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

Brian Ferolito1,6,*, Italo Faria do Valle1,2,6, Hanna Gerlovin1, Lauren Costa1, Juan P. Casas1,5, J. Michael Gaziano1,5, David R. Gagnon1,3, Edmon Begoli4, Albert-László Barabási2 & Kelly Cho1,5,6

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 expectancy1,2. Comorbidities suggest common mechanisms that underlie different diseases, which can be either genetic or environmental1. 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 mechanisms4,6–8, comorbidities driven by age9, gender9–11, demographic factors9 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 Biobank13,14; 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 projects\(^5\) 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 MVPs, 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 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 studies\(^4\). 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 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 studies\(^4\). 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 \(k\) (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 \((Pearson \rho = 0.672, p = 1.45 \times 10^{-107}\) \). 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 rs1384504 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 pathway\(^4\). The diseases in our network that contain the most edges with this variant are type 1 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 Louvain\(^22\), 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.

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.
findings 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 and to different levels of susceptibility to SARS-COV-2 infection. 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 the liver, and non-alcoholic fatty liver disease are also in this community, and previous studies report association among these diseases.

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, and vitiligo-related genes have been linked to skin cancer.

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, and several anti-diabetic drugs can promote neuronal survival and lead to clinical improvement of cognition and memory.

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, 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 study 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.

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.
The disease traits for which we identified the greatest number of novel disease relationships were, in descending order: glomerular filtration rate\(^41,42\), alcohol dependence\(^43\), peripheral arterial disease\(^40\), gout\(^44\), diabetes mellitus\(^42\), microalbuminuria\(^45\), urinary albumin to creatine ratio\(^45\), systolic blood pressure\(^46,47\), venous thromboembolism\(^40,48\), diastolic blood pressure\(^46,47\), and body height\(^49\) (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 anxiety\(^50\), anxiety disorder\(^50\), bipolar I disorder\(^51\), schizophrenia\(^51\), ankle brachial index\(^40\), and panic disorder\(^50\).

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 occur in atherosclerosis, such as endothelial activation, inflammatory cell infiltration, neovascularization, and collagen degradation\(^59\). However, most studies investigating the association of rheumatoid arthritis with PAD are small and cross-sectional and future research is needed\(^54-58\).

---

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).
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 American. 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%). 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 groups.

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 have constructed disease networks through the use of known disease genes
biological signal even amidst high levels of noise. For example, previous network-medicine studies\textsuperscript{4,7,61,63} used
ing in studies highly affected by literature bias. We believe that big data analysis has the power to identify true
availability on causal variants is very limited and specific to diseases of high clinical and research interest, result-
that arise due to technical artifacts or other biological factors, such as a linkage disequilibrium. However, data
throughput phenotypic engines to develop complex phenotypes using large clinical database\textsuperscript{71,72}. While MVP is a
and UK Biobank. More specifically in the VA, there is a nation-wide effort to harmonize and catalog phenotypic
studies that could avoid these limitations by using GWAS data from well-phenotyped cohorts such as the MVP
provide important insights for clinical practice. We also highlight that our study lays the foundations for future
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
go beyond genetics information only\textsuperscript{12}. 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\textsuperscript{47,61,63} 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 datasets\textsuperscript{69}. 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\textsuperscript{70,79}, Type II Diabetes and Alzheimer’s disease\textsuperscript{32–37}, and
Rheumatoid arthritis and PAD\textsuperscript{26–35}. Furthermore, previous studies\textsuperscript{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\textsuperscript{27,28}. 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\textsuperscript{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 30\textsuperscript{9}, 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 $1 \times 10^{-5}$. Additional criteria for inclusion in the catalog can be found elsewhere.

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. We then used the flags to replicate our network assembling pipeline and created two separate networks, “European” and “non-European”.

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 study29. 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

1. Dugoff, E. H., Canudas-Romo, V., Butteroff, 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.0b013e3181c05f34 (2014).
2. 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).
3. Gibson, G. Decanalization and the origin of complex disease. Nat. Rev. Genet 10, 134–140. https://doi.org/10.1038/nrg2502 (2009).
4. Goh, K. I. et al. The human disease network. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.0701361104 (2007).
5. 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).
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