Supplemental Methods:

Details of StringDB network generation:

In order to create the PPI network in Figure 3, we started with the *de novo* mutations published in each of the six exome studies [1-6] and limited these to events found in probands and intersecting exons or canonical splice sites. The network in Figure 3 was created using all genes with *de novo* truncating variants (defined as nonsense variants, frameshifting variants or variants likely affecting mRNA splicing) as well as six additional genes (*DLG4*, *GRIN2A*, *CASK*, *PSEN1*, *CHD7*, *NLGN1*) in which only missense variants have been observed thus far, but which have important neurobiological roles and/or disease association. In all, we included 158 genes, of which 157 could be identified in the StringDB database (*LTN1* was not found in any human interactions in StringDB).

Data from the StringDB interaction database version 9.05 [7] was used to create the edges of the PPI. We strictly limited our interactions to only human (organism ID 9606) interactions which were based on experimental evidence and an overall combined interaction confidence score of 400 or more. We did *not* include interactions solely based on any of the other StringDB interaction types, such as *in silico* text-mining, co-expression, etc. Overall, we included 85,678 interactions and 12,113 nodes in our analysis.

In order to create the network displayed in Figure 3, we took these steps:

1. Intersected the 157 identified genes based on the criteria above with the human, experimentally-validated StringDB interactions. These form the central red (truncating mutations) and blue nodes (selected missense mutations) in Figure 3, and are connected using thick black lines.
2. Found the two largest connected components. We observed that these were connected via the DLGAP1 protein (see main text for discussion) and added this node as a unfilled (white) node with dashed lines.
3. In addition, we surmised that our set of truncating mutations was likely incomplete, and that many ASD/ID genes may be excluded from the central network simply due to the fact that rare variants in these genes have not yet been discovered. Thus, we “grew” or expanded the network by allowing genes with truncating mutations to be included as “peripheral” nodes if they were within a distance of two (i.e., one intervening node) of the central network. These nodes are drawn as a lighter shade of red and have finely dashed edges. For this analysis, we excluded three proteins (SUMO1, SUMO2 and UBC) which had highly non-specific interactions in StringDB (sumosylation and ubiquitination).
4. We indicated which mutations have only been observed in studies of ID by using half-filled circles. The reciprocal (ASD-only) situation is not indicated due to the fact that there have been nearly ten-fold more ASD exomes sequenced than ID exomes.

5. Lastly, we scaled the sizes of the nodes based the number of times mutations in cases had been observed in each gene (including the mutations from the MIP resequencing data).

Estimating PPI significance:

In order to test if the PPI network of de novo mutations found in the six reviewed exome studies was significantly distinct from randomly formed networks of similar size, we performed two simulation studies. These two simulations were based on random sampling from the complete set of known PPI interactions (i.e., from StringDB) or from random permutation of the existing network. Both simulations were designed to take into account the highly variable degree distribution of interaction networks-- that is, some nodes are highly connected "hub nodes" while other proteins are scantly connected, if at all. The results of the simulations are described in Table S1, and each is described in more detail below.

Stratified node resampling:
For each iteration of the simulation, we randomly selected a stratified (based on degree distribution for the nodes with mutations) set of nodes from the complete StringDB interaction network (limited to interactions with "experimental" evidence and a minimum interaction score of 400). This ensured that the nodes we picked were similar in connectivity and that representation of "hub" nodes and "outlier" nodes was equivalent to that of the actual network. A new PPI graph was generated from each set of stratified random nodes, and the structural characteristics of these graphs were compiled into a null distribution. We primarily examined the average clustering coefficient and the total number of edges of the permuted graphs and compared these to the characteristics of the actual PPI networks. P-values were derived using the empirical distributions from 10,000 iterations of the simulation.

Edge swapping simulation:
In this simulation, we did not alter the set of nodes included in each PPI network, but instead permuted the edges found within the PPI network, thus preserving the degree distribution of the network. Specifically, in each iteration of the simulation, a random sampling of edges (where the number of sampled edges was equal to the total number of edges in the PPI network) in the network were swapped with another eligible edge:

```
u --- v          u     x
                 |     |
x ----y          x     y
```

After randomly swapping edges, we re-computed the average clustering coefficient, size of largest connected component and number of edges for the subgraph of the genes (nodes) with observed mutations and computed the empirical p-value as above. Due to the increased complexity and running time of this simulation, we performed only 1,000 iterations.

Table S1: Summary table of PPI network simulations

	Probands				**Siblings**			
	# of nodes	*# Edges*	*Average Clust. Coeff.*	*Largest Conn. Comp.*	*# of nodes*	*# Edges*	*Average Clust. Coeff.*	*Largest Conn. Comp.*
Missense + Truncating	775	**p < 10e-5** **-----** **p < 10e-4**	**p = 0.009** **-----** **p < 10e-4**	**p < 10e-5** **-----** **p = 0.02**	355	p = 0.10 ----- p = 0.13	p = 0.36 ----- p = 0.052	p = 0.10 ----- p = 0.36
Missense only	640	**p < 10e-5** **-----** **p < 10e-4**	**p = 0.002** **-----** **p < 10e-4**	**p < 10e-5** **-----** **p = 0.006**	317	p = 0.06 ----- p = 0.11	p = 0.26 ----- **p = 0.048**	p = 0.07 ----- p = 0.27
Synon. only	254	p = 0.062 ----- p = 0.08	p = 0.69 ----- p = 0.52	**p = 0.04** ----- p = 0.07	133	p = 0.28 ----- p = 0.315	p = 0.52 ----- p = 0.5	**p = 0.02** ----- p = 0.07
Truncating only	151	p = 0.59 ----- p = 0.53	p = 0.56 ----- p = 0.51	p = 0.89 ----- p = 0.86	39	p = 0.64 ----- p = 0.61	p = 0.50 ----- p = 0.50	p = 0.65 ----- p = 0.61
Missense + Trunc. ***(ASD Only)***	667	**p < 10e-5** **-----** **p < 10e-4**	**p = 0.003** **-----** **p < 10e-4**	**p = 0.0002** **-----** **p = 0.035**	346	p = 0.31 ----- p = 0.39	p = 0.48 ----- p = 0.1	p = 0.74 ----- p = 0.9

Top row p-values are from stratified node resampling simulation

-----

Bottom row p-values are from the edge-swap simulation

Nominally significant (p < 0.05) values highlighted **in bold**

Details of Hidden Species simulation in Figure 1:

In order to estimate the number of genes implicated in ASD under a *de novo/*rare variant model, we used mutations in probands from the four ASD exome studies and a reformulation of the "unseen species problem" (see [8] for review; [9] for application to *de novo* CNVs discovered in autism), where genes with severe *de novo* SNPs in probands are considered "observed species", and binned by their frequency of appearance (i.e., "singletons", "doubletons", etc.). For each category (truncating, truncating+missense), we find the distribution of the number of recurrently mutated genes (i.e., the bins and bin counts of a histogram function). All genes with more than one mutation are included, as is a fraction of the "singleton" mutations (those with only one observed mutation in the four studies). The recurrence counts are shown below:

Table S2: Recurrence of de novo mutations in 4 ASD studies

Recurrence in four exome studies	**Truncating only**	**Truncating + Missense**	**Truncating + Severe missense**
6	–	1	1
5	–	1	0
4	1	3	1
3	1	7	4
2	7	59	26
1	137	946	490

Given these frequencies and frequency counts, we estimated the total number of genes implicated in autism (the total number of species) using the Chao and Lee estimator implemented in the R package SPECIES [10]. The "Percentage of de novo singleton events considered pathogenic" refers to the fraction of the singletons (recurrence = 1) included in the frequency counts.

References:

1 O'Roak, B.J. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250

2 Sanders, S.J. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241

3 Iossifov, I. *et al.* (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299

4 Neale, B.M. *et al.* (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242–245

5 Rauch, A. *et al.* (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674–1682

6 de Ligt, J. *et al.* (2012) Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367, 1921–1929

7 Franceschini, A. *et al.* (2012) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 41, D808–D815

8 BungeFitzpatrick (1993) Estimating the Number of Species: A Review. *Journal of the American Statistical Association* 88, 364–374

9 Sanders, S.J. *et al.* (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70, 863–885

10 Wang, J.-P. 01-Apr-(2011), SPECIES: An R Package for Species Richness Estimation. *Journal of Statistical Software*. [Online]. Available: http://www.jstatsoft.org/. [Accessed: 30-Aug-2011]