Published in final edited form as:

Genet Epidemiol. 2017 December ; 41(8): 866–875. doi:10.1002/gepi.22078.

EVOLUTIONARILY DERIVED NETWORKS TO INFORM DISEASE PATHWAYS

BRITNEY E. GRAHAM^{1,2}, CHRISTIAN DARABOS^{1,3,4}, MINJUN HUANG¹, LOUIS J. MUGLIA⁵, JASON H. MOORE³, and SCOTT M. WILLIAMS^{1,6,*}

¹Department of Genetics, The Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, U.S.A

²Case Western Reserve University, Systems Biology and Bioinformatics, Cleveland, OH 44106, U.S.A

³Institute for Biomedical Informatics, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A

⁴Dartmouth College, Research Computing Services, Hanover, NH 03755, U.S.A

⁵Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center and March of Dimes Prematurity Research Center Ohio Collaborative, Cincinnati, OH 45229, U.S.A

⁶Case Western Reserve University, Department of Population and Quantitative Health Science, Cleveland, OH 44106, U.S.A

Abstract

Methods to identify genes or pathways associated with complex diseases are often inadequate to elucidate most risk because they make implicit and oversimplified assumptions about underlying models of disease etiology. These can lead to incomplete or inadequate conclusions. To address this, we previously developed human phenotype networks (HPN), linking phenotypes based on shared biology. However, such visualization alone is often uninterpretable, and requires additional filtering. Here, we expand the HPN to include another method, Evolutionary Triangulation (ET). ET utilizes the hypothesis that alleles affecting disease risk in multiple populations are distributed consistently with differences in disease prevalence and compares allele frequencies among populations and their relationship to phenotype prevalence. We hypothesized that combining these methods will increase our ability to detect genetic patterns of association in complex diseases. We combined HPN and ET to identify network patterns associated with Type 2 Diabetes Mellitus (T2DM), a leading cause of death worldwide. Fasting glucose, a continuous trait, was used as a proxy for T2DM and differs significantly among continental populations. The combined method identified several diabetes related traits and several phenotypes related to cardiovascular diseases, for which diabetes is a major risk factor. ET-HPN found more phenotypes related to our target and related phenotypes than the application of either method alone. Not only could we detect phenotype connections related to T2DM, but we also identified phenotypes that are distributed in

parallel to it, e.g. Amyotrophic Lateral Sclerosis. Our analyses showed that ET filtered HPN provides information that neither technique can individually.

Keywords

Evolution and Disease; Network Analyses; Complex Diseases; Type 2 Diabetes Mellitus; Knowledge-driven Network Filtering

1. Introduction

Identifying genes or pathways related to human phenotypes is a useful means to understand disease pathophysiology. For many simple genetic diseases (i.e., Mendelian), the underlying genetic causes have already been discovered or are relatively easy to identify. However, for more complex diseases, traditional analytical methods of establishing genetic links are often inadequate to explain a majority of genetic risk, as they are inherently unable to deal with the complexity of the likely pathophysiological models. One method of accounting for complexity in genetic modeling is to apply a network theory approach. This can provide powerful insights that are used to analyze large scale systems, such as the relationships between human diseases. For example, identification of common genetic risk factors among several diseases may provide novel insight into each disease. Networks can also be used to reduce complex systems to visually tractable formats, revealing connections that would not be observed when studying individual traits [Goh, et al. 2007]. Building on this concept, we previously developed human phenotype networks (HPN) as a means to quantitatively and qualitatively identify connections between traits or diseases based on shared genetic effects [Darabos, et al. 2013]. This has proven useful in demonstrating common processes among diverse phenotypes, illustrating the effect of pleiotropy, where one gene is implicated in several phenotypes [Darabos, et al. 2014a]. As our knowledge of genetics and population differences in disease expands, using this approach can further our understanding of disease pathophysiology and relationships among diseases, both those obviously related and *a priori* unrelated.

Despite the power of network analyses, resulting networks can be exceedingly complex and require filtering to uncover interpretable biological interactions. In the present study, we propose one means to trim the HPN, relying on another novel method we have developed, Evolutionary Triangulation (ET). ET utilizes population differences in disease and allelic frequencies to identify genes that are more likely to associate with diseases than randomly selected genes from the genome [Huang, et al. 2016]. ET compares allele frequencies among populations and their relationship to parallel variation in phenotype prevalences and is agnostic in the sense that it makes no *a priori* assumptions about gene function or relevance. ET draws on the hypothesis that evolutionary changes in allele frequencies have shaped disease risk through human migration into new environments. Because of this, risk allele distributions in populations often reflect the distributions of their associated diseases. ET uses evolutionary differentiation through comparisons of population structure among three populations with variable disease prevalence. Population selection based on patterns where two have a similar rate of disease that differs markedly from a third, gives us the ability to

filter out noise in genetic variants to find genetic risk factors most relevant to specific traits of interest, but does not necessarily allow a broader view of the interconnectivity of disease topology. A key feature of ET is that it requires data only from publicly available databases and exploits the ever-growing data on population distribution of phenotypes and allele frequencies.

Combining ET and HPN should define an intermediate space between the broadness of the HPN method and the specificity of the ET method that permits us to identify genetic relationships between multiple phenotypes. As a test of this hypothesis, we propose that combining these methods can provide a novel insight into the causes of Type 2 Diabetes Mellitus (T2DM) and related phenotypes. This compound method should also capture other diseases distributed similarly in the chosen populations. We applied this combination (ET-HPN) in an attempt to identify network patterns that associate with disease.

As a proof of principle we chose to use diabetes and its related phenotype, fasting glucose, as these traits have proven among the most intractable to genetic analyses [Franks 2011; Temelkova-Kurktschiev and Stefanov 2012]. Diabetes is not only complex with respect to genetic risks, but it has a high prevalence, approximately 9% of the world's adult population, and it varies significantly among continental populations. It is the seventh leading cause of death worldwide with serious complications, including chronic kidney disease [Stevens, et al. 2010], nonalcoholic fatty liver disease [Almobarak, et al. 2015], cardiovascular disease, eye disease and nerve damage [Aguirre, et al. 2013]. Most cases (90%) are T2DM. One of the main risk factors, obesity, is on the rise in many countries and has a worldwide prevalence of approximately 11% in men and 15% in women. 39% of the world's adult population is overweight [Mendis 2014].

Methods and Results

Human Phenotype Network

Network theory provides a way to model large and complex systems through a comprehensive set of analytical tools. These tools can be used to explore and characterize, in a graphical context, the relationships between entities. A network is composed of nodes and their connecting edges, where the degree of a node refers to the number of attached edges. The degree distribution of a network represents the probability of each node having a specific degree and shows important global network properties. For example, a normal or a Poisson distribution of degree frequency indicates that the topology of the network may be random [Newman 2010]. Biological networks tend to fall in a category of their own, neither fully random nor regular. HPNs are constructed by linking phenotypic traits and diseases based on their shared biology [Darabos, et al. 2013]. This method integrates data from two sources for the phenotype-to-SNP mapping: The National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) GWAS catalog, a manually curated list of all the National Institute of Health (NIH) funded GWAS studies (accessed May 2017), and the NIH's database of Genotypes and Phenotypes (*dbGaP*, <http://www.ncbi.nlm.nih.gov/gap>) [MacArthur, et al. 2017]. This approach has been proven to offer novel insights into the way diverse phenotypes are linked by genetic variations. The network is built by first extracting every phenotype that is associated with at least one SNP

from the NHGRI-EMBL-EBI GWAS catalog and from *dbGaP*, in at least one study. Each node in our HPN represents a phenotype. Each vertex is then annotated with genes associated with the identified SNPs as explained below. We then identify overlapping loci, and connect the associated traits in the HPN with an undirected edge. The edge weight is proportional to the number of genes shared by the two phenotypes connected by that edge [Newman 2010]. By building phenotype-SNP-loci associations, we link phenotypes with shared biology. We remove isolated nodes that have no connections to the rest of the network, as we are only interested in traits that have been associated with a locus, and their putative shared biology. Some of the phenotype categories consist of related or redundant traits that are collapsed into one node (Supplementary Table 1).

For the study of large-scale data sets, visualization alone is often uninterpretable, and requires filtering (Figure 1). Many methods exist, ranging from simply filtering out low degree nodes to more sophisticated methods, such as extracting the “backbone” of the network, the latter relying on the local structure of the graph to retain the most relevant information [Darabos, et al. 2014b; Serrano, et al. 2009]. These methods are data-driven and rely solely on the network topology, not on any specific information content of the network. Alternatively, one can use knowledge-driven filters that rely on specific prior knowledge contained in the networks or based on additional analyses. Those filters are context specific, as opposed to structure-driven. In this work, we used ET to more intelligently filter the HPN. Given the complexity of unfiltered HPNs, as in this case with 1,591 nodes and 149,810 edges (Figure 1), which are too dense to interpret, our approach provides a means to make networks more interpretable and suitable for analysis.

Evolutionary Triangulation

We previously developed a method that utilizes the hypothesis that, due to historical migration and changes in population structures, the alleles that affect disease risk in multiple populations are distributed consistently with differences in population prevalence [Huang, et al. 2016]. This would imply that risk alleles have a higher frequency in those populations where a disease has a higher prevalence and a lower frequency where the disease is rarer. ET can be used to filter genomic regions with respect to disease or phenotypic variation by removing most highly differentiated SNPs between any two populations that are unlikely to associate with relevant phenotypes. We assessed the genetic differences between populations by calculating unbiased pairwise estimates of F_{ST} , using HapMap allele frequency data for each SNP [Weir and Cockerham 1984].

Applying ET to type 2 diabetes and related traits

Because diagnostic criteria for T2DM can vary among epidemiological studies, affecting its prevalence estimates, we chose to use fasting glucose, a continuous trait that is highly correlated with T2DM prevalence and can be used as a proxy for its occurrence. In contrast to T2DM diagnosis this measure represents a more objective, yet highly correlated, criterion, as the diagnostic conditions may vary from study to study. This stringency theoretically allows us to reduce false positives. High fasting glucose (as defined by WHO standards of 7.0 mmol/l, [Association 2016] as it is more stringent than other criteria) is found in 5.1%–7.8% of Northern and Western European males (the trend is the same for women), with an

average of 5.9%. In contrast, high fasting glucose is more frequent in West African countries with a prevalence of 7.5%–9.7% and an average of 8.26%. Its prevalence in South Asian countries ranges from 9.4%–10.8%, with an average of 9.76%. High fasting glucose prevalence is higher amongst Mexicans and lower in Northern Europeans and Japanese [Mendis 2014]. We accessed the high fasting glucose prevalence data from the World Health Organization's report on the global burden of Type 2 Diabetes, which uses fasting glucose levels as a diagnostic criteria to predict T2D [WHO 2016]. In contrast to T2DM prevalence numbers that may be based on variable diagnostic criteria, these data represent a more objective measure of T2DM related phenotypes.

Allele frequency data for these populations was derived from *HapMap* Project Phase III data, which included 113 Utah residents with Northern and Western European ethnicity (CEU), 88 Gujarati Indians from Houston (GIH), 113 Yoruba from Ibadan, Nigeria (YRI), 113 Japanese in Tokyo (JPT) and 86 with Mexican ancestry in Los Angeles, California (MEX) [Aguirre, et al. 2013; International HapMap, et al. 2010]. We performed two comparisons to select ET SNPs. The first was CEU_GIH_YRI and the second was MEX_CEU_JPT. We overlapped high F_{ST} (CEU_GIH), high F_{ST} (CEU_YRI) and low F_{ST} (GIH_YRI) for the first comparison and overlapped high F_{ST} (MEX_CEU), high F_{ST} (MEX_JPT) and low F_{ST} (CEU_JPT) for the second. For both comparisons, we first applied a stringent threshold, where the 95th percentile of F_{ST} represented high differentiation and the 5th represented high similarity. We then applied less stringent thresholds, i.e. the 90th percentile, the 85th percentile and the 80th percentile representing high differentiation, and the corresponding 10th, 15th and 25th percentiles representing similarity. Different F_{ST} thresholds produced different numbers of ET SNPs [Huang, et al. 2016]. Using these variable thresholds, we were able to adjust the number of ET SNPs for filtering the HPN.

Filtering the HPN using the ET SNPs subsets

During the creation of the HPN, each trait is annotated with the list of risk-associated SNPs integrated from the GWAS Catalog and *dbGaP*. Since the output of ET consists of lists of SNPs, this enables us to filter the HPN by cross-referencing the phenotypes associated with the ET SNP lists with the HPN trait list, and then retain only the phenotypes that have been associated with at least one ET SNP. This allows us to identify sub-networks that are most relevant to our target phenotypes.

ET-HPN Fasting Glucose Network

The number of nodes increased as less stringent ET thresholds were used (Table 1). The 90th and 95th percentiles of the CEU_YRI_GIH comparison have only one phenotype each and, clearly, no edges, giving us no relevant information. The 85th percentile HPN has 20 nodes and 162 edges. The 80th percentile contains 53 phenotypes (nodes) linked by 1,000 edges (Figure 2). The network for the 75th percentile contains 96 nodes and 3,407 edges (Supplemental Figure 1). In this study, we focus on the 80th percentile, as it provides a manageable picture of the filtered HPN in contrast to the unfiltered network without eliminating all complexity from the network (Figure 1 versus Figure 2). This somewhat arbitrary threshold serves to demonstrate the utility of the method, but a more systematic approach, which is outside the scope of the present study, will be developed in the future.

Here, we identified 26 diabetes related traits as well as 17 phenotypes related to cardiovascular diseases (Table 2). While most of the phenotypes were related to either T2DM or cardiovascular disease, many were related to both, which is not unexpected as T2DM is a risk factor for Coronary Artery Disease (Table 2 and Supplemental Table 1) [Lincoff 2003]. We also identified Amyotrophic Lateral Sclerosis (ALS) and melanoma related phenotypes (Figure 2). The latter was expected as melanoma is related to skin color, which is, in turn, related to continental ancestry. Both of these diseases are distributed similarly to T2DM. The prevalence of ALS is higher in people with European ancestry [Cronin, et al. 2007].

Intersection of comparisons

For the second comparison networks, MEX_CEU_JPT, the 95th, 90th and 85th percentiles had 0, 2 and 3 nodes respectively. The 80th percentile network contained 88 phenotypes (nodes) and 2,794 edges (Supplemental Figure 2), twenty-three of which were shared with the CEU_YRI_GIH derived network (Figure 3). These include: Activated partial thromboplastin time, Amyotrophic lateral sclerosis, Arteries, Bipolar disorder, Blood metabolite levels, Blood metabolite ratios, Body Height, Body Mass Index, Body Weights and Measures, Blood Pressure, Coronary Artery Disease, Elbow, Erythrocyte indices, Formal thought disorder in schizophrenia, Glucose levels, HDL cholesterol, Height, Hematological parameters, Metabolite levels, Myocardial infarction, Self-reported allergy, Uric acid and Waist circumference. These results emphasize the high degree of overlap between different ET-HPN networks and provide complementary evidence that the connections defined are both real and biologically relevant to connected diseases.

Discussions & Future Work

Through a combination of ET and HPN, we were able to develop subnetworks of genetic connections, enriching for our target disease when compared to using ET or HPN alone. This gives us more leverage in identifying areas for future study, as we may be able to facilitate detection of clusters of related phenotypes that are more likely to occur together, providing support for the idea that combining methods can elucidate novel and important connections that singular approaches cannot. It is often difficult to establish an accurate risk factor assessment for diseases with complex backgrounds. Some risk factors are genetic, whereas some are environmental or behavioral, or both. Our composite method allows us to understand links between T2DM related phenotypes and common genetic factors. We have shown that, if diseases are similarly distributed, they are more likely to be both pleiotropic and associated, as in the case of T2DM risk factors and ALS. Association and pleiotropy can inform each other and ET and HPN are more informative used in conjunction than either one is alone.

Type 2 Diabetes Mellitus

In our previous study using ET without HPN, we were unable to identify any genes known to be associated with T2DM or related phenotypes, although we used more than one set of population comparisons that differed with respect to T2DM prevalence [Huang, et al. 2016]. In contrast, using HPN without ET generated a network too complex to interpret. Our

present study identified 26 traits related directly to T2DM and several more related to cardiovascular disease. Two of the major and most telling phenotypes in our network, glycosylated hemoglobin and glucose levels, are used as diagnostic criteria for T2DM [Nemesure, et al. 2006]. Many other traits identified were also related to T2DM, including body mass index [Bays, et al. 2007], body weight and height [Costanzo, et al. 2015], waist circumference [InterAct, et al. 2012], high serum uric acid which is an independent risk factor for T2DM [Lv, et al. 2013], and vitamin D sufficiency may protect against T2DM [Dalgard, et al. 2011]. Blood trace elements, such as increased copper levels, contribute to T2DM [Zargar, et al. 1998]. Our network also showed a connection between T2DM and bipolar disorders, which can be confirmed with current literature [Regenold, et al. 2002]. Metabolite levels are indicative of diabetic nephropathy [Pena, et al. 2014]. Chronic Hepatitis C infection is highly correlated with T2DM [Gastaldi, et al. 2017]. T2DM is a risk factor for Barrett's esophagus and is hence a possible indirect association with esophageal cancer [Iyer, et al. 2013]. Cystic Fibrosis is not specifically T2DM related, but Cystic Fibrosis Related Diabetes shares similarity with both T2DM and Type 1 Diabetes [O'Shea and O'Connell 2014]. Many of these traits are closely linked, but not necessarily directly. Rather, by using ET-HPN we were able to identify phenotype nodes that were indirectly linked via one or more intermediates. This adds substantially to our ability to tie seemingly disparate phenotypes together and provide a map of common disease risk factors that would otherwise be extremely difficult to find.

Consistent with our network, recent work indicates a possible protective effect of T2DM in the development of ALS [Kioumourtoglou, et al. 2015; Sun, et al. 2015]. Although the mechanism is still unknown, a Danish nationwide cohort study suggests that both T2DM and obesity are protective against the development of ALS and also increase survival time for ALS patients. Our network has defined genetic links between ALS and T2DM risk factor thereby possibly generating novel hypotheses, regarding the mechanistic connections between these, hitherto, unrelated phenotypes (genes and phenotypes shown in Supplemental Table 2). Furthermore, our second analysis highlighted this connection again, which suggests the strength of the method. We postulate that the identification of these edges may provide useful information in understanding, if not predicting, both novel use of existing drugs and predictions of adverse drug reactions.

Cardiovascular Disease

Our network contains both coronary artery disease and myocardial infarction as well as many platelet function related traits. The ET-HPN contains HDL cholesterol, which is associated with lower risk of cardiovascular disease [Rohatgi, et al. 2014]. Our network also contains blood trace elements (Cu levels) [Carpenter, et al. 2013; Zargar, et al. 1998], circulating vasoactive peptide levels [Henning and Sawmiller 2001], platelet function and related traits [Gregg and Goldschmidt-Clermont 2003], arteries and erythrocyte indices [Andresdottir, et al. 2003], which are all blood related phenotypes. Additionally, our network includes glucose levels [Shaye, et al. 2012] and body weight [Costanzo, et al. 2015], which are related to both T2DM and CVD. Finally, we found neuroblastoma [Mulrooney, et al. 2009], calcium [Reid 2013] and hypertension [Carretero and Oparil 2000], which are all associated with cardiovascular disease. Interestingly, glycosylated hemoglobin is more

strongly associated with cardiovascular disease than is fasting glucose [Selvin, et al. 2010]. Other phenotypes also show relation to both T2DM and cardiovascular disease including HDL cholesterol, which has an inverse association in cancer risk among patients with T2DM [Zhao, et al. 2014]. Partial thromboplastin time is also related to cardiovascular disease and to T2DM [Zhao, et al. 2011], as is low blood platelet count [Demirtunc, et al. 2009]. In addition to its relationship to T2DM, ALS may effect cardiovascular health [Dalla Vecchia, et al. 2015]. Atrial fibrillation is also related to both cardiovascular disease and T2DM [Odutayo, et al. 2016] [Sun and Hu 2010]. These phenotypes are listed in Table 2.

Additional Phenotypes

Our ET-HPN analysis identified a number of other traits known to be differentially and appropriately distributed in our study populations. Four of the phenotypes in our network, self-reported allergy, basophils, asthma and hay fever and IgE levels are related to allergy and are connected in the ET-HPN. Although the allergy related results were more prominent in the 75th percentile threshold analyses (Supplemental Figure 1), they are all more common in people of African and South Asian descent than in those of Northern European ancestry, providing additional support that this method can detect results that may lead to the better understanding of multiple, related disease processes. We also found a cluster of interconnected nodes in the network that are related to sun exposure. Three of these phenotypes are related to melanoma development, suntan, perceived skin darkness and freckles [Greinert, et al. 2015]. Basal cell carcinoma and cutaneous squamous cell carcinoma also appear in same cluster of the network. Both cancers are related to sun exposure [Armstrong and Cust 2017]. Vitamin D levels also appears in the network, but in a different location [Thorne and Campbell 2008]. These connections are consistent with skin pigmentation differences among Europeans, Africans and South Asian. In retrospect, this is not unexpected as the melanoma related phenotypes, basal cell carcinoma and cutaneous squamous cell carcinoma are all related to sun exposure.

There is no way of determining a “universal” threshold for ET, as each disease has a different etiology and a single general threshold would likely decrease precision for many phenotypes. Each triangulation may have a vastly different number of SNPs, so the cutoff is a parameter entirely dependent upon the populations under study and the diseases of interest. One of our future areas of study will be to investigate the thresholds for different population comparisons. In our present study, the number of nodes increased from too few to be useful for the 95th and 90th percentiles to sixteen nodes for the 85th percentile. The number of nodes then doubles from the 85th to the 80th percentiles and again from the 80th to the 75th percentiles. We briefly looked at the 75th percentile and the number of nodes was, again, roughly double that of the 80th percentile. There is no clear indication at which point the background noise in the data becomes too large to provide more useful information. We plan, in the future, to investigate a possible inflection point at which the cutoff provides adequate information versus too much noise, by investigating the number of nodes for each ET cutoff, combined with the other important network properties.

A limiting factor in this type of research is incomplete knowledge of the relationships between diseases. Although many of the connections in the network are novel, they may

eventually be validated as our knowledge increases. An example of this is our identification of a genetic relationship between ALS and T2DM, which, after searching the literature, appears to be a valid connection [Kioumourtoglou, et al. 2015; Sun, et al. 2015].

Although this was an exploratory study, our results provide interesting connections that our approach can bring to the forefront that would otherwise be difficult to identify on a large scale, effectively automating what would take much longer to determine if done manually. It not only quickly identifies pleiotropic relationships but has the advantage that it connects secondary and tertiary connections that would otherwise be extremely difficult to find. These connections may be important to discovering co-occurring diseases and key pathways that affect disease risk. In conclusion, our analyses illustrated that HPN filtered by ET provides information that neither technique is able to alone and can effectively determine complex relationships among even distantly related diseases and phenotypes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JHM, SMW and CD acknowledge the support of the National Institute of Health (NIH) grants LM010098, LM009012. JHM was also supported by NIH grants AI116794, DK112217 and ES013508. LJM, SMW, BEG, and MH acknowledge the support of the March of Dimes Prematurity Research Center Ohio Collaborative grant. The computations in this work were performed on the Discovery HPC supported by the Research Computing group, ITS at Dartmouth College. No authors of this paper have a conflict of interest related to this work.

References

- Aguirre F, Brown A, Cho NH, Dahlquist G, Dodd S, Dunning T, Hirst M, Hwang C, Magliano D, Patterson C. IDF diabetes atlas. 2013
- Almobarak AO, Barakat S, Suliman EA, Elmadhoun WM, Mohamed NA, Abobaker IO, Noor SK, Bushara SO, Ahmed MH. Prevalence of and predictive factors for nonalcoholic fatty liver disease in Sudanese individuals with type 2 diabetes: Is metabolic syndrome the culprit? Arab J Gastroenterol. 2015
- Andresdottir MB, Sigfusson N, Sigvaldason H, Gudnason V. Erythrocyte sedimentation rate, an independent predictor of coronary heart disease in men and women: The Reykjavik Study. Am J Epidemiol. 2003; 158(9):844–51. [PubMed: 14585762]
- Armstrong BK, Cust AE. Sun exposure and skin cancer, and the puzzle of cutaneous melanoma: A perspective on Fears et al. Mathematical models of age and ultraviolet effects on the incidence of skin cancer among whites in the United States. American Journal of Epidemiology 1977. 2017; 105:420–427. Cancer Epidemiol.
- Association AD. Diagnosing Diabetes and Learning About Prediabetes. 2016
- Bays HE, Chapman RH, Grandy S, Group SI. The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. Int J Clin Pract. 2007; 61(5):737–47. [PubMed: 17493087]
- Carpenter WE, Lam D, Toney GM, Weintraub NL, Qin Z. Zinc, copper, and blood pressure: Human population studies. Med Sci Monit. 2013; 19:1–8. [PubMed: 23291705]
- Carretero OA, Oparil S. Essential hypertension. Part I: definition and etiology. Circulation. 2000; 101(3):329–35. [PubMed: 10645931]
- Costanzo P, Cleland JGF, Pellicori P, Clark AL, Hepburn D, Kilpatrick ES, Perrone-Filardi P, Zhang J, Atkin SL. The Obesity Paradox in Type 2 Diabetes Mellitus: Relationship of Body Mass Index to

- Prognosis A Cohort Study The Obesity Paradox in Type 2 Diabetes Mellitus. *Annals of Internal Medicine*. 2015; 162(9):610–618. [PubMed: 25938991]
- Cronin S, Hardiman O, Traynor BJ. Ethnic variation in the incidence of ALS: a systematic review. *Neurology*. 2007; 68(13):1002–7. [PubMed: 17389304]
- Dalgard C, Petersen MS, Weihe P, Grandjean P. Vitamin D status in relation to glucose metabolism and type 2 diabetes in septuagenarians. *Diabetes Care*. 2011; 34(6):1284–8. [PubMed: 21515843]
- Dalla Vecchia L, De Maria B, Marinou K, Sideri R, Lucini A, Porta A, Mora G. Cardiovascular neural regulation is impaired in amyotrophic lateral sclerosis patients. A study by spectral and complexity analysis of cardiovascular oscillations. *Physiol Meas*. 2015; 36(4):659–70. [PubMed: 25798998]
- Darabos C, Desai K, Cowper-Sallari R, Giacobini M, Graham B, Lupien M, Moore J. Inferring Human Phenotype Networks from Genome-Wide Genetic Associations. In: Vanneschi, L. Bush, W., Giacobini, M., editors. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer; Berlin Heidelberg: 2013. p. 23–34.
- Darabos C, Harmon SH, Moore JH. Using the bipartite human phenotype network to reveal pleiotropy and epistasis beyond the gene. *Pac Symp Biocomput*. 2014a:188–99. [PubMed: 24297546]
- Darabos C, White MJ, Graham BE, Leung DN, Williams SM, Moore JH. The multiscale backbone of the human phenotype network based on biological pathways. *BioData Min*. 2014b; 7(1):1. [PubMed: 24460644]
- Demirtunc R, Duman D, Basar M, Bilgi M, Teomete M, Garip T. The relationship between glycemic control and platelet activity in type 2 diabetes mellitus. *J Diabetes Complications*. 2009; 23(2):89–94. [PubMed: 18358749]
- Franks PW. Gene x environment interactions in type 2 diabetes. *Curr Diab Rep*. 2011; 11(6):552–61. [PubMed: 21887612]
- Gastaldi G, Goossens N, Clement S, Negro F. Current level of evidence on causal association between hepatitis C virus and type 2 diabetes: A review. *J Adv Res*. 2017; 8(2):149–159. [PubMed: 28149650]
- Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007; 104(21):8685–90. [PubMed: 17502601]
- Gregg D, Goldschmidt-Clermont PJ. Cardiology patient page. Platelets and cardiovascular disease. *Circulation*. 2003; 108(13):e88–90. [PubMed: 14517153]
- Greinert R, de Vries E, Erdmann F, Espina C, Auvinen A, Kesminiene A, Schuz J. European Code against Cancer 4th edition: Ultraviolet radiation and cancer. *Cancer Epidemiol*. 2015
- Henning RJ, Sawmiller DR. Vasoactive intestinal peptide: cardiovascular effects. 2001
- Huang M, Graham BE, Zhang G, Harder R, Kodaman N, Moore JH, Muglia L, Williams SM. Evolutionary triangulation: informing genetic association studies with evolutionary evidence. *BioData Min*. 2016; 9(1):1–18. [PubMed: 26734074]
- InterAct C, Langenberg C, Sharp SJ, Schulze MB, Rolandsson O, Overvad K, Forouhi NG, Spranger J, Drogan D, Huerta JM, et al. Long-term risk of incident type 2 diabetes and measures of overall and regional obesity: the EPIC-InterAct case-cohort study. *PLoS Med*. 2012; 9(6):e1001230. [PubMed: 22679397]
- International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467(7311):52–8. [PubMed: 20811451]
- Iyer PG, Borah BJ, Heien HC, Das A, Cooper GS, Chak A. Association of Barrett's esophagus with type II Diabetes Mellitus: results from a large population-based case-control study. *Clin Gastroenterol Hepatol*. 2013; 11(9):1108–1114 e5. [PubMed: 23591277]
- Kioumourtoglou MA, Rotem RS, Seals RM, Gredal O, Hansen J, Weisskopf MG. Diabetes Mellitus, Obesity, and Diagnosis of Amyotrophic Lateral Sclerosis: A Population-Based Study. *JAMA Neurol*. 2015
- Lincoff AM. Important triad in cardiovascular medicine: diabetes, coronary intervention, and platelet glycoprotein IIb/IIIa receptor blockade. *Circulation*. 2003; 107(11):1556–9. [PubMed: 12654616]
- Lv Q, Meng XF, He FF, Chen S, Su H, Xiong J, Gao P, Tian XJ, Liu JS, Zhu ZH, et al. High serum uric acid and increased risk of type 2 diabetes: a systemic review and meta-analysis of prospective cohort studies. *PLoS One*. 2013; 8(2):e56864. [PubMed: 23437258]

- MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 2017; 45(D1):D896–D901. [PubMed: 27899670]
- Mendis, S. Global Status Report On Noncommunicable Diseases 2014. World Health Organization; 2014.
- Mulrooney DA, Yeazel MW, Kawashima T, Mertens AC, Mitby P, Stovall M, Donaldson SS, Green DM, Sklar CA, Robison LL, et al. Cardiac outcomes in a cohort of adult survivors of childhood and adolescent cancer: retrospective analysis of the Childhood Cancer Survivor Study cohort. *BMJ : British Medical Journal.* 2009; 339:b4606. [PubMed: 19996459]
- Nemesure B, Wu SY, Hennis A, Leske MC, Barbados Eye Study G. Hypertension, type 2 diabetes, and blood groups in a population of African ancestry. *Ethn Dis.* 2006; 16(4):822–9. [PubMed: 17061733]
- Newman, MEJ. *Networks : an introduction.* Oxford: Oxford University Press; 2010.
- O'Shea D, O'Connell J. Cystic fibrosis related diabetes. *Curr Diab Rep.* 2014; 14(8):511. [PubMed: 24915888]
- Odutayo A, Wong CX, Hsiao AJ, Hopewell S, Altman DG, Emdin CA. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis. *BMJ.* 2016; 354:i4482. [PubMed: 27599725]
- Pena MJ, Lambers Heerspink HJ, Hellemons ME, Friedrich T, Dallmann G, Lajer M, Bakker SJ, Gansevoort RT, Rossing P, de Zeeuw D, et al. Urine and plasma metabolites predict the development of diabetic nephropathy in individuals with Type 2 diabetes mellitus. *Diabet Med.* 2014; 31(9):1138–47. [PubMed: 24661264]
- Regenold WT, Thapar RK, Marano C, Gavirneni S, Kondapavuluru PV. Increased prevalence of type 2 diabetes mellitus among psychiatric inpatients with bipolar I affective and schizoaffective disorders independent of psychotropic drug use. *J Affect Disord.* 2002; 70(1):19–26. [PubMed: 12113916]
- Reid IR. Cardiovascular Effects of Calcium Supplements. *Nutrients.* 2013; 5(7):2522–2529. [PubMed: 23857224]
- Rohatgi A, Khera A, Berry JD, Givens EG, Ayers CR, Wedin KE, Neeland IJ, Yuhanna IS, Rader DR, de Lemos JA, et al. HDL Cholesterol Efflux Capacity and Incident Cardiovascular Events. *New England Journal of Medicine.* 2014; 371(25):2383–2393. [PubMed: 25404125]
- Selvin E, Steffes MW, Zhu H, Matsushita K, Wagenknecht L, Pankow J, Coresh J, Brancati FL. Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults. *New England Journal of Medicine.* 2010; 362(9):800–811. [PubMed: 20200384]
- Serrano MA, Boguna M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci U S A.* 2009; 106(16):6483–8. [PubMed: 19357301]
- Shaye K, Amir T, Shlomo S, Yechezkel S. Fasting glucose levels within the high normal range predict cardiovascular outcome. *Am Heart J.* 2012; 164(1):111–6. [PubMed: 22795290]
- Stevens PE, Schernthaner G, Raptis S, Wanner C, Scherhag A, Lameire N. Characteristics, Cardiovascular Comorbidity and Medicines Management in Patients with Type 2 Diabetes and CKD: Results of the IRIDIEM Study. *Kidney and Blood Pressure Research.* 2010; 33(2):119–128. [PubMed: 20424472]
- Sun Y, Hu D. The link between diabetes and atrial fibrillation: cause or correlation? *J Cardiovasc Dis Res.* 2010; 1(1):10–1. [PubMed: 21188083]
- Sun Y, Lu CJ, Chen RC, Hou WH, Li CY. Risk of Amyotrophic Lateral Sclerosis in Patients With Diabetes: A Nationwide Population-Based Cohort Study. *J Epidemiol.* 2015; 25(6):445–51. [PubMed: 25947580]
- Temelkova-Kurktschiev T, Stefanov T. Lifestyle and genetics in obesity and type 2 diabetes. *Exp Clin Endocrinol Diabetes.* 2012; 120(1):1–6. [PubMed: 21915815]
- Thorne J, Campbell MJ. The vitamin D receptor in cancer. *Proc Nutr Soc.* 2008; 67(2):115–27. [PubMed: 18412986]
- Weir BS, Cockerham C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution.* 1984; 38(6):1358–1370. [PubMed: 28563791]

- WHO. Prevalence of raised fasting blood glucose*, ages 18+, 2010–2014 (age standardized estimate). 2016
- Zargar AH, Shah NA, Masoodi SR, Laway BA, Dar FA, Khan AR, Sofi FA, Wani AI. Copper, zinc, and magnesium levels in non-insulin dependent diabetes mellitus. *Postgrad Med J*. 1998; 74(877): 665–8. [PubMed: 10197198]
- Zhao W, Guan J, Horswell R, Li W, Wang Y, Wu X, Hu G. HDL cholesterol and cancer risk among patients with type 2 diabetes. *Diabetes Care*. 2014; 37(12):3196–203. [PubMed: 25216507]
- Zhao Y, Zhang J, Zhang J, Wu J. Diabetes mellitus is associated with shortened activated partial thromboplastin time and increased fibrinogen values. *PLoS One*. 2011; 6(1):e16470. [PubMed: 21297995]

Figure 1.
The Unfiltered Human Phenotype Network. The HPN contains 1,591 nodes and 149,810 edges. Node colors represent computationally determined modules.

Figure 2.

The ET Filtered HPN. This network was filtered using the 80th percentile ET threshold and contained 53 nodes and 1000 edges. The majority of the nodes are diabetes or cardiovascular disease related. Node colors represent computationally determined modules.

Figure 3.
Network intersection. The intersection of the CEU_GHI_YRI and MEX_JPT_CEU networks at the 80th percentile.

Statistics for ET-filtered networks from the 80th–95th percentile for CEU-YRI-GIH populations.

Table 1

Statistics	HPN	75 th	80 th	85 th	90 th	95 th
Nodes	1,591	96	53	20	1	1
Edges	149,810	3,407	1000	162	0	0
Ave. Degree	188.322	70.979	37.736	16.2	0	0
Ave. Weighted Degree	544.963	1,652.854	1,013.547	575.5	0	0
Ave. Path Length	1.941	1.254	1.279	1.147	0	0
Network Diameter	4	3	3	2	0	0
Modularity	0.193	0.053	0.066	0.106	NA	NA
Density	0.118	0.747	0.726	0.853	NA	NA
Connected Components	8	1	1	1	1	1
Ave. Clustering Coefficient	0.777	0.891	0.891	0.925	NA	NA

Table 2Network Phenotypes from 80th percentile

Phenotype Category	Phenotype
Diabetes related phenotypes	Activated partial thromboplastin time
	Amyotrophic lateral sclerosis
	Atrial fibrillation [†]
	Barrett's esophagus
	Bipolar disorder
	Blood metabolite levels
	Blood metabolite ratios
	Blood platelets
	Blood trace elements (Cu levels)
	Body height
	Body mass index [†]
	Body weights and measures [†]
	Chronic hepatitis C infection
	Cystic fibrosis
	Formal thought disorder in schizophrenia
	Glucose levels [†]
	Glycosylated hemoglobin A
	HDL cholesterol [†]
	Height
	Hematological parameters [†]
	Metabolite levels
	Partial thromboplastin time [†]
	Platelet function and related traits
	Uric acid
	Vitamin D [*]
	Waist circumference
Cardiovascular related phenotypes	Arteries
	Atrial fibrillation [†]
	Blood pressure
	Body mass index [†]
	Body Weights and Measures [†]
	Glucose levels [†]
	Calcium
	Circulating vasoactive peptide levels

Phenotype Category	Phenotype
	Coronary artery disease
	Electrocardiography
	Erythrocyte indices
	HDL cholesterol [†]
	Hematological parameters [†]
	Hypertension
	Myocardial infarction
	Neuroblastoma
	Partial thromboplastin time [†]
Sun exposure related skin phenotypes	Basal cell carcinoma
	Cutaneous squamous cell carcinoma
	Freckles
	Perceived skin darkness
	Suntan
	Vitamin D [*]
Asthma related phenotypes	Asthma and hay fever
	Basophils
	IgE levels
	Self-reported allergy
Additional phenotypes	Attention Deficit Disorder with Hyperactivity
	Crohn's disease
	Elbow
	Essential tremor
	Fast beta electroencephalogram
	Neonatal lupus
	Temporal lobe

[†]Phenotypes related to both Cardiovascular disease and T2DM

^{*}Vitamin D is also related to melanoma