genes

MDPI

Article

Enriching Human Interactome with Functional Mutations to Detect High-Impact Network Modules Underlying Complex Diseases

Hongzhu Cui [1,*], Suhas Srinivasan [2] and Dmitry Korkin [1,2,3,*]

[1] Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA 01609, USA
[2] Data Science Program, Worcester Polytechnic Institute, Worcester, MA 01609, USA; ssrinivasan@wpi.edu
[3] Computer Science Department, Worcester Polytechnic Institute, Worcester, MA 01609, USA
* Correspondence: hcui2@wpi.edu (H.C.); dkorkin@wpi.edu (D.K.)

Received: 29 September 2019; Accepted: 11 November 2019; Published: 15 November 2019

Abstract: Rapid progress in high-throughput -omics technologies moves us one step closer to the datacalypse in life sciences. In spite of the already generated volumes of data, our knowledge of the molecular mechanisms underlying complex genetic diseases remains limited. Increasing evidence shows that biological networks are essential, albeit not sufficient, for the better understanding of these mechanisms. The identification of disease-specific functional modules in the human interactome can provide a more focused insight into the mechanistic nature of the disease. However, carving a disease network module from the whole interactome is a difficult task. In this paper, we propose a computational framework, Discovering most IMpacted SUbnetworks in interactoMe (DIMSUM), which enables the integration of genome-wide association studies (GWAS) and functional effects of mutations into the protein–protein interaction (PPI) network to improve disease module detection. Specifically, our approach incorporates and propagates the functional impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on PPIs to implicate the genes that are most likely influenced by the disruptive mutations, and to identify the module with the greatest functional impact. Comparison against state-of-the-art seed-based module detection methods shows that our approach could yield modules that are biologically more relevant and have stronger association with the studied disease. We expect for our method to become a part of the common toolbox for the disease module analysis, facilitating the discovery of new disease markers.

Keywords: protein–protein interaction network; module detection; GWAS; network propagation; functional annotation; complex diseases

1. Introduction

Since the first evidence of the genetic complexity of cancer [1], numerous research efforts have been dedicated to deciphering the mechanistic nature of polygenic diseases. Due to the rapid advancement of the next generation sequencing (NGS) technologies, including whole-genome [2] and whole-exome sequencing [3], and most recently single-cell transcriptomics [4], we have been able to sequence and analyze thousands of genomes at a much lower cost. As result, the high-throughput experiments produce large amounts of genomic data at the exponentially increasing rate. These data have also transformed the design of genome-wide association studies and enabled us to do comprehensive analyses of the genotype–phenotype relationships [5]. Most importantly, the studies have provided us with an extensive list of susceptible alleles and genes associated with complex diseases, as well as with catalogs of disease-relevant mutations [6]. The list includes the majority of

common and many rare complex diseases. Furthermore, it is expected that a comprehensive catalog of nearly all human genomic variations will be available soon [7].

A typical large-scale NGS-based study reports several million genetic variants [5]. However, not all mutations, even with statistically significant correlations with the disease, would contribute to the disease phenotype [8]. For example, in cancer genomics, many mutations are defined as "passenger" mutations. Unlike the "driver" mutations, which induce the clonal expansion, the passenger mutations do not provide any functional advantage to the development of cancer cells [9, 10]. Thus, distinguishing between the functional and non-functional mutations is usually the first step in genetics studies. Additionally, computational approaches for functional annotation are increasingly important, since many variants are not previously described in the literature [6, 11]. There are a plethora of functional annotation tools for genetic variants [6]. Many tools focus on the annotation of single nucleotide variants (SNVs), the most common type of genetic variation, which is easier to capture and analyze. Recently, we have developed a new computational method, the non-synonymous SNP INteraction effect predictor tool (SNP-IN) tool [12], which predicts the effects of non-synonymous SNVs on protein–protein interaction (PPI), given the interaction's structure or structural model. The accurate and balanced performance of the SNP-IN tool makes it apt for the functional annotation of non-synonymous SNVs. In particular, the SNP-IN tool may be helpful for characterizing the edgetic profiles of SNVs. Edgetics, or edgotype, is a recently proposed concept in network biology [13, 14]. Unlike the traditional view that a genetic variation causes a complete loss of gene product, the edgetic perturbation model treats such variation as interaction-specific 'edgetic' perturbation: the variation may cause the removal or addition of specific interactions while other edges remain unperturbed. To support this new model, the SNP-IN tool can be used as an *in silico* edgetic profiling tool.

The need of integrating Genome-wide Association Studies (GWAS) and the functional impact of the disease-associated mutations with the systems data is supported by the increasing body of evidence that large-scale biological systems and cellular networks underlie the majority, if not all, of complex genotype–phenotype relationships in diseases [15-18]. Understanding the biological network is essential in studying a genetic disease, because such a disease is likely a result of the disruption and rewiring of the complex intracellular molecular network, rather than a dysfunction of a single gene [19]. Along with the increasing availability of the high-throughput human protein interactomics data [20, 21], new computational approaches have been developed. In particular, network propagation has recently emerged as a prominent approach in network biology [22]. In network propagation, genes/proteins of interest correspond to the nodes in the biological network, the edges represent pair-wise protein–protein interactions, and the information is propagated through the edges to nearby nodes in an iterative fashion. Thus, it could amplify the weaker disease association signals from the genes interacting with the "seed" genes that carry the stronger source signal [22]. Network propagation have been applied for various purposes, including predicting gene function, identifying disease related subnetworks, and drug target prediction [22, 23].

Module identification is a central problem in network biology [24]. A topological module is a set of genes (nodes) with dense interactions between each other; these groups of nodes are also referred to as communities, or clusters, in network science [25]. A topological module can be functional, since the constituting proteins are often shown to pertain to the same biological function or be involved in a similar biological process. A disease module is a sub-network of proteins enriched with the disease-relevant proteins and responsible for the disease phenotype. Various module identification approaches have been proposed, presenting a wide range of theoretical perspectives and implementations. These methods primarily come in two different flavors. The first group of methods identify the modules in a biological network by relying exclusively on the network's topology. This is a challenging task due to the lack of information about specific genes/proteins contributing to biological functionality or disease phenotype. The recent open community DREAM challenge [24] provided a good review and benchmark of existing methods falling in this category. Methods from the second group start with the "seed genes", and gradually extract additional genes in the network to grow the module. For example, DIseAse MOdule Detection (DIAMOnD) [26] is a disease module

detection algorithm that utilizes known seed genes to identify disease modules according to the number of connections to the seed proteins. The algorithm outputs a connected disease module with a list of candidate disease-associated proteins ranked by their connectivity significance.

Discovering biologically relevant modules (disease modules or functional modules) is a challenging task [27, 28]. To tackle the mechanistic intricacy underlying complex diseases, it is necessary to couple disparate sources of data, each informing about a different aspect of the biological function. Computational approaches integrating molecular networks with different types of -omics data have demonstrated considerable power in bioinformatics studies [29]. For instance, integrating PPI networks with the gene expression profiles to identify sets of genes that participate in a biological function has been helpful to reveal the functional modularity of the network [30]. Suprisingly, integrating interactomics data with GWAS data has not yet gained wide attention in the bioinformatics community, and functional annotation information regarding genetic variants and mutated genes are not included during such integrations.

In this work, we developed a novel algorithmic framework named DIMSUM (Discovering most IMpacted SUbnetworks in interactoMe) to identify functional disease module in the human interactome. The DIMSUM framework includes three major steps: (1) network annotation, (2) network propagation, and (3) subnetwork extraction. Our approach benefits from integrating GWAS data, functional annotation information, and the protein–protein interaction network. We evaluated our approach using a set of eight complex diseases against two state-of-the-art seed-based module detection methods: DIAMOnD and Seed Connector Algorithm (SCA). From the set of complex diseases, we also carried out two case studies: the first study centered around genes associated with coronary artery disease, and the second one focusing on joint analysis of Schizophrenia and Bipolar Disorder. The evaluations results show that DIMSUM outperforms both DIAMOnD and SCA, because the discovered modules have stronger association with the disease and are more biologically relevant with the seed-gene pool.

2. Materials and Methods

At the preprocessing stage, we carried out the human interactome construction, GWAS data collection, and data processing (Figure 1). At the same time, we mapped the mutations to the structures of affected PPIs, and applied our SNP-IN tool [12] to characterize mutation-induced rewiring effects on the PPIs [15]. The computational workflow of DIMSUM consists of 3 main stages. In the first stage, we updated the node and edge weights of a fully annotated human PPI network: the node weight reflects the association with a disease, while the edge weight reflects the cumulative damage made to the corresponding interaction. Second, we applied a network propagation strategy with a goal to boost the signal for the genes from the GWAS study with weak association, thus increasing the pool of candidate genes. The last stage includes sub-network extraction: we proposed an iterative procedure to find the disease module with the greatest impact on the disease.

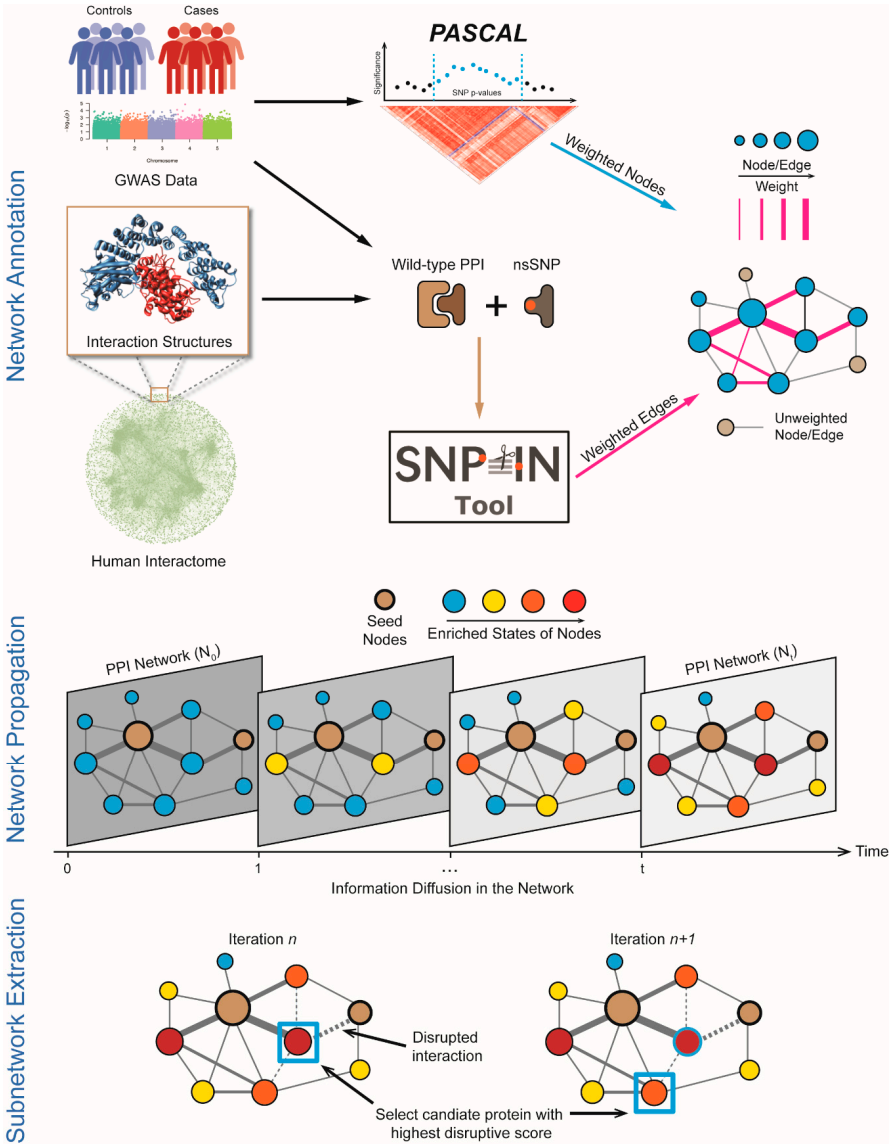

Figure 1. Basic workflow of Discovering most IMpacted SUbnetworks in interactoMe (DIMSUM) computational framework. The module detection framework contains three major steps: network annotation; network propagation and subnetworks extraction. In Network Annotation stage, we collect GWAS data and use Pascal tool to aggregate single nucleotide polymorphisms (SNP) summary statistics. Then we apply the non-synonymous SNP INteraction effect predictor tool (SNP-IN) tool to properly assess SNP's impact on the protein–protein interactions. Finally, we integrate this information with the human interactome to generate a fully weighted network. In Network Propagation stage, we apply a network propagation procedure to implicate genes most likely to be influenced by disruptive SNPs. Finally, in Subnetwork Extraction stage, we apply an iterative strategy to identify the most impacted module.

2.1. Human Interactome Construction

In this work, we utilized two different protein–protein interaction data sources to construct the human interactome. The first data source is the High-quality INTeractomes (HINT) database [31]. It integrates several databases and filters out low-quality and erroneous interactions. The other source is the Human Reference Protein Interactome Mapping Project (HuRI) [21]. HINT is a manually curated repository of PPIs mainly from the literature, whereas HuRI is a primary source of experimentally validated PPIs. We considered these two PPI sources as they complement each other,

hence we constructed the interactome by combining both. The current release of HINT interactome (Version 4) contains 63,684 interactions. Combining all the three proteome-scale human PPI datasets released from HuRI at different stages of the project, we obtained 76,537 interactions. In total, we generated a human interactome consisting of 105,087 interactions, where 35,134 interactions existed in both data sources.

2.2. GWAS Data Collection and Processing

A distinctive feature of our method was that it integrated GWAS data into the interactome to improve the disease module detection. To do so, we compiled eight publicly available GWAS datasets. The datasets spanned a broad range of diseases, including neurodegenerative, metabolic, and psychiatric disorders. For the GWAS integration, the dataset was only required to have pre-calculated summary statistics, no individual level information was needed. To generate the seeds for the later stage of network propagation, we computed gene scores by aggregating SNP *p*-values from GWAS studies using the Pascal tool (Pathway scoring algorithm) [32]. Integrating SNP *p*-values from GWAS studies has proven itself as a powerful method to improve statistical power.

Pascal is a fast and rigorous computational tool developed to aggregate SNP summary statistics into gene scores with the high power, while absolving the need to access the original, individual-level, genotypic data. It can be considered as an alternative to the traditional *p*-value estimation approaches, including chi-squared statistics (SOCS) and the maximum of chi-squared statistics (MOCS) [33], which measure the average and the strongest associations of signals per gene, respectively. Pascal relies on the assumption that a pairwise correlation matrix of the contributing genotypes underlying the null distributions of the MOCS and SOCS statistics can be estimated from the ethnicity-matched, publicly available genotypic data. To calculate gene scores, we selected the sum of chi-squared statistics (SOCS) of all SNPs from genes of interest. To properly correct for linkage disequilibrium (LD) correlation structure in GWAS data, we used the European population of the 1000 Genomes Project [34], because GWAS studies in this work are predominantly the European cohorts. The disease-associated genes with significant *p*-values that Pascal produces were then defined as the seed genes for our approach. The final list of seed genes was acquired after correcting for multiple testing using Bonferroni correction.

2.3. Functional Annotation of nsSNV with the SNP-IN Tool

To properly assess the functional damage caused by mutations with respect to protein–protein interactions, we applied our recently developed SNP-IN tool (non-synonymous SNP INteraction effect predictor tool) [12]. The SNP-IN tool predicts the rewiring effects of nsSNVs on PPIs, given the interaction's experimental structure or accurate comparative model. More specifically, the SNP-IN tool formulates this task as a classification problem. There are three classes of edgetic effects predicted by the SNP-IN tool: beneficial, neutral, and detrimental. The effects are assigned based on the difference between the binding free energies of the mutant and wild-type complexes ($\Delta\Delta G$). Specifically, $\Delta\Delta G = \Delta\Delta G_{mt} - \Delta\Delta G_{wt}$, where $\Delta\Delta G_{mt}$ and $\Delta\Delta G_{wt}$ are the mutant and wild-type binding free energies correspondingly. The beneficial, neutral, or detrimental types of mutations are then determined by applying two previously established thresholds to $\Delta\Delta G$ [35, 36]:

$$\textit{Beneficial: } \Delta\Delta G < -0.5 \text{ kcal/mol}$$

$$\textit{Neutral: } -0.5 \text{ kcal/mol} \leq \Delta\Delta G < 0.5 \text{ kcal/mol}$$

$$\textit{Detrimental: } \Delta\Delta G \geq 0.5 \text{ kcal/mol.}$$

The annotation workflow begins with processing the GWAS data. Most mutations in GWAS datasets come with only dbSNP RefSNP cluster ID's (rs#). The variant data is preprocessed using ANNOVAR [37] to retrieve SNV locations on the genes and the corresponding residue change information. For each mutated gene and the corresponding SNP, we collected all their interacting partners in the merged interactome (see Section 2.1). Because the SNP-IN tool is a structure-based

classifier, we needed both the mutation information and interaction structure as an input. There are several different cases when generating the PPI structure in which a mutated gene is involved (Supplementary Figure S1). First, if a PPI already has a native structure, it is extracted from a protein data bank (PDB) [38]. In the corresponding PDB file, we first identified the interacting subunit pair for each PPI using the 3did database [39]. The 3did database maintains information regarding the two interacting domains with physical interfaces. If there was no native structure for a protein–protein interaction, two options were explored. First, if a structural template for such interaction (i.e., a homologous protein interaction complex) existed, a comparative model of this interaction could be obtained [40]. When a full-length PPI could not be modeled, we only modeled the domain–domain interaction that included the domain containing the mutation. Homology modeling was done through Interactome3D [41], a web service for structural modeling of PPI network.

2.4. Network Annotation and Network Propagation

The human interactome is next represented as the graph $G = (V, E)$, where V is the set of nodes representing the genes and E is the set of edges representing the protein–protein interactions. After applying the Pascal tool to the GWAS studies, we obtained p-values of disease associated genes. These values, together with the functional annotations of the corresponding mutations from the SNP-IN tool, were used to weight the nodes and edges in the network. Thus, G is a graph with weighted nodes and edges. Specifically, for a disease-associated gene i, the corresponding node was weighted with $-\log(P_i)$, where P_i is the p-value obtained from Pascal tool. A node corresponding to a gene that is not listed in the GWAS study was assigned a zero weight. The edge was weighted according to the damage accumulated on the corresponding PPI by the disruptive SNVs. Specifically, a PPI between genes i and j was weighted with the total number of disruptive SNVs on both genes targeting the same interaction. In other words, the node weight reflected the "relevance" of the gene to the disease, and the edge weight reflected the accumulated "damage" on the interaction.

Next, we applied the network propagation strategy to implicate other core disease genes affected by the perturbations of disruptive SNVs. Let $F: V \rightarrow \Re$ represent a function reflecting the relevance of gene i to a specific complex disease. The goal of the network propagation was to prioritize the genes that were not showing significant association based on the GWAS study, but were expected to have possible relevance to the disease. We imposed two constraints on the prioritization function F: the computed function should be (1) smooth and (2) compliant with the prior knowledge. Smoothness of the function was defined by the assignment of similar values to the interaction partners (nodes) of disease gene i. The compliance with the prior knowledge implies that the difference between the final computed value for a disease gene $F(i)$ and the initial value. The values of F can be iteratively obtained as follows:

$$F_{t+1} = \alpha W' F_t + (1 - \alpha) Y, \quad (1)$$

where Y is the prior knowledge defined as the node weight, α is a parameter reflecting the importance of the two constraints mentioned above with the default value $\alpha = 0.5$, W' is a $|V|*|V|$ matrix whose values are determined by the edge weights and that is defined as a normalized form of network edge weight matrix W. Formally, we introduced a diagonal matrix D, such that $D(i, i)$ was the sum of row i of W. We then set $W' = D^{-1/2}WD^{-1/2}$. F_t was initialized as Y. The equation could be solved iteratively and guaranteed to converge to the system's solution [42]. This iterative algorithm could be considered as propagating the prior information from some nodes through the network. The disease genes first sent the signal to their neighbors, and every node then propagated the received signal to its neighbors (Figure 1). There were also additional constraints on the weighted network. First, when we propagated the information through the network, both the node and edge weights should have been in (0,1) range [42]. In this work, the seed genes in the network carried the largest weights. Thus, we normalized the node weight:

$$p_i' = \frac{p_i - p_{min}}{p_{max} - p_{min}}, \quad (2)$$

where p_{max} and p_{min} are the maximum and minimum of the initial node weights, i.e., –log(P_i). After network propagation, we de-normalized the node weight at the subnetwork extraction step (see Section 2.5 next). For the edge weight, the higher weight meant that the information flow was more likely to go through that edge. Thus, the weight was converted to (0,1) range using a sigmoid transformation:

$$w_{i,j}' = \frac{1}{1 + e^{w_{i,j}}}, \quad (3)$$

where the $w_{i,j}$ is the original weight of the edge. After convergence, the disease association information from seed genes was diffused into the interactome, and all the nodes were weighted with their relevance to the disease.

2.5. Sub-Network Extraction

Finally, we extracted the sub-network with the greatest "impact" on the disease etiology. To do so we defined disease-associated genes with significant *p*-values obtained from Pascal as the seed genes. The goal was to extract a sub-network containing all the seed genes while maintaining the greatest impact. Intuitively, the impact was defined based on the "severity" of the network damage caused by the disruptive SNVs located on the genes with high relevance to the disease. The relevance was reflected by the node weight after the network propagation procedure, while the network damage was determined based on the edge weights reflecting the total number of disruptive mutations occurring in each interaction. The subnetwork extraction was then formulated as an iterative procedure:

I. Assume that a seed gene set $\{g_1, g_2, ..., g_k\}$ induces a subnetwork with initial size N_0. For all immediate interacting partners of the seed gene set that are not in the subnetwork, define an impact score:

$$S(i) = p * \sum_j w_{i,j} \quad (4)$$

where p is the updated *p*-value after the network propagation and $\sum_j w_{i,j}$ is the total number of mutations that disrupt the PPIs between the candidate gene *i* and every gene *j* in the module gene set. Thus, the impact score combines the disease relevance and potential disruption to existing subnetwork caused by the candidate gene.

II. All the immediate interaction neighbors of the seed gene set, not included in the subnetwork, are ranked according to their impact score.

III. Select the gene with the largest impact score. If there are multiple candidates with the same impact score, randomly pick one gene to break the tie. Add the gene with the biggest impact to the set of seed genes and increase the size of the induced subnetwork by 1: $N_{t+1} = N_t + 1$.

Given a number of maximum iterations, steps I–III were repeated in a loop until the number of added genes equals maximum iteration number. The code for network propagation and sub-network extraction is available at https://github.com/hcui2/DIMSUM.

2.6. Validation and GO Analysis

To validate the performance of our module detection method, we compared it against two seed-based module detection methods, a widely used method called DIAMOnD [26], and a recently published method called SCA [43] For both methods, we used the same seed genes that were used for DIMSUM. For each method, we also limited the number of candidate genes forming a disease module to 100. To validate the disease association of the predicted candidate genes, we first compiled a list of known disease-related genes from two databases, Human Gene Mutation Database (HGMD) [44] and Online Mendelian Inheritance in Man (OMIM) [45]. The candidate genes supported by the literature were considered as true positives.

Furthermore, to compare the biological relevance of candidate genes to the seed genes we performed Gene Ontology (GO) enrichment analysis [46] on the two gene sets. In the GO enrichment analysis, we used the third level of the GO hierarchy as a trade-off between the too general and well-populated GO terms at the second level, and specific but not well-populated terms at the fourth level. The GO enrichment was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [37], and GO terms with a p-value ≤ 0.01 were selected. During the analysis, we first identified significantly enriched GO terms within the seed genes. We then checked how many of the GO terms significantly enriched in the pool of candidate genes were identical to the enriched GO terms in the pool of seed genes. The higher number of the identical GO terms suggests the stronger relevance between the candidate genes and the seed genes.

3. Results

3.1. Seed Genes Generated from GWAS Datasets

We collected eight GWAS datasets from various public sources. The collected GWAS data cover a diverse range of eight complex diseases, including Alzheimer's, bipolar disorder, coronary artery disease, macular degeneration, osteoporosis, rheumatoid arthritis, schizophrenia, and type 2 diabetes mellitus (Supplementary Table S1). The GWAS datasets were only required to contain SNP-phenotype association summary statistics, no individual level genotype information was used. These studies were predominantly from European cohorts, with the total number of SNPs reported in each study varying from 2 million to 12 million (Supplementary Table S1).

Our subnetwork detection strategy benefited from integrating the GWAS dataset and network data. Specifically, we derived the seed genes for network propagation from the GWAS dataset (see Section 2.2 in Methods). The length of the gene list generated from Pascal varied for each disease due to different sizes of GWAS datasets. The number of seeds for each disease ranged from 26 to 301 (Supplementary Table S2).

3.2. Functional Predictions from the SNP-IN Tool

The lack of functional knowledge for SNPs obtained from GWAS studies limits our understanding of the mechanistic processes that underlie diseases. Although there is a plethora of functional annotation tools for SNPs, most of them provide with annotations of generic putative deleterious effects of SNPs [6]. In particular, they do not provide the means to determine how SNPs disrupt protein–protein interactions, while such information could lead to a better understanding of how SNPs rewire the human interactome and help identify the impacted subnetworks responsible for the disease. Our recently developed SNP-IN tool accurately predicts how mutations affect the PPIs, given the interaction's structure (see Section 2.3 in Methods). Given that the sizes of GWAS datasets considered in this work varied, the prediction coverage of non-synonymous SNPs also varied for different diseases, ranging from 547 to 8323 (Supplementary Table S3). Previous studies reported high percentage of disease-associated mutations that affected PPIs [15, 17, 47]. Specifically, our latest study [15] showed that out of all pathogenic mutations collected from the ClinVar database, 76.2% were predicted to have a disruptive effect on PPIs. In the present work, on average, 51.1% of the annotated mutations from all eight GWAS studies were predicted to have detrimental effects on PPIs. The percentages of disruptive mutations in this study were lower than our previous work, which could be attributed to the fact that some of mutations detected in the GWAS studies were random mutations or passenger mutations without significant functional impact. Nevertheless, the current study showed that a considerable amount of mutations occurring in a disease could rewire the human interactome centered around the disease. In addition, these results reaffirmed our previous findings that the beneficial mutations, strengthening the PPIs, were rare in the human genome. The reported beneficial mutations are less than 1% percent in all cases (Supplementary Table S3). Given such a low percentage, we discard this beneficial mutation during network propagation.

3.3. Network Annotation and Network Propagation

A key idea of our approach is in integrating GWAS data and functional annotation data with the human interactome data to improve the network module detection. Specifically, we utilized the GWAS study results and functional annotations from the SNP-IN tool to properly weight the network. The node weight reflected the relevance of the gene to the disease, while the edge weight represented the cumulative damage imposed on the corresponding interaction (see Section 2.4 in Methods). Once the fully weighted network was generated, we examined the topological properties of the genes carrying disruptive mutations and the damaged interactions in the human interactome. In particular, between the eight datasets we compared the distribution of the node degrees in the human interactome for disease-associated genes. The average node degree for the genes carrying disruptive mutations among the eight diseases ranged from 14 to 24 (Figure 2a). Compared to the average node degree of the human interactome, all disease networks showed increased average degree suggesting that disease-associated mutations tend to disrupt genes occupying a central spot in the human interactome, rather than lying on the periphery. The results also suggested that mutations captured in complex diseases were more likely to cause the network rewiring than random mutations.

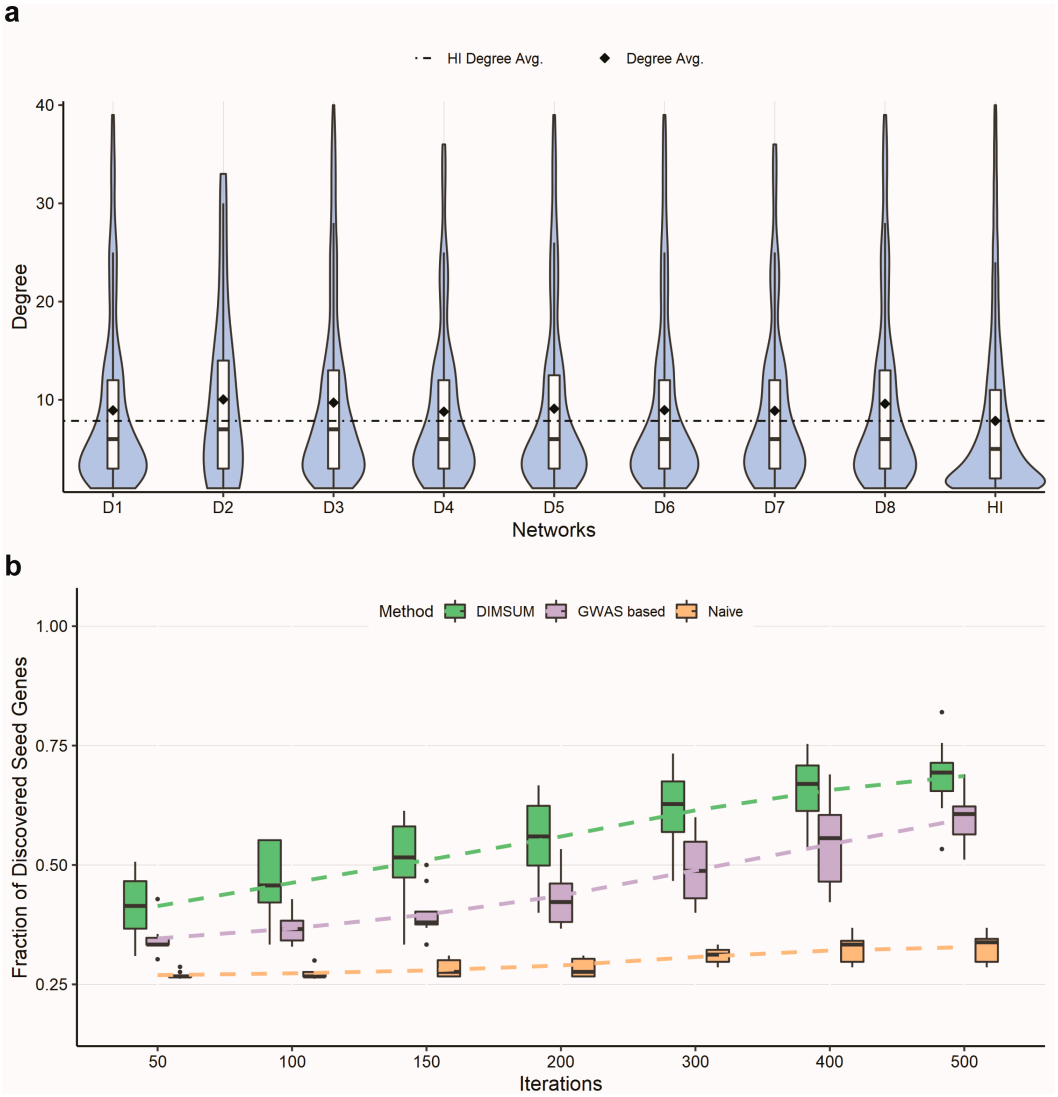

Figure 2. Annotated networks and comparison of DIMSUM against a naïve and a GWAS based network propagation procedures. (**a**) The first eight violin plots represent the node degree distributions of disruptively mutated genes for eight complex diseases; the last violin plot is the node

degree distribution of all genes in the human interactome (HI); the avg. degree of the disrupted genes is much greater than the avg. degree of HI, showing that highly connected genes are disrupted. (**b**) Comparison of the seed genes discovered when randomly selecting 25% of the seed gene pool as seeds. Each box plot represents the fraction of discovered seed genes across all eight diseases from DIMSUM, a GWAS based and a naïve network propagation at different iterations. DIMSUM performs significantly better than the other two approaches at the initial 50 iterations and improves drastically with increasing iterations.

Following annotation of the interactome with GWAS and functional data, we adopted a network propagation strategy to implicate protein interactions most likely to be influenced by disruptive SNVs and proposed a novel subnetwork extraction algorithm to find the mutation-specific module with the most "impact". To determine if our protocol benefited from integrating GWAS data and functional annotation into the interactome, we compared our protocol against a naïve network propagation solution on the basic human interactome, i.e., without integration GWAS or functional annotation data [42, 48]. We also compared it against another network propagation strategy with only GWAS data integrated. We used three selection ratios, 25%, 50%, and 75%, to randomly pick the number of seeds and compared the numbers of remaining seeds that could be rediscovered after applying either DIMSUM or naïve propagation. The edges were assigned with the same weight value 1 for each edge for both the naïve and GWAS based propagation approaches. After propagation, we selected genes with the highest node score to add to the module (Figure 2b and Supplementary Figure S2). The results demonstrated that our protocol had a substantially higher fraction of discovered seed genes compared to the naïve and the GWAS based network propagation strategy. As an additional experiment, we used the entire set of seeds for the network propagation in both DIMSUM and naïve approach, and then examined the top 100 discovered genes. When checking the overlap between these two gene sets, we found that the set of overlapping genes consisted of six genes on average across eight diseases. The results suggested that our method emphasized the genes with the greatest functional impact on the interactome and were not driven exclusively by the information propagated from the seeds. In other words, the genes extracted in the last step of our method indicated strong association with the disease and also reflected the severe damage caused by the mutations.

3.4. Comparison Against DIAMOnD and SCA

To validate the performance of our methods, we compared our method against two seeds-based module detection methods, DIAMOnD [26] and SCA [43]. DIseAse MOdule Detection (DIAMOnD) is one of the most popular methods for module detections. It was developed based on the observation that the connectivity significance is a more predictive quantity characterizing the module's interaction patterns, rather than connection density. The core idea of the algorithm was that, given a set of disease genes as seeds, it ranked all the candidates connected to the seeds based on their connectivity significance and added them to the existing seed set. Seed Connector Algorithm (SCA) is a recently developed seeds-based module detection method. SCA was built on the idea of seed connectors, which served as "bridges" of different network branches that were induced by seed genes. It selected a gene that maximally increased the size of the largest connected component of the subnetworks as the seed connector, and added them to the existing module.

We compiled a list of known disease-related genes for eight GWAS datasets as our benchmark (Supplementary Table S4). We then manually checked which of the added genes were supported with the literature evidence. We found that the lists of seed genes across eight diseases had between five and 29 of the disease genes supported by literature (Table 1). We next checked whether DIMSUM outperformed the other two methods in terms of the prediction accuracy in discovering the genes for our concurrence of the added gene list and the disease gene list from supported by literature. We observed that our method outperformed both DIAMOnD and SCA in all eight cases but one (osteoporosis), where DIMSUM did not find a match while both DIAMOnD and SCA found a single gene match. In fact, the predictions from DIAMOnD and SCA barely found any literature-supported disease genes (Table 1). The results also showed that except for the genes with very strong statistical

signals from GWAS studies, implicating other disease genes through a network-based approach is a challenging task. As to the agreement between the methods, the overlap of the resulted modules between every two methods is very low (Supplementary Table S5). This suggests that the algorithm design determines the way the disease module grows, and different algorithm favor different genes.

Table 1. Comparison of the number of disease associated genes in the detected modules with literature evidence between three methods: DIAMOnD, SCA and DIMSUM.

Disease ID	Disease Name	Seeds Match	DIAMOnD Match	SCA Match	DIMSUM Match
D1	Coronary artery disease	8	1	0	10
D2	Diabetes mellitus, Type 2	14	0	1	2
D3	Macular degeneration	17	0	0	3
D4	Osteoporosis	5	1	1	0
D5	Alzheimer's disease	8	0	0	3
D6	Rheumatoid arthritis	19	0	0	4
D7	Bipolar disorder	12	0	0	8
D8	Schizophrenia	29	0	0	14

We next carried out GO enrichment analysis on each gene set using all three categories of GO terms at the third level of GO hierarchy. GO annotation was then used to find how many genes from the newly obtained module shared the same GO terms with the seed genes. The results (Figure 3a) showed that DIMSUM on average yielded a higher number of GO terms compared to both DIAMOnD and SCA. The identical GO term number is further normalized with the total number of enriched terms (Supplementary Table S6). Although DIMSUM has the most enriched GO terms above the threshold, the normalized ratio is still higher than the rest two. So, taking the total number of enriched terms into account does not negate our conclusion; on the contrary, it shows that DIMSUM extracts groups of genes that are more functionally coherent. However, our method did not always have the highest number of shared GO terms for an individual disease. Interestingly, our analysis also showed that the dominant GO terms for genes extracted by all three methods fell in the Biological Process category.

Next, we examined the topological properties of the modules generated from the three methods. To quantify the structural difference of the modules generated from three methods, we focused on two topological properties. First, we calculated the connection density of the disease modules. Previous studies [26] showed that the connection density was not the primary quantity to characterize the connection patterns among disease proteins. It was further argued that, in biological networks, the paths through low-degree nodes bore stronger indications of functional similarity than the paths that went through the high-degree nodes, or hubs [49]. These findings suggested a good strategy for a module detection method should reduce the density of the detected modules and mitigate the influence of the hubs in the human interactome. When comparing these three methods, we found that DIAMOnD had the highest connection density in the detected networks, whereas modules generated from SCA and DIMSUM had much lower density (Figure 4 and Supplementary Figure S3). We also observed that SCA favored the genes with extremely high degrees, as indicated by the genes in the long tails (Figure 3b). The low density of the modules and node degree distributions from DIMSUM suggested that our method was not biased towards interaction hubs, and thus was expected to extract genes with similar functions from the rest of the interactome more efficiently than DIAMOnD.

Another interesting distinction between the methods was the fact that DIAMOnD and SCA had a tendency to grow a single giant globular component, whereas our method typically built a major component accompanied with a set of smaller, "satellite", components. In particular, we found some of the disease-associated seed genes occurring in the small satellite subnetworks from the modules obtained by DIMSUM, but not DIAMOnD or SCA. This observation suggests that the smaller subnetworks play equally important role in defining the disease phenotype.

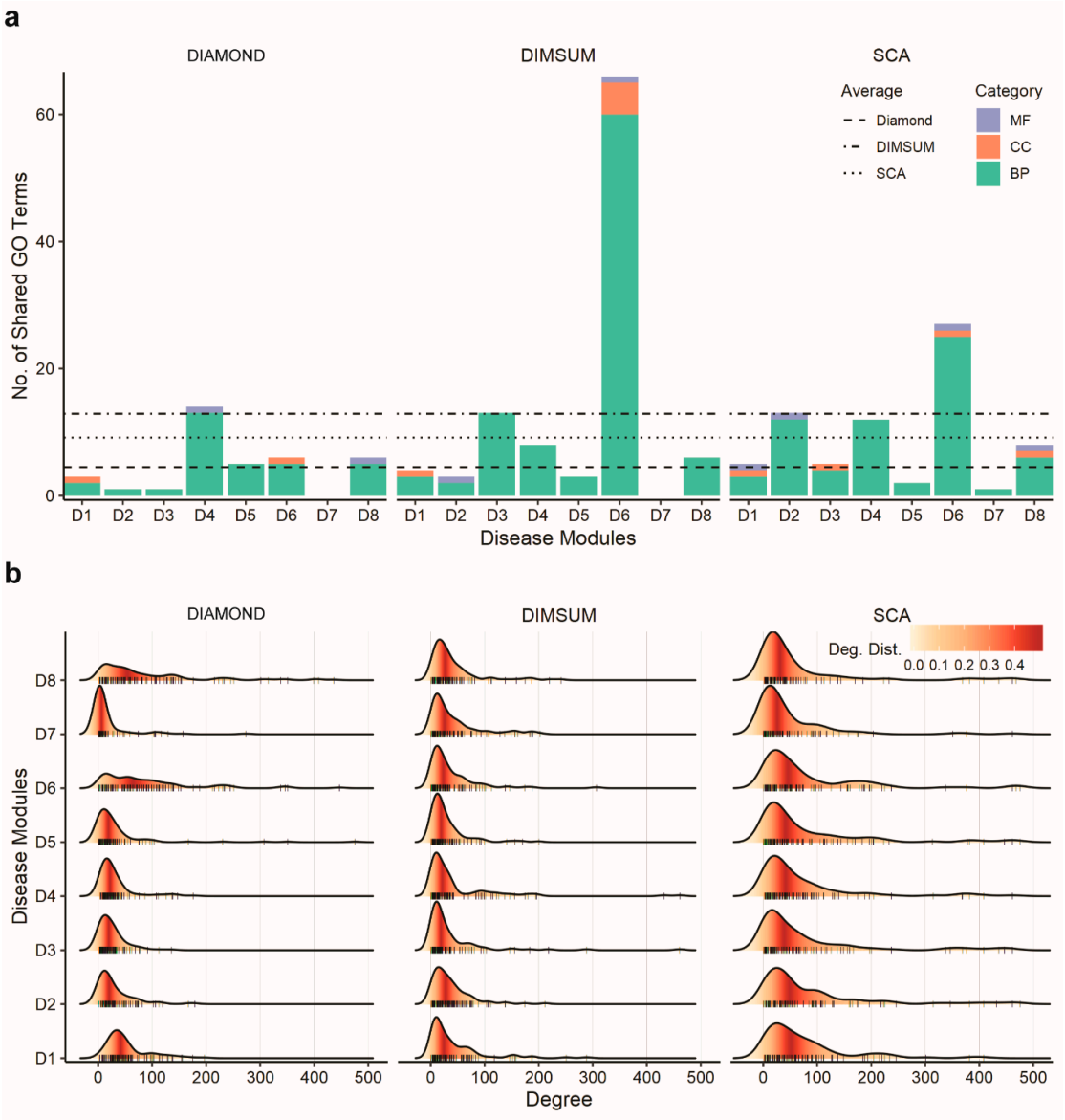

Figure 3. Comparison of biological relevance and topological properties of disease modules detected using three methods. (**a**) Bar graphs representing the number of the GO terms enriched in the added genes overlapped with the seed genes for DIAMOnD, SCA, and DIMSUM for the eight disease modules. The three GO term categories are Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). The average of significant and identical GO terms from DIMSUM is higher than from DIAMOnD or SCA. (**b**) Degree distribution of the added genes during module detection for the eight diseases by each method. When building a module, DIMSUM avoids bias towards always including the nodes of high degree and has lower node degrees for all eight modules.

3.5. Case Study 1: Coronary Artery Disease

As the first case study, we considered an application of DIMSUM to extract a network module centered around coronary artery disease (CAD) and compared the module to those ones derived by SCA and DIAMOnD using the same set of seed genes (Figure 4a,b and Supplementary Figure S3). The CAD GWAS dataset was obtained from the CARDIoGRAMplusC4D Consortium [40]. After pre-processing and applying the Pascal tool, we were able to curate a seed gene pool consisting of 37 genes, 24 of which could be mapped to the human interactome. The seed genes were spread across

the entire interactome, with very few direct interactions between each other. For each of the three methods, we extracted 100 genes in addition to the original seed gene set to form a functional module. We first validated the obtained new genes from each of the three modules against known CAD-associated genes that were collected from literature. DIMSUM outperformed both DIAMOnD and SCA (Table 1): DIAMOnD had only one and SCA reported no known CAD-related genes.

The CAD disease module from DIAMOnD formed a clique-like structure (Figure 4b). There were dense connections inside the largest connected component, the property typically not observed in a functional module [26]. Besides, the largest component originated from only seven seed genes. During the later stage of the DIAMOnD algorithm, extraction of additional genes to form the disease module was determined by several genes, including *FAM209A*, *STX1A*, and *CREB3L1*, which were not seeds and which were added in the early steps of the method run, rather than from the initial seed gene pool. We surveyed the literature and did not find a strong link between these non-seed genes and CAD. We also observed that the rest of the seed genes became isolated and separated from the largest component. On the contrary, SCA generated a globular structure for the disease module, in which the largest connected component (LCC) includes most of the seed genes. This is not surprising, as SCA is specifically designed to add a "seeds connector" to grow the LCC maximally, rather than the genes with functional importance. The addition of the seeds connector, therefore, was biased towards the hubs in the interactome. SCA tended to add many genes with high values of node degrees, as indicated by the long tails of the degree distributions. This phenomenon was not only demonstrated in the case of CAD module, but was also evident for the other diseases we studied (Figure 3b). However, recent work revealed that proteins connected to the high-degree hubs were less likely to have similar functions, compared to the proteins that interacted with a protein of significantly lower node degree [24].

Finally, when examining the disease module generated from DIMSUM we found that DIMSUM module included most of the seed CAD-related genes (Table 1 and Figure 4). In addition, there was a core component in the module set which was topologically different compared to the core components generated by DIAMOnD and SCA: it did not form a highly dense clique and it was not biased toward the hubs with very high node degrees (Figure 4c). In addition to the core component, the DIMSUM functional module included small satellite subnetworks that harbored several functionally important genes known to be associated with CAD but not reported in the seed gene pool. Three of these satellite modules contained five genes associated with CAD, namely: *PHB*, *BCAS3*, *GOSR2*, *APOH*, and *PCSK9*.

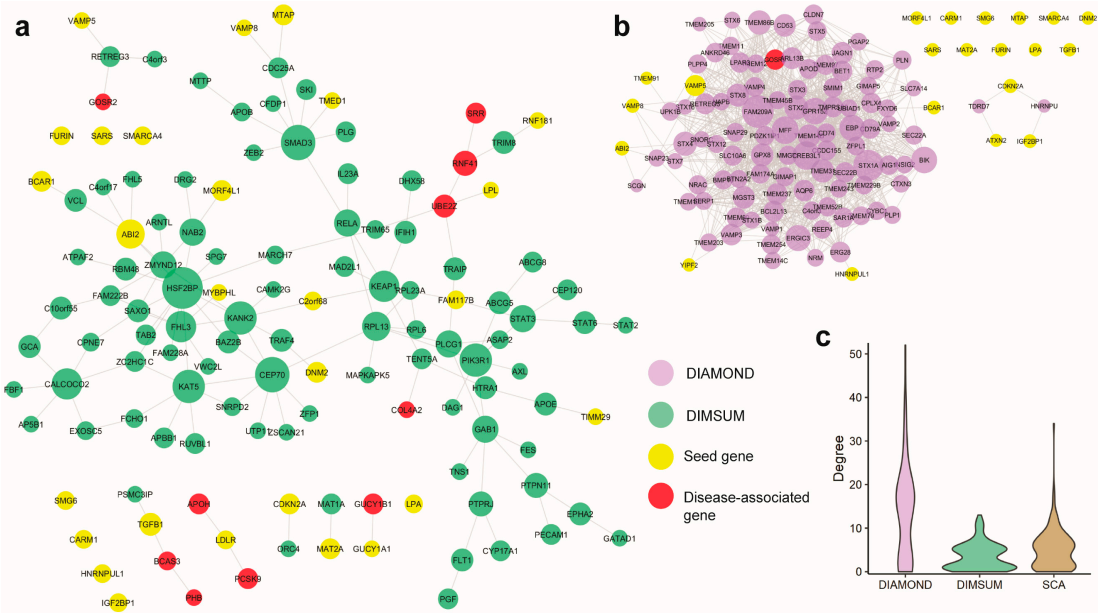

Figure 4. Comparison of Coronary Artery Disease (CAD) modules discovered by the three methods. (**a**) Largest connected component and satellite components detected by DIMSUM; yellow nodes represent the seed genes, and red nodes represent disease-associated genes that are supported by literature. (**b**) Large and high-density module detected by DIAMOnD. DIMSUM identifies ten CAD associated genes, whereas DIAMOnD identifies only one. (**c**) Degree distribution of the modules generated by each method, shows DIMSUM does not tend to grow a highly dense clique and it is not biased toward the hubs with very high node degrees.

3.6. Case Study 2: Schizophrenia and Bipolar Disorder

In the second case study, we use DIMSUM to find if two psychiatric disorders, schizophrenia and bipolar disorder, that shared symptoms also shared functional modules. Bipolar disorder (BPD), is a mental disorder, also known as manic-depressive illness, that causes unusual shifts in mood, energy and activity levels, often resulting in periods of depression or mania [50, 51]. Schizophrenia (SCZ) is a chronic and severe mental disorder that is represented by abnormal behavior and an altered notion of reality where the patients hear voices or see objects/persons that are not real [52]. While schizophrenia is not as common as other mental disorders, the symptoms can be very disabling. Schizophrenia and bipolar disorder had many common traits previously documented [53, 54].

The DIMSUM algorithm was supplied with 76 seed genes for Schizophrenia and 15 seed genes with Bipolar Disorder extracted and processed from two GWAS studies [55, 56]. There was no overlap between the seed gene sets for the two diseases. For each disease, additional 100 genes were extracted by DIMSUM to form the disease-centered module. We first queried the genes from the obtained BPD module against a list of BPD-associated genes from another recently published GWAS of the Psychiatric Genomic Consortium Bipolar Disorder Working Group [57]. As a result, we identified four genes from the module that were not among the initial set of seed genes but were found in the above GWAS study by Bipolar Disorder Working Group: *RIMS1*, *ERBB2*, *STK4*, and *MAD1L1*. These four genes have been previously shown to play functional roles in a number of neurological disorders [58-60]. For example, *RIMS1* is a *RAS* superfamily member, and the encoded protein regulates synaptic vesicle exocytosis [61]. Mutations occurring on *RIMS1* genes have been suggested to play a central role in cognition [62]. In addition to BPD, it is associated with autism spectrum disorder, neurodevelopmental disorders, and intellectual disability [58, 63].

To determine disease-associated genes in the SCZ module we relied on a recent study that categorized the disease associated genes under three tiers based on diagnosis, polygenic risk scores (PRS) and those reported by the Psychiatric Genomic Consortium (PGC) [64]. In total, we found that 33 genes among the 100 added genes were present in the PGC gene set, of which four were in Tier 1 (*CDK2AP1*, *MDK*, *ZFYVE21*, and *RRAS*). Genes in this Tier were found to be significantly associated with both diagnosis and PRS. Perhaps the most interesting of these four was *MDK*, a gene associated with many important neurological processes, e.g., cerebral cortex development, behavioral fear response, short-term memory, and regulation of behavior [65, 66]. Eight more genes belonged to Tier 2, i.e., associated with diagnosis but not PRS.

Finally, we determined if the BPD and SCZ modules shared any genes—or more importantly, submodules—in common. It had been established that bipolar disorder and schizophrenia shared a large overlap of genetic risk loci and often exhibited similar symptoms like mania and depression [54]. Thus, in spite of the missing overlap between the two seed genes sets between those disease, the modules enriched with more disease-related genes could share common genes. Intriguingly, the BPD and SCZ modules were found to share a smaller sub-module of 10 genes connected with each other and with other BPD and SCZ genes (Figure 5a). Out of 10 genes found shared between BPD and SCZ modules, four genes (*HIST1H3A*, *PBX4*, *MAD1L1*, and *NRAS*) were known to be strongly associated to both disorders (Figure 5a). We conjectured that the common genes between the BPD and SCZ modules could provide insights into the phenotypic similarities between the two diseases. To support this hypothesis, we revisited the functional predictions from the SNP-IN tool involving these ten genes. We found that while the mutations occurring in both diseases were quite diverse, a small group of mutations targeted the same subnetwork centering around *HIST1H3A* and *HIST1H4A*

(Figure 5b). Both of these genes were the core units of the nucleosome, implicated in a number of neuro-psychiatric disorders [67, 68]. Most of the mutations occurring in this subnetwork were predicted by the SNP-IN tool to disrupt the corresponding PPIs (Figure 5c, Supplementary Table S7). Thus, different mutation frequencies and combinations for BPD and SCZ could give rise to different rewiring, or edgetic, effects of the *HIST1H3A*/*HIST1H4A* centric subnetwork. These results lead us to suggest that the different rewiring patterns of the same subnetwork could explain the phenotypic similarity but underlie differences in symptomatic severity between the two diseases.

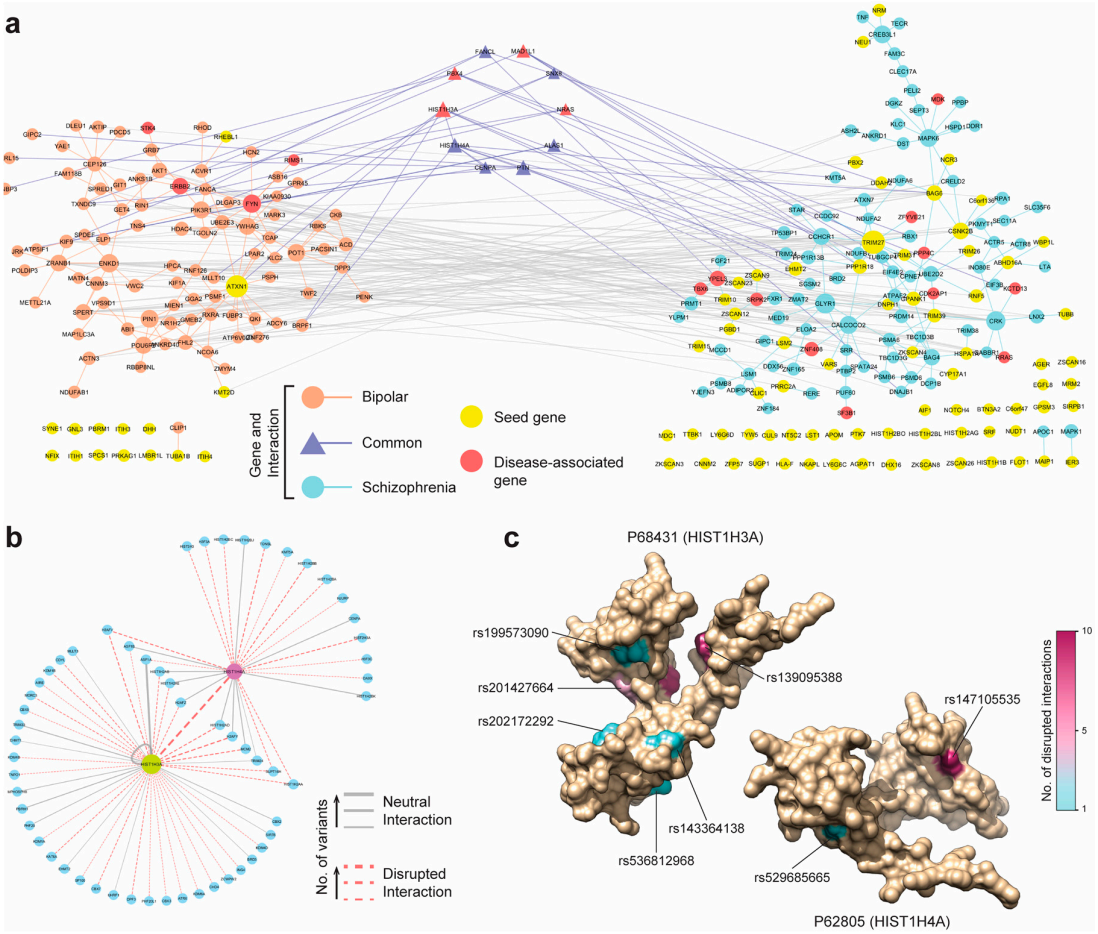

Figure 5. Analysis of Bipolar disorder and Schizophrenia modules discovered by DIMSUM. (**a**) The BPD and SCZ modules are represented by the left and right components respectively, and the genes common to both disorders are in the small central component. DIMSUM discovers a total of nineteen disease-associated genes in both modules. (**b**) The rewiring of the subnetwork centered around the shared histone genes *HIST1H3A* and *HIST1H4A*. The dash lines indicate the accumulative damage based on the SNP-IN predictions. (**c**) Protein structures for the histones HIST1H3A and HIST1H4A on the left and right respectively. Disruptive mutations occurring on these two genes are observed in both BPD and SCZ; HIST1H3A carries six disruptive mutations and HIST1H4A carries two. The color of each mutated residue corresponds to the number of interactions it disrupts.

4. Discussion

In this work, we proposed a computational framework for functional disease module detection, DIMSUM, which integrates GWAS datasets with the human interactome, propagating the functional impact of nsSNVs. Our module detection approach first annotates the network with the functional information from genes and the associated mutations, followed by the network propagation to determine new genes associated with the same disease and, finally, subnetwork extraction. We

assessed our approach using a set of eight complex diseases and comparing the performance of DIMSUM against two state-of-the-art seed-based module detection methods, DIAMOnD and SCA. The integration of multiple data types within a single computational framework allowed us to improve the effectiveness of module detection. In particular, the evaluation results showed that DIMSUM outperformed both DIAMOnD and SCA: our approach was able to yield modules with stronger disease association and greater biological relevance.

Integrating biomolecular networks across various types of data, including -omics profiles, GWAS, and functional annotation, have proven to be powerful for the detection and interpretation of biological modules [30]. GWAS investigates the entire genome and identifies the genomic loci related to a disease, providing a "macro view" of the underlying genetic architecture. On the other hand, leveraging functional annotation tools like the SNP-IN tool, enables a "micro view" by examining the specific and localized mechanistic effects of mutations and providing insights into disease etiology. Our computational framework facilitates joint interpretation of the biological information originating from those two different perspectives. Furthermore, the network propagation procedure allows to interpret the list of candidate genes into a genome-wide spectrum of gene scores, reflecting the disease association signal. This ability to amplify the signal from the seed gene pool has been previously proven helpful when identifying the genetic modules that underlie human diseases [22].

The case studies of eight complex diseases showed that the final modules obtained by DIMSUM typically include a large connected component containing most of the genes associated with a disease. This capacity of the algorithm to merge the initially isolated seed genes into a connected core component may be useful for elucidating the molecular mechanisms that are often carried out by functioning of molecular complexes and pathways, rather than isolated proteins. Furthermore, the submodules disconnected from the largest component were also found to harbor a considerable number of genes related to the disease. Examination of the discovered modules and findings from the case studies prompted us with a hypothesis that underlie the importance of the system-wide variation effects for complex genetic diseases. Specifically, we hypothesize that disease phenotypes observed in the complex diseases, such as coronary artery disease and schizophrenia, may be a consequence of rewiring of an orchestrated functional module system rather than the abnormal functioning of the independent genes. Such functional module system would consist of a core module and several smaller satellite modules. The core unit is mainly responsible for the disease, while rewiring of the smaller satellite modules could also contribute to the disease progress and disease phenotype diversity. We further hypothesize that some satellite modules could correspond to a specific symptom node in the recently proposed symptom network for the psychiatric disorders [69]. Thus, the specific rewiring of the functional module system could help explaining the disease subtypes or different symptom combinations among patients.

Supplementary Materials: The following are available online. Figure S1: Structure-based prediction of SNP's effect on PPI when applying the SNP-IN tool.; Figure S2: Comparison of DIMSUM against GWAS based and naïve network propagation procedure.; Figure S3: Coronary artery disease (CAD) module discovered by the SCA algorithm; Table S1: Description of the 8 GWAS datasets curated for this work; Table S2: Seeds generated by the Pascal tool from eight GWAS datasets of complex diseases; Table 3: SNP-IN tool annotation results for the eight GWAS datasets; Table S4: Disease-gene association data curated from OMIM, HGMD; Table S5: Overlapping genes between the discovered modules from all three methods; Table S6: Total number of enriched GO terms for eight GWAS datasets from all three methods; Table S7: HIST1H4A-HIST1H3A centered PPI subnetwork and associated disruptive mutations.

Author Contributions: H.C. conceived the idea. H.C., S.S. and D.K. designed the experiment(s), H.C. and S.S. conducted the experiment(s). H.C., S.S. and D.K. analyzed the results, and H.C., S.S. and D.K. wrote the manuscript. All authors reviewed the manuscript.

Funding: This research was funded by National Science Foundation (1458267) and National Institute of Health (LM012772-01A1) to D.K.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anderson, D.E. Ceinical characteristics of the genetic variety of cutaneous melanoma in man. Cancer **1971,** *28*, 721–725.
2. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46.
3. Ozsolak, F.; Milos, P.M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **2011**, *12*, 87–98.
4. Kolodziejczyk, A.A., Kim, J.K.; Svensson, V.; Marioni, J.C.; Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **2015**, *58*, 610–620.
5. Zeggini, E. Next-generation association studies for complex traits. *Nat. Genet.* **2011**, *43*, 287–288.
6. Cui, H.; Dhroso, A.; Johnson,N.; Korkin, D.The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods* **2015**, *79*, 18–31.
7. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
8. McCarthy, M.I.; Abecasis, G.R.; Cardon, L.R.; Goldstein, D.B.; Little, J.; Ioannidis, J.P. ; Hirschhorn, J.N. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **2008**, *9*, 356–369.
9. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719–724.
10. Gonzalez-Perez, A.; Mustonen, V.; Reva, B.; Ritchie, G.R.; Creixell, P.; Karchin, R.; Vazquez, M.; Fink, J.L.; Kassahn, K.S.; Pearson, J.V. Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **2013**, *10*, 723–729.
11. Alexander, R.P.; Fang, G.; Rozowsky, J.; Snyder, M.; Gerstein, M.B. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **2010**, *11*, 559–571.
12. Zhao, N.; Han, J.G.; Shyu, C.-R.; Korkin, D. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.* **2014**, *10*, e1003592.
13. Zhong, Q.; Simonis, N.; Li, Q.R.; Charloteaux, B.; Heuze, F.; Klitgord, N.; Tam, S.; Yu, H.; Venkatesan, K.; Mou, D. Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **2009**, *5*, 321.
14. Sahni, N.; Yi, S.; Zhong, Q.; Jailkhani, N.; Charloteaux, B.; Cusick, M.E.; Vidal, M. Edgotype: A fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* **2013**, *23*, 649–657.
15. Cui, H.; Zhao, N.; Korkin, D. Multilayer view of pathogenic SNVs in human interactome through in silico edgetic profiling. *J. Mol. Biol.* **2018**, *430*, 2974–2992.
16. Ideker, T.; Sharan, R. Protein networks in disease. *Genome Res.* **2008**, *18*, 644–652.
17. Cui, H.; Korkin, D. Effect-specific analysis of pathogenic SNVs in human interactome: Leveraging edge-based network robustness. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016.
18. Wang, X.; Wei, X.; Thijssen, B.; Das, J.; Lipkin, S.M.; Yu, H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **2012**, *30*, 159–164.
19. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–6854.
20. Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, S.J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R. A proteome-scale map of the human interactome network. *Cell* **2014**, *159*, 1212–1226.
21. Luck, K.; Kim, D.K.; Lambourne, L.; Spirohn, K.; Begg, B.E.; Bian, W.; Brignall, R.; Cafarelli, T.; Campos-Laborie, F.J.; Charloteaux, B. A reference map of the human protein interactome. *bioRxiv* **2019**, doi:10.1101/605451.

22. Cowen, L.; Ideker, T.; Raphael, B.J.; Sharan, R. Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet.* **2017**, *18*, 551–562.
23. Csermely, P.; Korcsmáros, T.; Kiss, H.J.; London, G.; Nussinov,R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol. Ther.* **2013**, *138*, 333–408.
24. Choobdar, S.; Ahsen, M.E.; Crawford, J.; Tomasoni, M.; Fang, T.; Lamparter, D.; Lin, J.; Hescott, B.; Hu, X.; Mercer. J. Assessment of network module identification across complex diseases. *Nat. Methods* **2019**, *16*, 843–852.
25. Girvan, M.; Newman, M.E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826.
26. 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.* **2015**, *11*, e1004120.
27. Tripathi, S.; Moutari, S.; Dehmer, M.; Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinform.* **2016**, *17*, 129.
28. Vlaic, S.; Conrad, T.; Tokarski-Schnelle, C.; Gustafsson, M.; Dahmen, U.; Guthke, R.; Schuster, S.ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Sci. Rep.* **2018**, *8*, 433.
29. Zhang, D.; Cui, H.; Korkin, D.; Wu, Z. Incorporation of protein binding effects into likelihood ratio test for exome sequencing data. In *BMC Proceedings*; BioMed Central: London, UK, 2016.
30. Mitra, K.; Carvunis, A.-R.; Ramesh, S.K.; Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **2013**, *14*, 719–732.
31. Das, J.; Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **2012**, *6*, 92.
32. Lamparter, D.; Marbach, D.; Rueedi, R.; Kutalik, Z.; S. Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **2016**, *12*, e1004714.
33. Wang, L.; Jia, P.; Wolfinger, R.D.; Chen, X.; Grayson, B.L.; Aune, T.M.; Zhao, Z. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics* **2011**, *27*, 686–692.
34. Consortium, G.P. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *491*, 56–65.
35. Benedix, A.; Becker, C.M.; de Groot, B.L.; Caflisch, A.; Böckmann, R.A. Predicting free energy changes using structural ensembles. *Nat. Methods* **2009**, *6*, 3–4.
36. Kamisetty, H.; Ramanathan, A.; Bailey-Kellogg, C.; Langmead, C.J. Accounting for conformational entropy in predicting binding free energies of protein-protein interactions. *Proteins Struct. Funct. Bioinform.* **2011**, *79*, 444–462.
37. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164–e164.
38. Sussman, J.L.; Lin, D.; Jiang, J.; Manning, N.O.; Prilusky, J.; Ritter, O.; Abola, E.E. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1998**, *54*, 1078–1084.
39. Mosca, R.; Ceol, A.; Stein, A.; Olivella, R. ; Aloy,P. 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **2013**, *42*, D374–D379.

40. Russell, R.B.; Alber,F.; Aloy,P.; Davis, F.P.; Korkin,D.; Pichaud, M.; Topf, M.; Sali,A. A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 313–324.
41. Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding structural details to protein networks. *Nat. Methods* **2013**, *10*, 47–53.
42. Vanunu, O.; Magger, O.; Ruppin, E.; Shlomi,T.; Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **2010**, *6*, e1000641.
43. Wang, R.-S.; Loscalzo, J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J. Mol. Biol.* **2018**, *430*, 2939–2950.
44. Stenson, P.D.; Ball, E.V.; Mort, M.; Phillips, A.D.; Shiel, J.A.; Thomas, N.S.; Abeysinghe, S.; Krawczak, M.; Cooper, D.N. Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.* **2003**, *21*, 577–581.
45. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2005**, *33* (Suppl. 1), D514–D517.
46. Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **2006**, *23*, 401–407.
47. Sahni, N.; Yi, S.; Taipale, M.; Bass, J.I.F.; Coulombe-Huntington, J.; Yang, F.; Peng, J.; Weile, J.; Karras, G.I.; Wang, Y. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **2015**, *161*, 647–660.
48. Qian, Y.; Besenbacher, S.; Mailund, T.; Schierup, M.H. Identifying disease associated genes by network propagation. In *BMC Systems Biology*; BioMed Central: London, UK, 2014.
49. Cao, M.; Zhang, H.; Park, J.; Daniels, N.M.; Crovella, M.E.; Cowen, L.J.; Hescott, B. Going the distance for protein function prediction: A new distance metric for protein interaction networks. *PLoS ONE* **2013**, *8*, e76339.
50. Craddock, N.; Sklar, P. Genetics of bipolar disorder. *Lancet* **2013**, *381*, 1654–1662.
51. Belmaker, R. Bipolar disorder. *N. Engl. J. Med.* **2004**, *351*, 476–486.
52. Fazel, S.; Gulati, G.; Linsell, L.; Geddes, J.R.; Grann, M. Schizophrenia and violence: Systematic review and meta-analysis. *PLoS Med.* **2009**, *6*, e1000120.
53. Lichtenstein, P.; Yip, B.H.; Björk, C.; Pawitan, Y.; Cannon, T.D.; Sullivan, P.F.; Hultman, C.M. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study. *Lancet* **2009**, *373*, 234–239.
54. Purcell, S.M.; Wray, N.R.; Stone, J.L.; Visscher, P.M.; O'Donovan, M.C.; Sullivan, P.F.; Sklar, P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **2009**, *460*, 748–752.
55. Ripke, S.; Sanders, A.R.; Kendler, K.S.; Levinson, D.F.; Sklar, P.; Holmans, P.A.; Lin, D.-Y.; Duan, J.; Ophoff, R.A.; Andreassen, O.A. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **2011**, *43*, 969–976.
56. Sklar, P.; Ripke, S.; Scott, L.J.; Andreassen, O.A.; Cichon, S.; Craddock, N.; Edenberg, H.J.; Nurnberger Jr, J.I.; Rietschel, M.; Blackwood, D. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **2011**, *43*, 977–983.
57. Stahl, E.A.; Breen, G.; Forstner, A.J.; McQuillin, A.; Ripke, S.; Trubetskoy, V.; Mattheisen, M.; Wang, Y.; Coleman, J.R.; Gaspar, H.A. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **2019**, *51*, 793–803.

58. Pinto, D.; Delaby, E.; Merico, D.; Barbosa, M.; Merikangas, A.; Klei, L.; Thiruvahindrapuram, B.; Xu, X.; Ziman, R.; Wang, Z. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **2014**, *94*, 677–694.
59. Pirooznia, M.; Wang, T.; Avramopoulos, D.; Potash, J.B.; Zandi, P.P.; Goes, F.S. High-throughput sequencing of the synaptome in major depressive disorder. *Mol. Psychiatry* **2016**, *21*, 650–655.
60. Fabbri, C.; Serretti, A. Pharmacogenetics of major depressive disorder: Top genes and pathways toward clinical applications. *Curr. Psychiatry Rep.* **2015**, *17*, 50.
61. Castillo, P.E.; Schoch, S.; Schmitz, F.; Südhof, T.C.; Malenka, R.C. RIM1α is required for presynaptic long-term potentiation. *Nature* **2002**, *415*, 327–330.
62. Sisodiya, S.M.; Thompson, P.J.; Need, A.; Harris, S.E.; Weale, M.E.; Wilkie, S.E.; Michaelides, M.; Free, S.L.; Walley, N.; Gumbs,C. Genetic enhancement of cognition in a kindred with cone–rod dystrophy due to RIMS1 mutation. *J. Med Genet.* **2007**, *44*, 373–380.
63. Stessman, H.A.; Xiong, B.; Coe, B.P.; Wang, T.; Hoekzema, K.; Fenckova, M.; Kvarnung, M.; Gerdts, J.; Trinh, S.; Cosemans, N. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* **2017**, *49*, 515–526.
64. Radulescu, E.; Jaffe, A.E.; Straub, R.E.; Chen, Q.; Shin, J.H.; Hyde, T.M.; Kleinman, J.E.; Weinberger, D.R. Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol. Psychiatry* **2018**, doi:10.1038/s41380-018-0304-1.
65. Winkler, C.; Yao, S. The midkine family of growth factors: Diverse roles in nervous system formation and maintenance. *Br. J. Pharmacol.* **2014**, *171*, 905–912.
66. Muramatsu, T. Midkine: A promising molecule for drug development to treat diseases of the central nervous system. *Curr. Pharm. Des.* **2011**, *17*, 410–423.
67. Rao, J.; Keleshian, V.; Klein, S.; Rapoport, S. Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients. *Transl. Psychiatry* **2012**, *2*, e132.
68. Sharma, R.P.; Rosen, C.; Kartan, S.; Guidotti, A.; Costa, E.; Grayson, D.R.; Chase, K. Valproic acid and chromatin remodeling in schizophrenia and bipolar disorder: Preliminary results from a clinical population. *Schizophr. Res.* **2006**, *88*, 227–231.
69. Borsboom, D. A network theory of mental disorders. *World Psychiatry* **2017**, *16*, 5–13.