Coherence of Molecular Mechanisms In Major Human Disease and Traits

Dozmorov, Mikhail G.[1+] (mikhail.dozmorov@vcuhealth.org)
Cresswell, Kellen G.[1+] (cresswellkg@vcu.edu)
Bacanu, Silviu-Alin[2] (silviu-alin.bacanu@vcuhealth.org)
Craver, Carl[4] (ccraver@wustl.edu)
Reimers, Mark[3] (reimersm@msu.edu)
Kendler, Kenneth S.[2] (kenneth.kendler@vcuhealth.org)

[1] - Department of Biostatistics, Virginia Commonwealth University
[2] - Virginia Institute for Psychiatric and Behavior Genetics and the Department of Psychiatry, Virginia Commonwealth University
[3] - Neuroscience Program and Dept. Biomedical Engineering Michigan State
[4] - Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis
[+] - equal contribution

Abstract

Background. Complex phenotypes such as height and intelligence, are thought to be a product of the collective effects of multiple phenotype-associated genes, or, more precisely, interactions among their protein products. High/low degree of interactions is suggestive of coherent/random molecular mechanisms, respectively. Comparing the degree of interactions may help to better understand the coherence of molecular mechanisms underlying complex phenotypes and the potential for therapeutic intervention. However, direct comparison of the degree of interactions is difficult due to different sizes and configurations of phenotype-associated gene networks.

Methods. We introduce a measure of network coherence as a slope of internal vs. external distributions of the degree of interactions. The internal degree distribution is defined by interaction counts within a phenotype-specific gene network, while the external degree distribution counts interactions with other genes in the whole protein-protein interaction (PPI) network. We present a novel method for normalizing the coherence estimates, making them directly comparable.

Results. Using STRING and BioGrid PPI databases, we compared the coherence of 116 phenotype-associated gene sets from GWAScatalog against size-matched KEGG pathways (the reference for high coherence) and random networks (the lower limit of coherence). We observed a range of coherence estimates for each category of phenotypes. Metabolic traits and diseases were the most coherent, while psychiatric disorders and intelligence-related traits were the least coherent.

Conclusions. We present a general-purpose method for estimating and comparing the coherence of phenotype-associated gene networks that accounts for the network size and shape differences. Our results highlight gaps in our current knowledge of genetics and molecular mechanisms of complex phenotypes and suggest priorities for future GWASs.

Keywords: GWAS, network, degree, coherence

Introduction

Genome-wide association studies (GWAS) have significantly advanced our understanding of complex phenotypes by identifying disease- and trait-associated genetic markers and suggesting corresponding genes [1, 2]. However, GWAS findings explain only a fraction of

heritability [3–5]. Furthermore, the corresponding genes are often spread across different chromosomes with no known connection to one another, hindering understanding of the molecular mechanisms. These limitations of current generation GWAS might stem from the yet incomplete knowledge of genetic determinants of complex phenotypes, the potential heterogeneity of complex phenotypes and/or the causes other than genetics [6–8].

Many studies have shown that genetically-driven complex phenotypes are often associated with functionally related genes that are more likely to form networks of interacting protein products [9–15]. This observation has been verified systematically for a large number of diseases [16], thus confirming a fundamental hypothesis of the interactome-based approach to understanding human phenotypes, namely that disease- and trait-associated genes tend to form interaction modules [17]. Such interaction modules may be viewed as connected subnetworks within the full interactome and may contain all molecular determinants of a certain phenotype. Consequently, methods to identify phenotype-associated networks of functionally-related genes have been developed [10, 13, 18].

Network properties, such as connectivity (aka degree distribution) [15, 19–21], can be compared to better understand the relationships among phenotypes [22–24]. Networks formed by interactions among phenotype-associated genes, or, more precisely, their protein products, are thought to have high connectivity [10, 11, 17]. This intuition reflects a well-known “guilt-by-association” principle that genes (or, more precisely, their products) that form an interaction network are more likely to share similar functions [25], or co-expression patterns [11, 26]. Although this view has been criticized [27], several studies consistently observed phenotype-associated genes to be either connected as a single network or to participate in common phenotype-specific subnetworks [16, 18, 28–30]. Therefore, high connectivity among phenotype-specific genes may indicate coherent molecular mechanisms that could be targeted therapeutically. In terms of graph theory, we are asking whether a phenotype-specific network is a community - a highly connected subgraph relatively well-separated from the rest of the network [31]. However, the direct comparison of connectivity across phenotype-specific networks is hindered by the fact that different phenotypes are associated with different numbers of genes forming networks of different sizes and configurations.

This study presents a novel measure of network coherence as a slope of the internal versus external degree distributions. The internal degree distribution is defined as gene-specific interaction counts within a protein-protein interaction (PPI) network formed by phenotype-associated genes, while the external degree distribution counts interactions with other genes in the whole PPI network. We used selected gene sets from the MSigDb database [32] as a reference for highly coherent networks, while sets of randomly sampled genes were used as the reference for the absence of coherence. Using MSigDb and random networks matched in size to phenotype-specific networks, we derived a normalized measure of coherence that can be compared across phenotypes. We hypothesized that the level of coherence may inform us about similarities and differences among phenotypes. Using two PPI databases (STRING [33] and BioGrid [34]), we compared coherence of 133 phenotypes from GWAScatalog [35, 36]. Our results show the tendency of complex phenotypes, such as intelligence-related traits, to have low coherence and metabolic traits and diseases to have high coherence. Our method enables direct comparison of coherence measures and highlights gaps in the current understanding of molecular mechanisms of many phenotypes, e.g., Major Depressive Disorder having the lowest coherence in the already low-coherence “Psychiatric disease” category.

Results

Phenotype-specific networks of protein-protein interactions (PPIs)

We collected 133 phenotype-associated gene lists from the NHGRI-EBI GWAS catalog [36] (see Methods). They were grouped into seven disease categories ("Autoimmune", "Cancer", "Cardiovascular disease", "Eye disease", "Metabolic disease", "Neurologic", "Psychiatric") and five trait categories ("Anthropometric trait", "Cardiovascular trait", "Eye trait", "Intelligence", "Metabolic trait", Table 1, Supplementary Table S1).

We created networks of phenotype-associated genes using PPI information from STRING [33] and BioGrid [34] databases (see Methods). We opted for the use of three types of PPI data. First, STRING data filtered by interaction confidence score above 500 (middle of the bimodal distribution of the score, 0-1000 range, referred hereafter as "STRING filtered") was used as the primary source of curated PPI data. This filtering step ensures that only high-confidence PPIs were selected. The advantage of using filtered data is the reliance on high-quality PPIs. The disadvantage is that the number of PPIs may be insufficient for forming phenotype-associated networks; consequently, phenotypes with genes without PPIs were removed from the analysis (see Methods) leaving 116 phenotypes that could be analyzed. Second, full STRING data was used to maximize the use of PPI information at the expense of potential noise (referred to hereafter as "STRING"). The advantage of using complete data is that more trait-associated genes will have PPI information and, hence, can be analyzed. The disadvantage, however, is that the results may be less reliable due to the presence of noisy PPIs having low confidence scores. Third, BioGrid PPI data, which has been curated to keep only high-confidence interactions, was used. The use of different data sources (STRING and BioGrid) and filtering (full or filtered STRING data) was intended to increase the generalizability of our conclusions.

Table 1. Summary statistics of the analyzed categories. Average coherence estimates are shown. "NA" values indicate that a category lacked phenotypes with a sufficient number of PPIs. Sorted by "String Filtered" average coherence.

Category	Number of SNPs	Number of Genes	Total Diseases	Mean Biogrid Coherence	Mean String Coherence	Mean String Filtered Coherence
Trait						
Eye trait	158	109	5	NA	0.724	1.050
Metabolic trait	790	646	10	0.738	0.779	0.863
Anthropometric trait	2090	1771	15	0.460	0.637	0.691
Cardiovascular trait	728	604	15	0.555	0.675	0.625
Intelligence	181	177	4	0.180	0.395	0.405
Disease						
Metabolic disease	331	269	4	0.470	0.825	0.907
Cancer	670	570	15	0.490	0.689	0.723
Neurologic	372	333	7	NA	0.493	0.718

Autoimmune	1383	1184	18	0.425	0.820	0.701
Eye disease	229	202	5	0.330	0.630	0.663
Cardiovascular disease	301	259	7	0.380	0.694	0.660
Psychiatric	678	611	11	0.540	0.552	0.642

Degree distribution as a measure of coherence

A typical network consists of nodes connected by edges [21]. In genomics, nodes typically represent genes and edges correspond to some measure of interactions, e.g., gene co-expression or interactions between protein products [37, 38]. In terms of phenotypes, an intuitive expectation is that they will be represented by coherent networks of functionally related genes, similar gene expression profiles, shared genomic variants, higher PPI interactions, and higher co-morbidity [16, 17, 38–42], reviewed in [24, 43]. In terms of graph theory, a coherent network is a community consisting of a group of nodes that are relatively highly connected to each other but sparsely connected to other nodes in the global network [31, 44].

Numerous network properties have been defined to describe network structures [23, 38]. A degree of a node, or connectivity, is one of the most fundamental characteristics defining the number of other nodes connected to a given node. A collection of degrees of all network nodes forms a degree distribution. Comparing degree distributions among networks is an intuitive way to gain an understanding of network similarities and differences in terms of connectivity (aka coherence) of the corresponding genes [15, 21]. We hypothesized that comparing degree distributions of networks formed by phenotype-associated gene sets may inform us about similarities and differences among phenotypes in terms of coherence estimates, informing us about the underlying molecular mechanisms.

Internal vs. external degree distributions as a measure of coherence

Given that phenotype-specific networks occur within the global network of PPIs, we developed a metric of coherence based on the level of gene interactions. This metric is inspired by previous work showing that communities within a large network have high internal but low external levels of interactions (degrees) [45, 46]. Thus, for each phenotype-associated PPI network, we considered the relationship between its internal and external degree distributions. The internal degree distribution is defined by considering edges within an isolated network formed by phenotype-associated genes. The external degree distribution is defined by considering edges to other genes in the whole PPI network. Intuitively, a highly coherent network is expected to have a large number of internal edges but a relatively small number of external edges. An extreme example of high coherence would be fully internally connected network with zero interactions with external genes. Conversely, a low-coherence network is expected to have a relatively low number of internal edges similar to the number of external edges.

To compare coherences, we estimated coherence as the slope of a line through a scatterplot of internal vs. external degree distributions. That is, we visualized internal (X-axis) vs. external (Y-axis) degrees for each phenotype-specific gene on a single plot and fit a regression line through it, enforcing the fit through the origin. A network with the highest coherence (full internal connectivity, zero external connections) would be represented by a horizontal line. Conversely, a network with the lowest coherence (zero internal connectivity, full external connectivity) would be represented by a vertical line. Consequently, slopes of internal vs. external degree distributions of phenotype-specific networks would represent the corresponding levels of

coherence. In summary, comparing internal vs. external degree distributions represents a viable metric to quantify and compare the molecular coherence of phenotype-associated networks.

Networks of KEGG/REACTOME pathways and GO Cellular Component collection vs. randomly selected genes serve as references for high vs. low coherence, respectively

The coherence of phenotype-associated gene sets should be measured with respect to the realistic references of high and low coherence. To establish a reference of high coherence, we considered networks from collections of the Molecular Signatures Database (MSigDB v6.2) [32]. MSigDB contains sets of genes having various types of functional relationships. We found that pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [47] had the smallest average slope corresponding to the highest level of coherence (Supplementary Table S2). Networks assembled from Gene Ontology - Cellular Component (GOCC) collection and REACTOME pathways followed, representing alternative sources of networks with high coherence. This is expected as genes expressed in the same cellular components (GOCC genes) or participating in the same metabolic pathways (REACTOME genes) are presumed to interact more frequently. In our study, KEGG networks were used in parallel with GOCC and REACTOME (referred hereafter as MSigDb networks) as a reference for high coherence. Conversely, as a reference for low coherence, we randomly sampled genes from a pool of genes in a given PPI database. These random gene lists represent expected network coherence arising by chance. The slopes of MSigDb and random networks represent reference levels for high and low coherence, respectively.

Network size and configuration hinders direct comparison of degree distributions

Direct comparison of degree distributions of the networks formed by phenotype-specific genes is hindered by the fact they depend on network size and configuration [48]. An intuitive example is to consider a ring and a fully connected network of size 10 and 20 nodes, respectively (Figure 1). The degree distributions of ring networks can be directly compared regardless of network size (vectors of “2” of length 10 and 20, respectively, Figure 1A). Although the fully connected networks are similar in that they are fully connected (i.e., similarly coherent), their degree distributions differ (all nodes in the 10-node network have a degree “9”, while all nodes in the 20-node network have degree “19”, Figure 1B). These observations highlight the difficulty in comparing degree distributions of phenotype-associated gene networks due to two facts: 1) the network sizes differ, and 2) the network configuration (e.g., fully connected, ring, or intermediate connectivity) is unknown.

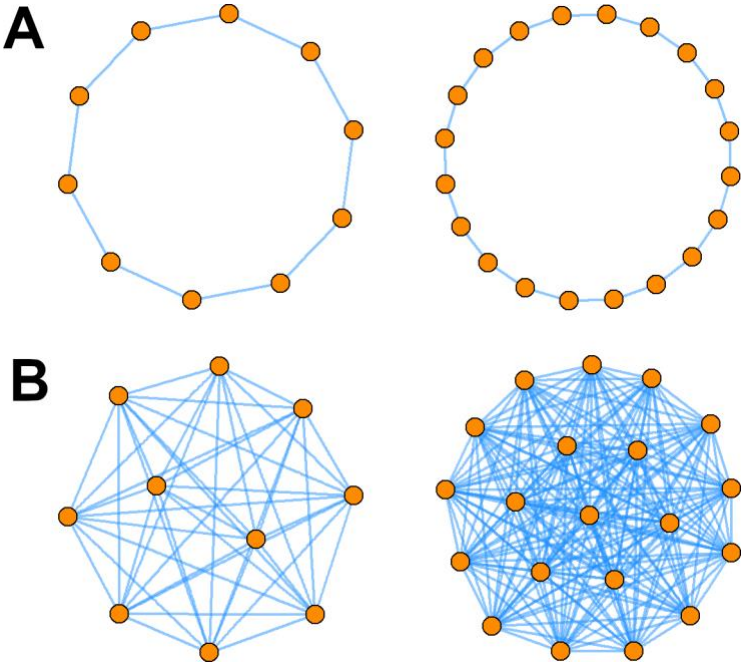

Figure 1. Degree distributions are affected by network size and configuration. Degree distributions of A) ring networks and B) fully connected networks containing 10 and 20 nodes. While degree distributions of similarly coherent ring networks can be directly compared, degree distributions of fully connected networks depend on network size.

To observe whether network properties affect coherence estimation in experimental settings, we investigated the effect of network size on the slopes of KEGG and random networks. We observed substantial negative correlation between network sizes and the slopes of KEGG and random networks (Pearson Correlation Coefficient (PCC) -0.58/-0.86, respectively, Figure 2, Supplementary Table S2). These observations confirm the notion that coherence estimation using internal vs. external degree distributions should be controlled for network size.

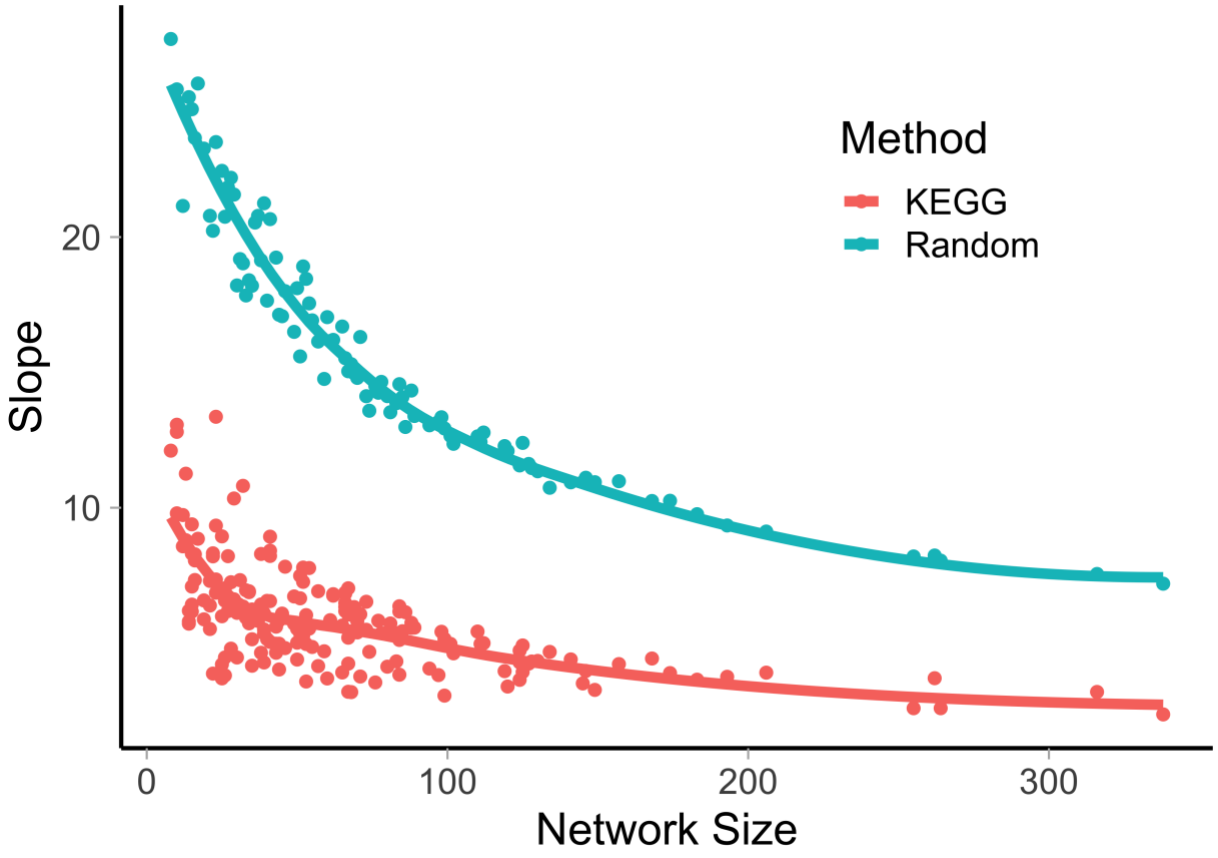

Figure 2. Network size is inversely associated with coherence. Slopes of internal vs. external degree distributions (aka coherence) for KEGG and random networks are plotted against network size.

Network size-independent estimation of coherence

Given the dependence of degree distributions on network size, we designed a simple strategy to account for it. Briefly, for each phenotype-specific network of a given size, we created size-matched references of high and low coherence. Specifically, we selected 30 MSigDb networks (10 from KEGG, GOCC, and REACTOME collections) and 100 random networks matched in size to the corresponding network of phenotype-specific genes. Consequently, we use median slopes of internal vs. external degree distributions created by these size-matched references to normalize the slopes of phenotype-specific networks to the [0, 1] range (see Methods), where 0/1 correspond to random/high coherence, respectively. We found that the use of size-matched references indeed alleviates the dependency of coherence on network size (Average PCC = 0.01, Supplementary Figure S1) and allows for direct comparison of normalized coherences across phenotypes.

Results derived from different PPI data are largely consistent

In addition to size dependency, coherence estimates depend on the choice of the PPI database used to build networks and estimate the corresponding internal vs. external degree distributions. To evaluate the consistency of coherence estimates, we assessed the pairwise correlation of normalized coherences obtained using different PPI databases. Correlation between coherence estimates obtained using “STRING” and “STRING filtered” was high (PCC = 0.80,

Supplementary Figure S1). Expectedly, coherence estimates using Biogrid database were less similar to that of “STRING” and “STRING filtered”; however, the overall correlation remained high. Notably, coherence estimates using “STRING filtered” database were most similar to that of “Biogrid” (PCC = 0.58), indicating that filtering step indeed removes “noisy” PPIs and improves quality of coherence estimates. Overall, these results demonstrate consistency in coherence estimates using different PPI databases.

Metabolic and intelligence-related traits as examples of networks with overall highest/lowest coherence

We estimated coherence of 49 traits from five categories (Figure 3, Supplementary Table S3, Supplementary Figure S2, 6). Among anthropometric traits, Weight and Body Mass Index had the highest level of coherence. Other examples of traits with high coherence include blood pressure-related traits (cardiovascular) and glucose-related traits (metabolic). Notably, the metabolic trait category had the highest overall coherence (mean coherence 0.86, Table 1), and these results were consistent when using other PPI databases (Supplementary Table S1). Several notable examples of low coherence include Obesity-related traits, Height (anthropometric traits), and blood count-related traits (cardiovascular). Traits in the Intelligence category had low overall coherence (mean coherence = 0.41).

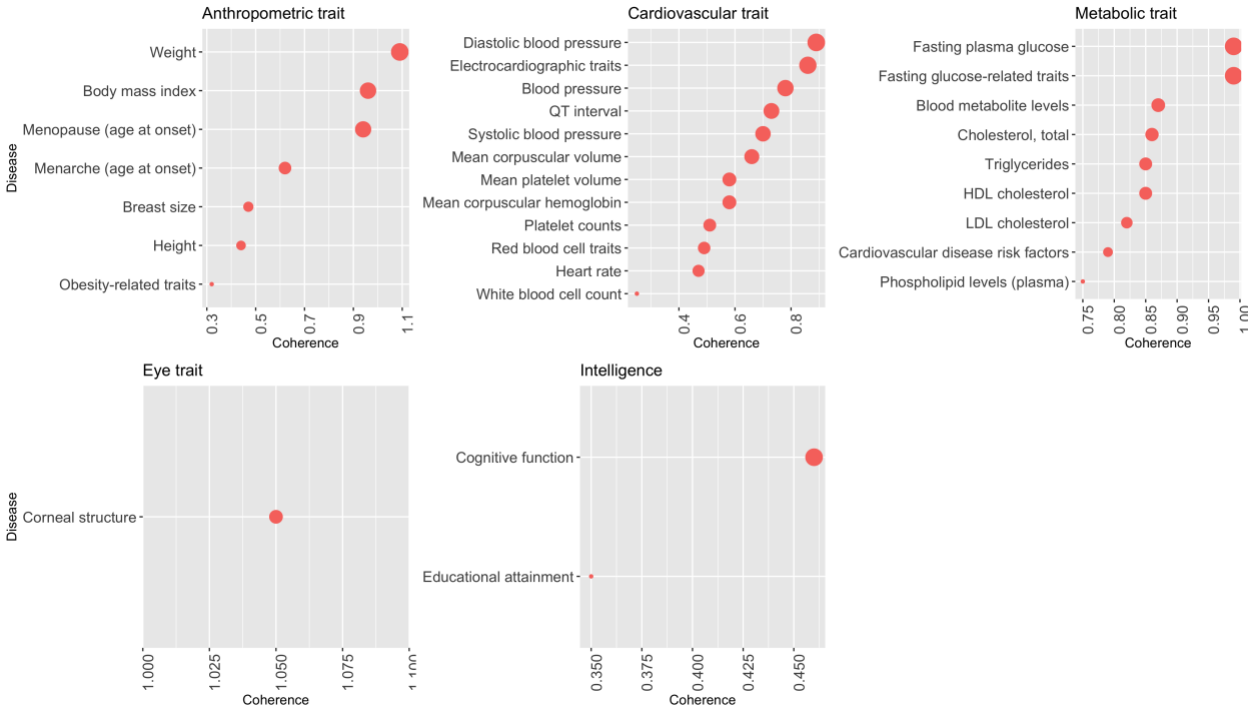

Figure 3. Coherence estimates of traits. Size of dots represents the level of normalized coherence (X-axis) for individual traits (Y-axis).

Metabolic diseases and cancer networks have high coherence

We further estimated coherence of 67 diseases from seven categories (Figure 4, Supplementary Figure S4, 8, Supplementary Table S3). Metabolic diseases had high overall coherence (average coherence 0.91, Table 1), with “Obesity” and “Type 2 diabetes” being the most coherent (1.01 and 0.91, respectively). Similarly, the average coherence for cancer

diseases was 0.72, with “Testicular germ cell tumor” and “Pancreatic cancer” being highly coherent (0.92 and 0.74, respectively).

Intermediate coherence of autoimmune, neurologic, and eye diseases

“Rheumatoid arthritis”, “Systemic Lupus Erythematosus”, and “Systemic sclerosis” were the most coherent in the “Autoimmune” category (0.86, 0.80, 0.79, respectively). On the other end of the coherence spectrum, “Allergic rinitis” and “Atopic dermatitis” were the least coherent (0.51, 0.60, respectively). The average coherence of autoimmune diseases was high, 0.70 (Table 1), with all autoimmune diseases, except “Allergic rinitis”, having coherence estimates > 0.60 (Supplementary Table S3). Diseases in the “Neurologic” category were similarly highly coherent (average coherence 0.72). However, the average coherence of neurologic disorders using STRING database was 0.49, suggesting that, with larger number of networks being estimated the overall coherence of neurologic diseases is low. “Alzheimer’s disease” and “Parkinson’s disease” were the most coherent (0.88 and 0.84, respectively), while “Migraine” was the least coherent (0.54). Diseases in the “Eye disease” category were more heterogeneous, with coherence ranging from 0.35 for “Corneal astigmatism” to 1.01 for “Refractive error”. The average coherence of diseases in the “Eye disease” category was relatively high (0.66).

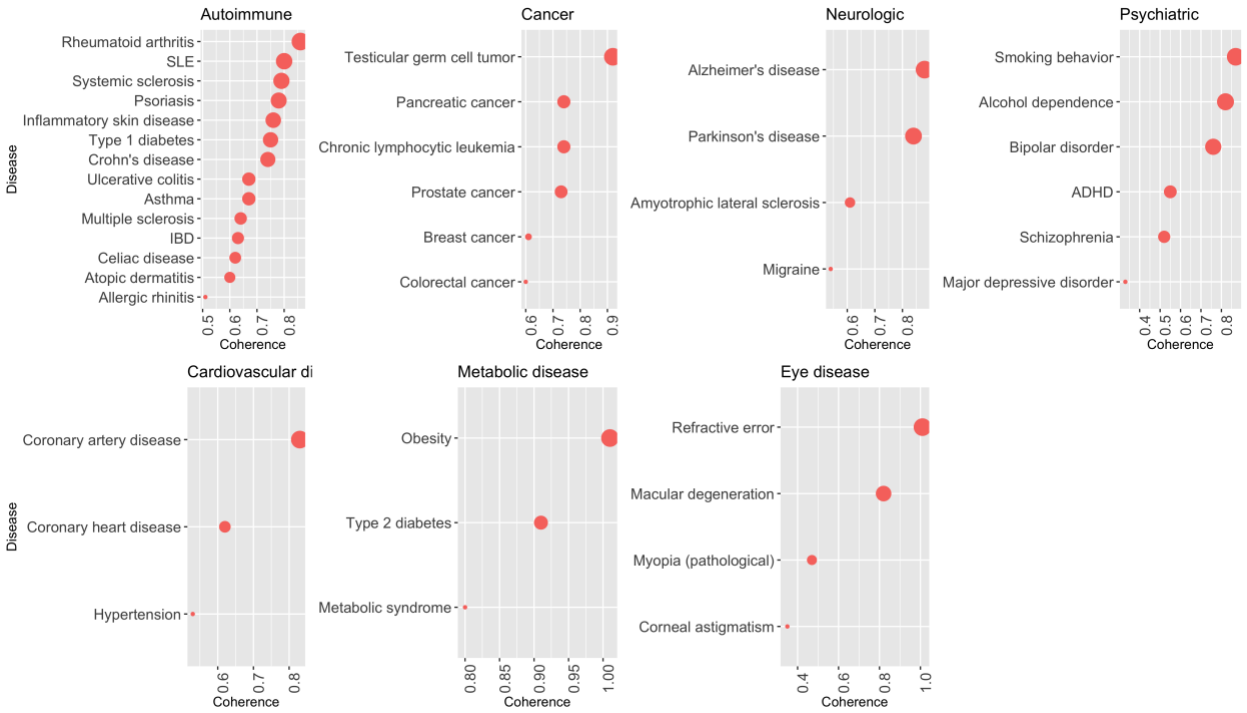

Figure 4. Coherence estimates of diseases. Size of dots represents the level of normalized coherence (X-axis) for individual diseases (Y-axis).

Psychiatric and cardiovascular diseases as examples of low coherence

Diseases in “Psychiatric” and “Cardiovascular” categories showed the lowest average level of coherence (0.64 and 0.66, respectively, Table 1). The corresponding coherence estimates were highly heterogeneous. Among cardiovascular diseases, “Coronary artery disease” showed the highest level of coherence (0.83), while “Hypertension” had the lowest coherence (0.53). Among psychiatric disorders, “Smoking behavior” and “Alcohol dependence” had high coherence (0.87

and 0.82, respectively). Notably, coherences of “Schizophrenia” and “Major depressive disorder” networks (0.52 and 0.33, respectively) were at the lower end of coherence estimates among all phenotypes. Other psychiatric disorders, such as “Bipolar disorder”, “ADHD”, showed intermediate coherence estimates (0.76 and 0.55, respectively).

Coherence of most phenotypes is significantly larger than random networks

Given the wide range of phenotype-specific coherence estimates, a natural question is whether they are significantly larger than the coherence of random networks. We assessed the significance of coherence estimates using a permutation test (see Methods), individually for each PPI database. Permutation p-values obtained using Biogrid and STRING filtered databases were highly correlated (PCC = 0.64, Supplementary Figure S1), as would be expected for high-quality PPIs. Permutation p-values using STRING databases were predominantly significant (p-value < 0.001) due to a large number of PPIs that enabled phenotype-specific networks to be consistently different from random. Permutation p-values inversely correlated with coherence estimates (average PCC = -0.23), as would be expected for less coherent networks that resemble random networks. This dependency was especially pronounced for database-specific estimates, e.g., Biogrid coherence and p-value estimates correlated with PCC = -0.69 (STRING filtered PCC = -0.53, Supplementary Figure S1). These observations confirmed that coherence and p-value estimates remained consistent when using different PPI databases.

The majority of phenotype-specific networks (~85%) were significantly more coherent than random networks (STRING filtered p-value < 0.001, Supplementary Table S3). Among less significant phenotypes were “White blood cell count” (p-value = 0.085) and other blood count-related phenotypes, “Corneal astigmatism” (p-value = 0.04), “Major depressive disorder” (p-value = 0.03). When using Biogrid databases, “Educational attainment”, “Myopia (pathological)”, and “Celiac disease” were among the non-significant phenotypes (p-values 0.30, 0.20, 0.16, respectively). Interestingly, “Schizophrenia” network, despite being among the least coherent, was consistently significantly different from random (p-value < 0.0001). Overall, the significance estimation of coherence supports our observation that phenotype-specific networks in “Intelligence” and “Psychiatric disease” categories generally have low coherence that is frequently statistically insignificant. In summary, these results confirm our ability to identify coherence of phenotype-specific networks and test their significance over random networks.

Discussion

We developed a general-purpose method to measure and compare the coherence estimates of the molecular mechanisms in networks of genes. Intuitively, gene networks with high internal and low external number of interactions (degrees) would be considered highly coherent; consequently, coherence is defined as the slope through the internal and external degree distribution scatterplot. We normalized coherences (slopes) to the range of high and random coherence, exemplified by MSigDb and random networks. To decouple coherence from its dependence on network size, we used MSigDb and random networks matched in size to phenotype-specific networks. Applied to the analysis of 116 disease- and trait-associated gene sets, we found that metabolic diseases and traits had overall high coherence, while psychiatric and cardiovascular diseases generally had lower coherence. These results were consistent when using different PPI databases. Our method allows one to quantify and compare the estimated coherence of molecular mechanisms across phenotypes, and will only improve as more PPI information becomes available.

The coherence measure allows us to gain insights into the genetic component of various phenotypes, as revealed by the current state of the corresponding GWASs. The high coherence of metabolic diseases and traits arguably reflects the strong genetic component driving molecular mechanisms of metabolism-specific gene networks. On the contrary, the low coherence of a phenotype suggests either insufficient knowledge of genetics (as represented in current-generation databases) or the true lack of such a strong genetic influence, (and correspondingly, the heightened importance in these cases of non-genetic, and possibly environmental, factors. If the incoherence results from insufficient knowledge, this should be remedied as novel genes are identified through larger-scale GWASs and meta-analyses. Yet, the low coherence of a phenotype is expected to remain low with increasing gene discovery if, in contrast, the absence of coherence is due to the relative importance of non-genetic causes. Our results highlight the low coherence of the molecular mechanisms of intelligence-related traits and cardiovascular and psychiatric disorders, but our method alone cannot decide whether insufficient knowledge or the weakness of the genetic component ultimately explains the lack of internal connectivity characteristic of low-coherence traits.

Our approach is similar in spirit to the well-researched community detection problem [31, 44, 48, 49]. However, we approach this problem from a different angle by asking whether a given network as a whole can be considered a community within the global PPI network. This may be considered a drawback as it has been shown that many networks consist of miniature communities known as motifs [50], or graphlets [51]. While the investigation of motif enrichment [50] is a viable approach to characterize a network, our goal was to develop a unified measure of coherence that can be compared across phenotype-specific networks.

Our method relies on the ability to link the phenotype-associated genetic variants to genes thought to be affected by them. Typically, genes are linked to genetic variants by proximity (nearest gene), a strategy adopted by the NHGRI-EBI GWAS catalog [36]. However, >88% of phenotype-associated genetic variants are located in regulatory elements outside of protein-coding regions [35]. The pilot project by ENCODE showed that fewer than 10% of all interactions between promoters and regulatory regions involved the closest promoter by linear distance [52], questioning the validity of nearest gene mapping strategy. The development of chromatin conformation capture technologies now allows for understanding long-distance interactions among genomic regions on a genome-wide scale [53, 54], helping to prioritize disease-associated gene-variant associations [55]. Consequently, tools are being developed to consider long-distance interactions when mapping SNPs to genes (3DSNP [56], FUMA [57], PINES [58], HUGIn [59]), and studies redefining disease-associated genes using long-distance interactions started to emerge [60]. Our future goal includes considering long-distance variant-gene interactions in defining phenotype-associated gene sets and comparing their network properties.

The definition of coherence as the slope between internal versus external degree distributions makes our measure dependent on the choice of PPI data [61]. Our limited knowledge of protein-protein interactions limits the coherence estimation to phenotypes that have a sufficient number of genes with PPI annotations, especially for phenotypes with low numbers of associated genes. Furthermore, networks formed by phenotype-associated genes should be sufficiently diverse to avoid cases in which all genes in a network have the same degree. As the different PPI databases vary significantly in the number of PPI annotations, some phenotype-associated gene sets have sufficient PPI annotations only when using one but not the other database. This problem is best illustrated by the BioGrid PPI database that has considerably fewer PPIs; hence, fewer phenotypes could be analyzed. Although our results show good correspondence between coherence estimates when using different PPI databases, care should be exercised when selecting the PPI database for network analyses [61].

In addition to internal versus external degree distributions, other gene-centric network metrics can be used to estimate coherence [48]. For example, the centrality measure has been used to demonstrate that highly central hub genes are functionally essential and highly conserved [62]. A large study of PPIs (BioPlex) compared centrality distributions among several disease categories. Cancers and immunological disease networks tended to have high centrality, while nervous system, congenital, neonatal, and hereditary disease networks had low centrality [63]. Metabolic pathways were shown to have more duplicated copies of highly central genes making the networks tolerant to loss-of-function mutations [62]. These observations of centrality measure parallel our estimation of network coherence. Our future work will incorporate centrality and other gene-centric network metrics into our framework to further refine our ability to compare the coherence of phenotype-associated molecular mechanisms.

Conclusion

Our study investigated the coherence of the molecular mechanisms of genes associated with genomic variants of 116 phenotypes. We developed a network coherence measure driven by the relationship between internal and external connectivity (degree distribution) that is robust to the difference in network sizes. Using well-curated lists of genomic variants and the associated genes, we built disease-specific networks using protein-protein interaction information from the STRING and BioGrid databases. We found a range of coherence estimates in each phenotype category, with metabolic phenotypes being the most coherent. Psychiatric disorders had low coherence, with schizophrenia and major depressive disorder being among the least coherent. In summary, we provide a general-purpose method for quantifying and comparing the coherence of molecular networks.

Methods

Data sources

Phenotype-associated gene lists were obtained from the NHGRI-EBI catalog of genome-wide association studies (GWAS catalog, `gwascat` v.2.14.0 R package) [35, 36]. “MAPPED_GENE” column was used to extract trait-associated genes. To organize traits into categories, Experimental Factor Ontology (EFO) IDs were extracted from the “MAPPED_TRAIT_URI” column. Categories were mapped to EFO IDs using the `ontoCAT` v.1.12.0 R package and the EFO database (accessed 12/04/2018). In the case of a trait being mapped into two categories, the most representative category was manually assigned. Non-canonical gene names were converted to common gene symbols using `alias2Symbol` function from the `limma` R package. Gene names which could not be mapped to gene symbols were excluded from the analysis.

Human protein-protein interactions (PPIs) were obtained from STRING database v.11.0 [33] and BioGrid database v.3.5.174 [34]. The STRING data for the tax ID “9606” (human) was used, and Ensembl protein IDs were converted to gene names using BioMart. Either the complete dataset (11,761,040 entries) or data filtered by “combined_score” > 500 (range 0-1000) (1,373,946 entries, the main PPI dataset) was used. The BioGrid data was filtered to include “TaxID interactor” type equal to “9606”, all “Interaction Types” were included, totaling 465,660 interactions. These three PPI databases were used in parallel to evaluate disease coherence.

Phenotype-associated gene networks were formed by mapping trait-associated genes to PPIs. Genes without PPI annotations (e.g., C5orf4, LOC100507462), antisense transcripts (e.g., CDKN2B-AS), readthrough (e.g., NPHP3-ACAD11), non-protein-coding transcripts (e.g., LINC00478) were excluded. Genes that do not have PPIs with each other but interact with other genes in the global PPI network were kept for external degree calculation. Non-Zero interacting

genes were defined as genes having at least one interaction with other phenotype-specific genes. Consequently, phenotypes with less than 10 non-zero interacting genes were omitted. Additionally, phenotypes with genes having identical internal degrees (hence, no internal degree distribution) were filtered due to the inability to derive a fit through the internal vs. external degree distributions (see below). These precautions were set to avoid situations where a small number of genes in a network can cause instability in the results.

Comparing internal vs. external degree distributions as a measure of coherence

Phenotype-specific **internal degree distribution** was defined as the number of interactions between phenotype-associated genes, or, more precisely, their interacting protein products. Phenotype-specific **external degree distribution** was defined as the number of interactions between genes in a phenotype-associated network and all other genes in the PPI network. The internal vs. external degree distributions were plotted on a scatterplot. The square root of the degree was taken for more informative representation. For each phenotype, the internal vs. external degree distribution plot was fit with a regression line through the origin to capture the slopes. Slopes were used as a measure of coherence. Lower slopes being associated with higher coherence, and vice versa. Degree distributions of the network nodes were obtained using the `igraph` R package v. 1.0.1.

Normalization of coherence

To alleviate the dependency of degree distributions from the size of phenotype-associated gene sets and the content of a selected PPI database, we normalized slopes of phenotype-specific networks to the range of slopes formed by degree distributions of highly coherent and random networks. Specifically, we select 30 MSigDb networks (10 from KEGG, GOCC, and REACTOME collections, β_{min} median slope) and 1,000 random networks (β_{max} median slope) matched in size to the corresponding network of phenotype-specific genes. For each phenotype i with slope β_i, we calculate the normalized slope $\beta_{i\ norm} = \frac{\beta_i - \beta_{max}}{\beta_{min} - \beta_{max}}$, which gives a measure of coherence on the range of [0,1]. Higher values indicate a higher level of coherence similar to that of MSigDb networks. These values represent normalized coherence and can be compared across phenotypes and PPI databases.

Although using median slopes of MSigDb and random networks is expected to be a robust reference for high/random coherence, some phenotype-specific networks may have coherence larger than 1, that is, being more coherent than the median coherence of MSigDb networks. These situations may occur in smaller networks that have a higher chance to contain genes with unusually high internal degree distributions. Such outliers would inflate coherence estimations. Similar outliers may be observed in larger networks. For these reasons, coherence estimates for excessively large networks (“Obesity-related traits”, 779 genes, “Height”, 418 genes) were excluded from calculating the correlation between coherences obtained using different PPI databases (Supplementary Figure S1).

Permutation test of slopes

To test whether each slope is significantly different than random, we used a permutation test. For phenotype i with the gene network of size g_i that contains n_i number of non-zero interactions, we define β_i as the slope derived by regressing the external degree distribution on the internal degree distribution. We then generate networks of randomly selected genes of size g_i to ensure size-matched topology. Each random network will have a relatively few number of non-zero interactions; therefore, we collect genes non-zero interacting genes from the size-

matched random networks until we have $\geq n_i$ of them, and combine them into a single random network. This procedure is repeated until we have $n = 10{,}000$ random networks of size g_i having n_i non-zero interactions. For each random network j, we identify its slope β_r representing the internal vs. external degree distribution regression line of random coherence. We calculate the permutation p-value as $p_i = \frac{(\sum_{j=1}^{n} I(\beta_{rj} > \beta_i) + 1}{n+1}$ [64]. Intuitively, this approach estimates whether the slope of the phenotype-specific network β_i is consistently larger than those of random networks. All analyses and visualizations were performed in the R/Bioconductor computing environment v.3.4.0 [65].

Availability of data and materials

The code and the data supporting the conclusions of this article are available in the https://github.com/dozmorovlab/disease_coherence GitHub repository.

Abbreviations

GWAS - Genome-Wide Association Study; KEGG - Kyoto Encyclopedia of Genes and Genomes; MSigDb - Molecular Signatures Database; PPI - Protein-Protein Interaction; SNP - Single Nucleotide Polymorphism

Acknowledgements

This work was supported by a Genetics and Human Agency Award from the Templeton Foundation to KSK, MR, CC, SB. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Templeton Foundation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MD, KK, CC, and MR conceived the study. KC and MD performed the experiments. MD primarily wrote the manuscript. KK, MR, CC, and SB participated in manuscript writing. All authors read and approved the final manuscript.

Supplementary material

Supplementary Table S1. Summary statistics of the analyzed phenotype categories. Minimum, mean, median, maximum of coherence estimates, the number of genes, SNPs, the total number of phenotype networks. Each worksheet corresponds to the PPI database used.

Supplementary Table S2. Summary of networks from MSigDB categories. Minimum, mean, median, maximum of the slopes, the number of genes, and the total number of networks are shown for each collection, along with the correlation between network size and slope. "CP:/KEGG/Reactome/Biocarta" - canonical pathways, "GOBP/GOMF/GOCC" - gene ontology biological processes, molecular functions, cellular component, "TFT" - transcription factor targets, "CGN" - cancer gene neighborhoods, "CM" - cancer modules, "MIR" - microRNA targets, "CGP" - chemical and genetic perturbations.

Supplementary Table S3. Phenotype-specific coherence estimates. The PPI database (STRING, STRING filtered, Biogrid) are specified in the corresponding column names. For each phenotype category, the results are sorted by the "String Filt Coherence" column in descending

order. “NA” indicates the corresponding value cannot be estimated due to lack of sufficient number of genes annotated with PPIs.

Supplementary Figure S1. Normalization of coherence estimates alleviates its network size dependence. Red/blue gradient and numbers represent Pearson correlation coefficients for each pairwise comparison of the number of SNPs, genes, and the normalized coherence estimates using the corresponding PPI databases.

Supplementary Figure S2. Coherence estimates of traits using Biogrid as a reference PPI database. Size of dots represents the level of normalized coherence (X-axis) for individual traits (Y-axis). Plots are faceted by categories. Missing entries indicate that, for a given trait, a network could not be built and the coherence cannot be estimated.

Supplementary Figure S3. Coherence estimates of traits using STRING as a reference of PPI database. See legend for Supplementary Figure S2.

Supplementary Figure S4. Coherence estimates of diseases using Biogrid as a reference PPI database. See legend for Supplementary Figure S2.

Supplementary Figure S5. Coherence estimates of diseases using STRING as a reference PPI database. See legend for Supplementary Figure S2.

References

1. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008;322:881–8.

2. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516–7.

3. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50.

4. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456:18–21.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

6. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228–37.

7. Turkheimer E. Genome wide association studies of behavior are social science. In: Philosophy of behavioral biology. Springer; 2012. pp. 43–64.

8. Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Børglum AD, Breen G, et al. Psychiatric genomics: An update and an agenda. Am J Psychiatry. 2018;175:15–27.

9. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402 6761 Suppl:C47–52.

10. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in alzheimer’s disease. Proc Natl Acad Sci U S A. 2004;101:15148–53.

11. Huang R, Wallqvist A, Covell DG. Comprehensive analysis of pathway or functionally related gene expression in the national cancer institute's anticancer screen. Genomics. 2006;87:315–28.

12. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007;25:309–16.

13. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392–404.

14. Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet. 2009;17:1231–40.

15. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A. 2008;105:4323–8.

16. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347:1257601.

17. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci U S A. 2007;104:8685–90.

18. Akula N, Baranova A, Seto D, Solka J, Nalls MA, Singleton A, et al. A network-based approach to prioritize results from genome-wide association studies. PLoS One. 2011;6:e24220.

19. Ghiassian SD, Menche J, Barabási A-L. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11:e1004120.

20. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A. 2008;105:9880–5.

21. Barabási A-L, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

22. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. Cell. 2011;144:986–98.

23. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.

24. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Brief Funct Genomics. 2011;10:280–93.

25. Farber CR. Systems-level analysis of genome-wide association data. G3 (Bethesda). 2013;3:119–29.

26. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91:243–8.

27. Gillis J, Pavlidis P. "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol. 2012;8:e1002444.

28. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21:1109–21.

29. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.

30. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res. 2008;18:1150–62.

31. Fortunato S. Community detection in graphs. Physics reports. 2010;486:75–174.

32. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

33. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43 Database issue:D447–52.

34. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The biogrid interaction database: 2017 update. Nucleic Acids Res. 2017;45:D369–79.

35. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.

36. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic Acids Res. 2017;45:D896–901.

37. Gonçalves JP, Francisco AP, Moreau Y, Madeira SC. Interactogeneous: Disease gene prioritization using heterogeneous networks and full topology scores. PLoS One. 2012;7:e49634.

38. Wang Q, Liu W, Ning S, Ye J, Huang T, Li Y, et al. Community of protein complexes impacts disease association. Eur J Hum Genet. 2012;20:1162–7.

39. Zhang S, Zhang S-H, Wu C, Li X, Chen X, Jiang W, et al. From phenotype to gene: Detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. FEBS Lett. 2010;584:3635–43.

40. Driel MA van, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14:535–42.

41. Hamaneh MB, Yu Y-K. DeCoaD: Determining correlations among diseases using protein interaction networks. BMC Res Notes. 2015;8:226.

42. Jia P, Zheng S, Long J, Zheng W, Zhao Z. DmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 2011;27:95–102.

43. Dozmorov MG. Disease classification: From phenotypic similarity to integrative genomics and beyond. Brief Bioinform. 2018.

44. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc Natl Acad Sci U S A. 2004;101:2658–63.

45. Lancichinetti A, Radicchi F, Ramasco JJ. Statistical significance of communities in networks. Phys Rev E. 2010;81:046110. doi:10.1103/PhysRevE.81.046110.

46. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences. 100:12123. doi:10.1073/pnas.2032324100.

47. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

48. Leskovec J, Lang KJ, Mahoney M. Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th international conference on world wide web. New York, NY, USA: ACM; 2010. pp. 631–40. doi:10.1145/1772690.1772755.

49. Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment. 2005;2005:P09008.

50. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. Science. 2002;298:824–7.

51. Przulj N. Biological network comparison using graphlet degree distribution. Bioinformatics. 2007;23:e177–83.

52. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.

53. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.

54. Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013;504:306–10.

55. Pan DZ, Garske KM, Alvarez M, Bhagat YV, Boocock J, Nikkola E, et al. Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from gwas. Nat Commun. 2018;9:1512.

56. Lu Y, Quan C, Chen H, Bo X, Zhang C. 3DSNP: A database for linking human noncoding snps to their three-dimensional interacting genes. Nucleic Acids Res. 2017;45:D643–9.

57. Watanabe K, Taskesen E, Bochoven A van, Posthuma D. Functional mapping and annotation of genetic associations with fuma. Nat Commun. 2017;8:1826.

58. Bodea CA, Mitchell AA, Day-Williams AG, Runz H, Sunyaev SR. Phenotype-specific information improves prediction of functional impact for noncoding variants. 2016. doi:10.1101/083642.

59. Martin JS, Xu Z, Reiner AP, Mohlke KL, Sullivan P, Ren B, et al. HUGIn: Hi-c unifying genomic interrogator. Bioinformatics. 2017;33:3793–5.

60. Fang H, ULTRA-DD Consortium, De Wolf H, Knezevic B, Burnham KL, Osgood J, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. Nat Genet. 2019;51:1082–91.

61. Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, et al. How helpful are the protein-protein interaction databases and which ones? bioRxiv.:566372. doi:10.1101/566372.

62. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013;9:e1002886.

63. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545:505–9.

64. Phipson B, Smyth GK. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9:Article39.

65. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S. Bioinformatics and computational biology solutions using r and bioconductor. Springer Science & Business Media; 2006.