




OPEN

Visualizing novel connections and genetic similarities across diseases using a network-medicine based approach

Brian Ferolito^{1,6}, Italo Faria do Valle^{1,2,6}, Hanna Gerlovin¹, Lauren Costa¹, Juan P. Casas^{1,5}, J. Michael Gaziano^{1,5}, David R. Gagnon^{1,3}, Edmon Begoli⁴, Albert-László Barabási² & Kelly Cho^{1,5,6}

Understanding the genetic relationships between human disorders could lead to better treatment and prevention strategies, especially for individuals with multiple comorbidities. A common resource for studying genetic-disease relationships is the GWAS Catalog, a large and well curated repository of SNP-trait associations from various studies and populations. Some of these populations are contained within mega-biobanks such as the Million Veteran Program (MVP), which has enabled the genetic classification of several diseases in a large well-characterized and heterogeneous population. Here we aim to provide a network of the genetic relationships among diseases and to demonstrate the utility of quantifying the extent to which a given resource such as MVP has contributed to the discovery of such relations. We use a network-based approach to evaluate shared variants among thousands of traits in the GWAS Catalog repository. Our results indicate many more novel disease relationships that did not exist in early studies and demonstrate that the network can reveal clusters of diseases mechanistically related. Finally, we show novel disease connections that emerge when MVP data is included, highlighting methodology that can be used to indicate the contributions of a given biobank.

Disease comorbidity, or the co-occurrence of diseases within a single individual, is a major clinical problem, posing challenges in prognosis and treatment, increasing health care costs, and reducing life expectancy^{1,2}. Comorbidities suggest common mechanisms that underlie different diseases, which can be either genetic or environmental³. Recent network-medicine based approaches have systematically studied the relationships across hundreds of diseases, using either molecular or clinical data. For example, Goh et al.⁴ created a network in which diseases are connected if they are associated to the same gene or genetic variant, and Hidalgo et al.⁵ built a network that mapped all correlations observed in the medical records of millions of patients. These approaches have the power to reveal insights that are not apparent when diseases are studied in isolation, offering a holistic approach to investigate diseases and how they are related. In fact, network-medicine based approaches have highlighted groups of disorders connected to the same molecular and metabolic mechanisms^{4,6–8}, comorbidities driven by age⁹, gender^{9–11}, demographic factors⁵ or by the same environmental triggers¹².

Recent advances in technology and computing power have allowed an exponential growth of data obtained by profiling thousands of patients. Large genomics initiatives across the world, such as the UK Biobank^{13,14}, Kaiser Permanente Research Program on Genes, Environment, and Health¹⁵, China Kadoorie Biobank¹⁶, and others, have profiled millions of patients through Genome-Wide Association Studies (GWAS), increasing our ability to investigate and understand the molecular and genetic origins of diseases. The Million Veteran Program¹⁷ (MVP)

¹VA Boston Healthcare System, Massachusetts Veterans Epidemiology and Research Information Center, (MAVERIC), 150 S. Huntington Avenue, Boston 02130, USA. ²Center for Complex Network Research, Department of Physics, Northeastern University, Boston 02115, USA. ³School of Public Health, Department of Biostatistics, Boston University, Boston 02215, USA. ⁴Oak Ridge National Laboratory, Oak Ridge 37830, USA. ⁵Brigham and Women's Hospital, Division of Aging, Department of Medicine, Harvard Medical School, Boston 02115, USA. ⁶These authors contributed equally: Brian Ferolito, Italo Faria do Valle and Kelly Cho. ✉email: brian.ferolito@va.gov

is one of such initiatives, which covers 825,000 patients in the United States from diverse ancestry backgrounds. At the current state, over 35 MVP research projects¹⁸ cover a wide range of high priority research areas including cardiovascular disease, mental health, substance abuse, cardiometabolic disease, urogenital disorders, diseases of the nervous system, cancer, pharmacogenomics, metabolism, infectious disease, and pain.

Mega-biobank repositories, as MVP's, contribute to the larger knowledgebase of genetic and disease mechanisms, and there is interest in being able to isolate the important and novel contributions of such initiatives to better target future efforts and research. Network-based approaches allow for the comparison of connected components that can also be further leveraged to focus causal inference towards genetic druggable targets, as well as, identifying pathways that are unique due to population stratification or genetic ancestries. We start by summarizing the current knowledge of genetic variants present in the GWAS Catalog, a curated public repository of genetic variant-phenotype associations from GWAS studies¹⁹. From the GWAS Catalog, we built a network where nodes represent single conditions and links represent shared genetic variants between a pair of diseases. We identified clusters of diseases based on the patterns of shared variants and compared the identified clusters with classical disease organization based on an anatomical system. We apply the novelty-comparison method to discover novel disease relationships for conditions, due to MVP's contribution, such as peripheral arterial disease, diabetes mellitus, and gout. Additionally, we show that these findings provide not only a high-level overview of our current understanding of genetic relationships among diseases, but also indicate new directions for further in-depth investigation, especially within particular ancestries, possibly offering new strategies for disease treatment and prevention.

Results

GWAS Catalog phenotypic network. We started by characterizing disease relationships arising from shared genetic variants among several diseases. To achieve this, we retrieved data from the GWAS Catalog, a curated public repository of variant-phenotype associations from eligible GWAS studies¹⁹. As of July 1st, 2020, the repository consisted of 3985 publications representing 113,841 genetic variants for 4298 unique traits. In this study, we focused only on 2764 disease-related traits from the full GWAS catalog, which included data from MVP as well as other sources. We eliminated many traits not directly associated with diseases from the analysis (See Methods). We then built a network in which nodes represent traits and links (or edges) connect traits that share variants. Each link contains a normalized measure of variant overlap between disease pairs (Jaccard Index), with its statistical significance being measured by the Fisher's Exact Test followed by Benjamini-Hochberg multiple testing correction, and links with $q > 0.05$ are filtered from the network. The final network contains 810 traits and 4980 links (Fig. 1). Node information and edge list for the Phenotypic Network can be found in Supplementary Tables 1 and 2, respectively.

In the overall phenotypic network, the traits with highest connectivity (k) were body mass index ($k = 154$), body height ($k = 146$), and systolic blood pressure ($k = 103$) as these are common anthropometric measurements included in large number of analyses. Specifically, for diseases, the most connected were schizophrenia ($k = 89$), type II diabetes mellitus ($k = 88$), and asthma ($k = 70$) (Table 1). As commonly observed in biological networks, our phenotypic network has a power law degree distribution, resulting in a network with a few nodes connected to many others, while most nodes have only a few connections (Fig. 2). The trait categories with the highest degree nodes were hematological and body measurements (Fig. 3). The pairs of traits with the highest overlap of genetic variants were systolic and diastolic blood pressure (1535); adolescent idiopathic scoliosis and scoliosis (1368); and basophil and neutrophil count (1076). We observed a high correlation between disease connectivity and total number of variants (Pearson $\rho = 0.866$, $p = 2.5 \times 10^{-245}$) as well as disease connectivity and number of studies for the disease (Pearson $\rho = 0.672$ and $p = 1.45 \times 10^{-107}$). These correlations with disease connectivity indicate that increased genetics data availability may make it more feasible to discover disease relationships not known before.

This can be demonstrated by comparing our results to previous disease networks. For example, Goh et al.⁴ mapped disease relationships using data from the Online Mendelian Inheritance in Man (OMIM) database. The authors report 7 diseases connected to schizophrenia and 11 connected to asthma, while our results report 89 and 70 connections, respectively.

Our results also highlight variants that connect the greatest number of disease pairs (Table 2). For example, the variant rs3184504 is shared between 641 disease pairs. This Single Nucleotide Polymorphism (SNP) is a missense variant found in the SH2B3 gene, which is a negative regulator of cytokine signaling, and an important component of the hematopoiesis pathway²⁰. The diseases in our network that contain the most edges with this variant are type I diabetes mellitus, rheumatoid arthritis, multiple sclerosis, inflammatory bowel disease, colorectal cancer, and prostate carcinoma.

Disease clusters. The identification of groups of diseases that are mechanistically related can offer insights about disease comorbidity and lead to better strategies for disease treatment and prevention. Here, we leveraged the patterns of connections in the Phenotypic Network to reveal diseases that are closely related. We applied the community detection algorithm Louvain²¹, which seeks to find groups of nodes more connected among themselves than with the rest of the network. We highlight that this method considers only the pattern of connections in the network and does not take disease classification into account. The largest connected component of our network is comprised of 22 communities with the remaining 39 communities occurring in isolated nodes. We focus our discussion of the communities on disease-related traits, i.e. not considering all traits classified in the following categories: other measurement, biological process, body measurement, lipid or lipoprotein measurement, response to drug, and hematological measurement (see Methods). Our results are consistent with previous

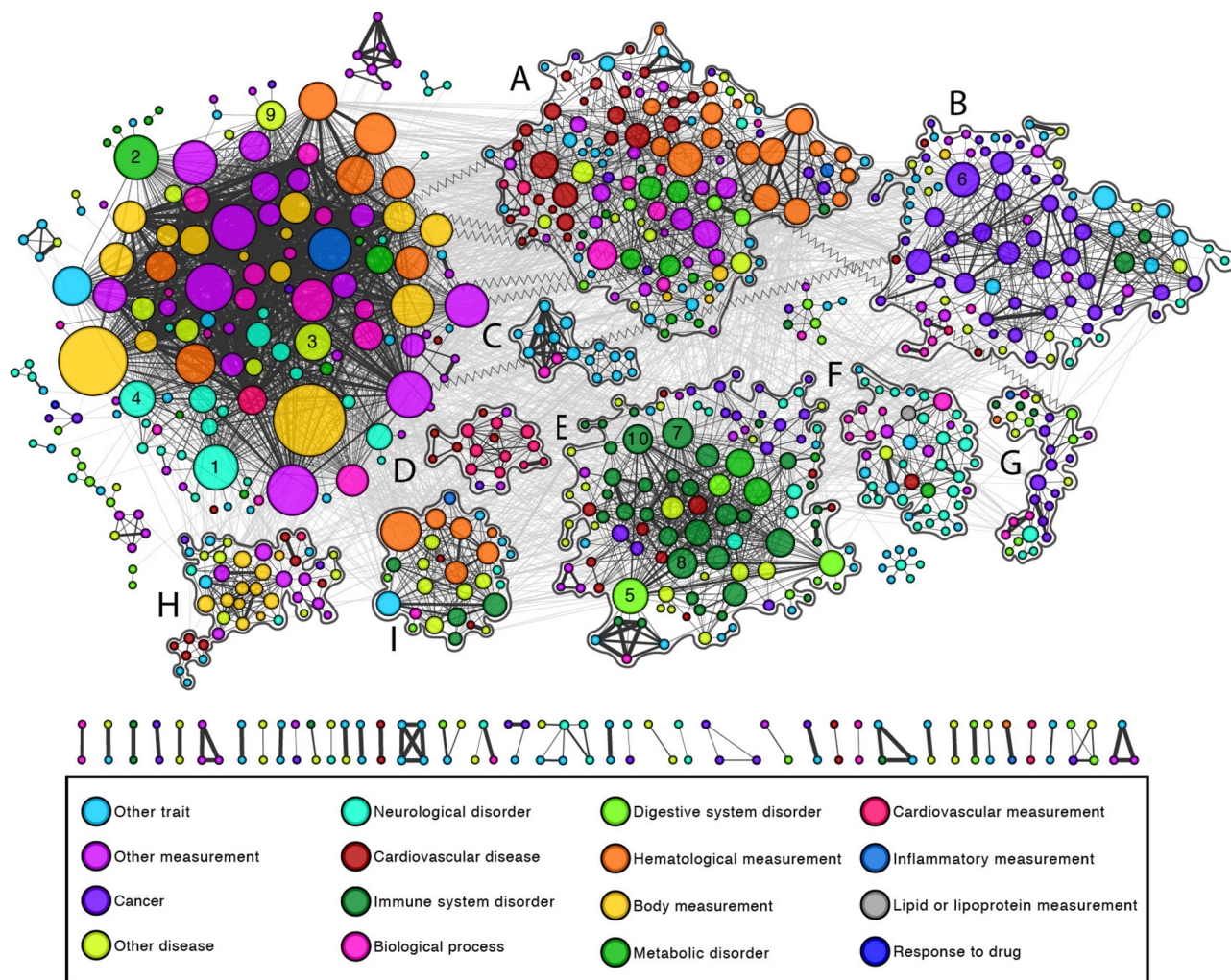


Figure 1. Phenotypic network assembled from GWAS catalog. Network in which nodes are traits that are connected with others to which they share genetic variants in the GWAS Catalog. The network communities detected are highlighted and labeled (A)–(H). Node colors represent disease categories and node size reflects connectivity in the network. The top high degree nodes are labeled 1–10 and their respective names are shown in Table 1. Only significant edges are shown (FDR < 0.05), the edge width indicates the overlap of variants between a pair of phenotypes (Jaccard Index), and lighter shade edges connect nodes in different communities.

Node	Degree	Centrality	Total variants	Responsible variants	Studies
Schizophrenia	89	0.141092	2497	1102	74
Type II diabetes mellitus	88	0.126487	1817	693	120
Asthma	70	0.119657	1617	870	66
Unipolar depression	68	0.119174	1763	954	64
Crohn's disease	67	0.090566	810	630	40
Breast carcinoma	64	0.097225	1046	225	66
Rheumatoid arthritis	56	0.058468	498	165	44
Ulcerative colitis	54	0.080451	692	595	27
Chronic obstructive pulmonary disease	52	0.1032	961	612	24
Psoriasis	52	0.068427	534	433	18

Table 1. High degree nodes of the phenotypic network. Table showing the top 10 most connected nodes, their corresponding eigenvector centrality, the total number of variants found for that trait, the number of those variants that are shared with other traits, and the number of unique papers reported for the traits in the database.

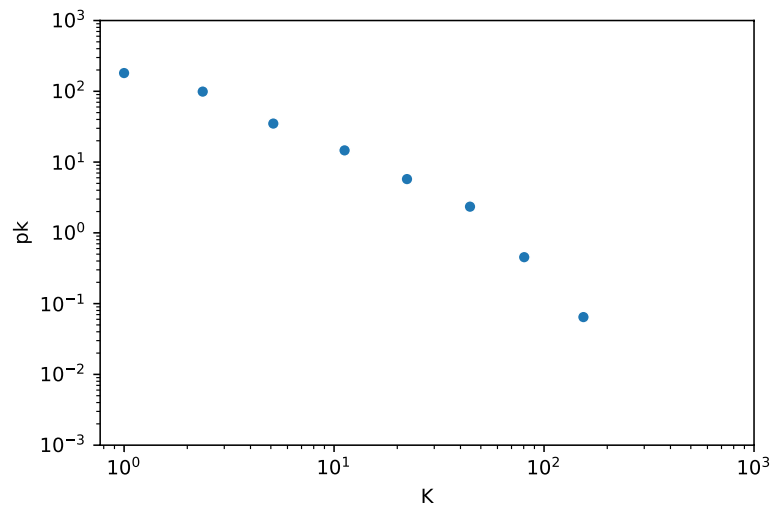


Figure 2. Degree distribution of phenotypic network. Log-binned degree distribution of the phenotypic network using a log–log scale. A power-law distribution which is a feature of scale-free networks. K represents the average degree of the bin where bin has size 2^{n-1} . pK is obtained from the number of nodes found in the bin divided by the width of the bin.

findings⁴ that clusters tend to aggregate diseases that share underlying mechanisms such as cancer, neurological, cardiovascular, and immune system disorders (Fig. 1).

Community E, the community with the most disease-related traits ($n = 105$) is characterized by disorders of the immune system and the most connected diseases in the community are Crohn's disease, rheumatoid arthritis, ulcerative colitis, psoriasis, and lupus (Fig. 4). It also highlights conditions classically characterized in other disease groups (e.g., cancer, neurological disorders) that are known to be related to the immune system, for example, cancers associated with immune cells, such as B-cell or Hodgkin's lymphoma. Interestingly, COVID-19 is also in this community, connected only to Type I Diabetes by the common variant rs657152 in the ABO gene. Indeed, studies have reported relationship association of the ABO blood groups with type I diabetes^{22,23} and to different levels of susceptibility to SARS-COV-2 infection^{24–26}. It is important to note that our data are limited to GWAS studies added to the GWAS catalog before June 30th, 2020, which were the early stages of the pandemic, and therefore more connections may be discovered with additional research.

Community A, the second community with most disease-related traits nodes ($n = 90$) is characterized by diseases of the vascular system (Fig. 5). The most connected nodes in the community were coronary heart disease, stroke, coronary artery disease, metabolic syndrome, cardiovascular disease, hypertriglyceridemia, gout, chronic kidney disease, diabetes mellitus, and atrial fibrillation. Peripheral arterial disease (PAD), cirrhosis of liver, and non-alcoholic fatty liver disease are also in this community, and previous studies report association among these diseases^{27,28}.

Community B, the third biggest community ($n = 85$) is characterized by several types of cancer, such as breast and ovarian serous carcinoma. This community also contains skin-related traits, such as vitiligo, sunburn, skin and hair pigmentation, and skin cancer. Retrospective studies in Taiwan and Korea have found increased risk of different types of cancer in patients with vitiligo^{29,30}, and vitiligo-related genes have been linked to skin cancer³¹.

Finally, the network shows that Type II Diabetes is in the same community as several neurological disorders, such as Alzheimer's disease and schizophrenia. In fact, previous studies show that Type II Diabetes is linked to Alzheimer's disease and dementia^{32–37}, and several anti-diabetic drugs can promote neuronal survival and lead to clinical improvement of cognition and memory³⁸.

Altogether, these results demonstrate the intricate molecular relationships among diseases and how a network-based approach can help identify groups of diseases with shared underlying mechanisms. These communities might offer insights on specific comorbidity patterns observed in patients, as well as highlight genetic variants for future functional in-depth research.

Novel disease relationships emerging from MVP findings. Large and representative cohorts allow for the discovery of new genetic variants associated with different conditions, especially amongst minority populations with diverse ancestries. In particular, the MVP cohort contains higher percentages of minority groups that are usually underrepresented in genetic studies^{17,39}, which lead to the discovery of variants not observed in more homogeneous populations. For example, PAD had 167 variants reported in the GWAS Catalog from non-MVP sources, but an MVP study⁴⁰ found 18 loci that were novel at the time of the publication. Out of these novel loci, four (rs2107595, rs505922, rs6025, rs7903146) were also observed for duodenal ulcers, glycosuria, large artery stroke, and ischemic stroke, revealing molecular links between diseases that were not observed before. Therefore, we sought to characterize the new relationships among diseases that emerge when genetic data from MVP studies obtained from the GWAS Catalog is integrated in the analysis. We analyzed the subnetwork formed only by edges exclusively created from MVP data, which contains 196 traits and 297 edges (Fig. 6).

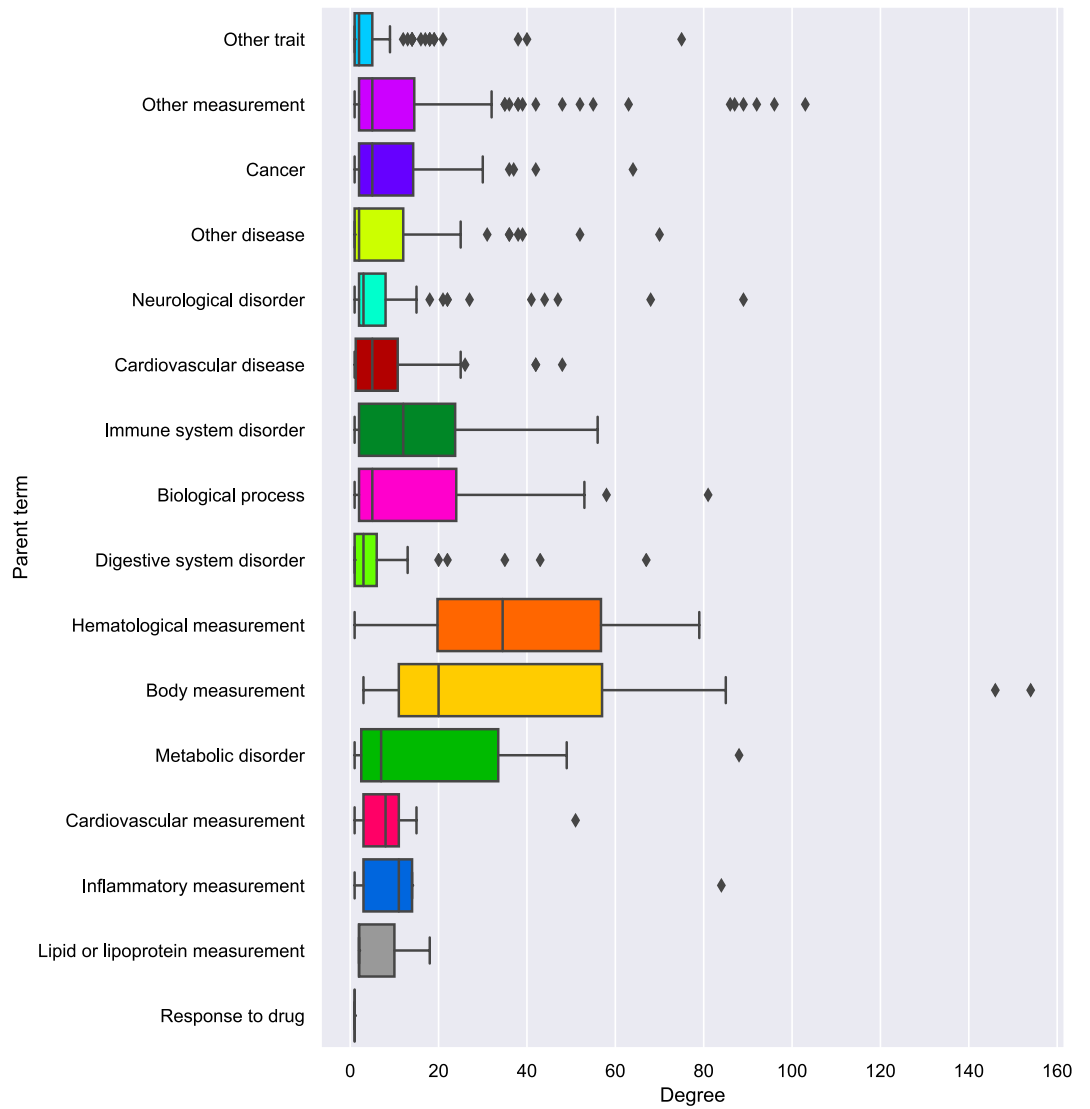


Figure 3. Degree distribution by trait category. Trait categories are defined by the EFO ontology system parent terms.

Variant	Chromosome	Edges	Gene
RS3184504	12	641	ATXN2, SH2B3
RS1260326	2	533	GCKR
RS12075	1	443	ACKR1, CADM3-AS1
RS516246	19	322	FUT2
RS8040868	15	311	CHRNA3
RS10830963	11	307	MTNR1B
RS2476601	1	278	AL137856.1, PTPN22
RS701428	22	276	LINC00896—RTN4R
RS3919627	3	276	AC092042.3, KRBOX1, AC099329.2, CYP8B1, ACKR2
RS700750	7	274	AC011294.1

Table 2. Top variants found in the phenotypic network. Table showing the variants responsible for creating the greatest number of edges in the Phenotypic Network. Information includes the number of edges and the gene associated with that variant. The gene-variant relationships are acquired from the GWAS Catalog. For variants occurring in intergenic regions, both the upstream and downstream genes are shown.

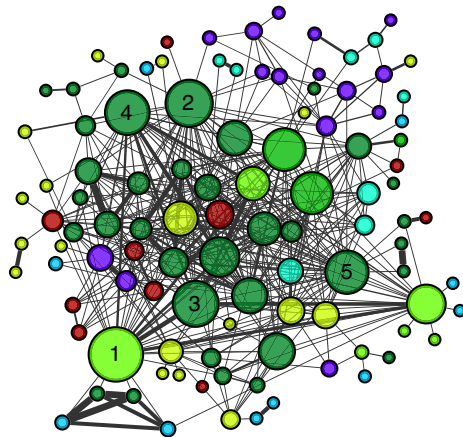


Figure 4. Network community ‘E’ characterized by immune-related disorders. Focused subgraph of community E from the Phenotypic Network. The most connected diseases in the community are Crohn’s disease (1), rheumatoid arthritis (2), ulcerative colitis (3), psoriasis (4), and lupus (5).

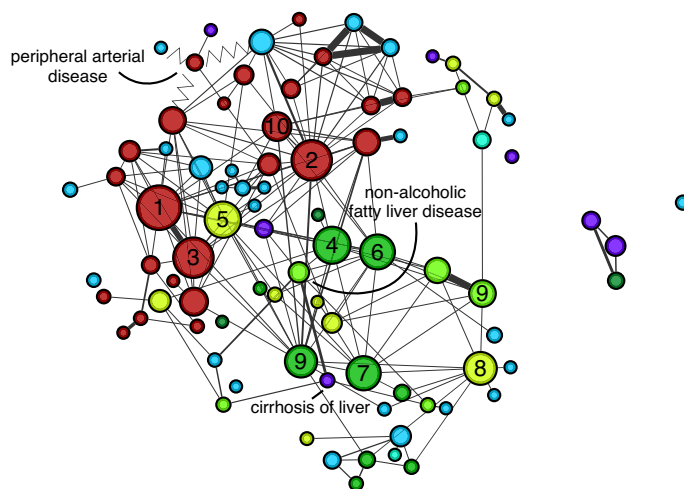


Figure 5. Network community ‘A’ characterized by vascular disorders. After removal of traits unrelated to diseases from the visualization, the most connected nodes in the community were coronary heart disease (1), stroke (2), coronary artery disease (3), metabolic syndrome (4), cardiovascular disease (5), hypertriglyceridemia (6), gout (7), chronic kidney disease (8), diabetes mellitus (9), and atrial fibrillation (10).

The disease traits for which we identified the greatest number of novel disease relationships were, in descending order: glomerular filtration rate^{41,42}, alcohol dependence⁴³, peripheral arterial disease⁴⁰, gout⁴⁴, diabetes mellitus⁴², microalbuminuria⁴⁵, urinary albumin to creatine ratio⁴⁵, systolic blood pressure^{46,47}, venous thromboembolism^{40,48}, diastolic blood pressure^{46,47}, and body height⁴⁹ (Fig. 5). Glomerular filtration rate was the trait with the most novel edges, in which two MVP studies^{41,42} found 664 variants that created 19 new connections in the network. Traits evaluated by MVP studies that did not produce novel connections in the network were anxiety⁵⁰, anxiety disorder⁵⁰, bipolar I disorder⁵¹, schizophrenia⁵¹, ankle brachial index⁴⁰, and panic disorder⁵⁰. MVP publications found in the GWAS Catalog, the Phenotypic Network, and the MVP Novel Network can be found in Supplementary Table 3.

Glomerular filtration rate and gout represented the disease pair with greatest number of shared neighbors ($n = 10$) in the novel disease network (Fig. 6). Five of these traits—lung adenocarcinoma, intelligence, squamous cell carcinoma, lung carcinoma and malaria—were connected not only to glomerular filtration rate and gout, but also to diabetes mellitus.

Our network also showed novel edges connecting rheumatoid arthritis (RA) to PAD and glomerular filtration rate (GFR). Previous studies have highlighted supporting evidence of the association between RA and GFR^{52,53} and RA and PAD^{54–58}. Indeed, RA has pathological processes that also occurs in atherosclerosis, such as endothelial activation, inflammatory cell infiltration, neovascularization, and collagen degradation⁵⁹. However, most studies investigating the association of rheumatoid arthritis with PAD are small and cross-sectional and future research is needed^{54–58}.

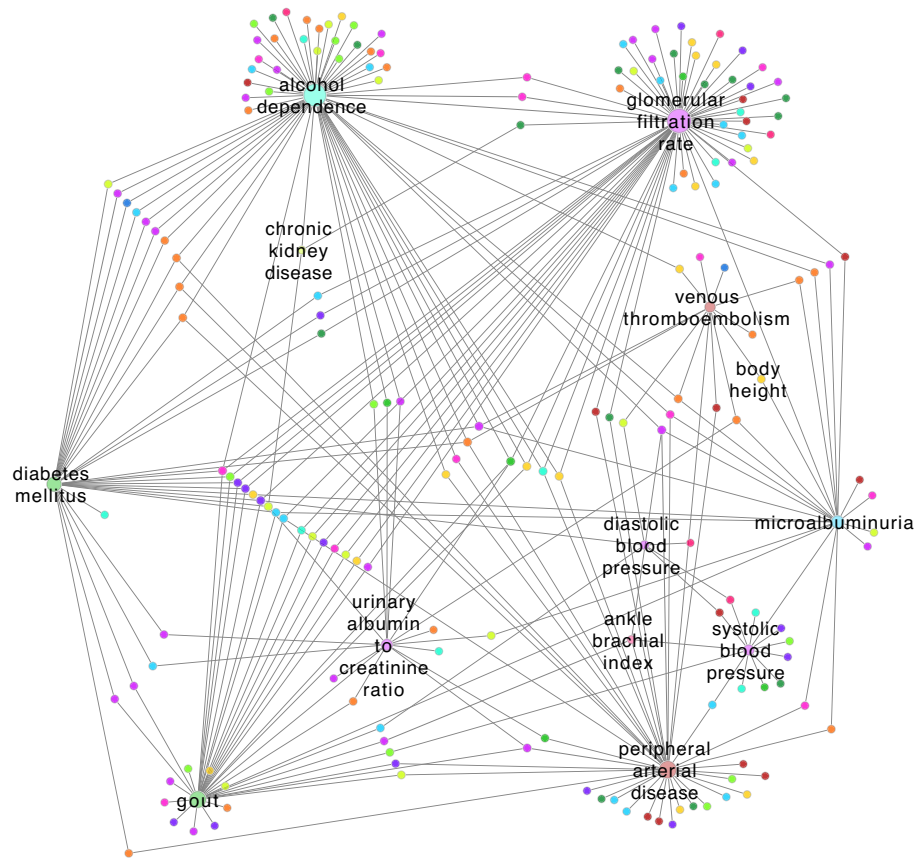


Figure 6. Disease connections that emerge from MVP data. Subgraph containing 196 traits and 297 edges that were formed only by the inclusion of genetic variant associations from the Million Veteran Program.

These results (found in Supplementary Tables 4 and 5) highlight that genetics data revealed by MVP studies can help identify relationships among diseases that were not known before, indicating areas for future research related to disease mechanism, treatment, and prevention.

Disease relationships driven by ancestry. It is well known that there exists some bias in genetic studies research, for which populations with European ancestry are over-represented in relation to other populations, such as Afro-American and Native American³⁹. Therefore, we demonstrate these methods have the ability to characterize the landscape of disease-disease relationships driven by ancestry through distinguishing studies and GWAS results by separating European-only studies from all others.

We found that the community clusters profiled in the separate genetic networks are considerably different, with over 90% of nodes having less than a 0.4 correlation coefficient (Fig. 7). For example, we observed that hypertension, which had large difference in degree between the European and non-European networks (93 and 41, respectively), had an inverse correlation (-0.22), demonstrating that it has a different profile of disease relationships in the two networks. In fact, blood pressure is a trait that has been found to be highly heritable, with substantial differences in blood pressure control rates between non-Hispanic white adults (55.7%) and non-Hispanic Blacks (48.5%)⁶⁰. Therefore, GWAS studies with more diverse populations may allow the discovery of novel anti-hypertensive therapeutics by identifying new gene targets based on loci that have similar effect sizes across race/ethnic groups⁴⁷.

Next, we explored the novel contributions that MVP has made by highlighting which edges in the European and non-European networks only occur in the presence of MVP publications. We found that, despite a large difference in size between the input data for these networks (155,760 and 47,749 SNP-trait associations, respectively), the graphs induced by the edges that only occur in MVP publications were relatively comparable in size (162 European edges vs 116 non-European edges). These results suggest that MVP has more heterogeneous population enabling investigation of both European and non-European based genetic relationships of diseases and their comorbidities.

Discussion

In this study, we provide an overview of the relationship among phenotypes that share strong SNP-trait associations. We assembled a network of published genetic variants available through the GWAS Catalog repository to visualize novel connections and to investigate new insights gained through findings from numerous studies to-date. While recent studies^{7,61–63} have constructed disease networks through the use of known disease genes

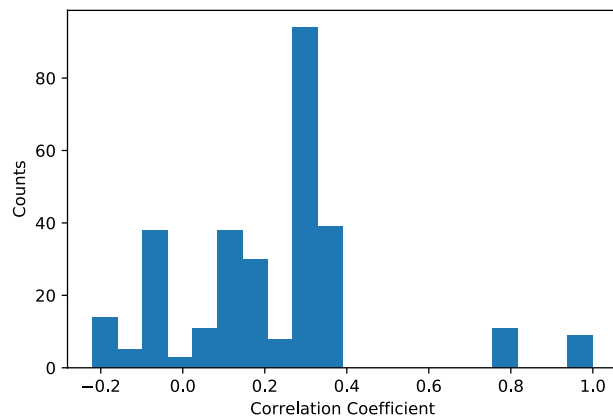


Figure 7. Community correlation between notes of different ancestry networks. Histogram of Pearson product-moment correlation coefficients for shared community members between the same trait in different ancestry networks. The count represents number of traits that have the given correlation coefficient.

from sources such as the Online Mendelian Inheritance in Man (OMIM) and various GWAS databases, these networks are typically smaller in size and utilized as part of a further analysis such as exploring drug efficacy⁶¹, drug repurposing⁶³ or revealing disease relationships based on expression levels⁶² or the interactome⁷. Our network reveals novel associations between diseases and provides a mechanistic approach to categorize diseases in different groups. Finally, we mapped the new disease associations that emerge only when we included variants from MVP studies contained within the GWAS Catalog. We believe that our results offer insights to better understand comorbidity patterns observed in patients and have the potential to reveal mechanistic links between diseases with further investigation. Additionally, the identification of diseases that share genetic similarities offers the opportunity to investigate possible drug-repurposing strategies for identification of new indications for existing drugs^{61,63,64}.

We highlight that our approach relies only on genetic information, but diseases often manifest through multifaceted mechanisms including other clinical factors and shared environmental exposure^{12,65}. Other approaches to evaluate disease relationships rely on connecting diseases that tend to co-occur in patients⁵ or for which patients usually show similar gene expression profiles^{66–68}. Indeed, following the strategy from Klimek et al.¹², a multi-layer network approach—where in each layer diseases are connected based on a different set of features (e.g., genetic variant or disease co-occurrence)—might distinguish driving forces in disease relationships that go beyond genetics information only¹². We bring to attention that GWAS data may include non-causal variants that arise due to technical artifacts or other biological factors, such as a linkage disequilibrium. However, data availability on causal variants is very limited and specific to diseases of high clinical and research interest, resulting in studies highly affected by literature bias. We believe that big data analysis has the power to identify true biological signal even amidst high levels of noise. For example, previous network-medicine studies^{4,7,61,63} used GWAS-derived variants and were able to recover true disease-disease and disease drug relationships with high levels of predictive power. Machine learning-based models are also able to leverage on (non-causal or not) genetic variants to help reveal missing heritability and epistatic interactions on GWAS-based datasets⁶⁹. Indeed, we also demonstrate that the proposed methodology identifies true biological signal by being able to recover clinically relevant disease relationships such as cancer and vitiligo^{29,30}, Type II Diabetes and Alzheimer's disease^{32–37}, and Rheumatoid arthritis and PAD^{54–58}. Furthermore, previous studies^{63,70} identified predictions that leverage GWAS-based variants and further validated observations with experimental and clinical data.

The results presented here aggregate the top hits from 3,985 studies found in the GWAS Catalog. Therefore, heterogeneity might exist in the definition of phenotypes across different studies. For example, the network contains 15 traits related to diabetes (Supplementary Table 6), containing broad definitions, such as diabetes mellitus, and more specific ones, such as type 2 diabetes nephropathy and diabetes mellitus type 2 associated cataract. However, we believe that, even in the presence of these variations, the general patterns observed here provide important insights for clinical practice. We also highlight that our study lays the foundations for future studies that could avoid these limitations by using GWAS data from well-phenotyped cohorts such as the MVP and UK Biobank. More specifically in the VA, there is a nation-wide effort to harmonize and catalog phenotypic mapping and algorithms where MVP is a major contributor. In addition, MVP has applied several advanced high-throughput phenotypic engines to develop complex phenotypes using large clinical database^{71,72}. While MVP is a diverse cohort, it's comprised of predominantly older men by design. However, due to the large size of the cohort, there are a substantial number in sub populations covering the rest of the general demographics. For instance, in a prior version of the MVP cohort (19.2), while women represented only 9.8% of the total cohort, there were still 64,658 individuals. Also, past MVP GWAS have found their results are able to be replicated^{40,42,47,49,50,73,74}. Finally, our current study included only a part of the genetics data available in MVP and the GWAS Catalog by including studies added to the GWAS catalog before June 30th, 2020. Our results merit further investigation of more integrated network as the MVP and other major biobanks and cohorts continue to grow and produce next generation genetic discoveries.

Methods

Data. GWAS Catalog (version 1.0.2¹⁹) data was obtained and downloaded in July 2020 with a freeze on studies added on or before June 30, 2020, ensuring that the dataset used for analyses remained consistent and static. The GWAS catalog database included study information (i.e. lead author, study name, PubMedID, ancestry, study type), traits (mapped to ontology terms), and genetic variants that met the p-value threshold of 1×10^{-5} . Additional criteria for inclusion in the catalog can be found elsewhere¹⁹.

The ontological system Experimental Factor Ontology (EFO)⁷⁵ is used in the GWAS catalog to provide a level of consistency in the description of the traits. We used the EFO to map traits to their corresponding EFO categories (e.g. digestive system disorder, hematological measurements) and when multiple EFO terms could be mapped to the same trait, we assigned the trait to each possible term.

As our primary aim was to observe relatedness among diseases, we performed filtering steps to reduce the number of traits not directly related to diseases. We performed a regular expression search and removed all nodes with the keywords: "measurement" or "response to (medication/treatment)". This step removed 1,686 EFO terms or potential network nodes from consideration. It was important for us to retain as many disease nodes as possible and for this reason, we limited the number of keywords that would trigger trait elimination. We also removed from the network data 21 EFO terms that independently provided no meaning outside the context of their respective phenotype, such as "age at onset" and "age at diagnosis".

Traits related to the following EFO terms are determined not to be disease-related and therefore are not labeled in figures: other measurement, biological process, body measurement, lipid or lipoprotein measurement, response to drug, and hematological measurement.

Finally, for each study we obtained the trait and corresponding EFO term, the PubMedID, and the genetic variants. We used the PubMedIDs to differentiate studies belonging to research contributions of MVP.

Network analysis. The network was created by using traits as nodes and by edges (or links) connected pairs of traits with shared variants. For each edge we calculate the normalized overlap (Jaccard Index) of variants between the pair of traits and applied the Fisher's exact test to assess the statistical significance of the overlap followed by Benjamini–Hochberg multiple testing correction. We performed community detection in the resulting network using the Louvain algorithm and the statistical significance of each community was evaluated following the strategy based on modularity and size, as proposed by Kojaku et al.⁷⁶. The network analyses were performed with the Python packages 'networkx'⁷⁷ and 'community'²¹, statistical tests were performed with 'Scipy'⁷⁸ and 'statsmodels'⁷⁹ packages, and network visualization was performed with Cytoscape⁸⁰.

Once the full disease related network was created from the GWAS catalog, we differentiated the networks for which there was no contribution from MVP studies from the network for which there was. We use the former to highlight the novel disease-disease relationships that emerge when MVP data is included.

To investigate the contribution of ancestry to our network we annotated the association data using a framework created by the GWAS Catalog team which contains ancestral categories for a given study³⁹. Using this separate file provided by the GWAS Catalog to roll up more granular classifications into broader categories. For instance, ancestries labeled as "Sub-Saharan African" or "African unspecified" were collapsed into the category "African". We then created indicator flags for each row in the catalog that highlights whether a study contained either European or non-European populations based on its study accession. These flags were not mutually exclusive. We then used the flags to replicate our network assembling pipeline and created two separate networks, European and non-European.

We then ask whether the diseases tend to have the same or different pattern of disease-disease connections in the European and non-European networks. We achieve this by representing each disease present in both networks ($n = 300$) with a vector of 0's and 1's, with 1's indicating other conditions to which a disease is connected to in the same network and 0's otherwise. By comparing the vectors of each disease in both networks, we were able to assess the extent to which their community profiles are similar or different.

Data used in this study are all publicly available from the GWAS Catalog which follows the General Data Protection Regulation (GDPR) as described on their website. The GWAS Catalog, a repository of summary statistics curated by the European Molecular Biology Laboratory, follows a time and release protocol where data is reviewed by a Data Access Committee before being released to the public. These research activities were approved by VA Central IRB #18-38.

Data availability

All data used was publicly available and downloaded from the GWAS catalog. More information can be found in the contents section of the Supplementary file.

Received: 21 April 2022; Accepted: 26 August 2022

Published online: 01 September 2022

References

- Dugoff, E. H., Canudas-Romo, V., Buttorff, C., Leff, B. & Anderson, G. F. Multiple chronic conditions and life expectancy: A life table analysis. *Med. Care* **52**, 688–694. <https://doi.org/10.1097/MLR.000000000000166> (2014).
- Cortaredona, S. & Ventelou, B. The extra cost of comorbidity: Multiple illnesses and the economic burden of non-communicable diseases. *BMC Med.* **15**, 1–11. <https://doi.org/10.1186/s12916-017-0978-2> (2017).
- Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev. Genet.* **10**, 134–140. <https://doi.org/10.1038/nrg2502> (2009).
- Goh, K. I. et al. The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* <https://doi.org/10.1073/pnas.0701361104> (2007).
- Hidalgo, C. A., Blumm, N., Barabasi, A. L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353. <https://doi.org/10.1371/journal.pcbi.1000353> (2009).

6. Lee, D. S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci.* <https://doi.org/10.1073/pnas.0802208105> (2008).
7. Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (N. Y.)* **347**, 1257601. <https://doi.org/10.1126/science.1257601> (2015).
8. Park, J., Lee, D.-S., Christakis, N. A. & Barabási, A.-L. The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* <https://doi.org/10.1038/msb.2009.16> (2009).
9. Chmiel, A., Klimek, P. & Thurner, S. Spreading of diseases through comorbidity networks across life and gender. *New J. Phys.* **16**, 115013. <https://doi.org/10.1088/1367-2630/16/11/115013> (2014).
10. Jeong, E., Ko, K., Oh, S. & Han, H. W. Network-based analysis of diagnosis progression patterns using claims data. *Sci. Rep.* <https://doi.org/10.1038/s41598-017-15647-4> (2017).
11. Westergaard, D., Moseley, P., Sorup, F. K. H., Baldi, P. & Brunak, S. Population-wide analysis of differences in disease progression patterns in men and women. *Nat. Commun.* **10**, 1–14. <https://doi.org/10.1038/s41467-019-08475-9> (2019).
12. Klimek, P., Aichberger, S. & Thurner, S. Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks. *Sci Rep* **6**, 39658. <https://doi.org/10.1038/srep39658> (2016).
13. Allen, N. *et al.* UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126. <https://doi.org/10.1016/j.hlpt.2012.07.003> (2012).
14. Sudlow, C. *et al.* 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. <https://doi.org/10.1371/journal.pmed.1001779> (2015).
15. Hedderson, M. M. *et al.* The Kaiser Permanente Northern California research program on genes, environment, and health (RPGEH) pregnancy cohort: study design, methodology and baseline characteristics. *BMC Pregn. Childbirth* **16**, 381. <https://doi.org/10.1186/s12884-016-1150-2> (2016).
16. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666. <https://doi.org/10.1093/ije/dyr120> (2011).
17. Gaziano, J. M. *et al.* Million veteran program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223. <https://doi.org/10.1016/j.jclinepi.2015.09.016> (2016).
18. MVP. *Current MVP Publications*, <https://www.mvp.va.gov/pwa/sites/default/files/2021-06/MVP%20Publications_2021-04-REEF.pdf> (2022).
19. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012. <https://doi.org/10.1093/nar/gky1120> (2019).
20. *SH2B3 SH2B adaptor protein 3 [Homo sapiens (human)]*, <<https://www.ncbi.nlm.nih.gov/gene/10019>> (2022).
21. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* <https://doi.org/10.1088/1742-5468/2008/10/p10008> (2008).
22. Li-Gao, R. *et al.* Genome-wide association study on the early-phase insulin response to a liquid mixed meal: Results from the NEO study. *Diabetes* **68**, 2327–2336. <https://doi.org/10.2337/db19-0378> (2019).
23. Parente, E. B. *et al.* Relationship between ABO blood groups and cardiovascular disease in type 1 diabetes according to diabetic nephropathy status. *Cardiovasc. Diabetol.* **19**, 68. <https://doi.org/10.1186/s12933-020-01038-z> (2020).
24. Goel, R. *et al.* ABO blood group and COVID-19: A review on behalf of the ISBT COVID-19 Working Group. *Vox Sang* **116**, 849–861. <https://doi.org/10.1111/vox.13076> (2021).
25. Wu, B. B., Gu, D. Z., Yu, J. N., Yang, J. & Shen, W. Q. Association between ABO blood groups and COVID-19 infection, severity and demise: A systematic review and meta-analysis. *Infect. Genet. Evol.* **84**, 104485. <https://doi.org/10.1016/j.meegid.2020.104485> (2020).
26. Severe Covid, G. G. Genomewide association study of severe Covid-19 with respiratory failure. *N. Engl. J. Med.* (2020).
27. Lin, S. Y. *et al.* Risk of acute coronary syndrome and peripheral arterial disease in chronic liver disease and cirrhosis: A nationwide population-based study. *Atherosclerosis* **270**, 154–159. <https://doi.org/10.1016/j.atherosclerosis.2018.01.047> (2018).
28. Zhu, W. *et al.* Peripheral artery disease and risk of fibrosis deterioration in nonalcoholic fatty liver disease: A prospective investigation. *Biomed. Environ. Sci.* **33**, 217–226. <https://doi.org/10.3967/bes2020.031> (2020).
29. Kim, H. S. *et al.* The incidence and survival of melanoma and nonmelanoma skin cancer in patients with vitiligo: a nationwide population-based matched cohort study in Korea. *Br. J. Dermatol.* **182**, 907–915. <https://doi.org/10.1111/bjd.18247> (2020).
30. Li, C. Y. *et al.* Cancer risks in vitiligo patients: A nationwide population-based study in Taiwan. *Int. J. Environ. Res. Public Health* <https://doi.org/10.3390/ijerph15091847> (2018).
31. Wu, W. *et al.* Inverse relationship between vitiligo-related genes and skin cancer risk. *J. Invest. Dermatol.* **138**, 2072–2075. <https://doi.org/10.1016/j.jid.2018.03.1511> (2018).
32. Leibson, C. L. *et al.* Risk of dementia among persons with diabetes mellitus: a population-based cohort study. *Am. J. Epidemiol.* **145**, 301–308. <https://doi.org/10.1093/oxfordjournals.aje.a009106> (1997).
33. Luchsinger, J. A., Tang, M. X., Shea, S. & Mayeux, R. Hyperinsulinemia and risk of Alzheimer disease. *Neurology* **63**, 1187–1192. <https://doi.org/10.1212/01.wnl.0000140292.04932.87> (2004).
34. Luchsinger, J. A., Tang, M. X., Stern, Y., Shea, S. & Mayeux, R. Diabetes mellitus and risk of Alzheimer's disease and dementia with stroke in a multiethnic cohort. *Am. J. Epidemiol.* **154**, 635–641. <https://doi.org/10.1093/aje/154.7.635> (2001).
35. Ott, A. *et al.* Association of diabetes mellitus and dementia: the Rotterdam Study. *Diabetologia* **39**, 1392–1397. <https://doi.org/10.1007/s001250050588> (1996).
36. Ott, A. *et al.* Diabetes mellitus and the risk of dementia: The Rotterdam Study. *Neurology* **53**, 1937–1942. <https://doi.org/10.1212/wnl.53.9.1937> (1999).
37. Peila, R., Rodriguez, B. L. & Launer, L. J. Type 2 diabetes, APOE gene, and the risk for dementia and related pathologies: The Honolulu-Asia Aging Study. *Diabetes* **51**, 1256–1262. <https://doi.org/10.2337/diabetes.51.4.1256> (2002).
38. Patrone, C., Eriksson, O. & Lindholm, D. Diabetes drugs and neurological disorders: new views and therapeutic possibilities. *Lancet Diabetes Endocrinol* **2**, 256–262. [https://doi.org/10.1016/S2213-8587\(13\)70125-6](https://doi.org/10.1016/S2213-8587(13)70125-6) (2014).
39. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21. <https://doi.org/10.1186/s13059-018-1396-2> (2018).
40. Klarin, D. *et al.* Genome-wide association study of peripheral artery disease in the Million Veteran Program. *Nat. Med.* **25**, 1274–1279. <https://doi.org/10.1038/s41591-019-0492-5> (2019).
41. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972. <https://doi.org/10.1038/s41588-019-0407-x> (2019).
42. Hellwege, J. N. *et al.* Mapping eGFR loci to the renal transcriptome and phenotype in the VA Million Veteran Program. *Nat. Commun.* **10**, 3842. <https://doi.org/10.1038/s41467-019-11704-w> (2019).
43. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* **10**, 1499. <https://doi.org/10.1038/s41467-019-09480-8> (2019).
44. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet.* **51**, 1459–1474. <https://doi.org/10.1038/s41588-019-0504-x> (2019).
45. Teumer, A. *et al.* Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat. Commun.* **10**, 4130. <https://doi.org/10.1038/s41467-019-11576-0> (2019).

46. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425. <https://doi.org/10.1038/s41588-018-0205-x> (2018).
47. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62. <https://doi.org/10.1038/s41588-018-0303-9> (2019).
48. Lindstrom, S. *et al.* Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* **134**, 1645–1657. <https://doi.org/10.1182/blood.2019000435> (2019).
49. Fang, H. *et al.* Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.* **105**, 763–772. <https://doi.org/10.1016/j.ajhg.2019.08.012> (2019).
50. Levey, D. F. *et al.* Reproducible genetic risk loci for anxiety: Results from approximately 200,000 participants in the million veteran program. *Am. J. Psychiatry* **177**, 223–232. <https://doi.org/10.1176/appi.ajp.2019.19030256> (2020).
51. Harvey, P. D. *et al.* Genome-wide association study of cognitive performance in US veterans with schizophrenia or bipolar disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **183**, 181–194. <https://doi.org/10.1002/ajmg.b.32775> (2020).
52. Hill, A. J., Thomson, R. J., Hunter, J. A. & Traynor, J. P. The prevalence of chronic kidney disease in rheumatology outpatients. *Scott. Med. J.* **54**, 9–12. <https://doi.org/10.1258/rsmmj.54.2.9> (2009).
53. Madrakhimov, S. B., Shukurov, U. M. & Ubaydullaev, S. A. Traditional cardiovascular risk factors and glomerular filtration rate in patients with rheumatoid arthritis. *Eur. Heart J.* <https://doi.org/10.1093/ehjci/ehaa946.3153> (2020).
54. Alkaabi, J. K., Ho, M., Levison, R., Pullar, T. & Belch, J. J. Rheumatoid arthritis and macrovascular disease. *Rheumatol. (Oxford)* **42**, 292–297. <https://doi.org/10.1093/rheumatology/keg083> (2003).
55. Henke, P. K., Sukheepod, P., Proctor, M. C., Upchurch, G. R. Jr. & Stanley, J. C. Clinical relevance of peripheral vascular occlusive disease in patients with rheumatoid arthritis and systemic lupus erythematosus. *J. Vasc. Surg.* **38**, 111–115. [https://doi.org/10.1016/s0741-5214\(03\)00074-0](https://doi.org/10.1016/s0741-5214(03)00074-0) (2003).
56. Liang, K. P. *et al.* Incidence of noncardiac vascular disease in rheumatoid arthritis and relationship to extraarticular disease manifestations. *Arthritis Rheum.* **54**, 642–648. <https://doi.org/10.1002/art.21628> (2006).
57. Sedrakyan, S. *et al.* Evaluation of the risk of getting peripheral artery disease in rheumatoid arthritis and the selection of appropriate diagnostic methods. *Cureus* **12**, e9782. <https://doi.org/10.7759/cureus.9782> (2020).
58. Stamatelopoulos, K. S. *et al.* Subclinical peripheral arterial disease in rheumatoid arthritis. *Atherosclerosis* **212**, 305–309. <https://doi.org/10.1016/j.atherosclerosis.2010.05.007> (2010).
59. Keech, S. & Bruce, I. N. Atherosclerosis in rheumatoid arthritis: is it all about inflammation?. *Nat. Rev. Rheumatol.* **11**, 390–400. <https://doi.org/10.1038/nrrheum.2015.40> (2015).
60. Franceschini, N., Chasman, D. I., Cooper-DeHoff, R. M. & Arnett, D. K. Genetics, ancestry, and hypertension: Implications for targeted antihypertensive therapies. *Curr. Hypertens Rep.* **16**, 461. <https://doi.org/10.1007/s11906-014-0461-9> (2014).
61. Guney, E., Menche, J., Vidal, M. & Barabasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331. <https://doi.org/10.1038/ncomms10331> (2016).
62. Yang, J., Wu, S.-J., Dai, W.-T., Li, Y.-X. & Li, Y.-Y. The human disease network in terms of dysfunctional regulatory mechanisms. *Biol. Direct* **10**, 60. <https://doi.org/10.1186/s13062-015-0088-z> (2015).
63. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691. <https://doi.org/10.1038/s41467-018-05116-5> (2018).
64. Aguirre-Plans, J. *et al.* Proximal pathway enrichment analysis for targeting comorbid diseases via network endopharmacology. *Pharmaceuticals* **11**, 61 (2018).
65. Bagley, S. C., Sirota, M., Chen, R., Butte, A. J. & Altman, R. B. Constraints on biological mechanism from disease comorbidity using electronic medical records and database of genetic variants. *PLoS Comput. Biol.* **12**, e1004885. <https://doi.org/10.1371/journal.pcbi.1004885> (2016).
66. Anand, R. & Chatterjee, S. Tracking disease progression by searching paths in a temporal network of biological processes. *PLoS ONE* **12**, e0176172. <https://doi.org/10.1371/journal.pone.0176172> (2017).
67. Sánchez-Valle, J. *et al.* Unveiling the molecular basis of disease co-occurrence: towards personalized comorbidity profiles. *bioRxiv* <https://doi.org/10.1101/431312> (2018).
68. Yang, J. *et al.* DNetDB: The human disease network database based on dysfunctional regulation mechanism. *BMC Syst. Biol.* **10**, 36. <https://doi.org/10.1186/s12918-016-0280-5> (2016).
69. Okser, S., Pahikkala, T. & Aittokallio, T. Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives. *BioData Mining* **6**, 5. <https://doi.org/10.1186/1756-0381-6-5> (2013).
70. do Valle, Í. F. *et al.* Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols. *Nat. Food* **2**, 143–155. <https://doi.org/10.1038/s43016-021-00243-7> (2021).
71. Liao, K. P. *et al.* High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J. Am. Med. Inf. Assoc.* **26**, 1255–1262. <https://doi.org/10.1093/jamia/ocz066> (2019).
72. Zhang, Y. *et al.* High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat. Protoc.* **14**, 3426–3444. <https://doi.org/10.1038/s41596-019-0227-6> (2019).
73. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579. <https://doi.org/10.1038/s41588-019-0519-3> (2019).
74. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523. <https://doi.org/10.1038/s41588-018-0222-9> (2018).
75. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118. <https://doi.org/10.1093/bioinformatics/btq099> (2010).
76. Kojaku, S. & Masuda, N. A generalised significance test for individual communities in networks. *Sci. Rep.* **8**, 7351. <https://doi.org/10.1038/s41598-018-25560-z> (2018).
77. Hagberg, A. A., Schult, D. A. & Swart, P. J. in *7th Python in Science Conference (SciPy 2008)* (2008).
78. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272. <https://doi.org/10.1038/s41592-019-0686-2> (2020).
79. Seabold, S. & Perktold, J. in *9th Python in Science Conference* (2010).
80. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. <https://doi.org/10.1101/gr.1239303> (2003).

Acknowledgements

The project was supported by Department of Veterans Affairs, Office of Research and Development, Million Veteran Program Core (#MVP000; <https://www.research.va.gov/>). This manuscript has been in part co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy, and under a joint program (MVP CHAMPION), between the U.S. Department of Energy (DOE), and the U.S. Department of Veterans Affairs (VA).

Author contributions

B.R.F., I.F.V., K.C., and H.G. designed the study. B.R.F., H.G., D.G., and L.C. gathered, cleaned, and assembled the data. B.R.F., I.F.V., H.G., and K.C. wrote the manuscript. All authors read and approved the manuscript.

Competing interests

A.L.B. is co-scientific founder of Scipher Medicine, Inc., which applies network medicine strategies to biomarker development and personalized drug selection and Foodome, Inc. that apply data science to health, and DataPolis, that explores the implications of human mobility. All other authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-19244-y>.

Correspondence and requests for materials should be addressed to B.F.

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022