Published in final edited form as:

Methods Mol Biol. 2017 ; 1613: 371–402. doi:10.1007/978-1-4939-7027-8_15.

Comprehensive analyses of tissue-specific networks with implications to psychiatric diseases

Guan Ning Lin¹, Roser Corominas^{1,2}, Hyun-Jun Nam¹, Jorge Urresti¹, and Lilia M. Iakoucheva^{1,*}

¹Department of Psychiatry, University of California San Diego, La Jolla, CA 92093, USA

Abstract

Recent advances in genome sequencing and “omics” technologies are opening new opportunities for improving diagnosis and treatment of human diseases. The precision medicine initiative in particular aims at developing individualized treatment options that take into account individual variability in genes and environment of each person. Systems biology approaches that group genes, transcripts and proteins into functionally meaningful networks will play crucial role in the future of personalized medicine. They will allow comparison of healthy and disease-affected tissues and organs from the same individual, as well as between healthy and disease-afflicted individuals. However, the field faces a multitude of challenges ranging from data integration to statistical and combinatorial issues in data analyses. This chapter describes computational approaches developed by us and the others to tackle challenges in tissue-specific network analyses, with the main focus on psychiatric diseases.

Keywords

Psychiatric diseases; autism; genetics; gene expression; protein-protein interactions; alternatively spliced isoforms; copy number variants; *de novo* mutations; network analyses; systems biology

1. Introduction to psychiatric disease networks

Recent large-scale genetic studies of patients and family cohorts have begun to unravel the genetic architecture of psychiatric disorders, including Autism Spectrum Disorders (ASD), Schizophrenia (SCZ) and Intellectual Disability (ID) [1]. Hundreds to thousands of genetic *loci* have been identified as putative risk factors for these diseases, with only a handful of them being strongly implicated as causative. To understand how this overwhelming number of identified genetic risk factors contributes to abnormal functioning of the brain and ultimately leads to disease phenotypes, it is necessary to adopt rigorous data-driven framework that operates at the system or network levels [2]. Over the past decade, rapid progress has been made in our understanding that biological networks formed by complex sets of interactions between numerous genes, transcripts and proteins play important role in

*Correspondence: lilyak@ucsd.edu.

²Current address: Genetics Unit, Universitat Pompeu Fabra, Hospital del Mar Research Institute (IMIM), and Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER) Instituto de Salud Carlos III, Barcelona, Spain.

deciphering disease phenotypes [3,4]. In particular, gene expression networks have been increasingly used to obtain systematic views about an immensely complex molecular landscape across brain development [5–10]. However, a dramatic increase in high-throughput experimental and computational data created a need for further improvement of effective network analytical techniques in order to unravel the molecular basis of brain disorders. This chapter describes computational approaches developed by us and the others for analyzing brain-specific biological networks related to psychiatric disorders (Figure 1). The approaches described below are generally applicable to other human diseases for which genetic, transcriptomic and protein interaction data are readily available.

2. Gene-level networks for psychiatric disorders

2.1. Construction of protein-protein interaction networks

In order to build a protein-protein interaction (PPI) network relevant to a specific disease, it is necessary to first select a set of the disease risk factors, and then obtain a set of PPIs connecting these factors. The list of disease candidate genes could be obtained either by literature curation of multiple studies with diverse sources of experimental evidence, or by extracting relevant genes from the high-throughput genetic studies, such as whole exome sequencing (WES) [11–13] or whole genome sequencing (WGS) of patients or family cohorts [14,15], or from the genome-wide association studies (GWAS) [16]. The set of PPIs for these genes can be obtained experimentally [17,18], predicted computationally [19], or alternatively downloaded from public databases such as BioGRID [20], HPRD [21], IntAct [22] and similar. The literature-curated protein interaction databases aim to aggregate all known interactions between proteins from multiple experimental sources. However, individual experiments generally focus on a selected subset of target proteins, and typically use a specific method for data collection, such as yeast two-hybrid (Y2H) system, tandem affinity purification, or co-immunoprecipitation followed by mass-spectrometry proteomics. In addition, the majority of PPIs are not collected in a tissue-specific manner. This complicates interpretation of the results relevant to the disease networks that likely operate in a tissue-specific manner. Therefore, selection of appropriate control (or background) networks for the analyses are crucial for obtaining meaningful insights into specific disease mechanisms.

2.2. Selection of control networks for the analyses

The PPIs from the public databases are intrinsically biased towards highly studied proteins, for example those implicated in cancer. These well-studied proteins accumulate more interactions than less studied ones, and consequently tend to become hubs in the PPI networks. Although hub proteins may be highly relevant to some disease networks, they may not have similar strong relevance to other disease networks. In order to minimize the biases introduced by well-studied proteins, one needs to carefully select subsets of compatible background data for network analyses in order to draw meaningful conclusions about a specific disease.

Other important sources of network biases are the intrinsic properties of the genes/proteins within the network. It has been noted that the number of interactions of a protein is

correlated with the length of a protein, with longer proteins having a greater number of interactions [23]. Thus, when using genetic data to construct and analyze networks, it is important to take into account the properties of the genes, particularly gene length and GC content [24]. It is especially important to ensure that the control networks chosen for the comparison have similar properties with the disease network under consideration. One possible way to control for biases is to limit the control or background networks to proteins that are most similar to the subset of studied proteins from the disease network, as detailed below.

2.2.1. Experimentally compatible control protein interaction networks—In order to control for experimental biases of the PPIs networks, one could limit the PPIs to only those interactions that are obtained through the high-throughput systematic screens [25,26,17]. These and similar studies identify thousands of PPIs by systematically testing large number of human proteins against the entire human ORFeome [27] using consistent experimental procedures with follow-up pair-wise verification by independent methods [28]. Although such selection strategy may decrease the number of PPIs that could be used for network construction, compared to those available from large public databases such as STRING [29] or InWeb [30], selection of unbiased PPIs would likely increase confidence and reliability of the resulting networks.

In cases, when the PPIs for building disease network were obtained through a new experimental screen, it is essential to use PPIs generated in a similar manner as a control for network analyses. In our recent work [31], we constructed an Autism Spliceform Interaction Network (ASIN) by screening over 400 isoforms of autism risk factors (baits) against ~15,000 human ORFs (preys) using a systematic yeast-two-hybrid (Y2H) screen. We used a similar, but much larger independent control dataset, the Human Interactome (HI) [17] for comparison with ASIN. HI is a set of ~14,000 high-quality binary human PPIs obtained in an unbiased reciprocal screen of ~15,000 human ORFs [17]. Both, the ASIN and the HI, share the same prey search space and were generated in the same laboratory using similar experimental pipelines. Therefore, HI serves as a perfect control network for enrichment analysis in ASIN [31].

Alternatively, in the absence of an experimental control dataset, a set of PPIs with comparable quality curated from the published literature or public databases can be used. To assemble such a dataset, one should restrict the curated data to only human high-quality interactions and preferably only those obtained using the same experimental method (e.g. binary physical interactions) [32,33]. For example, in case of ASIN, in addition to HI control network, we have also used a human binary literature-curated interaction (LCI) dataset consisting of ~40,000 PPIs. The LCI was assembled by updating a previously used PPI dataset [32] to include all newly reported PPIs from major databases with the identical filtering criteria. To verify that the control networks are comparable to ASIN, we tested for gene lengths and GC content biases and observed no statistically significant differences between the ASIN and both control datasets [31].

2.2.2. Randomized control networks—Another common strategy for analyzing disease networks when biologically comparable control networks are unavailable is to create

randomized networks and then estimate the null distribution of the test statistic (*i.e.* number of interacting partners of certain nodes, network connectivity and others). The statistics of the null distribution can be estimated from a set of randomized networks (*e.g.* 10,000 random networks with comparable properties). Such randomized networks could be generated by randomly selecting one protein of an interacting pair and replacing it with another protein from a randomly selected interaction pair, and repeating the replacement X number of times (where X is at least four times the total number of nodes in the network) to achieve complete randomization. Each randomized network preserves the degree distribution of the corresponding real network. As in all statistical hypothesis tests, the significance of a permutation test is represented by its P-value, which is estimated by the probability of obtaining an observed value at least as extreme as the test statistic given that the null hypothesis is true. In our work [31], we applied this strategy to create random control networks for assessing ASIN quality using different functional annotations (*i.e.* co-expression and co-regulation).

2.3. Comparing properties of disease and control networks

After the disease and control networks are constructed, it is sometimes desirable to compare the properties of these two networks. This could be achieved through calculating various parameters: (1) **network degree** is the number of interacting partners (*i.e.* the number of edges) of a network protein (*i.e.* a node); (2) **network shortest path** is the minimal number of protein-protein interactions (*i.e.* edges) needed to connect two nodes (*i.e.* the minimal number of edges traversed); (3) **network clustering coefficient** (C_x) calculates the degree of interconnectivity of the partners of one protein. The clustering coefficient is defined as:

$$C_x = \frac{2p_x}{(N_x(N_x - 1))}$$

where N_x is the number of neighbors of x and p_x is the number of the connected pairs between all neighbors of x , which ranges from zero (the partners of the node are not connected) to one (the partners are fully connected); and (4) **network betweenness centrality** is a measure that could be used for ranking a protein node within a network considering the number of shortest paths that pass through that node. The betweenness centrality is calculated as follows:

$$B(v) = \sum_{s=1}^n \sum_{p=1}^{s-1} \frac{n_{sd}(v)}{n_{sd}}$$

where n_{sd} is the total number of shortest paths from a protein s to a protein d , and $n_{sd}(v)$ is the number of the paths that traverse the node v . This value is then normalized by dividing the number of all possible edges between all proteins in the network (excluding s): $((n-1)(n-2))/2$. The parameters that are statistically different between the disease and carefully chosen control networks could represent biologically important findings.

2.4. Obtaining biological insights from disease networks

The disease networks could be useful in many aspects, from predicting the pathways disrupted by the disease mutations, to discovering new disease risk factors. The latter task could be accomplished by ranking binding partners (*i.e.* preys) of the confident disease-implicated proteins. The prey ranking method can be applied successfully when an unbiased control network with similar properties to the disease network is available. This method involves estimation of preys' biological connectivity. First, the number of confident disease proteins, to which a prey binds in the disease network is calculated. Then, the disease proteins are mapped to the control network, and binding enrichment is subsequently calculated using one-tailed Fisher's exact test with 5% FDR correction (Figure 2). This type of ranking is sensitive to network degree variations; therefore, the ranking should be repeated by decreasing/increasing the degree of each node tested in the disease network and for each partner in the control networks. The subset of preys consistently ranked high after applying these corrections could be potentially relevant new disease risk factors (Figure 2).

Another approach for analyzing disease networks is to compare their properties to those of the compatible control networks. Connectivity among binding partners of the disease risk factors (*i.e.* prey connectivity) could be compared using the control networks, thereby identifying preys that serve as biological linkers of the disease genes. First, the binding partners of the disease proteins from the disease network are mapped to the prey space of the control network (Figure 3). Then, the statistical significance is calculated by comparing connectivity of the binding partners in the control network with the empirical null distribution of connectivity of 10,000 sets of partners randomly selected from the control network. For example, we observed that in the background unbiased control network HI, significantly more ASIN preys were connected to each other with the shorter path length [31] (Figure 3). This suggests that ASD risk factors form a highly connected group and may converge on similar functions or processes.

2.4.1. Cell-type and tissue-specific networks—When performing network analyses of a disease affecting a particular tissue or organ, it is important to restrict the networks to only those genes/proteins that are expressed in the tissue of interest. Given the lack of tissue-specific PPI data, other methods could be utilized to construct tissue-specific networks. The results of such restrictive type of analyses are more biologically meaningful, as the PPIs identified experimentally by the *in vitro* assays in principle could occur only if both interacting genes/proteins are expressed in the specific tissue and/or developmental point.

Collaborative and consortium-level efforts, including Illumina Body Map 2.0 (GEO accession number GSE30611), ENCODE [34], and GTEx [35], collected transcriptomes of various human tissues and cell lines. Recently, a large dataset of brain transcriptome have been generated [5,36] and deposited into the BrainSpan (<http://www.brainspan.org/>). Other brain-related resources include quantification of the epigenetic landscape in CNS tissues and cell types by the Roadmap Epigenomics Mapping Consortium [37], integration of genetic variation with gene expression in the brain by the Genotype-Tissue Expression (GTEx) project [38] (<http://www.gtexportal.org/>), as well as datasets from the BrainCloud [39] and CommonMind Consortia (<http://www.commonmind.org/>). All these datasets are

applicable for the construction and analyses of the brain-specific disease networks, as described below.

2.4.2. Co-expression network analyses—Gene expression has long been used to elucidate biological and functional relationships between human genes. Gene co-expression analysis in particular was designed to identify shared patterns of expression across different experiments, tissues, or species [40–43]. Co-expression network analysis uses gene expression as a proxy for the biological and functional state of the system under investigation.

One popular approach for analyzing gene co-expression networks is to identify Topological Overlap (TO) between functional modules or subnetworks that are relevant to the disease. The module discovery using the Weighted Gene Co-expression Network Analysis (WGCNA) package [44] is a widely used method for this purpose. WGCNA aims to identify modules of genes that are highly correlated based on their expression patterns, with the goal to facilitate the identification of potential therapeutic targets or biomarkers. The usefulness of this method has been widely demonstrated as evidenced by multiple publications including diseases such as cancer, Alzheimer, and ASD among many others [45,46,10].

In order to perform WGCNA analyses, the pair-wise correlation matrix is computed for each gene pair, and an adjacency matrix is calculated by raising the correlation matrix to a power of 10 using the scale-free topology criterion. Modules are defined as branches of the clustering tree and are characterized based on the expression of the module eigengene (ME), or the first principle component of the module. To obtain moderately large and distinct modules, the minimum module size could be set at 5 genes and the minimum height for merging modules at 0.25. Genes are then assigned to a module if they have a high module membership ($kME > 0.7$). In ASIN, we have applied the WGCNA method to our PPI network genes and successfully detected five modules, with two of them enriched in brain-relevant functions [31] (Figure 4).

2.4.3. Integration of co-expression and protein interaction networks—Despite several recent efforts to generate genome-wide complete interactomes [17,18], the PPI datasets are still largely incomplete and lack tissue specificity, especially with regards to the brain tissue. In the absence brain-specific PPIs, one approach for building brain-relevant networks is to integrate brain-specific RNA expression (or co-expression) data with the PPIs [9]. The aim for integrating transcriptional (RNA-seq) and translational (PPIs) information is to render the networks more biologically relevant and dynamic, by restricting PPIs to only those that have a higher probability to occur within the brain tissue. Since two types of independent lines of evidence are used for network construction, the Pearson Correlation Coefficient ($PCC \geq 0.5$) could be used to identify co-expressed gene pairs. Then, the edges between two genes/proteins are drawn only when these two nodes are co-expressed ($PCC \geq 0.5$), and there is also a reported binary PPI between these nodes. More stringent PCC cut-offs (0.7–0.9) could also be used if desired.

Using this strategy, we have recently generated dynamic spatio-temporal networks by integrating gene expression data from the developing human brain (obtained from

BrainSpan) with the PPIs for genes from different Copy Number Variants (CNVs) implicated in various psychiatric diseases [9]. Based on the constructed networks, we observed that different CNVs showed distinct spatio-temporal signatures. This suggested that the dynamic CNV networks can reveal changes during brain development.

2.4.4. Integration of GO annotations with protein interaction networks—Another way to render networks more biologically relevant is by adding functional information extracted from public databases such as Gene Ontology (GO), KEGG Pathways, and others. Although the information about biological processes, functions or pathways is incomplete and could only be applied to a limited number of genes, using the currently available annotations together with the PPIs has a potential to identify novel gene functions, modules and connections.

One interesting strategy that we used in our work is to utilize GO Biological Process annotations for identification of functional modules within networks, and then integrate PPI information to construct co-GO networks [31]. GO database can be downloaded from the GO website (<http://geneontology.org/page/lead-database-downloads>), and we suggest to exclude inferred from Electronic Annotation (IEA) entries as less confident since they were not annotated by a human curator. Three GO branches, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC), could be used for the analyses. GO annotations should first be filtered based on information content (IC). The IC of a GO term t is defined as:

$$IC(t) = -\ln[p(t)]$$

where $p(t)$ is the fraction of genes annotated with term t or its descendants. GO terms with $IC < 0.95$ (*i.e.* those shared by more than 5% of all the annotated genes in one GO branch) should be discarded to avoid the “shallow annotation problem”. After filtering, G-SESAME method [47] could be implemented to calculate the similarity score of gene pairs in each GO branch. Once the functional modules are defined based on GO, the PPI information should be combined to either detect novel connections or to support existing functional links.

We applied this strategy to the ASIN network [31] with the aim to test the functionality of the autism risk factors that we have initially selected for the analyses. We investigated how known autism risk factors are functionally related, and whether adding newly discovered PPIs could provide new functional insights. We constructed the co-GO networks as described earlier, and the majority of the candidate genes were grouped into three functional modules (Figure 5). The new PPI data provided novel, previously unknown links for connecting these modules (Figure 5).

2.4.5. Functional network enrichment using independent sources of evidence

—In addition to general information extracted from public databases such as gene expression, PPIs, GO and others, the independent sources of information that are specific to a particular disease and are curated by the expert investigator could also be incorporated into the networks. The expertise of the investigator in a specific disease is a crucial part of this process.

With the advent of whole exome (WES) and whole genome (WGS) sequencing technologies, genetic data for many psychiatric and other human disorders is becoming available. Integrating genetic data into the networks and testing for genetic enrichment of the disease networks in genes mutated in the patients is a feasible way to identify key functional modules or disease pathways and to prioritize disease genes.

For example, when analyzing autism disease networks [31], we used the following lines of evidence to perform the enrichment analyses: (1) genes present in the *de novo* autism CNVs; (2) genes impacted by the *de novo* mutations from autism patients; (3) genes preferentially expressed in the brain; (4) post-synaptic density (PSD) genes [48]; (5) genes annotated with psychiatric phenotypes in the Online Mendelian Inheritance of Men (OMIM) database; (6) Fragile X Mental Retardation Protein (FMRP) targets [49]; (7) genes that are sensitive to mutations [50]; and (7) neuronal marker genes [51]. It is important to verify that there is no correlation between multiple sources of evidence, by plotting the annotated properties against each other and testing for correlation (*e.g.*, calculating correlations between different sources of evidence as previously shown [31]). Using the described strategy we were able not only better understand functional relevance of the autism network that we have built, but also predict new protein partners that may be relevant to this disease.

3. CNV-level networks for psychiatric disorders

Copy number variants comprising large deletions and duplications that span hundreds of thousands of DNA bases are the most widespread structural variation in the human genome [52]. Multiple studies have consistently demonstrated that CNVs play a major role in psychiatric disorders [53]. *De novo* CNV discovery and CNV burden studies have been successful in identifying precise genomic regions and even individual CNV genes conferring high risk for multiple neuropsychiatric disorders [54–60]. The CNV-level networks integrated with protein interaction networks have been previously used by us and others to investigate the mechanisms of psychiatric diseases [59,31,9]. The first step is to assemble the list of the disease-relevant CNVs defined by the genomic coordinates, and then to extract the list of genes spanning these CNVs' boundaries. The next step is to identify functional links among the CNV genes, which could be represented by co-expression, protein interactions or other functional measures.

3.1. Co-expression networks of CNV genes

Gene co-expression can be used to investigate how different CNVs conferring high risk for psychiatric disorders are related to each other in the context of brain development. The knowledge of when and where CNV genes are co-expressed in the brain will help to find the potential convergence points of different CNVs and to identify convergent molecular pathways that are disrupted by different CNVs. We used this approach to investigate 11 CNVs, each strongly implicated in more than one neuropsychiatric disorder (ASD, SCZ, ID or bipolar disorder). Using 169 genes from 11 CNVs, we constructed CNV-CNV co-expression networks using BrainSpan gene expression data [5] (Figure 6). Co-expressed gene pairs were defined as those with the pairwise PCC of ≥ 0.7 . The weight of the edges connecting different CNV nodes was calculated as a fraction of co-expressed gene pairs

between CNVs normalized by the total fraction of co-expressed gene pairs between all CNVs within the network. Specifically, the edge weight ($W_{a,b}$) is defined as:

$$W_{a,b} = \left(\frac{C_{a,b}}{T_{a,b}} \right) / \left(\frac{C_{\text{all}}}{T_{\text{all}}} \right)$$

$C_{a,b}$: The number of co-expressed gene pairs between CNV a and CNV b

C_{all} : The number of co-expressed gene pairs among all CNVs

$T_{a,b}$: The number of total gene pairs between CNV a and CNV b

T_{all} : The number of total gene pairs among all CNVs

The CNV-CNV co-expression networks demonstrate period-specific connectivity among different CNVs. For example, in the brain developmental period P2 (early fetal: 8–9 post conception weeks), 2p16.3 CNV (spanning *NRXN1* gene) has strong connections with many other CNVs, e.g. 15q12.2, 15q13.3 and 7q11.23 (Figure 6). However, the same CNV has only one strong connection with 15q11.2–13.1 in the brain developmental period P8 (Neonatal-early infancy: 0–5 months). Thus, the period-specific connectivity can provide insight into functional relationships between CNV genes implicated in multiple psychiatric disorders.

3.1.1. CNV genes prioritization using co-expression—The CNVs have been implicated as causative mutations for multiple psychiatric disorders, and some of them have demonstrated phenotypic effect in humans and model organisms. For example, the 16p11.2 CNV is associated with macrocephaly in deletion carriers and with microcephaly in duplication carriers in humans [61] and model organisms [62,63]. However, the mechanisms by which gene dosage changes within CNVs cause these diseases and specific phenotypes are still unknown in majority of cases.

One previous study investigated transcriptomic effect of CNVs in human lymphoblastoid cell lines (LCLs) from different human populations [64]. The authors demonstrated that the changes of gene copy number within CNVs explained approximately 20% of transcriptional variation [64]. Other studies [65,66] investigated the transcriptional changes caused by the deletion and duplication of 16p11.2 and 22q11.2 CNVs, respectively. They show that CNV genes affected by the deletion demonstrate ~30% expression level decrease, whereas those affected by the duplication demonstrate ~30% increase. Therefore, one could expect *cis*-effect of CNVs on expression to be approximately at the level of ~20–30%, usually in the same direction as CNV gene dosage changes. This approximation could be used to simulate

the effect of CNVs on gene expression in order to prioritize the genes within CNVs that are most sensitive to dosage changes. Below, we simulated expression levels of 169 genes within 11 high-risk CNVs implicated in psychiatric disorders (Figure 7a).

To estimate relative expression level changes of CNV genes, we first calculated the percentile rank of the expression levels of each CNV gene by comparing its expression with expression levels of all human genes during 12 periods of brain development (Figure 7b). For example, *MAPK3* and *KCTD13*, two genes within the 16p11.2 CNV, are highly expressed in the brain with more than 80th percentile rank in most brain developmental periods. In contrast, *CHDIL* and *GJA5*, two genes within the 1q21.1 CNV, are lowly expressed in the brain, with less than 20th percentile rank (Figure 7c). After the simulation of expression level changes due to deletion or duplication (~30% difference from the original values), the percentile ranks of these two genes were largely reduced or increased due to deletion or duplication, respectively (Figure 7d). These opposing effects of CNVs on expression may be different for different CNV genes, depending on their initial level of expression under the normal copy number conditions. Although the ranking of the CNV genes based on a specified 80th and 20th percentile threshold is quite arbitrary, it is reasonable to speculate that highly and lowly expressed genes would be influenced by the deletions and duplications to a different degree. When more experimental gene expression data for either CNV carriers or CRISPR cell lines with specific CNVs becomes available, it would be possible to model transcriptomic effects of different CNVs and to prioritize genes that have the largest *cis*- and even *trans*-effects. The availability of such models will improve our understanding of the functional mechanisms behind high-risk CNVs implicated in psychiatric disorders.

3.1.2. CNV genes prioritization using expression data from patients and controls—One caveat that greatly impairs studies of psychiatric diseases is inaccessibility of the brain tissues for molecular investigation. The availability of postmortem brain tissues from the carriers with specific genetic mutations is also scarce, especially for diseases with early onset such as autism. Therefore, the investigators frequently rely on the peripheral tissues (*i.e.* blood) for gene expression studies. We have used gene expression dataset derived from lymphoblast cell lines (LCLs) of autism patients and controls (GSE37772) [65] to identify gene pairs that are highly co-expressed and interacting at the protein level with the genes from the 16p11.2 CNV deletion and duplication carriers [9]. To reduce noise, only the probes with evidence of robust expression (detection $p < 0.05$ in at least 50% of 439 samples) were used for this study.

Briefly, for each network pair, a list of ‘partner-alike genes’ was assembled by selecting genes that are highly co-expressed ($SCC > 0.7$) with each partner of the 16p11.2 genes in the healthy control siblings (Figure 8). Then, the expression profiles of the ‘partner-alike genes’ from the control siblings were compared with those for the same ‘partner-alike genes’ from the 16p11.2 deletion and duplication carriers. The network pair was considered to have a significantly reduced expression in the deletion carriers, or a significantly increased expression in the duplication carriers, if its co-expression correlation coefficient was lower than 5th percentile or higher than 95th percentile of the background distribution of the correlation coefficients for the ‘partner-alike genes’ (Figure 8). We observed that some

highly co-expressed interacting pairs, including *KCTD13-Cul3*, have a significantly reduced expression in the lymphoblasts of deletion carriers, whereas other pairs have a significantly increased expression in the duplication carriers. This allowed identification of gene pairs that are dysregulated by the 16p11.2 CNV deletions and duplications in the ASD patients [9].

3.2. CNV-level protein interaction networks

The CNV-level networks could provide additional insights into functional relationships between different mutations involved in human diseases. In addition to the gene-level networks, individual genes from the same CNV could be combined into the same node and the CNV networks could be constructed (Figure 1). Various types of functional data could be used to draw the edges between different CNVs. Below, we describe the construction of the CNV networks using protein interaction data derived from the Y2H studies.

To generate CNV-level protein interaction network, each protein-encoding gene of interest is first mapped to either a specific CNV (for example, the CNV implicated in a disease), or to a non-CNV region. The genes that overlap CNVs are denoted as “baits”, and their non-CNV interacting partners as “preys”. Two types of CNV networks can be built: (1) CNV-prey networks, with CNV nodes as baits and interacting partners as preys; and (2) CNV-CNV networks, with CNV nodes as both, baits and preys.

To build a CNV-prey network, all baits from the same CNVs are merged to create a new network CNV node, and all edges become links between the CNV node and interacting prey proteins. This type of network allows identification of new disease risk factors that link more than expected number of disease CNVs. In order to generate compatible background network for enrichment analysis, random genomic regions need to be generated to mimic real CNVs. Two alternative randomization procedures of human genomic regions can be used: (1) either preserving the genomic size of CNVs and the total number of interactions for each CNV; or (2) by randomly selecting genomic regions (*i.e.* random CNVs) with the same number of genes (controlling for gene length and GC content) and similar interaction degree distribution as observed in the real CNV network. Likewise, the CNV-CNV network is constructed by merging all network genes into the new CNV nodes and preserving only edges between different CNV nodes. Next, the null distribution of the test statistic can be estimated by generating a large set of simulated CNV networks (~10,000).

3.3. Biological applications for the CNV-level networks

In our hands, the CNV-prey network has been a useful tool for detecting new autism risk factors [31]. By generating the CNV-prey network, we were able to identify 25 network preys that bind to at least two different *de novo* CNVs, with some of them linking 5 or even 6 different CNVs. The null distribution of 10,000 random CNV-prey networks with simulated CNVs could find no random networks with any prey binding to more than 4 CNVs. The preys that connected significantly more CNVs than expected by chance represent potentially new autism risk factors [31]. In the same study we also generated CNV-CNV network that directly connected 27 *de novo* autism CNVs into a single connected component. The largest connected component of the control network generated from randomly simulated CNVs with the same number of genes and PPIs as the real CNV

network comprised only 8 CNVs, a significantly smaller number than in the real network (Figure 9). The autism CNV-CNV network connected several individually rare autism CNVs with each other at the protein level, thereby pointing towards common molecular networks shared among different ASD patients.

4. Isoform-level networks for psychiatric disorders

4.1. Introduction to isoform-isoform networks

The majority of the multi-exon human genes undergo alternative splicing or use alternative promoters to increase proteomic diversity [67,68]. The brain, in particular, is one of the most complex tissues with finely regulated mechanisms of alternative splicing [69,70,35]. As a result, it is important to consider the diversity of interaction patterns of protein variants encoded by different isoforms of the same gene. Until now, only a limited number of studies, including one by us, have considered this variability in interaction networks, and these studies were also based on various subsets of human genes [71,72,31]. Recently, we have performed a global systematic screen of over a thousand of protein isoforms for interactions [73]. We demonstrated that the majority of isoform pairs, encoded by the same gene, share less than 50% of their interactions. In the global context of interactome network maps, alternative isoforms tend to behave like distinct proteins rather than minor variants of each other. We also showed that interaction partners specific to alternative isoforms tend to be expressed in a highly tissue-specific manner and belong to distinct functional modules [73].

To build an isoform-level PPI network, it is necessary to obtain isoform-level experimental data. Public databases are still scarce in any type of information at the isoform level, with the most comprehensive database recently designed by us (see <http://isoform.dfci.harvard.edu/>). Alternatively, the data could be generated experimentally for a selected subset of disease gene candidates by cloning their splicing isoforms from a relevant tissue and then testing them for PPIs either with the Y2H system [31], or by using cloned isoforms for pull-down experiments combined with mass-spectrometry [74]. It should be noted, however, that the latter approach does not provide binary interaction information that is critical for building accurate isoform-level networks.

While building isoform-level networks using Y2H, it is crucial to plan the experiments in a way that minimizes false positive and false negative interactions. In our recent study [73], each cloned isoform of a gene was tested for interactions against the entire human ORFeome (first-pass, ~15,000 ORFs), and subsequently each isoform was retested in triplicate against a union of all interaction partners for all isoforms of the same gene from the first-pass. This ensured that the observed lack of interaction for a specific isoform is a true negative event.

Using the described methods, it is possible to construct two types of isoform-level networks. First, if isoforms of only disease candidates (i.e. baits) are available for testing, then the isoform-level network would be unidirectional. However, if the isoform clones are also available for the interaction partners (i.e. preys), then bidirectional isoform networks could be constructed (Figure 10).

We observed that having isoform-level information for both interacting partners increases the number of interactions and the level of depth and complexity of the networks. For example, when we included the isoform-level PPI information in our autism network, it was expanded by 30% compared to the network constructed from only a single reference isoform of each gene [31].

4.2. Evaluating interaction profile similarity of different isoforms encoded by the same gene

More than 90% of multi-exon human genes encode multiple isoforms. These isoforms may share and/or have unique interacting partners at the protein level. The differences in interaction profiles of isoforms can be calculated using the Jaccard distance. This value measures the dissimilarity between sample sets, *i.e.* the differences in interaction profiles of multiple isoforms of a protein (Figure 11). The Jaccard distance $D_J(A,B)$ is calculated as follows:

$$D_J(A,B)=1 - J(A,B)=1 - \frac{|A \cup B| - (A \cap B)}{|A \cup B|}$$

where $J(A,B)$ is the Jaccard Score. When all isoforms interact with exactly the same partners, the Jaccard distance is 0; alternatively, when isoforms share no interacting partners the Jaccard distance is 1. In our recent study we observed that ~16% of the isoform pairs shared no interacting partners, whereas ~63% shared less than half of their interacting partners [73]. This suggests that different protein isoforms encoded by the same gene have strikingly different functional properties, and that they can behave like completely different proteins in interaction networks. We also observed that different isoforms can participate in different cellular pathways, thereby further emphasizing the importance of isoform-networks for improving our understanding of human disease pathways.

4.3. Isoform-level networks of co-expressed and interacting proteins

Many of the isoform variants are differentially regulated, and their expression is often restricted to certain organs, tissues or cells [68]. The differences in isoform expression have also been observed across tissues or organs development [75]. A given isoform may even exhibit dominant negative effects over other isoforms encoded by the same gene, be up- or down-regulated instead of being constitutively active, or even have opposing cellular functions [76]. Traditionally, and similar to protein interaction networks, the isoform-level expression data are less abundant in the public databases than the gene-level data. The majority of previous publications that use isoform-level expression data are typically only focused on a selected subset of genes of interested [77,78].

Although both, isoform-level PPI and isoform-level expression data, started to accumulate during last couple of years, the direct integration of these two types of datasets to create the isoform-level networks of co-expressed and physically interacting proteins still pose a great challenge. This is mainly due to the fact that these datasets are usually generated separately in different laboratories, and often different sets of isoforms are selected for the experiments.

4.3.1. Integration of isoform-level co-expression and protein interaction data—

As a feasible alternative to collecting both types of data, one can computationally estimate the expression levels of isoforms by quantifying it from the available tissue-specific RNA-Seq data, extracted from the Human Body Map (GEO accession number GSE30611), BrainSpan [5] (<http://www.brainspan.org/>) or GTEx [35] (<http://www.gtexportal.org/>). The dataset of known isoforms for each gene could be assembled from public databases, such as RefSeq, GenCode, UCSC known genes, and others.

When experimental isoform-level PPI networks are available for the genes of interest [31,73], one can first computationally quantify the expression levels of the network isoforms using tissue-specific RNA-seq data and the tools, such as TopHat [79], RSEM [80] or eXpress [81]. Next, co-expression between all interacting isoforms can be calculated by a correlation measurement (*i.e.* Pearson's correlation coefficient, Spearman's rank correlation coefficient, Mutual information or Euclidean distance). Alternatively, in the absence of the isoform-level PPI network for the genes of interest, one can start with a literature-curated gene-level PPI network. Then, all gene pairs in the network could be expanded to include all possible isoform pairs. Since isoform expression levels could be quantified using RNA-seq data, the isoform-level co-expression network can be generated by calculating co-expression coefficient for all isoform pairs. Typically, we connect a pair of isoforms with an edge in the network when co-expression coefficient between them is ≥ 0.5 (or ≥ 0.7 if a more stringent network is desired), and if they are also shown to physically interact (*i.e.* the PPI is reported). Figure 12 shows the example of two different isoforms of a *CTBP1* gene that are co-expressed and interacting with different isoforms of four other genes during brain development. Note that the isoform-level interaction networks change depending on the brain developmental period (P2–P13). Other examples of spatio-temporal networks are shown in our recent publication [9].

5. Summary

This chapter describes approaches to data integration that were developed over the years in our laboratory for building and analyzing disease-relevant networks. Heterogeneous data sources that include genetic, gene expression, GO, pathway annotations, disease variants (DNMs, CNV) could be used to build disease networks. Here, we used three levels of data abstraction (gene, CNV and isoform) (Figure 1) to demonstrate how to gain insights into molecular mechanisms of psychiatric diseases using tissue-specific dynamic networks. In addition, we also described how to use appropriate controls to eliminate potential biases, as well as summarized the advantages and limitations of the developed approaches.

Overall, integrating co-expression data with protein-protein interactions networks proved particularly useful in our hands for inferring disease pathways and functional modules, and leading to improved understanding of autism biology [31,9]. Other studies also used networks to discover disease mechanisms by integrating multiple data types (*e.g.*, RNA expression, genetic variation, epigenetics, or PPIs) and data sets (*e.g.*, human post-mortem brain, mouse data, *in vitro* assays) [82]. Data integration is becoming the method of choice for many human diseases, because data from diverse studies can reveal unexpected and distinct relationships between genes, RNAs and proteins that are invisible when only a single

dataset is used for the analyses. Developing reliable approaches to data integration is especially important in light of the precision medicine initiative. Large patient-derived datasets such as WGS, transcriptomic, metabolomic, and proteomic, will soon be produced with the aim of predicting disease susceptibility, or the course of disease development. Integrating these datasets is currently presenting enormous challenges that likely will be solved during the next decade of research. We hope that approaches described in this chapter will be contributing to the bright future of precision medicine.

References

- Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci*. 2014; 17(6):782–790. DOI: 10.1038/nn.3708 [PubMed: 24866044]
- Parikshak NN, Gandal MJ, Geschwind DH. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat Rev Genet*. 2015; 16(8):441–458. <http://www.nature.com/nrg/journal/v16/n8/abs/nrg3934.html#supplementary-information>. DOI: 10.1038/nrg3934 [PubMed: 26149713]
- Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007; 104(21):8685–8690. [PubMed: 17502601]
- Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell*. 2011; 144(6):986–998. S0092-8674(11)00130-9 [pii]. DOI: 10.1016/j.cell.2011.02.016 [PubMed: 21414488]
- Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, Johnson MB, Krsnik Z, Mayer S, Fertuzinhos S, Umlauf S, Lisgo SN, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE, Sestan N. Spatio-temporal transcriptome of the human brain. *Nature*. 2011; 478(7370):483–489. nature10523 [pii]. DOI: 10.1038/nature10523 [PubMed: 22031440]
- Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, Reilly SK, Lin L, Fertuzinhos S, Miller JA, Murtha MT, Bichsel C, Niu W, Cotney J, Ercan-Sencicek AG, Gockley J, Gupta AR, Han W, He X, Hoffman EJ, Klei L, Lei J, Liu W, Liu L, Lu C, Xu X, Zhu Y, Mane SM, Lein ES, Wei L, Noonan JP, Roeder K, Devlin B, Sestan N, State MW. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*. 2013; 155(5):997–1007. S0092-8674(13)01296-8 [pii]. DOI: 10.1016/j.cell.2013.10.020 [PubMed: 24267886]
- Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*. 2013; 155(5):1008–1021. S0092-8674(13)01349-4 [pii]. DOI: 10.1016/j.cell.2013.10.031 [PubMed: 24267887]
- Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, Rippey C, Shahin H, Nimgaonkar VL, Go RC, Savage RM, Swerdlow NR, Gur RE, Braff DL, King MC, McClellan JM. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*. 2013; 154(3):518–529. S0092-8674(13)00831-3 [pii]. DOI: 10.1016/j.cell.2013.06.049 [PubMed: 23911319]
- Lin GN, Corominas R, Lemmens I, Yang X, Tavernier J, Hill DE, Vidal M, Sebat J, Iakoucheva LM. Spatiotemporal 16p11.2 Protein Network Implicates Cortical Late Mid-Fetal Brain Development and KCTD13-Cul3-RhoA Pathway in Psychiatric Diseases. *Neuron*. 2015; 85(4):742–754. S0896-6273(15)00036-7 [pii]. DOI: 10.1016/j.neuron.2015.01.010 [PubMed: 25695269]
- Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*. 2011; 474(7351):380–384. nature10110 [pii]. DOI: 10.1038/nature10110 [PubMed: 21614001]
- Iossifov I, O’Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paepers B, Nickerson DA, Dea J, Dong S, Gonzalez LE, Mandell JD, Mane SM, Murtha MT, Sullivan CA, Walker MF, Waqar Z, Wei L, Willsey AJ, Yamrom B, Lee YH, Grabowska E, Dalkic E, Wang Z, Marks S, Andrews P, Leotta A, Kendall J,

- Hakker I, Rosenbaum J, Ma B, Rodgers L, Troge J, Narzisi G, Yoon S, Schatz MC, Ye K, McCombie WR, Shendure J, Eichler EE, State MW, Wigler M. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*. 2014; 515(7526):216–221. [nature13908 \[pii\]](#). DOI: 10.1038/nature13908 [PubMed: 25363768]
12. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, Autism Sequencing C. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron*. 2012; 76(6):1052–1056. DOI: 10.1016/j.neuron.2012.12.008 [PubMed: 23259942]
 13. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Shih-Chen F, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell NG, Carracedo A, Chahrouh MH, Chiocchetti AG, Coon H, Crawford EL, Curran SR, Dawson G, Duketis E, Fernandez BA, Gallagher L, Geller E, Guter SJ, Hill RS, Ionita-Laza J, Jimenez Gonzalez P, Kilpinen H, Klauck SM, Kolevzon A, Lee I, Lei I, Lei J, Lehtimäki T, Lin CF, Ma'ayan A, Marshall CR, McInnes AL, Neale B, Owen MJ, Ozaki N, Parellada M, Parr JR, Purcell S, Puura K, Rajagopalan D, Rehnstrom K, Reichenberg A, Sabo A, Sachse M, Sanders SJ, Schafer C, Schulte-Ruther M, Skuse D, Stevens C, Szatmari P, Tammimies K, Valladares O, Voran A, Li-San W, Weiss LA, Willsey AJ, Yu TW, Yuen RK, Cook EH, Freitag CM, Gill M, Hultman CM, Lehner T, Palotie A, Schellenberg GD, Sklar P, State MW, Sutcliffe JS, Walsh CA, Scherer SW, Zwick ME, Barrett JC, Cutler DJ, Roeder K, Devlin B, Daly MJ, Buxbaum JD. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*. 2014; 515(7526):209–215. [nature13772 \[pii\]](#). DOI: 10.1038/nature13772 [PubMed: 25363760]
 14. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillio J, Gadamski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*. 2012; 151(7):1431–1442. S0092-8674(12)01404-3 [pii]. DOI: 10.1016/j.cell.2012.11.019 [PubMed: 23260136]
 15. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, Chrysler C, Nalpathamkalam T, Pellecchia G, Liu Y, Gazzellone MJ, D'Abate L, Deneault E, Howe JL, Liu RS, Thompson A, Zarrei M, Uddin M, Marshall CR, Ring RH, Zwaigenbaum L, Ray PN, Weksberg R, Carter MT, Fernandez BA, Roberts W, Szatmari P, Scherer SW. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat Med*. 2015; 21(2):185–191. [nm.3792 \[pii\]](#). DOI: 10.1038/nm.3792 [PubMed: 25621899]
 16. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, Mowry BJ, Thapar A, Goddard ME, Witte JS, Absher D, Agartz I, Akil H, Amin F, Andreassen OA, Anjorin A, Anney R, Anttila V, Arking DE, Asherson P, Azevedo MH, Backlund L, Badner JA, Bailey AJ, Banaschewski T, Barchas JD, Barnes MR, Barrett TB, Bass N, Battaglia A, Bauer M, Bayes M, Bellivier F, Bergen SE, Berrettini W, Betancur C, Bettecken T, Biederman J, Binder EB, Black DW, Blackwood DH, Bloss CS, Boehnke M, Boomsma DI, Breen G, Breuer R, Bruggeman R, Cormican P, Buccola NG, Buitelaar JK, Bunney WE, Buxbaum JD, Byerley WF, Byrne EM, Caesar S, Cahn W, Cantor RM, Casas M, Chakravarti A, Chambert K, Choudhury K, Cichon S, Cloninger CR, Collier DA, Cook EH, Coon H, Cormand B, Corvin A, Coryell WH, Craig DW, Craig IW, Crosbie J, Cuccaro ML, Curtis D, Czamara D, Datta S, Dawson G, Day R, De Geus EJ, Degenhardt F, Djurovic S, Donohoe GJ, Doyle AE, Duan J, Dudbridge F, Duketis E, Ebstein RP, Edenberg HJ, Elia J, Ennis S, Etain B, Fanous A, Farmer AE, Ferrier IN, Flickinger M, Fombonne E, Foroud T, Frank J, Franke B, Fraser C, Freedman R, Freimer NB, Freitag CM, Friedl M, Frisen L, Gallagher L, Gejman PV, Georgieva L, Gershon ES, Geschwind DH, Giegling I, Gill M, Gordon SD, Gordon-Smith K, Green EK, Greenwood TA, Grice DE, Gross M, Grozeva D, Guan W, Gurling H, De Haan L, Haines JL, Hakonarson H, Hallmayer J, Hamilton SP, Hamshere ML, Hansen TF, Hartmann AM, Hautzinger M, Heath AC, Henders AK, Herms S, Hickie IB, Hipolito M, Hoefels S, Holmans PA, Holsboer F, Hoogendijk WJ, Hottenga JJ, Hultman CM, Hus V, Ingason A, Ising M, Jamain S, Jones EG, Jones I, Jones L, Tzeng JY, Kahler AK, Kahn RS, Kandaswamy R, Keller MC, Kennedy JL, Kenny E, Kent L, Kim Y, Kirov GK, Klauck SM, Klei L, Knowles JA, Kohli MA, Koller DL, Konte B, Korszun A, Krabbendam L, Krasucki R, Kuntsi J, Kwan P, Landen M, Langstrom N, Lathrop M, Lawrence J, Lawson WB, Leboyer M, Ledbetter DH, Lee PH, Lencz T, Lesch KP, Levinson DF, Lewis CM, Li J, Lichtenstein P, Lieberman JA, Lin DY, Linszen DH, Liu

C, Lohoff FW, Loo SK, Lord C, Lowe JK, Lucae S, Macintyre DJ, Madden PA, Maestrini E, Magnusson PK, Mahon PB, Maier W, Malhotra AK, Mane SM, Martin CL, Martin NG, Mattheisen M, Matthews K, Mattingsdal M, McCarroll SA, McGhee KA, McGough JJ, McGrath PJ, McGuffin P, McInnis MG, McIntosh A, McKinney R, McLean AW, McMahon FJ, McMahon WM, McQuillin A, Medeiros H, Medland SE, Meier S, Melle I, Meng F, Meyer J, Middeldorp CM, Middleton L, Milanova V, Miranda A, Monaco AP, Montgomery GW, Moran JL, Moreno-De-Luca D, Morken G, Morris DW, Morrow EM, Moskvina V, Muglia P, Muhleisen TW, Muir WJ, Muller-Myhsok B, Murtha M, Myers RM, Myin-Germeys I, Neale MC, Nelson SF, Nievergelt CM, Nikolov I, Nimgaonkar V, Nolen WA, Nothen MM, Nurnberger JI, Nwulia EA, Nyholt DR, O'Dushlaine C, Oades RD, Olincy A, Oliveira G, Olsen L, Ophoff RA, Osby U, Owen MJ, Palotie A, Parr JR, Paterson AD, Pato CN, Pato MT, Penninx BW, Pergadia ML, Pericak-Vance MA, Pickard BS, Pimm J, Piven J, Posthuma D, Potash JB, Poustka F, Propping P, Puri V, Quedstedt DJ, Quinn EM, Ramos-Quiroga JA, Rasmussen HB, Raychaudhuri S, Rehnstrom K, Reif A, Ribases M, Rice JP, Rietschel M, Roeder K, Roeyers H, Rossin L, Rothenberger A, Rouleau G, Ruderfer D, Rujescu D, Sanders AR, Sanders SJ, Santangelo SL, Sergeant JA, Schachar R, Schalling M, Schatzberg AF, Scheftner WA, Schellenberg GD, Scherer SW, Schork NJ, Schulze TG, Schumacher J, Schwarz M, Scolnick E, Scott LJ, Shi J, Shilling PD, Shyn SI, Silverman JM, Slager SL, Smalley SL, Smit JH, Smith EN, Sonuga-Barke EJ, St Clair D, State M, Steffens M, Steinhausen HC, Strauss JS, Strohmaier J, Stroup TS, Sutcliffe JS, Szatmari P, Szelinger S, Thirumalai S, Thompson RC, Todorov AA, Tozzi F, Treutlein J, Uhr M, van den Oord EJ, Van Grootheest G, Van Os J, Vicente AM, Vieland VJ, Vincent JB, Visscher PM, Walsh CA, Wassink TH, Watson SJ, Weissman MM, Werge T, Wienker TF, Wijsman EM, Willemsen G, Williams N, Willsey AJ, Witt SH, Xu W, Young AH, Yu TW, Zammit S, Zandi PP, Zhang P, Zitman FG, Zollner S, Devlin B, Kelsoe JR, Sklar P, Daly MJ, O'Donovan MC, Craddock N, Sullivan PF, Smoller JW, Kendler KS, Wray NR. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet.* 2013; ng.2711 [pii]. doi: 10.1038/ng.2711

17. Rolland T, Tasan M, Charleaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruysinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K, Walsh J, Cusick ME, Xia Y, Barabasi AL, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M. A proteome-scale map of the human interactome network. *Cell.* 2014; 159(5):1212–1226. S0092-8674(14)01422-6 [pii]. DOI: 10.1016/j.cell.2014.10.050 [PubMed: 25416956]
18. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F, Hyman AA, Mann M. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell.* 2015; 163(3):712–723. DOI: 10.1016/j.cell.2015.09.053 [PubMed: 26496610]
19. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature.* 2012; 490(7421):556–560. nature11503 [pii]. DOI: 10.1038/nature11503 [PubMed: 23023127]
20. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Regulj T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 2015; 43(Database issue):D470–478. DOI: 10.1093/nar/gku1204 [PubMed: 25428363]
21. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadrans S, Chaerkady R, Pandey A. Human Protein Reference Database--2009 update. *Nucleic Acids Res.* 2009; 37(Database issue):D767–772. DOI: 10.1093/nar/gkn892 [PubMed: 18988627]

22. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014; 42(Database issue):D358–363. DOI: 10.1093/nar/gkt1115 [PubMed: 24234451]
23. Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* 2015; 16:202.doi: 10.1186/s13059-015-0772-4 [PubMed: 26424194]
24. Raychaudhuri S, Korn JM, McCarroll SA, Altshuler D, Sklar P, Purcell S, Daly MJ. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* 2010; 6(9) doi:e1001097 [pii] 10.1371/journal.pgen.1001097.
25. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature.* 2005; 437(7062):1173–1178. [PubMed: 16189514]
26. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell.* 2005; 122(6): 957–968. [PubMed: 16169070]
27. Collaboration OR. The ORFeome Collaboration: a genome-scale human ORF-clone resource. *Nat Methods.* 2016; 13(3):191–192. DOI: 10.1038/nmeth.3776 [PubMed: 26914201]
28. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M. An empirical framework for binary interactome mapping. *Nat Methods.* 2009; 6(1):83–90. [PubMed: 19060904]
29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015; 43(Database issue):D447–452. DOI: 10.1093/nar/gku1003 [PubMed: 25352553]
30. Lage K, Mollgard K, Greenway S, Wakimoto H, Gorham JM, Workman CT, Bendtsen E, Hansen NT, Rigina O, Roque FS, Wiese C, Christoffels VM, Roberts AE, Smoot LB, Pu WT, Donahoe PK, Tommerup N, Brunak S, Seidman CE, Seidman JG, Larsen LA. Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol Syst Biol.* 2010; 6:381.doi: 10.1038/msb.2010.36 [PubMed: 20571530]
31. Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, Rodriguez M, Tam S, Trigg SA, Fan C, Yi S, Tasan M, Lemmens I, Kuang X, Zhao N, Malhotra D, Michaelson JJ, Vacic V, Calderwood MA, Roth FP, Tavernier J, Horvath S, Salehi-Ashtiani K, Korkin D, Sebat J, Hill DE, Hao T, Vidal M, Iakoucheva LM. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat Commun.* 2014; 5:3650. ncomms4650 [pii]. doi: 10.1038/ncomms4650 [PubMed: 24722188]
32. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M. Literature-curated protein interaction datasets. *Nat Methods.* 2009; 6(1):39–46. [PubMed: 19116613]
33. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol.* 2012; 30(2):159–164. nbt.2106 [pii]. DOI: 10.1038/nbt.2106 [PubMed: 22252508]

34. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57–74. DOI: 10.1038/nature11247 [PubMed: 22955616]
35. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015; 348(6235):648–660. DOI: 10.1126/science.1262110 [PubMed: 25954001]
36. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, Arnold JM, Bennet C, Bertagnolli D, Brouner K, Butler S, Caldejon S, Carey A, Cuhaciyan C, Dalley RA, Dee N, Dolbeare TA, Facer BA, Feng D, Fliss TP, Gee G, Goldy J, Gourley L, Gregor BW, Gu G, Howard RE, Jochim JM, Kuan CL, Lau C, Lee CK, Lee F, Lemon TA, Lesnar P, McMurray B, Mastan N, Mosqueda N, Nalwai-Cecchini T, Ngo NK, Nyhus J, Oldre A, Olson E, Parente J, Parker PD, Parry SE, Stevens A, Pletikos M, Reding M, Roll K, Sandman D, Sarreal M, Shapouri S, Shapovalova NV, Shen EH, Sjoquist N, Slaughterbeck CR, Smith M, Sodt AJ, Williams D, Zollei L, Fischl B, Gerstein MB, Geschwind DH, Glass IA, Hawrylycz MJ, Hevner RF, Huang H, Jones AR, Knowles JA, Levitt P, Phillips JW, Sestan N, Wohnoutka P, Dang C, Bernard A, Hohmann JG, Lein ES. Transcriptional landscape of the prenatal human brain. *Nature*. 2014; 508(7495):199–206. nature13185 [pii]. DOI: 10.1038/nature13185 [PubMed: 24695229]
37. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shores N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518(7539):317–330. nature14248 [pii]. DOI: 10.1038/nature14248 [PubMed: 25693563]
38. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013; 45(6):580–585. DOI: 10.1038/ng.2653 [PubMed: 23715323]
39. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahoul AG, Herman MM, Weinberger DR, Kleinman JE. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*. 2011; 478(7370):519–523. DOI: 10.1038/nature10524 [PubMed: 22031444]
40. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003; 302(5643):249–255. 1087447 [pii]. DOI: 10.1126/science.1087447 [PubMed: 12934013]
41. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005; 4 Article17.
42. Henegar C, Tordjman J, Achard V, Lacasa D, Cremer I, Guerre-Millo M, Poitou C, Basdevant A, Stich V, Viguier N, Langin D, Bedossa P, Zucker JD, Clement K. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. *Genome Biol*. 2008; 9(1):R14.doi: 10.1186/gb-2008-9-1-r14 [PubMed: 18208606]
43. Prifti E, Zucker JD, Clement K, Henegar C. Interactional and functional centrality in transcriptional co-expression networks. *Bioinformatics*. 2010; 26(24):3083–3089. DOI: 10.1093/bioinformatics/btq591 [PubMed: 20959383]
44. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, Lee Y, Scheck AC, Liao LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A*. 2006; 103(46):17402–17407. 0608396103 [pii] 10.1073/pnas.0608396103. [PubMed: 17090670]

45. Presson AP, Yoon NK, Bagryanova L, Mah V, Alavi M, Maresh EL, Rajasekaran AK, Goodglick L, Chia D, Horvath S. Protein expression based multimarker analysis of breast cancer samples. *BMC Cancer*. 2011; 11:230.doi: 10.1186/1471-2407-11-230 [PubMed: 21651811]
46. Levine AJ, Miller JA, Shapshak P, Gelman B, Singer EJ, Hinkin CH, Commings D, Morgello S, Grant I, Horvath S. Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. *BMC Med Genomics*. 2013; 6:4.doi: 10.1186/1755-8794-6-4 [PubMed: 23406646]
47. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. *Bioinformatics*. 2007; 23(10):1274–1281. btm087 [pii] 10.1093/bioinformatics/btm087. [PubMed: 17344234]
48. Bayes A, van de Lagemaat LN, Collins MO, Croning MD, Whittle IR, Choudhary JS, Grant SG. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci*. 2011; 14(1):19–21. doi:nn.2719 [pii] 10.1038/nn.2719. [PubMed: 21170055]
49. Darnell JC, Van Driesche SJ, Zhang C, Hung KY, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW, Licatalosi DD, Richter JD, Darnell RB. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*. 2011; 146(2):247–261. doi:S0092-8674(11)00655-6 [pii] 10.1016/j.cell.2011.06.013. [PubMed: 21784246]
50. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Cook EH Jr, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ. A framework for the interpretation of de novo mutation in human disease. *Nat Genet*. 2014; 46(9):944–950. DOI: 10.1038/ng.3050 [PubMed: 25086666]
51. Pasca SP, Portmann T, Voineagu I, Yazawa M, Shcheglovitov A, Pasca AM, Cord B, Palmer TD, Chikahisa S, Nishino S, Bernstein JA, Hallmayer J, Geschwind DH, Dolmetsch RE. Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nat Med*. 2011; 17(12):1657–1662. nm.2576 [pii]. DOI: 10.1038/nm.2576 [PubMed: 22120178]
52. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. *Science*. 2004; 305(5683):525–528. [PubMed: 15273396]
53. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*. 2012; 148(6):1223–1241. S0092-8674(12)00277-2 [pii]. DOI: 10.1016/j.cell.2012.02.039 [PubMed: 22424231]
54. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, Vorstman JA, Thompson A, Regan R, Pilorge M, Pellecchia G, Pagnamenta AT, Oliveira B, Marshall CR, Magalhaes TR, Lowe JK, Howe JL, Griswold AJ, Gilbert J, Duketis E, Dombroski BA, De Jonge MV, Cuccaro M, Crawford EL, Correia CT, Conroy J, Conceicao IC, Chiocchetti AG, Casey JP, Cai G, Cabrol C, Bolshakova N, Bacchelli E, Anney R, Gallinger S, Cotterchio M, Casey G, Zwaigenbaum L, Wittemeyer K, Wing K, Wallace S, van Engeland H, Tryfon A, Thomson S, Soorya L, Roge B, Roberts W, Poustka F, Moug S, Minshew N, McInnes LA, McGrew SG, Lord C, Leboyer M, Le Couteur AS, Kolevzon A, Jimenez Gonzalez P, Jacob S, Holt R, Guter S, Green J, Green A, Gillberg C, Fernandez BA, Duque F, Delorme R, Dawson G, Chaste P, Caffe C, Brennan S, Bourgeron T, Bolton PF, Bolte S, Bernier R, Baird G, Bailey AJ, Anagnostou E, Almeida J, Wijsman EM, Veland VJ, Vicente AM, Schellenberg GD, Pericak-Vance M, Paterson AD, Parr JR, Oliveira G, Nurnberger JI, Monaco AP, Maestrini E, Klauck SM, Hakonarson H, Haines JL, Geschwind DH, Freitag CM, Folstein SE, Ennis S, Coon H, Battaglia A, Szatmari P, Sutcliffe JS, Hallmayer J, Gill M, Cook EH, Buxbaum JD, Devlin B, Gallagher L, Betancur C, Scherer SW. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am J Hum Genet*. 2014; 94(5):677–694. DOI: 10.1016/j.ajhg.2014.03.018 [PubMed: 24768552]
55. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Rocanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. Rare structural variants disrupt multiple genes in

- neurodevelopmental pathways in schizophrenia. *Science*. 2008; 320(5875):539–543. [PubMed: 18369103]
56. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH Jr, Dobyns WB, Christian SL. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet*. 2008; 17(4):628–638. [PubMed: 18156158]
 57. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. De novo rates and selection of large copy number variation. *Genome Res*. 2010; 20(11):1469–1481. DOI: 10.1101/gr.107680.110 [PubMed: 20841430]
 58. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, Buja A, Krieger A, Yoon S, Troge J, Rodgers L, Iossifov I, Wigler M. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*. 2011; 70(5):886–897. doi:S0896-6273(11)00396-5 [pii] 10.1016/j.neuron.2011.05.015. [PubMed: 21658582]
 59. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*. 2011; 70(5):898–907. doi:S0896-6273(11)00439-9 [pii] 10.1016/j.neuron.2011.05.021. [PubMed: 21658583]
 60. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney JF 3rd, Klei L, Mandell JD, Moreno-De-Luca D, Poultnery CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, Lowe JK, Mane SM, Martin DM, Morrow EM, Talkowski ME, Sutcliffe JS, Walsh CA, Yu TW, Autism Sequencing C, Ledbetter DH, Martin CL, Cook EH, Buxbaum JD, Daly MJ, Devlin B, Roeder K, State MW. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*. 2015; 87(6):1215–1233. DOI: 10.1016/j.neuron.2015.09.016 [PubMed: 26402605]
 61. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Rocanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimäki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nothen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet*. 2009; 41(11):1223–1227. doi:ng.474 [pii] 10.1038/ng.474. [PubMed: 19855392]
 62. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, Reymond A, Sun M, Sawa A, Gusella JF, Kamiya A, Beckmann JS, Katsanis N. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature*. 2012; 485(7398):363–367. nature11091 [pii]. DOI: 10.1038/nature11091 [PubMed: 22596160]
 63. Horev G, Ellegood J, Lerch JP, Son YE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M, Mills AA. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci U S A*. 2011; 108(41):17076–17081. 1114042108 [pii]. DOI: 10.1073/pnas.1114042108 [PubMed: 21969575]
 64. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007; 315(5813):848–853. DOI: 10.1126/science.1136678 [PubMed: 17289997]
 65. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, Cai C, Ou J, Lowe JK, Hurles ME, Devlin B, State MW, Geschwind DH. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet*. 2012; 91(1):38–55. S0002-9297(12)00267-4 [pii]. DOI: 10.1016/j.ajhg.2012.05.011 [PubMed: 22726847]

66. Rees E, Kirov G, Sanders A, Walters JT, Chambert KD, Shi J, Szatkiewicz J, O'Dushlaine C, Richards AL, Green EK, Jones I, Davies G, Legge SE, Moran JL, Pato C, Pato M, Genovese G, Levinson D, Duan J, Moy W, Goring HH, Morris D, Cormican P, Kendler KS, O'Neill FA, Riley B, Gill M, Corvin A, Wellcome Trust Case Control C, Craddock N, Sklar P, Hultman C, Sullivan PF, Gejman PV, McCarroll SA, O'Donovan MC, Owen MJ. Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol Psychiatry*. 2014; 19(1):37–40. DOI: 10.1038/mp.2013.156 [PubMed: 24217254]
67. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 2008; 40(12):1413–1415. doi:ng.259 [pii] 10.1038/ng.259. [PubMed: 18978789]
68. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456(7221):470–476. [PubMed: 18978772]
69. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biol*. 2004; 5(10):R74. doi: 10.1186/gb-2004-5-10-r74 [PubMed: 15461793]
70. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, Johnson R, Segre AV, Djebali S, Niarchou A, Wright FA, Lappalainen T, Calvo M, Getz G, Dermizakis ET, Ardlie KG, Guigo R. Human genomics. The human transcriptome across tissues and individuals. *Science*. 2015; 348(6235):660–665. 348/6235/660 [pii]. DOI: 10.1126/science.aaa0355 [PubMed: 25954002]
71. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*. 2012; 46(6):871–883. doi:S1097-2765(12)00484-4 [pii] 10.1016/j.molcel.2012.05.039. [PubMed: 22749400]
72. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, Wrana JL, Blencowe BJ. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*. 2012; 46(6):884–892. doi:S1097-2765(12)00482-0 [pii] 10.1016/j.molcel.2012.05.037. [PubMed: 22749401]
73. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, Spirohn K, Begg BE, Duran-Frigola M, MacWilliams A, Pevzner SJ, Zhong Q, Trigg SA, Tam S, Ghamsari L, Sahni N, Yi S, Rodriguez MD, Balcha D, Tan G, Costanzo M, Andrews B, Boone C, Zhou XJ, Salehi-Ashtiani K, Charloteaux B, Chen AA, Calderwood MA, Aloy P, Roth FP, Hill DE, Iakoucheva LM, Xia Y, Vidal M. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*. 2016; 164(4):805–817. DOI: 10.1016/j.cell.2016.01.029 [PubMed: 26871637]
74. Liu C, Song X, Nisbet R, Gotz J. Co-immunoprecipitation with tau isoform-specific antibodies reveals distinct protein interactions, and highlights a putative role for 2N tau in disease. *J Biol Chem*. 2016; doi: 10.1074/jbc.M115.641902
75. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet*. 2011; 12(10):715–729. DOI: 10.1038/nrg3052 [PubMed: 21921927]
76. Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. *Mol Cell*. 2005; 19(1):1–13. DOI: 10.1016/j.molcel.2005.05.026 [PubMed: 15989960]
77. Liu J, McClelland M, Stawiski EW, Gnad F, Mayba O, Haverty PM, Durinck S, Chen YJ, Klijn C, Jhunjhunwala S, Lawrence M, Liu H, Wan Y, Chopra V, Yaylaoglu MB, Yuan W, Ha C, Gilbert HN, Reeder J, Pau G, Stinson J, Stern HM, Manning G, Wu TD, Neve RM, de Sauvage FJ, Modrusan Z, Seshagiri S, Firestein R, Zhang Z. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nat Commun*. 2014; 5:3830. doi: 10.1038/ncomms4830 [PubMed: 24807215]
78. Barrett CL, DeBoever C, Jepsen K, Saenz CC, Carson DA, Frazer KA. Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy. *Proc Natl Acad Sci U S A*. 2015; 112(23):E3050–3057. DOI: 10.1073/pnas.1508057112 [PubMed: 26015570]
79. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25(9):1105–1111. DOI: 10.1093/bioinformatics/btp120 [PubMed: 19289445]

80. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323.doi: 10.1186/1471-2105-12-323 [PubMed: 21816040]
81. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013; 10(1):71–73. DOI: 10.1038/nmeth.2251 [PubMed: 23160280]
82. Consortium PAwgotICG, Consortium. Pathway and network analysis of cancer genomes. Nat Methods. 2015; 12(7):615–621. DOI: 10.1038/nmeth.3440 [PubMed: 26125594]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Schematic representation of the multilayer analyses of disease networks leading to the identification of the disease-relevant pathways

Three layers of network complexity are considered (left panels): top, the CNV-level network, where proteins encoded by genes from the same copy number variant (CNV) are grouped into one network node and the interactions of these proteins are merged; middle, the gene-level network, where each network node represents one gene/protein; bottom, the isoform-level network, where a new layer of complexity is added by splitting gene nodes into multiple splicing isoform nodes. (Right panels) Various types of analyses carried out on the networks. Examples of disease-relevant pathways are shown at the bottom and represent potential new disease biomarkers or drug targets.

Figure 2. Schematic representation of the binding enrichment analysis of disease network

The top grids represent disease-related network (left) and the human interactome (HI) as a control network (right). The Y2H screens share the same prey search space (~15,000 ORFs). The lower figure shows how the binding enrichment is calculated for two preys (A and B) present in both, disease and control networks. Prey A, but not prey B, interacts with a greater number of disease baits than expected from the control HI network. The counts should be normalized by the bait search space. For each prey, F_{Disease} is the fraction of all disease network baits binding to the common preys, and F_{HI} is the fraction of all control HI baits binding to the common preys.

Figure 3. Disease network connectivity analyses

Interacting partners of the disease genes are tightly connected at the protein level. (Left panel) A significantly higher number of interactors of the disease proteins are connected when mapped to the control background network compared to random control. (Right panel) Shorter paths among the interacting partners of the disease network can be detected by mapping them to the control network and comparing the path length to the empirical null distribution of connectivity of 10,000 sets of preys randomly selected from the control network.

Figure 4. Co-expression network analysis using WGCNA

Network analysis dendrogram shows the modules based on the co-expression topological overlap of genes within the network. Color bars below the dendrogram give information on module membership; the enriched GO terms for largest module (M1) are shown below the dendrogram with the turquoise horizontal bar indicating the significance of the enrichments.

Figure 5. Protein-protein interactions create a more comprehensive connectivity map among disease candidate genes

The disease network is clustered into three distinct subnetworks (colored spheres) consisting of functionally related groups of genes (grey ovals) based on the shared co-GO annotations (dotted edges). The newly identified PPIs (red edges) link these subnetworks into a single connected component.

Figure 6. The CNV-CNV co-expression networks

Each CNV is shown as a circle with different color and size. The size of the circle reflects the number of genes in each CNV. The edge represents the normalized fraction of co-expressed gene pairs between different CNVs. PCW: post conception weeks; M: months; Y: years.

Figure 7. Simulation of the expression level changes of CNV genes

(a) Simulated expression level changes of highly- and lowly-expressed CNV genes as a result of duplications and deletions. (b) Percentile ranking of the expression levels of all human genes. The CNV genes were defined as highly- or lowly-expressed based on the 80th and 20th percentiles cut-off relative to all human genes. (c) *MAPK3* and *KCTD13* are highly expressed CNV genes (>80th percentile rank), whereas *CHD1L* and *GJA5* are lowly expressed CNV genes (<20th percentile rank). (d) Examples of simulated expression changes of *KCTD13* and *CHD1L* across developmental periods as a result of the deletion and duplication.

Figure 8. Overview of the method for identifying co-expressed and interacting pairs with altered expression in the 16p11.2 deletion and duplication carriers

(a) Left panel: The procedure for identification of the 'partner-alike genes' to build the expected distributions of correlation coefficients for control, 16p11.2 deletion and duplication carriers. Upper right panel: Pairwise Spearman Correlation Coefficients (SCCs) were calculated between CNV genes (G1, G2, G3) and their partners (P1, P2, P3), as well as CNV genes with partner-alike genes (A, C, E). Lower right panel: The distributions of SCCs between CNV genes and 'partner-alike genes' in the control (gray line), deletion (red line) and duplication (blue line) carriers as well as SCCs between G1 and P1 in the same datasets.

(b) The distribution of SCCs for KCTD13-CUL3 pair in ten randomly selected sets (six

samples each) of control subjects (multicolored lines), six deletion carriers (red line) and six duplication carriers (blue line). Reprinted from [9].

Figure 9. The CNV-CNV network links CNV nodes into a single connected component
 Genes (small black circles) within the same CNV node are grouped into nodes (larger grey circles), edges (blue lines) correspond to PPIs. The disease CNV-CNV network connects all 27 CNVs into a single connected component.

Figure 10. Isoform-level protein-protein interactions network

Bidirectional isoform-level network can only be constructed when isoform-level information for both, baits and preys, is available. Each isoform is shown as a small dark circle with the number inside, and isoforms from the same gene are grouped into grey ellipses. The red edges represent isoform-level protein-protein interactions. Different isoforms may have different interaction patterns.

Figure 11. The Jaccard distance measures interaction profile dissimilarity of protein isoforms encoded by the same gene

The network genes with two or more isoforms with PPIs can be arranged in order of increasing fraction of differences in interacting partners (Jaccard distance). The PPI matrices are shown for three proteins, A, B and C. The blue squares within a matrix represent positive interactions, the grey squares represent negative interactions. In network representations, nodes corresponding to multiple isoforms of the same gene are shown as grey circles, and their interacting partners are shown as grey triangles. The dark grey edges are interactions of one isoform, and the red edges are interactions of the other isoform of the same gene.

Figure 12. The example of different isoforms of the *CTBP1* gene interacting with different isoform partners during human brain development

(a) Expression values of two different isoforms of *CTBP1* gene across different brain developmental periods (P2–P13, fetal to adult). (b) Exon-intron structure of *CTBP1* isoforms. (c) Static PPIs between *CTBP1* isoforms and the isoforms of their interacting partners. (d) The heatmap shows co-expression between interacting isoform pairs across brain developmental periods (P2–P13). Green color represents negative correlation and red color represents positive correlation. The dynamic networks below the heatmap are constructed by integrating co-expression with PPIs; note network changes in different brain developmental periods.