with at least 10 total carriers with hepatic fat quantifications available were associated with hepatic fat using a linear regression model adjusted for age, sex, and principal components (PC) of ancestry (PC1-5 in Africans, PC1-10 in Europeans). For targeted gene burden analyses of genes nominated by the single variant discovery, gene burdens with at least 5 total carriers with hepatic fat quantifications available were associated with hepatic fat. For all gene burdens, we used an additive genetic model to aggregate variants as previously described.[22] These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. Additionally, trans-ancestral cosmopolitan analyses were also performed, adjusted for age, sex, and cosmopolitan PC1-10.

We conducted a PheWAS for the gene burden of pLOF variants in *LMF2*, where we focused on a subset of 162 Phecodes in the "digestive" group, leading to a Bonferroni-corrected significance threshold of p=0.05/162=3.09E-04. The gene burden PheWAS analysis was performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis.

5.3.7. Replication analyses in the UK Biobank (UKB)

Replication analyses were conducted in the UK Biobank (UKB). For replication studies in the UKB, we focused on 9,071 individuals with both exome sequences (after removing samples with evidence of 1st and 2nd-degree relatedness, high missingness, and dissimilar reported and genetically determined sex) and quantitated hepatic fat available based on liver proton density fat fraction (PDFF) derived from MRI. Individuals with ICD10 diagnosis codes indicating chronic hepatitis B or C or alcohol-related conditions or dependence were excluded (N=22), leading to a total sample size of 9,049 for analyses. Single variants and gene burdens with at least 5 total carriers with hepatic fat quantifications available selected based on discovery in PMBB were associated with hepatic fat using a linear regression model adjusted for age, sex, and PC1-10 of ancestry. Similarly, for targeted gene burden analyses of genes nominated by the single variant discovery in PMBB, gene burdens with at least 5 total carriers with hepatic fat quantifications available were associated with hepatic fat in UKB. These analyses were performed in individuals of European ancestry, accompanied by trans-ancestral cosmopolitan analyses. Liver PDFF values from UKB were log-transformed to normalize their distribution for regression analyses. For replication studies in the UKB, International Classification of Diseases Tenth Revision (ICD-10) diagnosis codes and liver PDFF values derived from abdominal MRI scans were downloaded. Access to the UKB data for this project was from application 32133.

5.3.8. Statistical analyses

To associate hepatic fat phenotypes or genotypes with serum laboratory measurements in PMBB, we used a linear regression model adjusted for age, sex, and PCs of genetic ancestry (PC1-5 in Africans, PC1-10 in Europeans). These analyses were performed across all ancestries (cosmopolitan, PC1-10) and/or separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. All statistical analyses, including PheWAS and hepatic fat associations, were completed using R version 3.5 (Vienna, Austria). Minimum, median, and maximum values for hemoglobin A1C, alkaline phosphatase, ALT, AST, and triglycerides were log-normalized for regression analyses.

5.3.9. Analysis of publicly available expression datasets

We interrogated microarray and RNA-seq data publicly available on the NCBI GEO platform (https://www.ncbi.nlm.nih.gov/geo/).[71] We compared *Lmf2* expression in liver endothelial cells from cirrhotic livers of rats versus control (GSE1843).[127] We also compared *Lmf2* expression in a hepatic steatosis mouse model due to hepatocyte-specific *Sirt1* deficiency versus control (GSE14921).[128] Differential expression for each dataset was interrogated using the GEO2R software using a moderated *t* statistic after log transformation was applied to the data.

5.4. Results

5.4.1. Hepatic fat extracted from clinical CT scans is highly significantly associated with a range of cardiometabolic diseases and traits

Among exome-sequenced individuals in PMBB (Table 5.1), we conducted a phenome-wide association study (PheWAS) of the quantitative trait of median hepatic fat to interrogate the clinical diagnosis phenotypes associated with hepatic fat (Appendix D, Figure S1). Hepatic fat quantity was associated with increased risk for “Chronic liver disease and cirrhosis” (p=1.70E-45) and “Other chronic nonalcoholic liver disease” (Phecode representing NAFLD; p=8.89E-42) at phenome-wide significance. Hepatic fat also showed phenome-wide significant associations with increased risk for cardiometabolic comorbidities such as “Type 2 diabetes” (p=9.42E-30), “Obesity” (p=3.31E-18), and “Hypertension” (p=1.64E-13). Additionally, “Viral hepatitis” (p=1.17E-05) and “Alcoholic liver damage” (p=3.50E-15) were associated with increased hepatic fat at phenome-wide significance. Hepatic fat was also highly significantly associated with the quantitative trait of BMI (p=1.94E-42; Appendix D, Table S1), consistent with the known relationship between obesity and hepatic steatosis.

We also analyzed the association of hepatic fat with several clinical laboratory quantitative traits (Appendix D, Table S1). We found that hepatic fat values were significantly positively associated with serum ALT, AST, alkaline phosphatase, hemoglobin A1C, random glucose, and random triglycerides, and significantly inversely associated with HDL cholesterol, LDL cholesterol, and total cholesterol.

Basic demographics	
Total population, N	10283
Female, N (%)	4551 (44.3)
Median age, years	69
Genetically informed ancestry	
AFR, N (%)	2814 (27.4)
AMR, N (%)	110 (1.1)
EAS, N (%)	103 (1.0)
EUR, N (%)	7096 (69.0)
SAS, N (%)	84 (0.8)
Phecodes	
Chronic liver disease and cirrhosis, N (%)	1000 (9.7)
Other chronic nonalcoholic liver disease, N (%)	872 (8.5)
Type 2 Diabetes, N (%)	3018 (29.3)
Obesity, N (%)	2998 (29.2)
Essential hypertension, N (%)	6089 (59.2)
Alcoholic liver damage, N (%)	129 (1.3)
Viral hepatitis, N (%)	611 (5.9)
Portal hypertension, N (%)	135 (1.3)
Liver replaced by transplant, N (%)	260 (2.5)

Table 5.1. PMBB discovery cohort characteristics. Basic demographic characteristics and representative Phecodes identified by PheWAS of median hepatic fat in PMBB. Each characteristic is labeled with count data and percent prevalence where appropriate. Individuals were determined to be a case for a Phecode if they had the corresponding ICD diagnosis on two or more dates, while controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as those under control exclusion criteria based on Phecode mapping protocols were not considered. AFR-African, AMR-Mixed American, EAS-East Asian, EUR-European, SAS-South Asian

5.4.2. Exome-wide analyses of single coding variants identify novel variants associated with hepatic fat

We conducted a univariate exome-wide analysis of hepatic fat for all nonsynonymous coding variants of sufficient frequency (MAF>0.1% in gnomAD) (Figure 5.1; Appendix D, Figure S2). Among 120,315 total variants with at least 10 carriers, we identified 91 variants in 86 genes with exome-wide significant (P<4.2E-07) or suggestive (P<9.9E-05) associations with hepatic fat (Appendix D, Table S2). Among these included variants previously reported to be associated with hepatic fat: *PNPLA3* variants I148M, K434E, and S453I; *TM6SF2* E167K; *SAMM50* D110G; *NCAN* P92S; *PARVB* W37R; and APOE4 (C130R). Additional positive control associations were found below the significance threshold (Appendix D, Table S3), including *GCKR* L446P, *MTARC1* T165A, and *TM6SF2* L156P. 27 of the 91 single variants were African ancestry-specific variants (African/European MAF ratio > 10 in gnomAD; Appendix D, Table S2).

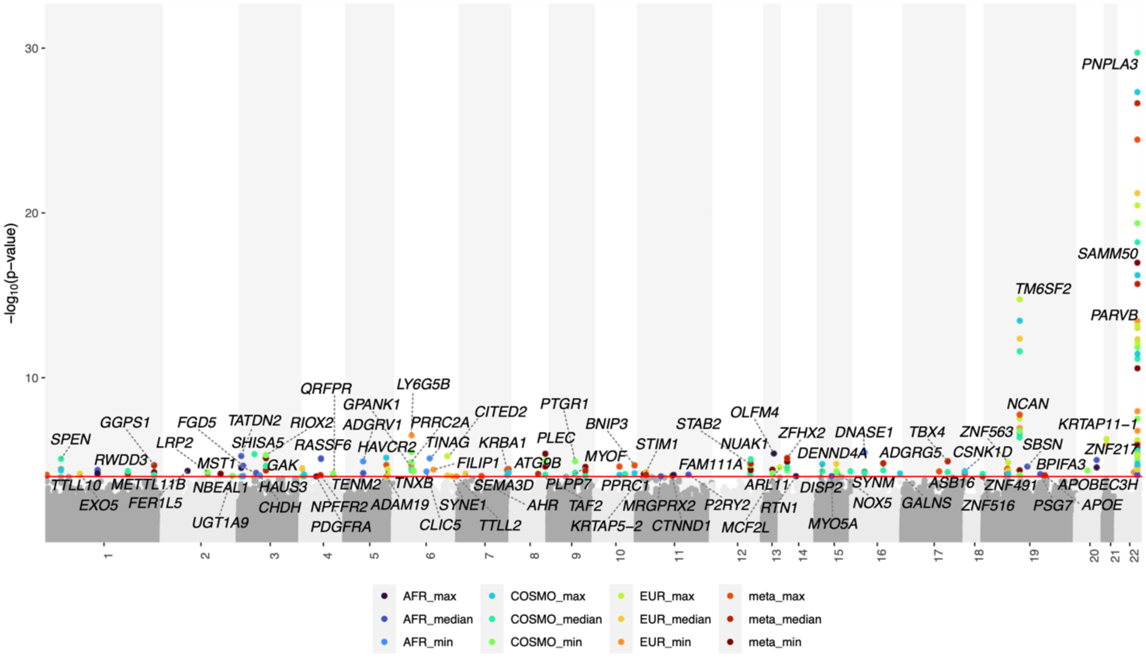

Figure 5.1. Manhattan plot of single variant discovery in PMBB. Manhattan plot showing the results of the exome-wide single variant discovery analysis in PMBB for coding variants of sufficient frequency (MAF>0.1%, N≥10). The x axis represents the exome and is organized by chromosomal location. The

location of each single variant along the x axis corresponds to the genomic location for each variant according to Genome Reference Consortium Human build 38 (GRCh38). The association of each single variant with hepatic fat is plotted vertically above each variant. Each point is color-coded according to the key for the type of analysis conducted, and the height of each point represents the $-\log_{10}$(p value) of the association. Each variant is annotated with its corresponding gene name. The red line represents the suggestive significance threshold at p=9.9E-05 to account for multiple hypothesis testing.

For replication, we tested the association of these 91 variants with liver proton density fat fraction (PDFF) measurements derived from abdominal MRI scans in the UKB (Appendix D, Table S4). Not only were the methods of hepatic fat quantitation different, but there were notable differences in the distribution of hepatic fat between PMBB and UKB (Appendix D, Figure S3). Furthermore, there was insufficient power in UKB for replication of the 27 African ancestry-specific variants, as 20 of 27 variants had N<5 in the cosmopolitan analyses. Despite this, in addition to replicating a number of the previously known associations, we also replicated novel associations of H600Y in *FGD5* and a 6-bp deletion (S198_G199del) in *CITED2* with hepatic fat.

To test whether rare predicted deleterious coding variants in the 86 unique genes containing the 91 single variants were also associated with differences in hepatic fat, we aggregated rare (MAF≤0.1% in gnomAD) pLOF variants into gene burdens for targeted associations with hepatic fat in PMBB (Appendix D, Table S5). A burden of rare pLOF variants in 8 genes was significantly associated with hepatic fat, including genes with previously described common variation associated in differences in hepatic fat such as *PNPLA3* and *PARVB*, as well as new findings like *PTGR1*. We also aggregated the combination of rare pLOF and rare predicted deleterious missense (pDM) variants (REVEL≥0.5) per gene for targeted gene burden association with hepatic fat in PMBB (Appendix D, Table S5). We found 11 additional genes associated with differences in hepatic fat in PMBB by adding rare pDM variants to pLOFs. Additionally, the associations for *DISP2*, *PARVB*, *PTGR1*, and *QRFPR* were strengthened or remained significant with the addition of carriers for rare pDM variants. Of note, the combined

CITED2 gene burden was nominally associated with increased median hepatic fat although underpowered (N=3, beta=9.723, p=0.0276).

We also tested the burden of rare pLOFs in the 86 genes for association with hepatic fat in UKB (Appendix D, Table S6) and found 3 concordant gene burdens, namely *ADAM19*, *EXO5*, and *SEMA3D*. Additionally, the combined burden of rare pLOFs and rare pDM variants (Appendix D, Table S6) revealed 3 additional significant genes, namely *ADGRG5*, *GPS1*, and *NOX5*. The associations for *EXO5* and *SEMA3D* were strengthened or remained nominally significant with the addition of carriers for rare pDM variants.

5.4.3. A gene burden of rare pLOFs in LMF2 is associated with increased hepatic fat in PMBB and UKB

We also performed an exome-wide rare pLOF gene burden analysis for association with hepatic fat in PMBB. We aggregated rare (MAF≤0.1% in gnomAD) pLOFs per gene: among 4187 genes with at least 10 carriers for rare pLOFs who have hepatic fat quantifications available (Appendix D, Figure S4), there were 26 genes that had exome-wide significant (P<1.2E-05) or suggestive (P<9.9E-04) associations with hepatic fat (Figure 5.2; Appendix D, Table S7). For these 26 genes, we attempted replication in UKB using a similar rare pLOF gene burden approach. We found that the *LMF2* pLOF gene burden had a significant association with hepatic fat (beta=0.429, p=5.79E-03, N=5) and importantly in the same direction (increased) as we observed in PMBB.

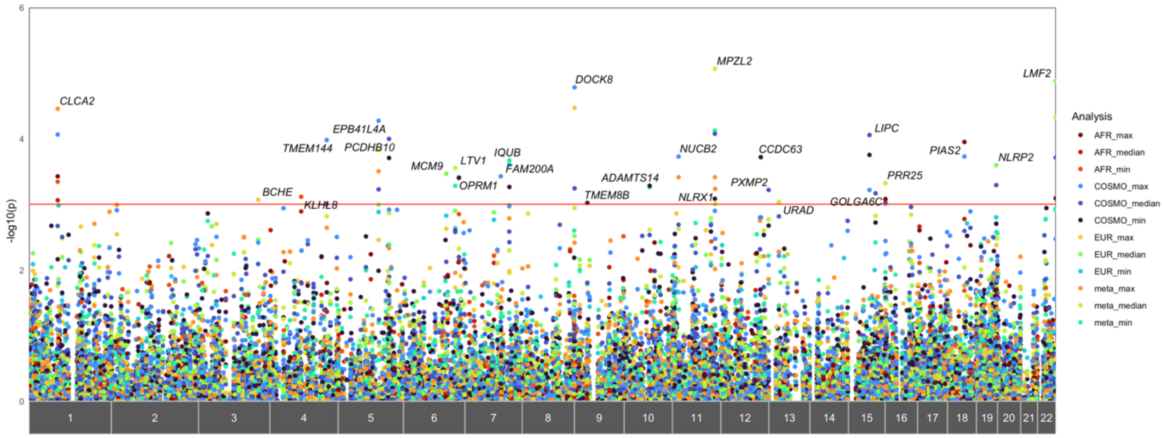

Figure 5.2. Manhattan plot of rare pLOF gene burden discovery in PMBB. Manhattan plot showing the results of the exome-wide gene burden discovery analysis in PMBB aggregating rare (MAF≤0.1%) predicted loss-of-function (pLOF) variants per gene (N≥10). The x axis represents the exome and is organized by chromosomal location. The location of each gene along the x axis corresponds to the genomic location for each variant according to Genome Reference Consortium Human build 38 (GRCh38). The association of each gene burden with hepatic fat is plotted vertically above each gene. Each point is color-coded according to the key for the type of analysis conducted, and the height of each point represents the $-\log_{10}$(p value) of the association. Each gene is annotated with its gene name. The red line represents the suggestive significance threshold at p=9.9E-04 to account for multiple hypothesis testing.

We conducted an analysis of the association of the gene burden of rare pLOFs in *LMF2* in PMBB (N=105 het carriers) with Phecodes in the “digestive” group (520 to 579.8). In addition to a significant association with the NAFLD Phecode, the burden of pLOF variants in *LMF2* had significant associations with an array of biliary-related Phecodes such as “Cholangitis”, “Calculus of bile duct”, “Cholelithiasis with acute cholecystitis”, and “Primary biliary cirrhosis” (Appendix D, Table S8). Furthermore, the *LMF2* pLOF gene burden was associated with increased serum alkaline phosphatase, total cholesterol levels, and BMI (Appendix D, Table S9), as well as a trend to an increase in triglycerides (beta=0.0474, p=0.0957). There were no significant associations with HDL, LDL, ALT, or AST levels.

In a primary analysis of publicly available data, we found that *Lmf2* expression was significantly decreased in liver endothelial cells from cirrhotic livers of rats versus control

(Appendix D, Figure S5), where cirrhosis was induced via inhalation of CCl_4. Similarly, in a hepatic steatosis mouse model due to hepatocyte-specific *Sirt1* deficiency leading to impaired PPARα signaling and decreased fatty acid beta-oxidation, *Lmf2* liver expression was also significantly decreased versus control (Appendix D, Figure S6).

5.5. Discussion

We present an exome-wide analysis of coding variants for association with hepatic fat quantitated from abdominal CT scans via machine learning in a medical biobank. Our single variant discovery analyses confirmed many previously described single variants associated with hepatic fat or NAFLD risk which support the validity of our data and analysis pipeline, including I148M, K434E, and S453I in *PNPLA3*,[122,129] E167K and L156P in *TM6SF2*,[124,130] D110G in *SAMM50*,[131] P92S in *NCAN*,[132] W37R in *PARVB*,[133] C130R in *APOE*,[134] L446P in *GCKR*,[135] and T165A in *MTARC1*.[136] We also conducted replication studies in UKB looking for consistent directions of effect by linking exomes to hepatic fat based on liver PDFF extracted from abdominal MRI scans, given the strong linear correlation between hepatic fat extracted from non-contrast CT and PDFF quantifications derived from MRI.[137,138]

Among our new significant single variants not previously associated with hepatic steatosis in humans, we found that H600Y in *FGD5* was associated with increased hepatic fat in PMBB, which replicated in UKB. *FGD5* encodes a protein which activates CDC42[139] and is also expressed in the liver.[140] *CDC42* is a member of the Rho GTPase family and plays important roles in the regulation of the cytoskeleton as well as cell proliferation, polarity, and transport. Importantly, liver-specific knockout of *Cdc42* in mice has been shown to lead to excessive hepatic accumulation of lipids during liver regeneration after partial hepatectomy, likely due to impaired cytoskeletal organization and intracellular trafficking in hepatocytes.[141] Additionally, we found that a 6-bp deletion in *CITED2*, a coactivator of HNF4α, was also associated with increased hepatic fat in PMBB, which also replicated in UKB. A burden of rare predicted deleterious coding variants in *CITED2* was also associated with increased hepatic fat although underpowered. *Cited2* has been shown to be essential for mouse fetal liver development, and knockout of *Cited2* in fetal liver leads to disrupted sinusoidal architecture and accumulation of lipid droplets in the sinusoidal space.[142]

There is a substantial gap of knowledge regarding the clinical implications of genetic variants overrepresented among Africans due to the lack of ancestral diversity in populations previously studied, especially regarding AFR-predominant genetic variants that modulate hepatic fat.[116] In the PMBB discovery cohort, 27.4% of individuals with WES linked to hepatic fat quantifications were of African ancestry, and interestingly we identified 27 single variants enriched among individuals of African ancestry which were significant in the PMBB discovery. These variants represented a challenge for replication in UKB given that there are only 70 individuals who identified as having African, Caribbean or any other Black ethnic background in the subset of individuals with MRI-derived hepatic fat linked to WES data in UKB. The previously described AFR-predominant S453I variant in *PNPLA3* and the new S198_G199del variant in *CITED2*, among the more common of the 27 AFR-predominant variants, were able to be replicated via cosmopolitan analyses in UKB. However, the rest of the AFR-predominant single variants were in very low numbers in UKB and failed to replicate, likely due to lack of power, with 20 of 27 variants having N<5 and thus not even being included in replication studies. Our findings suggest that larger experiments of this type in ethnically diverse cohorts are essential for improving our understanding of the contribution of ancestry-specific genetic variation to the regulation of intrahepatic fat.

For significant single variants, we also tested whether a gene burden of rare predicted deleterious coding variants might also be associated with differences in intrahepatic fat. While phenotype-first approaches to identifying pLOF carriers in genes nominated by GWAS loci for hepatic fat among NAFLD cases have been difficult,[129] our study used a genome-first approach to identify pLOF carriers in *PNPLA3*, *PARVB*, and *SAMM50* for associations with hepatic fat. We report for the first time the associations of pLOFs in *PNPLA3* and *PARVB* as well as pLOFs and pDM variants in *PARVB* and *SAMM50* with decreased hepatic fat in PMBB. The *PNPLA3* pLOF gene burden's protective association is consistent with previous studies suggesting that accumulation of PNPLA3 protein on lipid droplets causes hepatic steatosis, and that depletion of the protein may be a potential strategy for therapeutic intervention.[143] While additional functional work would be needed to validate the associations for the *PARVB* and *SAMM50* gene burdens, our study suggests a protective role for haploinsufficiency in genes that were identified via common variant associations with incompletely described functionalities with respect to gene products.

The targeted gene burden associations for genes nominated by the single variant discovery also shed light to genes with previously undescribed relationships to hepatic fat. Both pLOFs and pDM variants in *PTGR1* were associated with increased hepatic fat in PMBB. *PTGR1* encodes an enzyme involved in the inactivation of the chemotactic factor leukotriene B4 and also has its highest expression in the liver.[144] Notably, leukotriene B4 has been shown to promote insulin resistance in mouse hepatocytes,[145] suggesting that haploinsufficiency of *PTGR1* could lead to unsuppressed activation of leukotriene B4 and thus development of insulin resistance in the liver. Additionally, we found that pLOF variants in *ADAM19* were associated with decreased hepatic fat in UKB, consistent with the direction of effect of its respective single variant from discovery (G660D), a pDM variant in *ADAM19* (REVEL=0.683). Importantly, *ADAM19* has been suggested to be pro-obesogenic and enhance insulin resistance, and neutralization may be a potential therapeutic approach to treating obesity and T2D.[146] We also identified gene burdens of pLOFs as well as pLOFs combined with pDM variants in *QRFPR* (*GPR103*), the receptor for neuropeptide 26RFa (encoded by *QRFP*), as being associated with increased hepatic fat. 26RFa and *QRFPR* work both in the hypothalamic nuclei to control feeding behavior as well as in the gut and pancreatic islets.[147] Specifically, 26RFa increases insulin sensitivity and prevents pancreatic beta cell death and apoptosis, and disruption leads to dysregulation of glucose homeostasis and a deficit in insulin production by pancreatic islets.[148-150] Taken together, our results suggest that haploinsufficiency of *QRFPR* increases risk for hepatic steatosis through a mechanism of insulin resistance.

We also report the first rare variant gene burden discovery analysis for exome-wide gene-based associations with hepatic fat in a medical biobank. We identified a burden of rare pLOF variants in *LMF2* as being associated with increased hepatic fat in PMBB, which replicated in UKB. Importantly, this association would have been missed in previous studies focusing on common variants, and common variants in *LMF2* were also not exome-wide significant in our single variant discovery. *LMF2* is a paralog of *LMF1* and is the ancestral gene, and *Lmf1* emerged in echinoderms after losing an internal segment of the DUF1222 domain and gaining a new C-terminal tail with lipase maturation activity. While *LMF2* does not have this C-terminus, it may share a common ancestral cellular function with *LMF1*, possibly the maintenance of endoplasmic reticulum homeostasis. While *LMF2* lacks description in the literature, a study of adolescents undergoing bariatric surgery with a high prevalence of NAFLD showed that *LMF2* was one of the top down-regulated genes in the livers of those with non-alcoholic steatohepatitis (NASH) versus not NAFLD control subjects.[151] Similarly, hepatic *Lmf2* expression was decreased in cirrhotic rat livers and steatotic mouse livers versus controls.[127,128] While additional description of the function of *LMF2* is needed, our study suggests that *LMF2* haploinsufficiency may contribute to an increase in hepatic fat.

While many single variants successfully replicated in UKB following discovery analyses in PMBB, we noticed a relative lack of significant findings when interrogating rare variant gene burdens in UKB regarding replication of significant genes from the gene burden discovery as well as targeted analyses of genes nominated by the single variant discovery. While the number of samples with exome sequences linked to hepatic fat in UKB was comparable to PMBB, the distribution of hepatic fat was substantially skewed toward lower values in UKB compared to the normal distribution in PMBB. This might be expected, given the UKB is a population-based biobank that is widely recognized to have a “healthy volunteer selection bias”,[117] and we have previously described the relative lack of replication for rare variant gene burdens in UKB compared to medical biobanks.[22] Thus, our study suggests that additional experiments of this type linking exome sequencing to imaging-derived hepatic fat quantifications in medical biobanks are warranted to interrogate the impact of rare variation on differences in hepatic fat.

We recognize that while CT imaging provides the opportunity to quantify hepatic fat through rapid and scalable imaging, there are some limitations. In particular, the presence of iron, copper, glycogen, fibrosis, or edema may confound attenuation values and lead to errors in hepatic fat quantification.[152] However, by conducting replication studies in UKB by interrogating PDFF derived from MRI, we suggest that replicated signals are less likely to be confounded by these factors, given that PDFF is a more accurate measure with less confounding variables.[153] Thus, while the magnitude of effect size for genetic associations may differ between CT-derived and MRI-derived hepatic fat quantifications, we suggest that there is value in interrogating multiple imaging modalities for increased specificity.

In conclusion, by linking exome sequences to hepatic fat quantifications in a medical biobank, our study not only adds to the breadth of knowledge regarding single coding variants and their associations with hepatic fat and NAFLD risk through our exome-wide single variant discovery analyses, but also provides gene-based associations which support the genes nominated by single variants and suggests mechanisms of haploinsufficiency by which these genes may affect the regulation of intrahepatic fat. Furthermore, our study demonstrates the feasibility and value of aggregating rare predicted deleterious coding variants into gene burdens on an exome-wide scale for association with hepatic fat quantifications for the discovery of new genes which may regulate intrahepatic fat and confer risk for or protect from NAFLD. We suggest that much larger experiments of this type which link genetic sequencing to quantitated hepatic fat will lead to many new insights into genes which regulate hepatic fat and modulate risk for NAFLD and other metabolic comorbidities.

CHAPTER 6. Conclusions and future directions

The analyses of this dissertation clearly demonstrate the value of using a genome-first approach to interrogating rare coding variation to describe how loss of gene function may impact human disease. By linking whole-exome sequences to electronic health records (EHR) in the Penn Medicine BioBank (PMBB), this dissertation first determined key methodologies for a genome-first approach to selecting rare variants for gene burden association tests to establish a platform by which the upcoming analyses could apply these approaches to provide new insights into the role of many genes in human biology and phenotypes. Importantly, the significance of assessing both predicted loss-of-function (pLOF) and predicted deleterious missense (pDM) variants for gene burden testing was emphasized through the explorations of variant selection methods. Then, by applying these approaches, the analyses of this dissertation demonstrated the utility of a genome-first approach to conducting gene burden association tests to describe expected and unsuspected clinical implications of loss-of-function in both well-described and novel genes. More specifically, Chapters 2 and 3 took advantage of known 'positive control' gene-phenotype associations to assess the performance of rare variant selection methods for gene burden association testing, and in doing so also revealed novel pleiotropic relationships for *LMNA* as well as major gene-specific differences in ideal gene burdening approaches for two genes both intimately tied to the pathogenesis of the same disease, namely *MYBPC3* and *MYH7*. Then, Chapters 4 and 5 demonstrated how application of a genome-first approach to rare variant gene burden testing on an exome-wide scale allowed for identification of several novel gene-phenotype relationships.

While PMBB is a healthcare-based population with enrichment for rare diseases which allowed for the potential to substantially gain new insights into the biology of human phenotypes and diseases in this dissertation, a major takeaway and suggestion from these analyses is that application of this genome-first approach to even larger numbers of individuals are likely to be highly informative, providing the statistical power required to uncover unexplored relationships between rare genetic variation and human disease phenotypes. Already there are efforts underway to conduct rare variant-based exome-by-phenome-wide association studies in larger populations, as shown by studies in the first 45K+ individuals with exome sequencing linked to disease phenotypes in the UK Biobank (UKB)[21,118] as well as updated analyses in the larger, more recent ~200K individuals with exome sequencing in UKB.[154] However, given that the UKB is widely recognized to have a 'healthy volunteer selection bias'[117] as we also demonstrated through comparisons of disease prevalence between multiple medical biobanks versus UKB in Chapter 4, there are clearly many more opportunities ahead within the realm of investigating the clinical implications of rare coding variation in sicker populations such as healthcare-based medical biobanks to uncover additional insights into novel gene biology and function, as we still do not know how loss-of-function in most human genes impact disease.

In addition to being enriched for disease, PMBB has the advantage of being ancestrally diverse. The analyses of this dissertation were empowered to conduct genetic association studies in both African and European ancestry due to the relatively high prevalence of individuals of African ancestry in PMBB (19.9% to 27.4% across Chapters 2-5). There has been a significant gap of knowledge regarding the clinical implications of genetic variants overrepresented among individuals of Africans ancestry due to previous genetic association studies lacking ancestral diversity in the populations that have been analyzed, usually being limited to individuals of European ancestry.[116] Importantly, this dissertation addressed this gap of knowledge by not only including individuals of African ancestry for all analyses performed but also identifying several African ancestry-specific single variants which are robustly associated with diseases and traits. For example, Chapter 4 identified 16 rare predicted deleterious single variants which are African ancestry-specific and that replicated associations with the same disease in which a pLOF gene burden was associated in the exome-by-phenome-wide discovery. Additionally, Chapter 5 identified 27 single variants enriched among individuals of African ancestry which were significantly associated with differences in intrahepatic fat in the PMBB discovery analyses. As might be expected, most of these African-specific variants neither exist in the GWAS catalog nor have been previously mentioned in the published literature. These analyses suggest that it is imperative for future studies in this field to address the need for more diversity in genomic experiments of this type by analyzing ethnically diverse cohorts to improve our understanding of the contribution of previously missed ancestry-specific variation to human disease.

In the 2020 National Human Genome Research Institute's (NHGRI) Strategic Vision, the NHGRI made ten bold predictions for human genomics that might come true by 2030,[155] many of which are relevant to future directions for these types of studies. Prediction #2 states that "the biological function(s) of every human gene will be known," which is a major goal and impetus for conducting larger studies of this type in diverse, sick populations. One obvious place to start is for all Mendelian genes with known phenotype associations but remaining enigmatic gene biology and clinical implications. Application of a genome-first approach to rare variants like the analyses performed in Chapters 2-4 for all such Mendelian genes could cover significant distance toward reaching that goal. For example, Chapter 4 confirmed that a gene burden of rare pLOFs in *CFTR* was significantly associated with cystic fibrosis (CF), a recessive Mendelian condition caused by biallelic variants in *CFTR*. But in addition, we found that the *CFTR* pLOF gene burden was also significantly associated with bronchiectasis independent of a CF diagnosis and occurred in individuals without a second *CFTR* variant, strengthening previous suggestions of predisposition to bronchiectasis due to haploinsufficiency of *CFTR*.[89] These kinds of analyses of Mendelian and clinically actionable genes could also help clarify the interpretation and relevance of genomic variants encountered in the clinical setting, essentially heading toward achieving prediction #7 of "rendering the diagnostic designation 'variant of uncertain significance (VUS' obsolete." Furthermore, such studies could aid in the translation of important findings from genomic research into the clinic such that the regular use of genomic information will become mainstream in all clinical settings.

The analyses of this dissertation demonstrated how a genome-first approach to interrogating rare loss-of-function variants can link haploinsufficiency in a gene to conferment of risk for or protection from disease. However, the relative scarcity of homozygotes or compound heterozygotes for rare loss-of-function variants in these analyses represents a major limitation of the genome-first approach in outbred populations like PMBB. This relative scarcity essentially mandated the rare variant analyses of this dissertation to explore gene-phenotype associations based on haploinsufficiency. However, other studies focusing on consanguineous, founder, or bottleneck populations such as the Pakistan Risk of Myocardial Infarction Study (PROMIS)[156] and FinnGen[157] are likely to be more successful in identifying homozygotes for ascertaining phenotypes that are ‘recessive’ as are many linked to Mendelian genes, though statistical power for interrogation of rare disease phenotypes will continue to be a challenge.

This dissertation also demonstrated how quantitative imaging-derived phenotypes (IDP) that are correlated with disease risk can provide information beyond that captured in binary EHR diagnoses. For example, Chapters 2 and 3 used echocardiographic data to show that carriers of predicted deleterious variants in *LMNA*, *MYBPC3*, and *MYH7* have abnormal cardiac morphologies which may not necessarily pass the threshold for clinical diagnosis of disease but certainly warrant clinical follow-up for preventive actions. As another example, Chapter 5 demonstrated how exome-wide analyses of hepatic fat quantifications, which are correlated with risk for non-alcoholic fatty liver disease (NAFLD) as well as various other comorbidities such as type 2 diabetes and cardiovascular disease, shed light to new genetic variants which are associated with differences in hepatic fat but may not be captured when analyzing binary NAFLD diagnosis rates. Another advantage of analyzing the genetic architecture of quantitative IDPs, which was not specifically assessed in this dissertation, is the potential to interrogate directions of effect. Future studies which prove that loss-of-function variants in a certain gene lead to one direction of effect for a given IDP and subsequently discover another variant in the same gene that is robustly associated with the opposite direction of effect for the same IDP could identify novel gain-of-function variants.

Importantly, this concept could extend to genomic analyses of any clinical quantitative phenotype, which leads to another major future direction for analyzing rare variation. If one could identify that loss-of-function in a gene leads to protection from disease, this could be directly translated to curative therapies. For example, gain-of-function variants in *PCSK9* were discovered to be associated with very high circulating LDL cholesterol levels, which led to the subsequent finding through targeted sequencing of individuals with very low circulating LDL levels that loss-of-function variants in *PCSK9* lower LDL.[158] This finding ultimately led to the development of PCSK9 inhibitors, which are monoclonal antibodies that antagonize PCSK9 protein and lead to increased LDL receptors and thus therapeutically decreased circulating LDL cholesterol levels.[159] Additionally, identification of protective gain-of-function variation could also lead to development of curative therapies, although identification of such variants as well as therapeutic translation through genomic modifications may be more difficult. Analyzing quantitative traits such as IDPs may allow such discovery of protective gain-of-function variants to be more feasible due to the ability to compare directions of effect for the associations of different coding variants in the same gene with a single quantitative phenotype. For example, knowing that loss-of-function variants in *APOA5* lead to hypertriglyceridemia,[160] one could systematically search for single coding variants in *APOA5* that are associated with decreased serum triglyceride levels to uncover potentially novel gain-of-function variants which are protective for cardiovascular disease. Thus, systematic analyses of quantitative traits in large biobanks could help in identifying novel gain-of-function variants throughout the human genome. Furthermore, recent advances in *in silico* predictions of three-dimensional protein folding may accelerate the identification of gain-of-function variants,[161] which could one day allow for genome-first approaches to exome-wide analyses of the clinical implications of human gain-of-function variants. This may make NHGRI’s bold prediction #10 for

2030 feasible—that “breakthrough discoveries will lead to curative therapies involving genomic modifications for dozens of genetic diseases.”

The success in identifying compelling findings based on the limited number of individuals included in the analyses of this dissertation (~11K to ~44K in PMBB) is promising, in that there is hope that the rate of new discovery for gene-phenotype relationships will accelerate over the next decade as the approaches detailed in this dissertation are applied to much larger populations and thus potentiate many new insights into gene biology and function through increased statistical power. As our understanding of the biological and clinical relevance of rare variation becomes more complete, genomic testing becomes more routine in the clinical setting, and a better understanding of the genomic architecture of human disease becomes translated to curative and even preventive therapeutics, human health disparities and outcomes are likely to improve with the hope that these future services are equitably offered, as it should be a human right to appropriately act upon the genome one is innately inherited with for the benefit of one’s health.

BIBLIOGRAPHY

1 Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**, 443-453, doi:10.1056/NEJMoa072366 (2007).

2 Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645, doi:10.1038/ng.120 (2008).

3 Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17**, 692-704, doi:10.1038/nrc.2017.82 (2017).

4 Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**, 199-215, doi:10.1016/j.ajhg.2015.06.009 (2015).

5 Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870-1879, doi:10.1001/jama.2014.14601 (2014).

6 Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872-877, doi:10.1016/j.cell.2014.02.002 (2014).

7 Mefford, H. C. Genotype to phenotype-discovery and characterization of novel genomic disorders in a "genotype-first" era. *Genet Med* **11**, 836-842, doi:10.1097/GIM.0b013e3181c175d2 (2009).

8 Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, doi:10.1126/science.aaf6814 (2016).

9 Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet* **17**, 129-145, doi:10.1038/nrg.2015.36 (2016).

10 Park, J. *et al.* A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes. *Genet Med*, doi:10.1038/s41436-019-0625-8 (2019).

11 Zhang, X., Basile, A. O., Pendergrass, S. A. & Ritchie, M. D. Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics* **20**, 46, doi:10.1186/s12859-018-2591-6 (2019).

12 Verma, A. *et al.* Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals. *Am J Hum Genet* **104**, 55-64, doi:10.1016/j.ajhg.2018.11.006 (2019).

13 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23, doi:10.1016/j.ajhg.2014.06.009 (2014).

14 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).

15 Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* **100**, 267-280, doi:10.1016/j.ajhg.2017.01.004 (2017).

16 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

17 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

18 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).

19 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362, doi:10.1038/nmeth.2890 (2014).

20 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885, doi:10.1016/j.ajhg.2016.08.016 (2016).

21 Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun* **11**, 542, doi:10.1038/s41467-020-14288-y (2020).

22 Park, J. *et al.* Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. *Nat Med* **27**, 66-72, doi:10.1038/s41591-020-1133-8 (2021).

23 Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494, doi:10.1371/journal.pgen.1004494 (2014).

24 Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, doi:10.1126/science.aaf7000 (2016).

25 Worman, H. J. & Bonne, G. "Laminopathies": a wide spectrum of human diseases. *Exp Cell Res* **313**, 2121-2133, doi:10.1016/j.yexcr.2007.03.028 (2007).

26 Benedetti, S. *et al.* Phenotypic clustering of lamin A/C mutations in neuromuscular patients. *Neurology* **69**, 1285-1292, doi:10.1212/01.wnl.0000261254.87181.80 (2007).

27 Capell, B. C. & Collins, F. S. Human laminopathies: nuclei gone genetically awry. *Nat Rev Genet* **7**, 940-952, doi:10.1038/nrg1906 (2006).

28 Genschel, J. & Schmidt, H. H. Mutations in the LMNA gene encoding lamin A/C. *Hum Mutat* **16**, 451-459, doi:10.1002/1098-1004(200012)16:6<451::AID-HUMU1>3.0.CO;2-9 (2000).

29 Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* **18**, 906-913, doi:10.1038/gim.2015.187 (2016).

30 Dewey, F. E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N Engl J Med* **377**, 211-221, doi:10.1056/NEJMoa1612790 (2017).

31 Staples, J. *et al.* Profiling and Leveraging Relatedness in a Precision Medicine Cohort of 92,455 Exomes. *Am J Hum Genet* **102**, 874-889, doi:10.1016/j.ajhg.2018.03.012 (2018).

32 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

33 Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110, doi:10.1038/nbt.2749 (2013).

34 Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-2376, doi:10.1093/bioinformatics/btu197 (2014).

35 Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832-838, doi:10.1016/j.ajhg.2010.04.005 (2010).

36 Verma, A. *et al.* A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics* **19**, 120, doi:10.1186/s12859-018-2135-0 (2018).

37 Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3, doi:10.1186/1471-2164-14-S3-S3 (2013).

38 Douville, C. *et al.* Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat* **37**, 28-35, doi:10.1002/humu.22911 (2016).

39 Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586, doi:10.1038/ng.3703 (2016).

40 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

41 MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476, doi:10.1038/nature13127 (2014).

42 Marian, A. J. Causality in genetics: the gradient of genetic effects and back to Koch's postulates of causality. *Circ Res* **114**, e18-21, doi:10.1161/CIRCRESAHA.114.302904 (2014).

43 Burkett, E. L. & Hershberger, R. E. Clinical and genetic issues in familial dilated cardiomyopathy. *J Am Coll Cardiol* **45**, 969-981, doi:10.1016/j.jacc.2004.11.066 (2005).

44 Newton-Cheh, C. Should Identifying a Titin Truncating Variant Change the Management of Patients With Dilated Cardiomyopathy? *J Am Coll Cardiol* **70**, 2275-2277, doi:10.1016/j.jacc.2017.09.020 (2017).

45 Hasselberg, N. E. *et al.* Lamin A/C cardiomyopathy: young onset, high penetrance, and frequent need for heart transplantation. *Eur Heart J*, doi:10.1093/eurheartj/ehx596 (2017).

46 Ellepola, C. D., Knight, L. M., Fischbach, P. & Deshpande, S. R. Genetic Testing in Pediatric Cardiomyopathy. *Pediatr Cardiol*, doi:10.1007/s00246-017-1779-2 (2017).

47 Anselme, F. *et al.* Implantable cardioverter-defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. *Heart Rhythm* **10**, 1492-1498, doi:10.1016/j.hrthm.2013.06.020 (2013).

48 Taylor, M. R. *et al.* Natural history of dilated cardiomyopathy due to lamin A/C gene mutations. *J Am Coll Cardiol* **41**, 771-780 (2003).

49 Lee, J. M. *et al.* Modulation of LMNA splicing as a strategy to treat prelamin A diseases. *The Journal of clinical investigation* **126**, 1592-1602, doi:10.1172/JCI85908 (2016).

50 Thong, K. M. *et al.* Cosegregation of focal segmental glomerulosclerosis in a family with familial partial lipodystrophy due to a mutation in LMNA. *Nephron Clin Pract* **124**, 31-37, doi:10.1159/000354716 (2013).

51 Imachi, H. *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* **35**, 18-21, doi:10.1007/s12020-008-9127-1 (2009).

52 Burke, M. A., Cook, S. A., Seidman, J. G. & Seidman, C. E. Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy. *J Am Coll Cardiol* **68**, 2871-2886, doi:10.1016/j.jacc.2016.08.079 (2016).

53 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524, doi:10.1016/j.cell.2018.12.015 (2019).

54 Thompson, A. D. *et al.* Computational prediction of protein subdomain stability in MYBPC3 enables clinical risk stratification in hypertrophic cardiomyopathy and enhances variant interpretation. *Genet Med*, doi:10.1038/s41436-021-01134-9 (2021).

55 Helms, A. S. *et al.* Effects of MYBPC3 loss-of-function mutations preceding hypertrophic cardiomyopathy. *JCI Insight* **5**, doi:10.1172/jci.insight.133782 (2020).

56 Sabater-Molina, M., Perez-Sanchez, I., Hernandez Del Rincon, J. P. & Gimeno, J. R. Genetics of hypertrophic cardiomyopathy: A review of current state. *Clinical genetics* **93**, 3-14, doi:10.1111/cge.13027 (2018).

57 Helms, A. S. *et al.* Spatial and Functional Distribution of MYBPC3 Pathogenic Variants and Clinical Outcomes in Patients With Hypertrophic Cardiomyopathy. *Circ Genom Precis Med* **13**, 396-405, doi:10.1161/CIRCGEN.120.002929 (2020).

58 Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* **19**, 192-203, doi:10.1038/gim.2016.90 (2017).

59 Seeger, T. *et al.* A Premature Termination Codon Mutation in MYBPC3 Causes Hypertrophic Cardiomyopathy via Chronic Activation of Nonsense-Mediated Decay. *Circulation* **139**, 799-811, doi:10.1161/CIRCULATIONAHA.118.034624 (2019).

60 Klaassen, S. *et al.* Mutations in sarcomere protein genes in left ventricular noncompaction. *Circulation* **117**, 2893-2901, doi:10.1161/CIRCULATIONAHA.107.746164 (2008).

61 Witjas-Paalberends, E. R. *et al.* Mutations in MYH7 reduce the force generating capacity of sarcomeres in human familial hypertrophic cardiomyopathy. *Cardiovasc Res* **99**, 432-441, doi:10.1093/cvr/cvt119 (2013).

62 Glazier, A. A., Thompson, A. & Day, S. M. Allelic imbalance and haploinsufficiency in MYBPC3-linked hypertrophic cardiomyopathy. *Pflugers Arch* **471**, 781-793, doi:10.1007/s00424-018-2226-9 (2019).

63 Cirino, A. L., Seidman, C. E. & Ho, C. Y. Genetic Testing and Counseling for Hypertrophic Cardiomyopathy. *Cardiol Clin* **37**, 35-43, doi:10.1016/j.ccl.2018.08.003 (2019).

64 Bagnall, R. D. *et al.* Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients With Hypertrophic Cardiomyopathy. *J Am Coll Cardiol* **72**, 419-429, doi:10.1016/j.jacc.2018.04.078 (2018).

65 Mital, S. *et al.* Enhancing Literacy in Cardiovascular Genetics: A Scientific Statement From the American Heart Association. *Circ Cardiovasc Genet* **9**, 448-467, doi:10.1161/HCG.0000000000000031 (2016).

66 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

67 Fiorillo, C. *et al.* MYH7-related myopathies: clinical, histopathological and imaging findings in a cohort of Italian patients. *Orphanet J Rare Dis* **11**, 91, doi:10.1186/s13023-016-0476-1 (2016).

68 Haggerty, C. M. *et al.* Genomics-First Evaluation of Heart Disease Associated With Titin-Truncating Variants. *Circulation* **140**, 42-54, doi:10.1161/CIRCULATIONAHA.119.039573 (2019).

69 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

70 QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe (2006).

71 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

72 Harrison, P. F., Pattison, A. D., Powell, D. R. & Beilharz, T. H. Topconfects: a package for confident effect sizes in differential expression analysis provides a more biologically useful ranked gene list. *Genome Biol* **20**, 67, doi:10.1186/s13059-019-1674-7 (2019).

73 Wagner, A. H. *et al.* Exon-level expression profiling of ocular tissues. *Exp Eye Res* **111**, 105-111, doi:10.1016/j.exer.2013.03.004 (2013).

74 Libby, R. T. *et al.* Inherited glaucoma in DBA/2J mice: pertinent disease features for studying the neurodegeneration. *Vis Neurosci* **22**, 637-648, doi:10.1017/S0952523805225130 (2005).

75 Howell, G. R., Walton, D. O., King, B. L., Libby, R. T. & John, S. W. Datgan, a reusable software system for facile interrogation and visualization of complex transcription profiling data. *BMC Genomics* **12**, 429, doi:10.1186/1471-2164-12-429 (2011).

76 Yang, W. *et al.* Generation of iPSCs as a Pooled Culture Using Magnetic Activated Cell Sorting of Newly Reprogrammed Cells. *PLoS One* **10**, e0134995, doi:10.1371/journal.pone.0134995 (2015).

77 Chavali, V. R. M. *et al.* Dual SMAD inhibition and Wnt inhibition enable efficient and reproducible differentiations of induced pluripotent stem cells into retinal ganglion cells. *Sci Rep* **10**, 11828, doi:10.1038/s41598-020-68811-8 (2020).

78 Verkuil, L. *et al.* SNP located in an AluJb repeat downstream of TMCO1, rs4657473, is protective for POAG in African Americans. *Br J Ophthalmol* **103**, 1530-1536, doi:10.1136/bjophthalmol-2018-313086 (2019).

79 Campbell-Thompson, M. *et al.* Network for Pancreatic Organ Donors with Diabetes (nPOD): developing a tissue biobank for type 1 diabetes. *Diabetes Metab Res Rev* **28**, 608-617, doi:10.1002/dmrr.2316 (2012).

80 Wang, Y. J. *et al.* Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* **65**, 3028-3038, doi:10.2337/db16-0405 (2016).

81 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

82 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).

83 Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* **16**, 983-986, doi:10.1038/s41592-019-0535-3 (2019).

84 Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344, doi:10.1016/j.cels.2016.08.011 (2016).

85 Schwartz, G. W. *et al.* TooManyCells identifies and visualizes relationships of single-cell clades. *Nat Methods* **17**, 405-413, doi:10.1038/s41592-020-0748-5 (2020).

86 Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40, doi:10.1186/s12859-019-2599-6 (2019).

87 Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N. & Lippincott, M. F. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am J Hum Genet* **103**, 522-534, doi:10.1016/j.ajhg.2018.08.016 (2018).

88 Ciesielski, T. H. *et al.* Diverse convergent evidence in the genetic analysis of complex disease: coordinating omic, informatic, and experimental evidence to better identify and validate risk factors. *BioData Min* **7**, 10, doi:10.1186/1756-0381-7-10 (2014).

89 Casals, T. *et al.* Bronchiectasis in adult patients: an expression of heterozygosity for CFTR gene mutations? *Clinical genetics* **65**, 490-495, doi:10.1111/j.0009-9163.2004.00265.x (2004).

90 Haufroid, V. & Hantson, P. CYP2D6 genetic polymorphisms and their relevance for poisoning due to amfetamines, opioid analgesics and antidepressants. *Clin Toxicol (Phila)* **53**, 501-510, doi:10.3109/15563650.2015.1049355 (2015).

91 Stoetzel, C. *et al.* BBS10 encodes a vertebrate-specific chaperonin-like protein and is a major BBS locus. *Nat Genet* **38**, 521-524, doi:10.1038/ng1771 (2006).

92 Elbedour, K., Zucker, N., Zalzstein, E., Barki, Y. & Carmi, R. Cardiac abnormalities in the Bardet-Biedl syndrome: echocardiographic studies of 22 patients. *Am J Med Genet* **52**, 164-169, doi:10.1002/ajmg.1320520208 (1994).

93 Ji, H. L. *et al.* delta ENaC: a novel divergent amiloride-inhibitable sodium channel. *Am J Physiol Lung Cell Mol Physiol* **303**, L1013-1026, doi:10.1152/ajplung.00206.2012 (2012).

94 Battaglia, A. Del 1p36 syndrome: a newly emerging clinical entity. *Brain Dev* **27**, 358-361, doi:10.1016/j.braindev.2004.03.011 (2005).

95 Gronich, N., Kumar, A., Zhang, Y., Efimov, I. R. & Soldatov, N. M. Molecular remodeling of ion channels, exchangers and pumps in atrial and ventricular myocytes in ischemic cardiomyopathy. *Channels (Austin)* **4**, 101-107, doi:10.4161/chan.4.2.10975 (2010).

96 Bowl, M. R. *et al.* A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nat Commun* **8**, 886, doi:10.1038/s41467-017-00595-4 (2017).

97 Ingham, N. J. *et al.* Mouse screen reveals multiple new genes underlying mouse and human hearing loss. *PLoS Biol* **17**, e3000194, doi:10.1371/journal.pbio.3000194 (2019).

98 Liu, H. *et al.* Characterization of transcriptomes of cochlear inner and outer hair cells. *J Neurosci* **34**, 11085-11095, doi:10.1523/JNEUROSCI.1690-14.2014 (2014).

99 Gilling, C. E. & Carlson, K. A. The effect of OTK18 upregulation in U937 cells on neuronal survival. *In Vitro Cell Dev Biol Anim* **45**, 243-251, doi:10.1007/s11626-009-9175-8 (2009).

100 Cacciottolo, M. *et al.* Muscular dystrophy with marked Dysferlin deficiency is consistently caused by primary dysferlin gene mutations. *Eur J Hum Genet* **19**, 974-980, doi:10.1038/ejhg.2011.70 (2011).

101 Bonventre, J. A. *et al.* Fer1l6 is essential for the development of vertebrate muscle tissue in zebrafish. *Mol Biol Cell* **30**, 293-301, doi:10.1091/mbc.E18-06-0401 (2019).

102 Burgess, R. W. *et al.* Evidence for a conserved function in synapse formation reveals Phr1 as a candidate gene for respiratory failure in newborn mice. *Mol Cell Biol* **24**, 1096-1105, doi:10.1128/mcb.24.3.1096-1105.2004 (2004).

103 Wan, H. I. *et al.* Highwire regulates synaptic growth in Drosophila. *Neuron* **26**, 313-329, doi:10.1016/s0896-6273(00)81166-6 (2000).

104 Zhen, M., Huang, X., Bamber, B. & Jin, Y. Regulation of presynaptic terminal organization by C. elegans RPM-1, a putative guanine nucleotide exchanger with a RING-H2 finger domain. *Neuron* **26**, 331-343, doi:10.1016/s0896-6273(00)81167-8 (2000).

105 Laizure, S. C., Herring, V., Hu, Z., Witbrodt, K. & Parker, R. B. The role of human carboxylesterases in drug metabolism: have we overlooked their importance? *Pharmacotherapy* **33**, 210-222, doi:10.1002/phar.1194 (2013).

106 Bergamaschi, D. *et al.* iASPP oncoprotein is a key inhibitor of p53 conserved from worm to human. *Nat Genet* **33**, 162-167, doi:10.1038/ng1070 (2003).

107 Howell, G. R. *et al.* Molecular clustering identifies complement and endothelin induction as early events in a mouse model of glaucoma. *The Journal of clinical investigation* **121**, 1429-1444, doi:10.1172/JCI44646 (2011).

108 Wilson, A. M. *et al.* Inhibitor of apoptosis-stimulating protein of p53 (iASPP) is required for neuronal survival after axonal injury. *PLoS One* **9**, e94175, doi:10.1371/journal.pone.0094175 (2014).

109 Nickells, R. W. Apoptosis of retinal ganglion cells in glaucoma: an update of the molecular pathways involved in cell death. *Surv Ophthalmol* **43 Suppl 1**, S151-161, doi:10.1016/s0039-6257(99)00029-6 (1999).

110 Snow, B. E. *et al.* GTPase activating specificity of RGS12 and binding specificity of an alternatively spliced PDZ (PSD-95/Dlg/ZO-1) domain. *J Biol Chem* **273**, 17749-17755, doi:10.1074/jbc.273.28.17749 (1998).

111 Cui, S. *et al.* The antagonist of CXCR1 and CXCR2 protects db/db mice from metabolic diseases through modulating inflammation. *Am J Physiol Endocrinol Metab* **317**, E1205-E1217, doi:10.1152/ajpendo.00117.2019 (2019).

112 Mori, M. *et al.* Transcriptional regulation of the cartilage intermediate layer protein (CILP) gene. *Biochem Biophys Res Commun* **341**, 121-127, doi:10.1016/j.bbrc.2005.12.159 (2006).

113 Zhang, C. L. *et al.* Cartilage intermediate layer protein-1 alleviates pressure overload-induced cardiac fibrosis via interfering TGF-beta1 signaling. *J Mol Cell Cardiol* **116**, 135-144, doi:10.1016/j.yjmcc.2018.02.006 (2018).

114 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).

115 Pinard, A., Jones, G. T. & Milewicz, D. M. Genetics of Thoracic and Abdominal Aortic Diseases. *Circ Res* **124**, 588-606, doi:10.1161/CIRCRESAHA.118.312436 (2019).

116 Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31, doi:10.1016/j.cell.2019.02.048 (2019).

117 Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026-1034, doi:10.1093/aje/kwx246 (2017).

118 Zhao, Z. *et al.* UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *Am J Hum Genet* **106**, 3-12, doi:10.1016/j.ajhg.2019.11.012 (2020).

119 Nassir, F., Rector, R. S., Hammoud, G. M. & Ibdah, J. A. Pathogenesis and Prevention of Hepatic Steatosis. *Gastroenterol Hepatol (N Y)* **11**, 167-175 (2015).

120 Byrne, C. D. & Targher, G. NAFLD: a multisystem disease. *J Hepatol* **62**, S47-64, doi:10.1016/j.jhep.2014.12.012 (2015).

121 Eslam, M. & George, J. Genetic contributions to NAFLD: leveraging shared genetics to uncover systems biology. *Nat Rev Gastroenterol Hepatol* **17**, 40-52, doi:10.1038/s41575-019-0212-0 (2020).

122 Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461-1465, doi:10.1038/ng.257 (2008).

123 Speliotes, E. K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* **7**, e1001324, doi:10.1371/journal.pgen.1001324 (2011).

124 Kozlitina, J. *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **46**, 352-356, doi:10.1038/ng.2901 (2014).

125 MacLean, M. T. *et al.* Linking abdominal imaging traits to electronic health record phenotypes. *medRxiv* (2020).

126 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

127 Tugues, S. *et al.* Microarray analysis of endothelial differentially expressed genes in liver of cirrhotic rats. *Gastroenterology* **129**, 1686-1695, doi:10.1053/j.gastro.2005.09.006 (2005).

128 Purushotham, A. *et al.* Hepatocyte-specific deletion of SIRT1 alters fatty acid metabolism and results in hepatic steatosis and inflammation. *Cell Metab* **9**, 327-338, doi:10.1016/j.cmet.2009.02.006 (2009).

129 Donati, B. *et al.* The rs2294918 E434K variant modulates patatin-like phospholipase domain-containing 3 expression and liver damage. *Hepatology* **63**, 787-798, doi:10.1002/hep.28370 (2016).

130 Ehrhardt, N. *et al.* Hepatic Tm6sf2 overexpression affects cellular ApoB-trafficking, plasma lipid levels, hepatic steatosis and atherosclerosis. *Hum Mol Genet* **26**, 2719-2731, doi:10.1093/hmg/ddx159 (2017).

131 Kitamoto, T. *et al.* Genome-wide scan revealed that polymorphisms in the PNPLA3, SAMM50, and PARVB genes are associated with development and progression of nonalcoholic fatty liver disease in Japan. *Human genetics* **132**, 783-792, doi:10.1007/s00439-013-1294-3 (2013).

132 Gorden, A. *et al.* Genetic variation at NCAN locus is associated with inflammation and fibrosis in non-alcoholic fatty liver disease in morbid obesity. *Hum Hered* **75**, 34-43, doi:10.1159/000346195 (2013).

133 Kleinstein, S. E. *et al.* Whole-Exome Sequencing Study of Extreme Phenotypes of NAFLD. *Hepatol Commun* **2**, 1021-1029, doi:10.1002/hep4.1227 (2018).

134 De Feo, E. *et al.* A case-control study on the effects of the apolipoprotein E genotypes in nonalcoholic fatty liver disease. *Mol Biol Rep* **39**, 7381-7388, doi:10.1007/s11033-012-1570-7 (2012).

135 Santoro, N. *et al.* Variant in the glucokinase regulatory protein (GCKR) gene is associated with fatty liver in obese children and adolescents. *Hepatology* **55**, 781-789, doi:10.1002/hep.24806 (2012).

136 Emdin, C. A. *et al.* A missense variant in Mitochondrial Amidoxime Reducing Component 1 gene and protection against liver disease. *PLoS Genet* **16**, e1008629, doi:10.1371/journal.pgen.1008629 (2020).

137 Graffy, P. M. & Pickhardt, P. J. Quantification of hepatic and visceral fat by CT and MR imaging: relevance to the obesity epidemic, metabolic syndrome and NAFLD. *Br J Radiol* **89**, 20151024, doi:10.1259/bjr.20151024 (2016).

138 Kramer, H. *et al.* Accuracy of Liver Fat Quantification With Advanced CT, MRI, and Ultrasound Techniques: Prospective Comparison With MR Spectroscopy. *AJR Am J Roentgenol* **208**, 92-100, doi:10.2214/AJR.16.16565 (2017).

139 Kurogane, Y. *et al.* FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. *Arterioscler Thromb Vasc Biol* **32**, 988-996, doi:10.1161/ATVBAHA.111.244004 (2012).

140 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015).

141 Yuan, H. *et al.* Hepatocyte-specific deletion of Cdc42 results in delayed liver regeneration after partial hepatectomy in mice. *Hepatology* **49**, 240-249, doi:10.1002/hep.22610 (2009).

142 Qu, X. *et al.* Cited2, a coactivator of HNF4alpha, is essential for liver development. *The EMBO journal* **26**, 4445-4456, doi:10.1038/sj.emboj.7601883 (2007).

143 BasuRay, S., Wang, Y., Smagris, E., Cohen, J. C. & Hobbs, H. H. Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. *Proc Natl Acad Sci U S A* **116**, 9521-9526, doi:10.1073/pnas.1901974116 (2019).

144 Schmidt, T. *et al.* ProteomicsDB. *Nucleic Acids Res* **46**, D1271-D1281, doi:10.1093/nar/gkx1029 (2018).

145 Li, P. *et al.* LTB4 promotes insulin resistance in obese mice by acting on macrophages, hepatocytes and myocytes. *Nat Med* **21**, 239-247, doi:10.1038/nm.3800 (2015).

146 Weerasekera, L. *et al.* ADAM19: A Novel Target for Metabolic Syndrome in Humans and Mice. *Mediators Inflamm* **2017**, 7281986, doi:10.1155/2017/7281986 (2017).

147 Chartrel, N. *et al.* The Neuropeptide 26RFa (QRFP) and Its Role in the Regulation of Energy Homeostasis: A Mini-Review. *Front Neurosci* **10**, 549, doi:10.3389/fnins.2016.00549 (2016).

148 Granata, R. *et al.* RFamide peptides 43RFa and 26RFa both promote survival of pancreatic beta-cells and human pancreatic islets but exert opposite effects on insulin secretion. *Diabetes* **63**, 2380-2393, doi:10.2337/db13-1522 (2014).

149 Prevost, G. *et al.* Neuropeptide 26RFa (QRFP) is a key regulator of glucose homeostasis and its activity is markedly altered in obese/hyperglycemic mice. *Am J Physiol Endocrinol Metab* **317**, E147-E157, doi:10.1152/ajpendo.00540.2018 (2019).

150 El-Mehdi, M. *et al.* Glucose homeostasis is impaired in mice deficient in the neuropeptide 26RFa (QRFP). *BMJ Open Diabetes Res Care* **8**, doi:10.1136/bmjdrc-2019-000942 (2020).

151 Xanthakos, S. A. *et al.* High Prevalence of Nonalcoholic Fatty Liver Disease in Adolescents Undergoing Bariatric Surgery. *Gastroenterology* **149**, 623-634 e628, doi:10.1053/j.gastro.2015.05.039 (2015).

152 Reeder, S. B. & Sirlin, C. B. Quantification of liver fat with magnetic resonance imaging. *Magn Reson Imaging Clin N Am* **18**, 337-357, ix, doi:10.1016/j.mric.2010.08.013 (2010).

153 Reeder, S. B., Hu, H. H. & Sirlin, C. B. Proton density fat-fraction: a standardized MR-based biomarker of tissue fat concentration. *J Magn Reson Imaging* **36**, 1011-1014, doi:10.1002/jmri.23741 (2012).

154 Jurgens, S. J. *et al.* Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. *bioRxiv* (2020).

155 Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683-692, doi:10.1038/s41586-020-2817-4 (2020).

156 Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235-239, doi:10.1038/nature22034 (2017).

157 Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323-328, doi:10.1038/s41586-019-1457-z (2019).

158 Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* **34**, 154-156, doi:10.1038/ng1161 (2003).

159 Chaudhary, R., Garg, J., Shah, N. & Sumner, A. PCSK9 inhibitors: A new era of lipid lowering therapy. *World J Cardiol* **9**, 76-91, doi:10.4330/wjc.v9.i2.76 (2017).

160 Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102-106, doi:10.1038/nature13917 (2015).

161 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710, doi:10.1038/s41586-019-1923-7 (2020).