LETTERS
https://doi.org/10.1038/s41591-020-1133-8

nature medicine

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

Joseph Park[1,2,3], Anastasia M. Lucas[1,3], Xinyuan Zhang[1,3], Kumardeep Chaudhary[4,5,6], Judy H. Cho[4,5,6], Girish Nadkarni[4,5,6], Amanda Dobbyn[4,5,6], Geetha Chittoor[7], Navya S. Josyula[7], Nathan Katz[2], Joseph H. Breeyear[8], Shadi Ahmadmehrabi[1], Theodore G. Drivas[2], Venkata R. M. Chavali[9], Maria Fasolino[1,10], Hisashi Sawada[11], Alan Daugherty[11,12], Yanming Li[13,14], Chen Zhang[13,14], Yuki Bradford[1,3], JoEllen Weaver[15], Anurag Verma[1,3], Renae L. Judy[16], Rachel L. Kember[1], John D. Overton[17], Jeffrey G. Reid[17], Manuel A. R. Ferreira[17], Alexander H. Li[17], Aris Baras[17], Scott A. LeMaire[13,14], Ying H. Shen[13,14], Ali Naji[16], Klaus H. Kaestner[1,10], Golnaz Vahedi[1,10], Todd L. Edwards[8], Jinbo Chen[18], Scott M. Damrauer[16], Anne E. Justice[7], Ron Do[4,5,6], Marylyn D. Ritchie[1,3] and Daniel J. Rader[1,2,15] ✉

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 (EHRs) could enhance our understanding of the contribution of rare genetic variation to human disease[1]. By leveraging 10,900 whole-exome sequences linked to EHR data in the Penn Medicine Biobank, we addressed the association of the cumulative effects of rare predicted loss-of-function variants for each 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 the Penn Medicine Biobank, as well as in three other medical biobanks and the population-based UK Biobank. 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 predicted loss-of-function 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.

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[2]. This approach can be applied to 'medical' biobanks consisting of health care 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[3]. Genome-first approaches utilizing PheWAS have primarily focused on individual common variants of modest effect[4]. 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[5]. Aggregation of multiple rare variants in a gene (that is, 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[6].

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

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Bio Phenomics Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Department of Population Health Sciences, Geisinger, Danville, PA, USA. [8]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [9]Scheie Eye Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [10]Penn Epigenetics Institute, University of Pennsylvania, Philadelphia, PA, USA. [11]Saha Cardiovascular Research Center, University of Kentucky, Lexington, KY, USA. [12]Department of Physiology, University of Kentucky, Lexington, KY, USA. [13]Division of Cardiothoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX, USA. [14]Department of Cardiovascular Surgery, Texas Heart Institute, Houston, TX, USA. [15]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [16]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [17]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA. [18]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: rader@pennmedicine.upenn.edu

Table 1 | Demographics and disease prevalence of the PMBB discovery cohort

Basic demographics	
Total population, *n*	10,900
Female, *n* (%)	4,432 (40.7)
Median age (at biobank entry), years	67.0
Genetically informed ancestry	
AFR, *n* (%)	2,172 (19.9)
AMR, *n* (%)	304 (2.8)
EAS, *n* (%)	79 (0.7)
EUR, *n* (%)	8,198 (75.2)
SAS, *n* (%)	114 (1.0)
Cardiovascular phenotypes	
Essential hypertension, *n* (%)	6,441 (59.1)
Ischemic heart disease, *n* (%)	5,008 (45.9)
Myocardial infarction, *n* (%)	1,640 (15.0)
Cardiomyopathy, *n* (%)	1,976 (18.1)
Congestive heart failure; nonhypertensive, *n* (%)	3,695 (33.9)
Heart transplant/surgery, *n* (%)	518 (4.8)
Cardiac dysrhythmias, *n* (%)	5,784 (53.1)
Atrial fibrillation and flutter, *n* (%)	3,782 (34.7)
Cerebrovascular disease, *n* (%)	1,706 (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* (%)	2,799 (25.7)
Overweight, obesity and other hyperalimentation, *n* (%)	2,275 (20.9)
Hyperlipidemia, *n* (%)	6,231 (57.2)
Hypercholesterolemia, *n* (%)	2,034 (18.7)
Hypothyroidism, *n* (%)	1,314 (12.1)
Gout and other crystal arthropathies, *n* (%)	811 (7.4)
Gastrointestinal phenotypes	
Esophagitis, GERD and related diseases, *n* (%)	2,526 (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, *n* (%)	2,135 (19.6)
End-stage renal disease, *n* (%)	510 (4.7)
Kidney replaced by transplant, *n* (%)	283 (2.6)
Neuropsychiatric phenotypes	
Mood disorders, *n* (%)	1,353 (12.4)
Anxiety, phobic and dissociative disorders, *n* (%)	1,322 (12.1)
Delirium, dementia and amnestic and other cognitive disorders, *n* (%)	123 (1.1)
Respiratory phenotypes	
Chronic airway obstruction, *n* (%)	1,314 (12.1)
Asthma, *n* (%)	920 (8.4)
Obstructive sleep apnea, *n* (%)	1,623 (14.9)
Respiratory failure, insufficiency and 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)

Demographic information and clinical phenotypic counts for all individuals with WES linked to EHRs in the PMBB. Clinical phenotypes were defined by phecodes (Methods). Data are represented as count data with percentage prevalence in the population in parentheses, where appropriate. AFR, Africa; AMR, the Americas; EAS, East Asia; EUR, Europe; SAS, South Asia; GERD, gastroesophageal reflux disease.

potential to uncover new pleiotropic relationships between the gene and human disease[7,8]. We applied rare predicted loss-of-function (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 (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.

We interrogated a dataset of 10,900 individuals with WES data in the PMBB (Table 1) for carriers of rare (minor allele frequency (MAF) $\leq 0.1\%$ in the Genome Aggregation Database (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 for each gene was on a negative exponential distribution (Extended Data Fig. 1). We chose to interrogate genes with at least 25 heterozygous carriers for rare pLOF variants ($n = 1,518$ genes), for which we show that statistical power to detect an association is sufficient as a function of effect size and the number of cases of the associated phenotype (Extended Data Fig. 2). We collapsed rare pLOF variants into gene burdens across these 1,518 genes for ExoPheWAS analyses with 1,000 binary phecodes with at least 20 cases (Fig. 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[9], we found that our associations roughly deviated from the fitted expected distribution at an observed $P < 10^{-6}$ (Extended Data Fig. 3). We identified 97 gene burdens with phenotype associations at $P < 10^{-6}$ (Fig. 2 and Supplementary Table 1). We addressed potential inflation issues regarding small sample sizes using Firth's penalized likelihood approach and found that beta and significance estimates were consistent with exact logistic regression (Supplementary Table 1).

We evaluated the robustness of the significant gene–phenotype associations identified using 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 the discovery cohort (Fig. 1). We utilized the rare exonic variant ensemble learner (REVEL), an ensemble method for predicting the pathogenicity of missense variants[10], to define predicted deleterious missense variants (REVEL score ≥ 0.5), given the success of the tool in identifying likely pathogenic variants for gene burden association studies[7]. First, we separately collapsed rare (MAF $\leq 0.1\%$), REVEL-informed predicted deleterious missense variants to test discovery-driven associations with their corresponding phenotypes (Supplementary Table 2). We also interrogated single variants, including both pLOF variants and predicted deleterious missense (REVEL ≥ 0.5) variants, in the 97 genes identified in the discovery cohort that were of sufficient frequency (MAF > 0.1%) and therefore were not included in either of the gene burden analyses (Supplementary Table 3).

We also endeavored to replicate our significant ExoPheWAS discovery analysis associations (Fig. 1) using a separate cohort of 6,432 African Americans in the PMBB who were exome sequenced (PMBB2; Supplementary Tables 4–6), as well as two additional medical biobanks with WES data linked to EHR phenotypes, namely BioMe (Mount Sinai; Supplementary Tables 7–9) and DiscovEHR (Geisinger Health System; Supplementary Tables 10–12), as well as the population-based UK Biobank (UKB; Supplementary Tables 13–15). 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 nonoverlapping rare (MAF ≤ 0.1%) REVEL-predicted deleterious missense variants and (3) single pLOF or REVEL-predicted deleterious missense variants with MAF > 0.1% for association with their discovery phenotypes. Finally, we further interrogated a targeted list of univariate replications in BioVU (Vanderbilt; Supplementary Table 16).

We identified a total of 26 robust genes using the diverse convergent evidence (DiCE) approach[11] for ranking associations using a combination of the number of significant replications and functional validation (Table 2 and Supplementary Table 17). Five of these genes can be considered positive control gene–disease associations. A gene burden of rare pLOF variants 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 variant 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 (Supplementary Table 18). 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 was replicated in all interrogated cohorts. While a predisposition to bronchiectasis due to haploinsufficiency of *CFTR* has been suggested[12], our finding strengthens this observation. *TTN* is a known dilated cardiomyopathy gene that was replicated convincingly across other cohorts. *MYBPC3* is a known hypertrophic cardiomyopathy (HCM) gene that was replicated in BioMe and DiscovEHR, but not in the UKB, where HCM had a case-control ratio of an order of magnitude lower than the medical biobanks (Supplementary Table 18). 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 the discovery cohort ($P = 1.29 \times 10{-}4$) but due to inadequate power did not meet our significance threshold. Finally, *CYP2D6* is a P450 enzyme known to metabolize opioids[13]; we found that *CYP2D6* was significantly associated with adverse effects of therapeutic opiate use.

We identified 20 robust genes with new disease associations that had at least two additional replications beyond the discovery experiment, and one strongly supported by the DiCE analysis (Table 2 and Supplementary Tables 2–17). Some have previous 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[14]. *BBS10* is expressed in the heart[15] and cardiac abnormalities have been reported in Bardet–Biedl syndrome, including hypertrophy of the interventricular septum[16], 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 the PMBB compared with non-carriers and found increased left ventricular outflow tract stroke volume, consistent with cardiac hypertrophy (Supplementary Table 19). 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*)[17], there is an association between 1p36 deletions (which contain *SCNN1D*) and congenital heart defects[18], and decreased expression of δENaC may contribute to disrupted sodium and potassium homeostasis in ischemic heart diseases[19]. The association between rare pLOF variants in *ZNF175* and tinnitus (additionally, hearing loss barely missed the 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 an abnormal Preyer reflex (auditory startle response)[20], as well as raised auditory brainstem response thresholds[21]. *Zfp719* is expressed in inner and outer hair cells of the mouse ear[22], and human *ZNF175* has a suggested role in neurotrophin production and neuronal survival[23].

Rare pLOF variants 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[24]. Importantly, loss of the zebrafish ortholog *Fer1l6* has been shown to lead to deformation of striated muscle and delayed cardiac development[25]. Similarly, pLOF variants in *MYCBP2*, an E3 ubiquitin-protein ligase critical in neuromuscular development in mice[26], *Drosophila*[27] and *Caenorhabditis elegans*[28], 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[26]. We found that *MYCBP2* showed significantly decreased expression in various lower extremity muscle tissues in tibial muscular dystrophy in humans (Extended Data Fig. 4). Our findings suggest that haploinsufficiency in *FER1L6* or *MYCBP2* increases the risk of developing dystrophic skeletal muscle.

Rare pLOF variants in *CES5A* were robustly associated with abnormal coagulation. Upon further investigation of EHR laboratory data in the PMBB, we found that carriers of rare pLOF variants in *CES5A* had increased international normalized ratios ($\beta = 8.2$, $P = 2.13 \times 10^{-2}$; $n = 5{,}275$) and partial thromboplastin times ($\beta = 13.9$, $P = 2.07 \times 10^{-2}$; $n = 3{,}786$) compared to non-carriers. Through chart review, we found an enrichment of gastrointestinal bleeding episodes following use of antiplatelet 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 antiplatelet medications aspirin and clopidogrel[29]. Given its predominant expression in the liver[15], it is thus plausible that haploinsufficiency of *CES5A* predisposes to adverse effects of antiplatelet medications.

Another finding was that rare pLOF variants in *PPP1R13L*, one of the most evolutionarily conserved inhibitors of p53 (ref. [30]), 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 (Supplementary Table 20). 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.

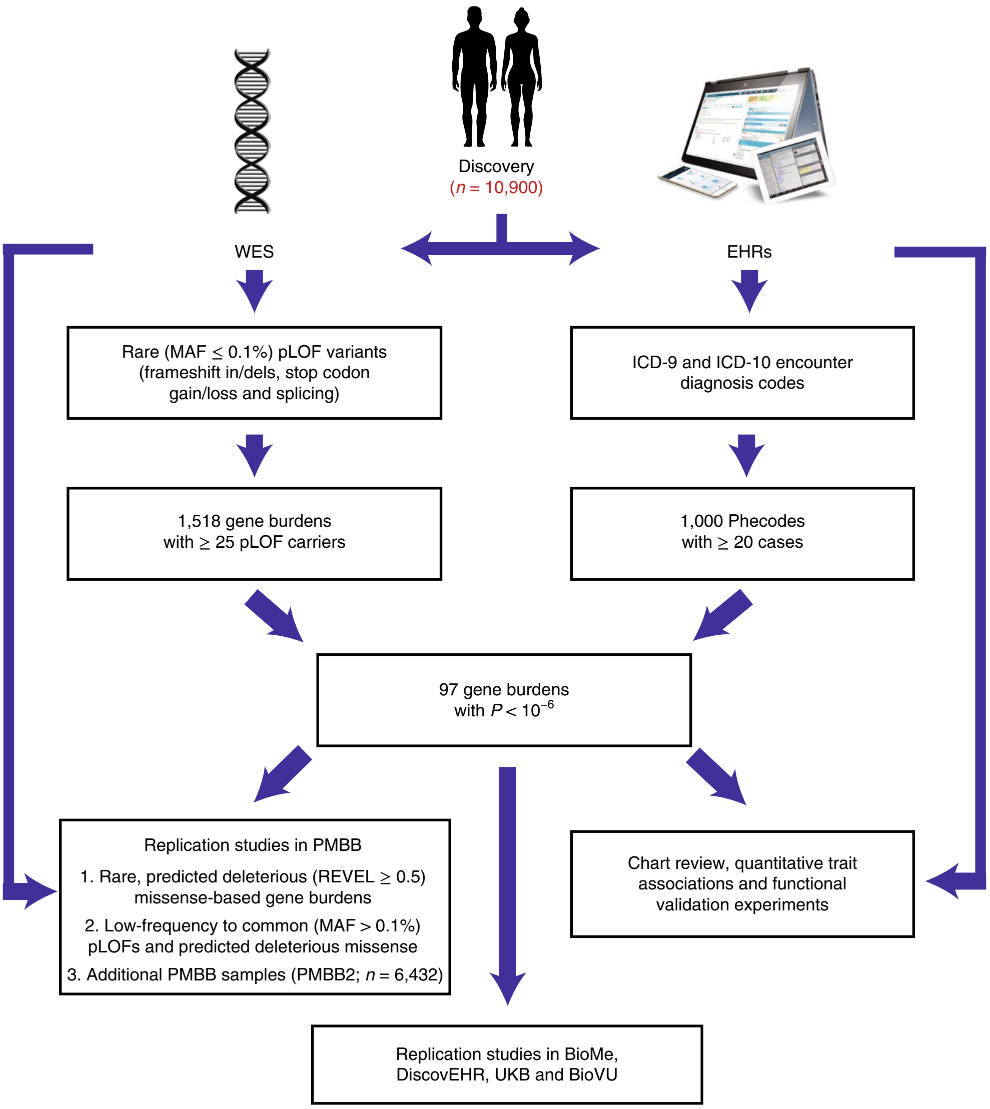

Discovery
(n = 10,900)
WES
EHRs
Rare (MAF ≤ 0.1%) pLOF variants (frameshift in/dels, stop codon gain/loss and splicing)
ICD-9 and ICD-10 encounter diagnosis codes
1,518 gene burdens with ≥ 25 pLOF carriers
1,000 Phecodes with ≥ 20 cases
97 gene burdens with P < 10−6
Replication studies in PMBB
1. Rare, 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, UKB and BioVU

Fig. 1 | Flow chart for exome-by-phenome-wide association analysis using electronic health record phenotypes. Flow diagram outlining the primary methodologies used for conducting the ExoPheWAS and for evaluation of the robustness of the associations, indicating that 97 genes had associations at a significance level of $P<10^{-6}$ using logistic regression. The pathways starting with short descending arrows represent the 'discovery phase', in which pLOF-based gene burdens were studied on an exome-by-phenome-wide scale in 10,900 individuals from the 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 UKB ($n=32{,}268$), as well as univariate analyses in BioVU ($n=66{,}400$). The DNA helix image is from Pixabay. The male and female silhouettes are from Freepik. The electronic health records image is from eClinicalWorks with permission.

We investigated whether *Ppp1r13l* is differentially expressed in the mouse ONH in glaucoma by comparing microarray gene expression datasets of the ONH[31]. We found *Ppp1r13l* expression was highest during late-early to moderate stages of glaucoma (Extended Data Fig. 5a). Additionally, inhibition of *PPP1R13L* has been shown to exacerbate RGC death following axonal injury[32]. We found that the PPP1R13L protein was 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 (Extended Data Fig. 5b). Using human induced pluripotent stem cell-derived RGCs (iPSC-RGCs), we found that oxidative stress markedly upregulated *PPP1R13L* expression (Extended Data Fig. 5c) 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[33]. Our results are consistent with the concept that haploinsufficiency of *PPP1R13L* in RGCs increases the visual consequences of primary open-angle glaucoma.

Another interesting finding was that rare pLOF variants in *RGS12* were associated with type 1 diabetes mellitus (T1D) and its complications. In the PMBB, carriers of rare pLOF variants in *RGS12* had higher median values for random serum glucose than non-carriers ($\beta=16.9$, $P=2.91\times10^{-2}$; $n=5{,}389$). *RGS12*, an inhibitor of signal transduction in G-protein signaling, contains an N-terminal PDZ

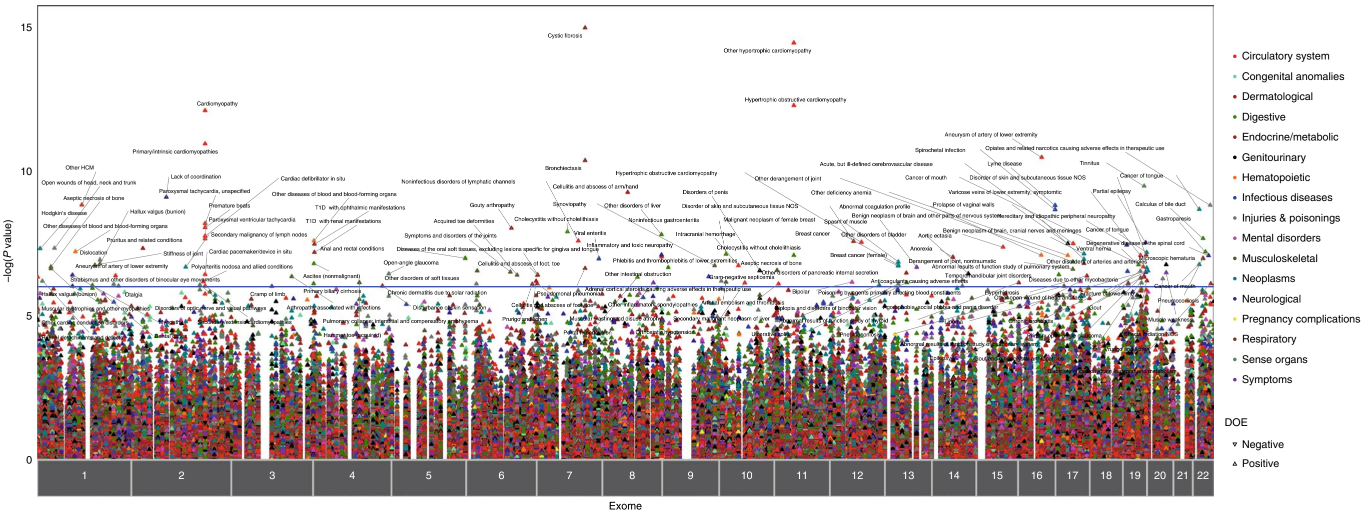


Fig. 2 | ExoPheWAS plot exhibits the landscape of gene–phenotype associations across the exome and phenome in the PMBB. Plot showing the results of the ExoPheWAS in the PMBB for 1,518 gene burdens of rare (MAF ≤ 0.1%) pLOF variants. The *x* axis represents the exome and is organized by chromosomal location. The location of each gene along the *x* axis corresponds to the genomic location for each gene according to the 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 $-\log_{10}(P \text{ value})$ 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 $P = 10^{-6}$ to account for multiple hypothesis testing. NOS, not otherwise specified.

domain that selectively binds to and represses the macrophage interleukin (IL)-8 receptor CXCR2 (ref. [34]). Activation of macrophage CXCR2 by IL-8 is proinflammatory, 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 signaling pathways in diabetes[35]. To further investigate *RGS12* in T1D, we generated single-cell RNA-sequencing (RNA-seq) data in human pancreatic islets from individuals with T1D and controls collected by the Human Pancreas Analysis Program (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 T1D compared to controls, there was a substantial reduction of expression of *RGS12* in peri-islet CD45$^+$ macrophages in T1D (Extended Data Fig. 6). 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 T1D.

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[36]. However, *CILP* is also expressed in the cardiovascular system[15], and has been shown to be involved in cardiac remodeling in response to pressure overload[37]. 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 (Extended Data Fig. 7a,b). Single-cell RNA-seq of human aorta confirmed that *CILP* was localized to aortic fibroblasts (Extended Data Fig. 7c,d). Importantly, *CILP* has been reported to modulate *TGFB1* signaling and *IGF1*-induced proliferation[38], and dysregulated transforming growth factor (TGF)-β signaling has been shown to contribute to the pathogenesis of thoracic aortic aneurysm formation[39]. 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-β 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-β ($\log_2$ fold change = 1.964, $P = 3.60 \times 10^{-29}$; Extended Data Fig. 7e), confirming its role in a functional feedback loop with TGF-β as similarly seen in the context of chondrocyte metabolism[36]. Furthermore, *CILP* was differentially coexpressed with *IGF1*, as well as genes implicated in aortic ectasia including *SMAD3*, *ACTA2*, *MYH11* and *ELN* (Extended Data Fig. 7e)[39]. 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-β signaling.

There has been a substantial 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[40]. In the PMBB discovery cohort, 19.9% of individuals were of African ancestry, 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 specific to African ancestry that replicated associations with the same disease in which a pLOF gene burden was associated in the discovery study (Supplementary Table 21). None of these rare variants exist in the genome-wide association study 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.

An important 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 the PMBB, as well as in two other medical biobanks with WES data. These provided more replication than the UKB, which is a population-based

Table 2 | List of robust exome-by-phenome-wide significant gene–phenotype associations

Gene	Phecode description	Discovery *P*	Replications (*n*)	Clinical/experimental evidence
Positive control associations				
BRCA2	Breast cancer	1.72×10^{-7}	4	✓
CFTR	Bronchiectasis	2.27×10^{-7}	10	✓
	Pseudomonal pneumonia	4.21×10^{-11}	5	✓
	Cystic fibrosis	1.05×10^{-15}	1	✓
CYP2D6	Opiates and related narcotics causing adverse effects in therapeutic use	1.50×10^{-9}	3	✓
MYBPC3	Hypertrophic cardiomyopathy	3.49×10^{-15}	5	✓
TTN	Cardiomyopathy	7.83×10^{-13}	10	✓
	Cardiac conduction disorders	6.45×10^{-9}	10	✓
	Cardiac dysrhythmias	1.77×10^{-8}	12	✓
New associations				
ABCA10	Benign neoplasm of brain, cranial nerves and meninges	7.26×10^{-8}	2	
	Abnormal results of function study of pulmonary system	1.54×10^{-7}	3	
BBS10	Hypertrophic cardiomyopathy	2.89×10^{-8}	1	✓
CES5A	Abnormal coagulation profile	8.10×10^{-8}	5	
CILP	Aortic ectasia	4.29×10^{-8}	3	✓
CTC1	Temporomandibular joint disorders	3.76×10^{-7}	3	
DNAH6	Lack of coordination	7.93×10^{-10}	2	
DNHD1	Aseptic necrosis of bone	2.67×10^{-7}	4	
EFCAB5	Prolapse of vaginal walls	3.19×10^{-8}	3	
EPPK1	Phlebitis and thrombophlebitis of lower extremities	9.19×10^{-8}	3	
FER1L6	Muscular wasting and disuse atrophy	7.18×10^{-7}	3	✓
FLG2	Stiffness of joint	1.76×10^{-7}	2	
MYCBP2	Spasm of muscle	2.08×10^{-7}	2	✓
PPP1R13L	Primary open-angle glaucoma	7.29×10^{-7}	2	✓
RGS12	Type 1 diabetes	6.48×10^{-8}	5	✓
RTKN2	Orthostatic hypotension	7.24×10^{-7}	5	
SCNN1D	Cardiac conduction disorders	4.52×10^{-7}	5	
TGM6	Lipoma	2.77×10^{-7}	4	
TRDN	Acquired toe deformities	3.90×10^{-7}	3	
WDR87	Ventral hernia	1.70×10^{-7}	4	
ZNF175	Tinnitus	3.24×10^{-10}	3	✓
ZNF334	Microscopic hematuria	1.69×10^{-7}	3	

List of genes among 97 pLOF-based gene burdens with phenotype associations at $P<10^{-6}$ in the PMBB discovery cohort that were most robust according to the 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 the PMBB (the PMBB2 cohort), as well as in BioMe, DiscovEHR and the UKB. Targeted single variants that showed successful replication in the 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 the PMBB, as well as the number of total replications and existence of clinical/experimental evidence, fully detailed in Supplementary Table 17. Only associations with at least two total check marks in Supplementary Table 17, where each successful mode of replication in a particular biobank (for example, pLOF burden in BioMe) or the existence of clinical/experimental evidence is labeled with a checkmark, were deemed robust and therefore included here. Previously known associations were considered to represent positive controls. Positive control and new associations are each ranked alphabetically by gene name.

biobank that is widely recognized to have a ‘healthy volunteer selection bias’ (ref. [41]) and has lower prevalence of the specific diseases than the medical biobanks (Supplementary Table 18). This may be one factor explaining the relative lack of new findings in gene burden studies using the UKB for discovery[42,43]. 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 new 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 the discovery of new gene–disease relationships. Our compelling findings based on initial discovery in <11,000 whole-exome sequences 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-1133-8.

Received: 6 May 2020; Accepted: 13 October 2020;
Published online: 11 January 2021

References

1. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, https://doi.org/10.1126/science.aaf6814 (2016).
2. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).
3. 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 (2016).
4. Verma, A. et al. Human-disease phenotype map derived from PheWAS across 38,682 individuals. *Am. J. Hum. Genet.* **104**, 55–64 (2019).
5. 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 (2019).
6. 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 (2014).
7. Park, J. et al. A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes. *Genet. Med.* https://doi.org/10.1038/s41436-019-0625-8 (2019).
8. Haggerty, C. M. et al. Genomics-first evaluation of heart disease associated with titin-truncating variants. *Circulation* **140**, 42–54 (2019).
9. 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 (2018).
10. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
11. 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 (2014).
12. Casals, T. et al. Bronchiectasis in adult patients: an expression of heterozygosity for *CFTR* gene mutations? *Clin. Genet.* **65**, 490–495 (2004).
13. Haufroid, V. & Hantson, P. CYP2D6 genetic polymorphisms and their relevance for poisoning due to amfetamines, opioid analgesics and antidepressants. *Clin. Toxicol.* **53**, 501–510 (2015).
14. Stoetzel, C. et al. *BBS10* encodes a vertebrate-specific chaperonin-like protein and is a major BBS locus. *Nat. Genet.* **38**, 521–524 (2006).
15. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
16. 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 (1994).
17. Ji, H. L. et al. δENaC: a novel divergent amiloride-inhibitable sodium channel. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **303**, L1013–L1026 (2012).
18. Battaglia, A. Del 1p36 syndrome: a newly emerging clinical entity. *Brain Dev.* **27**, 358–361 (2005).
19. 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* **4**, 101–107 (2010).
20. Bowl, M. R. et al. A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nat. Commun.* **8**, 886 (2017).
21. Ingham, N. J. et al. Mouse screen reveals multiple new genes underlying mouse and human hearing loss. *PLoS Biol.* **17**, e3000194 (2019).
22. Liu, H. et al. Characterization of transcriptomes of cochlear inner and outer hair cells. *J. Neurosci.* **34**, 11085–11095 (2014).
23. 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 (2009).
24. 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 (2011).
25. Bonventre, J. A. et al. *Fer1l6* is essential for the development of vertebrate muscle tissue in zebrafish. *Mol. Biol. Cell* **30**, 293–301 (2019).
26. 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 (2004).
27. Wan, H. I. et al. Highwire regulates synaptic growth in *Drosophila*. *Neuron* **26**, 313–329 (2000).
28. 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 (2000).
29. 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 (2013).
30. Bergamaschi, D. et al. iASPP oncoprotein is a key inhibitor of p53 conserved from worm to human. *Nat. Genet.* **33**, 162–167 (2003).
31. Howell, G. R. et al. Molecular clustering identifies complement and endothelin induction as early events in a mouse model of glaucoma. *J. Clin. Invest.* **121**, 1429–1444 (2011).
32. 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 (2014).
33. Nickells, R. W. Apoptosis of retinal ganglion cells in glaucoma: an update of the molecular pathways involved in cell death. *Surv. Ophthalmol.* **43**, S151–S161 (1999).
34. 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 (1998).
35. 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 (2019).
36. Mori, M. et al. Transcriptional regulation of the cartilage intermediate layer protein (CILP) gene. *Biochem. Biophys. Res. Commun.* **341**, 121–127 (2006).
37. Zhang, C. L. et al. Cartilage intermediate layer protein-1 alleviates pressure overload-induced cardiac fibrosis via interfering TGF-β1 signaling. *J. Mol. Cell. Cardiol.* **116**, 135–144 (2018).
38. 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 (2019).
39. Pinard, A., Jones, G. T. & Milewicz, D. M. Genetics of thoracic and abdominal aortic diseases. *Circ. Res.* **124**, 588–606 (2019).
40. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
41. 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 (2017).
42. 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 (2020).
43. 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 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Setting and study participants. All individuals who were recruited for the 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 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 the PMBB, replication analyses were conducted using the WES dataset from an additional set of independent individuals of African American ancestry in the PMBB (PMBB2), BioMe, DiscovEHR, UKB, as well as imputed genotype data in BioVU, for evaluation of the robustness of gene–phenotype associations identified in the PMBB. For replication analyses in BioMe, DiscovEHR and BioVU, each study was approved by the institutional review board in the institution of each respective biobank. Access to the UKB for this project was from application 32133.

Genetic sequencing. This PMBB study dataset included a subset of 11,451 individuals in the PMBB who have undergone WES. For each individual, we extracted DNA from stored buffy coats and then obtained exome sequences generated by the Regeneron Genetics Center. These sequences were mapped to GRCh37 as previously described[7]. Furthermore, for subsequent phenotypic analyses, we removed samples with low exome sequencing coverage (less than 75% of targeted bases achieving 20× coverage), high missingness (greater than 5% of targeted bases), high heterozygosity, dissimilar reported and genetically determined sex, genetic evidence of sample duplication, and cryptic relatedness (closer than third-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 the 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 third-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 third-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 who were exome sequenced on the IDT platform and a separate set of 59,133 individuals of European ancestry who were 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 third-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 the UKB, we interrogated 34,629 individuals of European ancestry (based on reported genetic ancestry grouping from the UKB) with diagnosis codes according to the tenth revision of the International Classification of Diseases (ICD-10) available among the 49,960 individuals who had WES data as generated by the functional equivalence pipeline. We focused our replication efforts on 32,268 individuals after removing samples with poor genotype quality, individuals closer than third-degree relatives and those with dissimilar reported and genetically determined sex. The PLINK files for exome sequencing provided by the 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 the PMBB, PMBB2 and/or UKB. We interrogated these variants for an 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 third-degree relatives and those with dissimilar reported and genetically determined sex. These sequences were mapped to GRCh37.

Variant annotation and selection for association testing. For all cohorts analyzed, genetic variants were annotated using ANNOVAR (version 2018Apr16)[44] as pLOF or missense variants according to the NCBI Reference Sequence database. These pLOF variants were defined as frameshift insertions/deletions, gain/loss of stop codon or disruption of canonical splice site dinucleotides. Predicted deleterious missense variants were defined as those with REVEL[10] scores ≥ 0.5. MAF for each variant was determined per non-Finnish European, African and Latino minor allele frequencies reported by gnomAD (v2)[45]. pLOF and REVEL-informed missense variants were selected for gene burden testing or univariate association analyses for every ancestry group in each cohort according to the corresponding ancestry-specific MAF thresholds for each ancestry (rare variants with MAF $\leq 0.1\%$ for gene burden testing; single variants with MAF $> 0.1\%$).

Clinical data collection. ICD-9 and ICD-10 disease diagnosis codes and procedural billing codes, medications and clinical imaging and laboratory measurements were extracted from patient EHRs for the PMBB. ICD-10 encounter diagnoses were mapped to ICD-9 using 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 (phecodes) with Phecode Map 1.2 using the R package 'PheWAS' [46]. Participants were determined as having 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 20 March 2019; all units were converted to their respective clinical traditional units. The 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 01 January 2010 until 09 September 2016; outliers for each echocardiographic parameter (less than Q1 − 1.5 × interquartile range (IQR) or greater than Q3 + 1.5 × IQR) were removed. Similarly, the minimum, median and maximum values for each parameter were recorded for each participant 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 the 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.

Association studies. A PheWAS approach was used to determine the phenotypes associated with rare (MAF $\leq 0.1\%$ in gnomAD) pLOF variants carried by individuals in the PMBB for the discovery experiment[47]. Each disease phenotype was tested for an association with each gene burden or single variant using a logistic regression model adjusted for age, age squared, sex and the first ten principal components (PCs) of genetic ancestry. We used an additive genetic model to collapse variants for each gene using the fixed threshold approach[48]. 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 pLOF variants 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). Power analyses were conducted using QUANTO (version 1.2.4)[49].

We further evaluated the robustness of our gene–phenotype associations in the same PMBB discovery cohort by (1) associating the aggregation of rare (MAF $\leq 0.1\%$) predicted deleterious missense variants in gene burden association tests and (2) testing pLOF variants and predicted deleterious missense variants with MAF > 0.1 in univariate association tests. We ensured that individuals were nonoverlapping across rare pLOF variants and rare deleterious missense and single variant groups. Rare deleterious missense gene burdens and single variants were analyzed for an association with the specific phenotype identified in the pLOF-based gene burden discovery, together 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 (for example, 123.6). Notably, we checked for the presence of mutual carriers between pLOF-based gene burdens for each gene 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 associations among phenotypes were not being driven by mutual carriers. All association studies in the PMBB were based on a logistic regression model adjusted for age, age squared, sex and the first ten 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 the discovery phase in the PMBB. A specific set of single variants was further replicated in BioVU. Association statistics were calculated similarly to the PMBB, such that each disease phenotype was tested for an association with each gene burden or single variant using a logistic regression model adjusted for age, age squared, sex and the first ten 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 analyzed in the meta-analysis. In DiscovEHR, the summary statistics obtained from running the logistic regression model separately in individuals of European ancestry on the IDT compared with VCRome platforms were analyzed in the meta-analysis. In BioVU, the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry were analyzed in the meta-analysis. All association analyses for the PMBB, PMBB2, BioMe, DiscovEHR, UKB and BioVU were completed using R (version 3.3.1 or later).

Undercalling of variants in the UK Biobank. Given the undercalling of variants largely limited to ~3.25% of the exome target regions in the functional equivalence pipeline data, we found that 3 of the 97 genes that had associations with $P<10^{-6}$ from the discovery phase overlapped 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 a call rate of at least 65% for these three genes, with the understanding that undercalling per variant is random for each individual.

Statistical analyses of clinical measurements. To compare available measurements for echocardiographic parameters and serum laboratory values between carriers of predicted deleterious variants and genotypic controls in the PMBB, we utilized linear regression adjusted for age, age squared, sex and the first ten PCs of genetic ancestry in individuals of European ancestry only. These analyses were conducted with the minimum, median and maximum values 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).

Chart review to validate robust gene–phenotype associations. To confirm our curated list of robust exome-by-phenome-wide significant associations, we performed manual chart review of EHR data 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 $P>10^{-6}$. 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 grouped 'other' diagnoses that did not fit into disease-specific ICD codes (for example, 'other diseases of blood and blood-forming organs').

We discovered, on chart review, that individuals with the phecode 'hypertrophic obstructive cardiomyopathy' or 'other hypertrophic cardiomyopathy' in the PMBB were participants with HCM who were being assigned one of the codes due to the lack of a single ICD diagnosis code for HCM. Thus, we defined a new phenotype for HCM encompassing cases for either phecode, and we repeated the association with the pLOF gene burdens of *MYBPC3* (positive control) and *BBS10* (new), confirming their associations as exome-by-phenome-wide significant (Supplementary Table 22).

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/)[50]. To investigate the new association between *CILP* and aortic ectasia, we interrogated 11 different microarray and RNA-seq datasets of human fibroblasts from various tissues treated with TGF-β (GSE1724, GSE65069, GSE64192, GSE39394, GSE79621, GSE68164, GSE97833, GSE97823, GSE135065, GSE125519 and GSE40266). Differential expression for each dataset was interrogated using the GEO2R software using a moderated *t* statistic. Meta-analysis of differential expression across the datasets was achieved using the Fisher's combined probability test and visualized 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[51].

We also analyzed microarray data from muscle biopsies in participants with tibial muscular dystrophy compared to controls (GSE42806) to validate the new association between *MYCBP2* and muscle spasms. Differential expression was interrogated using the GEO2R software via a moderated *t* statistic.

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 OTDB (https://genome.uiowa.edu/otdb/)[52]. 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 values, where individual gene expression values are normalized with its expression in other tissues.

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 provided an interactive way to analyze RNA-seq data obtained from RGCs isolated from the retina and ONH of a 9-month-old female D2 mouse, which is an age-dependent model of ocular hypertension/glaucoma, and an D2-*Gpnmb*$^+$ mouse, which does not develop high IOP/glaucoma[53]. For transcriptomic studies, four distinct groups were compared based on axonal degeneration and gene expression patterns. The transcriptome of D2 group 1 was identical to the control strain (D2-*Gpnmb*$^+$), while D2 groups 2–4 exhibited increasing levels of molecular changes relevant to axonal degeneration when compared to the control group. We used the Datgan software to assess the differential expression of *Ppp1r13l* in the retina[54].

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 sections ($n=3$) obtained from normal cadaver eyes of a 68-year-old donor with a commercially available antibody, anti-PPP1R13L (51141-1-AP, Proteintech). Antigen retrieval was performed in 1× citrate buffer (Life Technologies) warmed to 95 °C for 30 min. Sections were allowed to cool to room temperature and subsequently blocked in 10% normal goat serum with 1% BSA in 1× TBS buffer for 1 h. The retinal distribution of PPP1R13L protein was visualized by incubating the retinal section with rabbit polyclonal anti-PPP1R13L antibody at a 1:300 dilution overnight at 4 °C, followed by chicken anti-rabbit IgG conjugated with Alexa Fluor 594 (A21442, Life Technologies) at a 1:3,000 dilution. Nuclei were stained with the use of Vectashield DAPI in the mounting medium. The images were captured using a Zeiss Imager Z1 fluorescence microscope equipped with AxioVS40 software (version 4.8.1.0).

Human induced pluripotent stem cell-derived retinal ganglion cell 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 human embryonic stem/iPS cell pluripotency PCR with reverse transcription (RT–PCR) kit[55]. The iPSC-RGCs for our studies were derived using small molecules to inhibit bone morphogenetic protein, TGF-β (SMAD) and Wnt signaling to differentiate RGCs from iPSCs. The iPSCs were differentiated into pure iPSC-RGCs with structural and functional features characteristic of native RGCs based on a previous protocol[56].

Evaluating oxidative stress in 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 medium. The cells were collected 24 h 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[57]. Mean expression levels (± s.e.m.) were calculated by analyzing at least three independent samples with replica reactions and were 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 to an untreated control to obtain normalized gene expression. A two-tailed unpaired Student's *t*-test was used for statistical analysis.

Single-cell RNA sequencing of human pancreatic islets in type 1 diabetes and controls. Pancreatic islets were procured from the Human Pancreas Analysis Program consortium under the Human Islet Research Network (https://hirnetwork.org/) with approval from the University of Florida Institutional Review Board (201600029) and the United Network for Organ Sharing. A legal representative for each donor provided informed consent before organ retrieval. For 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. The individuals with T1D (50% female) had a median age of 29.5 years and median body mass index (BMI) of 21.25. The control individuals (60% female) had a median age of 13 years and median BMI of 17.3. All individuals were of European ancestry. Organs were recovered and processed as previously described[58]. Pancreatic islets were cultured and dissociated into single cells as previously described[59]. 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 single-cell RNA-seq data. About 3,000 cells were targeted for recovery for each 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 HiSeq 4000: 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 (https://satijalab.org/seurat/)[60] was used for filtering, uniform manifold approximation and projection (UMAP) generation and initial clustering. Genes were kept that were in 0.01% of cells (three 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 did not eliminate any of the 35,134 cells. nFeature, nCount, percent.mt, nFeature versus nCount and percent.mt versus nCount plots were generated to ascertain the lenient filtering criteria of $200 > \text{nFeature} < 7{,}500$, $\text{percent.mt} < 30$ and $\text{nCount} < 100{,}000$, which led to the filtering out of 66 cells (35,066 cells remaining). Data were then log normalized, and the top 2,000 variable genes were detected using the 'vst' selection method. The data were then linearly transformed, and principal-component analysis (PCA) was carried out on the scaled data, using the 2,000 variable genes as input. To determine the dimensionality of the data (that is, how many PCs to choose when clustering), we used 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 PC. Based on these two approaches, 14 PCs with a resolution of 2 were used to cluster the cells, and nonlinear dimensionality reduction (UMAP) was used with 14 PCs to visualize the dataset.

DoubletFinder (v2.0)[61] was used to demarcate and remove potential doublets in the data as previously described, with the following details: paramSweep_v3, doubletFinder_v3, and 14 PCs were used to determine the neighborhood size (pK; no ground-truth). The following parameters were used when running doubletFinder_v3: PCs = 14, pN = 0.25, pK = 0.005, nExp = nEx_poi.adj and sct = FALSE. The doublets had a 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 the standard deviation explained by each PC 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 were 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[62]. In brief, a cell-type marker file with 17 different cell types was compiled using various resources[59,60,63], 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 using '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[64]. Briefly, the raw data from the 34,259 cells were not filtered and were normalized by total count and gene normalized 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 (v3.24.3) 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 a false discovery rate control to determine differentially expressed genes[65].

Single-cell RNA sequencing 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 were housed in standard ventilated cages that each had a floor area of 65 square inches and contained a maximum of four mice. Room temperatures were maintained at 70 ± 2 °F. Normal humidity for animal holding rooms ranged from 30% to 70%. The standard light timer was set on a 14-h light cycle with the lights coming on at 06:00 and off at 20:00.

Ascending aortic samples were harvested from Mef2c-Cre;Rosa26R$^{\text{mT/mG}}$ male mice ($n = 5$) and were pooled in Hanks' Balanced Salt Solution (HBSS; 14175095, Thermo Fisher Scientific) with 10% FBS. 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^{-1} collagenase type II (LS004176, Worthington); 0.15 mg ml^{-1} collagenase type XI (C7657, Sigma-Aldrich); 0.24 mg ml^{-1} hyaluronidase type I (H3506, Sigma-Aldrich); 0.1875 mg ml^{-1} elastase (LS002290, Worthington); 2.38 mg ml^{-1} HEPES (H4034, Sigma-Aldrich)) in HBSS with Ca^{2+}/Mg^{2+} (14025092, Thermo Fisher Scientific) for 60 min at 37 °C. The cell suspension was filtered through a 40-µm cell strainer (CLS431750-50EA, Sigma-Aldrich), centrifuged at 300*g* for 10 min and resuspended using cold HBSS (14175095) with 5% FBS. Cells were stained with DAPI 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 paired-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 5,000 genes were filtered out for exclusion of non-cell or cell aggregates. Cells with more than 10% mitochondrial genes were also excluded. Data were then normalized and processed into scaled data. PCA and nonlinear dimensional reduction using *t*-distributed stochastic neighbor embedding were performed to create clusters and for data visualization. The 'FindAllMarkers' function in Seurat was used to identify the conserved marker genes in each cluster.

Single-cell RNA-seq of human aorta. The protocol for collecting human aortic tissue samples for the 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 three controls (two females and one male, heart transplant recipient or lung transplant donor) and eight individuals with an ascending thoracic aortic aneurysm (four females and four males). For each sample, a piece of aortic tissue (1–2 cm^2) was torn into thin layers and cut into small pieces in HBSS (without Ca^{2+} and Mg^{2+}; Gibco) with 10% FBS. Small pieces of tissue were then moved to an enzyme cocktail prepared with 3 mg ml^{-1} collagenase type II (LS004176, Worthington Biochemical), 0.15 mg ml^{-1} collagenase type XI (H3506, Sigma), 0.25 mg ml^{-1} soybean trypsin inhibitor (LS003571, Worthington), 0.1875 mg ml^{-1} elastase lyophilized (LS002292, Worthington), 0.24 mg ml^{-1} hyaluronidase type I (H3506, Sigma) and 2.38 mg ml^{-1} HEPES (H4034, Sigma) in HBSS (with Ca^{2+} and Mg^{2+}) (14025092, Thermo Fisher Scientific) and were digested in a 37 °C water bath for 1–2 h. Tissue dissociation was examined under a microscope. Cell suspensions were collected by using a 40-µm cell strainer (CLS431750-50EA, Corning), centrifuged at 300*g* for 10 min and resuspended in HBSS (without Ca^{2+} and Mg^{2+}; 14175095, Thermo Fisher) with 5% FBS, followed with incubation on ice for 30 min. 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) staining.

Single-cell suspensions were submitted to the 10x Genomics Chromium System (10x Genomics), followed by the construction of 3′ gene expression v3 libraries and sequencing on an Illumina NovaSeq 6000. Raw FASTQ data alignment was processed using Cell Ranger 3.0, with GRCh38 as a reference. Mapped UMI counts were loaded into R for further analysis. The single-cell sequencing data were filtered using Seurat 3.0 with the following criteria: gene count per cell of >200 and <4,000 (or 5,000), 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 scaling. The conserved (marker) genes for each cluster were identified using the function 'FindAllMarkers' in Seurat. For reclustering, the UMI counts of cells of interest were extracted and analyzed similarly to clusters identified in multiple combined datasets.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All summary statistics for significant gene–phenotype associations from the discovery phase in the PMBB, as well as significant replications from each replication cohort are fully detailed in Supplementary Tables 1–16. Data for the individual rare pLOF and missense variants in significant genes that were used for gene burden analyses in the PMBB discovery cohort are also included in Supplementary Tables 23 and 24. In addition, a list of all of the single variants that were used for replication analyses across all the cohorts are provided in Supplementary Table 25. Each variant in Supplementary Tables 23–25 is annotated with information regarding genomic location, variant effect, amino acid change, REVEL score (for missense) and MAF in gnomAD, as well as in the PMBB discovery cohort. Additionally, up-to-date summary data for genetic variants captured using WES in the PMBB can be accessed via the PMBB Genome Browser (https://pmbb.med.upenn.edu/allele-frequency/). Individual-level data are not publicly available due to research participant privacy concerns; however, requests from accredited researchers for access to individual-level data relevant to this manuscript can be made by contacting the corresponding author. Additionally, public expression datasets were obtained from the OTDB (https://genome.uiowa.edu/otdb/), Glaucoma Discovery Platform (http://glaucomadb.jax.org/glaucoma/), and NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/). From NCBI GEO, we interrogated 11 different microarray and RNA-seq datasets of human fibroblasts from various tissues treated with TGF-β (GSE1724, GSE65069, GSE64192, GSE39394, GSE79621, GSE68164, GSE97833, GSE97823, GSE135065, GSE125519 and GSE40266), as well as microarray data from muscle biopsies in participants with tibial muscular dystrophy (GSE42806).

References

44. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
45. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
46. 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 (2014).
47. 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 (2013).
48. Price, A. L. et al. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
49. Gauderman, W. J., Morrison, J. M. & Morrison, W. G. J. QUANTO 1.1: a computer program for power and sample size calculations for genetic-epidemiology studies. http://hydra.usc.edu/gxe (2006).
50. 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).
51. 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 (2019).
52. Wagner, A. H. et al. Exon-level expression profiling of ocular tissues. *Exp. Eye Res.* **111**, 105–111 (2013).
53. Libby, R. T. et al. Inherited glaucoma in DBA/2J mice: pertinent disease features for studying the neurodegeneration. *Vis. Neurosci.* **22**, 637–648 (2005).
54. 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 (2011).
55. Yang, W. et al. Generation of iPSCs as a pooled culture using magnetic activated cell sorting of newly reprogrammed cells. *PLoS ONE* **10**, e0134995 (2015).
56. 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 (2020).
57. 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 (2019).
58. 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 (2012).
59. Wang, Y. J. et al. Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* **65**, 3028–3038 (2016).
60. 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 (2018).
61. 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 (2019).
62. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
63. 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 (2016).
64. Schwartz, G. W. et al. TooManyCells identifies and visualizes relationships of single-cell clades. *Nat. Methods* **17**, 405–413 (2020).
65. 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 (2019).

Acknowledgements

The PMBB is funded by the Perelman School of Medicine at the University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA Award Number UL1TR001878. We thank D. Birtwell, H. Williams, P. Baumann and M. Risman for informatics support regarding the PMBB. We thank the staff of the Regeneron Genetics Center for whole-exome sequencing of DNA from PMBB participants. We thank S. Rathi for help with the real-time PCR experiment on iPSC-RGCs and J. He for help with iPSC-RGC cultures. We thank S. Dudek for assistance with the PMBB Genome Browser. Research reported in this paper was supported by grants from the National Human Genome Research Institute of the National Institutes of Health under award number F30HG010442 (to J.P.); the National Eye Institute of the National Institutes of Health under award number R21EY028273-01A1, BrightFocus Foundation, Lisa Dean Moseley Foundation and Research to Prevent Blindness, F.M. Kirby Foundation and The Paul and Evanina Bell Mackall Foundation Trust (to V.R.M.C); American Heart Association SFRN in Vascular Disease under award numbers 18SFRN33960114 and 18SFRN33960163 (to S.A.L. and A.D.) and National Institutes of Health under award number 1R01HL143359 (to Y.H.S. and S.A.L.); Sarnoff Cardiovascular Research Foundation (to S.A.); Institute for Translational Medicine and Therapeutics Transdisciplinary Program in Translational Medicine and Therapeutics (to R.L.K.)

Author contributions

The study was conceived and designed by J.P., M.D.R and D.J.R. Association analyses for the study were conducted by J.P., A.M.L., X.Z., K.C., J.H.C., G.N., A.D., G.C., N.S.J., N.K., J.H.B, M.A.R.F., A.H.L. and A.E.J. Association data interpretation was performed by J.P., A.M.L., N.K., Y.B., A.V., J.C., S.M.D., T.L.E., A.E.J., R.D., M.D.R. and D.J.R. Chart review was conducted by J.P., S.A. and T.G.D. Experiments specific to *PPP1R13L* and glaucoma were performed by V.R.M.C. Single-cell RNA-seq of human pancreas was performed by M.F., A.N., K.H.K. and G.V. Single-cell RNA-seq of mouse and human aortae was performed by H.S., A.D., Y.L., C.Z., S.A.L. and Y.H.S. Data acquisition for association analyses was performed by J.P., X.Z., J.W., A.V., R.L.J., R.L.K., J.D.O., J.G.R., A.B., S.M.D., A.E.J., R.D., M.D.R. and D.J.R. The manuscript was written by J.P., M.D.R. and D.J.R., and revised by all authors.

Competing interests

The authors declare no competing interests.

Additional information

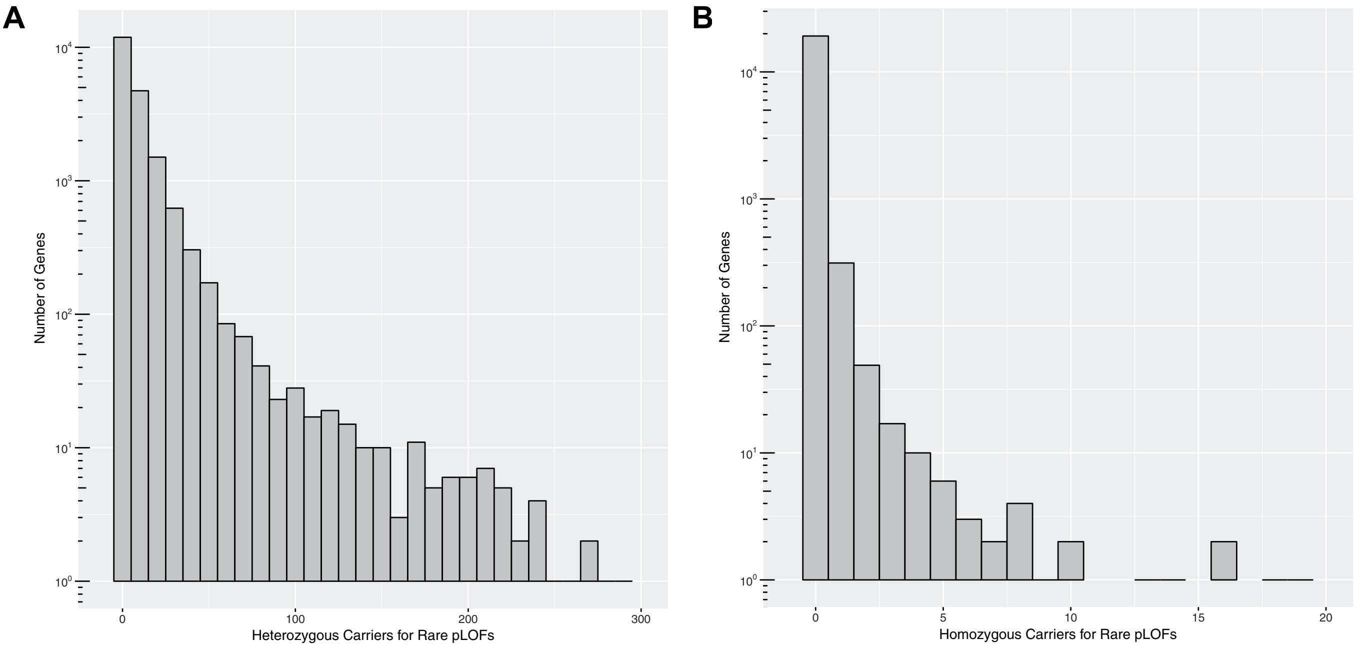
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-1133-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-1133-8.

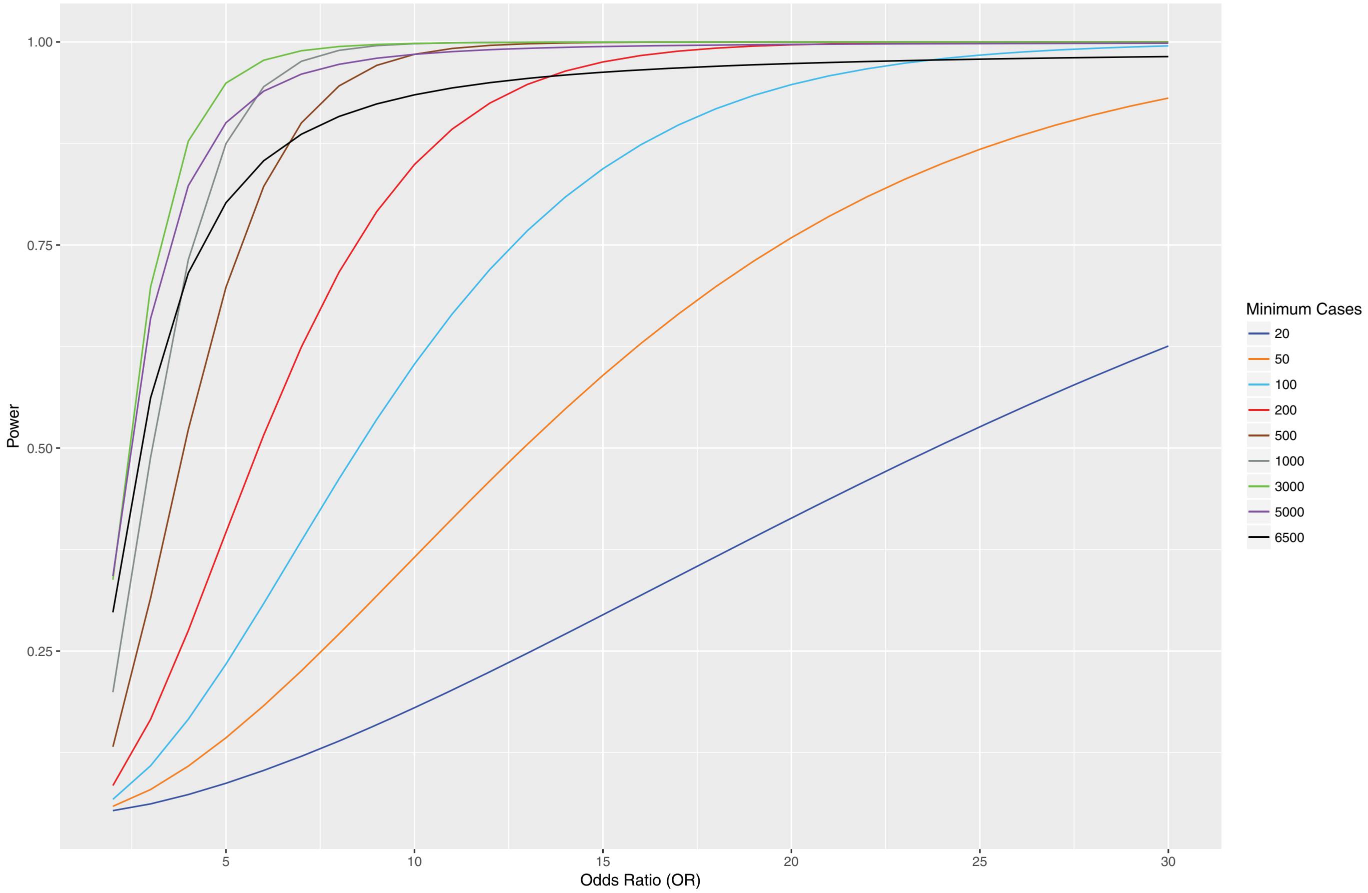
Correspondence and requests for materials should be addressed to D.J.R.

Peer review information Michael Basson and Kate Gao were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

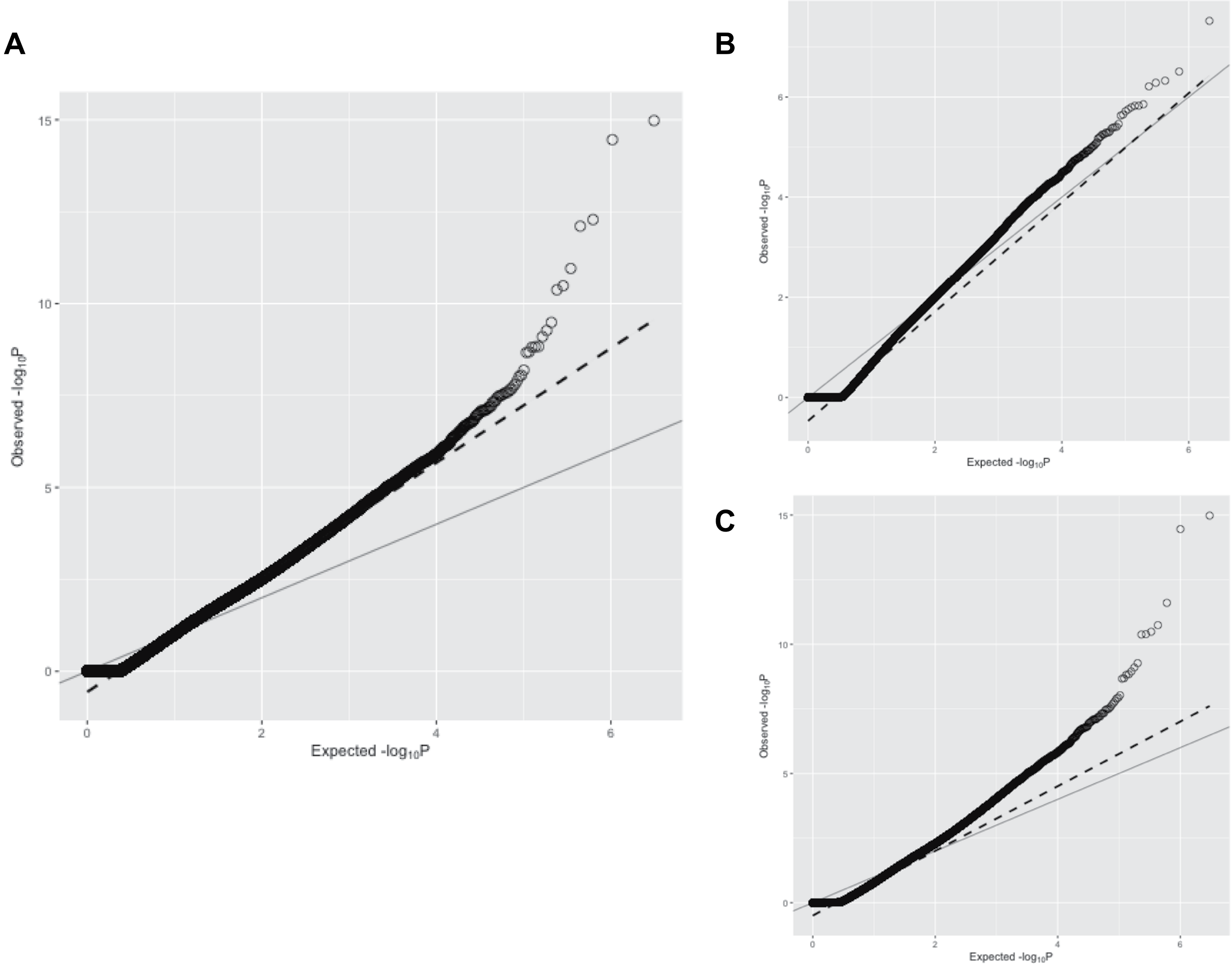
Reprints and permissions information is available at www.nature.com/reprints.



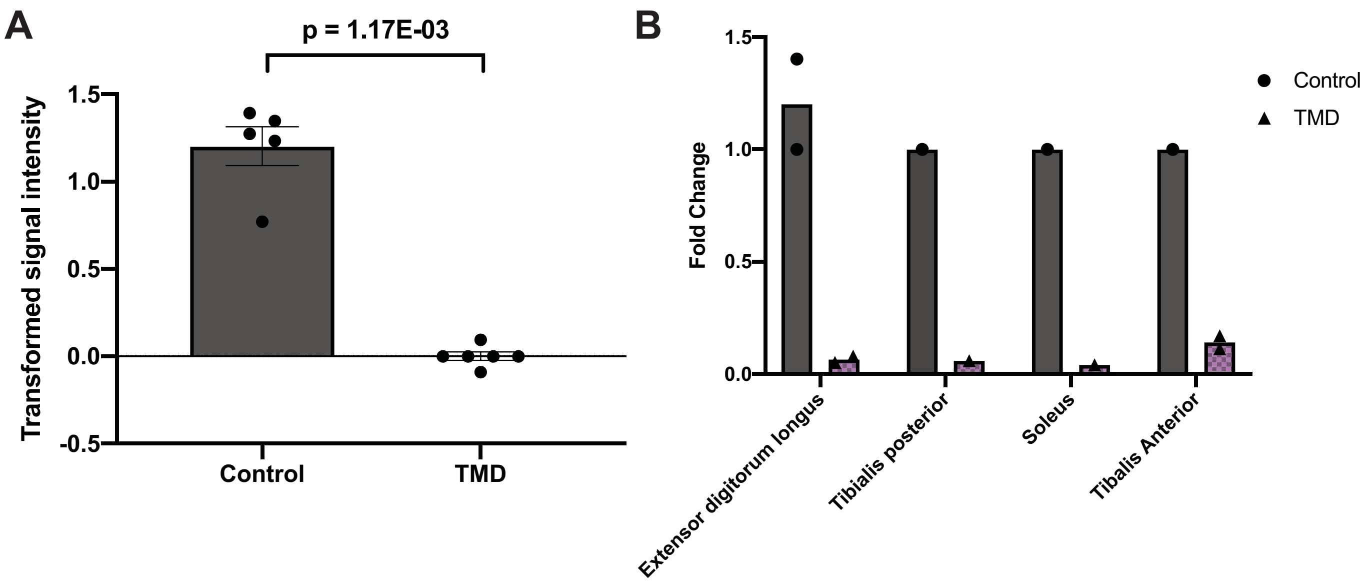
Extended Data Fig. 1 | Distribution of number of carriers for rare predicted loss-of-function (pLOF) variants per gene in the Penn Medicine Biobank. **a**, Histogram plot for the distribution of number of heterozygous carriers for rare (MAF ≤ 0.1%) pLOF variants per gene in the Penn Medicine Biobank's (PMBB) exome sequenced cohort. The x-axis represents number of heterozygous pLOF carriers per gene in bin widths of 10, and the log-scaled y-axis represents the number of genes with the x-axis-specified number of heterozygous carriers. **b**, Histogram plot for the distribution of number of homozygous carriers for rare pLOF variants per gene in the PMBB's exome sequenced cohort. The x-axis represents number of homozygous pLOF carriers per gene in bin widths of one, and the log-scaled y-axis represents the number of genes with the x-axis-specified number of homozygous carriers.



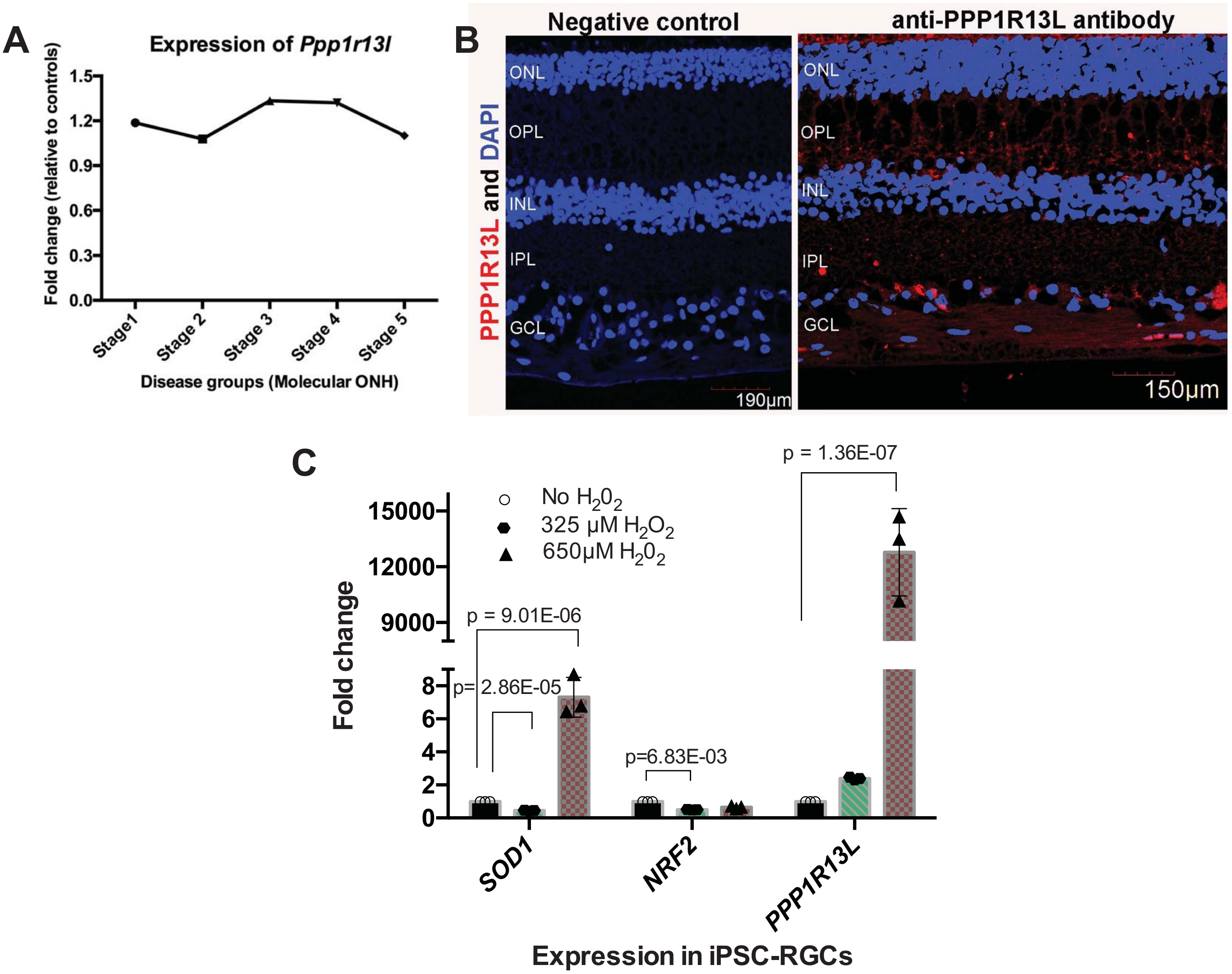
Extended Data Fig. 2 | Power analyses for association of gene burdens with at least 25 heterozygous carriers for rare pLOF variants with phenotypes of various case counts. Power analyses for association of gene burdens collapsing rare pLOF variants with 25 heterozygous carriers (allele frequency = 25/2N ≈ 0.001, where N = 2172 (AFR) + 8198 (EUR)) with phenotypes having various case counts. Phenotype case counts range from 20 to 6500 to reflect the range of case counts for phecodes in the Penn Medicine Biobank discovery cohort, and the power of the gene burden association with each phenotype as a function of odds ratio (OR=exp(beta)) is plotted on separate lines per the plot legend.



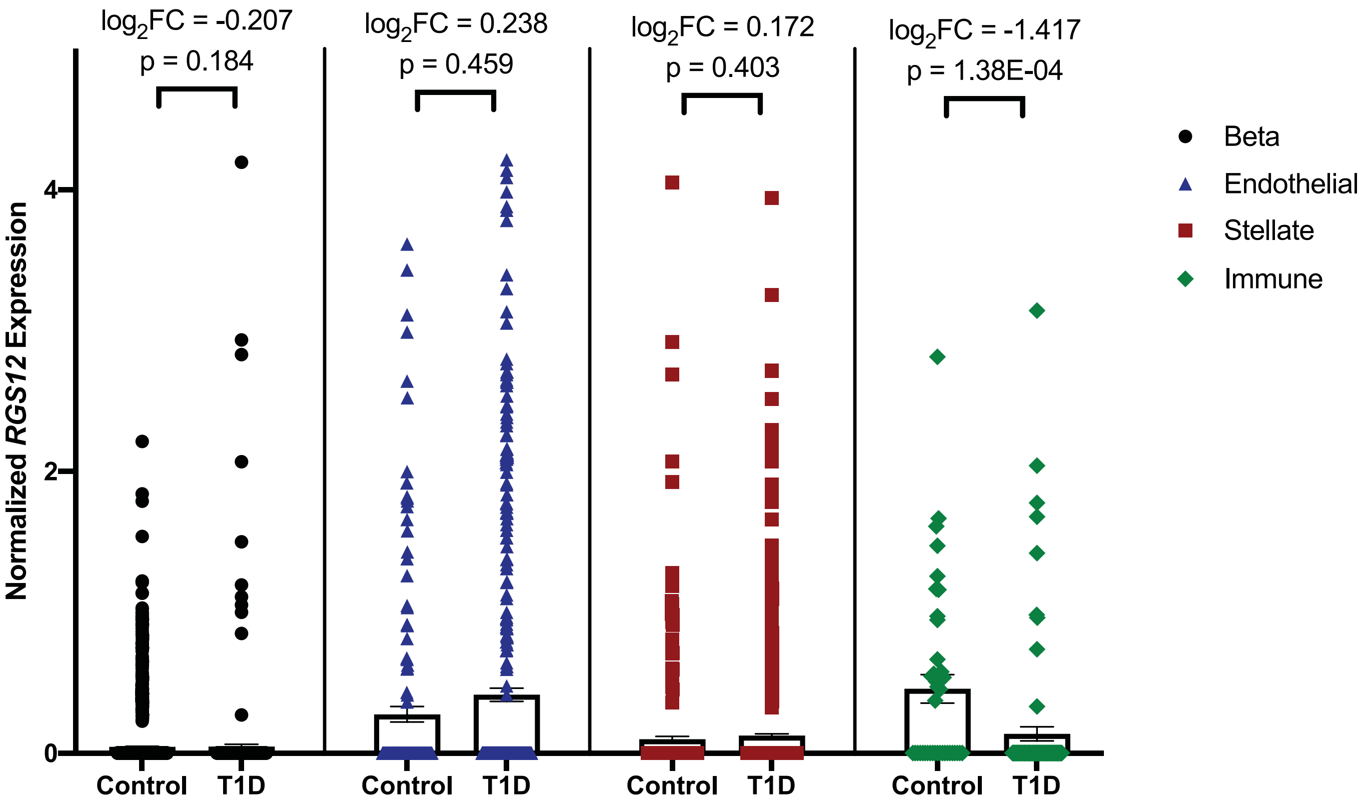
Extended Data Fig. 3 | Quantile-quantile plot of gene burden testing results from discovery phase of exome-by-phenome-wide association studies in the Penn Medicine Biobank. a, Quantile-quantile plot of p values from all exome-by-phenome-wide associations using gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants per gene in the Penn Medicine Biobank (PMBB). The x-axis represents the expected $-\log_{10}$(p value) under the uniform distribution of p values. The y-axis represents the observed $-\log_{10}$(p value) from the discovery phase of the exome-by-phenome-wide gene burden association studies collapsing rare pLOF variants in the PMBB. Each point represents an association between one of 1518 gene burdens and one of 1000 phecodes via logistic regression. The solid line shows the relationship between the expected and observed p values under the uniform p value distribution. The dashed line represents the observed fit line between the 50th and 95th percentile of gene burden associations, and the slope of this line is $\lambda_{\Delta 95} = 1.558$. **b**, AFR-specific QQ plot of p values from all exome-by-phenome-wide associations using gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants per gene in the PMBB. Data is presented in a similar manner to panel A. The slope of the fitted line is the AFR-specific $\lambda_{\Delta 95} = 1.09$. C) EUR-specific QQ plot of p values from all exome-by-phenome-wide associations using gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants per gene in the PMBB. Data is presented in a similar manner to panel A. The slope of the fitted line is the EUR-specific $\lambda_{\Delta 95} = 1.251$.



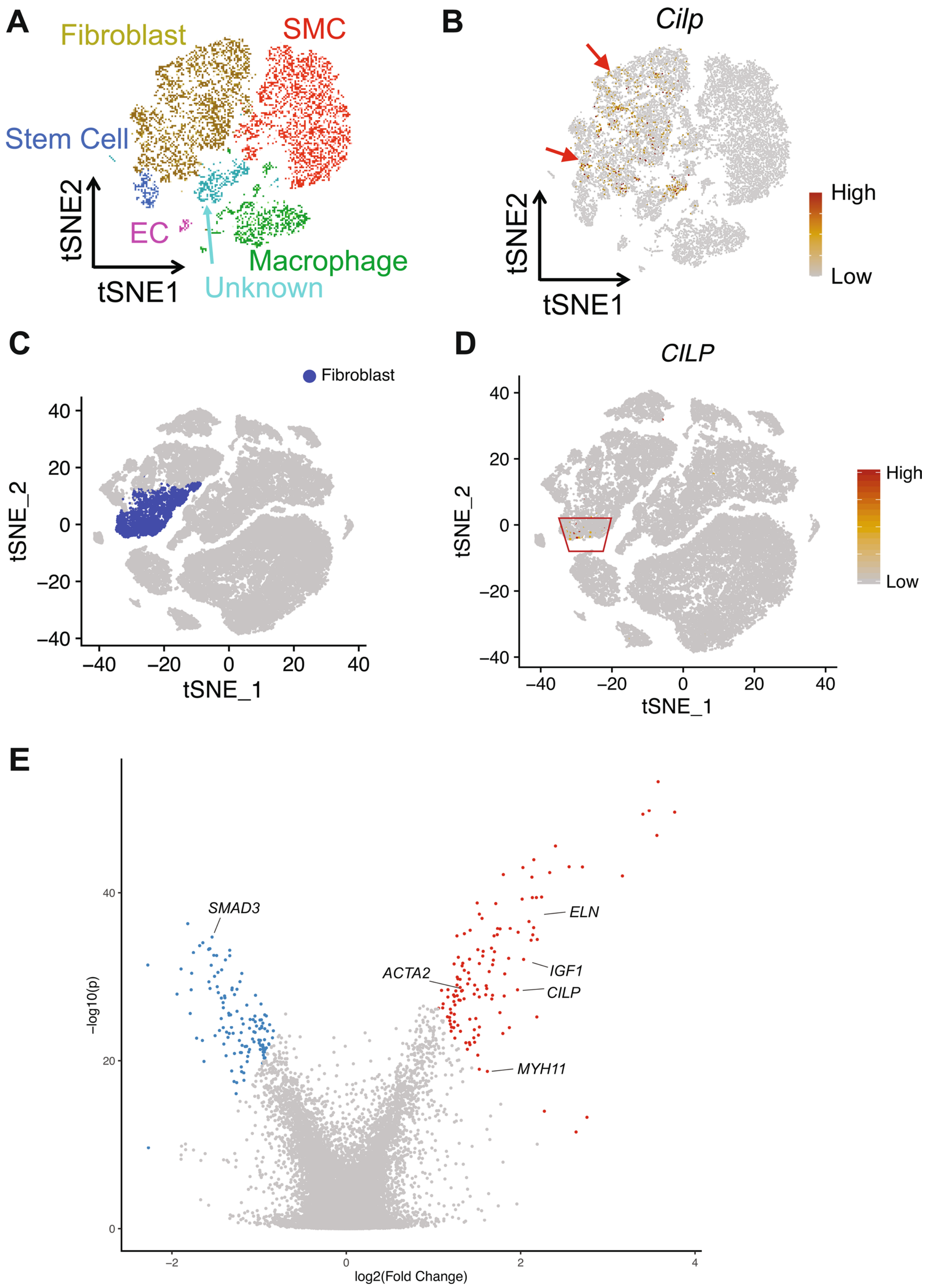
Extended Data Fig. 4 | *MYCBP2* is downregulated in tibial muscular dystrophy. a, Comparison of *MYCBP2* expression levels in human distal lower extremity muscles in tibial muscular dystrophy (TMD; N=6 independent muscle samples) versus healthy controls (N=5 muscle samples). Data is presented as mean transformed signal intensity, and error bars denote SEM. Transformed signal intensity values were obtained from GEO Series GSE42806, which are baseline-transformed and MAS5.0-normalized signal intensities, and individual values are plotted overlaying the bar plot. Statistical comparison was based on a moderated t-statistic, and p values were adjusted by Benjamini & Hochberg (FDR) correction. **b**, Comparison of *MYCBP2* expression levels in each distal lower extremity muscle included in the comparison in Extended Data Fig. 4a. Data is presented as a bar plot showing mean fold change as compared to a single control sample, and individual values are plotted overlaying the bar plot. Fold changes were calculated based on inverse log-transformed signal intensity values from each lower extremity muscle, including extensor digitorum longus (N=2 independent TMD samples, 2 independent control samples), tibialis posterior (N=1 TMD sample, 1 control sample), soleus (N=1 TMD sample, 1 control sample), and tibialis anterior (N=2 TMD samples, 1 control sample).



Extended Data Fig. 5 | Functional validation for the association between *PPP1R13L* and primary open-angle glaucoma. a, Differential expression profile of *Ppp1r13l* transcript in mouse optic nerve head (ONH) with varying stages of intraocular pressure (IOP)-induced glaucoma. Data represent the fold change in *Ppp1r13l* expression between different stages of D2 mice (glaucoma, N=50 mice) and D2 Gpnmb+ samples (control, N=10 mice). **b**, Localization of PPP1R13L protein in the human retina. Shown is the distribution of PPP1R13L by immunohistochemical localization in the retina from normal 68-year-old donor eyes. Overlay of images from DAPI (blue; nuclei) and PPP1R13L (red) in adult human retinal layers are shown on the right. The left represents primary antibody control. Scale bars are shown in each image. The experiment was performed twice independently with consistent results. ONL, outer nuclear layer; OPL, outer plexiform layer; IPL, inner plexiform layer; GCL, ganglion cell layer. **c**, Relative expression of *PPP1R13L* transcript in response to oxidative stress in induced pluripotent stem cell-derived retinal ganglion cells (iPSC-RGCs). A two-tailed unpaired Student's *t* test was used for statistical analysis, and significant p values are shown. Expression of *PPP1R13L* in iPSC-RGCs is shown under increasing concentrations of H_2O_2 treatment (N=3 independent experiments per condition). Plotted are the mean fold changes in comparison to no H_2O_2, error bars represent standard error of the mean (SEM), and individual values are plotted overlaying the bar plot.



Extended Data Fig. 6 | Single-cell RNA-seq of human pancreatic cells shows that *RGS12* is not differentially expressed in pancreatic exocrine and endocrine cells, but is reduced in type 1 diabetic peri-islet macrophages. Comparison of *RGS12* expression levels in type 1 diabetes (T1D) versus control in pancreatic beta (endocrine; N=2 T1D donors (410 cells), N=6 control donors (1573 cells)), endothelial (N=5 T1D donors (441 cells), N=6 control donors (166 cells)), stellate (exocrine; N=5 T1D donors (910 cells), N=6 control donors (356 cells)), and peri-islet immune (CD45+ macrophages; N=5 T1D donors (95 cells), N=4 control donors (40 cells)) cells based on single-cell RNA-seq. Differential expression of *RGS12* in each cell type was determined by edgeR, which 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. Data is presented as bars representing mean normalized *RGS12* expression and error bars representing standard error of the mean (SEM). Individual points are plotted overlaying their respective bar plots. Differential expression as determined by edgeR are displayed for each cell type as $\log_2$ fold change and p values adjusted by FDR correction.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | *CILP* is expressed in aortic adventitial fibroblasts, and is downregulated in human fibroblasts in response to treatment with TGF-β. a, t-SNE plot of aortic single cells in mice. Colors denote 6 cell types: smooth muscle cell (SMC), fibroblast, endothelial cell (EC), macrophage, stem cell, unknown. **b**, Relative expression of *Cilp* in all cells projected onto a t-SNE plot based on single-cell RNA-seq. The red arrows indicate where *Cilp* is expressed. **c**, t-SNE plot of aortic single cells in humans, with fibroblasts highlighted. **d**, Relative expression of *CILP* in all cells projected onto a t-SNE plot based on single-cell RNA-seq. The red box indicates where *CILP* is expressed. **e**,Volcano plot displaying differential expression of genes from meta-analysis of microarray and RNA-seq data for human fibroblasts treated with TGF-β (see Methods or the Reporting Summary for details about the datasets used). Meta-analysis of differential expression across the datasets was achieved using the Fisher's combined probability test. The x-axis represents meta-analyzed $\log_2$(fold change), and the y-axis represents meta-analyzed $-\log_{10}$(p value). The top 1% of differentially expressed genes across all datasets are labeled in red (upregulation) or blue (downregulation).

nature research

Corresponding author(s): Daniel J Rader

Last updated by author(s): Aug 31, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Variant annotation was performed using ANNOVAR (version 2018Apr16). Phenotypes were mapped to Phecodes using the R package "PheWAS" via R version 3.3.1 or later. Images of the retinal distribution of PPP1R13L protein were captured using AxioVS40 software version 4.8.1.0. For scRNA-seq of mouse and human aortas, raw FASTQ data alignment was processed using Cell Ranger 3.0. For scRNA-seq of human pancreas, expression data was generated using Cell Ranger 2.1.0.
Data analysis	All statistical analyses of de-identified clinical data were performed using R version 3.3.1 or later. PheWAS was performed using the R package "PheWAS" via R version 3.3.1 or later. Power analyses were conducted using QUANTO version 1.2.4. Public expression datasets from GEO were analyzed using GEO2R and R version 3.6.1. Meta-analysis of differential expression across datasets and its visualization were achieved using the R package "MetaVolcanoR 1.0.1". Public expression data from the Glaucoma Discovery Platform were analyzed using Datgan. For scRNA-seq of human pancreas, Seurat 3.0.2 was used for filtering, UMAP generation, and initial clustering. DoubletFinder 2.0 was used to demarcate and remove potential doublets in the data. Garnett was used for initial cell classification. TooManyCells 2.0.0.0 was then used to cluster and visualize the single cells. Differential genes were found using edgeR 3.24.3. For scRNA-seq of mouse aortas, Seurat 3.1.4 was used for filtering, normalization, and clustering. For scRNA-seq of human aortas, Seurat 3.0 was used for filtering, normalization, and clustering.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All summary statistics for significant gene-phenotype associations from the discovery phase in PMBB as well as significant replications from each replication cohort are fully detailed in the Supplementary Information (Table S1-S16). Data for the individual rare pLOF and missense variants in significant genes that were used for gene burden analyses in the PMBB discovery cohort are also included in the Supplementary Information (Tables S23-S24). In addition, a list of all of the single variants that were used for replication analyses across all the cohorts are provided in the Supplementary Information (Table S25). Each variant in Tables S23-25 is annotated with information regarding genomic location, variant effect, amino acid change, REVEL score (for missense), and minor allele frequency in gnomAD as well as in the PMBB discovery cohort. Additionally, up-to-date summary data for genetic variants captured via whole-exome sequencing in PMBB can be accessed via the Penn Medicine Biobank Genome Browser (https://pmbb.med.upenn.edu/biobank/allele-frequency/). Individual-level data are not made publicly available due to research participant privacy concerns; however, requests from accredited researchers for access to individual-level data relevant to this manuscript can be made by contacting the corresponding author. Additionally, public expression datasets were obtained from the Ocular Tissue Database (https://genome.uiowa.edu/otdb/), Glaucoma Discovery Platform (http://glaucomadb.jax.org/glaucoma), and the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). From NCBI GEO, 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) as well as microarray data from muscle biopsies in tibial muscular dystrophy patients (GSE42806).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Our discovery experiment in Penn Medicine Biobank (PMBB) included a subset of 10,900 individuals in the PMBB who have undergone whole-exome sequencing following quality control measures. We chose to interrogate gene burdens with at least 25 carriers, which is computationally equivalent to a single variant with minor allele frequency ~ 0.1%. Given that a gene burden of rare variants is expected to have significantly higher effect sizes than common variants, we show through power analyses (Extended Data Figure 2) that this minimum is sufficient for discovery of significant associations across variable odds ratios.

For replication studies in other biobanks, including BioMe (N=23,989), DiscovEHR (N=85,450), UK Biobank (N=32,268), and BioVU (N=66,400), the sample size was opportunistic and was determined by the availability of samples in each biobank at the time of genetic sequencing.

For qRT-PCR of iPSC-RGCs, we did not use statistical methods to predetermine sample size but are confident our sample numbers give enough confidence to interpret significance when fold change is compared against untreated control sample given the clear differences seen in Extended Figure S5C as well as consistency seen from replication..

For scRNA-seq of human pancreas, sample size calculation was not performed since all high-quality single cells from all donor samples that were collected by HPAP at the time of the preparation of this manuscript with high quality scRNA-seq data were used in this analysis.

For scRNA-seq of mouse aorta, sample size was not predetermined. Instead, given that 3 to 5 samples need to be pooled to establish a cDNA library for scRNA-seq studies in mouse aortas, the present study pooled 5 ascending aortas from mice. Similarly, for scRNA-seq of human aorta, sample size was not predetermined. Instead, ascending aorta samples from 11 individuals were pooled for this study.

Data exclusions

For the discovery experiment in PMBB, on the genotypic side, from a total of 11,451 individuals with whole-exome sequencing in PMBB, 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 following pre-established protocols for quality control of exome sequencing data. On the phenotypic side, 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. These pre-established phenotypic exclusion criteria were implemented to add a level of stringency to increase sensitivity for defining cases for disease phenotypes. For replication analyses in all other biobanks, similar exclusion criteria were implemented with regard to both genotypic and phenotypic aspects.

Replication

Evaluation of the robustness of the associations we found during the discovery phase with exome-by-phenome-wide significance (p<E-06) is a central theme of this paper. Firstly, in the same cohort, we interrogated other non-overlapping predicted deleterious variants in the same genes to test whether their association with the originally associated phenotype on discovery withheld. Then, we repeated the same experiments in multiple replication cohorts: 1) a new (non-overlapping) set of 6,432 exomes in African-Americans in the Penn Medicine Biobank (PMBB2); 2) 23,989 exomes (6,470 African, 8,735 European, 8,784 Hispanic) from Mount Sinai's BioMe; 3) 85,450 exomes (European)

from the Geisinger Health System's DiscovEHR cohort; 4) 32,268 European exomes from the UK Biobank; and 5) 66,400 genotypes (10,456 African, 55,944 European) from Vanderbilt's BioVU. Successful replication attempts are detailed in Table S17.

For immunohistochemical localization of PPP1R13L in the retina, the experiment was performed twice independently with consistent results. For qRT-PCR of iPSC-RGCs, we performed the qRT-PCR using RNA derived from a triplicate experiment for each concentration of hydrogen peroxide.

For scRNA-seq of human pancreas, all high-quality single cells from all donor samples that were collected by HPAP at the time of the preparation of this manuscript with high quality scRNA-seq data were used in this analysis. For each donor, a single scRNA-seq experiment was conducted using all high-quality single cells, and thus was carried out without replication.

For scRNA-seq of mouse and human aortas, all high quality single cells with high quality scRNA-seq data were used in this study. A single scRNA-seq experiment was conducted using all high-quality single cells for scRNA-seq of mouse aorta, and thus was carried out without replication. Similarly, a single scRNA-seq experiment was conducted using all high-quality single cells for scRNA-seq of human aorta for each sample, and thus was carried out without replication

Randomization

Individuals in PMBB, BioMe, DiscovEHR, UK Biobank, and BioVU were allocated into experimental groups based on their genotypic and phenotypic statuses, which were variables determined prior to this study. Each disease phenotype was tested for association with each gene burden or single variant adjusted for age, age^2, sex and the first ten principal components of genetic ancestry as covariates.

For transcriptional expression studies in iPSC-RGCs, samples were allocated into experimental groups based on their exposure to a particular concentration of hydrogen peroxide treatment.

For scRNA-seq of human pancreas, participants were allocated into the groups type 1 diabetes (T1D) vs. control based on their medical charts and C-peptide measurements in accordance with the American Diabetes Association guidelines.

For scRNA-seq of mouse aorta, there was no allocation of samples into experimental groups given that the mouse scRNA-seq study was performed using one group of normal aortas from wild type mice. For scRNA-seq of human aorta, control ascending aortic tissue samples were obtained from recipients of heart transplants or lung donors, and diseased aortic tissue samples were obtained from patients with sporadic ATAA excluding those who had ascending aortic dissection, an heritable form of aortopathy (e.g., Marfan syndrome, Loeys-Dietz syndrome, a first-degree relative with ATAA, bicuspid aortic valve), or ATAA related to infection, aortitis, trauma, or isolated pseudoaneurysm. However, gene expression analyses for this study were conducted on the entire cohort regardless of disease state.

Blinding

The investigators were blinded to group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
□	☒	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

anti-PPP1R13L (Cat# 51141-1-AP, Proteintech, IL, USA), chicken anti-rabbit IgG conjugated with Alexa Fluor 594 (Cat# A21442, Life Technologies, Carlsbad, CA)

Validation

Specificity of the antibody staining was determined when compared with immunostaining patterns observed with no antibody and a secondary antibody controls.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

The de-identified patient-derived iPSCs from Caucasian individuals were procured from the Human iPSC core facility, University of Pennsylvania.

Authentication

The hiPSCs 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. Detailed methods about iPSC charecterization and generation are described in PMID:26281015.

The iPSCs were differentiated into pure iPSC-RGCs cells with structural and functional features characteristic of native RGC cells based on a novel methodology we developed (PMID: 32678240)

Mycoplasma contamination

All the iPSC lines used for the study are negative for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Details on the mice used for single cell RNA sequencing in this study, including housing conditions, have been included in the Methods section of this manuscript. Mef2c-Cre strain was provided by Dr. Alan Daugherty, University of Kentucky; Rosa26-mTmG mice are available from the Jackson Laboratory (JAX) as strain B6.129(Cg)-Gt(ROSA)26Sortm4(ACTB-tdTomato,-EGFP)Luo/J(stock number 007676). Eight weeks old double-heterozygous Mef2c-Cre;Rosa26-mTmG male mice were used. All of the mice were on a C57BL/6 background.

Wild animals

No wild animals were used in this study.

Field-collected samples

No field-collected samples were used in this study.

Ethics oversight

Mice were maintained in the Center for Comparative Medicine at Baylor College of Medicine, and procedures were performed according to a protocol AN-4195 approved by the Institutional Animal Care and Use Committee at Baylor College of Medicine. All animal experiments complied with the National Institutes of Health guide for the care and use of Laboratory animals (NIH Publications No. 8023, revised 1978).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics

A subset of 10,900 individuals in PMBB were analyzed in the discovery phase of this study, as described in Table 1. They have a median age of 67, and have a variety of disease phenotypes as described in Table 1. The two most prevalent ancestries, Europeans (75.2%) and Africans (19.9%), were analyzed for exome-by-phenome-wide association analyses.

For replication analyses in BioMe, the cohort consisted of 23,989 individuals with 59% being female and having a median age of 61. 6,470 individuals were of African ancestry, 8,735 were of European ancestry, and 8,784 were of Hispanic ancestry. They have a variety of disease phenotypes as described in Table S18.

For replication analyses in DiscovEHR, the cohort consisted of 85,450 individuals of European ancestry with a median age of 59. They have a variety of disease phenotypes as described in Table S18.

For replication analyses in UK Biobank, the cohort consisted of 32,268 individuals of European ancestry with a median age of 59. They have a variety of disease phenotypes as described in Table S18.

For replication analyses in BioVU, the cohort consisted of 66,400 genotype individuals with a median age of 56 and 56.4% female. 10,456 individuals are of African ancestry and 55,944 individuals are of European ancestry.

The retinal eye sections in this study were obtained from a normal cadaver eye globe from a 68 year-old, female, Caucasian donor.

For scRNA-seq of human pancreas cell types in type 1 diabetes (T1D) versus control, 5 individuals with T1D and 6 control individuals were recruited for this study. 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.

For scRNA-seq of human aorta, 11 individuals were recruited for this study. 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). 54.5% were female and the median age was 63. 81.8% were of White race, 9.1% Black, and 9.1% Latino.

Recruitment

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. All genotypic and phenotypic data were de-identified prior to analyses.

For replication analyses in BioMe, samples were ascertained based on the patient population of Mount Sinai who enrolled at various Mount Sinai clinical sites in and around the New York City region.

For replication analyses in DiscovEHR, all samples were drawn from MyCode participants, whom provided informed consent that allows their clinical and genomic data to be used for health research (PMID: 26866580 and PMID: 28008009).

For replication analyses in UK Biobank, access to the UK Biobank for this project was from Application 32133.

For replication analyses in BioVU, all samples were drawn from participating Vanderbilt clinic patients who provided consent.

For scRNA-seq of human pancreas, pancreatic islets were procured from the HPAP consortium under Human Islet Research Network (https://hirnetwork.org/). To avoid self-selection bias in donor selection, all high-quality single cells from all donor samples that were collected by HPAP at the time of the preparation of this manuscript with high quality scRNA-seq data were used in this analysis. To avoid self-selection bias in cell clustering, two independent algorithms were employed, Seurat and TooManyCells. To avoid self-selection biases in defining cell types, we used Garnett, a regression-based classifier to assist in the automation of cell type classification

For scRNA-seq of human aorta, 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). Furthermore, two control samples came from heart transplant recipients. Although those patients did not have aortic aneurysms, they may have exhibited molecular or cellular changes in the ascending aorta related to their cardiac disease. However, as this dataset was used to identify gene expression in particular aortic cell types regardless of disease to mirror the scRNA-seq studies of mouse aorta, these differences are unlikely to change the conclusions of the results.

Ethics oversight

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.

Replication analyses in BioMe were approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai.

For replication analyses in DiscovEHR, the Geisinger IRB and MyCode Governing Board both reviewed and approved the use of MyCode data for this study.

All replication analyses in BioVU were approved under Vanderbilt IRB #200350.

For replication analyses in UK Biobank, access to the UK Biobank for this project was from Application 32133.

The donor eye tissue used for retinal eye sections in this study are exempt from IRB approvals (Exception 4).

For scRNA-seq of human pancreas, 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 scRNA-seq of human aorta, the protocol for collecting human aortic tissue samples 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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.