Please cite this article in press as: Verma et al., Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals, The American Journal of Human Genetics (2019), https://doi.org/10.1016/j.ajhg.2018.11.006

ARTICLE

Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals

Anurag Verma,[1,2] Lisa Bang,[3] Jason E. Miller,[1] Yanfei Zhang,[4] Ming Ta Michael Lee,[4] Yu Zhang,[5] Marta Byrska-Bishop,[3,7] David J. Carey,[6] Marylyn D. Ritchie,[1,2] Sarah A. Pendergrass,[3] Dokyoon Kim,[2,3,*] and the DiscovEHR Collaboration

Phenome-wide association studies (PheWASs) have been a useful tool for testing associations between genetic variations and multiple complex traits or diagnoses. Linking PheWAS-based associations between phenotypes and a variant or a genomic region into a network provides a new way to investigate cross-phenotype associations, and it might broaden the understanding of genetic architecture that exists between diagnoses, genes, and pleiotropy. We created a network of associations from one of the largest PheWASs on electronic health record (EHR)-derived phenotypes across 38,682 unrelated samples from the Geisinger's biobank; the samples were genotyped through the DiscovEHR project. We computed associations between 632,574 common variants and 541 diagnosis codes. Using these associations, we constructed a "disease-disease" network (DDN) wherein pairs of diseases were connected on the basis of shared associations with a given genetic variant. The DDN provides a landscape of intra-connections within the same disease classes, as well as interconnections across disease classes. We identified clusters of diseases with known biological connections, such as autoimmune disorders (type 1 diabetes, rheumatoid arthritis, and multiple sclerosis) and cardiovascular disorders. Previously unreported relationships between multiple diseases were identified on the basis of genetic associations as well. The network approach applied in this study can be used to uncover interactions between diseases as a result of their shared, potentially pleiotropic SNPs. Additionally, this approach might advance clinical research and even clinical practice by accelerating our understanding of disease mechanisms on the basis of similar underlying genetic associations.

Introduction

Pleiotropy occurs when a given locus (e.g., a SNP or gene) influences two or more different phenotypes or traits. The phenome-wide association study (PheWAS) is an important tool that has the strength to identify associations between genetic variants and clinical phenotypes and also the potential to reveal pleiotropic associations among diseases.[1–4] Although pleiotropy often refers to a common molecular mechanism, PheWASs can identify statistical associations between a single variant and multiple phenotypes. They can also provide the basis for a statistical approach to identifying cross-phenotype associations, which can then be verified as true pleiotropic effects.[1] Over the past decade, associations from hundreds of genome-wide association studies (GWASs) have accumulated in the EBI GWAS Catalog.[5] Although a GWAS typically investigates a single phenotype at a time, the accumulated associations from many studies (such as those in the EBI GWAS Catalog) provide the opportunity to investigate cross-phenotype associations.[6,7] More recently, PheWASs have shown success in identifying cross-phenotype associations within the same study populations.[8,9]

Electronic health records (EHRs) are a powerful resource for studying individual outcomes via multiple longitudinal data elements, such as disease diagnoses, laboratory measures, medications, and other health-related information. EHR data have been useful in population health research; more importantly, linking EHR data with genomics data enables us to examine the genetic architecture of various disease outcomes and traits. PheWASs have been an effective tool to mine genetic associations for candidate SNPs or genome-wide variants;[10] hence, PheWASs provide the ability to identify cross-phenotype associations in which one SNP is associated with multiple diseases or traits. While investigating such cross-phenotype associations at a genome-wide scale, researchers might uncover potential hidden connections between diseases, especially when two diseases share associations with two or more SNPs that are located in different regions of the genome (Figure 1). One way to examine these connections is by creating a network of diseases in which pairs of diseases are connected on the basis of their shared associations with one or more SNPs. The strength of the network approach is that it condenses the complex links between SNPs and diseases and reveals links between diseases that would be hard to identify by just looking at disease associations at a single locus, such as when one only considers cross-phenotype association with a SNP.

Previous networks based on gene-disease associations, such as the Human Disease Network, used gene-disease

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA; [2]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA; [3]Biomedical and Translational Informatics Institute, Geisinger, Danville, PA 17821, USA; [4]Genomic Medicine Institute, Geisinger, Danville, PA 17821, USA; [5]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA; [6]Weis Center for Research, Geisinger, Danville, PA 17821, USA
[7]Present address: New York Genome Center, New York, NY 10013, USA
*Correspondence: dkim@geisinger.edu
https://doi.org/10.1016/j.ajhg.2018.11.006.

Please cite this article in press as: Verma et al., Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals, The American Journal of Human Genetics (2019), https://doi.org/10.1016/j.ajhg.2018.11.006

associations cataloged in the Online Mendelian Inheritance in Man (OMIM) database.[11] Other studies have used summary statistics from the GWAS catalog and/or the Genetic Associations Database (GAD) to investigate SNP-phenotype associations by using network-based analyses.[6] However, because these networks are based on summary statistics from disparate studies they have several critical limitations. First, differences in disease phenotype definitions across different studies can impact the interpretation of the association results,[12] leading to a high false positive rate. Second, in most cases, networks constructed from summary statistics are limited to providing individual-level genotype and phenotype information. This limitation can restrict the design of follow-up studies conducted on the new hypotheses derived from the network analyses.

In this study, we circumvented these limitations by drawing on association results from a single-source EHR that used consistent phenotype definitions and by using a single genotyping platform. We employed genetic associations between 625,325 SNPs and 541 ICD-9 (internal classification of diseases, ninth revision) diagnosis codes from a PheWAS of 38,668 unrelated individuals in Geisinger's biobank. A disease-disease network (DDN) was constructed from the 31,017 PheWAS association results (p value $< 1 \times 10^{-4}$).[13]

The DDN revealed thousands of connections between hundreds of diseases, and it also provided a high-level view of disease connections, including known and previously unreported disease links; therefore, to identify relevant disease connections from this dense network, we focused on three broad research goals. One of the key goals was to gain a bird's-eye view of disease connections characterized by underlying genetic associations. More specifically, when we grouped the diseases into disease classes, we asked which diseases share strong links within a disease class, as well as across different disease classes. The second goal was to integrate functional knowledge of the genome with genetic associations to ascertain biologically relevant findings. We integrated epigenomic knowledge into the DDN and examined the changes on the basis of tissue specificity. A number of recent studies have used EHR data alone to identify disease correlations and comorbidities,[14,15] so our last goal was to explain some of the disease correlations and comorbidities due to shared genetics. We compared the PheWAS-derived DDN to a separate network of diseases identified via an orthogonal EHR-only approach without genetics. Additionally, we used network statistics to mine the DDN for clusters of diseases with known links to one another in order to generate new hypotheses.

These disease connections can serve as the basis for new hypotheses to test for comorbidities and pleiotropy. With regard to testing new hypotheses, one of the most significant advantages of our approach is the single-source EHR linked to genomic data; it provides an opportunity to revisit individual-level genotype and phenotype data to design more targeted studies and ask more specific questions.

Material and Methods

Cross-Phenotype Associations

To construct the DDN, we used the genetic associations, identified through the PheWAS approach, that were reported in in our previous study to comprehensively test for associations between 625,325 SNPs and 541 EHR-based phenotypes.[13] As part of MyCode initiative, individuals agreed to provide blood and DNA samples for broad, future research, including genomic analyses as part of the Regeneron-Geisinger DiscovEHR collaboration and linking to data in the Geisinger EHR under a protocol approved by the Geisinger Institutional Review Board. The association testing was performed on genotype and phenotype data from 38,668 unrelated individuals. We used 31,017 associations with a p value $< 1 \times 10^{-4}$ to generate a network between disease diagnoses derived from ICD-9 phenotypes.[13]

Construction of the Network

Disease-Disease Network

In a bipartite network, the edges (E) are only formed between two distinct node groups. Different network objects, commonly represented as circles or dots, are referred to as nodes, and the connections drawn between these nodes are referred to as edges. The two nod groups in our DDN are diseases (D) and SNPs (S), and these two groups can be represented in a network by $N = (D,S,E)$, where E is an edge between two nodes. We also accounted for the linkage disequilibrium (LD) correlations between the SNPs in the association results used for construction of the network. Therefore, S can be either a SNP or an LD haplotype block shared between the two diseases (D). One can further compress the information in a bipartite network by projecting the network for each group of nodes (D or S), such that the nodes in the projection for one group will form an edge if they share at least one node with the other group. We constructed a bipartite network projection of diseases on the basis of shared SNP associations identified in the PheWAS analysis. In the DDN, nodes represent disease diagnoses, and two nodes are connected to each other when they share one or more SNP or an LD haplotype block (Figure 1 and Table S1). Further, we divided the ICD-9 codes into broader disease classes based on the ICD-9 categories reclassified by Rassekh et al.[16] We used a software called Gephi to construct and visualize the DDN (see Web Resources).

To evaluate the strength of the associations, we applied the hypergeometric test (SciPy implementation) to calculate the probability that an ICD-9 code shared associated SNPs with another ICD-9 code as a result of pure chance. The hypergeometric test is a generalization of Fisher's exact test for the one-tailed case and has been applied to gene-set enrichment tests,[17–19] gene-GO term-association tests,[20] and quantification of mosaicism,[21] among other tests. Because our genetic association data come from a single source, the number of SNPs associated with each disease can be compared, and thus this method surpasses some of the limitations of GWASs or literature-based networks. Given a population of N SNPs, wherein K is associated with given ICD-9 code 1 and n is associated with given ICD-9 code 2, the probability that strictly k SNPs are associated with both ICD-9 codes is given by the probability mass function as follows:

$$p = \frac{\binom{N-K}{n-k}\binom{K}{k}}{\binom{N}{n}}$$

The integral of this function is called the cumulative distribution function (CDF). To get the probability that k or more SNPs are

Please cite this article in press as: Verma et al., Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals, The American Journal of Human Genetics (2019), https://doi.org/10.1016/j.ajhg.2018.11.006

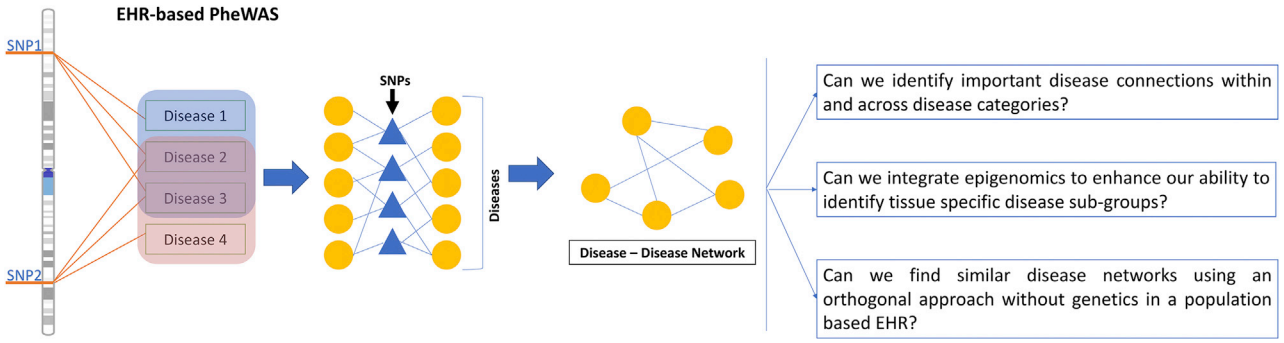


Figure 1. Overview of Network Construction
The cross-phenotype associations from a PheWAS analysis were used to construct the network of diseases. In the construction of the bipartite network, diseases (represented by yellow circles) and SNPs (represented by blue triangles) formed an edge if there was an association identified between them. Then, the bipartite network projection for the diseases was used for constructing a disease-disease network (DDN).

associated with both ICD-9 codes, we took (1 – CDF), the complementary cumulative distribution function (CCDF). Generally, the p value for a disease-disease association will be lower if the number of common SNP associations (k) is higher than the number of SNPs associated with each disease.

Network Statistics

Network statistics allow for the descriptive characterization of a network graph and the identification of meaningful connections. In this study, we applied various network analysis approaches to the DDN to identify the most crucial disease nodes, as well as to automate the extraction of disease cluster subnetworks. We used the statistical packages available as plug-ins within Gephi to perform all of the network analytics.

Hub Diseases

Hub nodes are those that have significantly more edges than other nodes. These nodes are important because they play a critical role in the centrality of the network. There are a number of ways to measure centrality of a network and, hence, identify hub disease nodes. In this case, we used a measure called betweenness centrality to identify such nodes in the DDN. Betweenness centrality for a given node (n_i) is calculated on the basis of the number of shortest paths between two other nodes (n_j, n_k) in the network and the number of times these paths pass through the node (n_i). We computed the betweenness centrality for all pairs of nodes across the whole network. The mathematical notation of betweenness centrality is as follows:

$$C_B(n_i) = \sum_{j,k} \frac{g_{j,k}(n_i)}{g_{j,k}}$$

$g_{j,k}$ *Shortest path linking node j and k*

$g_{j,k}(n_i)$ *Number of paths passing through node i*

The nodes with a high betweenness centrality value tend to be most important for keeping the network connected. We used this measure to change the representation of the nodes in the network by scaling the node size based on its betweenness centrality. In this way, we were able to visually identify the most important disease nodes in the network on the basis of network statistics.

Community Detection

Community detection is an approach used in network analytics to partition a large, densely connected network into smaller subnetworks.[22,23] Various community-detection methods can algorithmically identify meaningful subnetworks. These methods have most commonly been applied in social network analyses for the detection of structure in social interactions.[24] We used Louvain's method,[22,25] which is implemented in Gephi as the "modularity" feature, to partition the DDN and detect subnetworks, or communities, of diseases (see Web Resources). The communities detected had varying types of disease nodes. We used the identified disease communities to further investigate the biological interpretation of disease connections in the DDN.

Tissue-Specific Functional Annotation

To investigate the tissue-specific disease connections in the network, we used annotations from the 15 chromatin state models available on the Roadmap Epigenomics website to assign chromatin states to different tissues.[26] Using posterior probability, we assigned the most probable chromatin states for 127 different tissues, defined via posterior probability, to every 200 base-pair window across the genome. We also consolidated the 127 different tissues into 27 functional groups of tissues; for example, we used four different adipose tissues for the chromatin-state prediction, but we consolidated these into one group called "adipose tissue."[27] To calculate the most probable chromatin state for each functional tissue category, we averaged the posterior probabilities.[27] The chromatin-state prediction provides the annotations for the most active to the most quiescent regions of the non-coding genome. In this study, we focused on the active regulatory elements, such as enhancers, promoters, and active transcription start site (TSS); as a proof-of-concept, we only analyzed enhancer-state annotations.

The chromosome base pair position of each SNP was mapped onto the annotated chromatin states of the 27 functional groups of tissues. We considered variants to belong inside enhancer regions when a chromosome base pair position mapped onto either of the three enhancer states: enhancer (Enh), genic enhancer (EnhG), and bivalent enhancer (EnhBiv). Then, a total of seven DDNs were constructed from the associations between SNPs in enhancer regions. For visualization, we overlaid the networks created for each tissue onto the original DDN we had constructed.

Please cite this article in press as: Verma et al., Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals, The American Journal of Human Genetics (2019), https://doi.org/10.1016/j.ajhg.2018.11.006

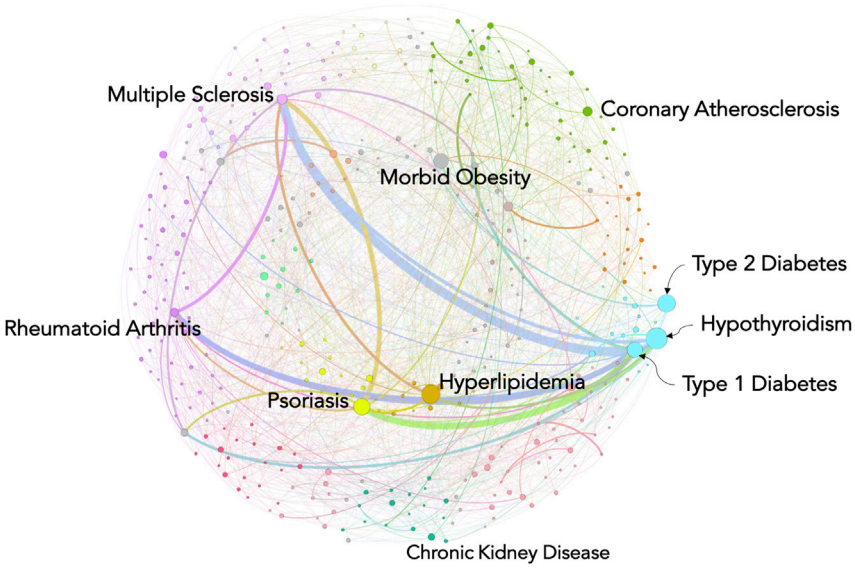


Figure 2. Disease-Disease Network
Using the cross-phenotype associations from an EHR-based PheWAS, we generated the disease-disease network (DDN). In this network, nodes represent the diseases, and the edges (lines) between the nodes represent shared genetic associations between pairs of diseases. The color of the node represents the broader disease category to which it belongs. The size of the node indicates the importance of the node in the network; importance was based on the betweenness centrality measure. The bigger nodes have higher betweenness centrality, and these nodes are referred to as hub nodes. The width of the edges (lines) represents the number of shared variants or variants in an LD block.

Results

Disease-Disease Network

Using the cross-phenotype associations found in the EHR-based PheWAS analysis, we constructed a disease-disease network (DDN) in order to understand the genetic similarities between human diseases (Figure 1). The network consists of 385 ICD-9-based disease diagnoses (which we obtained from an original 541 ICD-9 codes by using a threshold of $p < 1 \times 10^{-4}$) acting as nodes and the 1,398 edges connecting them. As shown in Figure 2, we classified ICD-9 codes into 15 broad disease classes, labeled with different colors. The DDN provides a bird's-eye view of the interconnections between the diseases on the basis of shared genetic associations. Many interconnections, including those between endocrine, musculoskeletal, and neurological disorders, were observed across classes. The strongest connections (indicated by the thickness of the network lines in Figure 2), which are based on the highest number of shared genetic variants, were between autoimmune disorders such as type 1 diabetes (MIM: 222100), rheumatoid arthritis (MIM: 180300), psoriasis (MIM: 177900), and multiple sclerosis (MIM: 126200) (Figure 2). These links are consistent with previous findings suggesting that these autoimmune diseases are determined by shared genetic components, indicating similar pathogenic mechanisms, even if completely different tissue types are affected in each disorder.[28–31] This could indicate that there are shared genetic pathways linking multiple SNPs to the same diseases. This could also be a reflection of a high correlation between disease occurrences.

Diseases Connected to the Most Other Diseases

Next, we focused on the disease nodes with the highest number of direct connections with other diseases in the network. The degree property (K) of the network represents the number of neighbors for each node. We observed that on average each disease shares direct links with seven other diseases ($K = 7$, Figure 3). With links to 32 diseases, hypothyroidism had the highest degree property ($K = 32$) in the network. In hypothyroidism, a disorder of the endocrine system, the thyroid gland does not produce enough thyroid hormones, and this deficiency can lead to the development of other diseases. Some comorbidities observed in the DDN were morbid obesity,[32,33] type 2 diabetes mellitus (MIM: 125853),[34] vitamin D deficiency,[35] hypertensive heart disease,[36] thyroid cancer, and rheumatoid arthritis.[37] On the other end of the scale, five diseases (blepharitis; "acute, but ill-defined, cerebrovascular disease; hyposmolality and/or hyponatremia; pain in joint; and goiter) had links to only one neighboring disease ($K = 1$). Thus, representing cross-phenotype associations in the form of networks enabled visualization of complex interconnections between different diseases.

Hub Diseases in the DDN

To further characterize the DDN, we applied different network statistics to identify disease nodes necessary for the cohesiveness of the network. Such nodes are also commonly referred to as hub nodes (see Material and Methods). We used a betweenness centrality measure to identify hub nodes, which are represented in the DDN by larger nodes (Figure 2). We identified many hub nodes in different disease classes across the DDN; the highest number were in endocrine disorders and included hypothyroidism, type 1 diabetes, and type 2 diabetes (Figure 2). Other main hub nodes that we observed in the DDN were psoriasis, morbid obesity, multiple sclerosis, rheumatoid arthritis, coronary atherosclerosis, and chronic kidney disease.

Identifying Biologically Relevant Subnetworks via Epigenomics

These results demonstrate that community detection is a good approach to visualizing the global and local structures

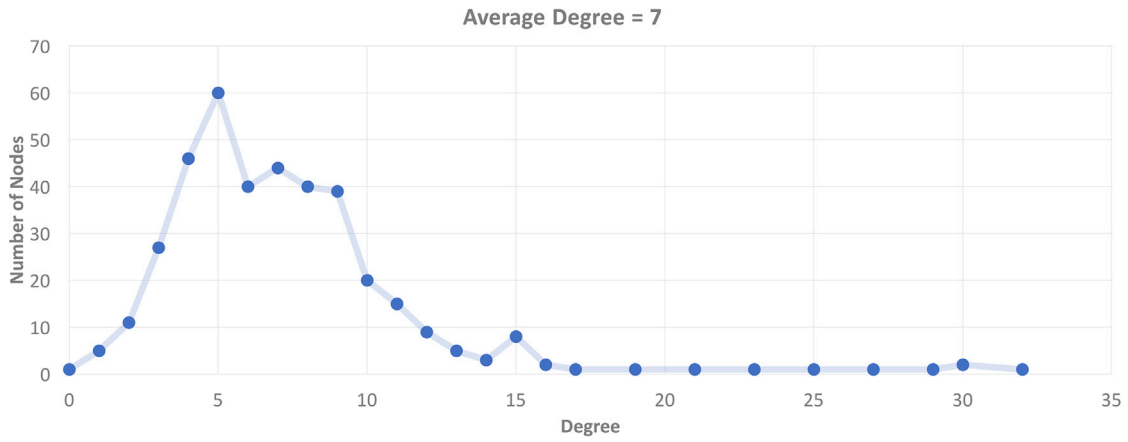


Figure 3. Disease Neighbors
In a network, the degree property is the number of direct connections between one node and other nodes. This plot presents the distribution of degrees observed in the DDN.

of disease interaction. To further test whether the disease nodes and the connections between them are relevant to molecular mechanisms of disease, we incorporated chromatin-state annotations from the Roadmap Epigenomics Consortium and used them to extract biologically relevant subnetworks by using a similar approach. We only considered SNPs within enhancer regions for specific tissues for the current analysis. Seven tissue-specific DDNs were constructed from the shared variants in enhancer regions. The largest observed subnetwork where SNPs were in active enhancer regions was in the liver. The associated diseases for this tissue included 19 diseases, such as cirrhosis of the liver, chronic non-alcoholic liver disease, hyperlipidemia, morbid obesity, essential hypertension, and cardiovascular diseases, among others (Table S2). For adipose tissue, there were eight diseases in the subnetwork, including links between cardiovascular, nutritional, endocrine, and autoimmune diseases (Figure 4). Only two of the nodes in this subnetwork were connected to each other. Within the adipose subnetwork, we observed connections between cardiovascular diseases such as peripheral vascular disease, myocardial infarction, coronary artery disease, and abdominal aneurysm. Supporting these connections, previous studies have reported known links between increased gene expression in adipose tissue and cardiovascular diseases.[24,25] The second node was for type 1 diabetes, which had connections to psoriasis and Raynaud syndrome. Psoriasis and type 1 diabetes are both autoimmune diseases, and they share associations with the variation in the human leukocyte antigen (HLA) region. Numerous studies have identified strong connections between the pathogenesis of these autoimmune diseases and variations in HLA.[38,39]

Community Detection

EHR data provide a vast amount of information pertaining to diseases. Machine-learning approaches are being applied to longitudinal EHR data so that predictive models of disease correlations, risk predictions, and comorbidities can be developed.[40–42] EHR-based predictive models can be used for combining disease connections into a network similar to the DDN. To compare the DDN with networks from longitudinal EHR data, we applied a probabilistic relationship model to ICD-9 diagnoses derived from the same Geisinger longitudinal EHR data (unpublished data). These prediction models were developed under an Ising model framework,[43] and all the predictions were based on EHR data alone. The Ising model is a type of Markov random field (MRF) graphical model for binary data.[44] It provides an approximation of the full joint-probability distribution across hundreds of ICD-9 codes. Thus, it can help to uncover patterns of dependencies between ICD-9 codes that result from either shared genetic or environmental architecture. This predictive algorithm generated a graphical model of disease states for 500 ICD-9 codes; this model is a representation of similarities between ICD-9 codes. Then we evaluated whether we observed the same links that we identified in the PheWAS-derived DDN.

Rather than comparing all the disease connections, which would be computationally intensive, we applied the community-detection method in Gephi to the DDN in order to find subnetworks algorithmically. The method found nine communities; as shown in Figure 5, the number of diseases in each community varied between clusters of 2 and 102.

Next, we selected one community that encompassed 20 diseases and showed connections between different disease classes, such as nutritional, neurological, cardiovascular, skin, and digestive-system disorders (Figure 6A). We compared this subnetwork of the DDN with the network derived from probabilistic graphical model of disease state, wherein disease state is defined as the status of all ICD-9 code diagnoses in an individual's EHR. We used the Ising model framework to develop the probabilistic graphical model of disease state. We checked to see whether we could observe some of the links we identified in our DDN subnetwork (identified via community detection) in the Ising

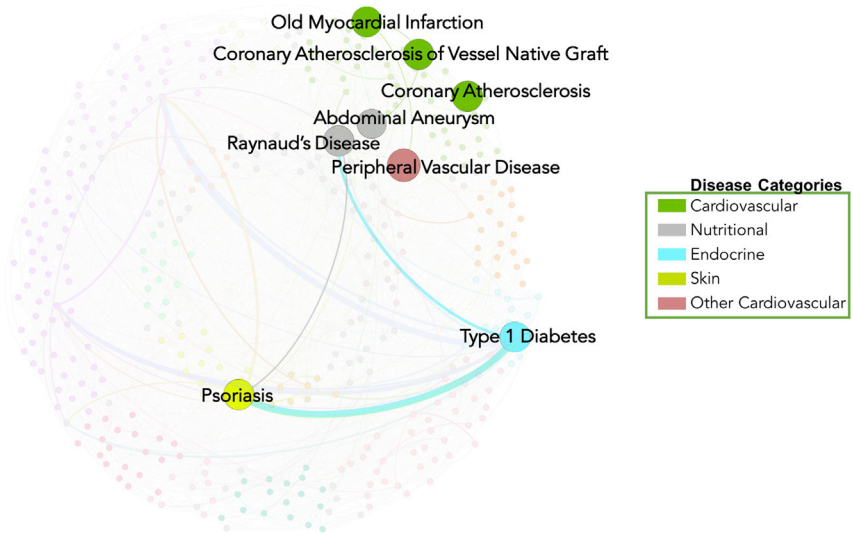


Figure 4. Diseases with Shared Enhancers in Adipose Tissue
The highlighting of disease nodes in the network indicates that the shared SNPs between these diseases are located in the enhancer region of the nearby gene.

model of disease state (Figure 6B). Through this independent investigation, we identified direct and indirect connections between ICD-9 codes in the Ising model network; these connections were similar to those found in the DDN. Thus, we demonstrated a probabilistic dependence between these diagnosis codes in line with what we see in our network. When we compared the morbid obesity associated with diseases directly neighboring one another in both the DDN and the Ising model (Figure 6), we found many similarities. Specifically, the comorbidities that showed direct links to morbid obesity in both networks were sleep apnea,[45] lumbago,[46] and edema.[47] These results suggest that the probabilistic dependencies observed between these diseases in the Ising model network can probably be explained by the shared genetic architecture that was identified through the DDN. In the DDN, we also found links between morbid obesity and cardiovascular diseases (coronary atherosclerosis and intermediate coronary syndrome), which are known comorbidities.[45] Other interesting links with morbid obesity were bariatric-surgery-associated conditions, such as post-gastric absorption and post-surgical non-absorption. It is possible that these connections might be due to a diagnosis correlation that arose in the EHR when an individual underwent bariatric surgery because of their pre-existing condition of morbid obesity. Gout was also a comorbidity of morbid obesity.[45] However, these diseases were connected indirectly through another comorbidity: sleep apnea. With this example, we highlight the core strength of EHR-based studies, which allow us to answer similar questions about disease relationships with different methods and thereby provide more robustness to the findings.

Discussion

In this study, we generated and evaluated a network of cross-phenotype associations derived from an EHR-based PheWAS. In contrast to previous disease networks, which were built of summary statistics from disparate studies, the DDN benefits from utilizing a single source of EHR data. The network analyses performed on the DDN have illuminated deeper structures within and across disease classes. For example, autoimmune diseases are caused by dysfunctional immune systems that attack the healthy cells in a variety of organs. Type 1 diabetes, rheumatoid arthritis, and multiple sclerosis were some of the common autoimmune conditions within the DDN. Although these conditions have distinct symptoms, previous findings have shown strong evidence that complex interactions occur between these diseases as a result of shared genetic architecture.[48,49] The identification of these previously known findings regarding these autoimmune diseases provides support for the network approach of investigating cross-phenotype associations derived from PheWASs.

In this study, the SNPs linking these autoimmune diseases mapped to 19 genes, variations in all of which were associated with increased risk of autoimmune disease (Table S2). Two genes, *C6orf10* (chromosome 6 open reading frame 10 [MIM: 618151]) and *TAP2* (transporter 2, ATP-binding cassette, subfamily B [MIM: 170261]), were the only two genes linked to three autoimmune diseases: type 1 diabetes, rheumatoid arthritis, and multiple sclerosis. Of the 19 genes, *C2* (complement component 2 [MIM: 613927]), *HCG26* (HLA complex group 26 [HGNC: 29671]), and *PSMB8* (proteasome subunit beta 8 [MIM: 177046]) had no previously known associations with autoimmune diseases. However, we replicated the findings of a genetic study of one of the largest European American cohorts (UK Biobank), which revealed associations between rheumatoid arthritis and multiple sclerosis.[54] Additionally, we performed a gene ontology (GO) enrichment analysis with genes shared between type 1 diabetes, multiple sclerosis, and rheumatoid arthritis. Notably, many immune-system-process-related GO terms were identified (Table S3). Using epigenomics, we found that a variant in *HCG26*, one of the 19 genes, is located in the enhancer region targeting *LTA* (lymphotoxin alpha [MIM: 153440]); the variant was identified in multiple tissues by the fine mapping approach described in *Verma et al.*.[13] (dbSNP: rs2523663). *LTA* is a protein-coding gene that encodes cytokines produced by lymphocytes in the immune system (see NCBI in Web Resources). Cytokines play an important role in the pathogenesis of various autoimmune disorders,

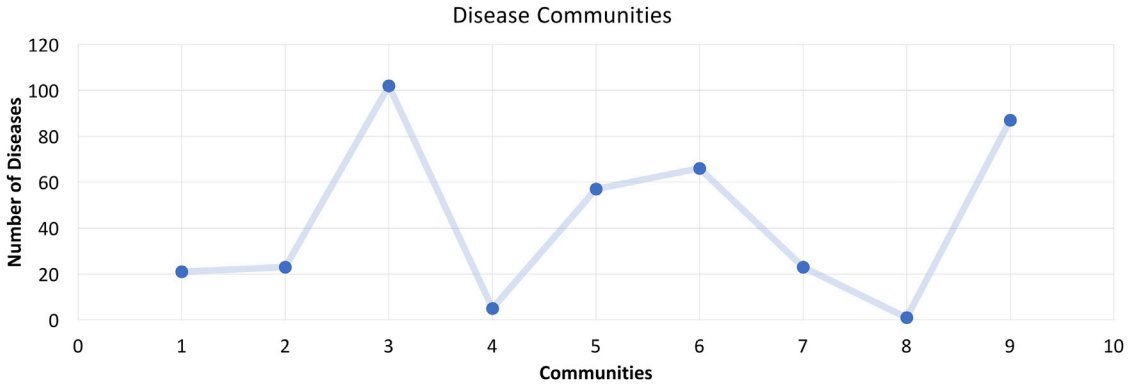


Figure 5. Disease Communities
The plot shows the distribution of community disease connections, which were identified by community detection. The x axis shows the total number of communities identified, and the y axis shows the number of disease nodes in each community.

and cytokine-inhibiting agents are key drug targets for type 1 diabetes and multiple sclerosis.[50–53] Because of the many key genes shared between connected diseases, along with the epigenetic regulation, cytokine-inhibiting agents may offer intervention strategies to satisfy the unmet medical needs that still exist in those connected diseases.

Additionally, we identified previously unreported disease connections by using the DDN approach. For example, we found that links between morbid obesity and its known comorbidities can be explained by shared genetic associations. These comorbidities were not present in the Human Disease Network (Figure S1). This inconsistency might be explained by differences in the phenotypes used to construct the network. We also demonstrated similarities between networks formed from two distinct predictive algorithms from the same EHR system. Taken together, these results suggest that the probabilistic dependencies observed between certain diseases (e.g., morbid obesity, sleep apnea, lumbago, gout, venous insufficiency, and edema) in the Ising model can be explained by shared genetic architecture identified via our disease-disease network. With this example, we highlight the core strength of EHR-based studies: the ability to apply different approaches, such as using genetic and/or phenotypic information, in order to arrive at a stronger conclusion.

The potential strength of the DDN is to identify disease connections that were not expected. From the DDN generated in this study, we found that hyperlipidemia was linked to not only atherosclerosis, but also many immune-related diseases, such as type I diabetes, psoriasis, hypothyroidism, and multiple sclerosis, as well as other immune-mediated diseases, such as allergic rhinitis, blepharitis, acute bronchitis, and herpes. These unexpected observations indicate the non-canonical role of the immune system in lipid-metabolizing disorders and/or the pathogenic role of hyperlipidemia in immune responses.

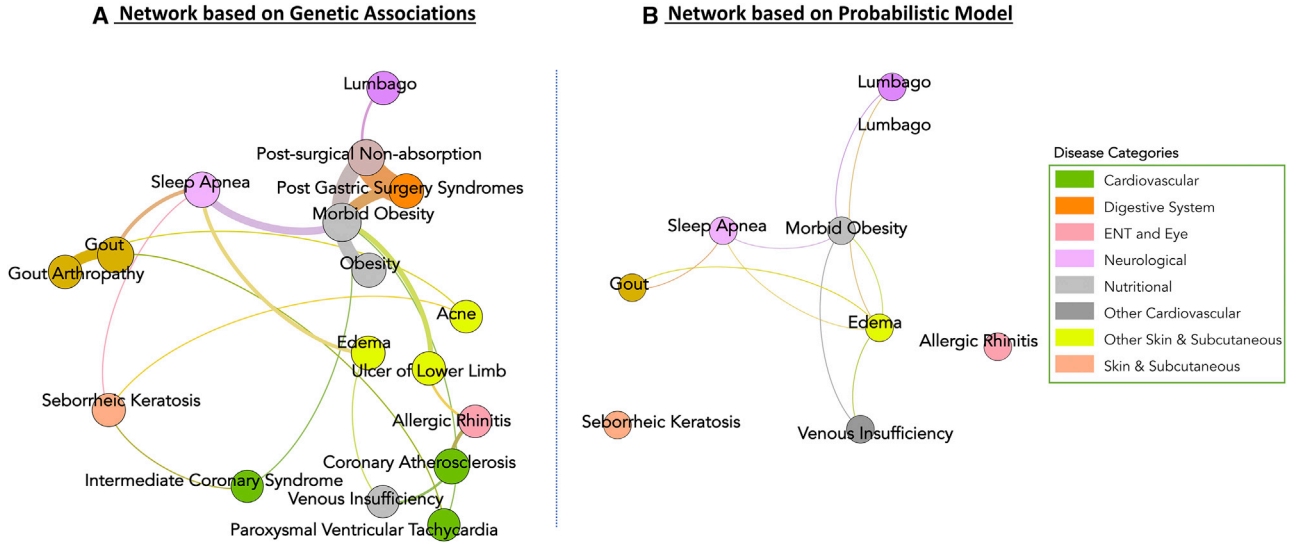


Figure 6. Comparison of Disease-Disease Network Construction through Two Orthogonal Approaches
The figure illustrates the similarities between the disease network that was constructed on the basis of genetic associations (the DDN) (A) and the probabilistic model created from longitudinal EHR data (the Ising model) (B).

Indeed, lymphotoxin (LT) and LIGHT, two tumor necrosis factor cytokine family members that are primarily expressed on lymphocytes, are critical regulators of key enzymes that control lipid metabolism in mouse models.[54] Although further studies are warranted to infer the causality of these associations, our DNN confirmed the shared genetic risk of hyperlipidemia and immune diseases.

In conclusion, community detection is a powerful method to identify and visualize cross-phenotype associations from analysis of PheWASs. It uncovered previously unreported shared links in known interactions between diseases, as well as other unreported connections between diseases. It also provided a way to generate new hypotheses to guide further targeted investigation into comorbidities, pleiotropy, and epistasis. Although we explored interconnections between multiple diseases in an EHR-based population, this approach can also be applied to many publicly available resources with summary-level and individual-level data on multiple phenotypes. Networks similar to the ones generated here could be adapted from NHANES, UK BioBank,[55] GERA,[56] eMERGE,[57] and the Million Veteran Program, among other populations. Furthermore, we plan to extend the network analysis by including associations between genetic variants and clinical laboratory measures in EHR. This work provides new avenues by which network-based methods can be applied to large, gene-trait-based studies to uncover the genetic underpinnings of disease.

Lastly, an interactive visualization tool of the disease-disease network is available (see Web Resources).

Supplemental Data

Supplemental Data include one figure and three tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.11.006.

Acknowledgments

This work was supported by the National Library of Medicine (NLM) R01 NL012535. This project is also funded, in part, by a grant provided by the Pennsylvania Department of Health (#SAP 4100070267). The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Declaration of Interests

The authors declare no competing interests.

Received: July 12, 2018
Accepted: November 12, 2018
Published: January 3, 2019

Web Resources

Disease-Disease Network Visualization Tool, https://www.biomedinfolab.com/software
eMERGE, https://emerge.mc.vanderbilt.edu
Gephi, https://gephi.org
Million Veteran Program, https://www.research.va.gov/mvp/
NCBI, https://www.ncbi.nlm.nih.gov/gene/4049
NHANES, https://www.cdc.gov/nchs/nhanes/index.htm
OMIM, http://www.omim.org/
Roadmap Epigenomics Project, http://www.roadmapepigenomics.org/data/
UK BioBank, https://www.ukbiobank.ac.uk

References

1. Tyler, A.L., Crawford, D.C., and Pendergrass, S.A. (2016). The detection and characterization of pleiotropy: Discovery, progress, and promise. Brief. Bioinform. *17*, 13–22.
2. Cronin, R.M., Field, J.R., Bradford, Y., Shaffer, C.M., Carroll, R.J., Mosley, J.D., Bastarache, L., Edwards, T.L., Hebbring, S.J., Lin, S., et al. (2014). Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. Front. Genet. *5*, 250.
3. Hall, M.A., Verma, A., Brown-Gentry, K.D., Goodloe, R., Boston, J., Wilson, S., McClellan, B., Sutcliffe, C., Dilks, H.H., Gillani, N.B., et al. (2014). 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.
4. Oetjens, M.T., Bush, W.S., Denny, J.C., Birdwell, K., Kodaman, N., Verma, A., Dilks, H.H., Pendergrass, S.A., Ritchie, M.D., and Crawford, D.C. (2016). Evidence for extensive pleiotropy among pharmacogenes. Pharmacogenomics *17*, 853–866.
5. MacArthur, J., Bowler, E., Cerezo, M., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.
6. Darabos, C., Grussing, E.D., Cricco, M.E., Clark, K.A., and Moore, J.H. (2015). A bipartite network approach to inferring interactions between environmental exposures and human diseases. Pac. Symp. Biocomput., 171–182.
7. Moore, J.H., and Williams, S.M. (2015). Epistasis: Methods and protocols (Humana Press).
8. Pendergrass, S.A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E.S., Goodloe, R., Ambite, J.L., Avery, C.L., Buyske, S., Bůžková, P., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet. *9*, e1003087.
9. Namjou, B., Marsolo, K., Caroll, R.J., Denny, J.C., Ritchie, M.D., Verma, S.S., Lingren, T., Porollo, A., Cobb, B.L., Perry, C., et al. (2014). Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to eosinophilic esophagitis. Front. Genet. *5*, 401.
10. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.
11. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA *104*, 8685–8690.
12. Richesson, R.L., Rusincovitch, S.A., Wixted, D., Batch, B.C., Feinglos, M.N., Miranda, M.L., Hammond, W.E., Califf, R.M., and Spratt, S.E. (2013). A comparison of phenotype

definitions for diabetes mellitus. J. Am. Med. Inform. Assoc. *20* (e2), e319–e326.

13. Verma, A., Lucas, A., Verma, S.S., Zhang, Y., Josyula, N., Khan, A., Hartzel, D.N., Lavage, D.R., Leader, J., Ritchie, M.D., and Pendergrass, S.A. (2018). PheWAS and beyond: The landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger. Am. J. Hum. Genet. *102*, 592–608.
14. Lingren, T., Chen, P., Bochenek, J., Doshi-Velez, F., Manning-Courtney, P., Bickel, J., Wildenger Welchons, L., Reinhold, J., Bing, N., Ni, Y., et al. (2016). Electronic health record based algorithm to identify patients with autism spectrum disorder. PLoS ONE *11*, e0159621.
15. Mayer, M.A., Gutierrez-Sacristan, A., Leis, A., De La Peña, S., Sanz, F., and Furlong, L.I. (2017). Using electronic health records to assess depression and cancer comorbidities. Stud. Health Technol. Inform. *235*, 236–240.
16. Rassekh, S.R., Lorenzi, M., Lee, L., Devji, S., McBride, M., and Goddard, K. (2010). Reclassification of ICD-9 codes into meaningful categories for oncology survivorship research. J. Cancer Epidemiol. *2010*, 569517.
17. Schork, A.J., Thompson, W.K., Pham, P., Torkamani, A., Roddey, J.C., Sullivan, P.F., Kelsoe, J.R., O'Donovan, M.C., Furberg, H., Schork, N.J., et al.; Tobacco and Genetics Consortium; Bipolar Disorder Psychiatric Genomics Consortium; and Schizophrenia Psychiatric Genomics Consortium (2013). All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. *9*, e1003449.
18. Adams, W.T., and Skopek, T.R. (1987). Statistical test for the comparison of samples from mutational spectra. J. Mol. Biol. *194*, 391–396.
19. Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.
20. Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257–258.
21. Markello, T.C., Carlson-Donohoe, H., Sincan, M., Adams, D., Bodine, D.M., Farrar, J.E., Vlachos, A., Lipton, J.M., Auerbach, A.D., Ostrander, E.A., et al. (2012). Sensitive quantification of mosaicism using high density SNP arrays and the cumulative distribution function. Mol. Genet. Metab. *105*, 665–671.
22. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.
23. Fortunato, S., and Barthélemy, M. (2007). Resolution limit in community detection. Proc. Natl. Acad. Sci. USA *104*, 36–41.
24. Nguyen, C., Varney, M.D., Harrison, L.C., and Morahan, G. (2013). Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms. Diabetes *62*, 2135–2140.
25. Clauset, A., Newman, M.E.J., and Moore, C. (2004). Finding community structure in very large networks. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *70*, 066111.
26. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.
27. Verma, S.S., Frase, A.T., Verma, A., Pendergrass, S.A., Mahony, S., Haas, D.W., and Ritchie, M.D. (2016). Phenome-wide interaction study (PheWIS) in aids clinical trials group data (ACTG). Pac. Symp. Biocomput. *2016*, 57–68.
28. Somers, E.C., Thomas, S.L., Smeeth, L., and Hall, A.J. (2009). Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder? Am. J. Epidemiol. *169*, 749–755.
29. Liao, K.P., Gunnarsson, M., Källberg, H., Ding, B., Plenge, R.M., Padyukov, L., Karlson, E.W., Klareskog, L., Askling, J., and Alfredsson, L. (2009). Specific association of type 1 diabetes mellitus with anti-cyclic citrullinated peptide-positive rheumatoid arthritis. Arthritis Rheum. *60*, 653–660.
30. Tettey, P., Simpson, S., Jr., Taylor, B.V., and van der Mei, I.A.F. (2015). The co-occurrence of multiple sclerosis and type 1 diabetes: Shared aetiologic features and clinical implication for MS aetiology. J. Neurol. Sci. *348*, 126–131.
31. Tseng, C.-C., Chang, S.-J., Tsai, W.-C., Ou, T.-T., Wu, C.-C., Sung, W.-Y., Hsieh, M.-C., and Yen, J.-H. (2016). Increased incidence of rheumatoid arthritis in multiple sclerosis: A nationwide cohort study. Medicine (Baltimore) *95*, e3999.
32. Sanyal, D., and Raychaudhuri, M. (2016). Hypothyroidism and obesity: An intriguing link. Indian J. Endocrinol. Metab. *20*, 554–557.
33. Michalaki, M.A., Vagenakis, A.G., Leonardou, A.S., Argentou, M.N., Habeos, I.G., Makri, M.G., Psyrogiannis, A.I., Kalfarentzos, F.E., and Kyriazopoulou, V.E. (2006). Thyroid function in humans with morbid obesity. Thyroid *16*, 73–78.
34. Wang, C. (2013). The relationship between type 2 diabetes mellitus and related thyroid diseases. J. Diabetes Res. *2013*, 390534.
35. Mackawy, A.M.H., Al-Ayed, B.M., and Al-Rashidi, B.M. (2013). Vitamin d deficiency and its association with thyroid disease. Int. J. Health Sci. (Qassim) *7*, 267–275.
36. Grais, I.M., and Sowers, J.R. (2014). Thyroid and the heart. Am. J. Med. *127*, 691–698.
37. Staykova, N.D. (2007). Rheumatoid arthritis and thyroid abnormalities. Folia Med. (Plovdiv) *49*, 5–12.
38. Granata, M., Skarmoutsou, E., Trovato, C., Rossi, G.A., Mazzarino, M.C., and D'Amico, F. (2017). Obesity, type 1 diabetes, and psoriasis: An autoimmune triple flip. Pathobiology *84*, 71–79.
39. Chen, H., Hayashi, G., Lai, O.Y., Dilthey, A., Kuebler, P.J., Wong, T.V., Martin, M.P., Fernandez Vina, M.A., McVean, G., Wabl, M., et al. (2012). Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. PLoS Genet. *8*, e1002514.
40. Wu, J., Roy, J., and Stewart, W.F. (2010). Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Med. Care *48* (6, Suppl), S106–S113.
41. Kennedy, E.H., Wiitala, W.L., Hayward, R.A., and Sussman, J.B. (2013). Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med. Care *51*, 251–258.
42. Jensen, P.B., Jensen, L.J., and Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. Nat. Rev. Genet. *13*, 395–405.
43. Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. Z. Für Physiol. *31*, 253–258.

44. Kindermann, R., and Snell, J.L. (1980). Markov Random Fields and Their Applications (American Mathematical Society).
45. Khaodhiar, L., McCowen, K.C., and Blackburn, G.L. (1999). Obesity and its comorbid conditions. Clin. Cornerstone *2*, 17–31.
46. Topsakal, S., Erurker, T., Akin, F., Yaylali, G.F., Yerlikaya, E., and Kaptanoglu, B. (2014). Heel pain and comorbid conditions in obese patients. J. Musculoskeletal Pain *22*, 38–42.
47. Todd, M. (2009). Managing chronic oedema in the morbidly obese patient. Br. J. Nurs. *18*, 1120–1124.
48. Roizen, J.D., Bradfield, J.P., and Hakonarson, H. (2015). Progress in understanding type 1 diabetes through its genetic overlap with other autoimmune diseases. Curr. Diab. Rep. *15*, 102.
49. Richard-Miceli, C., and Criswell, L.A. (2012). Emerging patterns of genetic overlap across autoimmune disorders. Genome Med. *4*, 6.
50. Moudgil, K.D., and Choubey, D. (2011). Cytokines in autoimmunity: Role in induction, regulation, and treatment. J. Interferon Cytokine Res. *31*, 695–703.
51. Li, P., Zheng, Y., and Chen, X. (2017). Drugs for autoimmune inflammatory diseases: From small molecule compounds to anti-TNF biologics. Front. Pharmacol. *8*, 460.
52. Rasouli, J., Ciric, B., Imitola, J., Gonnella, P., Hwang, D., Mahajan, K., Mari, E.R., Safavi, F., Leist, T.P., Zhang, G.-X., and Rostami, A. (2015). Expression of GM-CSF in T cells is increased in multiple sclerosis and suppressed by IFN-β therapy. J. Immunol. *194*, 5085–5093.
53. Rabinovitch, A., and Suarez-Pinzon, W.L. (2007). Roles of cytokines in the pathogenesis and therapy of type 1 diabetes. Cell Biochem. Biophys. *48*, 159–163.
54. Lo, J.C., Wang, Y., Tumanov, A.V., Bamji, M., Yao, Z., Reardon, C.A., Getz, G.S., and Fu, Y.-X. (2007). Lymphotoxin beta receptor-dependent control of lipid homeostasis. Science *316*, 285–288.
55. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.
56. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A., et al. (2015). Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. Genetics *200*, 1051–1060.
57. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M., et al.; eMERGE Team (2011). The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics *4*, 13.

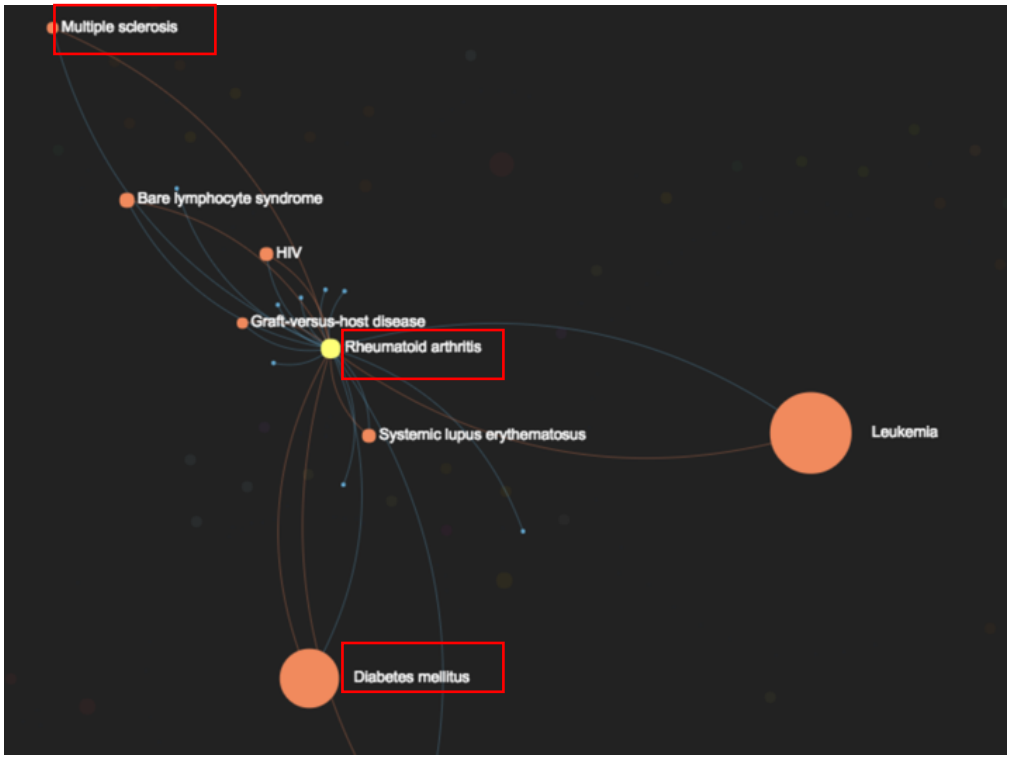
The American Journal of Human Genetics, Volume *104*

Supplemental Data

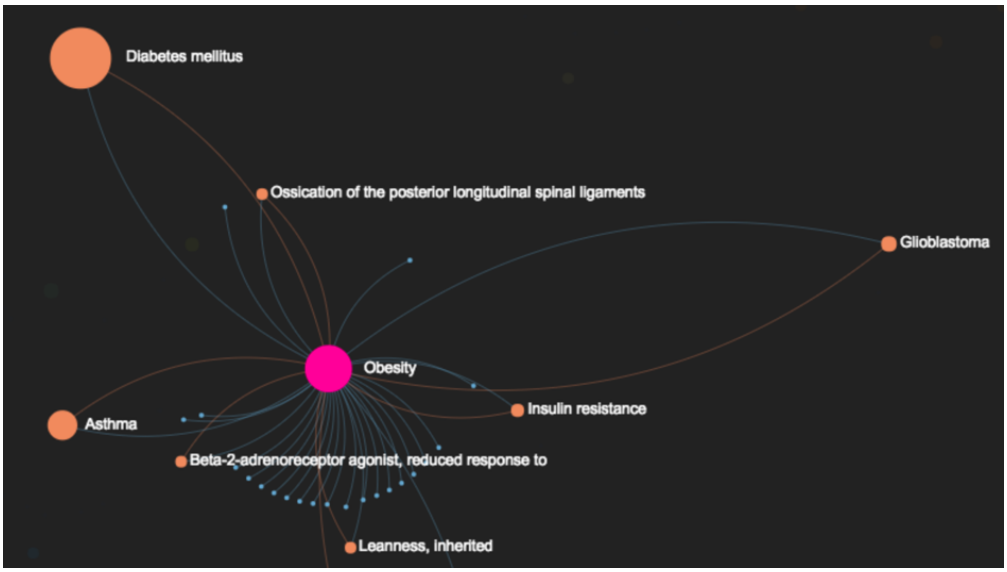
Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals

Anurag Verma, Lisa Bang, Jason E. Miller, Yanfei Zhang, Ming Ta Michael Lee, Yu Zhang, Marta Byrska-Bishop, David J. Carey, Marylyn D. Ritchie, Sarah A. Pendergrass, Dokyoon Kim, and the DiscovEHR Collaboration

Supplementary Information



(A)



(B)

Figure S1. Network similarities with human disease network. (A) The disease connection between the three highlighted autoimmune diseases (Red box) in HDN are same as observed in the DDN. (B) Obesity and its neighboring diseases in HDN and the highlighted red box represent the connections found in our network as well.

Gene	Diseases	Shared SNPs	Known GWAS Associations	Enhancer Region Variants	Tissues
PTPN22	Type 1 Diabetes Multiple sclerosis	1	Previously Known	exm85427	Breast Dnd41_TCell_Leukemia
BAG6	Type 1 Diabetes Multiple sclerosis	1	Previously Known	rs760293	Adrenal HepG2_Hepatocellular_Carcinoma Liver
BTNL2	Type 1 Diabetes Rheumatoid arthritis	2	Previously Known	rs3129953	Breast Cervix
C2	Type 1 Diabetes Multiple sclerosis	1	Novel	N/A	N/A
C6orf10	Type 1 Diabetes Multiple sclerosis Rheumatoid arthritis	4	Previously Known	N/A	N/A
CFB	Type 1 Diabetes Multiple sclerosis	1	Previously Known	rs1048709	Pancreas Placenta
HCG20	Type 1 Diabetes Rheumatoid arthritis	1	Previously Known	rs6920124	Blood Bone Cervix Dnd41_TCell_Leukemia GI Smooth Muscle HepG2_Hepatocellular_Carcinoma Stromal Connective Stem cells Thymus
HCG26	Type 1 Diabetes Multiple sclerosis	1	Novel	rs2523663	Blood Heart Spleen
HLA-DOB	Type 1 Diabetes Multiple sclerosis	1	Previously Known	rs2071469	Placenta Stromal Connective Stem cells
HLA-DQA1	Type 1 Diabetes Multiple sclerosis	1	Novel	N/A	N/A
HLA-DQA2	Type 1 Diabetes Multiple sclerosis	1	Previously Known	N/A	N/A
HLA-DQB1	Type 1 Diabetes Multiple sclerosis	1	Novel	N/A	N/A
HLA-DRA	Type 1 Diabetes Multiple sclerosis	1	Previously Known	N/A	N/A
LOC101929072	Type 1 Diabetes Rheumatoid arthritis	1	Previously Known	rs2251396	A549_EtOH_0.02pct_Lung_Carcinoma Blood Bone Breast Dnd41_TCell_Leukemia Fat Skin (Adipose Tissue)
LOC102725019	Type 1 Diabetes Multiple sclerosis	1	Previously Known	N/A	N/A
NOTCH4	Type 1 Diabetes Multiple sclerosis	2	Previously Known	N/A	N/A
PSMB8	Type 1 Diabetes Multiple sclerosis	1	Novel	N/A	N/A
TAP1	Type 1 Diabetes Multiple sclerosis	1	Previously Known	rs3198005	Breast Cervix
TAP2	Type 1 Diabetes Multiple sclerosis Rheumatoid arthritis	5	Previously Known	rs3819721 rs241426 rs3819714	Blood Breast Dnd41_TCell_Leukemia HepG2_Hepatocellular_Carcinoma Placenta Thymus Cervix

Table S2. Shared SNPs between disease network. Here we present all the SNPs shared between key hub nodes in the network i.e. Type 1 Diabetes, Multiple Sclerosis, and Rheumatoid Arthritis.

.

Gene ontology term	p-value	FDR q-value
Antigen processing and presentation ion of peptide antigen	$9.74e^{-16}$	$5.76e^{-12}$
MHC class II receptor activity	$2.87e^{-15}$	$6.47e^{-12}$
Immune system process	$3.53e^{-15}$	$6.47e^{-12}$
Antigen processing and presentation	$4.38e^{-15}$	$6.47e^{-12}$
MHC class II protein complex	$1.58e^{-14}$	$1.87e^{-11}$
Positive regulation of immune system process	$9.36e^{-14}$	$9.23e^{-11}$
Regulation of immune system process	$2.16e^{-13}$	$1.83e^{-10}$
MHC protein complex	$2.91e^{-13}$	$2.15e^{-10}$
Immune response	$9.9e^{-13}$	$6.51e^{-10}$
Regulation of immune resonse	$5.62e^{-12}$	$3.33e^{-9}$

Table S3. Gene Ontology enrichment analysis using shared genes between Type 1 Diabetes, Multiple Sclerosis, and Rheumatoid Arthritis. P-values were obtained from the hypergeometric distribution