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

Authors: Joseph Park, Anastasia M Lucas, Xinyuan Zhang, Kumardeep Chaudhary, Judy H Cho, Girish Nadkarni, Amanda Dobbyn, Geetha Chittoor, Navya S Josyula, Nathan Katz, Joseph H Breeyear, Shadi Ahmadmehrabi, Theodore G Drivas, Venkata RM Chavali, Maria Fasolino, Hisashi Sawada, Alan Daugherty, Yanming Li, Chen Zhang, Yuki Bradford, JoEllen Weaver, Anurag Verma, Renae L Judy, Rachel L Kember, John D Overton, Jeffrey G Reid, Manuel AR Ferreira, Alexander Li, Aris Baras, Regeneron Genetics Center, Scott A LeMaire, Ying H Shen, Ali Naji, Klaus H Kaestner, Golnaz Vahedi, Todd L Edwards, Jinbo Chen, Scott M Damrauer, Anne E. Justice, Ron Do, Marylyn D Ritchie, Daniel J Rader

Supplementary Table Legends

Table S1. Summary statistics for 97 gene burdens with associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank.

List of 97 gene burdens with phecode associations with p<E-06 from discovery phase of exome-by-phenome-wide association studies in Penn Medicine Biobank based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry. Also reported are the betas and p values resulting from meta-analysis of European- and African-specific summary statistics based on Firth's penalized likelihood model, also adjusted for age, age^2, sex, and the first ten principal components of ancestry.

Table S2. Evaluation of robustness via REVEL-informed missense-based gene burdens within Penn Medicine Biobank.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted deleterious missense (REVEL ≥ 0.5) variants in Penn Medicine Biobank (PMBB) among 97 genes for which predicted loss-of-function (pLOF)-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in PMBB. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry.

Table S3. Evaluation of robustness via univariate association studies within Penn Medicine Biobank.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥ 0.5) variants in Penn Medicine Biobank (PMBB) among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in PMBB. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry. Additionally, single variants are annotated by their genomic location according to GRCh37/hg19, as well as their rs identification codes if available.

Table S4: Replication via predicted loss-of-function (pLOF)-based gene burdens in PMBB2.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants in an independent cohort of African Americans in Penn Medicine Biobank (PMBB2) among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of African ancestry only in PMBB2.

Table S5: Replication via REVEL-informed missense-based gene burdens in PMBB2.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted deleterious missense (REVEL ≥ 0.5) variants in an independent cohort of African Americans in Penn Medicine Biobank (PMBB2) among 97 genes for which predicted loss-of-function (pLOF)-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of African ancestry only in PMBB2.

Table S6. Replication via univariate association studies in PMBB2.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥ 0.5) variants in an independent cohort of African Americans in Penn Medicine Biobank (PMBB2) among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of African ancestry only in PMBB2. Additionally, single variants are annotated by their genomic location according to GRCh38/hg38, as well as their rs identification codes if available.

Table S7: Replication via predicted loss-of-function (pLOF)-based gene burdens in BioMe.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants in BioMe among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European, African, and Hispanic ancestry.

Table S8: Replication via REVEL-informed missense-based gene burdens in BioMe.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted deleterious missense (REVEL ≥ 0.5) variants in BioMe among 97 genes for which predicted loss-of-function (pLOF)-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European, African, and Hispanic ancestry.

Table S9. Replication via univariate association studies in BioMe.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥

0.5) variants in BioMe among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European, African, and Hispanic ancestry.

Additionally, single variants are annotated by their genomic location according to GRCh38/hg38, as well as their rs identification codes if available.

Table S10: Replication via predicted loss-of-function (pLOF)-based gene burdens in DiscovEHR.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants in DiscovEHR among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in DiscovEHR.

Table S11: Replication via REVEL-informed missense-based gene burdens in DiscovEHR.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted deleterious missense (REVEL ≥ 0.5) variants in DiscovEHR among 97 genes for which predicted loss-of-function (pLOF)-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in DiscovEHR.

Table S12. Replication via univariate association studies in DiscovEHR.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥ 0.5) variants in DiscovEHR among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine

Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in DiscovEHR. Additionally, single variants are annotated by their genomic location according to GRCh38/hg38, as well as their rs identification codes if available.

Table S13: Replication via predicted loss-of-function (pLOF)-based gene burdens in UK Biobank.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted loss-of-function (pLOF) variants in UK Biobank (UKB) among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in UKB.

Table S14: Replication via REVEL-informed missense-based gene burdens in UK Biobank.

List of significantly replicated associations via gene burdens collapsing rare (MAF ≤ 0.1%) predicted deleterious missense (REVEL ≥ 0.5) variants in UK Biobank (UKB) among 97 genes for which predicted loss-of-function (pLOF)-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in UKB.

Table S15. Replication via univariate association studies in UK Biobank.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥ 0.5) variants in UK Biobank (UKB) among 97 genes for which pLOF-based gene burdens had associations with p<E-06 from exome-by-phenome-wide association studies in Penn Medicine

Biobank. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from running the logistic regression model in individuals of European ancestry only in UKB. Additionally, single variants are annotated by their genomic location according to GRCh38/hg38, as well as their rs identification codes if available.

Table S16. Replication via univariate association studies in BioVU.

List of significantly replicated associations via univariate analyses of low-frequency to common (MAF > 0.1%) predicted loss-of-function (pLOF) or predicted deleterious missense (REVEL ≥ 0.5) variants in BioVU among single variants significantly replicated in PMBB, PMBB2, and/or UKB. Summary statistics are based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry. Additionally, single variants are annotated by their genomic location according to GRCh37/hg19, as well as their rs identification codes if available.

Table S17. List of robust exome-by-phenome-wide significant gene-phenotype associations with DiCE labeling.

List of 26 genes among 97 gene burdens with phenotype associations at p<E-06 that were most robust according to a Diverse Convergent Evidence (DiCE) approach integrating successful replication as well as clinical or experimental evidence. For replication studies, gene-phenotype associations were evaluated for their robustness by interrogating pLOF-based gene burdens, REVEL-informed missense-based gene burdens, and single variants in Penn Medicine Biobank (PMBB), an independent cohort of African Americans in PMBB (PMBB2), BioMe, DiscovEHR, and UK Biobank (UKB). Targeted single variants that showed successful replication in PMBB, PMBB2, and UKB were additionally replicated in BioVU. Each association is labeled with the corresponding p value from the logistic regression studies in the discovery phase in PMBB, and checkmarks for successful replication in their respective cohorts (p<0.05) and/or supportive observational/experimental data. Gene-phenotype associations with at least two total checkmarks were deemed robust and were included in this table. Positive-control associations are listed on the top and are separated from novel associations by a double line. Positive-control and novel associations are each ranked alphabetically by gene name.

Table S18. Comparisons of case-control ratios for phecodes of interest across all interrogated WES datasets.

List of phecodes that had associations at p<E-06 from logistic regression analyses in the discovery phase of exome-by-phenome-wide association studies in Penn Medicine Biobank (PMBB). Phecodes are sorted by phecode number, and are each labeled with number of cases, number of controls, and case-control ratios in PMBB, an independent cohort of African Americans in PMBB (PMBB2), BioMe, DiscovEHR, and UK Biobank (UKB). *For phecodes that were not mapped to from any ICD-10 codes in Phecode Map 1.2b1, corresponding ICD-9 codes according to Phecode Map 1.2 were translated to ICD-10 codes, then mapped back to phecodes via Phecode Map 1.2b1. If the resulting phecode was different, then the case and control counts for the new phecode were listed. **ICD-10 R791, which corresponds to the ICD-9 that maps to phecode 286.9, does not map to any phecodes in Phecode Map 1.2b1. ***There were <20 cases in each DiscovEHR cohort analyzed.

Table S19. Echocardiographic measurements of cardiomyopathy-associated gene burdens in Penn Medicine Biobank.

Comparison of representative echocardiography parameters for cardiac size and functionality between heterozygous carriers of rare pLOF variants included in cardiomyopathy-associated gene burdens from the Penn Medicine Biobank (PMBB) discovery phase versus individuals in PMBB not carrying any of the respective gene's pLOF variants with echo data available. Data is represented as beta and p value for robust linear regression adjusted for age, sex, and the first four principal components of genetic ancestry. Analyses were limited to individuals of European ancestry only.

Table S20. Expression of *PPP1R13L* transcript in human ocular tissues.

Expression of *PPP1R13L* in human ocular tissues per the Ocular Tissue Database (OTDB) ranked from highest to lowest expression per ocular tissue. Expression values are represented as Probe Logarithmic Intensity Error (PLIER) values, where individual gene expression values are normalized to its expression in other tissues. Probe ID: 3865344.

Table S21. African-predominant single variants from significant replication studies.

List of significant single variants from replication studies from cohorts with individuals of African ancestry (PMBB, PMBB2, BioMe, and BioVU) predominantly found among Africans according to gnomAD. Each variant is labeled with its corresponding amino acid change, rs ID, the Phecode that was significantly associated via logistic regression, and in which cohort the replication was significant. "X" denotes $p<0.05$; "." denotes $p<0.2$; "-" denotes lack of genotypic power (*i.e.* number of minor alleles < 5 or the variant had high missingness); "†" denotes lack of phenotypic power (*i.e.* cases < 20).

Table S22. Confirmation of exome-by-phenome-wide significant associations with hypertrophic cardiomyopathy for *MYBPC3* and *BBS10*.

Summary statistics for association of rare pLOF-based gene burdens for *MYBPC3* and *BBS10* with hypertrophic cardiomyopathy phenotypes in Penn Medicine Biobank based on an exact logistic regression model adjusted for age, age^2, sex, and the first ten principal components of ancestry. The phenotype "hypertrophic cardiomyopathy" represents the combination of phecodes 425.11 and 425.12 such that individuals who are a case for phecode 425.11 or 425.12 are a case for hypertrophic cardiomyopathy. Reported are the betas and p values resulting from meta-analysis of the summary statistics obtained from running the logistic regression model separately in individuals of European and African ancestry.

Table S23. Information about rare pLOF variants captured via whole-exome sequencing in the PMBB discovery cohort used for exome-by-phenome-wide gene burden discovery analyses.

List of rare pLOF variants in the 97 significant genes used for exome-by-phenome-wide gene burden discovery analyses in the PMBB cohort. Each variant per row is annotated with gene name, genomic location according to GRCh37/hg19, rs ID if available, variant effect, gnomAD MAFs (African, Non-Finnish European, all ancestries), and MAF in the PMBB discovery dataset (all ancestries).

Table S24. Information about rare missense variants with REVEL≥0.5 captured via whole-exome sequencing in the PMBB discovery cohort used for gene burden replication analyses.

List of rare missense variants with REVEL≥0.5 in the 97 significant genes used for gene burden replication analyses in the PMBB cohort. Each variant per row is annotated with gene name, genomic location according to GRCh37/hg19, rs ID if available, variant effect, REVEL score, gnomAD MAFs (African, Non-Finnish European, all ancestries), and MAF in the PMBB discovery dataset (all ancestries).

Table S25. Information about single variants used for univariate replication analyses across all replication cohorts.

List of pLOF and missense (REVEL≥0.5) variants in the 97 significant genes used for univariate replication analyses across all replication cohorts. Each variant per row is annotated with gene name, genomic location according to GRCh37/hg19, rs ID if available, variant effect, REVEL

score for missense, gnomAD MAFs (African, Non-Finnish European, all ancestries), and MAF in the PMBB discovery dataset (all ancestries).