Published in final edited form as:
Pac Symp Biocomput. 2019 ; 24: 272–283.

Detecting potential pleiotropy across cardiovascular and neurological diseases using univariate, bivariate, and multivariate methods on 43,870 individuals from the eMERGE network

Xinyuan Zhang[*,1], **Yogasudha Veturi**[*,2], **Shefali Verma**[2], **William Bone**[1], **Anurag Verma**[2], **Anastasia Lucas**[2], **Scott Hebbring**[3], **Joshua C. Denny**[4], **Ian B. Stanaway**[5], **Gail P. Jarvik**[5], **David Crosslin**[5], **Eric B. Larson**[6], **Laura Rasmussen-Torvik**[7], **Sarah A. Pendergrass**[8], **Jordan W. Smoller**[9], **Hakon Hakonarson**[10], **Patrick Sleiman**[10], **Chunhua Weng**[11], **David Fasel**[11], **Wei-Qi Wei**[12], **Iftikhar Kullo**[13], **Daniel Schaid**[14], **Wendy K. Chung**[15], and **Marylyn D. Ritchie**[†,2]

[1] Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[2] Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[3] Center for Human Genetics, Marshfield Clinic, Marshfield, WI 54449, USA

[4] Department of Medicine, Vanderbilt University, Nashville, TN 37235, USA

[5] Departments of Medicine (Medical Genetics) and Genomic Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA

[6] Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, USA

[7] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

[8] Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA 17822, USA

[9] Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[10] Center for Applied Genomics, Children's Hospital of Philadelphia, PA 19104, USA

[11] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA

[12] Department of Biomedical Informatics in School of Medicine, Vanderbilt University, Nashville, TN 37230, USA

[13] Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN 55905, USA

[†] Corresponding author.
[*] Authors contributed equally to this work

[14] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

[15] Department of Pediatrics, Columbia University, New York, NY 10032, USA

Abstract

The link between cardiovascular diseases and neurological disorders has been widely observed in the aging population. Disease prevention and treatment rely on understanding the potential genetic nexus of multiple diseases in these categories. In this study, we were interested in detecting pleiotropy, or the phenomenon in which a genetic variant influences more than one phenotype. Marker-phenotype association approaches can be grouped into univariate, bivariate, and multivariate categories based on the number of phenotypes considered at one time. Here we applied one statistical method per category followed by an eQTL colocalization analysis to identify potential pleiotropic variants that contribute to the link between cardiovascular and neurological diseases. We performed our analyses on ~530,000 common SNPs coupled with 65 electronic health record (EHR)-based phenotypes in 43,870 unrelated European adults from the Electronic Medical Records and Genomics (eMERGE) network. There were 31 variants identified by all three methods that showed significant associations across late onset cardiac- and neurologic- diseases. We further investigated functional implications of gene expression on the detected “lead SNPs” via colocalization analysis, providing a deeper understanding of the discovered associations. In summary, we present the framework and landscape for detecting potential pleiotropy using univariate, bivariate, multivariate, and colocalization methods. Further exploration of these potentially pleiotropic genetic variants will work toward understanding disease causing mechanisms across cardiovascular and neurological diseases and may assist in considering disease prevention as well as drug repositioning in future research.

Keywords

Pleiotropy; Cardiovascular Diseases; Neurological Disorders; Univariate Analysis; Bivariate Analysis; Multivariate Analysis; Colocalization; eQTL

1. Introduction

Cognitive decline has been observed in nearly 42% of elderly individuals at five years after cardiac surgery[1]. Of late, there has been increasing clinical evidence suggesting a link between cardiovascular and neurological diseases. To facilitate efficient disease prevention and treatment for cardiovascular and neurological diseases, it is imperative to understand the underlying, often unexplained, disease-causing mechanisms across multiple phenotypes. Pleiotropy is a phenomenon that can explain the influence of a specific allele on two or more unrelated phenotypes. While there has been evidence of polygenic pleiotropy (where multiple variants are causally associated with multiple traits) among cardiovascular[2] and neurological diseases[3], recent work has also demonstrated a genetic basis for the link *between* these disease groupings. In particular, there has been evidence of genetic overlap *between* cardiovascular disease and (a) multiple sclerosis[4] as well as (b) schizophrenia[5]. Large-scale genomics data coupled with electronic health record (EHR) data can enhance our ability to uncover novel cross phenotype associations and potentially pleiotropic variants (cross-phenotype association could also be an artifact of linkage disequilibrium (LD) or disease co-morbidities rather than true pleiotropy)[6]. In this study, we sought to identify common genetic variants that contribute to the link between diseases of the circulatory and nervous system using 43,870 unrelated European adults and 65 disease phenotypes from the Electronic Medical Records and Genomics (eMERGE) network.

Statistical approaches to detect pleiotropy across multiple phenotypes can be univariate (CPMA[7], ASSET[8], MultiMeta[9], GPA[10], MTAG[11], etc.), bivariate, and multivariate (MTMM[12], MultiPhen[13], GEMMA[14], mvLMM[15], mvBIMBAM[16], etc.) in addition to network-based approaches, among others[17]. Univariate methods (e.g. Phenome wide association studies or PheWAS) are a powerful way to characterize the effect of a genetic variant on each phenotype independently, and potential pleiotropy can be detected when the same SNP is found to be significantly associated with multiple phenotypes. This method has shown great success in identifying potential pleiotropy in several clinical genomics studies[18–23]. However, a limitation of univariate analysis is that it tests only one trait at a time, so it cannot be a formal test of pleiotropy. In contrast, bivariate analysis has been shown to have higher power over univariate analysis by analyzing pairs of phenotypes simultaneously[24]. Furthermore, because bivariate analysis can be structured to test the association of a trait with a variant, while adjusting for another trait's association with the variant, bivariate analyses can be constructed to formally test pleiotropy, and extended to multivariate traits to perform sequential tests for pleiotropic effects[25,26]. In this study, we used a bivariate analysis approach using summary-statistics from univariate analysis to test the hypothesis of "joint association" of a SNP with a trait pair while accounting for correlation in z-scores between the trait pair[24]. The alternative hypothesis here is that *at least* one of the two traits is significantly associated with a SNP marker. This implementation of bivariate analysis has suggested potential pleiotropy as well as hinted at underlying disease-causing mechanisms in many recent studies[27,28]. Finally, multivariate analysis is designed to test the joint association between genotype with multiple phenotypes in a single regression model. Multivariate analysis has been shown to have increased power over univariate analysis in many scenarios, including when the genotype affects either a single phenotype or multiple correlated phenotypes[29,30]. We chose MultiPhen[13] to perform multivariate analysis because of its ability to handle binary phenotypes as well as its high power, as demonstrated via simulations[29]. In this paper, we refer to MultiPhen as multivariate analysis for the sake of convenience. Again, here the alternative hypothesis is that *at least one* of many traits is significantly associated with the SNP marker.

Since the "true" pleiotropic associations among cardiovascular diseases and neurological disorders are largely unknown, we applied three types of widely used methods to characterize the landscape of *potential* pleiotropy at genome-wide level[31,32]. To improve our confidence that the list of potential pleiotropic variants obtained across all three methods reflect a single causal variant instead of coincidental overlap, we performed statistical colocalization for these signals with gene expression datasets across all 48 available tissues from the Genotype-Tissue Expression (GTEx) consortium[33]. For instance, if a SNP colocalizes with an eQTL for traits A *and* B, it means that the same SNP associates with both: (a) gene expression and trait A, (b) gene expression and trait B. This can help us infer that the same SNP associates with both traits A and B and is likely pleiotropic. We found that many of the potentially pleiotropic signals associated with both disease groupings (diseases of the nervous and circulatory system) colocalized with eQTLs from the GTEx consortium (especially on chromosome 22) indicating that gene expression might be influencing risk of disease at those loci. This study is one of the first large-scale natural data applications and evaluation of univariate, bivariate, multivariate and colocalization methods in one comprehensive analysis. The overall study design is shown in Figure 1.

2. Methods

2.1. eMERGE network

In this study, we used data from the Electronic Medical Records and Genomics (eMERGE) network Phase III. The eMERGE network is a National Human Genome Research Institute (NHGRI) organized consortium to explore the utility of DNA biorepositories coupled with Electronic Health Record (EHR) systems for large-scale genomic research. The eMERGE network Phase III consists of 83,717 genotyped samples across multiple platforms that are imputed to Haplotype Reference Consortium 1.1 reference in genome build 37 covering ~39 million genetic variants. There are seven eMERGE adult sites included in our study: Marshfield Clinic Research Foundation, Vanderbilt University Medical Center, Kaiser Permanente Washington/University of Washington, Mayo Clinic, Northwestern University, Geisinger, and Harvard University.

2.2. Genotypic Data and Quality Control

eMERGE Phase III imputed genotypic data were cleaned following the “best-practice” quality control (QC) pipeline designed for imputed data[34]. We included genetic variants with genotype call rate > 99% and sample call rate > 99%. We selected common variants with minor allele frequency (MAF) > 0.05. To account for sample relatedness, we dropped one of each related pair of individuals with pi_hat > 0.25 (obtained from identity-by-descent estimation using PLINK[35]). We filtered out variants that had a linkage disequilibrium r^2 greater than 0.5 using a 100kb sliding window. We also filtered out the variants with a mean of imputation score less than or equal to 0.4. We further removed variants which have MAF difference greater than 0.1 compared to European population from 1000 Genomes Project[34]. After genotypic QC assessment and LD pruning, we had 54,942 unrelated individuals of European ancestry and 533,878 SNPs.

2.3. Phenotype Definition and Selection Criteria

2.3.1. Phenotype Definition—Cardiovascular and neurological phenotypes were defined using International Classification of Diseases, Ninth Revision (ICD-9) billing codes. We selected 98 ICD-9 codes from “Diseases of the circulatory systems” and “Diseases of nervous system and sense organs” as our primary phenotypes. Table 1 presents the major disease groups and corresponding ICD-9 codes. Of note, association analyses were performed using individual ICD-9 codes to define case/control status, and we used broader major disease categories for the purpose of presentation. The number of clinical visits per ICD-9 code per individual was used to define case-control status for each ICD-9 code: a case would be assigned if an individual had ≥ 3 instances; a control would be assigned if an individual had zero instances; an NA would be assigned if an individual had one or two instances[22].

2.3.2. Phenotype Selection Criteria—Our cohort comprised adults of European ancestry (age ≥ 25 years old) from eMERGE network Phase III. We only used ICD-9 codes with more than or equal to 200 cases so as to increase statistical power of association tests[36]. As a result, a total of 65 cardiovascular and neurological ICD-9 based diagnoses and 43,870 individuals were included in our final round of association analyses. Individuals who have both cardiovascular and neurological disease were counted as cases for both. The sample size distribution of the 65 phenotypes is shown in Figure 2.

2.4. Association Methods

2.4.1. Univariate Analysis—We performed univariate logistic regression using 65 ICD-9 based diagnoses with 533,878 variants. We adjusted logistic regression models for sex, age, eMERGE site, and the first six principal components. We used PLINK 1.90 software[35] to perform the first round of univariate analysis because of its high computational efficiency. The logistic regression models converged for 33 out of 65 phenotypes. The major reason contributing to the non-convergence was the low sample sizes corresponding to some of the sites when we adjusted for eMERGE site (7 levels) as a categorical covariate. To address this, we used PLATO 2.1.0[37] to perform the second round of logistic regression tests on the remaining 32 phenotypes with the same set of covariates as before. Since PLATO implements an increased number of iterations compared to PLINK to find the best solution for logistic models, the software achieved convergence for all the remaining models. It should be noted that when both PLINK and PLATO converge, the results are concordant; these tools have been extensively compared previously[37].

2.4.2. Bivariate Analysis—Bivariate analysis involved using summary-statistics (Z scores) from univariate analyses. We modeled our bivariate analysis protocol (with modifications) on the one followed by Siewert et al[27]. We first estimated mean and covariance of the Z-scores obtained from univariate analyses for each of the 2080 pairs of phenotypes using all the available *LD-pruned* SNPs. This was done to ensure a null bivariate normal distribution of Z scores for each pair of phenotypes and to satisfy the “independence” assumption for hypothesis testing. Subsequently, we applied a p-value threshold of 0.005 on the univariate GWAS results and filtered out any SNPs that did not meet this threshold. We also filtered out SNPs with MAF = 0.5 to remove ambiguity pertaining to which allele was chosen as the referent allele in univariate analyses. Finally, we identified a list of common SNPs and estimated a p-value for each of 2,080 “pairs” of phenotypes using a chi-squared test with two degrees of freedom. Although we conducted a reduced number of tests, it should be noted that we corrected for multiple comparisons using the original “unfiltered” SNP set in order to control our type I error rate well.

2.4.3. Multivariate Analysis—We performed multivariate analysis using MultiPhen 2.0.2 R package[13]. MultiPhen analyzes multiple phenotypes jointly by testing linear combinations of phenotypes against each SNP using reverse ordinal regression. We adjusted for the same set of covariates as we did for univariate tests. By default, MultiPhen excludes individuals with at least one NA out of 65 phenotypes. Under this scenario, the power of association tests would be limited as there would only be 7,535 individuals in total with extremely low case sample size per phenotype. Since we applied the "rule of three" to define a case, any person who had one or two instances of the occurrence of an ICD-9 code was set to missing (NA). Because we did not want to drop so many individuals, we needed to fill in an alternative value for the N/A. For the purposes of multivariate analyses, these missing values were replaced by 0.5 to retain comparable sample size with univariate and bivariate analysis (sensitivity analyses on top significant SNPs yielded comparable results -- see Discussion). These individuals are *likely* cases since they have the ICD code in their record one or two times. A detailed evaluation of this replacement strategy will be conducted in the future to determine if a more optimal imputation strategy exists. Finally, to increase computational efficiency of MultiPhen, we parallelized the runs by splitting the genome into chunks of 10Mb each.

2.5. Statistical Correction

We implemented two Bonferroni correction calculation strategies to adjust for multiple testing when comparing the statistical performance of three types of methods. The Bonferroni threshold was calculated by dividing the level of significance by the number of tests. In the first strategy ("method-specific Bonferroni") we calculate Bonferroni threshold separately for each method. The derived significant thresholds for univariate, bivariate, multivariate testing were 1.44×10^{-9} [0.05/65*533878], 4.50×10^{-11} [0.05/(2080*533878)], and 9.37×10^{-8}[0.05/533878], respectively. We used an overly conservative significance threshold for bivariate analyses due to potential non-independence of tests (even after LD pruning). In the second strategy ("family-wise Bonferroni") we calculate Bonferroni threshold based on the total number of tests across all three methods. The derived significant threshold was 4.36×10^{-11} [0.05/(65*533878+2080*533878+533878)], and the criteria was applied across all three methods. Again, this correction is overly conservative given the correlation across the tests and methods but offers good control of the type I error rate.

2.6. Colocalization

Finally, we performed colocalization analysis to have greater confidence in our assessment of pleiotropy. We first obtained a list of potentially pleiotropic variants that cleared the "family-wise Bonferroni" multiple comparison threshold for univariate, bivariate and multivariate methods and narrowed down this list to SNPs that were associated with at least one disease from both nervous and circulatory systems. Finally, we ensured that for any given SNP, if one of the two traits in this circulatory-nervous trait pair had a univariate p-value that did not meet the "family-wise Bonferroni" threshold, it had a univariate -log10 p-value of at least 3. We termed the final list of SNPs as our "lead" SNPs. To test if these signals were being influenced by gene expression as well as driven by the same underlying variant, we performed statistical colocalization analyses using the "coloc" R package[38] between these signals and eQTLs (across all 48 available tissues) from the GTEx consortium[33]. We first obtained a 200KB window on either side of a "lead" SNP and looked for whether the lead SNP (or one in close LD with it) was an eQTL in a given tissue. If it was not an eQTL, that lead SNP was ignored. If it was an eQTL for a given tissue, we identified the corresponding "eGene" and obtained summary statistics from GTEx for all gene-variant associations in that 200KB window (either side). Note that we only chose the eGene that had the smallest p-value for a given eQTL from GTEx. Finally, for each phenotype with which the lead SNP is significantly associated, we performed statistical colocalization between the SNP and the corresponding eQTL in that tissue. We set a coloc threshold of PP4/(PP3+PP4) > 0.8 to identify pleiotropic signals that are strongly influenced by gene expression. Here PP4 refers to the posterior probability that a single SNP associates with the phenotype as well as the gene expression whereas PP3 refers to the posterior probability of having two independent SNPs associate with either.

3. Results

3.1. Landscape of Univariate, Bivariate and Multivariate Associations

The landscape of univariate, bivariate, and multivariate association results is shown in Figure 3. There is an overall similar trend of association signals for univariate and bivariate analysis. We found that bivariate analysis identified more significant associations than univariate analysis when the correlation between phenotypes was low (less than 0.4). From the bottom half of Figure 3, we can see if the association signal from bivariate analyses comes from pairs of circulatory, nervous or circulatory-nervous traits. Black dots in Figure 3 represent the variants that passed “method-specific Bonferroni” significance from multivariate analysis. There are scenarios in which there is no significant association from univariate/bivariate analyses but significant results from multivariate analyses. Using “method-specific Bonferroni” threshold, univariate, bivariate, and multivariate methods detected 124, 108, and, 107 unique statistically significant SNPs, respectively; and there are 49 overlapping SNPs across three methods (data not shown). The number of variants detected at the more stringent “family-wise” threshold is given in Figure 4.

3.2. Variants associated with cardiovascular disease and neurological disorders

Among the 31 “family-wise Bonferroni” SNPs across all three methods, we obtained 9 unique variants that are significantly associated with at least one cardiovascular disease and one neurological disorder from bivariate analysis that also “colocalized” with eQTLs across a host of tissues with a coloc PP4/(PP3+PP4) probability threshold of at least 0.8. Table 2 shows a comprehensive summary of these identified 9 variants. Our colocalization analyses revealed whether there was a shared variant underlying our potentially pleiotropic signals and whether gene expression may be influencing disease risk at these loci. For instance, the SNP at chromosome 1 and position 36822024 colocalized with eQTLs in the same 35 tissues for “Muscular dystrophies and other myopathies”, “Pain” and “Other conditions of the brain” (neurological phenotypes) as well as “Heart failure”, “Essential hypertension”, “Cardiac dysrhythmias” and “Hypotension” (cardiovascular phenotypes) (eGenes: *EVA1B, TRAPPC3*). This means that rs10796883 influences 4 different cardiovascular disease categories, 3 different neurological disease categories as well as gene expression for *EVA1B* and *TRAPPC3* eGenes across 35 different tissues. Likewise, the variant on chromosome 22 position 22947156 colocalized with eQTLs in 4 tissues (Brain-cerebellum, testis, transformed fibroblasts, small intestine ileum) for 4 different neurological phenotypes as well as 9 other cardiovascular phenotypes (eGenes*: IGLV3–21, GGTLC2*). Please refer to Supplementary table 1 at https://ritchielab.org/files/PSB2019/Veturi/

Supplementary_Data_1.txt for a complete list of tissues in which each of the lead SNPs colocalizes with eQTLs.

4. Discussion

In this study, we conducted EHR-based univariate, bivariate, and multivariate analyses on 43,870 adults of European ancestry from the eMERGE network using 65 cardiovascular and neurological ICD-9 disease categories. The aim of this study was to detect pleiotropic genetic variants that influence diseases of the circulatory and nervous systems. We also evaluated the performance of three types of methods for detecting pleiotropy.

We observed 79, 108, and, 58 unique variants, respectively that were detected by univariate, bivariate, and multivariate methods and 31 that overlapped among the three methods using a “family-wise Bonferroni” significance threshold. Univariate analysis suggests direct association between genetic variant and phenotype; bivariate association can offer insights into whether a variant is associated with a pair of phenotypes, whereas multivariate analysis is powerful in detecting if a variant is associated with multiple phenotypes. We took the intersection of the significant genetic variants across the three methods as our list of potential pleiotropic variants. Our colocalization analyses revealed 9 SNP variants associated with at least one disease from both, nervous and circulatory system that cleared the “family-wise Bonferroni” threshold for multivariate and bivariate analyses. Since we were looking at trait pairs here, we ensured that at least one of the two traits had a univariate p-value that cleared the “family-wise Bonferroni” threshold while the other trait had a univariate -log10 p-value of at least 3. Note that we conducted sensitivity analyses for MultiPhen on identified potentially pleiotropic variants in Table 2 when missing values were imputed with 0 and 1 (i.e. treated as controls or cases) in addition to 0.5 and observed no change in significance. To cross-check overlap between methods, we also performed multivariate analysis restricted to a pair of bivariate significant traits for the 9 potentially pleiotropic variants in Table 2 and found 100% consensus between bivariate and multivariate methods. These 9 variants showed strong evidence of colocalization with eQTLs across a host of tissue types (see Supplementary table 1) from the GTEx consortium[33], especially on chromosome 22.

Our results replicated previous association signals as well as detected novel associations. SNP at chromosome 6 position 32569056 (rs9270779) has been directly implicated in autonomic nervous system and has been shown to be associated with heart rate response to exercise in females suggesting it could be pleiotropic for the two disease groupings of interest[39]. Also, the corresponding eGenes for this SNP, *HLA-DRB5* and *HLA-DRB9* from colocalization analysis have been previously shown to be associated with multiple sclerosis. Among the 31 total SNP hits, the one at chromosome 19 position 45416741 (rs438811) is correlated with rs445925 (r^2=0.341), which has been shown to be clinically relevant to cardiovascular phenotypes[40]. This SNP is also located in the *APOC1/APOE* region, which has been shown to be associated with Alzheimer’s disease[41]. Among novel potential pleiotropic variants identified by all three methods *and* colocalization analysis, 6 out of 9 variants locate on chromosome 22, suggesting its potential crucial contribution to the link between cardiovascular and neurological diseases. In particular, the eGene *FBXO7* has been associated with multiple sclerosis[42] as well as heart disease[43]. As part of future work, we will conduct pathway analyses or conditional analyses to have confidence in a singular pleiotropic association or shared biology between these disease groupings.

The limitations of this study are that (1) using only ICD-9 codes instead of both ICD-9 and ICD-10 codes may have reduced the number of cases in our data; (2) the use of disease category instead of disease code as phenotype might have reduced the specificity of detected associations. We are planning to incorporate ICD-9 and ICD-10 codes to define primary phenotypes and examine disease heterogeneity in the future; (3) sample size considerations led to some diagnosis codes being left out of analyses; (4) given our very conservative multiple comparison thresholds, we have likely reported only a fraction of all potential pleiotropic signals, leading to type II errors, and (5) we were unable to investigate how many additional associated variants obtained using bivariate analyses in comparison to univariate and multivariate were "true positives". One way to investigate this would be to test for statistical colocalization on top bivariate analyses hits[27]. However, this necessitates that summary statistics be obtained from independent datasets which was not the case with our data. Replication of these signals in independent cohorts in future can help us address this limitation.

In summary, we provide a framework for future pleiotropy analyses in EHR data. Our work expands the pleiotropy detection framework from univariate methods (e.g. PheWAS) to bivariate and multivariate methods in large-scale real-world EHR data to detect a broader net of potentially pleiotropic signals across cardiovascular and neurological disorders. We also utilize colocalization analyses to enhance our understanding of the influence of gene expression on these potentially pleiotropic variants and consequently on disease risk. In future, we will also try to replicate the partially overlapping SNP signals in independent cohorts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The eMERGE Network was initiated and funded by NHGRI through the following grants:

Phase III: U01HG8657 (Kaiser Permanente Washington (formerly known at GroupHealth) /University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine)

Phase II: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center).

If the project includes data from the eMERGE imputed merged Phase I and Phase II dataset, please also add U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers. And/or The PGRNSeq dataset (eMERGE PGx), please also add U01HG004438 (CIDR) serving as a Sequencing Center.

Phase I: U01-HG-004610 (Kaiser Permanente Washington /University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

References

1. Bruggemans EF Cognitive dysfunction after cardiac surgery: Pathophysiological mechanisms and preventive strategies. Neth Heart J 21, 70–73 (2012).
2. Webb TR et al. Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated with Coronary Artery Disease. Journal of the American College of Cardiology 69, 823–836 (2017). [PubMed: 28209224]
3. Ibanez L et al. Pleiotropic Effects of Variants in Dementia Genes in Parkinson Disease. Front. Neurosci 12, 633–10 (2018). [PubMed: 30254564]
4. Wang Y et al. Genetic overlap between multiple sclerosis and several cardiovascular disease risk factors. Mult Scler 22, 1783–1793 (2016). [PubMed: 26920376]
5. Andreassen OA et al. Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. Mol Psychiatry 20, 207–214 (2014). [PubMed: 24468824]
6. Ritchie MD Large-Scale Analysis of Genetic and Clinical Patient Data. Annu. Rev. Biomed. Data Sci 1, 263–274 (2018).
7. Cotsapas C et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS Genet 7, e1002254(2011). [PubMed: 21852963]
8. Bhattacharjee S et al. A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. The American Journal of Human Genetics 90, 821–835 (2012). [PubMed: 22560090]
9. Vuckovic D et al. MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. Bioinformatics 31, 2754–2756 (2015). [PubMed: 25908790]
10. Chung D et al. GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. PLoS Genet 10, e1004787(2014). [PubMed: 25393678]
11. Turley P et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet 50, 229–237 (2018). [PubMed: 29292387]
12. Korte A et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44, 1066–1071 (2012). [PubMed: 22902788]
13. O’Reilly PF et al. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. PLoS ONE 7, e34861–12 (2012). [PubMed: 22567092]
14. Zhou X et al. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods 11, 407–409 (2014). [PubMed: 24531419]
15. Furlotte NA et al. Efficient Multiple Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed-Model. Genetics 200, 114.171447–68 (2015).
16. Stephens M A Unified Framework for Association Analysis with Multiple Related Phenotypes. PLoS ONE 8, e65245(2013). [PubMed: 23861737]
17. Hackinger S et al. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 7, 170125–13 (2017). [PubMed: 29093210]
18. Verma A et al. PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. The American Journal of Human Genetics 102, 592–608 (2018). [PubMed: 29606303]
19. Pendergrass SA et al. Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet 9, e1003087–26 (2013). [PubMed: 23382687]

20. Bastarache L et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature Biotechnology 31, 1102–1110 (2013).

21. Hall MA et al. Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. PLoS Genet 10, e1004678–33 (2014). [PubMed: 25474351]

22. Verma A et al. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. BMC Medical Genomics 9, 1–7 (2016). [PubMed: 26729011]

23. Denny JC et al. Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. The American Journal of Human Genetics 89, 529–542 (2011). [PubMed: 21981779]

24. Liu Y et al. Powerful Bivariate Genome-Wide Association Analyses Suggest the SOX6 Gene Influencing Both Obesity and Osteoporosis Phenotypes in Males. PLoS ONE 4, e6827–8 (2009). [PubMed: 19714249]

25. Schaid DJ et al. Multivariate generalized linear model for genetic pleiotropy. Biostatistics 5, e553–18 (2017).

26. Schaid DJ et al. Statistical Methods for Testing Genetic Pleiotropy. Genetics 204, 116.189308–497 (2016).

27. Siewert KM et al. Bivariate GWAS scan identifies six novel loci associated with lipid levels and coronary artery disease. bioRxiv 1–27 (2018).

28. Medina-Gomez C et al. Bivariate genome-wide association meta-analysis of pediatric musculoskeletal traits reveals pleiotropic effects at the SREBF1/TOM1L2 locus. Nature Communications 8, 1–10 (2017).

29. Porter HF et al. Multivariate simulation framework reveals performance of multi-trait GWAS methods. Nature Publishing Group 7, 1–12 (2017).

30. Galesloot TE et al. A Comparison of Multivariate Genome-Wide Association Methods. PLoS ONE 9, e95923–8 (2014). [PubMed: 24763738]

31. Solovieff N et al. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics 14, 483–495 (2013).

32. Zhu Z et al. Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. PLoS ONE 13, e0193256(2018). [PubMed: 29494641]

33. Carithers LJ et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation and Biobanking 13, 311–319 (2015). [PubMed: 26484571]

34. Verma S et al. Imputation and quality control steps for combining multiple genome-wide datasets. Frontiers in genetics 5, 370(2014). [PubMed: 25566314]

35. Purcell S et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559–575 (2007). [PubMed: 17701901]

36. Verma A et al. A simulation study investigating power estimates in phenome-wide association studies. BMC Bioinformatics 19, 1–8 (2018). [PubMed: 29291722]

37. Hall MA et al. PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. Nature Communications 1–10 (2017).

38. Giambartolomei C et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet 10, e1004383–15 (2014). [PubMed: 24830394]

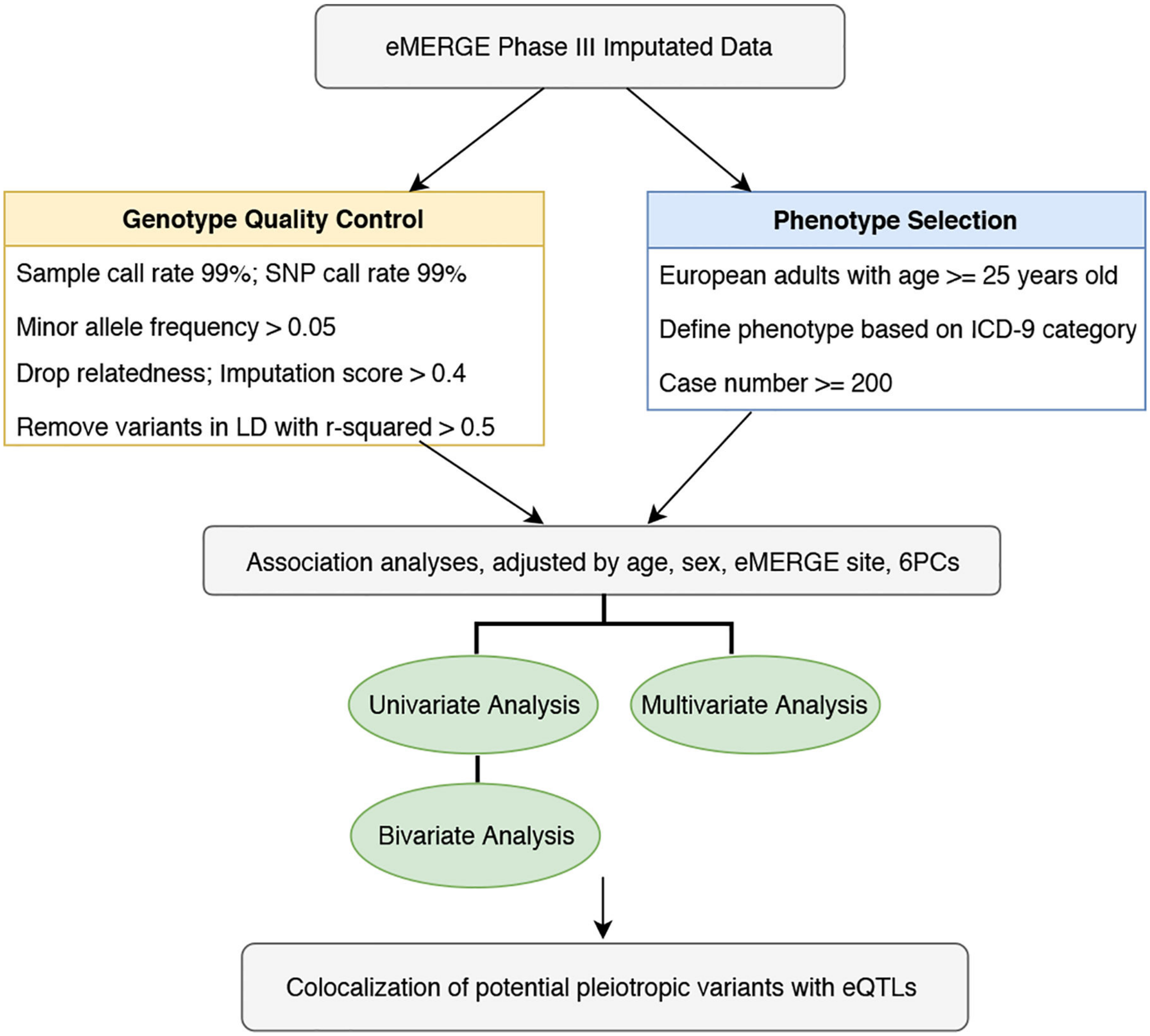
39. Ramirez J et al. Thirty loci identified for heart rate response to exercise and recovery implicate autonomic nervous system Nature Communications 9, 2041–1723 (2018).

40. Allen NB et al. Genetic loci associated with ideal cardiovascular health: A meta-analysis of genome-wide association studies. American Heart Journal 175, 112–120 (2016). [PubMed: 27179730]

41. Bertram L et al. Genome-wide association studies in Alzheimer’s disease. Human Molecular Genetics 18, R137–R145 (2009). [PubMed: 19808789]

42. Burchell VS et al. The Parkinson’s disease–linked proteins Fbxo7 and Parkin interact to mediate mitophagy. Nature Neuroscience 16, 1257–1265 (2013). [PubMed: 23933751]

43. Li Y et al. The Role of Proteasome in Heart Disease. Biochim Biophys Acta. 1809, 141–149 (2011). [PubMed: 20840877]


eMERGE Phase III Imputated Data

Genotype Quality Control

Sample call rate 99%; SNP call rate 99%

Minor allele frequency > 0.05

Drop relatedness; Imputation score > 0.4

Remove variants in LD with r-squared > 0.5

Phenotype Selection

European adults with age >= 25 years old

Define phenotype based on ICD-9 category

Case number >= 200

Association analyses, adjusted by age, sex, eMERGE site, 6PCs

Colocalization of potential pleiotropic variants with eQTLs

Figure 1.

Overview of the analysis plan

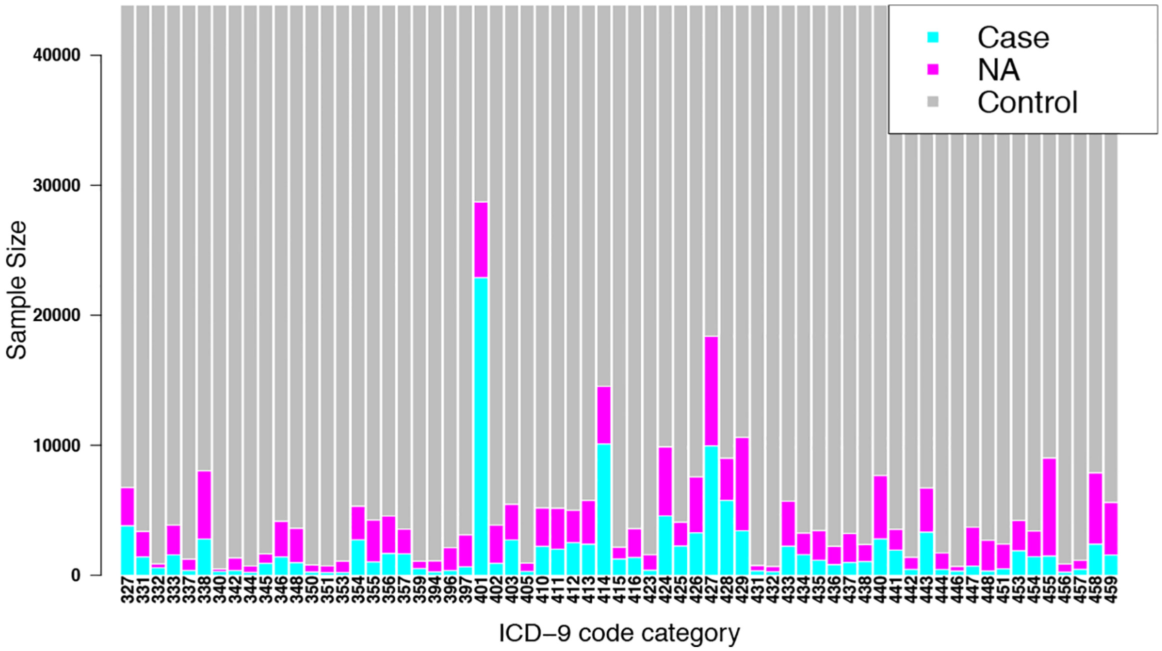


Figure 2.

Sample size distribution for 65 ICD-9 disease categories

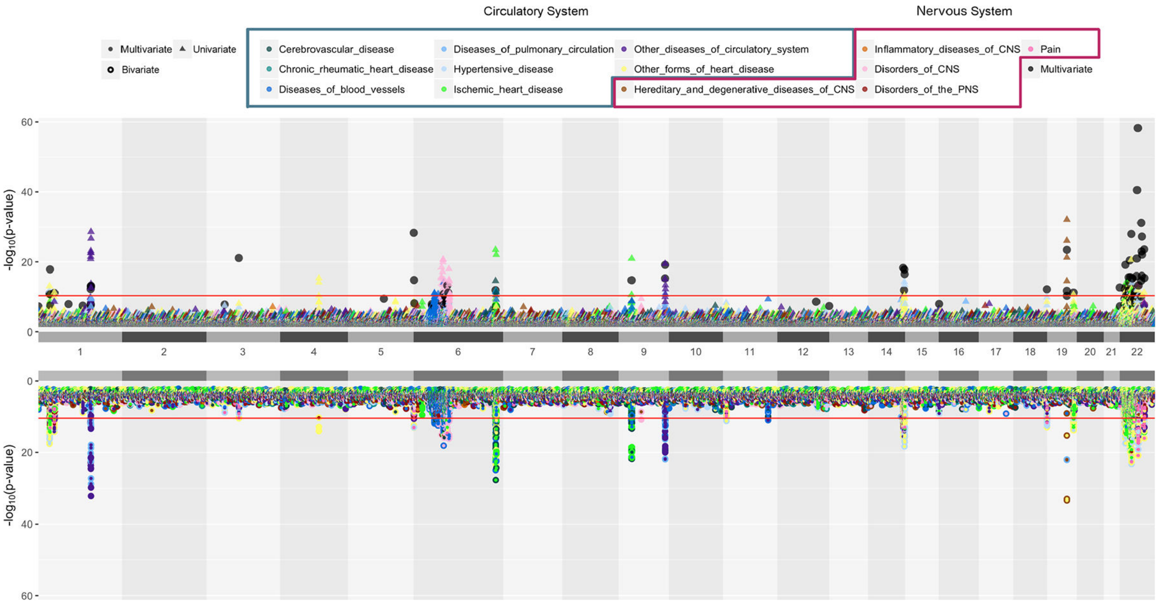


Figure 3.

Univariate, Bivariate and Multivariate Results

A position-by-position comparison of genetic associations for univariate, bivariate and multivariate methods using code modified from Hudson R package (https://github.com/anastasia-lucas/hudson). The horizontal axis represents genomic locations by chromosome and the vertical axis represents $-\log_{10}$(p-value). Colors represent major disease groups of circulatory and nervous systems. The top plot presents univariate results with p-value less than 0.01 in triangles and multivariate results that passed “method-specific Bonferroni” threshold in black dots. The bottom plot present bivariate analysis results in a two-colored circle, denoting the two phenotypes with which a variant is associated with. The red lines in both plots are the “family-wise Bonferroni” threshold.

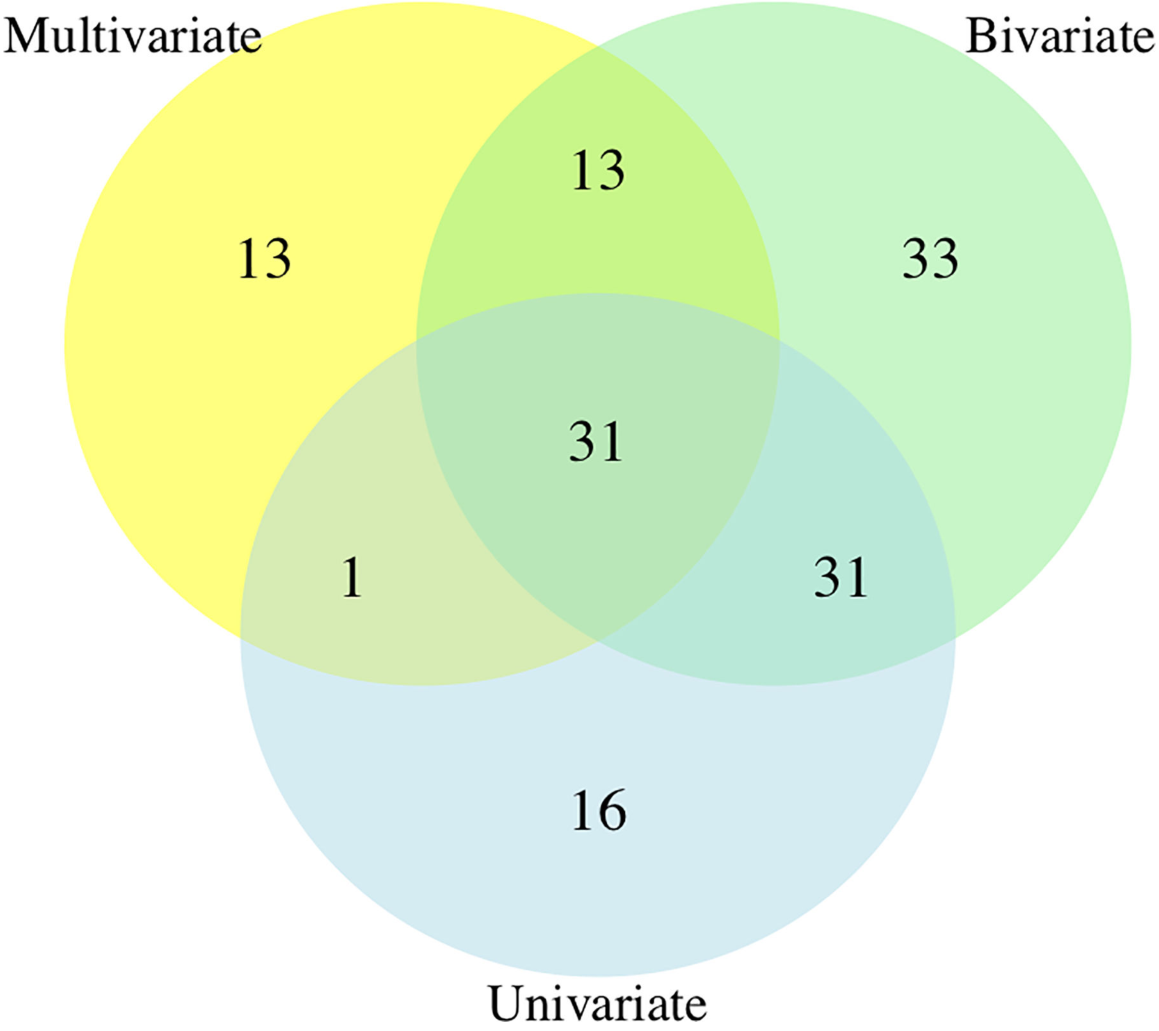


Figure 4.

Venn diagram of the number of SNPs obtained at a “family-wise Bonferroni” threshold

Table 1.

Major group and ICD-9 category of neurological disorders and cardiovascular diseases

	Major Group	ICD-9 Codes
Circulatory System	Chronic rheumatic heart disease	393–398
	Hypertensive disease	401–405
	Ischemic heart disease	410–414
	Diseases of pulmonary circulation	415–417
	Other forms of heart disease	420–429
	Cerebrovascular disease	430–438
	Diseases of blood vessels	440–449
	Other diseases of circulatory system	451–459
Nervous System	Inflammatory diseases of the central nervous system	320–327
	Hereditary and degenerative diseases of the central nervous system	330–337
	Pain	338
	Disorders of the central nervous system	340–349
	Disorders of the peripheral nervous system	350–359

Table 2.

Potential pleiotropic SNPs and their associated disease groups

SNP	Circulatory NeglogP(Uni-variate)	Nervous NeglogP(Uni-variate)	NeglogP (Bi-variate)	NeglogP (Multi variate)	Tissue count	eGenes
1:36822024 rs10796883	Cardiac_dysrhythmias(11.305)	Muscular dystrophies and other myopathies(4.921)	13.247	11.165	35	EVA1B, TRAPPC3
		Other conditions of brain(3.451)	12.030		35	EVA1B, TRAPPC3
		Pain(4.151)	12.363		35	EVA1B, TRAPPC3
	Essential hypertension(9.125)	Muscular dystrophies and other myopathies(4.921)	11.325		35	EVA1B, TRAPPC3
	Heart_failure(10.029)	Muscular dystrophies and other myopathies(4.921)	11.988		35	EVA1B, TRAPPC3
		Pain(4.151)	11.452		35	EVA1B, TRAPPC3
	Hypotension(8.660)	Muscular dystrophies and other myopathies(4.921)	10.699		35	EVA1B, TRAPPC3
6:32569056 rs9270779	Atherosclerosis(14.165)	Multiple sclerosis(6.355)	18.112	10.861	8	HLA-DRB5, HLA-DRB9
		Parkinson's disease(3.196)	15.097		11	HLA-DRB5, HLA-DRB9
	Occlusion_and_stenosis_of_precerebral_arteries(6.355)	Multiple sclerosis(5.913)	10.400		7	HLA-DRB5, HLA-DRB9
	Other peripheral vascular disease(6.355)	Multiple sclerosis(7.442)	11.787		4	HLA-DRB5, HLA-DRB9
14:106995720 rs716 0440	Cardiac_dysrhythmias(11.322)	Muscular dystrophies and other myopathies(4.394)	12.989	18.291	5	IGHV3–53,IGHV4–39, IGHV3–49
		Other conditions of brain(3.726)	12.420		5	IGHV3–53,IGHV4–39, IGHV3–49
		Pain(6.297)	14.259		5	IGHV3–53,IGHV4–39, IGHV3–49
	Essential hypertension(7.451)	Pain(6.297)	10.610		1	IGHV3–49
	Heart_failure(9.038)	Muscular dystrophies and other myopathies(4.394)	10.752		8	IGHV3–53,IGHV4–39, IGHV3–49, HOMER2P1
		Other conditions of brain(3.726)	10.469		6	IGHV3–53,IGHV4–39, IGHV3–49
		Pain(6.297)	12.465		5	IGHV3–53,IGHV4–39, IGHV3–49
	Hypertensive chronic kidney disease(8.116)	Pain(6.297)	11.623		5	IGHV3–53,IGHV4–39, IGHV3–49
	Hypotension(10.278)	Muscular dystrophies and other myopathies(4.394)	11.832		5	IGHV3–53,IGHV4–39, IGHV3–49
		Other conditions of brain(3.726)	11.252		5	IGHV3–53,IGHV4–39, IGHV3–49
		Pain(6.297)	13.004		5	IGHV3–53,IGHV4–39, IGHV3–49
	Ill-defined_descriptions_and_complications_of_ heart disease(7.610)	Pain(6.297)	11.224		1	
22:22876236 rs361535	Other_forms_of_chronic_ischemic_heart_disease(4.985)	Inflammatory and toxic neuropathy(14.211)	14.702	10.424	1	

SNP	Circulatory NeglogP(Uni-variate)	Nervous NeglogP(Uni-variate)	NeglogP (Bi-variate)	NeglogP (Multi variate)	Tissue count	eGenes
22:22947156 rs2097594	Cardiac_dysrhythmias(10.930)	Inflammatory and toxic neuropathy(3.011)	11.236	28.019	1	
		Muscular dystrophies and other myopathies(3.773)	12.116		1	
		Other conditions of brain(3.328)	11.738		1	
		Pain(5.622)	13.348		1	
	Cardiomyopathy(12.330)	Inflammatory and toxic neuropathy(3.011)	12.818		2	GGTLC2
		Muscular dystrophies and other myopathies(3.773)	13.768		2	IGLV3–21, GGTLC2
		Other conditions of brain(3.328)	13.507		1	GGTLC2
		Pain(5.622)	15.503		2	GGTLC2
	Essential_hypertension(10.187)	Muscular dystrophies and other myopathies(3.773)	11.380		2	BCRP4
		Other conditions of brain(3.328)	10.968			
		Pain(5.622)	12.386			
	Heart_failure(20.621)	Inflammatory and toxic neuropathy(3.011)	19.807		2	GGTLC2
		Muscular dystrophies and other myopathies(3.773)	20.963		3	IGLV3–21, GGTLC2
		Other conditions of brain(3.328)	21.000		2	GGTLC2
		Pain(5.622)	22.553		2	GGTLC2
	Hypertensive_chronic_kidney_disease(9.331)	Muscular dystrophies and other myopathies(3.773)	10.760		2	GGTLC2
		Pain(5.622)	12.119		2	GGTLC2
	Hypotension(9.778)	Muscular dystrophies and other myopathies(3.773)	10.883		2	GGTLC2
		Other conditions of brain(3.328)	10.491		2	GGTLC2
		Pain(5.622)	12.026		2	GGTLC2
	Ill-defined_descriptions_and_complications_of_heart_disease(10.665)	Inflammatory and toxic neuropathy(3.011)	10.863		2	GGTLC2
		Muscular dystrophies and other myopathies(3.773)	11.703		2	GGTLC2
		Other conditions of brain(3.328)	11.478		2	GGTLC2
		Pain(5.622)	13.385		2	GGTLC2
	Other_diseases_of_endocardium(10.340)	Inflammatory and toxic neuropathy(10.340)	11.032			
		Muscular dystrophies and other myopathies(10.340)	11.844			
		Other conditions of brain(10.340)	11.617			
		Pain(5.622)	13.627			

SNP	Circulatory NeglogP(Uni-variate)	Nervous NeglogP(Uni-variate)	NeglogP (Bi-variate)	NeglogP (Multi variate)	Tissue count	eGenes
	Other forms of chronic ischemic heart dis ease(11.873)	Inflammatory and toxic neuropathy(11.873)	11.335			
		Muscular dystrophies and other myopathies(11.873)	12.690			
		Other conditions of brain(11.873)	12.530			
		Pain(5.622)	14.168			
22:25420792 rs13056641	Cardiac_dysrhythmias(9.528)	Inflammatory and toxic neuropathy(4.159)	10.817	40.505	11	KIAA1671, SGSM1, CRYBB2, CRYBB3, IGLL3P
		Organic sleep disorders(4.166)	10.687		1	IGLL3P
		Pain(4.590)	11.247		6	KIAA1671, IGLL3P
	Essential_hypertension(12.162)	Inflammatory and toxic neuropathy(4.159)	12.620		16	KIAA1671, SGSM1, CRYBB2, CRYBB3, IGLL3P, BCRP3
		Organic sleep disorders(4.166)	12.521		1	IGLL3P
		Pain(4.590)	13.284		7	KIAA1671, IGLL3P
22:25436904 rs1040421	Angina pectoris(3.067)	Pain(13.338)	15.015	58.239	7	KIAA1671, SGSM1, IGLL3P
	Atherosclerosis(5.075)	Pain(13.338)	15.580		8	KIAA1671, SGSM1, IGLL3P
	Cardiac dysrhythmias(11.931)	Pain(13.338)	20.872		7	KIAA1671, SGSM1, IGLL3P
	Cardiomyopathy(4.939)	Pain(13.338)	15.904		8	KIAA1671, SGSM1, IGLL3P
	Conduction disorders(5.764)	Pain(13.338)	16.372		5	KIAA1671, SGSM1, IGLL3P
	Essential_hypertension(10.303)	Pain(13.338)	19.175		8	KIAA1671, SGSM1, IGLL3P
	Heart failure(7.101)	Pain(13.338)	17.129		8	KIAA1671, SGSM1, IGLL3P
	Hypertensive chronic kidney disease(7.426)	Pain(13.338)	17.404		8	KIAA1671, SGSM1, IGLL3P
	Hypotension(6.693)	Pain(13.338)	16.037		4	KIAA1671, SGSM1, IGLL3P
	Other diseases of endocardium(5.845)	Pain(13.338)	16.677		4	KIAA1671, SGSM1, IGLL3P
22:28250172 rs1997739	Cardiac_dysrhythmias(10.517)	Pain(4.966)	12.443	22.064	19	ZNRF3, TTC28-AS1
22:33079917 rs5749490	Cardiac_dysrhythmias(11.280)	Hereditary and idiopathic peripheral neuropathy(3.04 9)	11.884	23.601	9	FBXO7, SLC5A4-AS1
		Inflammatory and toxic neuropathy(3.958)	12.254		2	FBXO7, SLC5A4-AS1
		Mononeuritis of lower limb and unspecified site(3.1 53)	12.242		2	FBXO7, SLC5A4-AS1
		Pain(8.424)	16.011		9	FBXO7, SLC5A4-AS1
	Hypertensive chronic kidney disease(6.449)	Pain(8.424)	12.064		9	FBXO7, SLC5A4-AS1
	Hypertensive heart disease(4.191)	Pain(8.424)	10.592		10	FBXO7, SLC5A4-AS1
	Hypotension(8.197)	Pain(8.424)	12.959		3	FBXO7, SLC5A4-AS1

Notes: We left as missing in the table any eGene (Ensembl gene ID from GTEx) that did not have an HGNC symbol counterpart.