




ARTICLE OPEN



Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery

Paola Paci¹, Giulia Ficon^{2,3}, Federica Conte², Rui-Sheng Wang⁴, Lorenzo Farina¹ and Joseph Loscalzo¹

In this study, we integrate the outcomes of co-expression network analysis with the human interactome network to predict novel putative disease genes and modules. We first apply the SWitch Miner (SWIM) methodology, which predicts important (switch) genes within the co-expression network that regulate disease state transitions, then map them to the human protein–protein interaction network (PPI, or interactome) to predict novel disease–disease relationships (i.e., a SWIM-informed diseasome). Although the relevance of switch genes to an observed phenotype has been recently assessed, their performance at the system or network level constitutes a new, potentially fascinating territory yet to be explored. Quantifying the interplay between switch genes and human diseases in the interactome network, we found that switch genes associated with specific disorders are closer to each other than to other nodes in the network, and tend to form localized connected subnetworks. These subnetworks overlap between similar diseases and are situated in different neighborhoods for pathologically distinct phenotypes, consistent with the well-known topological proximity property of disease genes. These findings allow us to demonstrate how SWIM-based correlation network analysis can serve as a useful tool for efficient screening of potentially new disease gene associations. When integrated with an interactome-based network analysis, it not only identifies novel candidate disease genes, but also may offer testable hypotheses by which to elucidate the molecular underpinnings of human disease and reveal commonalities between seemingly unrelated diseases.

npj Systems Biology and Applications (2021)7:3; <https://doi.org/10.1038/s41540-020-00168-0>

INTRODUCTION

The new emerging paradigm of *network medicine* has been dramatically changing the way we define and analyze human diseases. Rather than view a disease as an independent entity, the network medicine approach recognizes the interplay of multiple molecular processes in expressing the pathophenotype^{1–3}. By proposing a holistic approach according to which characterizing the behavior of the network as a whole is essential for understanding disease complexity⁴, network medicine sets the stage for exploring disease complexity at the cellular and molecular levels, and for studying the relationships among apparently different pathophenotypes. A key goal of network medicine is to gain a better understanding of how perturbations propagate through the system by identifying and characterizing potential network modules that can be targeted for clinical intervention. Although introduced relatively recently, scientific research in the network medicine field has been growing rapidly as witnessed by the number of evolving methods designed to interrogate disease etiology, model molecular and genetic interactions, identify potential biomarkers, and design therapeutic interventions, including both drug discovery and drug repurposing^{1,5–18}.

The two key issues that each network-based algorithm has to address are the identification of the critical entities in the system under investigation (nodes), and the nature of the interactions between these entities (edges). This information depends on the study design, the phenotype under investigation, the biological

question of interest, the molecular entities measured, and the type and size of the available datasets. Thus, tools developed within the field of network medicine are highly customized to leverage these biomedical data with respect to the given biological or disease context. Several of these algorithms^{5–7} make use of the human protein–protein interaction (PPI) network, also denoted the human interactome, which is a network of proteins (nodes) in which the edges are the physical and functional interactions occurring between them. Despite the fundamental insights PPI networks provide about the topological features of specific protein interactions within them, their intrinsic immutable nature renders them void of context-specific information (i.e., cell-, tissue-, or disease-specificity). Moreover, the incompleteness of the current interactome and the partial knowledge of the number of genes associated with various diseases make mining disease-specific interactions via the PPI network a very demanding task. To overcome all of these shortcomings, novel *in silico* strategies are necessary to overlay the interactome with this additional, important biomedical information.

Gene expression networks (GENs) are context-specific by definition, as they directly leverage phenotype-specific gene expression data in network construction by calculating correlations between the expression profiles of each gene pair. The basic premise of this exercise is that, even though correlation is not causation, co-expressed genes are functionally coordinated in response to an external stimulus, implying that they may be part of the same complexes or pathways, and may influence each other or may be influenced by the same underlying mechanism(s).

¹Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy. ²Institute for Systems Analysis and Computer Science “Antonio Ruberti”, National Research Council, Rome, Italy. ³Fondazione per la Medicina Personalizzata, Via Goffredo Mameli, 3/1 Genova, Italy. ⁴Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, USA. ✉email: paci@diag.uniroma1.it

For example, SWItch Miner (SWIM)¹⁹ is a new, promising methodology that considers differentially expressed genes within the co-expression network framework in order to predict important genes affected by a disease of interest, and combines this information with a structured network of correlated patterns. Considering the topological properties of the nodes and assessing their functional roles according to their ability to convey information within and between modules in the network, SWIM identifies a small pool of genes (known as *switch genes*) that are associated with intriguing patterns of molecular co-abundance and play a crucial role in the observed phenotype (transitions). The phenotype-specific applications of SWIM are broad and include the identification of switch genes in both complex diseases and cancers^{19–23}, and as well as in grapevine berry maturation (*Vitis vinifera*)²⁴.

In cancer research, SWIM network analysis has been gainfully applied to a large panel of TCGA (The Cancer Genome Atlas) cancer datasets²⁵ in order to characterize disease etiologies and identify potential therapeutic targets¹⁹. SWIM has also been used to investigate glioblastoma multiforme and to uncover new insights into the molecular mechanism determining the stem-like phenotype of glioblastoma cells²⁰. [Stem-like cells determine tumor aggressiveness by sustaining tumor growth and causing relapse and metastasis by their resistance to conventional cancer therapies²⁶.] In particular, the role of FOSL1 was explored and found to be a repressor of a core of four master neurodevelopmental transcription factors whose induction is sufficient to reprogram fully differentiated glioblastoma cells into stem-like cells²⁷. This result could have a significant impact on personalized healthcare, since promoting differentiation and restraining tumor growth may support rational, personalized planning of disease prevention or treatment.

Recently, SWIM methodology has been successfully applied to chronic obstructive pulmonary disease (COPD)²², a severe lung disease characterized by progressive and incompletely reversible airflow obstruction. COPD is a heterogeneous and complex syndrome influenced by both genetic and environmental determinants, and is one of the main causes of morbidity and mortality worldwide. COPD switch genes appear to form localized connected subnetworks displaying an intriguingly common pattern of upregulation in COPD cases compared with controls. A more sophisticated analysis revealed that they were not only topologically related, but also functionally relevant to the observed phenotype as witnessed by their enrichment in the regulation of inflammatory and immune responses. The results obtained in COPD were compared with those obtained in the acute respiratory distress syndrome (ARDS), another severe lung disease with an inflammatory component. Interestingly, ARDS switch genes were different from COPD switch genes, but the major pathways affected in the two diseases were similar, emphasizing that different diseases often have common underlying mechanisms and share intermediate endophenotypes (convergent phenotypes)^{6,28}. Moreover, the two lists of switch genes, when mapped to the human interactome, appear to form non-overlapping modules and to be situated in different network neighborhoods. This observation demonstrates that even though different diseases can share similar endophenotypes, the molecular network determinants responsible for them are disease-specific. This observation is also fully consistent with the fundamental principles of network medicine, where disease proteins are assumed not to be randomly scattered, but agglomerate in specific regions of the molecular interactome, suggesting the existence of specific disease network modules for each disease.

Inspired by the results obtained by SWIM network analysis of cancers and COPD, here we investigated three other complex diseases for a more generalizable understanding of the highly interconnected nature of human diseases. Specifically, two cardiac

disorders, ischemic and non-ischemic cardiomyopathy, and one neurodegenerative disorder, Alzheimer's disease (AD), were analyzed. These new results, together with the previously obtained analyses from the application of SWIM to ten tumor types and COPD, were mapped to the human interactome in order to overlay the PPI network with disease information derived from SWIM-based disease correlation networks.

Our goal is to assess the utility of SWIM network analysis in classifying several different disorders and in understanding their complex interconnections in the human interactome. In particular, through the construction of a SWIM-informed human disease network (SHDN) by analogy with ref. ⁵, we test whether or not switch genes of a specific disease tend to localize in a critical module in the interactome that is functionally relevant to the observed phenotype.

RESULTS

Workflow of the analysis

In this study, we combined the topological properties of the human interactome with disease information derived from SWIM-based correlation network analysis. The baseline networks of our analysis are SWIM-based GENs and the outcome network is an SHDN, by analogy with a previous study⁵. The workflow of our study design is depicted in Fig. 1.

Identification of disease-specific switch genes

The SWIM algorithm was applied to a specific group of diseases of interest to build disease-specific GENs (Supplementary Data 1) and extract a list of switch genes for each disease through an accurate topological analysis (Supplementary Table 2). The analyzed human diseases were:

- (i) ten tumor types (i.e., BLCA, BRCA, CHOL, COAD, HNSC, KIRP, LUAD, LUSC, PRAD, and UCEC) available from TCGA, whose corresponding lists of switch genes were retrieved from our previous study¹¹;
- (ii) one pulmonary disease (COPD), whose corresponding list of switch genes was retrieved from our previous study¹⁴;
- (iii) ischemic cardiomyopathy (IC), whose list of switch genes was obtained by applying SWIM correlation network analysis to RNA-sequencing data from ischemic human failing versus non-failing control hearts; and
- (iv) non-ischemic cardiomyopathy (NIC), whose list of switch genes was obtained by applying SWIM correlation network analysis to RNA-sequencing data from non-ischemic human failing versus non-failing control hearts;
- (v) AD, whose list of switch genes was obtained by applying SWIM correlation network analysis to microarray expression data related to AD patients versus controls.

Identification of SWIM-informed disease modules

Actually, members (nodes, proteins, or genes) of a network module are more functionally and topologically related to each other than to other nodes in the network. Thus, the lists of switch genes for the 14 analyzed diseases were mapped onto the human interactome to investigate whether or not they tend to agglomerate in local neighborhoods and constitute statistically significant disease-specific modules. To do so, for each disease, the corresponding switch genes subnetwork was extracted and the following three metrics were computed: (i) the total number of interactions (edges); (ii) the size of the largest connected component (LCC); and (iii) the number of edges in the LCC. In order to complement these metrics with a measure of statistical significance, we computed *module significance*, which measures the probability that a given list of switch genes is localized within

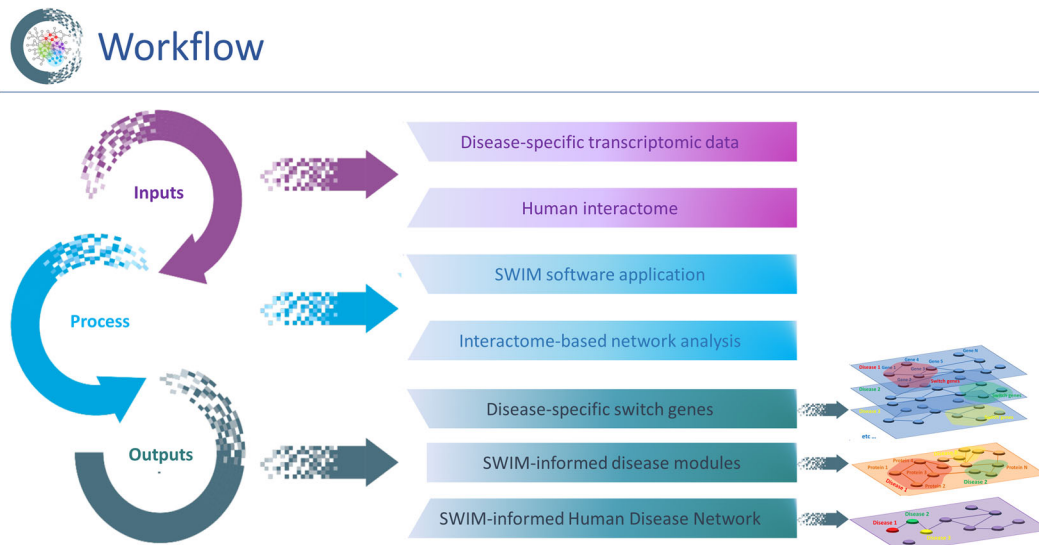


Fig. 1 Workflow of the analysis. Our analysis takes as input the transcriptomic data of the 14 diseases of interest and the human interactome network. The analysis begins with the application of the SWIM algorithm to construct the disease-specific Gene Expression Networks (GENs), and, hence, to identify the disease-specific switch genes. Next, by mapping the disease-specific switch genes to the human interactome, the analysis next proceeds with an interactome-based network analysis, which reveals the SWIM-informed disease modules and the SWIM-informed Human Disease Network (SHDN), in which nodes are now diseases, and a link occurs if the corresponding switch gene modules were found to overlap.

a certain network neighborhood more than expected by chance. For each disease, we randomly selected groups of proteins of the same size and degree distribution as the original list of switch genes in the human interactome. We then extracted the corresponding subnetwork and we computed: (i) the total number of its interactions; (ii) the size of the LCC; and (iii) the number of edges in the LCC. This procedure was repeated 1000 times. [As reported in the majority of state-of-the-art approaches^{29–37}, 1000 permutations is commonly used for estimating the power of a randomization test, showing it that can be considered as a reasonable number of permutations for a test at the 5% level of significance.] Finally, we derived three distributions for all three metrics corresponding to the subgraph induced by the random gene set. The three metrics calculated for the original list of switch genes were z-score-normalized with respect to the corresponding reference random distribution. Subsequently, the p -value for the given z statistic was calculated. We found that all of the analyzed sets of switch genes form statistically significant modules (i.e., all three metrics were statistically significant) in the human interactome that are disease-specific (Fig. 2 and Supplementary Data 3).

Overlap estimation of SWIM-informed disease modules

In order to evaluate the extent to which two disease-specific modules (A, B) of switch genes are in the immediate vicinity of each other in the human interactome, we leveraged the *module separation* parameter defined in Eq. (1) (cf. “Methods”) that measures the separation or overlap of two modules⁶, and we applied a degree-preserving randomization procedure to assess the statistical significance of each separation value. By analyzing the topological structure of the identified SWIM-informed disease modules, we found that three topologically different situations came to light:

1. two given modules overlap more than expected by chance (i.e., $s < 0$ and p -value < 0.05), hereafter denoted *cognate modules*;
2. two given modules separate more than expected by chance (i.e., $s > 0$ and p -value < 0.05), hereafter denoted *non-cognate modules*; and

3. there is insufficient evidence to support the hypothesis that two given modules overlap or separate more than expected by chance (i.e., p -value > 0.05), hereafter denoted *modules of uncertain overlap*.

In particular, we observed that diseases displaying a pathobiological similarity (such as cancers or cardiomyopathies) shared a substantial number of switch genes reflected by overlapping disease modules (cognate modules), whereas diseases characterized by different pathological phenotypes (such as inflammatory lung diseases and AD) showed specific switch genes reflected in non-overlapping disease modules (non-cognate modules). These two situations are presented in Fig. 3a, where the projection of disease-specific switch gene products (disease-specific switch proteins) on the PPI network, denoted the Disease Switch Gene Network (DSGN), is represented. In the DSGN, disease-specific switch proteins with their corresponding interactions are colored based on the disorder class to which they belong, and all cognate modules (sharing a substantial number of disease-specific switch proteins) are represented as one coalesced module colored with a less intense color corresponding to the disease class. It is worth noting that the specificity of the DSGN is twofold: it is constructed starting from genes (1) that are predicted to have a key role in transcriptional rewiring (co-expression analysis) for a tissue-specific experiment (the “specific” side of the DSGN), and (2) that are related by interactions on the PPI network that are identified using various techniques under different specific experimental and biological conditions (the “universal” part of the DSGN).

Functional enrichment analysis of overlapping SWIM-informed disease modules

DSGN provides an intuitive visualization of the phenotypic relatedness among diseases, clearly showing how, for example, the broad tumor disease module, encompassing several overlapping cancer-related modules, is well-separated from the COPD module, as well as from the cardiomyopathies disease module and the AD module (Fig. 3a). To provide a biological interpretation of these findings, we extracted the overlapping switch genes within the disease modules related to tumor and cardiomyopathies classes, and we performed a functional enrichment analysis by

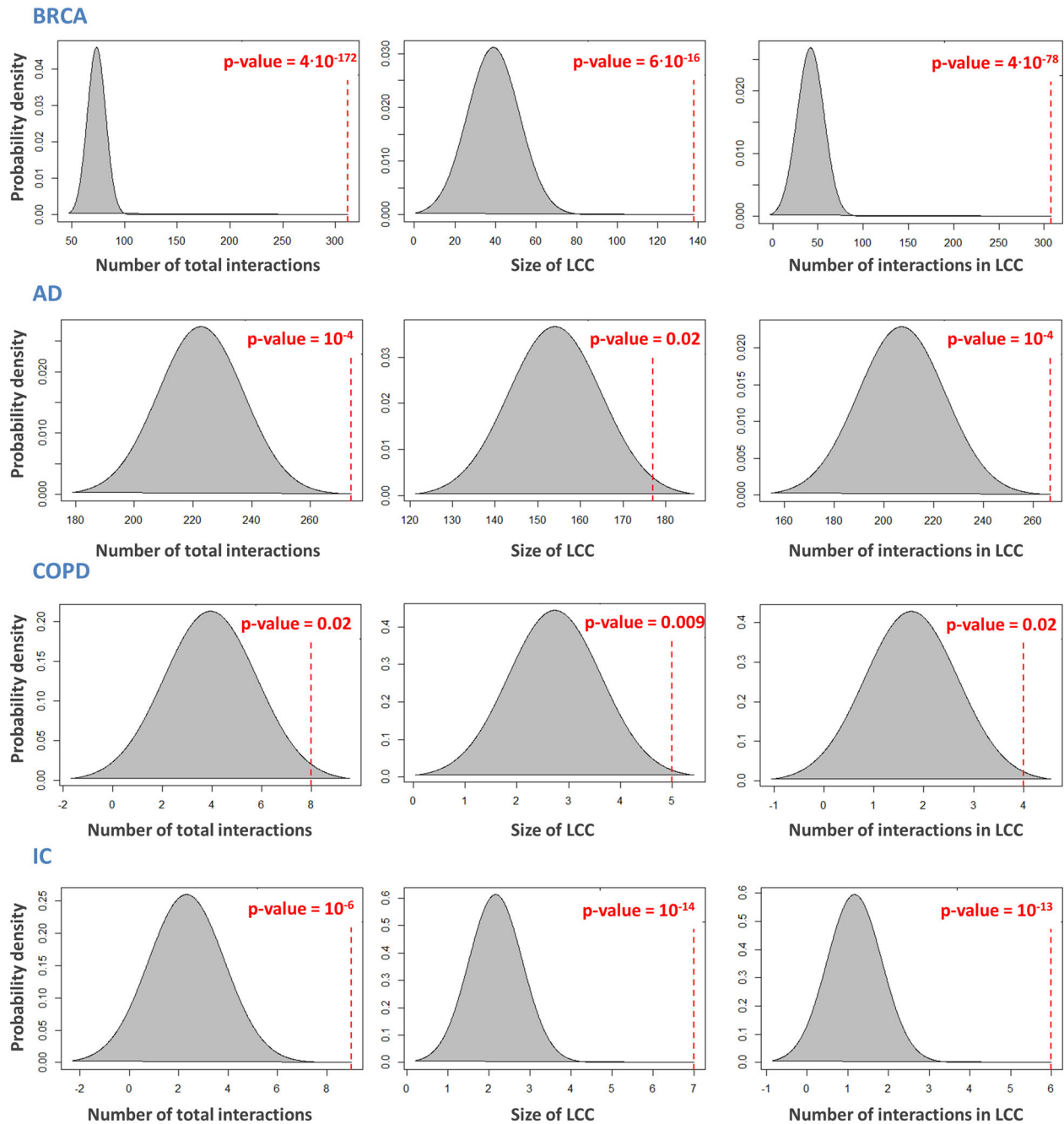


Fig. 2 Examples of SWIM-informed disease module. From left to right: Distribution of the number of total interactions, the size of the largest connected component (LCC), and the number of edges in the LCC in the subgraph induced by a randomly selected gene set of the same size and degree distribution as the original disease list of the switch genes in the human interactome. Dashed red lines correspond to the observed values of each metric computed for the list of switch genes mapped to the interactome. All p -values were calculated by using a one-tailed z test.

querying both KEGG pathways³⁸ and Gene Ontology (GO)³⁹ databases.

Among the tumor class, we found a prevalent set of 26 switch genes recurring across multiple tumors that were all over-expressed in tumor tissues (Supplementary Data 2) and appeared primarily involved in the regulation of cell cycle, which is a fundamental and tightly controlled process under physiological circumstances. Specifically, these tumor-recurring switch genes appeared functionally enriched (adjusted p -value < 0.05) in the cell cycle and progesterone-mediated oocyte maturation KEGG pathways, and include cyclin A2 (CCNA2), cyclin B2 (CCNB2), and

polo-like kinase 1 (PLK1); as well as in the G2/M phase transition and the mitotic spindle checkpoint GO biological processes, including the forkhead transcription factor (FOXM1), MYB proto-oncogene like 2 (MYBL2), the NIMA related kinase 2 (NEK2), the BUB1 mitotic checkpoint serine/threonine kinase B (BUB1B), Aurora B kinase (AURKB), centromere protein F (CENPF), and the dual specificity protein kinase (TTK). Moreover, by evaluating the enrichment of known binding motifs in their promoter regions, this set of 26 tumor-recurring switch genes appeared to be putatively co-regulated by the nuclear transcription factor Y (NF-Y) family (NF-YA, NF-YB) and the E2F transcription factor family

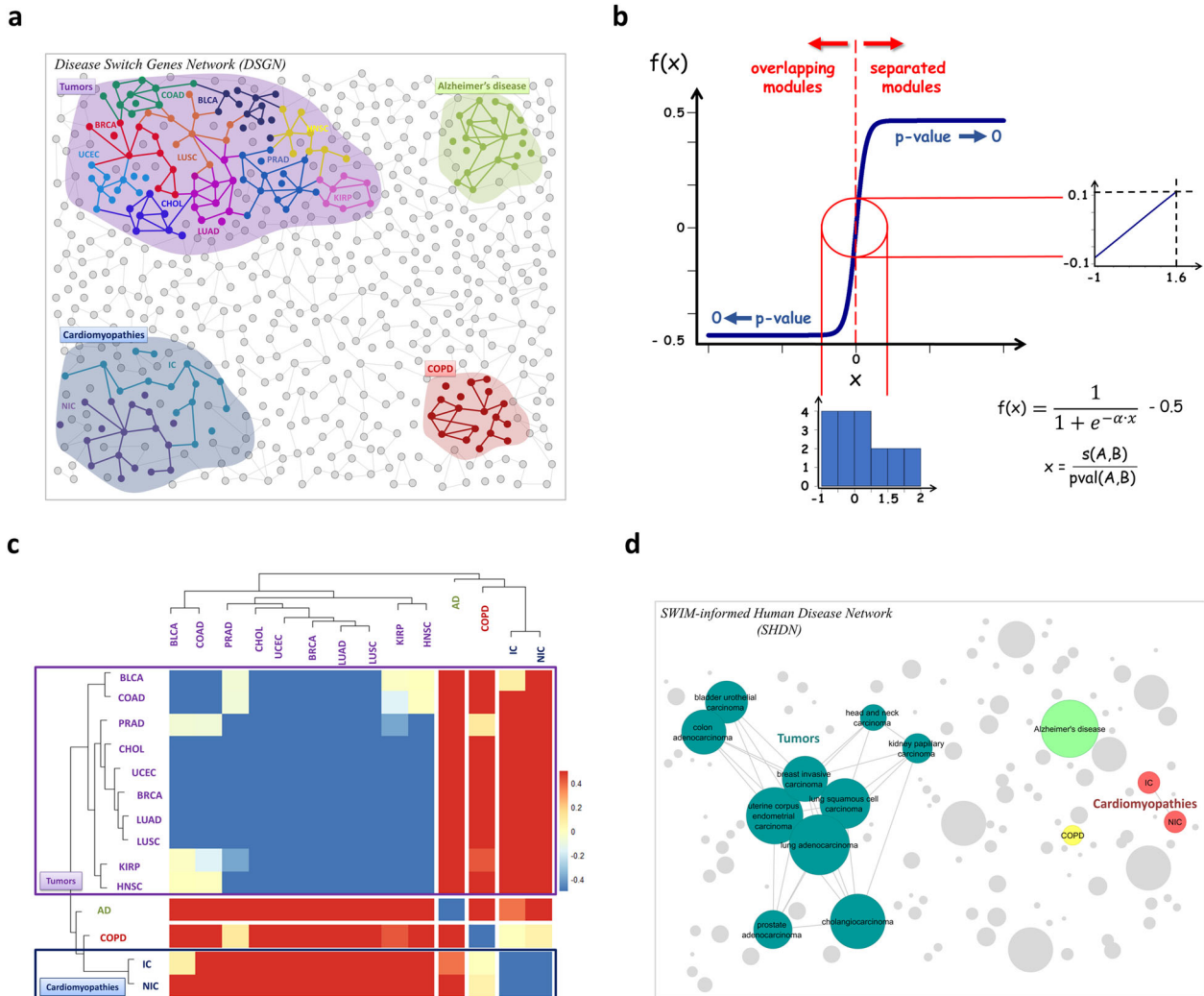


Fig. 3 Results of the analysis. **a** Disease Switch Gene Network (DSGN). Schematic representation of disease modules informed by switch genes in the human interactome. Switch gene products (switch proteins) were colored based on the disorder class to which they belong. Gray nodes and the corresponding links are theoretical and represent non-switch proteins hypothetically connected to switch proteins within the interactome. **b** Generalized measure of the module separation. The function $f(x)$ is the generalized version of module separation defined in Eq. (3) (cf "Methods"). This function approaches its maximum value when the disease modules are significantly well-separated (p -value < 0.05), whereas it approaches its minimum value when the disease modules significantly overlap (p -value < 0.05). The value of $\alpha = 0.3$ was chosen to have $f(x)$ close to zero for statistically insignificant p -values (i.e., $f(x)$ in $[-0.1, 0.1]$ as highlighted by the red circle). The blue bars represent the frequency of x values, ranging from -1 and 1.6 , Supplementary Data for statistically insignificant p -values, ranging from 0.09 and 1 (Supplementary Data 4). **c** SWIM-based disease dendrogram and symmetrical heatmap. The diseases modules identified by the disease-specific switch proteins in the human interactome are clustered by a complete linkage hierarchical clustering algorithm and by using the separation metric as a distance metric. Heatmap colors refer to the generalized separation metric, increasing from blue to red: shades of blue refer to *cognate disease modules* (i.e., $s < 0$, p -value < 0.05); shades of red refer to *non-cognate disease modules* (i.e., $s > 0$, p -value < 0.05); and shades of yellow refer to *uncertain disease modules* (i.e., p -value > 0.05). **d** SWIM-informed Human Disease Network (SHDN). In the SHDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs. Labeled nodes correspond to the 14 diseases analyzed in this study, while unlabeled nodes are artificial and represent other diseases or developmental endotypes to be investigated. The size of each node is proportional to the number of switch genes involved in the corresponding disorder. A link between two diseases occurs if they share a substantial number of switch genes. AD Alzheimer's disease, BLCA bladder urothelial carcinoma, BRCA invasive breast carcinoma, CHOL cholangiocarcinoma, COAD colon adenocarcinoma, COPD chronic obstructive pulmonary disease, HNSC head and neck squamous cell carcinoma, IC ischemic cardiomyopathies, KIRP kidney renal papillary cell carcinoma, LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, NIC non-ischemic cardiomyopathies, PRAD prostate adenocarcinoma, UCEC uterine corpus endometrial carcinoma.

(E2F4/E2F6), known to participate in the regulation of progression through the cell cycle.

By contrast, we found a set of 29 switch genes shared between the two cardiomyopathies that were all downregulated in the disease (Supplementary Data 2) and appeared functionally enriched (adjusted p -value < 0.05) in the cardiac muscle contraction KEGG pathway, including cardiac-type troponin T2 (TNNT2), myosin light chain 3 (MYL3), the subunits 5A and 7B of the cytochrome c oxidase (COX5A and COX7B),

the ubiquinol-cytochrome c reductase core protein 1 (UQCRC1), and cytochrome c1 (CYC1).

SWIM-based estimation of disease relationships

In order to distinguish better among the three topologically different situations (cognate, non-cognate, and modules of uncertain overlap), we combined the module separation defined in Eq. (1) and its statistical significance (p -value) into a *generalized*

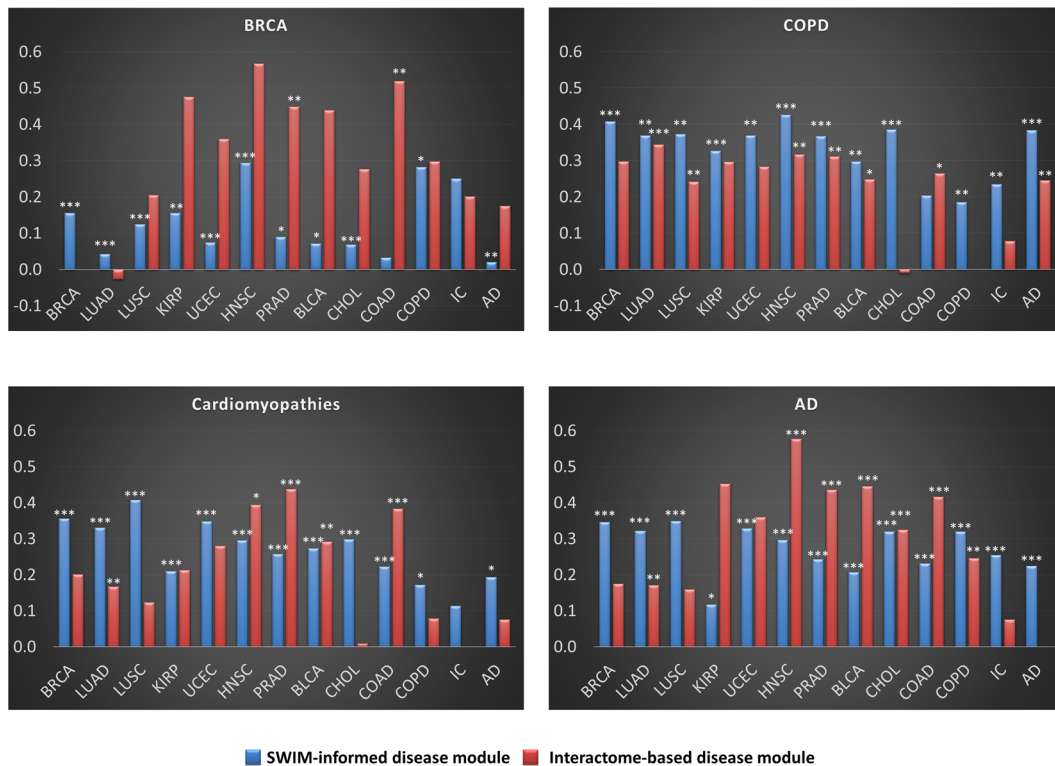


Fig. 4 Comparison between SWIM-informed and interactome-based disease modules. The plots show the values of the non-Euclidean separation distance (Eq. 1, cf. “Methods”) computed between the interactome-based disease module corresponding to the disease reported in the title and SWIM-informed (blue bars) or interactome-based (red bars) disease modules, for all the diseases included in this study. We applied a degree-preserving randomization procedure to assess the statistical significance of each separation value, and we calculated all p -values by applying a two-tailed z test. The stars flag levels of significance for three of the most commonly used levels: p -value < 0.05 is flagged with one star (*); p -value < 0.01 is flagged with two stars (**); and p -value < 0.001 is flagged with three stars (***)

measure of module separation defined in Eq. (3). This generalized version \tilde{s} of module separation (Fig. 3b) was used to elucidate better the relationships among diseases. In fact, we performed a hierarchical clustering of the 14 analyzed diseases using the module separation measure as a distance metric, with the generalized \tilde{s} coded by color scale as illustrated in the associated heatmap (Fig. 3c and Supplementary Data 4). Shades of yellow refer to poorly informative interactions between diseases, corresponding to modules of uncertain overlap in the interactome; shades of blue associate with strictly linked diseases, corresponding to overlapping modules in the interactome (cognate modules); and shades of red quantify the distance between diseases whose corresponding modules are distant from each other in the interactome (non-cognate modules). We found two main clusters: one including all tumor datasets (violet in Fig. 3c), and one including the two cardiomyopathies (dark blue in Fig. 3c) along with AD and COPD datasets as isolated branches, showing a direct relation between the pathobiological similarity of diseases and their relative distance in the human interactome.

Building the SHDN

Using the generalized separation measure, the SHDN was built, where each node corresponds to a distinct disorder and the occurrence of a link between two disorders depends on the extent to which their corresponding modules are in the immediate vicinity of each other (Fig. 3d). The underlying hypothesis is that disease modules closer to each other than to other network components are more likely to share common switch genes and etiology mechanisms.

To correct the problem of not having a fully connected network, we define a threshold on the values the generalized separation

measure can assume that directly reflects the existence or not of a link between two nodes. Thus, given a module, the generalized separation measure with another module must be less than the 75th percentile of the distribution of the negative values of the generalized separation measure between the given module and all others in order to produce a link between the two corresponding diseases. Nodes in this network are colored based on the disorder class to which they belong.

Comparison between SWIM-informed and interactome-based disease modules

In order to analyze how the SWIM-informed disease modules (i.e., nodes in SHDN) are related to the interactome-based disease modules derived from disease genes in databases, we computed the non-Euclidean separation distance (Eq. 1) between the SWIM-informed disease modules and the interactome-based disease modules for each disease included in this study. We first retrieved the lists of disease genes from the DisGeNET database⁴⁰, then built interactome-based disease modules and compared their distance from the modules in SHDN. We observed that SWIM-informed disease modules do not overlap with the interactome-based disease modules for the same disorder class (Fig. 4 and Supplementary Data 5), whereas they do among themselves. It is worth noting that although SWIM-informed disease modules do not overlap with interactome-based disease modules for the diseases analyzed, they do have lower separation values than those observed among interactome-based disease modules themselves. The lack of overlap between SWIM modules and interactome-based disease modules may reflect the partial knowledge of the number of genes associated with various diseases as well as the current incompleteness of the human

interactome. An alternative explanation may lie in the differences in scale and influence across the interactome. The SWIM modules, by virtue of their basis in correlation among switch genes whose expression could be regulated by common transcription factors, reflect actions that may also occur “at-a-distance” in the interactome (i.e., long-range interactions), likely reflecting concomitant modulation of functionally distinct and separate submodules that inform phenotype. By contrast, the interactome-based disease modules are, instead, strictly defined on the basis of physical interactions among disease proteins in proximity to one another a local neighborhood of the PPI (i.e., short-range interactions), without reflecting the longer-range influences of hierarchical regulatory features. With regard to diseases characterized by different pathological phenotypes (such as COPD vs Alzheimer’s or Cardiomyopathies), the separations between SWIM modules and interactome-based disease modules or among interactome-based disease modules are similar.

DISCUSSION

The present study allowed us to demonstrate the relevance of switch genes in the context of network medicine and, in particular, their relation to the definition of disease genes. As broadly established^{1,2}, disease genes have unique, quantifiable characteristics that distinguish them from other genes. From a network perspective, this observation translates into the verification that disease genes do not map randomly to the interactome, but, rather, manifest detectable correlations between their location and their network topology. This observation has led to a series of widely used hypotheses and organizing principles that tie the interactome to human diseases. These are summarized as follows: (i) the *local hypothesis*, according to which proteins involved in the same disease have an increased tendency to interact with each other; (ii) the *disease module hypothesis*, according to which proteins involved in the same disease show a tendency to cluster in connected subnetworks (or connected components), within which one of them is often much larger than the others (LCC); (iii) the *functional coherence hypothesis*, according to which genes in a disease module are often involved in the same biological process (es); and (iv) the *shared components hypothesis*, according to which related diseases are located in the same interactome neighborhood from which unrelated diseases are separated. We will next discuss how the results presented in this study support the validity and applications of those hypotheses with respect to switch genes.

Local hypothesis. Our work over the last decade demonstrated that switch genes appear consistently co-expressed in each disease studied thus far, showing coherent patterns of correlation that could presuppose possible co-regulation. Here, we have systematically demonstrated that these co-expression patterns turned into disease-specific subgraphs when mapped to the PPI network, whose nodes show a higher tendency to interact with each other more frequently than expected by chance (Fig. 2 and Supplementary Data 3). This observation confirms a fundamental hypothesis of interactome-based approaches to human disease, the *local hypothesis*, that genes associated with the same disease are not scattered randomly in the interactome, but aggregate in local, disease-specific neighborhoods.

Disease module hypothesis. By exploring the structural and topological properties of these disease-specific neighborhoods, we observed that they were composed of a dominant connected component, viz., to the LCC, whose size is significantly greater than expected by chance (Fig. 2 and Supplementary Data 3). This dominant component constitutes a highly connected and locally dense subgraph of the interactome, as witnessed by the number of its interactions, which are greater than expected by chance (Fig. 2 and Supplementary Data 3). We conclude that the LCC of each disease-specific subnetwork, built starting from switch genes,

corresponds to their specific disease modules, fulfilling the *disease module hypothesis*.

Functional coherence hypothesis. We next extracted the switch genes within the disease modules, and, performed a functional enrichment analysis by querying both KEGG pathways³⁸ and GO³⁹ databases. Among the tumor class, we found a prevalent set of 26 switch genes recurring across multiple tumors that were all overexpressed in tumor tissues and appeared primary involved in the regulation of cell cycle.

For cardiomyopathies, we found a prevalent set of 29 switch genes shared between the two cardiomyopathies that were all downregulated in the disease and appeared functionally enriched in the cardiac muscle contraction KEGG pathway. Among them, we found TNNT2, a tropomyosin-binding subunit of the troponin complex that is located on the thin filament of striated muscles and regulates muscle contraction in response to alterations in intracellular calcium ion concentration; MYL3, referred to also as the ventricular isoform, whose mutations have been identified as a cause of mid-left ventricular chamber-type hypertrophic cardiomyopathy; and the ubiquinol-cytochrome c oxidoreductase complex that is part of the mitochondrial electron transport chain, which drives oxidative phosphorylation, playing an important role in the mitochondrial respiratory chain. This observation confirms that disease-specific switch genes fulfill the *functional coherence hypothesis*, being involved in closely disease-related cellular functions.

Shared components hypothesis. We have shown that switch genes may also belong to several disease modules, implying that disease modules may overlap, and, thus, perturbations in one disease module can disrupt pathways of other interlinked disease modules, as well. By building the SHDN, in which nodes are diseases and a link occurs if they share a substantial number of common switch genes, we quantified and visualized the overlap between the disease-associated switch gene modules (Fig. 3d). Although the SHDN was generated independent of any a priori knowledge of disease category, the resulting network is visibly clustered according to major disease classes, where cancers and cardiomyopathies represent the most connected disease classes, in contrast to COPD and AD, which appear as individual disorders. Clustering of nodes of similar color (denoting the same disease class) reflected the fact that similar pathophenotypes have a higher likelihood of sharing genes than do pathophenotypes that belong to different disease classes (Fig. 3d). For example, cancers formed a tightly interconnected and easily detectable cluster, which was held together by a small group of genes that were associated with multiple cancers. Therefore, the SHDN clearly shows how network modules identified by switch genes are highly specific for each disease category and tend to group according to similar pathobiological phenotypes, implying that disease-associated switch gene modules fulfill the *shared components hypothesis*.

The set of 26 tumor-recurring switch genes across multiple tumors showed a marked functional annotation enrichment in cell-cycle-related terms, specifically regulation of the G2-to-M transition. Among them, we found FOXM1, which is a transcription factor with a crucial, central role in cancer development⁴¹. Indeed, FOXM1 overexpression was detected in a variety of human cancers and is associated with poor clinical prognosis^{42,43}; it drives the expression of critical genes involved in the regulation of different cancer hallmarks including high proliferation, invasion, drug resistance, and angiogenesis. In particular, a very recent study demonstrated that FOXM1 physically interacts with the architectural transcription factor HMGA1 to promote tumor angiogenesis cooperatively both in vitro and in vivo models⁴⁴.

Interestingly, among the positive nearest neighbors of FOXM1 in the GENs identified by SWIM methodology, we found both HMGA1 and several well-known pro-angiogenic factors^{45,46} such as tumor necrosis factors (TNFs), fibroblast growth factor (FGF), the

matrix metalloproteinases (MMPs), together with other genes involved in the regulation of angiogenesis such as ADM2, ESM1, E2F7, E2F8, and E2F2. Yet, the negative nearest neighbors of FOXM1 included genes functionally related to metabolic process, as a well appreciated mark of tumor transformation⁴⁷.

This set of 26 tumor-recurrent switch genes appeared coregulated by two major transcription factors (viz., E2F and NF-Y), already known to play key roles in cell cycle regulation and transformation. In particular, the role of NF-Y in controlling cell proliferation has been widely established based on the following findings^{48–51}: it controls the expression of several key regulators of the cell cycle; NF-Y silencing impairs G2/M progression and induces apoptosis; widespread activation of G2/M and anti-apoptotic genes requires NF-Y; NF-Y and mutant p53 physically interact, upregulating the expression of many cell-cycle-related genes in response to DNA damage; and NF-Y overexpression increases cell proliferation. Yet, E2F4 may function as an activator of genes implicated in positive regulation of the cell cycle, including MYBL2 (ref. ⁵²), whose overexpression in transgenic mice leads to the development of tumors, and mutated E2F4, which has been reported in various human tumors, providing evidence for its oncogenic activity^{53–55}.

Taken together, these findings suggest a model wherein NF-Y, in collaboration with E2F4 and/or MYBL2 complex, binds to and activates transcription of E2F/NF-Y-dependent switch genes accelerating the late phase of the cell cycle by promoting angiogenesis with a consequent increase of cancer progression, together with a rewiring of some metabolic pathways, hallmarks of the malignant transformation.

These results support the hypothesis that correlation-based network analysis may move toward causation highlighting functionally coordinated genes whose common perturbations in expression pattern and abundance may contribute to the pathobiological phenotype. In addition, this approach may aid in the identification of biologically significant PPIs (e.g., HMGA1–FOXM1 interaction) of the human interactome, which remains incomplete at the current time.

By definition, disease genes refer to genes with mutations that are known to have a phenotypic impact, e.g., sequence alterations that are causal for Mendelian diseases or variants that increase the susceptibility to complex diseases or cancers^{2,40,56–58}. However, fundamental insights toward the discovery of disease biomarkers can also stem from measuring transcript abundance or gene expression patterns for given phenotypes (case-control) across multiple samples, whose changes could reveal tissue/cell-specific co-expression relationships in the context of the disease^{2,3}.

Here, we demonstrated that switch genes simultaneously satisfy all of the widely used hypotheses and organizing principles formalized by the network medicine construct that tie the interactome to human diseases, in the same way as disease genes themselves do. Thus, the identification of switch genes could allow the systematic prediction of novel disease–gene associations whose perturbations in their expression pattern and abundance contribute to the pathobiological phenotype, as well.

Being context-specific by definition, switch genes can be used to integrate the human interactome with the cell-type or tissue-specific manifestations that characterize many diseases. The driving principle is to use tissue-specific expression information arising from switch genes to filter the global interactome for interactions that are feasible in a given tissue (i.e., both switch interaction partners are present). Furthermore, SWIM methodology may even help in the identification of biologically significant, yet unmapped, PPIs connecting proteins (i.e., switch gene products) on the basis of their co-expression profiles and network-based proximity. In this sense, SWIM supports the link prediction process aiding in increasing the coverage of the human interactome, which is incomplete at the current time⁶.

Taken together, these observations make the accurate identification of switch genes an important step toward a systematic understanding of the networked nature of human pathobiology. We believe that the SWIM-informed approach to protein interaction networks presented here, if broadly applied, would significantly catalyze innovation in the discovery of the determinants of human diseases.

METHODS

SWIM software

In order to identify disease-specific switch genes, we exploited the SWIM software, a program for gene co-expression network mining developed in MATLAB with a user-friendly Graphical User Interface (GUI) and freely downloadable¹⁹.

Consolidated human interactome

To build the comprehensive human interactome, we compiled human physical molecular interaction data from different sources, including PPIs, protein complexes, kinase–substrate interactions, and signaling pathways. PPIs from several high-throughput yeast-two-hybrid studies as well as high-quality PPIs from the literature were compiled from the CCSB Human Interactome^{59–63}. We also collected binary PPIs from other laboratories^{64,65}. A protein complex is a group of two or more associated polypeptide chains linked by non-covalent associations. Protein–protein co-complex interactions were compiled from different high-profile publications^{66–72}. In addition, we also incorporated experimental signaling interactions and kinase–substrate interactions, as well as high-quality literature-based signaling interactions involved in various biological pathways^{73–76}. This new version of the consolidated human interactome has 16,470 proteins and 233,957 interactions after incorporating the latest reference map of the human binary protein interactome⁷⁷.

Ischemic and non-ischemic cardiomyopathy dataset

This dataset is available through the GEO public repository at accession number GSE76293 published on February 10, 2014 (ref. ⁷⁸). Data include a complete RNA-sequencing transcriptome profiling from left ventricular apex tissue from human failing hearts and from non-failing control hearts, with a total of 40 samples: 16 ischemic subjects, 16 non-ischemic subjects, and 8 control heart subjects. High-throughput RNA-sequencing data correspond to normalized expression data created using the reads per kilobase of transcript per million mapped reads (RPKM) procedure to perform the normalization. By running SWIM on ischemic (or non-ischemic) subjects with respect to controls samples, we extracted a list of 81 switch genes, mapped to 68 proteins in the human interactome (less than the total number of switch genes owing to the incompleteness of the interactome).

Alzheimer's disease dataset

This dataset is available through the GEO public repository at accession numbers GSE63060 (batch 1) and GSE63061 (batch 2) published on August 05, 2015 (ref. ⁷⁹). Data include expression profiling by array related to AD and control samples (CTL) originating from the EU funded AddNeuroMed Cohort⁸⁰, which is a large cross-European AD biomarker study relying on human blood as the source of RNA. In particular, batch 1 (GSE63060) comes from array A-MEXP-1171-Illumina HumanHT-12 v3.0 Expression BeadChip and has a total of 249 samples (145 AD, 104 CTL); whereas batch 2 (GSE63061) comes from array A-GEOD-10558-Illumina HumanHT-12 V4.0 expression beadchip and has a total of 273 samples (139 AD and 134 CTL). The probe-sets were mapped to official gene symbols using the relative platform (GPL6947-13512 for GSE63060 and GPL10558-50081 for GSE63061) available from the GEO repository. Multiple probe measurements of a given gene were collapsed into a single gene measurement by considering the mean. By matching genes based on gene symbols, we created a single merged dataset with both batches; we ran Combat function from R/Bioconductor package SVA to correct for batch-specific effects. Finally, we obtained a data matrix of 19,460 gene symbols (rows) and 522 samples (columns) including 284 AD and 238 CTL. By running SWIM on AD subjects with respect to controls samples, we extracted a list of 375 switch genes, mapped to 301 proteins in the human interactome.

Table 1. Summary of TCGA datasets.

Acronym	Tumor name	No. of samples	No. of switch genes	No. of proteins in human interactome
BLCA	Bladder urothelial carcinoma	38 (19 matched-normal)	297	203
BRCA	Breast invasive carcinoma	206 (103 matched-normal)	257	223
CHOL	Cholangiocarcinoma	18 (9 matched-normal)	324	285
COAD	Colon adenocarcinoma	52 (26 matched-normal)	264	217
HNSC	Head and neck squamous cell carcinoma	82 (41 matched-normal)	109	96
LUAD	Lung adenocarcinoma	36 (18 matched-normal)	366	321
LUSC	Lung squamous cell carcinoma	76 (38 matched-normal)	274	254
KIRP	Kidney renal papillary cell carcinoma	46 (23 matched-normal)	133	119
PRAD	Prostate adenocarcinoma	104 (52 matched-normal)	229	177
UCEC	Uterine corpus endometrial carcinoma	14 (7 matched-normal)	395	297

TCGA datasets

A selection of ten tumor types were recovered from the original study¹⁹, where a collection of tumor expression data from high-throughput RNA- and miRNA-sequencing were downloaded from the TCGA data portal on December 6, 2014. High-throughput RNA-sequencing data correspond to level 3 data (i.e., normalized expression data) from RNASeq Version 2 created using MapSplice to do the alignment and RSEM to perform the quantification and normalization. MiRNA-sequencing data correspond to level 3 data (i.e., normalized expression data) created using the RPKM procedure to perform the normalization. In the original study¹⁹, only cancer datasets including at least seven patients with tumor and matched-normal samples (i.e., the matched-normal tissue is defined as the tissue that is adjacent to the tumor and taken from the same patient) for both RNA- and miRNA-sequencing experiments were retained for subsequent analysis. Our selection of ten tumor types, detailed in Table 1, corresponds to tumor types whose switch genes formed a statistically significant module in the human interactome (i.e., showing statistically significant *module significance* for all the three measurements: size of LCC, edges of LCC, and total number of interactions), and showed a statistically significant (p -value < 0.05) *module separation* measure in at least 70% of the comparisons.

COPD dataset

Data for COPD were recovered from the original study²² in which the SWIM software was applied to the COPD dataset. The dataset is available through the GEO public repository at accession number GSE76925 published on March 29, 2017 (ref.⁸¹). Data include microarray gene expression profiling of lung or airway tissue from subjects with COPD obtained using HumanHT-12 BeadChips (Illumina, San Diego, CA). A total of 111 COPD cases and 40 control smokers with normal lung function were collected; all subjects were ex-smokers. The probe-sets were mapped to official gene symbols using the platform GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip) available from the GEO repository. Multiple probe measurements of a given gene were collapsed into a single gene measurement by considering the mean. A list of 61 switch genes was extracted, mapped to 55 proteins in the human interactome.

Module separation

To evaluate disease–disease relationships, we computed the non-Euclidean separation distance, which measures the disease modules' overlap⁶, as follows:

$$s(A, B) = p_{AB} - \frac{p_{AA} + p_{BB}}{2} \quad (1)$$

where, $p(A, B)$ is the module proximity:

$$p(A, B) = \frac{1}{|A| + |B|} \left[\sum_{a \in A} \min_{b \in B} d(a, b) + \sum_{b \in B} \min_{a \in A} d(b, a) \right] \quad (2)$$

and $d(a, b)$ is the shortest distance between switch gene a of module A and switch gene b of module B . A positive value for the separation measure indicates that the two lists of switch genes mapped to proteins are topologically well-separated in the human interactome, whereas a negative value for the separation measure indicates that two switch gene sets are located in the same network neighborhood and, thus, form

overlapping modules with some switch genes belonging to the two disease modules simultaneously.

To evaluate the significance of the network separation parameter across two disease-specific modules (A, B) of switch genes, we built a reference distance distribution corresponding to the expected distance between two randomly selected groups of proteins of the same size and degree distribution as the original two sets of switch genes in the human interactome. The random selection was repeated 1000 times in order to build the reference distance distribution. The module separation measure across the two lists of switch genes was z -score-normalized by using the mean and the standard deviation of the reference distribution. Subsequently, the p -value for the given z statistic was calculated. A p -value < 0.05 indicates that the module separation in the human interactome of the two lists of switch genes is more (or less, see below) than expected by chance.

Generalized measure of module separation

The module separation defined in Eq. (1) and its statistical significance (p -value) were combined into a *generalized measure of module separation* modeled as the following sigmoidal function:

$$\tilde{s}(A, B) = \frac{1}{1 + e^{-\alpha \frac{s(A, B)}{pval(A, B)}}} - 0.5 \quad (3)$$

where $s(A, B)$ is the module separation, $pval(A, B)$ is the corresponding p -value, and α is a smoothing parameter: the greater the α , the steeper the function (Fig. 3a).

This generalization of $s(A, B)$ is a bounded function returning a value that monotonically increases from -0.5 to 0.5 , and explicitly considers the statistical significance (i.e., the p -value) of the observed module separation between each disease pair (A, B): the lower the p -value, the greater the absolute value of $\tilde{s}(A, B)$ with $|\tilde{s}|$ approaching 0.5 as the p -value approaches 0. In particular, we chose a quite small value of α ($=0.3$) in order to emphasize better the differences between negative (overlapping disease modules) and positive (well-separated disease modules) values of s when the corresponding p -value is small and statistically significant (i.e., p -value < 0.05). Note that for statistically insignificant p -values (i.e., p -value > 0.05), the $\tilde{s}(A, B)$ clearly shows its exponential behavior near zero (Fig. 3a).

Functional and motif enrichment analysis

The functional enrichment analysis was performed using EnrichR web tool⁸². Binding motif enrichment analysis in promoter regions (identified as genomic regions spanning from -450 to $+50$ nucleotides with respect to transcription start sites) was performed using Pscan⁸³, which employs the JASPAR 2018 motif collection⁸⁴. p -Values were adjusted with the Benjamini–Hochberg method, and a threshold equal to 0.05 was set to identify functional annotations and regulatory motifs significantly enriched among the selected switch gene lists.

DATA AVAILABILITY

The accession codes, unique identifiers, or web links for publicly available datasets are provided in Datasets subsection of "Methods". Data associated to figures are provided as supplementary material.

CODE AVAILABILITY

SWIM code is freely available at www.nature.com/articles/srep44797 (Supplementary Information).

Received: 4 June 2020; Accepted: 19 October 2020;

Published online: 21 January 2021

REFERENCES

- Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
- Caldera, M., Buphamalai, P., Müller, F. & Menche, J. Interactome-based approaches to human disease. *Curr. Opin. Syst. Biol.* **3**, 88–94 (2017).
- Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network medicine in the age of biomedical big data. *Front. Genet.* **10**, 294 (2019).
- Barabási, A.-L. Network medicine—from obesity to the “Diseaseome”. *N. Engl. J. Med.* **357**, 404–407 (2007).
- Goh, K.-I. et al. The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
- Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
- Ghiassian, S. D., 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.* **11**, e1004120 (2015).
- Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
- Fiscon, G., Conte, F., Farina, L. & Paci, P. Network-based approaches to explore complex biological systems towards network medicine. *Genes* **9**, 437 (2018).
- Conte, F. et al. A paradigm shift in medicine: a comprehensive review of network-based approaches. *Biochim. Biophys. Acta Gene Regul. Mech.* **1863**, 194416 (2020).
- Panebianco, V. et al. Prostate cancer screening research can benefit from network medicine: an emerging awareness. *Npj Syst. Biol. Appl.* **6**, 1–6 (2020).
- Silverman, E. K. et al. Molecular networks in network medicine: development and applications. *WIREs Syst. Biol. Med.* **12**, e1489, <https://doi.org/10.1002/wsbm.1489> (2020).
- Cacace, F., Farina, L., Germani, A. & Manes, C. Internally positive representation of a class of continuous time systems. *IEEE Trans. Autom. Control* **57**, 3158–3163 (2012).
- Cheng, F. et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691 (2018).
- Zhou, Y. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* **6**, 1–18 (2020).
- Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
- Wang, R.-S. & Loscalzo, J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J. Mol. Biol.* **430**, 2939–2950 (2018).
- Iida, M., Iwata, M. & Yamanishi, Y. Network-based characterization of disease-disease relationships in terms of drugs and therapeutic targets. *Bioinformatics* **36**, i516–i524 (2020).
- Paci, P. et al. SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Sci. Rep.* **7**, srep44797 (2017).
- Fiscon, G., Conte, F., Licursi, V., Nasi, S. & Paci, P. Computational identification of specific genes for glioblastoma stem-like cells identity. *Sci. Rep.* **8**, 7769 (2018).
- Falcone, R. et al. BRAFV600E-mutant cancers display a variety of networks by SWIM analysis: prediction of vemurafenib clinical response. *Endocrine* **64**, 406–413 (2019).
- Paci, P. et al. Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. *Sci. Rep.* **10**, 1–18 (2020).
- Fiscon, G., Conte, F. & Paci, P. SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines. *BMC Bioinform.* **19**, 436 (2018).
- Palumbo, M. C. et al. Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch genes that induce major transcriptome reprogramming during grapevine development. *Plant Cell*. <https://doi.org/10.1105/tpc.114.133710> (2014).
- McLendon, R. et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
- Gimple, R. C., Bhargava, S., Dixit, D. & Rich, J. N. Glioblastoma stem cells: lessons from the tumor hierarchy in a lethal cancer. *Genes Dev.* **33**, 591–609 (2019).
- Suva, M. L. et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594 (2014).
- Ghiassian, S. D. et al. Endophenotype network models: common core of complex diseases. *Sci. Rep.* **6**, 1–13 (2016).
- Anderson, M. J. & Legendre, P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Stat. Comput. Simul.* **62**, 271–303 (1999).
- Hayes, A. F. Cautions in testing variance equality with randomization tests. *J. Stat. Comput. Simul.* **59**, 25–31 (1997).
- Kennedy, P. E. Randomization tests in econometrics. *J. Bus. Econ. Stat.* **13**, 85–94 (1995).
- Marozzi, M. A bi-aspect nonparametric test for the two-sample location problem. *Comput. Stat. Data Anal.* **44**, 639–648 (2004).
- Shipley, B. A permutation procedure for testing the equality of pattern hypotheses across groups involving correlation or covariance matrices. *Stat. Comput.* **10**, 253–257 (2000).
- Wan, Y., Cohen, J. & Guerra, R. A permutation test for the robust sib-pair linkage method. *Ann. Hum. Genet.* **61**, 77–85 (1997).
- Smith, E. P. Randomization methods and the analysis of multivariate ecological data. *Environmetrics* **9**, 37–51 (1998).
- Bailer, A. J. Testing variance equality with randomization tests. *J. Stat. Comput. Simul.* **31**, 1–8 (1989).
- Pesarin, F. An Almost Exact Solution For The Multivariate Behrens-Fisher Problem. *Metron* **55**, 85–100 (1997).
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
- Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
- Piñero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
- Liao, G.-B. et al. Regulation of the master regulator FOXM1 in cancer. *Cell Commun. Signal.* **16**, 57 (2018).
- Koo, C.-Y., Muir, K. W. & Lam, E. W.-F. FOXM1: From cancer initiation to progression and treatment. *Biochim. Biophys. Acta Gene Regul. Mech.* **1819**, 28–37 (2012).
- Wierstra, I. FOXM1 (Forkhead box M1) in tumorigenesis: overexpression in human cancer, implication in tumorigenesis, oncogenic functions, tumor-suppressive properties, and target of anticancer therapy. *Adv. Cancer Res.* **119**, 191–419 (2013).
- Zanin, R. et al. HMGA1 promotes breast cancer angiogenesis supporting the stability, nuclear localization and transcriptional activity of FOXM1. *J. Exp. Clin. Cancer Res.* **38**, 313 (2019).
- Loizzi, V. et al. Biological pathways involved in tumor angiogenesis and bevacizumab based anti-angiogenic therapy with special references to ovarian cancer. *Int. J. Mol. Sci.* **18**, 1967 (2017).
- Bergers, G. & Benjamin, L. E. Tumorigenesis and the angiogenic switch. *Nat. Rev. Cancer* **3**, 401–410 (2003).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Benatti, P. et al. A balance between NF- κ B and p53 governs the pro- and anti-apoptotic transcriptional response. *Nucleic Acids Res.* **36**, 1415–1428 (2008).
- Caretti, G., Salsi, V., Vecchi, C., Imbriano, C. & Mantovani, R. Dynamic recruitment of NF- κ B and histone acetyltransferases on cell-cycle promoters. *J. Biol. Chem.* **278**, 30435–30440 (2003).
- Hu, Q. & Maity, S. N. Stable expression of a dominant negative mutant of CCAAT binding factor/NF- κ B in mouse fibroblast cells resulting in retardation of cell growth and inhibition of transcription of various cellular genes. *J. Biol. Chem.* **275**, 4435–4444 (2000).
- Gurtner, A. et al. NF- κ B dependent epigenetic modifications discriminate between proliferating and postmitotic tissue. *PLoS ONE* **3**, e2047 (2008).
- Lee, B.-K., Bhing, A. A. & Iyer, V. R. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res.* **39**, 3558–3573 (2011).
- Wang, D., Russell, J. L. & Johnson, D. G. E2F4 and E2F1 have similar proliferative properties but different apoptotic and oncogenic properties in vivo. *Mol. Cell. Biol.* **20**, 3417–3424 (2000).
- Khaleel, S. S., Andrews, E. H., Ung, M., DiRenzo, J. & Cheng, C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. *Breast Cancer Res.* **16**, 486 (2014).
- Souza, R. F. et al. Frequent mutation of the E2F4 cell cycle gene in primary human gastrointestinal tumors. *Cancer Res.* **57**, 2350–2353 (1997).
- Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
- Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
- Rath, A. et al. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).

59. Rual, J.-F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
60. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
61. Yu, H. et al. Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8**, 478–480 (2011).
62. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
63. Yang, X. et al. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805–817 (2016).
64. Stelzl, U. et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
65. Yachie, N. et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12**, 863 (2016).
66. Ewing, R. M. et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
67. Havugimana, P. C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
68. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907–909 (2012).
69. Huttlin, E. L. et al. The BioPlex Network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
70. Wan, C. et al. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).
71. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
72. Huttlin, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
73. Vinayagam, A. et al. A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8 (2011).
74. Túrei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
75. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
76. Newman, R. H. et al. Construction of human activity-based phosphorylation networks. *Mol. Syst. Biol.* **9**, 655 (2013).
77. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
78. Yang, K.-C. et al. Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support. *Circulation* **129**, 1009–1021 (2014).
79. Sood, S. et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* **16**, 185 (2015).
80. Lovestone, S. et al. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer’s disease. *Ann. N.Y. Acad. Sci.* **1180**, 36–46 (2009).
81. Morrow, J. D. et al. Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Sci. Rep.* **7**, 44232 (2017).
82. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
83. Zambelli, F., Pesole, G. & Pavesi, G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.* **37**, W247–W252 (2009).

84. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

ACKNOWLEDGEMENTS

The authors wish to thank Ms. Stephanie Tribuna for expert technical assistance. This work was supported, in part, by NIH grants HG007690, HL108630, and HL119145; and AHA grants D700382 and CV-19; and Rockefeller Foundation grant FOD-26 to J. Loscalzo; PRIN 2017 - Settore ERC LS2 - Codice Progetto 20178L3P38; and Sapienza University of Rome grant entitled “Network medicine-based machine learning and graph theory algorithms for precision oncology” - n. RM1181642AFA34C2.

AUTHOR CONTRIBUTIONS

P.P., L.F., and J.L.: concept and design. P.P., G.F., F.C., and R.W.: analysis of data. All authors contributed to interpretation of data, review, and approval of the final manuscript.

COMPETING INTERESTS

J.L. is scientific co-founder of Scipher Medicine, Inc., which uses network medicine analyses to identify disease biomarkers and potential therapies. The other authors have no competing interests to declare.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41540-020-00168-0>.

Correspondence and requests for materials should be addressed to P.P.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2021