Cell PRESS

Neuron
Article

Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses

Sarah R. Gilman,[1,3] Ivan Iossifov,[2,3,*] Dan Levy,[2] Michael Ronemus,[2] Michael Wigler,[2] and Dennis Vitkup[1,*]
[1]Center for Computational Biology and Bioinformatics and Department of Biomedical Informatics, Columbia University, 1130 St. Nicolas Ave, New York, NY 10032, USA
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[3]These authors contributed equally to this work
*Correspondence: iossifov@cshl.edu (I.I.), dv2121@columbia.edu (D.V.)
DOI 10.1016/j.neuron.2011.05.021

SUMMARY

Identification of complex molecular networks underlying common human phenotypes is a major challenge of modern genetics. In this study, we develop a method for *net*work-*b*ased *a*nalysis of *g*enetic associations (NETBAG). We use NETBAG to identify a large biological network of genes affected by rare de novo CNVs in autism. The genes forming the network are primarily related to synapse development, axon targeting, and neuron motility. The identified network is strongly related to genes previously implicated in autism and intellectual disability phenotypes. Our results are also consistent with the hypothesis that significantly stronger functional perturbations are required to trigger the autistic phenotype in females compared to males. Overall, the presented analysis of de novo variants supports the hypothesis that perturbed synaptogenesis is at the heart of autism. More generally, our study provides proof of the principle that networks underlying complex human phenotypes can be identified by a network-based functional analysis of rare genetic variants.

INTRODUCTION

The ongoing revolution in genomic and sequencing technologies has allowed researchers to routinely perform genome-wide association studies (GWAS) for multiple common human diseases and phenotypes (Frazer et al., 2007; Hardy and Singleton, 2009). Although these studies have successfully identified hundreds of significant associations, common polymorphisms reaching genome-wide significance usually explain a relatively small fraction of disease heritability (Goldstein, 2009). There is a growing consensus in genetics that the most valuable contribution of GWAS studies will be in the identification of functional pathways underlying the observed phenotypes (Hirschhorn, 2009). In addition, it is likely that a significant fraction of so-called missing heritability (Manolio et al., 2009), which has eluded association studies, is accounted for by rare single nucleotide mutations and structural genomic variations (McClellan and King, 2010).

A notable example of a disease with a very complex allelic architecture is autism—one of the most common neurological disorders (Geschwind, 2008). Autism spectrum disorders are characterized by impaired social interactions, abnormal verbal communication, restricted interests, and repetitive behaviors. Due in part to better detection strategies, the combined prevalence of ASD has been steadily increasing for several decades and is now approaching a staggering 1% in the human population. Although autism has a very strong genetic component, with an estimated heritability as high as 90% based on studies of monozygotic twins (Hyman, 2008), GWAS-based searches have implicated only a few genes that are associated with common polymorphisms reaching genome-wide significance (Wang et al., 2009; Weiss et al., 2009). In addition, the agreement between published findings remains poor (Manolio et al., 2009) and underlying genetic determinants for this disease still remain largely unknown. Importantly, there is growing evidence that rare sequence mutations and de novo copy number variations (CNVs) (Marshall et al., 2008; Moessner et al., 2007; Pinto et al., 2010; Sebat et al., 2007) significantly contribute to autism etiology (Zhao et al., 2007).

The main challenge in the analysis of rare genetic variations, such as de novo CNVs, is precisely their rarity, i.e., the fact that a vast majority of the observed genetic events are unique. Consequently, each rare variant by itself is not statistically significant, so an integrative conceptual framework is required to understand their overall functional impact. We hypothesized that recently obtained genome-wide de novo CNV data (Levy et al., 2011) could allow identification of the underlying biological pathways and processes if considered in the context of functional biological networks (Feldman et al., 2008; Iossifov et al., 2008). Here, we develop a method for *net*work-*b*ased *a*nalysis of *g*enetic associations (NETBAG) and demonstrate its utility in autism. The presented approach can determine whether the observed rare events en masse affect a significantly interconnected functional network of human genes.

RESULTS

NETBAG Method Overview

To implement our approach, we first built a background network that connects any pair of human genes with a weighted edge

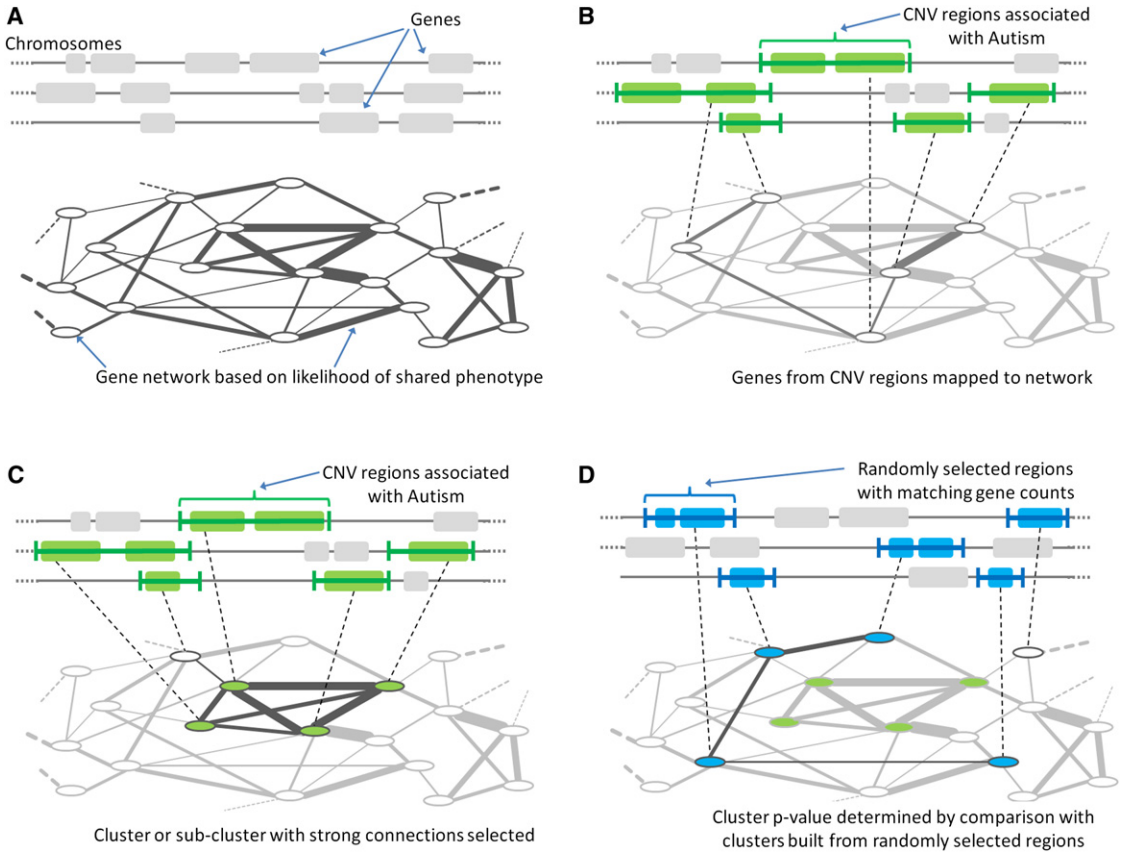

Figure 1. Outline of the NETwork-Based Analysis of Genetic Associations (NETBAG), the Method Used in Our Study to Identify Significant and Functionally Related Gene Networks Affected by De Novo CNV Events

(A) A background network of human genes is constructed in which nodes indicate genes and edges represent the likelihood that two genes participate in the same genetic phenotype.

(B) One or two genes are selected from each of de novo CNV region to form a cluster. The genes are mapped to the likelihood network and a combined score is calculated for each cluster based on interactions between its genes.

(C) A greedy search procedure is used to identify the cluster with maximal score.

(D) The significance of the cluster with maximum score is determined by comparing it to the distribution of maximal scores from randomly selected genomic regions with similar gene counts.

See Figure S1 for a further description of the NETBAG approach.

encapsulating our a priori expectation that the two genes participate in the same genetic phenotype (see Experimental Procedures and Supplemental Experimental Procedures). This background network was based on a combination of various functional descriptors, such as shared gene ontology (GO) annotations (Ashburner et al., 2000), functional pathways in KEGG (Kanehisa and Goto, 2000), shared interaction partners and coevolutionary patterns (see Experimental Procedures). Similar methods have been previously used to build functional networks in humans and several model organisms (Lee et al., 2004, 2008). In contrast to the aforementioned studies, edges in our network represent the likelihood that two genes participate in a similar genetic phenotype rather than necessarily share cellular functions. Importantly, no deliberate biases toward genes previously implicated in autism or biological functions related to nervous system were used in building the network. The likelihood network was assembled using a large set of known disease-gene associations that were carefully curated for our previous study (Feldman et al., 2008). This set contains 476 genes associated with 132 different genetic diseases (see Experimental Procedures).

Using the constructed network, we searched for functionally connected clusters of human genes affected by de novo CNVs (Figure 1). The genes within the observed CNV regions were first mapped to the nodes corresponding to these genes in the network (Figure 1B). Clusters of genes were assigned scores based on the strength of their connections, and a greedy search algorithm (see Experimental Procedures) was then used to find high-scoring clusters of genes within the CNV regions (Figure 1C). In this search procedure genes from any CNV region could be selected to be members of the growing cluster (Figure 1C), but to prevent large CNV regions from dominating clusters, we allowed no more than one or two genes from a given CNV to participate in a cluster (Figures 2A and 2B, respectively).

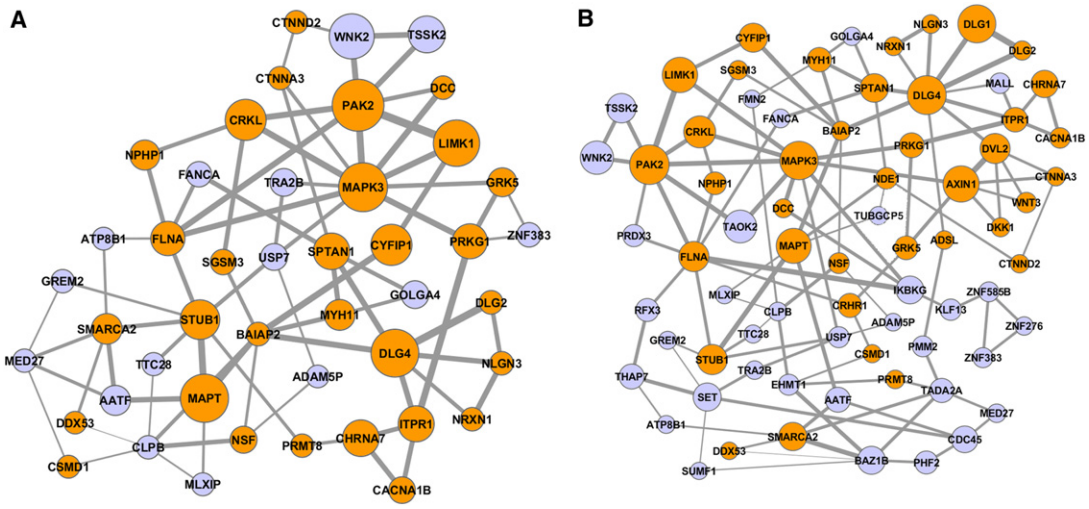

Figure 2. Gene Clusters Found using NETBAG Analysis of De Novo CNV Regions Observed in Autistic Individuals

(A) The highest scoring cluster obtained using the search procedure with up to one gene per each CNV region.

(B) The cluster obtained using the search with up to two genes per region. In the figure, genes (nodes) with known functions in the brain and nervous systems are colored in orange (see Table S2 for functional information about the genes forming the cluster). Node sizes represent the importance of each gene to the overall cluster score. Edge widths are proportional to the prior likelihood that the two corresponding genes contribute to a shared genetic phenotype. For clarity, we show only edges corresponding to the two strongest connections for at least one node.

See also Figure S2, Table S1, and Table S2.

A gene cluster was scored by considering the sum of log likelihood-based edges between all genes within the cluster. Such a scoring scheme is conceptually equivalent to calculating the expected likelihood that all genes within the cluster will participate in the same genetic phenotype. To account for the fact that functional interactions between genes in a cluster are not independent, we employed a previously developed de-weighting heuristic (Lee et al., 2004) described in the Experimental Procedures; similar results were obtained with or without the deweighting procedure (see Table S1). To calculate the p value for the resulting clusters, random events were generated with the same gene count or alternatively with the same genomic length, as in the observed de novo CNV dataset. The greedy algorithm was then applied to search for high-scoring clusters formed by genes from these random events. p values were assigned to clusters based on the distribution of scores in the randomized data clusters (see Experimental Procedures).

We and others have previously used various network-based methods to analyze genetic data from rare and common diseases (Feldman et al., 2008; Franke et al., 2006; Iossifov et al., 2008, 2009; Lango Allen et al., 2010; Raychaudhuri et al., 2009). NETBAG differs from the previous approaches in several important ways. Specifically, the underlying weighted network does not represent a molecular interaction network or a set of predefined functional pathways, but instead the prior likelihood that any pair of human genes is involved in the same genetic phenotype. NETBAG then defines a formal procedure for identifying strongly connected clusters among a large set of genetically perturbed genes and evaluating the genome-wide cluster significance. The relative importance of specific genes forming a cluster is then evaluated based on the contribution of genes to the overall cluster score. We are currently working on making the NETBAG method available as a web server; in the meantime, we will be happy to share the developed methodology with any interested parties.

Functional Gene Network Identified from Rare De Novo CNVs

The NETBAG approach was directly applied to the experimental CNV dataset described in the companion paper by Levy et al. (2011; this issue of *Neuron*). This set contained 75 rare de novo CNVs encompassing 746 unique human genes. For our analysis, we combined all overlapping events into a single region and removed all events that did not intersect any genes; we also removed six very large CNV events (length >5 mb). As a result, the final set used for our analysis contained 47 CNV regions from affected individuals intersecting 433 genes. In addition, Levy et al. (2011) also identified ultrarare CNVs inherited by autistic subjects but not their siblings; inherited CNVs were classified as ultrarare if at least one of the corresponding genes was not affected by any other event from the study. Applying the same preprocessing procedure resulted in 156 CNV regions with 419 genes associated with rare inherited events in autistic children.

Using the data described above, we identified statistically significant gene clusters affected by de novo CNV events associated with autistic individuals. Significant clusters detected using either one-gene-per-CNV (p value = 0.02) or the two-genes-per-CNV (p value = 0.02) clustering are shown in Figure 2. If genes forming the high scoring clusters were masked, no other significant clusters were detected in the data.

In contrast, no statistically significant clusters were obtained using ultrarare inherited CNVs from affected individuals (the

best cluster p value = 0.6) The absence of a significant cluster in the ultrarare inherited data set, which had a comparable number of genes to the de novo events, suggests that the inherited CNVs contain a significantly smaller fraction of casual genes, i.e., genes associated with autistic phenotype. This conclusion is also supported by the observation by Levy et al. (2011) that there is less bias in transmission of ultrarare inherited events. In other words, autistic children were almost as likely as their unaffected siblings to inherit an ultrarare event. This is in sharp contrast to the de novo events, which were nearly four times more frequent in the autistic children (7.9% in autistic children versus 2.0% in unaffected siblings).

The contribution of each gene to the cluster score, i.e., its functional connection to other cluster genes, is not uniform. To capture each gene's contribution to the cluster score we performed Markov Chain Monte Carlo (MCMC) simulations, sampling clusters based on their scores. The size of each gene (node) in Figure 2 is proportional to the each gene's membership in high-scoring clusters during the sampling simulations (see Experimental Procedures, Supplemental Experimental Procedures). Similar node sizes were also obtained based on the average connection strength from each gene to the other genes in the cluster (Pearson's r = 0.8, p value = $4*10^{-11}$).

Interestingly, we found that genes affected by de novo CNVs observed in females are significantly more important for the overall cluster score than genes affected by CNVs in males, i.e., female genes have stronger average connections with other genes from the identified network (see Figure S2; one-tail Mann-Whitney test, female > male, p value = 0.013). This observation is illuminating because one of the striking phenotypic characteristics of autism is the male-to-female incidence ratio of more than 5:1 for high-functioning ASD (Newschaffer et al., 2007). It has been previously suggested (Zhao et al., 2007) that stronger genetic perturbations are required, on average, to trigger an autistic phenotype in females than males due to currently unknown compensatory mechanisms. Two mechanisms may lead to stronger perturbations in females: CNVs encompassing a larger number of genes that are associated with ASD and CNVs intersecting individual genes that produce a more deleterious impact when disrupted. The analysis of de novo events in affected individuals lends support to both of these mechanisms: the CNVs in females are indeed significantly larger (with median of 10 genes per CNV in females, three genes per CNV in males, two-tail Mann Whitney, p value = 0.02), and genes derived from female CNVs are more functionally important for the network shown in Figure 2. Using simulations of random CNVs we also confirmed that the difference in the relative importance of female versus male nodes is unlikely (p = 0.024) to be a simple consequence of the larger CNV sizes in females (see Supplemental Information; Figure S2C). We believe that both of the aforementioned mechanisms are at play. Indeed, it would be surprising that stronger perturbation can be inflicted exclusively by larger CNVs and not disruption of high impact genes, and vice versa.

Functional Characterization of the Identified Network

Analysis of the established annotation resources, such as Swiss-Prot (UniProt Consortium, 2007), GeneCards (www.genecards.org), WikiGenes (www.wikigenes.org), and IHOP (Hoffmann and Valencia, 2004), suggests that a significant fraction of genes in the identified network either play a well-defined functional role in the brain or have been previously implicated in neurodegenerative and psychiatric disorders. Only ~25% (54 of a randomly selected 214) of all genes within the de novo CNV regions have been previously associated with brain-related phenotypes. However, when we consider genes in the identified clusters this proportion rises drastically (p value < 10^{-3}), to ~67% (Figure 2A; 30 out of 45) for the one-gene-per-CNV cluster or ~52% (Figure 2B; 38 out of 72) for the two-genes-per-CNV cluster (see Table S2 for functional description of cluster genes).

To characterize in more detail the specific biological processes related to the cluster in Figure 2A, we investigated the strength of functional interactions between the cluster genes and various gene ontology (GO) categories (Ashburner et al., 2000). GO categories represent a curated set of functionally related genes described by a controlled vocabulary. For human genes in each of 1454 GO categories we calculated their average log likelihood interaction score (using the background network) with the genes in the identified cluster (Figure 2). The GO-specific significance of these interaction scores was calculated by comparison with scores of randomly generated CNV events with the same gene count at in real data by Levy et al. (2011). A false discovery rate (FDR) procedure was used to correct for multiple hypothesis testing (see Experimental Procedures). The 25 GO categories with lowest Q values, indicating the highest connection significance to the autism associated cluster, are shown in Table 1 (see Table S3 for other significant GO categories). These GO categories are primarily related to actin network dynamics and reorganization, synaptogenesis, axonogenesis, cell-cell adhesion, small GTPase signaling, and neurite development. Consequently, the identified functional network is associated with a diverse collection of molecular and cellular processes essential for proper synaptogenesis and axon guidance. We note that is it not possible to obtain the same functional results by a statistical analysis of significantly overrepresented GO terms for all 433 gene within the de novo CNVs from affected individuals (see Supplemental Experimental Procedures for details). The significant GO terms presented in Table 1 specifically describe the functional connection of the network in Figure 2.

Using the same methodology, we found that the cluster in Figure 2A is strongly related to the set of genes previously implicated in autism (p value = 0.001; see Supplemental Experimental Procedures) and genes associated with intellectual disability phenotypes (p value = 0.017). The collections of genes responsible for these phenotypes were manually compiled recently by Pinto et al. (2010) through an extensive review of the literature and available databases. In spite of strong functional connections, the overlap between genes in the aforementioned sets and the genes identified in our analysis is relatively small (~3%). Thus, our study significantly expands the collection of genes implicated in ASD. The cluster genes are also strongly connected (p value = 0.013) to proteins identified experimentally by recent proteomic profiling of postsynaptic density (PSD) from human neocortex (Bayés et al., 2011).

Table 1. Gene Ontology (GO) Terms Highly Connected to the Functional Network in Figure 2A

Gene Ontology Term	GO Category	Q Value
GO:0007015: actin filament organization	Biological process	<0.01
GO:0030424: axon	Cellular component	<0.01
GO:0048469: cell maturation	Biological process	<0.01
GO:0007611: learning and or memory	Biological process	<0.01
GO:0044456: synapse part	Cellular component	<0.01
GO:0045202: synapse	Cellular component	<0.01
GO:0007163: establishment and or maintenance of cell polarity	Biological process	0.01
GO:0045216: intercellular junction assembly and maintenance	Biological process	0.01
GO:0019201: nucleotide kinase activity	Molecular function	0.01
GO:0005912: adherens junction	Cellular component	0.01
GO:0007409: axonogenesis	Biological process	0.01
GO:0016323: basolateral plasma membrane	Cellular component	0.01
GO:0030041: actin filament polymerization	Biological process	0.01
GO:0051258: protein polymerization	Biological process	0.01
GO:0021700: developmental maturation	Biological process	0.01
GO:0030863: cortical cytoskeleton	Cellular component	0.01
GO:0000904: cellular morphogenesis during differentiation	Biological process	0.01
GO:0005925: focal adhesion	Cellular component	0.01
GO:0030427: site of polarized growth	Cellular component	0.01
GO:0032271: regulation of protein polymerization	Biological process	0.01
GO:0031175: neurite development	Biological process	0.01
GO:0048666: neuron development	Biological process	0.01
GO:0030055: cell matrix junction	Cellular component	0.01
GO:0030832: regulation of actin filament length	Biological process	0.01
GO:0030036: actin cytoskeleton organization and biogenesis	Biological process	0.01

We defined a score that reflects the degree to which the network genes interact with gene sets assigned to various GO categories (1454 in total). For every GO category, a background distribution was built by generating clusters from randomly chosen CNV regions with the same gene count as observed in real data by Levy et al. (2011). These distributions were used to derive false discovery rate (FDR) corrected Q values accounting for the multiple hypothesis tests (one test for every GO gene set; see Experimental Procedures). See also Table S3.

DISCUSSION

At the core of the processes listed in Table 1 is the development and maturation of synaptic contacts in the brain. The functional relationships between proteins in the identified cluster can be better appreciated if considered in the context of molecular interactions involved in formation and maturation of the excitatory (glutamatergic) synapse (Figure 3). The excitatory synaptic connections are formed between axons and dendritic spines, which are complex and dynamic postsynaptic structures containing thousands of different proteins (Alvarez and Sabatini, 2007; Tada and Sheng, 2006). The formation, maturation and elimination of dendritic spines lie at the core of synaptic transmission and memory formation (Roberts et al., 2010; Yang et al., 2009). In Figure 3 the genes that are members of the identified network are shown in yellow, other functionally related genes within rare de novo CNV regions from Levy et al. (2011) are in blue and genes previously implicated or discussed in the context of autism are highlighted using orange borders. Although the picture shows a dense and interconnected web of molecular interactions, the processes depicted in the figure can be understood in terms of several signaling and structural pathways. Many of these pathways ultimately converge on the regulation of the growth and branching of the actin filament network, which is essential for spine structural remodeling and morphogenesis.

The initial contacts between axons and dendrites are mediated by specific adhesion-related proteins, such as neurexin and neuroligin (e.g., *NRXN1* and *NLGN3*, genes perturbed by rare de novo CNVs associated with ASD are underlined here and below) (Südhof, 2008). On the postsynaptic side of an excitatory synapse, the initial axon-dendrite contacts ultimately develop into a complex and dense structure, the postsynaptic density (PSD), dominated by several types of glutamate receptors (such as AMPA and NMDA), various scaffolding proteins (*DLG4/PSD95*, *DLG2*, *SHANK2/3*, *SynGAP1*, *DLGAP2*) and trafficking/signaling proteins (*CTNND2*). In total, the PSD contains many hundreds of distinct proteins (Bayés et al., 2011; Sheng and Hoogenraad, 2007). Information for activity-dependent regulation of spine morphology is passed through an intermediate level of signaling protein, such as Rho family (Linseman and Loucks, 2008) of small GTPases (RhoA/B, Cdc42, Rac1) to downstream targets (*LIMK1* and *PAK1/2/3*) connected to proteins modifying morphology of the actin network (*cofilin* and *Arp2/3*) (Blanchoin et al., 2000). The activity of the GTPases is regulated pre- or postsynaptically by many guanine exchange factors (GEFs), GDP dissociation inhibitors (GDIs, such as *GDI1*) and GTP-activating proteins (GAPs). Many other proteins shown in Figure 3, such as *FLNA*, *CTNNA3*, *DOCK8*, *SPTAN1*, *CYFIP1*, either bind directly to the actin network or mediate interaction of actin filaments with other proteins.

The WNT signaling pathway plays a crucial role in diverse processes associated with formation of neural circuits (Salinas and Zou, 2008). This pathway is also known to be directly involved in the regulation of dendrite morphogenesis (Rosso et al., 2005; Salinas et al., 1994). WNT signaling is accomplished through the canonical branch (*DVL*, *AXIN1*, beta-catenin) and the noncanonical branch (*DVL1/2/3*, Rac1, and JNK); both of these pathway branches converge on regulation of actin network morphogenesis. Similar to WNT, the reelin signaling also plays a prominent role in the context of autism phenotype and specifically dendritic spine morphogenesis (Fatemi et al., 2005; Niu et al., 2008). Signaling by secreted extracellular *RELN* protein acts though *VLDR* and Apoer2 receptors and the PI3K/Akt pathway (Jossin and Goffinet, 2007) regulating the mammalian

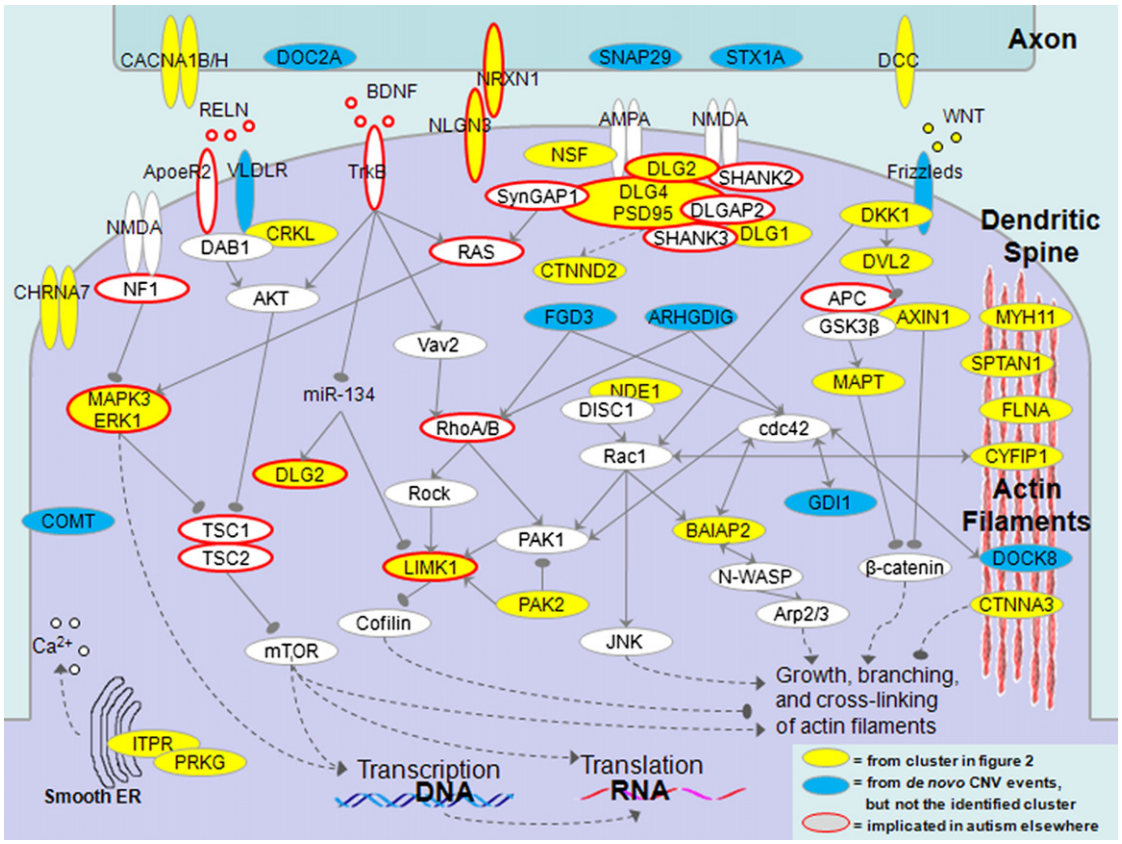

Figure 3. Genes Associated with the Morphogenesis of Dendritic Spines

Dendritic spines are dynamically forming protrusions from a neuron's dendrite which mediate excitatory connection to axons and determine synaptic strength. The proteins shown in the figure play crucial roles in formation of physical contacts between axons and dendrites, organization of postsynaptic density (PSD), and signaling processes controlling spine morphology. Many of the signaling pathways ultimately converge on the regulation of the growth and branching of the actin filament network, which is essential for spine structural remodeling. The proteins encoded by genes from the identified functional cluster (Figure 2) are shown in yellow, other genes hit by de novo CNV from Levy et al. (2011), in blue, and genes previously implicated or discussed in the context of autism are highlighted using orange borders. See also Table S4.

target of rapamycin (mTOR) pathway (Kumar et al., 2005; Shaw and Cantley, 2006). Another important pathway converging on mTOR involves *MAPK3/ERK*, which can be activated by Ras and NF1. mTOR integrates various inputs from upstream growth-related pathways, and is also known to regulate dendrite morphogenesis (Tavazoie et al., 2005).

Although we discussed proteins in the context of dendritic spine development (Figure 3 and text above), many of the aforementioned proteins also participate in diverse cellular processes and are reused in the context of axon guidance and neuron motility (Shen and Cowan, 2010). Such recycling of proteins is natural because actin network dynamics are essential for such processes as growth of axonal filopodia, which are used in searching for growth cone guidance cues (Tessier-Lavigne and Goodman, 1996). The presence of *DCC* protein in the identified network (Figures 2 and 3), also suggests an important role of perturbed axonal guidance in autism. Although *DCC* is also involved in dendrite development (Suli et al., 2006), this receptor and its signaling protein, netrin, are primarily essential for guiding axons to their final destinations (Tessier-Lavigne and Goodman, 1996). Several signaling pathways highlighted in Figure 3, such as the WNT and reelin pathways, also play prominent roles in neuron motility (Reiner and Sapir, 2005; Salinas and Zou, 2008). In addition, several specific proteins, such as *PAKs* and *LIMK*, which regulate the dynamics of actin network, are reused in axonal morphogenesis. Consequently, malfunction of many proteins shown in Figure 3 may influence autistic phenotypes through their role in either dendrite or axon signaling, or possibly a combination of these processes.

Considering the genes hit by rare de novo variants from the perspective of the functional molecular network (Figure 3) allowed us to investigate the likely morphological consequences of some CNVs. There is growing evidence that changes in dendritic spine morphology contribute to a number of neurological disorders (Halpain et al., 2005). A decrease in the density of dendritic spines in regions of the cerebral cortex has been linked to schizophrenia (Blanpied and Ehlers, 2004; Garey et al., 1998; Glantz and Lewis, 2000). On the other hand, an increase in spine size or density has been connected to Fragile X syndrome, a disorder frequently associated with autism (Fiala et al., 2002; Kaufmann

and Moser, 2000). Following the logic that CNV deletions should decrease while duplications increase the dosage of the affected genes, we can infer—based on the structure and regulatory logic of the functional network in Figure 3—the morphological effects of 13 gene perturbations on dendritic spines. Specifically, we found that in 11 out of 13 cases (~85%) the gene perturbations caused by the observed CNV events should increase either dendritic spine growth or their density (see Table S4). This result is consistent with recent findings that autistic individuals have increased spine density in portions of their cerebral cortex (Hutsler and Zhang, 2010; Woolfrey et al., 2009) and possibly a local brain overconnectivity (Scott-Van Zeeland et al., 2010).

Overall, the results of this study, the first to our knowledge, demonstrate that autism-associated rare de novo CNVs, observed in an unbiased genome-wide study, form a large and statistically significant functional network responsible for synaptogenesis, axon guidance, and related molecular processes. Therefore, our analysis strongly supports the hypothesis that autism is primarily a disease of synaptic and neuronal connectivity malfunction (Zoghbi, 2003). The identified functional network also reveals a striking genetic complexity of autism. The genetic events we observe affect the whole arc of molecular processes essential for proper synapse formation and function. Similar genetic complexity is already apparent in many cancers (Cancer Genome Atlas Research Network, 2008; Wood et al., 2007) and—as we and others believe—will be a hallmark of many other common human phenotypes and maladies (Wang et al., 2010). In spite of the observed complexity, our study provides an important proof of the principle that underlying functional networks responsible for common phenotypes can be identified by an unbiased analysis of multiple rare genetic perturbations from a large collection of affected individuals.

The functional network presented in Figure 3 contains approximately 70 genes, with about 40% of them perturbed by rare de novo CNVs observed by Levy et al. (2011). As more genetic data are analyzed it is likely that the network will grow in size and significance. Considering that up to a thousand (Sheng and Hoogenraad, 2007) distinct proteins are associated with postsynaptic density or that hundreds of different GAPs/GEFs modify activity of Rho GTPases that are associated with actin network remodeling, it is likely that many hundreds of genes could ultimately contribute to the autistic phenotype. This estimate, based on the functional network, is consistent with independent estimates based on recurrent mutations and the overall incidence of autism in the human population (Zhao et al., 2007; Levy et al., 2011). Deleterious variants in different genes contributing to autistic phenotype will almost certainly have different penetrance and vulnerabilities. The identification of the complete set of genes responsible for ASD and understanding their respective contributions to the phenotype will require analyses of next generation sequencing data coupled with investigation of underlying molecular networks.

EXPERIMENTAL PROCEDURES

Copy Number Variation (CNV) Data Set

In our analysis, we used the CNV data set obtained in a companion study by Levy et al. (2011). The data set contained 75 rare de novo CNV events from autistic children. Six very large CNV events, spanning more than 5 mb each, were not considered in our analysis. The initial CNV dataset contained several overlapping events, including a set of 10 events all within the region 16p11.2. Any overlapping CNVs were collapsed into single events to avoid double counting of genes. We ignored all CNV events that did not contain any annotated human gene based on the NCBI genome build 36. After aforementioned preprocessing steps, our final CNV set from autistic children contained 47 loci in total affecting 433 human genes; the average number of genes within each de novo CNV region was ~9, with the median of three genes per regions. Levy et al. (2011) also identified 157 ultrarare inherited CNVs transmitted between parents and autistic children. Inherited CNVs were classified as ultrarare if the genes within these events were not affected by any other inherited events in the study. We applied the same preprocessing steps to these inherited CNVs, resulting in 156 regions affecting 418 genes.

Building the Background Likelihood Network

To perform the NETBAG analysis, we built a background network connecting all pairs of human genes. Every gene pair in this network was assigned a score proportional to the log of the ratio of the likelihood that the two genes participate in the same genetic phenotype to the likelihood that they do not (see Supplemental Experimental Procedures). Importantly, although similar in spirit to integrative methods that have been used previously to build functional networks in several model species (Lee et al., 2004, 2008), the edges in our network represent the likelihood to participate in the same genetic phenotype rather than share a functional and molecular interaction.

The likelihood network was build using, as a positive gold standard, the carefully curated set of human genes compiled recently by Feldman et al. (2008). This set contains 476 human genes associated with 132 different genetic phenotypes. As a negative gold standard we used a set of randomly selected pairs of human genes that are not known to be associated with identical diseases phenotypes. Importantly, no genes previously implicated in ASD or any biologically related functions were used in the network construction. The likelihood score was derived based on naive Bayesian integration of various descriptors of proteins function: shared GO annotations, participation in the same KEGG pathways, shared protein domains in InterPro, direct protein-proteins interactions and shared interaction partners from multiple databases (BIND, BioGRID, DIP, HPRD, InNetDB, IntAct, BiGG, MINT, and MIPS), sequence homology between the gene pair calculated using BLAST (Altschul et al., 1997), and two measures of similarity in coevolutionary patterns: phylogenetic profile similarity and chromosomal coclustering across genomes (Chen and Vitkup, 2006). We cross-validated the quality of the background network by showing that it can be successfully used to prioritize (rank) genes, located within a chromosomal region, across a variety of genetic phenotypes (see Supplemental Experimental Procedures for details).

Search for High-Scoring Clusters Affected by CNVs

To score a cluster of genes in the network (Figure 1), we combined the scores for all gene pairs forming the cluster. The direct multiplication of the corresponding likelihoods (network edges) is conceptually equivalent to assuming that all connections within the cluster are independent; we refer to this procedure as the naive scoring scheme. Second, we applied a simple deweighting scheme used previously for functional data integration (Lee et al., 2004). For each gene forming the cluster the deweighting scheme scores the strongest connection in full, and then decreases the other connections in order of their strength in a linear fashion (see Supplemental Experimental Procedures). Effectively, the deweighting scheme gives more weight to strongest gene-gene connections within the cluster. The detected functional clusters were significant under both scoring schemes.

A greedy growth algorithm was used to find strongly connected clusters of genes located within CNV regions (Figure 1). Specifically, the search algorithm was started from every possible gene in CNV regions, then the gene with the strongest connection to the first gene was added. At all subsequent iterations, genes located within CNV regions that most increased the cluster score were added. Only one (results in Figure 2A) or two (Figure 2B) genes per each CNV region were allowed in the growing cluster. This growth procedure was run until no further genes could be added. For each cluster size, clusters obtained

by starting with each gene within CNV regions were compared and the cluster with the highest score was selected.

Evaluating Significance of Cluster Scores

We first determined the p value for the best cluster at each cluster size; we refer to this as the *local p value*. Local p values were calculated based on rerunning the greedy search algorithm using random human genome regions identical (either in length or gene number) to those observed by Levy et al. (2011). Second, to determine the most significant cluster across sizes, we compared the lowest local p value obtained from the real data, to the distribution of lowest local p values obtained in the 10,000 trails from the randomized regions. Effectively, this allowed us to assign a p value to our local p value; we refer to this as the *global p value*. The global p value is more stringent because it accounts for multiple hypotheses testing, arising due to different cluster sizes; in our manuscript we refer to global p value simply as p value. In the aforementioned calculation of local and global p values, we used two alternative randomization procedures for human genomic regions: we either preserved the genomic size of CNVs or the gene counts to the values observed in the real data. All randomized regions were generated using the NCBI human genome build 36 (hg18). The functional cluster identified in our work was significant under both randomization schemes (preserving length of CNVs or gene counts) and cluster scoring methods (naive and deweighted). The p values for different randomization procedures are given in Table S1.

In addition to the randomization of genomic regions we wanted to ensure that our results were not due to some general topological features of the background network. To explore this possibility, we randomly shuffled the background network while preserving the distribution of connection strengths for each gene (see Supplemental Experimental Procedures). We then repeated the NETBAG search using the de novo CNVs from affected children. This search using the shuffled network identified no significant clusters or GO terms.

Analysis of Genes Contribution to Cluster Score

Contributions of different genes to the score of the identified functional cluster (Figure 2) vary substantially. To capture that, we devised a formal method to assign weights to individual genes reflecting their contribution to high scoring clusters. The method is based on two distributions over clusters: p(C), in which clusters with high scores are assigned a high probability, and a uniform distribution, $p^u(C)$, in which all clusters are equally likely (See Supplemental Experimental Procedures). Each individual gene was then given a score equal to the ratio of the number of clusters that contain the gene sampled from p(C) to the number sampled from $p^u(C)$. As a result, the genes which were more frequently included in high-scoring clusters were assigned higher ratios. We used Markov-Chain Monte Carlo (MCMC) to sample 5 million clusters from each of the two distributions.

Calculation of Functional Relationships between the Cluster and Various GO Categories

To characterize the identified cluster we investigated its interactions with a collection of a priori defined functional sets of human genes. For this purpose, we utilized the 1454 gene sets corresponding to the gene ontology (GO) categories used in the MSigDB database (Subramanian et al., 2005). Using the background likelihood network, we calculated, for each gene set, its average interaction to the identified cluster shown in Figure 2. To determine the significance of the calculated interaction scores we built gene set-specific background distributions by generating random clusters from the randomized genomic regions with the same gene count as in Levy et al. (2011). We used the background distribution to assign an empirical p-value for every gene set, and then applied the FDR procedure to address the multiple hypotheses involved in testing all gene sets within the collection (see Supplemental Experimental Procedures).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.neuron.2011.05.021.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the Simons Foundation (SFARI award number SF51 to M.W.), the National Centers for Biomedical Computing (MAGNet) grant U54CA121852 to Columbia University. S.R.G. was supported by the training grant T32 GM082797. We are grateful to all of the families at the participating SFARI Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, B. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, E. Wijsman).

We would also like to sincerely thank Simons Foundation Autism Research Initiative for generous financial support, Linda Van Aelst, Thomas Jessell, Gerald Fischbach, Marian Carlson, Alan Packer, Barry Honig, Itsik Pe'er, Lauren DeMaria, and Stephen Johnson for helpful discussions.

Accepted: May 24, 2011
Published: June 8, 2011

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Alvarez, V.A., and Sabatini, B.L. (2007). Anatomical and physiological plasticity of dendritic spines. Annu. Rev. Neurosci. *30*, 79–97.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bayés, A., van de Lagemaat, L.N., Collins, M.O., Croning, M.D., Whittle, I.R., Choudhary, J.S., and Grant, S.G. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat. Neurosci. *14*, 19–21.

Blanchoin, L., Pollard, T.D., and Mullins, R.D. (2000). Interactions of ADF/cofilin, Arp2/3 complex, capping protein and profilin in remodeling of branched actin filament networks. Curr. Biol. *10*, 1273–1282.

Blanpied, T.A., and Ehlers, M.D. (2004). Microanatomy of dendritic spines: emerging principles of synaptic pathology in psychiatric and neurological disease. Biol. Psychiatry *55*, 1121–1127.

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Chen, L., and Vitkup, D. (2006). Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol. *7*, R17.

Fatemi, S.H., Snow, A.V., Stary, J.M., Araghi-Niknam, M., Reutiman, T.J., Lee, S., Brooks, A.I., and Pearce, D.A. (2005). Reelin signaling is impaired in autism. Biol. Psychiatry *57*, 777–787.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc. Natl. Acad. Sci. USA *105*, 4323–4328.

Fiala, J.C., Spacek, J., and Harris, K.M. (2002). Dendritic spine pathology: cause or consequence of neurological disorders? Brain Res. Brain Res. Rev. *39*, 29–54.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am. J. Hum. Genet. *78*, 1011–1025.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Garey, L.J., Ong, W.Y., Patel, T.S., Kanani, M., Davis, A., Mortimer, A.M., Barnes, T.R., and Hirsch, S.R. (1998). Reduced dendritic spine density on

cerebral cortical pyramidal neurons in schizophrenia. J. Neurol. Neurosurg. Psychiatry *65*, 446–453.

Geschwind, D.H. (2008). Autism: many genes, common pathways? Cell *135*, 391–395.

Glantz, L.A., and Lewis, D.A. (2000). Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. Arch. Gen. Psychiatry *57*, 65–73.

Goldstein, D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696–1698.

Halpain, S., Spencer, K., and Graber, S. (2005). Dynamics and pathology of dendritic spines. Prog. Brain Res. *147*, 29–37.

Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. N. Engl. J. Med. *360*, 1759–1768.

Hirschhorn, J.N. (2009). Genomewide association studies—illuminating biologic pathways. N. Engl. J. Med. *360*, 1699–1701.

Hoffmann, R., and Valencia, A. (2004). A gene network for navigating the literature. Nat. Genet. *36*, 664.

Hutsler, J.J., and Zhang, H. (2010). Increased dendritic spine densities on cortical projection neurons in autism spectrum disorders. Brain Res. *1309*, 83–94.

Hyman, S.E. (2008). A glimmer of light for neuropsychiatric disorders. Nature *455*, 890–893.

Iossifov, I., Zheng, T., Baron, M., Gilliam, T.C., and Rzhetsky, A. (2008). Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res. *18*, 1150–1162.

Iossifov, I., Rodriguez-Esteban, R., Mayzus, I., Millen, K.J., and Rzhetsky, A. (2009). Looking at cerebellar malformations through text-mined interactomes of mice and humans. PLoS Comput. Biol. *5*, e1000559. 10.1371/journal.pcbi.1000559.

Jossin, Y., and Goffinet, A.M. (2007). Reelin signals through phosphatidylinositol 3-kinase and Akt to control cortical development and through mTor to regulate dendritic growth. Mol. Cell. Biol. *27*, 7113–7124.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kaufmann, W.E., and Moser, H.W. (2000). Dendritic anomalies in disorders associated with mental retardation. Cereb. Cortex *10*, 981–991.

Kumar, V., Zhang, M.X., Swank, M.W., Kunz, J., and Wu, G.Y. (2005). Regulation of dendritic morphogenesis by Ras-PI3K-Akt-mTOR and Ras-MAPK signaling pathways. J. Neurosci. *25*, 11288–11299.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science *306*, 1555–1558.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat. Genet. *40*, 181–188.

Levy, D., Ronemus, M., Yamrom, B., Yoon-ha, L., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, this issue, 886–897.

Linseman, D.A., and Loucks, F.A. (2008). Diverse roles of Rho family GTPases in neuronal development, survival, and death. Front. Biosci. *13*, 657–676.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. (2008). Structural variation of chromosomes in autism spectrum disorder. Am. J. Hum. Genet. *82*, 477–488.

McClellan, J., and King, M.C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.

Moessner, R., Marshall, C.R., Sutcliffe, J.S., Skaug, J., Pinto, D., Vincent, J., Zwaigenbaum, L., Fernandez, B., Roberts, W., Szatmari, P., and Scherer, S.W. (2007). Contribution of SHANK3 mutations to autism spectrum disorder. Am. J. Hum. Genet. *81*, 1289–1297.

Newschaffer, C.J., Croen, L.A., Daniels, J., Giarelli, E., Grether, J.K., Levy, S.E., Mandell, D.S., Miller, L.A., Pinto-Martin, J., Reaven, J., et al. (2007). The epidemiology of autism spectrum disorders. Annu. Rev. Public Health *28*, 235–258.

Niu, S., Yabut, O., and D'Arcangelo, G. (2008). The Reelin signaling pathway promotes dendritic spine development in hippocampal neurons. J. Neurosci. *28*, 10339–10348.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., and Daly, M.J.; International Schizophrenia Consortium. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. *5*, e1000534. 10.1371/journal.pgen.1000534.

Reiner, O., and Sapir, T. (2005). Similarities and differences between the Wnt and reelin pathways in the forming brain. Mol. Neurobiol. *31*, 117–134.

Roberts, T.F., Tschida, K.A., Klein, M.E., and Mooney, R. (2010). Rapid spine stabilization and synaptic enhancement at the onset of behavioural learning. Nature *463*, 948–952.

Rosso, S.B., Sussman, D., Wynshaw-Boris, A., and Salinas, P.C. (2005). Wnt signaling through Dishevelled, Rac and JNK regulates dendritic development. Nat. Neurosci. *8*, 34–42.

Salinas, P.C., and Zou, Y. (2008). Wnt signaling in neural circuit assembly. Annu. Rev. Neurosci. *31*, 339–358.

Salinas, P.C., Fletcher, C., Copeland, N.G., Jenkins, N.A., and Nusse, R. (1994). Maintenance of Wnt-3 expression in Purkinje cells of the mouse cerebellum depends on interactions with granule cells. Development *120*, 1277–1286.

Scott-Van Zeeland, A.A., Abrahams, B.S., Alvarez-Retuerto, A.I., Sonnenblick, L.I., Rudie, J.D., Ghahremani, D., Mumford, J.A., Poldrack, R.A., Dapretto, M., Geschwind, D.H., and Bookheimer, S.Y. (2010). Altered functional connectivity in frontal lobe circuits is associated with variation in the autism risk gene CNTNAP2. Sci. Transl. Med. *2*, ra80.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Shaw, R.J., and Cantley, L.C. (2006). Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature *441*, 424–430.

Shen, K., and Cowan, C.W. (2010). Guidance molecules in synapse formation and plasticity. Cold Spring Harb. Perspect. Biol. *2*, a001842.

Sheng, M., and Hoogenraad, C.C. (2007). The postsynaptic architecture of excitatory synapses: a more quantitative view. Annu. Rev. Biochem. *76*, 823–847.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Südhof, T.C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. Nature *455*, 903–911.

Suli, A., Mortimer, N., Shepherd, I., and Chien, C.B. (2006). Netrin/DCC signaling controls contralateral dendrites of octavolateralis efferent neurons. J. Neurosci. *26*, 13328–13337.

Tada, T., and Sheng, M. (2006). Molecular mechanisms of dendritic spine morphogenesis. Curr. Opin. Neurobiol. *16*, 95–101.

Tavazoie, S.F., Alvarez, V.A., Ridenour, D.A., Kwiatkowski, D.J., and Sabatini, B.L. (2005). Regulation of neuronal morphology and function by the tumor suppressors Tsc1 and Tsc2. Nat. Neurosci. *8*, 1727–1734.

Tessier-Lavigne, M., and Goodman, C.S. (1996). The molecular biology of axon guidance. Science *274*, 1123–1133.

UniProt Consortium. (2007). The Universal Protein Resource (UniProt). Nucleic Acids Res. *35* (Database issue), D193–D197.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528–533.

Wang, K., Bucan, M., Grant, S.F., Schellenberg, G., and Hakonarson, H. (2010). Strategies for genetic studies of complex diseases. Cell *142*, 351–353, author reply 353–355.

Weiss, L.A., Arking, D.E., Daly, M.J., and Chakravarti, A.; Gene Discovery Project of Johns Hopkins & the Autism Consortium. (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802–808.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108–1113.

Woolfrey, K.M., Srivastava, D.P., Photowala, H., Yamashita, M., Barbolina, M.V., Cahill, M.E., Xie, Z., Jones, K.A., Quilliam, L.A., Prakriya, M., and Penzes, P. (2009). Epac2 induces synapse remodeling and depression and its disease-associated forms alter spines. Nat. Neurosci. *12*, 1275–1284.

Yang, G., Pan, F., and Gan, W.B. (2009). Stably maintained dendritic spines are associated with lifelong memories. Nature *462*, 920–924.

Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D.H., Law, K., Law, P., Qiu, S., Lord, C., Sebat, J., et al. (2007). A unified genetic theory for sporadic and inherited autism. Proc. Natl. Acad. Sci. USA *104*, 12831–12836.

Zoghbi, H.Y. (2003). Postnatal neurodevelopmental disorders: meeting at the synapse? Science *302*, 826–830.