PLOS ONE

RESEARCH ARTICLE

Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases

Anurag Verma[1,2], Anna O. Basile[2], Yuki Bradford[1], Helena Kuivaniemi[3,4], Gerard Tromp[3,4], David Carey[3], Glenn S. Gerhard[5], James E. Crowe, Jr[6], Marylyn D. Ritchie[1,2], Sarah A. Pendergrass[1]*

1 Biomedical and Translational Informatics Program, Geisinger Health System, Danville, Pennsylvania, United States of America, **2** Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, Pennsylvania, United States of America, **4** Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa, **5** Department of Medical Genetics and Molecular Biochemistry, Temple University School of Medicine, Philadelphia, Pennsylvania, United States of America, **6** The Vanderbilt Vaccine Center, Vanderbilt University, Nashville Tennessee, United States of America

* spendergrass@geisinger.edu

Citation: Verma A, Basile AO, Bradford Y, Kuivaniemi H, Tromp G, Carey D, et al. (2016) Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. PLoS ONE 11(8): e0160573. doi:10.1371/journal.pone.0160573

Editor: Yong-Gang Yao, Kunming Institute of Zoology, Chinese Academy of Sciences, CHINA

Received: April 19, 2016

Accepted: July 16, 2016

Published: August 10, 2016

Data Availability Statement: All relevant summary data are within the paper and Supporting Information files. The authors do not make the entirety of the genetic data and phenotypic data of these biorepositories publicly available as these are de-identified electronic health record data linked to genetic information. PheWAS uses a wide range of phenotypic information linked to that genetic data, and they do not want to risk any patient re-identification.

Abstract

We performed a Phenome-Wide Association Study (PheWAS) to identify interrelationships between the immune system genetic architecture and a wide array of phenotypes from two de-identified electronic health record (EHR) biorepositories. We selected variants within genes encoding critical factors in the immune system and variants with known associations with autoimmunity. To define case/control status for EHR diagnoses, we used International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes from 3,024 Geisinger Clinic MyCode® subjects (470 diagnoses) and 2,899 Vanderbilt University Medical Center BioVU biorepository subjects (380 diagnoses). A pooled-analysis was also carried out for the replicating results of the two data sets. We identified new associations with potential biological relevance including SNPs in tumor necrosis factor (*TNF*) and ankyrin-related genes associated with acute and chronic sinusitis and acute respiratory tract infection. The two most significant associations identified were for the *C6orf10* SNP rs6910071 and "rheumatoid arthritis" (ICD-9 code category 714) ($p_{METAL} = 2.58 \times 10^{-9}$) and the *ATN1* SNP rs2239167 and "diabetes mellitus, type 2" (ICD-9 code category 250) ($p_{METAL} = 6.39 \times 10^{-9}$). This study highlights the utility of using PheWAS in conjunction with EHRs to discover new genotypic-phenotypic associations for immune-system related genetic loci.

Funding: The BioVU data used in the analyses described were obtained from Vanderbilt University Medical Centers BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. Genome-wide genotyping was funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. The MyCode biobanking and genotyping at Geisinger Clinic was funded by Pennsylvania Commonwealth Universal Research Enhancement Program, the Ben Franklin Technology Development Fund of PA, Grants from the NIH (P30DK072488, R01DK088231 and R01DK091601), the Geisinger Clinical Research Fund, and a Grant-In-Aid from the American Heart Association.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Autoimmune diseases affect about 5% of the population and can lead to chronic inflammation targeting specific tissues [1]. The most common autoimmune diseases, such as rheumatoid arthritis (RA), multiple sclerosis, and type 1 diabetes mellitus (T1DM), have overlapping clinical, epidemiological and therapeutic features, but their genetic underpinnings and pathogenesis are still not fully understood [2]. Genome Wide Association Studies (GWAS) have discovered over 200 genetic loci associated with autoimmune diseases [2], elucidating biological pathways and potential drug targets for autoimmune disorders [3]. Comparison of results across GWAS shows a series of single nucleotide polymorphisms (SNPs) associated with multiple autoimmune diseases, suggesting the existence of variance in immune traits and pleiotropy [3]. For example, multiple genetic variants that reside within the region encompassing the human leukocyte antigen (HLA) system have been associated with several autoimmune diseases [4]. Although GWAS have identified multiple autoimmune disease susceptibility loci, the biological relationship between genetic variation within these loci and disease status has not been well characterized.

While genetic variation in immune function and inflammation contributes to susceptibility to autoimmune conditions, this variation may also impact a variety of other diseases and diagnoses. The immune system serves as a major defense network in fighting disease and infection. Genetic variation in immune function has been found to contribute to disease susceptibility in multiple classes of disorders [3]. For example, monocyte-specific expression quantitative trait loci (eQTLs) have been identified for genetic variants associated with neurodegenerative disorders such as Parkinson's and Alzheimer's diseases [5]. As a manifestation of immune function, inflammation also plays an important role in conditions beyond contagious or autoimmune diseases. For instance, inflammation has been implicated in multiple disorders including vascular diseases such as atherosclerosis [6] and congestive heart failure [7], neuropsychiatric diseases like autism [8], as well as metabolic traits and disorders such as obesity [9] and type 2 diabetes (T2DM) [10].

To examine potential associations across many phenotypes, Phenome-wide association studies (PheWAS) have been developed as a complementary approach to GWAS, using all available phenotypic information and genetic variation in order to estimate the association between genotype and phenotype [11]. PheWAS are dependent on comprehensive phenotypic information on large numbers of individuals; PheWAS to date have used electronic health record (EHR) International Classification of Diseases (ICD-9) billing codes to define case-control statuses for multiple diagnoses [12], data from epidemiological studies with hundreds to thousands of phenotypic measurements [13][11], as well as clinical trials data [14]. The PheWAS framework of evaluating the association between a wide array of phenotypes and markers permits the study of pleiotropy, compared to the GWAS framework of investigating association between a single trait and genetic markers, except when comparing results from multiple separate GWAS [15]. In this PheWAS, we used variants in immune-related genes which provided an opportunity to explore the association between immune system SNPs and phenotypes beyond specific autoimmune and immune system traits, such as diagnoses that may have an immune system involvement but are not specifically classified as an autoimmune/immune system trait.

The goal of this study was to identify associations between selected SNPs with known or possible associations with autoimmune disease and the immune system and a variety of diagnoses, evaluating and contrasting results across two separate EHR systems. We performed our PheWAS analysis using SNPs within genes encoding critical factors for the immune system and SNPs with known associations with autoimmunity, including a series of SNPs also found

on ImmunoChip, an array designed by investigators of 11 autoimmune and inflammatory diseases [16,17]. To explore associations between these SNPs and diagnoses, we used ICD-9 diagnosis codes to define case/control status from two sites within the Electronic Medical Record and Genomics (eMERGE) Network: Geisinger MyCode® and Vanderbilt BioVU. Highly significant results were investigated within the individual datasets, and replication of associations was also sought across the two different bio-repositories. The results of this study also demonstrate cross-phenotype associations that may be due to pleiotropy and identified complex networks that exist between immune related genetic variants and many different diagnoses.

Methods

Data Sets

We used de-identified EHR biorepository data linked to genotypic data and ICD-9 diagnosis code data from two sites in the eMERGE Network: Geisinger Health System's MyCode® and Vanderbilt University Medical Center's BioVU [18]. The MyCode dataset had a total of 3,024 individuals and the BioVU dataset had 2,899 individuals available for the study with *both* phenotypic and genotypic data (Table 1). Because a majority of subjects in MyCode® were of European ancestry (EA), we selected only EA subjects to seek replication with the BioVU data [19].

The Geisinger biorepository has had both general and targeted recruitment for specific diseases, such as obesity and abdominal aortic aneurysms (AAA). BioVU has consented using an opt-out approach, where individuals with discarded blood may or may not be added to the biorepository unless they indicate they would like to opt-out of BioVU [20]. Thus BioVU has no pre-selection for individuals with a specific disease phenotype.

Genotyping, Imputation & Quality Control

We summarize the genotyping, imputation, and quality control procedures in S1 Fig. Geisinger MyCode® subjects were genotyped using the Illumina HumanOmniExpress-12 v1.0 array, a total of 729,078 SNPs. Genotyping of BioVU subjects was performed using the Illumina 660 Quad array, a total of 558,590 SNPs. We used imputation for improved genomic coverage and overlap between datasets of immune related variants. We performed imputation using the IMPUTE2 algorithm [21] after phasing with SHAPEIT2 [22] using the 1,000 Genomes cosmopolitan reference panel, resulting in a total of 38,054,243 SNPs in 3,111 samples for MyCode® and 38,041,351 SNPs in 3,375 samples for BioVU [23].

Genotype Quality Control (QC) procedures were performed prior to association testing using the R programming statistical package [24] and PLINK software [25]. QC was performed on each dataset separately. The first step was to filter out the SNPs with poor imputation

Table 1. Summary of Data Sets Used for the Study.

EHR Site	Total Sample Size	% Male	Median Age (in decade)	Case Size Range	Genotyping Platform	Number of SNPs Pre-imputation	Number of SNPs Post-imputation	Number of SNPs after filtering	Number of Diagnosis Codes
Geisinger MyCode®	3024	53.0	40	Min = 11; Max = 1898; Median = 32	Illumina Human OmniExpress	729,078	38,054,243	95,448	477
Vanderbilt BioVU	2899	45.4	60	Min = 11; Max = 1056; Median = 31	Illumina 660	558,980	38,041,351	87,690	380

For additional information on the study design, see Figs 1 and S1.

doi:10.1371/journal.pone.0160573.t001

quality; SNPs with imputation quality scores > 0.9 were used for further analyses. Data were filtered further for 99% genotype and sample call rates and minor allele frequency (MAF) threshold of 1%. Also, related samples were removed using Identity by Descent (IBD) kinship coefficient estimates. We also performed principal component analysis (PCA), determining principle components to use to correct for population differences within the EA of these datasets, as association results for immune system genes in particular can be particularly affected by population substructure. After QC, the genotypic data consisted of 4,636,178 SNPs and 3,029 samples from MyCode® and 4,163,988 SNPs and 2,900 samples from BioVU.

Phenotype Data

To define case-control status for each ICD-9 code, a MySQL database was used to assemble the phenotypic data, consisting of 6,525 ICD-9 codes from the MyCode® dataset and 1,206 ICD-9 codes from BioVU. A case was defined as an individual with more than three instances of a specific ICD-9 code. Controls were defined as individuals not meeting the case criteria. More than ten case subjects were required for inclusion of a diagnosis in our study. Using these criteria there, were 3,024 samples and 477 ICD-9 codes in MyCode®, and 2,899 samples and 380 ICD-9 codes in BioVU. For replication of results across the two studies, there were a total of 50 exact ICD-9 code matches (i.e. 3, 4 and 5 digit ICD-9 code) across both datasets, and a total of 186 ICD-9 category (i.e. three digit ICD-9 code) matches across both datasets.

Selection of SNPs for the Study

While we used genetic data from Illumina arrays, we focused on SNPs related to the immune system, and chose from our array data SNPs present on ImmunoChip (Illumina) or known to be involved in the immune system within a specific set of genes (see S1 Table for a list of these 34 genes). The ImmunoChip is a custom genotyping array designed by Illumina with 195,806 SNPs for performing deep replication of associations with major autoimmune and inflammatory diseases including fine mapping of GWAS loci covering 11 major autoimmune diseases (*e.g.*, T1DM, autoimmune thyroid disease, celiac disease and multiple sclerosis), seronegative diseases (*e.g.*, ulcerative colitis, Crohn's disease, and psoriasis), and rheumatic diseases (*e.g.*, RA, ankylosing spondylitis and systemic lupus erythematosus) [17]. Also included on ImmunoChip are all the previously confirmed GWAS SNPs for which probes could be designed using data from the 1,000 Genomes Project [16]. From this study, we selected only SNPs from our genome-wide array data that were on the ImmunoChip array or within the genes we identified for involvement in the immune system. The SNP filtering process is shown in Fig 1.

Biofilter was the tool used to generate the list of SNPs for association testing. Biofilter is a software tool with an extensive database containing biological knowledge from publicly available repositories of biological data that can be used to annotate genomic information, as well as filter genomic information based on specific criteria [26,27]. First Biofilter 2.1 was used to annotate the post-QC SNPs of this study with gene information for any SNPs within Entrez-defined gene boundaries. Then Biofilter was used with the genotypic data from each EHR site to filter SNPs, maintaining only those SNPs matching within the 34 genes selected for their known involvement in the immune system (S1 Table). The SNP filtering step resulted in 95,448 SNPs from MyCode® and 87,690 SNPs from BioVU, with a total of 76,861 SNPs overlapping across the two datasets available for association testing (Table 1).

Association Testing and Identifying Replication

In both datasets, separately, associations were calculated using logistic regression with models adjusted for sex, age, and the first five principal components. In the MyCode® dataset, using

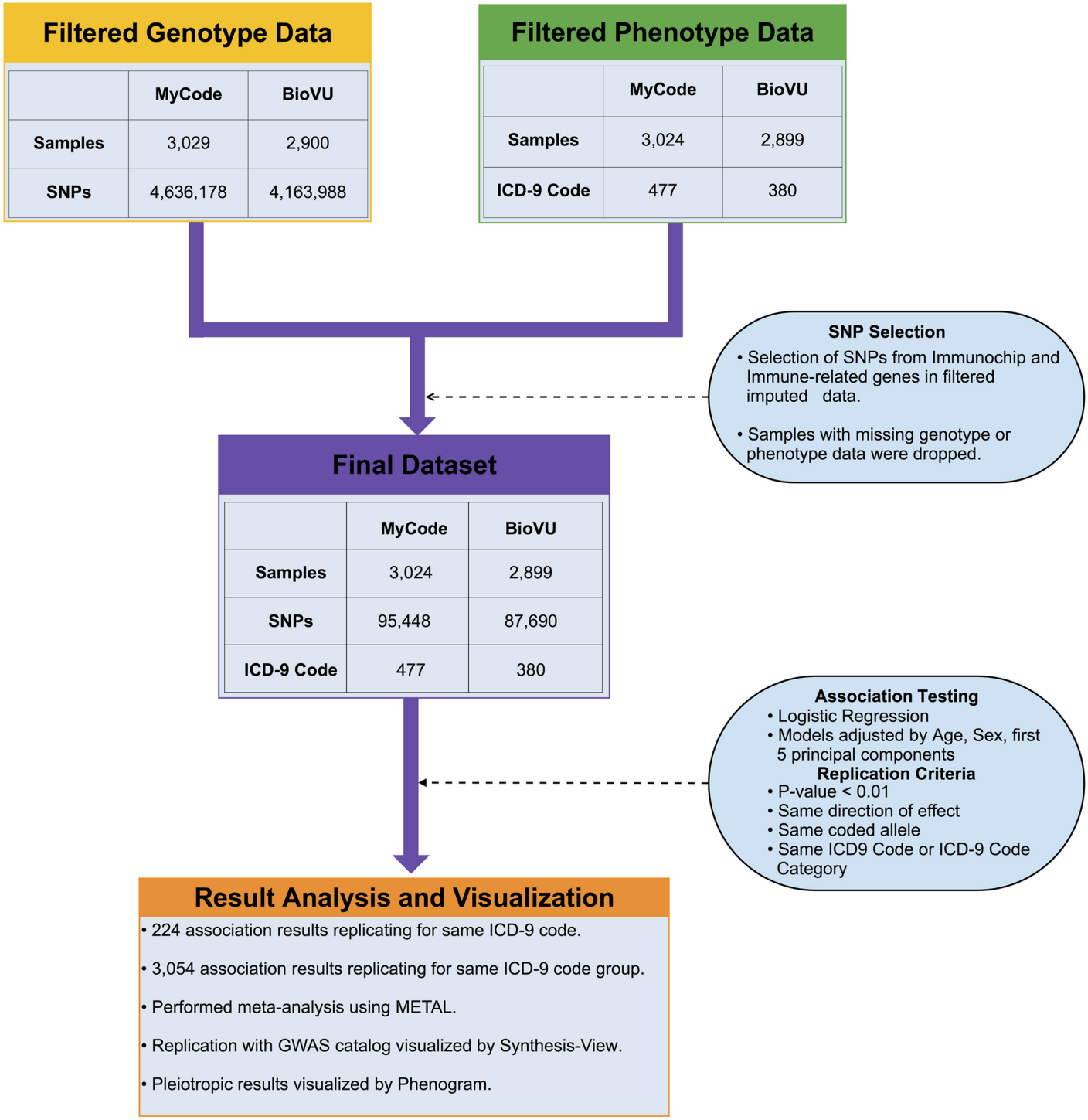

Fig 1. Overview of PheWAS with Immune Variants. This flow chart provides an overview of the steps taken to perform PheWAS between immune variants and ICD-9 diagnosis codes. The final testing dataset (purple) was formed by selecting SNPs from our array data that also exist on Immunochip and/or are within immune-related genes (yellow) and removing samples with missing genotypic or phenotypic data (green). Comprehensive associations were calculated between all final dataset SNPs and ICD-9 code based case/control status using logistic regression, with all models adjusted for age, sex and first five principal components. Replication was sought following both an exact ICD-9 code and a category ICD-9 code approach following the specified criteria. Pooled analysis was performed for both approaches using METAL. See S1 Fig for the full workflow from imputation through quality control, association testing, and replication for this study.

doi:10.1371/journal.pone.0160573.g001

logistic regression, the strength and significance of associations were evaluated between 95,448 SNPs and 477 clinical diagnoses. There were a total of 366,468 associations with $p < 0.01$. In the BioVU dataset, association testing was performed on 87,690 SNPs and 380 phenotypes. There were a total of 261,346 associations with $p<0.01$. We also compared our results to a Bonferroni corrected p-value threshold. A LD pruning approach was used to account for correlation between the SNPs and identified independent SNPs at $r^2 = 0.3$ [28]. For MyCode and BioVU association testing, the Bonferroni threshold was 4.73×10^{-9} [0.05/(22,138 x 477)] and 6.07×10^{-9} [0.05/(21,673 x 380)], respectively. There was only one result that passed the conservative Bonferroni threshold in MyCode and none of the results passed threshold in BioVU.

A MySQL database was used to organize all association results. This included inspection of results in single datasets where the ICD-9 code existed only in one or the other dataset. S2 Table shows the results of MyCode® and BioVU with $p<1 \times 10^{-4}$, where we could not seek replication across both datasets, as the ICD-9 codes were only specific to each study. For replication we used two approaches. In the first approach, association results replicating across both datasets for the same SNP and *exact ICD-9 diagnosis code*, with same direction of effect of the association were used. In the second approach, we also used the database to seek replication across both datasets, for the same SNP and *ICD-9 code category*, with the same direction of effect of the association. ICD-9 codes classify diagnoses; there are three digit ICD-9 codes that specify disease categories (e.g. code 405 for "secondary hypertension") that can be further subdivided using multiple four or five digit sub ICD-9 codes (e.g. 405.1 for "benign secondary hypertension", 405.11 "benign renovascular hypertension"). Wide variation exists across institutions in the way specific ICD-9 codes are applied, although three digit ICD-9 categories are used more consistently for diagnoses from institution to institution. We therefore analyzed results based on replication requiring the *exact* ICD-9 code used in the association (*i.e.* three, four, or five digit sub ICD-9 codes), as well as evaluating results based on replication requiring only the same three digits of the ICD-9 code category. S3 and S4 Tables show all of the results where the criteria for replication were met for the same ICD-9 code, and for the same ICD-9 code category, respectively.

We also annotated SNPs in replicating associations between BioVU and MyCode® with information from the NHGRI-EBI GWAS catalog [29,30] and GRASP [31,32], thus identifying any previously reported associations for these SNPs. We used a p-value threshold of 1×10^{-5} (default for NHGRI-EBI GWAS catalog) on the associations reported in both the sources.

Focusing on Immune System and Autoimmune Traits

We further explored results for associations with phenotypes/diagnoses more closely linked to the immune system or autoimmune disease. Thus, we filtered the results presented in S4 Table (all results where the criteria for replication were met for the same ICD-9 code category) for only ICD-9 categories related to the immune system or autoimmune disease by having three individuals identify any ICD-9 categories for removal that were too broad (such as "general symptoms, not otherwise specified", etc.), cancer related diagnoses, and diagnoses clearly related to accident or surgery (such as "hypothyroidism due to ablation"). In this way, we retained ICD-9 codes describing autoimmune reactions or clearly influenced by immune system variation. Table 2 lists the ICD-9 categories and descriptions selected through this process, as well as the genes in which the variants are located. This process resulted in 409 associations (S5 Table).

Visualization Tools

We used Synthesis-View [33], PhenoGram [34], and Cytoscape [35], to visualize the results. Synthesis-View was used to visualize the SNP-phenotype associations and to plot associations

Table 2. Immune-Related ICD-9 categories selected for further analysis.

ICD-9 General Classification	ICD-9 Code Category (Code: Category Description)	Nearest Genes
Endocrine, nutritional and metabolic diseases, and immunity disorders	250: Diabetes mellitus, Type 1	*LOC645266*, *MTCO3P1*, *HLA-DRA*, *HLA-DQB2*, *HLA-DOB*, *DDC*, *DDC*, *LOC100129427*, *HLA-DMA*, *HCG23*, *C6orf10*, *MIR588*, *CAST*, *EPHA5*, *LOC645321*, *SERPINB11*, *KC6*, *FLJ30679*, *PRELID1P1*, *RPS4XP9*, *THEMIS*, *HIST1H1T*
	273: Disorders of plasma protein metabolism	*LOC100996339*, *ESRRG*
Diseases of the nervous system	331: Other cerebral degenerations	*QRSL1*
	340: Multiple sclerosis	*MYT1L*, *IRF4*
	357: Inflammatory and toxic neuropathy	*GTDC2*, *RAB38*, *LOC100129160*
Diseases of the sense organs	373: Inflammation of eyelids	*USH2A*, *MAP4K4*
Diseases of the respiratory system	461: Acute sinusitis	*CNTNAP2*, *ANKS1A*
	465: Acute upper respiratory infections of multiple or unspecified sites	*RPL23AP54*, *ZZEF1*, *ANK3*
	466: Acute bronchitis and bronchiolitis	*CLSTN2*, *LOC100129949*
	472: Chronic pharyngitis and nasopharyngitis	*DAP3P2*, *DCAF17*, *LOC100287243*, *METTL8*, *TAF3*, *KIAA1217*
	473: Chronic sinusitis	*ANK3*, *TRPS1*, *KSR1*, *RBM17*
	477: Allergic rhinitis	*KCNK3*,*VAV3-AS1*,*PADI4*,*MED13*,*KIRREL3*,*TOX*,*FTLP7*
	482: Other bacterial pneumonia	*TLR6*, *LOC645481*
	491: Chronic bronchitis	*SDC4*,*RPS2P9*, *SYS1-DBNDD2*,*PITX2*,*SYS1*,*SYS1-DBNDD2*,*XCR1*
	492: Emphysema	*GABRA4*
	493: Asthma	*TPD52L1*, *CAST*, *WDR11*
	515: Postinflammatory pulmonary fibrosis	*SLC16A10*,*FRMD6-AS2*,*LOC100506923*,*ADIPOR1*
Diseases of the digestive system	556: Ulcerative colitis	*BTF3L4P3*
	571: Chronic liver disease and cirrhosis	*RSBN1*, *RFX3*, *CTIF*, *PHTF1*
	577: Diseases of pancreas	*LCE1B*, *NTRK3-AS1*
Diseases of the genitourinary system	584: Acute renal failure	*FAM205B*, *LACC1*, *DGKZP1*, *E2F3*, *FAM205A*
	585: Chronic kidney disease	*STXBP4*, *LOC100996324*, *CCDC148*, *RPL13AP25*, *BMP2*, *RRP15*
	586: Renal failure	*PLK2*, *LOC100507162*
	595: Cystitis	*SDK1*, *DNTTIP2*
Diseases of the skin and subcutaneous tissue	692: Contact dermatitis and other eczema	*MB21D2*
	695: Erythematous conditions	*SRRM4*, *PRDM15*, *VDR*, *INTS6*, *FGFR3P3*
Diseases of the musculoskeletal system and connective tissue	714: Rheumatoid arthritis and other inflammatory polyarthropathies	*MTCO3P1*, *HLA-DRA*, *NOTCH4*, *TNIP1*, *IL6*, *HLA-DRB1*, *C6orf10*, *HLA-DRB9*
	715: Osteoarthrosis and allied disorders	*BRD2*, *SEMA6A*, *FLJ42102*, *RNY4P22*, *TNFSF8*, *KIAA1919*, *BACH2*, *AKT3*, *SDHAP3*, *ZBTB38*, *CAMK1G*, *PRDM1*, *U2SURP*, *RETNLB*, *ANXA6*
	716: Other and unspecified arthropathies	*IL23R*, *GPX3*, *TRNAS13*, *ZNF192P2*, *AGPAT4*, *QKI*, *ENTPD7*, *GJD4*
	719: Other and unspecified disorders of joint	*TCF4*, *LOC100288337*, *TBX3*

doi:10.1371/journal.pone.0160573.t002

matching previously reported associations in the NHGRI-EBI GWAS catalog and GRASP. PhenoGram was used to visualize potentially pleiotropic SNPs by creating a chromosomal ideogram with lines denoting SNP locations and colored circles depicting phenotypes associated with those SNPs.

We also used Cytoscape 3.0 to visualize network diagrams from the significant PheWAS results. To produce the Cytoscape plots, we first used Biofilter to annotate the PheWAS result SNPs with the gene, or the closest gene to the SNP. We then used Biofilter to obtain the gene annotations from Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [36].

Functional Annotation

To obtain functional information, we annotated the SNPs used in this study with HaploReg V2 [37] and SNP and CNV Annotation Database (SCAN) [38]. HaploReg provides functional annotation of SNPs within LD blocks, and includes information on chromatin state in multiple cell types, regulatory motif alterations and sequence conservation. SCAN provides information from eQTL experiments with a list of genes whose expression is affected by the given SNP in Caucasian (CEU) and Yoruba (YRI) populations. We used expression data specific to the CEU population from SCAN.

Results

PheWAS in Two EHR Datasets Using ICD-9 Codes

Fig 1 provides an overview of our study to identify comprehensive associations between immune system related variants and ICD-9 based case/control diagnoses within MyCode® and BioVU data (further details available in Methods). S1 Fig shows the full workflow, from imputation through quality control, association testing and replication for this study. Table 1 provides a summary of the datasets used. Only subjects of EA with both phenotypic and genotypic data within these sites were selected for discovery and replication analyses.

Evaluating the two EHR datasets separately yielded a total of 366,468 associations ($p < 0.01$) between SNPs and ICD-9 codes in the MyCode® dataset and 261,346 associations ($p < 0.01$) in the BioVU dataset (see details on datasets in Methods). The most significant association in MyCode® passing our Bonferroni threshold of 4.73×10^{-9} was between rs41272317 in the *ACAD11* gene on chromosome 3 and "symptoms concerning nutrition metabolism and development" (ICD-9 code 783.21), with $p = 8.15 \times 10^{-12}$ ($\beta = 2.68$, cases = 45, controls = 2,979). In the BioVU PheWAS results, the most significant association result was between SNP rs870769 on chromosome 8 and "transient cerebral ischemia" (ICD-9 code 435), with $p = 7.41 \times 10^{-8}$ ($\beta = 2.2$, cases = 32, controls = 2,861), which did not pass our Bonferroni threshold. While these are most significant associations within each dataset independently, they did not replicate across both MyCode and BioVU.

Geisinger's genotyped cohort from MyCode® had a large number of cases with AAA or obesity, warranting further exploration for immune variation related to AAA and obesity. Inflammatory processes have been implicated in AAA as well as obesity [39]. We found an association within MyCode between "abdominal aortic aneurysm without mention of rupture" (ICD-9 code 441.4) and rs11084402 on chromosome 19 ($p = 2.22 \times 10^{-5}$, $\beta = 0.36$, cases = 778, controls = 2,246). There were no associations for AAA in the BioVU dataset.

We further explored results with diagnosis codes meeting the criteria for inclusion, but for associations where we could not seek replication because that ICD-9 code was not present in the other dataset. There were a total of 50 exact ICD-9 code matches across both datasets, and a total of 186 exact ICD-9 category matches across both datasets, which placed limitations on seeking replication across the two sets. For the results where we could not seek replication due to the ICD-9 codes not being present in the other dataset, we focused our attention on association results with the highest case numbers (hundreds to thousands of cases) with $p < 1 \times 10^{-4}$. While this p-value cutoff is less stringent than our Bonferroni cutoff, we chose a more exploratory p-value cutoff focused on the most highly suggestive and powered associations from the

PheWAS. S2 Table lists these results for MyCode® and BioVU with $p<1x10^{-4}$. In MyCode®, we found associations between SNPs and metabolic disorder traits, including with "essential primary hypertension" (ICD-9 code 401.9), "other and unspecified hyperlipidemia" (ICD9-code 272.4), and "morbid obesity" (ICD-9 code 278.01). Interestingly, there were related metabolic disorder traits in BioVU, but these did not replicate the MyCode® results. For example, among the highest case numbers in BioVU for $p<1x10^{-4}$ there were associations between SNPs and diagnoses including "essential hypertension" (ICD9-code 401), and "disorder of lipoid metabolism" (ICD-9 code 272).

PheWAS Results with Replication

To identify further robust associations, we also sought replication of results across the two datasets using two approaches: seeking results for the same SNP and *same ICD-9 code category (i.e.* truncating the ICD-9 code for each case/control status to the three digit ICD-9 code) with $p<0.01$ and the same direction of association, and also seeking results for the same SNP and *exact ICD-9 code* (*i.e.* the exact ICD-9 case/control status for each association, varying from three to five digits) with $p<0.01$ and the same direction of association (Fig 1). There were a total of 224 associations with exact ICD-9 code replication across the two datasets with $p<0.01$, for the same SNP, coded allele, and direction of effect (S3 Table). Of the 224 replicated associations with the exact same ICD-9 code, the most significant association in BioVU that replicated in MyCode® was between "soft tissue disorders" (ICD-9 code 729.1) and the *PLA2G2E* SNP rs1108975 with $p_{BioVU} = 3.29 \times 10^{-6}$ (Case-Control = 43/2,853), replicating in the MyCode® data with $p_{MyCode} = 3.29 \times 10^{-3}$ (Case/Control = 136/2,888). The most significant association in the MyCode® data, replicating in the BioVU dataset was between SNP rs11869607 and the diagnosis "deficiency anemias" (ICD-9 281.1) with $p_{MyCode} = 5.98 \times 10^{-5}$ (Case/Control = 21/3003) and in BioVU $p_{BioVU} = 3.07 \times 10^{-3}$ (Case/Control = 16/2,882).

We had a total of 3,054 results for the same SNP, coded allele, and direction of effect, when the replication criterion was based on requiring the same ICD-9 code category, (S4 Table) and association between *F5* SNP rs6427196 and "pulmonary embolus" (ICD-9 code category 453) was most significant with $p_{MyCode} = 1.3 \times 10^{-7}$ ($Case/Control_{MyCode} = 53/2,970$) and $p_{BioVU} = 5.72 \times 10^{-3}$ ($Case/Control_{BioVU} = 63/2,834$).

We used METAL [40] to perform a pooled-analysis for both sets of results meeting our criteria for replication across BioVU and MyCode® (S3 and S4 Tables). The most significant of these associations was between the diagnosis "myalgia and myostosis" (ICD-9 code 729.1) and the *PLA2G2E* SNP rs1108975 with $p_{METAL} = 8.99 \times 10^{-8}$ ($Case/Control_{MyCode} = 136/2,888$, $Case/Control_{BioVU} = 43/2,853$). Of the 3,054 results meeting our PheWAS replication criteria for the same ICD-9 code category, the most significant association was between the diagnosis "rheumatoid arthritis and other inflammatory polyarthropathies" (ICD-9 code category 714) and *C6orf10* SNP rs6910071 with a meta-analysis $p_{METAL} = 2.58 \times 10^{-9}$ ($Case-Control_{MyCode} = 60/2,964$, $Case-Control_{BioVU} = 81/2,818$). Another top association signal was between the *ATN1* SNP rs2239167 and "diabetes mellitus, type 2" (ICD-9 category code 250) with a $p_{METAL} = 6.39 \times 10^{-9}$ ($Case-Control_{MyCode} = 23/3,001$, $Case-Control_{BioVU} = 41/2,858$).

Matching Previously Reported GWAS Results

Results replicating in both biorepositories using the exact ICD-9 code and ICD-9 category-based PheWAS were evaluated for any matches to SNPs with previously reported associations for the same phenotypes with significance of p-value$<1x10^{-5}$ in the NHGRI-EBI GWAS catalog and GRASP. We found that GRASP included GWAS results from many more studies than the NHGRI-EBI catalog and most association in NHGRI-EBI were also reported in GRASP.

However, we report SNP associations from previous GWAS if it is reported in either NHGRI-EBI catalog or GRASP.

Of the SNPs in the 224 exact ICD-9 code replicating associations, we found a total of 10 SNPs with phenotypic associations also previously reported in existing GWAS. However, none of these SNPs were associated with the same phenotypes in our study compared to existing GWAS. For example, the *PARD3B* SNP rs1207421 is reported in the GWAS catalog to be associated with "knee osteoarthritis" (reported GWAS $p = 6 \times 10^{-6}$) [41]. In our study this SNP had a novel association with "scar conditions and fibrosis of skin" (ICD-9 code; 709.2), and was not associated with osteoarthritis.

In our results meeting the PheWAS criteria for replication in both datasets for the same ICD-9 code category, a total of 284 SNPs were also represented in either NHGRI-EBI GWAS catalog or the GRASP database. A total of 42 SNP-phenotype pairs matched identical associations reported in the GWAS catalogs, and five results had a diagnosis closely related to the phenotype reported in the GWAS catalogs. A few top SNP-phenotype pairs matching previously reported associations include the *C6orf10* SNP rs6910071 associated in our study with "rheumatoid arthritis and other inflammatory polyarthropathies" and reported previously to be associated with RA [42–44]. Also, an *F5* SNP rs6427196 and pulmonary embolus/DVT association ($p_{METAL} = 1.16 \times 10^{-8}$, $Case/Control_{MyCode} = 53/2{,}970$, $Case/Control_{BioVU} = 63/2{,}834$) that has been previously reported with venous thromboembolism [45,46,46,47]. A SNP rs2647044 downstream of *MTC03P1* associated with "diabetes mellitus type 1" ($p_{METAL} = 7.94 \times 10^{-7}$, $Case/Control_{MyCode} = 22/3{,}002$, $Case/Control_{BioVU} = 98/2{,}801$) in our study, was previously reported to be associated in GWAS with T1DM [48] and RA [42]. Finally, there was an association between rs660895 and "rheumatoid arthritis and other inflammatory polyarthropathies" ($p_{METAL} = 3.28 \times 10^{-7}$, $Case/Control_{MyCode} = 85/2{,}939$, $Case/Control_{BioVU} = 138/2{,}761$), and this SNP has shown previous association with RA [43]. Fig 2 shows a plot of the replicating associations found for our ICD-9 code category PheWAS for SNPs matching the exact or closely related ICD-9 code category description in previously reported studies.

Associations with Immune and Autoimmune Related Diagnoses

From the replicating category ICD-9 code results, we concentrated on the 441 SNP-ICD-9 code associations for immune- or autoimmune-related diagnoses, i.e. diagnoses more directly impacted by immune system variation. To assist developing a robust list of more specifically immune- or autoimmune-related ICD-9 codes, three separate researchers evaluated the selection of these ICD-9 code classes to reach consensus, resulting in 441 SNP-ICD-9 association results for further evaluation. In Table 2, we list the ICD-9 categories selected for this analysis as well as the genes with SNPs. S4 Table lists the association results meeting the criteria for significance and replication. Fig 3 presents the results for the 441 SNP-ICD-9 associations related to immune function.

Fig 3 shows the great diversity of the replicating results, ranging from autoimmune conditions, such as "rheumatoid arthritis" (RA) (ICD-9 code: 714), to contagious diseases, such as "bacterial pneumonia" (ICD-9 code: 482), to inflammatory diseases including "erythematous conditions" (ICD-9 code: 695). The diagnosis with the largest total number of replicating association results in Fig 3 was "osteoarthrosis" (ICD-9 code: 715) with 53 results in the MyCode® and BioVU datasets. "Diabetes mellitus, type I" (TIDM) (ICD-9 code: 250) and "chronic sinusitis" (ICD-9 code: 473) each had 49 replicating results in both datasets.

Within this subset of results, the statistically most significant results include associations with RA, T1DM, and allergic rhinitis and osteoarthrosis. RA and T1DM are both common autoimmune diseases, and inflammation has been implicated in the pathogenesis of

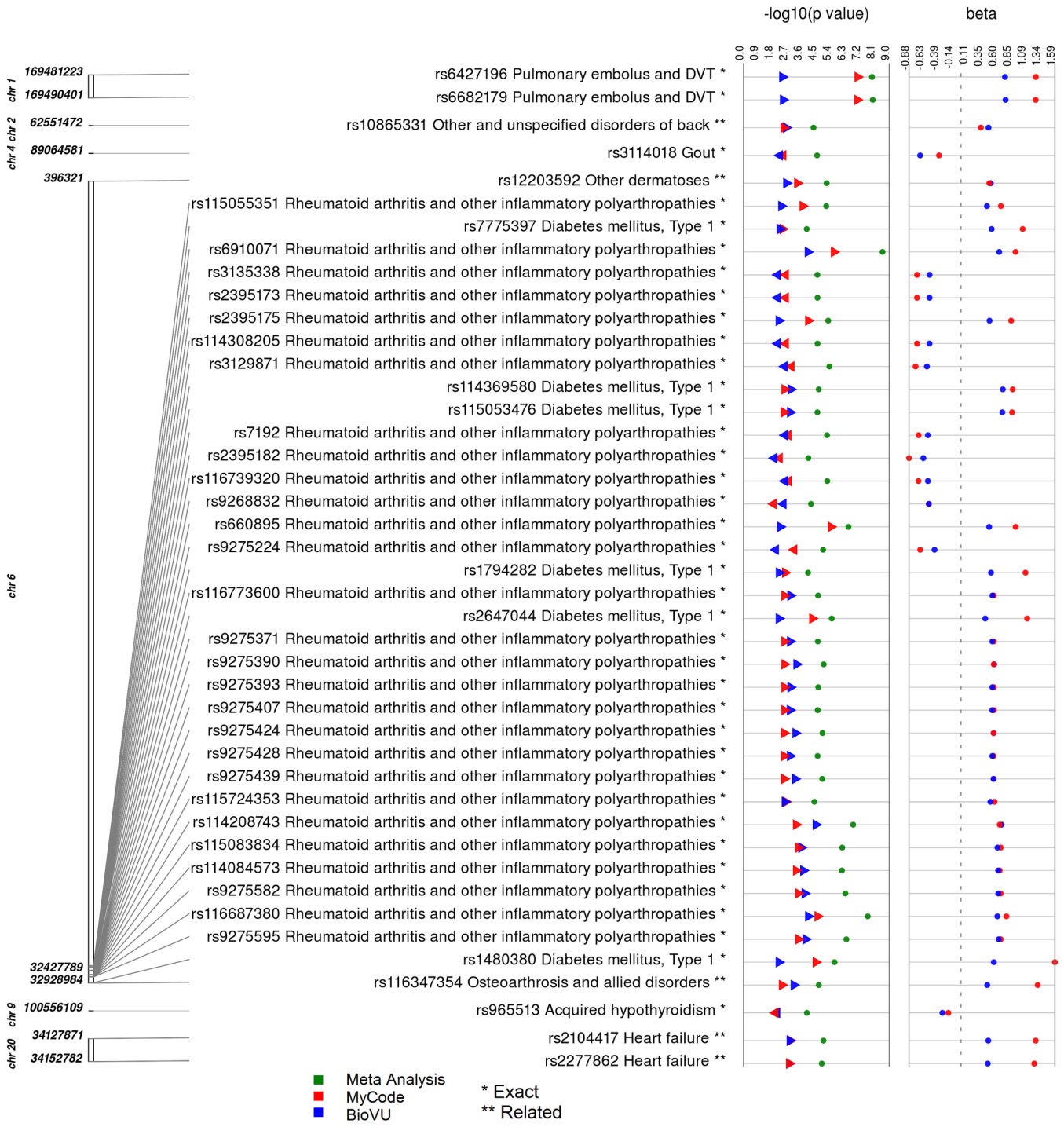

Fig 2. Synthesis view plot showing PheWAS results replicating across MyCode® and BioVU that have previously reported associations. The first track is the chromosomal location for each SNP. The next column lists the SNP identifier, the phenotype associated in our study, and the reported GWAS trait ($p<10^{-5}$). Results representing exact matches with the NHGRI-EBI GWAS catalog and GRASP are annotated with a single asterisk and the closely related traits are represented with a double asterisk. Blue symbols represent results from MyCode®, red symbols represent results from BioVU and green symbols are the pooled analysis results obtained using the program METAL.

doi:10.1371/journal.pone.0160573.g002

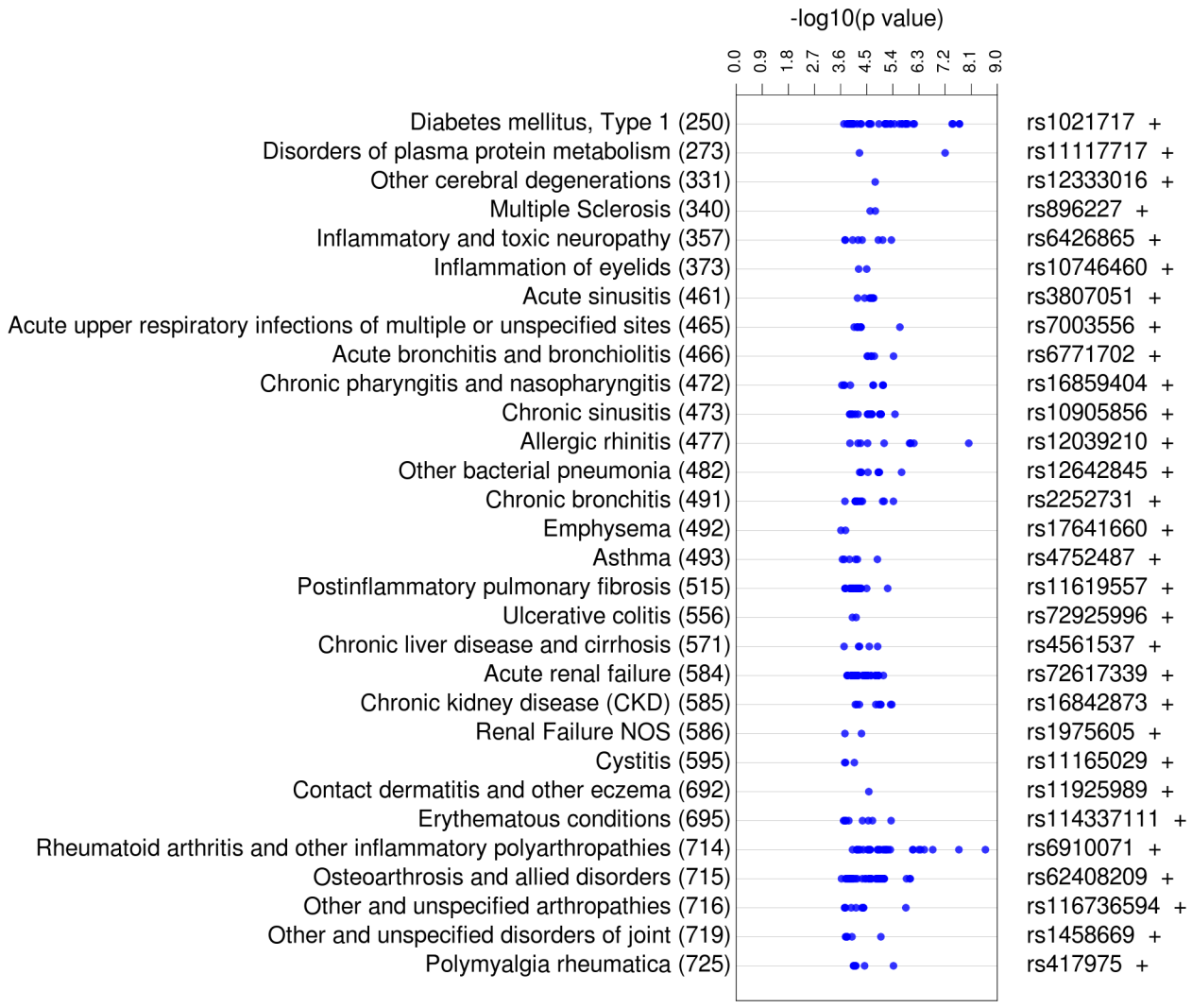

Fig 3. PheWAS View Plot of Meta-analysis Results with p<0.01 Replicating for the Same ICD-9 Category, Meeting Autoimmune and Immune-Related Diagnosis Criteria. The left track specifies the phenotype and ICD-9 Category code with which the SNP was associated. The next track indicates–$\log_{10}$(p-value) from the meta-analysis performed on all replicating SNPs with p<0.01. The last track indicates the SNP that had the most significant p-value, and the direction of effect of the association (+, positive; -, negative). The total number of associations between the SNPs and diagnoses was 409.

doi:10.1371/journal.pone.0160573.g003

osteoarthrosis [49]. The most significant replicating association from the pooled analysis was between rs6910071 and the diagnosis RA with $p_{METAL} = 2.58 \times 10^{-9}$, as mentioned above this SNP has also been associated with RA in previous GWAS.

Functional Annotation of Associated Variants within Genes

Of the 441 autoimmune and immune system related results, we next considered SNPs directly mapping to or within 50 kb of a gene to include promoter and regulatory regions, for potentially relevant genes. There were 233 associations of SNPs that mapped within genes. Herein, we again observed multiple variants mapped to genes with known relationships with associated phenotypes, particularly for RA and T1DM within the well-characterized HLA locus. This group included SNPs associated with T1DM within the *HLA-DMA*, and *HLA-DOB* genes, and

SNPs associated with RA mapped to *HLA-DRB9* and *HLA-DRB1*. The SNP rs1480380 associated with osteoarthrosis and T1DM in our study is within 50 kb of *HLA-DMA*. Nine SNPs associated with T1DM in our study were in the *HLA-DRA* gene, and eight SNPs associated with RA also mapped to *HLA-DRA*. It is important to consider that the HLA region on chromosome 6 is highly polymorphic and there could be variability in HLA alleles due to population stratification. We only used EA individuals in this study, thus we expect less variation in the HLA region compared to a cohort across multiple ancestries. Further, we compared the MAFs of HLA region SNPs with 1000 Genomes EA population and the frequencies were very close.

Functional Annotation of Associated Variants outside Protein-Coding Genes

Of the 441 replicating associations classified as autoimmune- or immune-related, there were 208 associations where SNPs mapped outside protein coding genes, with a total of 206 SNPs. We annotated the SNPs that did not map to genes with information on potential functionality using two public databases: HaploReg V2 and SCAN (S6 Table).

A total of 40 SNPs were associated with significantly altered gene expression in HapMap CEU lymphoblastoid cell lines in the SCAN database. The statistically most significant eQTL in SCAN for our 206 SNPs is rs2395182 and the expression of eight HLA locus genes as well as other genes including *RNASE2*, and *ZNF749*. In our study, this SNP was associated with "rheumatoid arthritis and other inflammatory polyarthropathies" (ICD-9 code: 714). This SNP is located 495 bp upstream of *HLA-DRA*. Another highly significant SCAN eQTL exists between the SNP rs1794282 and altered expression of *HLA-DQA1* and *HLA-DQA2*. In our study, SNP rs1794282 was associated with "type 1 diabetes" (ICD-9 code: 250).

We also used HaploReg to annotate the 206 SNPs that were outside gene boundaries, as well as any SNPs in LD ($r^2 > 0.8$) with the original SNPs. Of these 206 SNPs, 134 had altered regulatory motifs in HaploReg and are likely to influence transcriptional regulation. Sixteen of the SNPs were reported to be strong enhancers in one cell type, and 47 SNPs are weak enhancers in one or more cell types, supporting the potentially functional role of these variants.

Pleiotropy and Association Network

With the wide range of phenotypes explored in PheWAS, SNPs associated with more than one phenotype can be identified, indicating potential pleiotropy. In this study, 107 out of 2,770 SNPs had associations meeting our PheWAS criteria for the same ICD-9 code category that demonstrated potential pleiotropy. An instance of such potential pleiotropy is seen with SNP rs114369580 which was found to be associated with immune related disorders T1DM and gout. The plot on the results of the SNPs associated with more than one phenotype is shown in Fig 4.

We further explored the interrelations between association results as a network using Cytoscape. Fig 5 shows a sub-network of our results for potential pleiotropy. We used SNPs that met our PheWAS criteria for replication for the same ICD-9 code category, and annotated those SNPs with the nearest gene. Next we used Cytoscape to link together ICD-9 codes with genes, where those ICD-9 codes are associated with the SNPs within those genes. Next we annotated the genes using the KEGG [36] pathways, and added them to the network. In the network, the diagnoses of RA and T1DM link to *HLA-DRA*, a gene found in the rheumatoid arthritis and type I diabetes KEGG disease pathways. Another interesting pattern was between *IL6* (rs2069843, rs2069849, rs1548216, rs2069844) and RA (ICD-9: 714) and *IL23R* (rs10889675) and "arthropathy" (ICD-9: 716), where both genes are from the interleukin gene family and

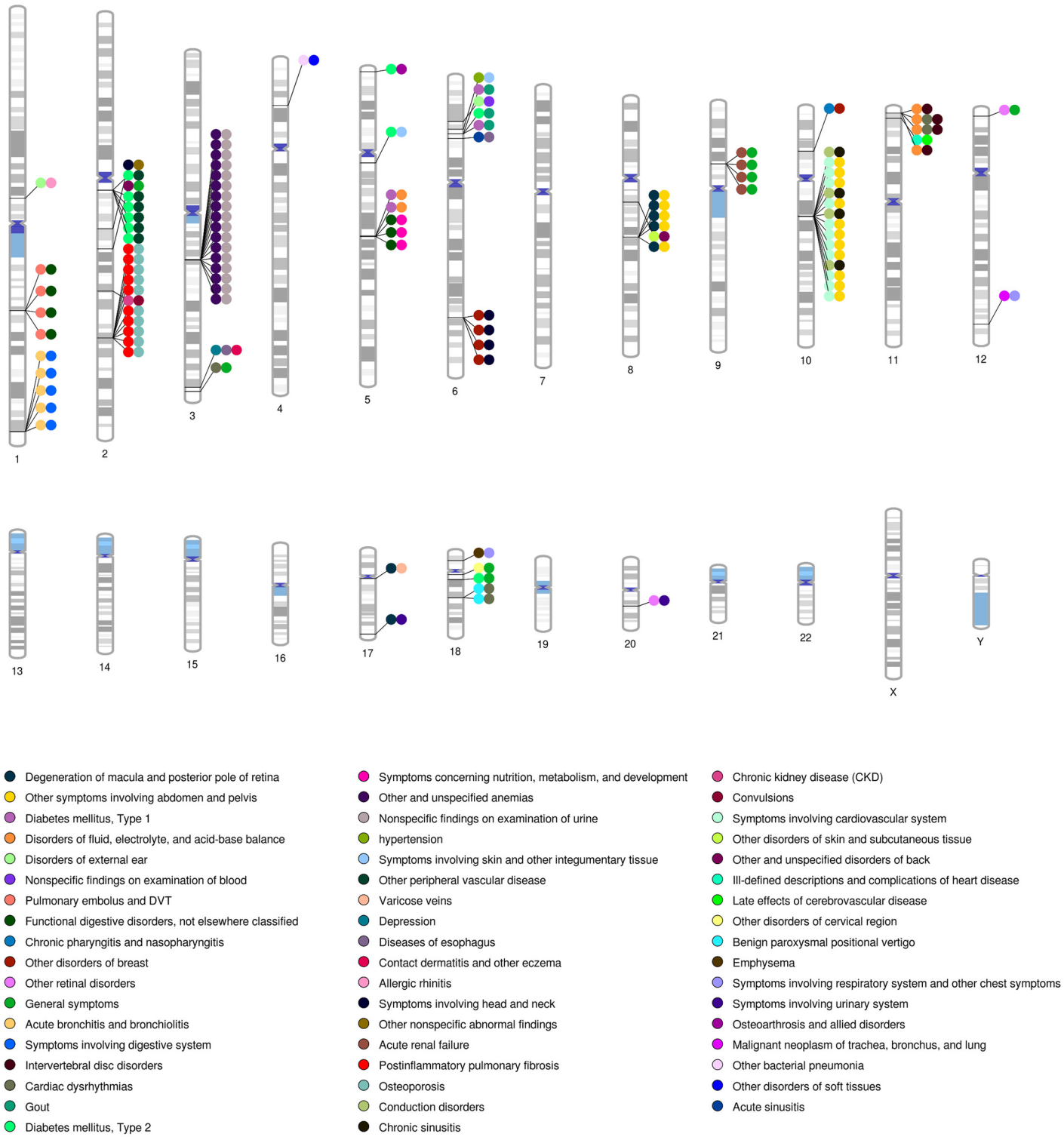

Fig 4. Pleiotropy: SNPs Associated with more than One Phenotype and Replicating across more than One Study for the Same ICD-9 Category. This chromosomal ideogram has lines indicating the location of the SNP, with filled colored circles indicating different ICD-9 code diagnoses associated with that particular SNP. When there are multiple pairs of the same phenotypes in the same region, this indicates regions where several SNPs in close proximity were associated with the same pairs of phenotypes.

doi:10.1371/journal.pone.0160573.g004

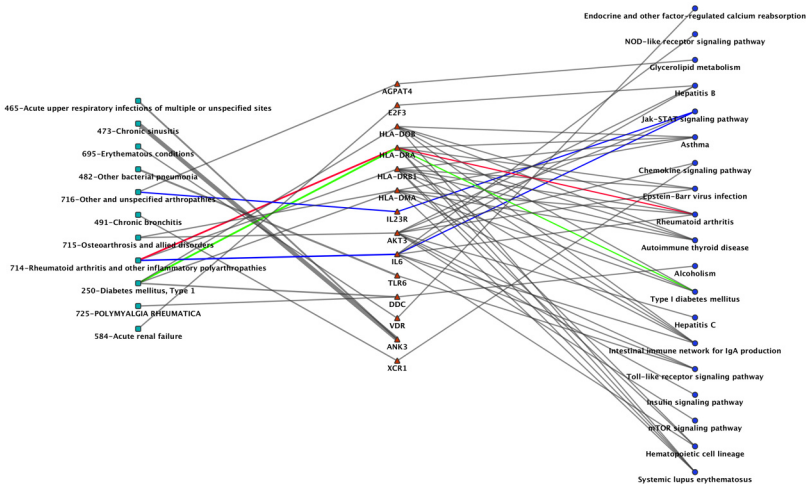

Fig 5. Cytoscape Network Showing the Connections between Phenotypes, the Genes with SNPs, and Pathways. In this network, green squares represent phenotype; red triangles represent genes; and blue circles are KEGG pathways. The colored lines highlight the link between phenotype and pathway. For the gene *HLA-DRA* with SNPs associated with "*714*: *rheumatoid arthritis*" and "*250*: *type 1 diabetes*" is present in the KEGG pathway of "*rheumatoid arthritis*" (red line) and "*type 1 diabetes"* (green line) respectively. Also, the blue edge shows the connection between *"714*: *rheumatoid arthritis"*, *"716*: *other specified arthropathies"* and the KEGG "*JAK-STAT signaling pathway*" through two interleukin genes, *IL23R* and *IL6*.

doi:10.1371/journal.pone.0160573.g005

found in a JAK-STAT signaling pathway. In Fig 5, the number of replicating association SNPs within a gene is represented by thickness of the lines. We identified that *ANK3* has most number of SNPs (a total of 65 SNPs) associated with chronic sinusitis (ICD-9: 473), Acute upper respiratory infections (ICD-9: 465).

Discussion

Using PheWAS we found a series of associations between SNPs within or in close proximity of genes with known involvement in the immune system, such as genes within the HLA locus, including genetic variants within *HLA-DRA* associated with a series of immune-relevant diagnoses. As a member of the HLA class of molecules, *HLA-DRA* is expressed in various antigen presenting cells, and has been implicated in both T1DM [50] and RA [51].

Four SNPs in both biorepositories with associations with RA are within *IL6*. The product of the *IL6* gene is an interleukin, both a pro-inflammatory cytokine and an anti-inflammatory myokine, with an important role in regulation of inflammation and hematopoiesis. Therapies targeting the IL6 signaling system have been found effective for the treatment of RA [52].

A different group of three SNPs associated with RA mapped to *TNIP1*, a gene encoding an A20-binding protein that has a role in autoimmunity through the regulation of NFκB activation. Previous studies have shown genetic variants in *TNIP1* associated with various autoimmune conditions including psoriasis, and systemic lupus erythematous [53].

The most significant association in MyCode® that replicated in BioVU was for the SNP rs6682179 mapped to the *F5* gene, which was associated with pulmonary embolus and deep vein thrombosis (ICD-9 code category 453). The *F5* gene encodes the coagulation factor V protein, a protein that circulates in the blood and is part of the blood coagulation cascade. The *F5* gene has many known mutations causing different blood coagulation disorders like factor V deficiency [54] and factor V Leiden thrombophilia [55]. We found that rs6682179 from our study is in linkage disequilibrium with rs2420371 and rs1018827 that have known

associations with plasma levels of natural anticoagulant inhibitors [56] and venous thrombosis [57], respectively.

Within our replicating results, a majority of the SNPs associated with chronic sinusitis (ICD-9: 473), and acute upper respiratory infection (ICD-9: 465) phenotypes map to genes that encode members of the ankyrin protein family. Specifically, several SNPs mapping to *ANK3* are highly represented in both the acute respiratory infection and the chronic sinusitis association results, and variants mapping to *ANKS1A* were associated with acute sinusitis. Ankyrin proteins are involved in cell migration and in mediating the attachment of proteins to the cytoskeleton. While this protein family is not implicated in immune-related disorders, it is of interest that related phenotypes are associated with SNPs mapping to the same class of genes.

While we were able to seek replication of association results across two separate EHRs, a challenge was the limited overlap between the two datasets for specific ICD-9 codes as well as ICD-9 categories. Thus, it will be worthwhile to seek replication in other data sets in the future, such as through other sites of eMERGE [58,59], or evaluate the results using another method such as permutation testing. This lack of overlap is partially due to the variation in coding practices for ICD-9 codes from medical institution to institution. The lack of overlap is also likely partly due to the targeted recruitment of individuals with specific diseases at Geisinger, as the MyCode® dataset is enriched for patients with obesity or AAA.

Another potential limitation in PheWAS is the multiple hypothesis-testing burden. We contrasted the significance of our results with a Bonferroni correction. We have correlated SNPs, and we also have correlated phenotypes within this study. The use of ICD-9 codes for case/control status is less well powered than traditional GWAS, due to lower case numbers. Also, the goal of PheWAS is to be exploratory, generating new hypothesis for further research. Thus we focused our evaluation within this manuscript on replication (when possible) across the two datasets and performed a meta-analysis for the results replicating across the two studies. We evaluated the SNP-phenotype associations by calculating individual SNP-phenotype associations; a future direction is evaluation and use of methods that combine information from multiple phenotypes for statistical testing.

We focused our associations on genetic variants within genes with evidence of involvement in autoimmunity and the immune system. We could have used a more general approach, seeking associations between autoimmune- and immune-traits and genome-wide SNPs, but this would have increased our multiple hypothesis testing further. Our more narrow search space will have missed genetic variants without previous evidence of association with autoimmunity and the immune system.

Our PheWAS analysis pipeline replicated previously published associations between SNPs and immune phenotypes. We also have identified a series of potential novel associations, and some of these results replicated with exact ICD-9 code or ICD-9 code category across two separate EHRs. Further studies are needed to confirm the biological validity of our potentially novel associations. Our results demonstrate potential pleiotropy, through cross-phenotype associations where individual SNPs are associated with more than one diagnosis. Further, our results show associations between inflammation/autoimmune related SNPs and disease/outcomes such as obesity, underscoring the impact of variation in the immune system on complex traits beyond direct connections to autoimmunity and the immune system.

Supporting Information

S1 Fig. Study design and analysis workflow.
(TIFF)

S1 Table. A list of immune related genes.
(XLS)

S2 Table. Results unique to each study with $p < 1x10^{-4}$.
(XLS)

S3 Table. Replicating Exact ICD-9 code results.
(XLSX)

S4 Table. Replicating Category ICD-9 code results.
(XLSX)

S5 Table. Autoimmune/Immune-related category ICD-9 code results.
(XLS)

S6 Table. Annotation via SCAN and HaploReg.
(XLS)

Author Contributions

Conceived and designed the experiments: AV AB HK GT JC MR SP.

Performed the experiments: YB.

Analyzed the data: AV AB SP.

Wrote the paper: AV AB HK DC GG JC MR SP.

References

1. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS Genet. 2011; 7: e1002254. doi: 10.1371/journal.pgen.1002254 PMID: 21852963
2. Cho JH, Gregersen PK. Genomics and the Multifactorial Nature of Human Autoimmune Disease. N Engl J Med. 2011; 365: 1612–1623. doi: 10.1056/NEJMra1100030 PMID: 22029983
3. Knight JC. Genomic modulators of the immune response. Trends Genet TIG. 2013; 29: 74–83. doi: 10.1016/j.tig.2012.10.006 PMID: 23122694
4. Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, et al. Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. Fisher EMC, editor. PLoS Genet. 2008; 4: e1000024. doi: 10.1371/journal.pgen.1000024 PMID: 18437207
5. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. Science. 2014; 344: 519–523. doi: 10.1126/science.1249547 PMID: 24786080
6. Libby P, Ridker PM, Maseri A. Inflammation and Atherosclerosis. Circulation. 2002; 105: 1135–1143. doi: 10.1161/hc0902.104353 PMID: 11877368
7. Anker SD, von Haehling S. Inflammatory mediators in chronic heart failure: an overview. Heart. 2004; 90: 464–470. doi: 10.1136/hrt.2002.007005 PMID: 15020532
8. Theoharides TC, Asadi S, Patel AB. Focal brain inflammation and autism. J Neuroinflammation. 2013; 10: 46. doi: 10.1186/1742-2094-10-46 PMID: 23570274
9. Badeanlou L, Furlan-Freguia C, Yang G, Ruf W, Samad F. Tissue factor-protease-activated receptor 2 signaling promotes diet-induced obesity and adipose inflammation. Nat Med. 2011; 17: 1490–1497. doi: 10.1038/nm.2461 PMID: 22019885
10. Donath MY, Shoelson SE. Type 2 diabetes as an inflammatory disease. Nat Rev Immunol. 2011; 11: 98–107. doi: 10.1038/nri2925 PMID: 21233852
11. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, et al. Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLOS Genet. 2013; 9. doi: 10.1371/journal.pgen.1003087

12. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010; 26: 1205–1210. doi: 10.1093/bioinformatics/btq126 PMID: 20335276
13. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, Wilson S, 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. 2014; 10: e1004678. doi: 10.1371/journal.pgen.1004678 PMID: 25474351
14. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, Daar ES, et al. Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. Open Forum Infect Dis. 2015; 2: ofu113–ofu113. doi: 10.1093/ofid/ofu113 PMID: 25884002
15. Park SH, Lee JY, Kim S. A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. BMC Syst Biol. 2011; 5 Suppl 2: S13. doi: 10.1186/1752-0509-5-S2-S13 PMID: 22784570
16. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 2013; 14: 661–673. doi: 10.1038/nrg3502 PMID: 23917628
17. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011; 13: 101. doi: 10.1186/ar3204 PMID: 21345260
18. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: Past, Present and Future. Genet Med Off J Am Coll Med Genet. 2013; 15: 761–771. doi: 10.1038/gim.2013.72
19. Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. Genet Med Off J Am Coll Med Genet. 2010; 12: 648–650.
20. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci. 2010; 3: 42–48. doi: 10.1111/j.1752-8062.2010.00175.x PMID: 20443953
21. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet. 2009; 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373
22. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013; 10: 5–6. doi: 10.1038/nmeth.2307 PMID: 23269371
23. Verma SS, de Andrade M, Tromp G, Kuivaniemi H, Pugh E, Namjou-Khales B, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet. 2014; 5. doi: 10.3389/fgene.2014.00370
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. Available: http://www.R-project.org/.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–575. doi: 10.1086/519795 PMID: 17701901
26. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput Pac Symp Biocomput. 2009; 368–379. PMID: 19209715
27. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, Moore C, et al. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min. 2013; 6: 25. doi: 10.1186/1756-0381-6-25 PMID: 24378202
28. Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao Y-T, et al. Addressing Population-Specific Multiple Testing Burdens in Genetic Association Studies: Population-Specific Genome-Wide Thresholds. Ann Hum Genet. 2015; 79: 136–147. doi: 10.1111/ahg.12095 PMID: 25644736
29. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42: D1001–1006. doi: 10.1093/nar/gkt1229 PMID: 24316577
30. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 2009; 106: 9362–9367. doi: 10.1073/pnas.0903103106 PMID: 19474294
31. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014; 30: i185–i194. doi: 10.1093/bioinformatics/btu273 PMID: 24931982

32. Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, Jensen N, et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. 2015; 43: D799–D804. doi: 10.1093/nar/gku1202 PMID: 25428361

33. Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. BioData Min. 2010; 3: 10. doi: 10.1186/1756-0381-3-10 PMID: 21162740

34. Wolfe D, Dudek S, Ritchie MD, Pendergrass SA. Visualizing genomic information across chromosomes with PhenoGram. BioData Min. 2013; 6: 18. doi: 10.1186/1756-0381-6-18 PMID: 24131735

35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13: 2498–2504. doi: 10.1101/gr.1239303 PMID: 14597658

36. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28: 27–30. PMID: 10592173

37. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2011; gkr917. doi: 10.1093/nar/gkr917

38. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, et al. SCAN: SNP and copy number annotation. Bioinforma Oxf Engl. 2010; 26: 259–262. doi: 10.1093/bioinformatics/btp644

39. Ghigliotti G, Barisione C, Garibaldi S, Fabbi P, Brunelli C, Spallarossa P, et al. Adipose tissue immune response: novel triggers and consequences for chronic inflammatory conditions. Inflammation. 2014; 37: 1337–1353. doi: 10.1007/s10753-014-9914-1 PMID: 24823865

40. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma Oxf Engl. 2010; 26: 2190–2191. doi: 10.1093/bioinformatics/btq340

41. Valdes AM, Loughlin J, Timms KM, van Meurs JJB, Southam L, Wilson SG, et al. Genome-wide association scan identifies a prostaglandin-endoperoxide synthase 2 variant involved in risk of knee osteoarthritis. Am J Hum Genet. 2008; 82: 1231–1240. doi: 10.1016/j.ajhg.2008.04.006 PMID: 18471798

42. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42: 508–514. doi: 10.1038/ng.582 PMID: 20453842

43. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. N Engl J Med. 2007; 357: 1199–1209. doi: 10.1056/NEJMoa073491 PMID: 17804836

44. Padyukov L, Seielstad M, Ong RTH, Ding B, Rönnelid J, Seddighzadeh M, et al. A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Ann Rheum Dis. 2011; 70: 259–265. doi: 10.1136/ard.2009.126821 PMID: 21156761

45. Tang W, Teichert M, Chasman DI, Heit JA, Morange P-E, Li G, et al. A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. Genet Epidemiol. 2013; 37: 512–521. doi: 10.1002/gepi.21731 PMID: 23650146

46. Germain M, Saut N, Oudot-Mellakh T, Letenneur L, Dupuy A-M, Bertrand M, et al. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. PLOS ONE. 2012; 7: e38538. doi: 10.1371/journal.pone.0038538 PMID: 22675575

47. Heit JA, Armasu SM, Asmann YW, Cunningham JM, Matsumoto ME, Petterson TM, et al. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. J Thromb Haemost JTH. 2012; 10: 1521–1531. doi: 10.1111/j.1538-7836.2012.04810.x PMID: 22672568

48. Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007; 448: 591–594. doi: 10.1038/nature06010 PMID: 17632545

49. Wang Q, Rozelle AL, Lepus CM, Scanzello CR, Song JJ, Larsen DM, et al. Identification of a central role for complement in osteoarthritis. Nat Med. 2011; 17: 1674–1679. doi: 10.1038/nm.2543 PMID: 22057346

50. Steck AK, Zhang W, Bugawan TL, Barriga KJ, Blair A, Erlich HA, et al. Do non-HLA genes influence development of persistent islet autoimmunity and type 1 diabetes in children with high-risk HLA-DR,DQ genotypes? Diabetes. 2009; 58: 1028–1033. doi: 10.2337/db08-1179 PMID: 19188433

51. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJM. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated

signalling pathways. Hum Mol Genet. 2011; 20: 3494–3506. doi: 10.1093/hmg/ddr248 PMID: 21653640

52. Park JY, Pillinger MH. Interleukin-6 in the pathogenesis of rheumatoid arthritis. Bull NYU Hosp Jt Dis. 2007; 65 Suppl 1: S4–10. PMID: 17708744

53. Kawasaki A, Ito S, Furukawa H, Hayashi T, Goto D, Matsumoto I, et al. Association of TNFAIP3 interacting protein 1, TNIP1 with systemic lupus erythematosus in a Japanese population: a case-control association study. Arthritis Res Ther. 2010; 12: R174. doi: 10.1186/ar3134 PMID: 20849588

54. Lippi G, Favaloro EJ, Montagnana M, Manzato F, Guidi GC, Franchini M. Inherited and acquired factor V deficiency. Blood Coagul Fibrinolysis Int J Haemost Thromb. 2011; 22: 160–166. doi: 10.1097/MBC.0b013e3283424883

55. Kujovich JL. Factor V Leiden thrombophilia. Genet Med Off J Am Coll Med Genet. 2011; 13: 1–16. doi: 10.1097/GIM.0b013e3181faa0f2

56. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, et al. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br J Haematol. 2012; 157: 230–239. doi: 10.1111/j.1365-2141.2011.09025.x PMID: 22443383

57. Germain M, Saut N, Greliche N, Dina C, Lambert J-C, Perret C, et al. Genetics of venous thrombosis: insights from a new genome wide association study. PLOS ONE. 2011; 6: e25581. doi: 10.1371/journal.pone.0025581 PMID: 21980494

58. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med Off J Am Coll Med Genet. 2013; 15: 761–771. doi: 10.1038/gim.2013.72

59. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; 4. doi: 10.1186/1755-8794-4-13