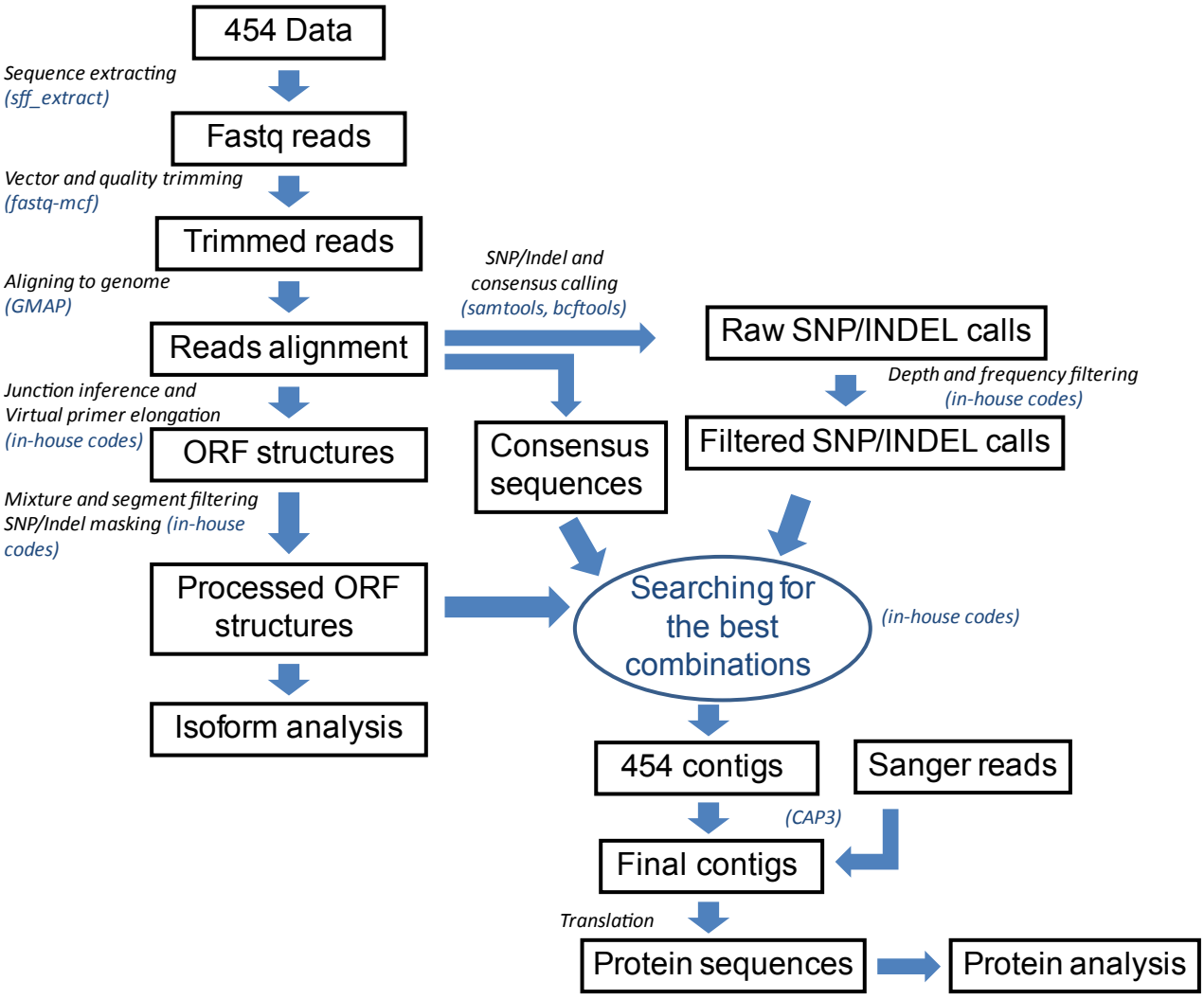
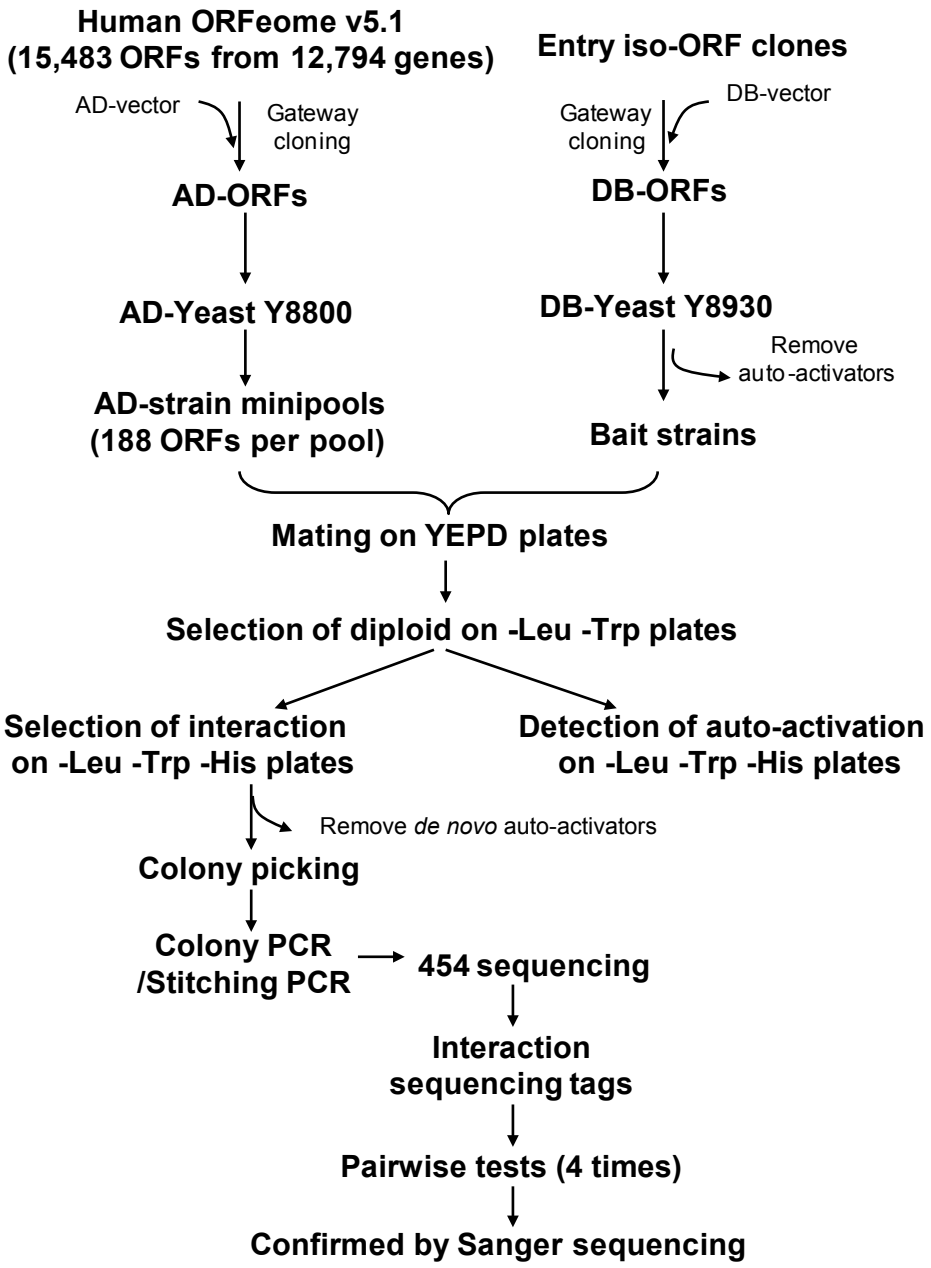
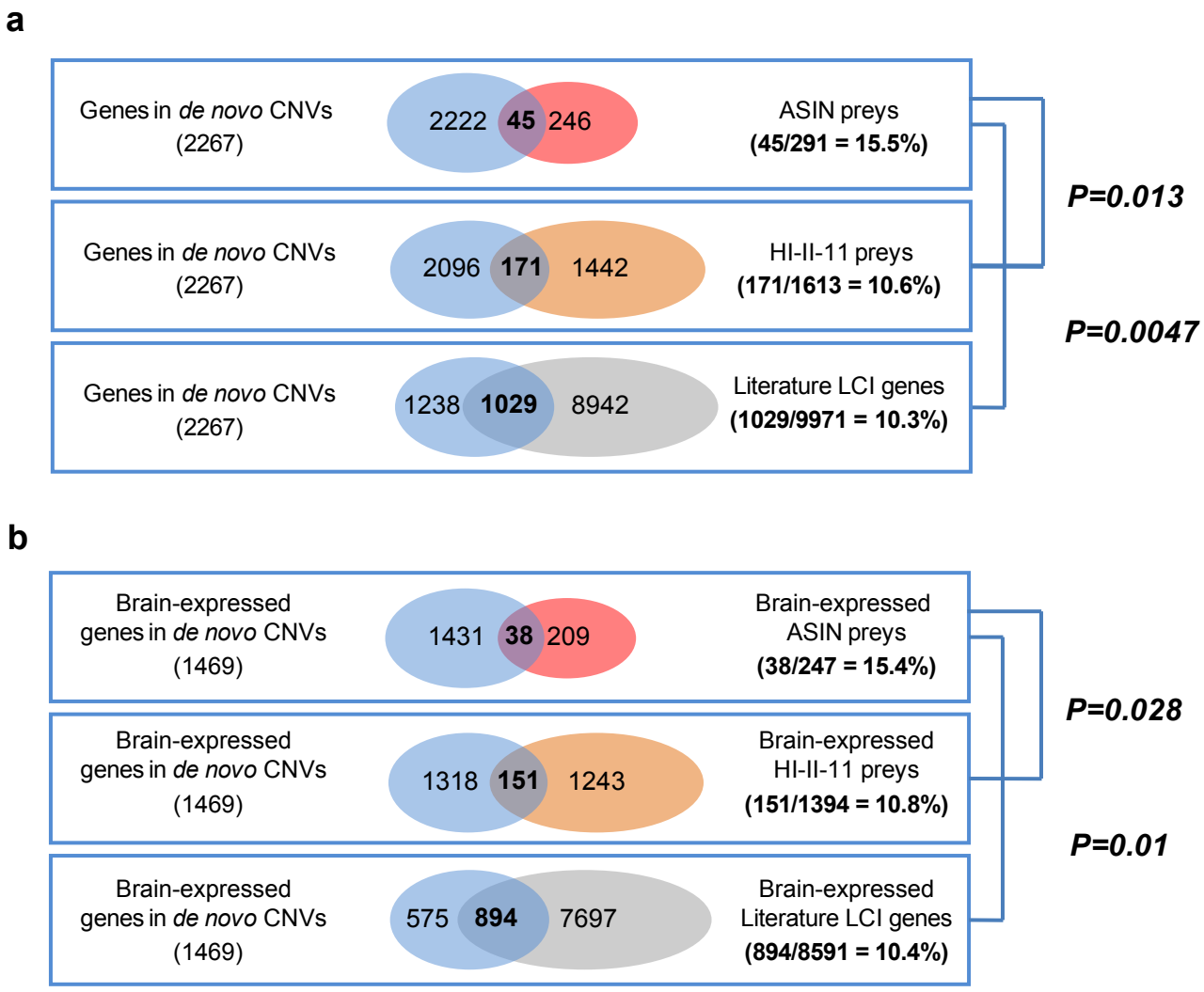


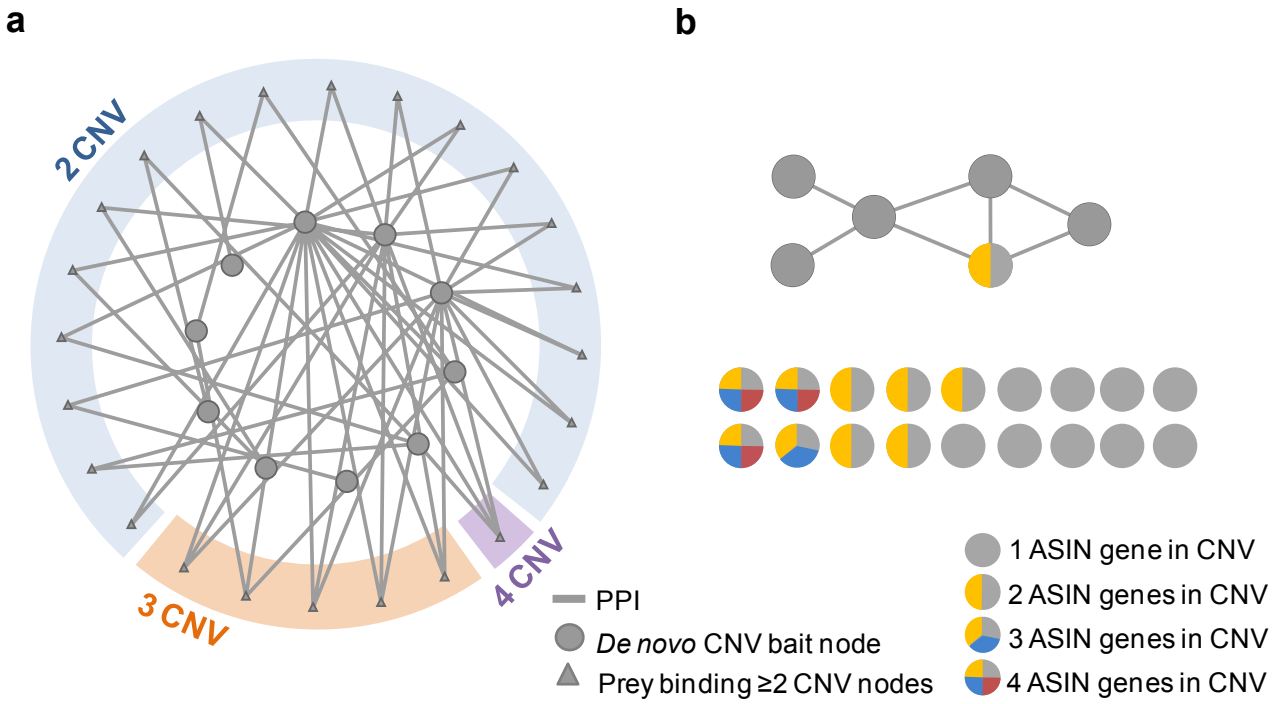
Supplementary Figure 1. The reference-based splicing isoform assembly pipeline. Raw 454 reads were aligned to human reference genome and isoforms were assembled. The final contigs were constructed by integrating additional Sanger sequencing reads.



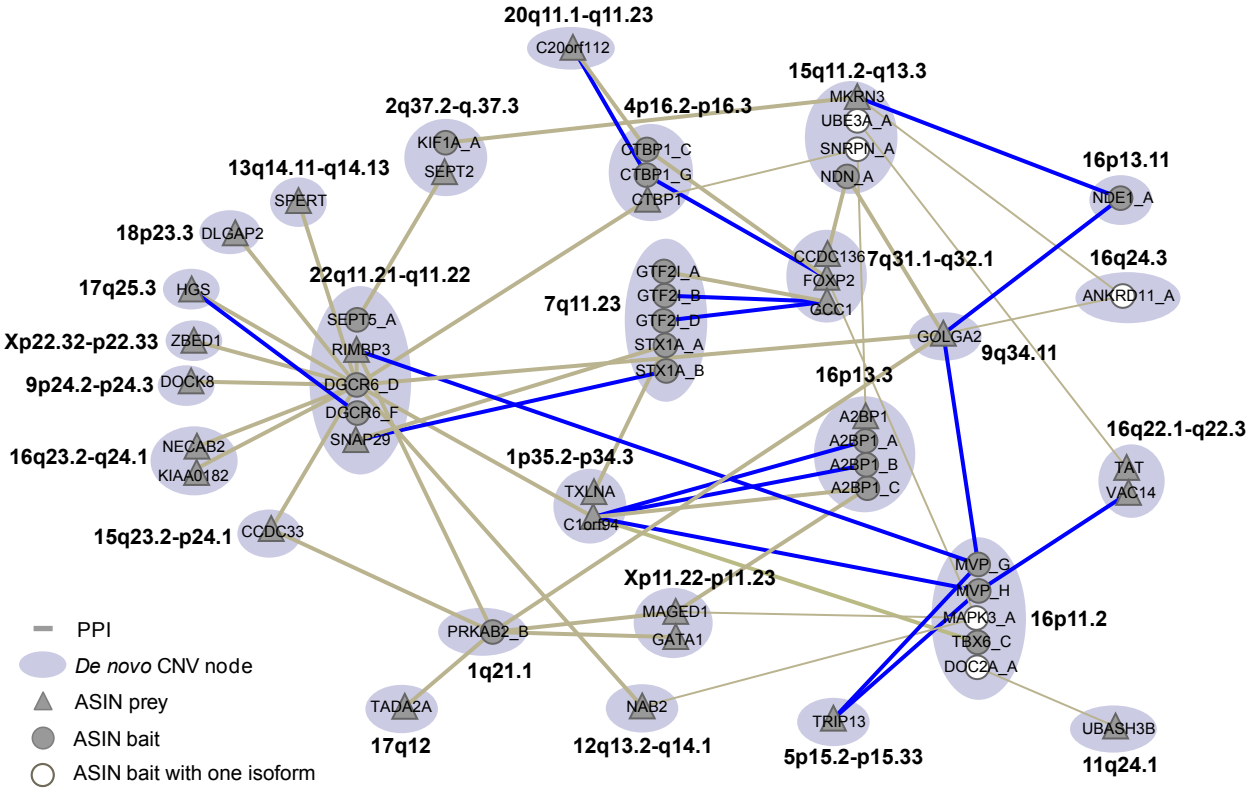
Supplementary Figure 2. The yeast two-hybrid screening pipeline. Isoform-ORFs were screened against minipools of the human ORFeome 5.1. Stitch-seq was performed to identify the interacting pairs. Positive interactions were retested in a pairwise format four times and confirmed by Sanger sequencing.



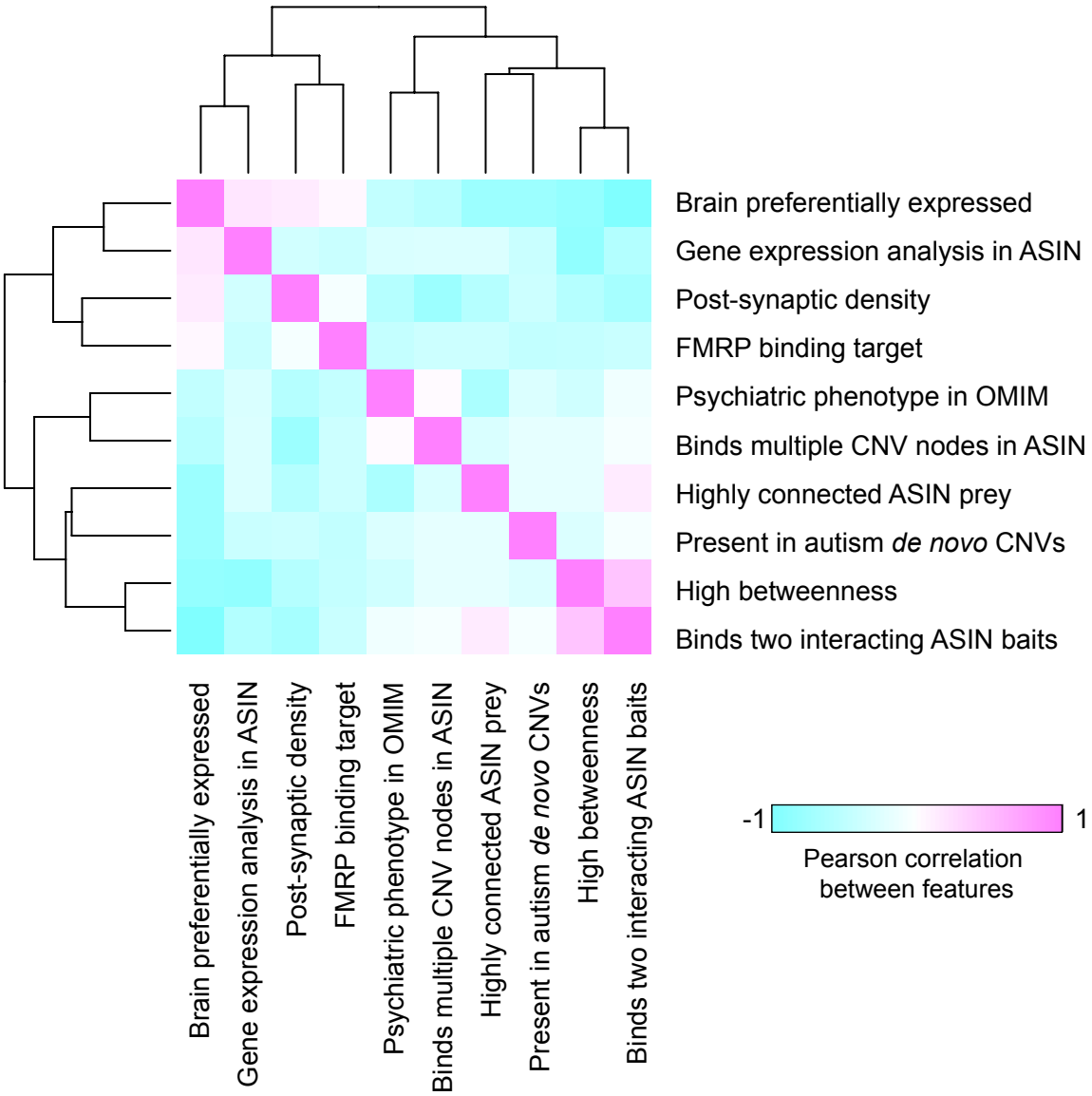
Supplementary Figure 3. ASIN is enriched in preys from *de novo* CNVs. All genes (a total of 2267) within 198 *de novo* validated CNVs identified in the ASD patients from genetic studies (**Supplementary Methods**) were mapped to the prey space of ASIN and HI-II-11 networks, as well as to all genes in the literature (LCI) network (because the bait-prey directionality is often unknown for LCI PPIs). **(a)** The Venn diagrams built for each dataset separately (ASIN, upper panel; HI-II-11 middle panel and LCI lower panel) demonstrates significant enrichment of ASIN preys in genes from the *de novo* CNVs compared to HI2-II-11 preys (Fisher's exact P=0.013) and LCI genes (Fisher's exact P=0.0047). **(b)** Restricting the analysis to the brain-expressed genes in each dataset still results in significant enrichments (ASIN *vs.* HI-II-11 P=0.028; ASIN *vs.* LCI P=0.01).



Supplementary Figure 4. Control CNV-prey and CNV-CNV networks constructed using LCI PPIs. All control networks were constructed from the randomly selected genomic regions with the same number of genes with PPIs and the same number of interacting partners (preys) as in ASIN but connected by the LCI interactions (see **Methods**). **(a)** This CNV-prey control network was selected from 10,000 control networks to represent an example with the greatest number of preys (only 1) that interact with four CNV nodes using LCI PPIs. No networks with preys that bind to 5, 6 or higher number of CNV nodes were observed. Preys binding to one CNV node are not shown. **(b)** This CNV-CNV control network was selected from 10,000 control networks to represent the network that connects the largest number of CNV nodes using LCI PPIs. All networks were visualized using Cytoscape.

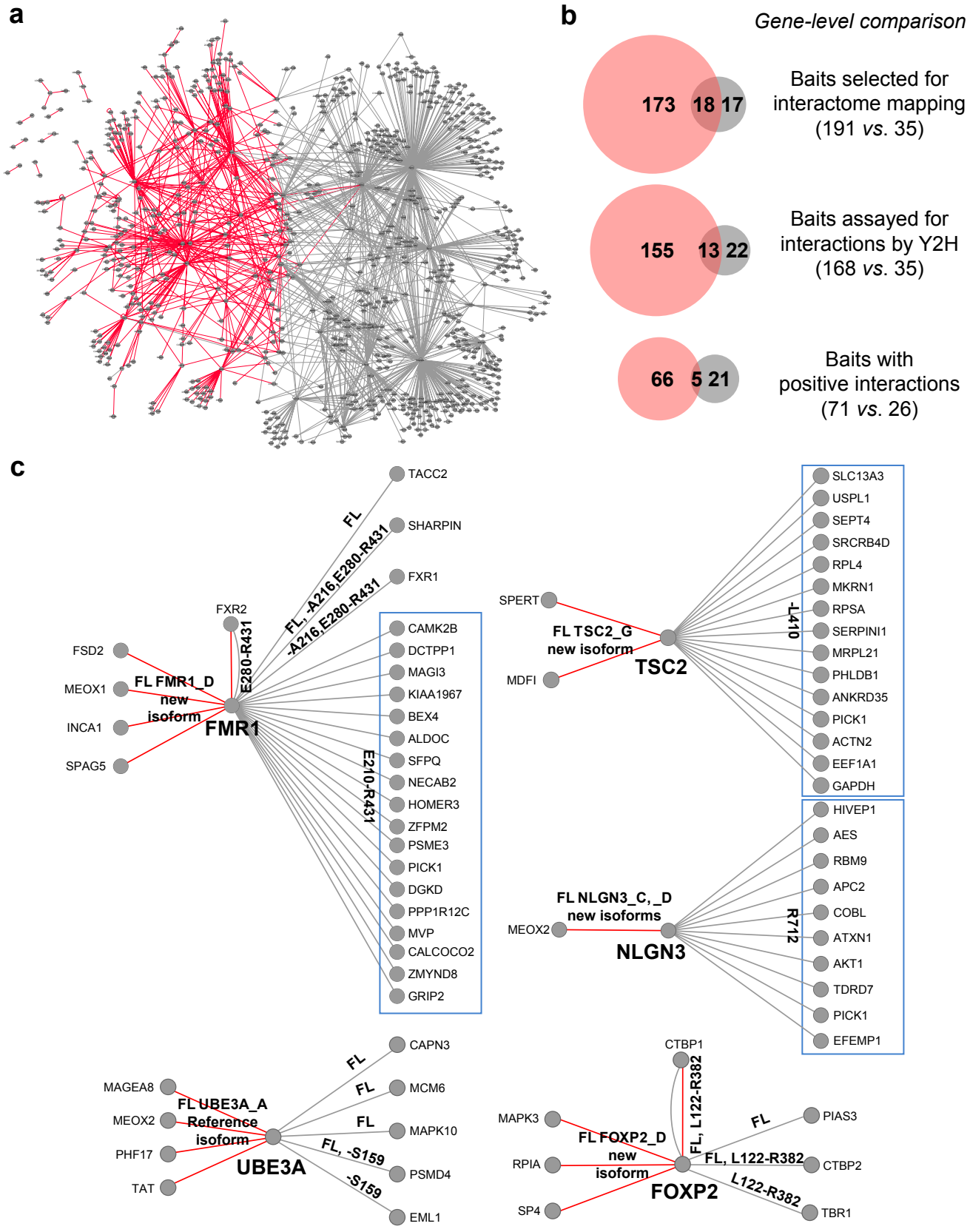


Supplementary Figure 5. The isoform-resolved ASIN CNV-CNV network. The isoform-resolved CNV-CNV network connects ASIN bait isoforms with preys via 56 interactions. 34% of interactions are supported exclusively by the non-reference isoforms of the genes that have at least two isoforms cloned in this study. Gray circles are ASIN baits with at least two cloned isoforms; white circles are baits with only one isoform cloned; gray triangles are preys. Transparent puple circles grouping the genes represent different CNVs. Each line represents an ASIN interaction; blue lines are interactions supported exclusively by the non-reference isoforms; tan lines – exclusively by the reference isoforms; thin tan lines are interactions of baits with only one isoform. The network was visualized using Cytoscape.



Supplementary Figure 6. The correlation between different sources of evidence used for prey ranking. Five different sources of evidence from the ASIN analyses and five from the public domain were plotted against each other and tested for correlation. 98% of the PCCs values are <0.36, with exception of one pair, "High network betweenness" and "binds two interacting ASIN baits" (PCC=0.57). The median PCC=-0.11 indicates low or no correlation between the evidence sources.

Supplementary Figure 7



Supplementary Figure 7. Comparison of ASIN and Sakai interactomes. (a) ASIN and Sakai interactomes are complementary. ASIN protein-protein interactions (PPIs) are shown in red; Sakai[1] interactions are shown in gray. There is almost no overlap between interactions detected in these two studies. **(b)** Venn diagram shows the number of baits shared by ASIN (red) and Sakai (gray) studies. Only five genes with positive interactions are common between these two studies. **(c)** Graphical representation of the isoform–level PPIs of the five baits with positive interactions from ASIN and Sakai interactomes. Only two interactions are shared between ASIN and Sakai interactomes. This is likely due to different isoforms/fragment sequences tested in each study, as indicated next to each interaction. ASIN PPIs are shown in red, Sakai PPIs are shown in gray, FL – full-length protein, fragments sizes (in amino acid residues) are shown next to each interaction of the isoform/fragment (extracted from Sakai). For example, FMR1 interaction with FXR2 is supported by both, ASIN and Sakai despite the differences in the FMR1 bait sequences (ASIN interaction was identified by testing new brain-expressed cloned full-length isoform FMR1_D, whereas Sakai interaction was identified by testing an FMR1 fragment spanning the following amino acids (E280-R431). All boxed interactions were detected using indicated protein fragments (E210-R431 for FMR1; -L410 for TSC2 and R712 for NLGN3).

Supplementary Tables

Supplementary Table 1. Comparison of ASIN preys with HI-II-11 preys and LCI genes. Enrichment analysis of ASIN preys from the *de novo* autism CNVs and those annotated with psychiatric phenotype in OMIM. All p-values were calculated using the one-sided Fisher's exact test.

Evidences	Number of ASIN preys	Number of HI-II-11 preys	Fisher's exact p-value (ASIN *vs.* HI-II-11)	Number of LCI genes	Fisher's exact p-value (ASIN *vs.* LCI)
Autism *de Novo* CNVs	45 out of 291 (15.5%)	171 out of 1613 (10.6%)	**0.013**	1029 out of 9971 (10.3%)	**0.0047**
Psychiatric phenotypes in OMIM	15 out of 291 (5.2%)	42 out of 1613 (2.6%)	**0.02**	291 out of 9971(2.9%)	**0.028**
	Restricted to brain-expressed genes				
Autism *de Novo* CNVs	38 out of 247 (15.4%)	151 out of 1394 (10.8%)	**0.028**	894 out of 8591(10.4%)	**0.01**
Psychiatric phenotypes in OMIM	15 out of 247 (6.1%)	40 out of 1394 (2.9%)	**0.012**	275 out of 8591(3.2%)	**0.016**

Supplementary Table 2. Comparison of ASIN CNV-prey and CNV-CNV networks with 10,000 control networks. The table lists median, range and empirical p-values for comparison of ASIN with 10,000 control networks generated using randomly selected genomic CNV regions with the same number of genes with PPIs and the same number of preys as ASIN CNVs but connected by PPIs from HI-II-11 or from literature LCI.

	CNV nodes in ASIN	**10,000 simulated CNV-prey networks with HI-II-11 PPIs**			**10,000 simulated CNV-prey networks with LCI PPIs**		
		Median	Range	Empirical p-value	Median	Range	Empirical p-value
CNV-prey networks							
# of preys binding to at least 2 CNV nodes	55	21	[1,38]	<0.0001	8	[0,46]	<0.0001
# of preys binding to at least 3 CNV nodes	27	3	[0,17]	<0.0001	0	[0,14]	<0.0001
# of preys binding to at least 4 CNV nodes	12	0	[0,3]	<0.0001	0	[0,1]	<0.0001
# of preys binding to at least 5 CNV nodes	4	0	[0,0]	<0.0001	0	[0,0]	<0.0001
# of preys binding to at least 6 CNV nodes	3	0	[0,0]	<0.0001	0	[0,0]	<0.0001
CNV-CNV networks							
Size of the largest connected component	27	2	[1,12]	<0.0001	1	[1,7]	<0.0001

To ensure high quality of control networks, millions of them were generated in order to randomly select a set of 10,000 networks with exactly the same parameters as ASIN CNV-preys network. Two sets of 10,000 networks each were created: one using interactions from the HI-II-11 dataset, and the other using literature LCI PPs. These networks were used to estimate the statistical significance of the ASIN preys binding to multiple (≥2) unique CNVs in CNV-prey network.

Supplementary Methods

1. High-throughput isoform assembly pipeline

Sequencing reads assembly

The 454 GS FLX sequencing data were assembled and analyzed to identify full-length alternatively spliced ORFs using the in-house reference-based assembly pipeline (**Supplementary Fig. 1**). Raw 454 sequencing data were converted into fastq format, fastq-mcf was used to trim vector sequences and low quality base calls. The processed reads were aligned to hg19 human reference genome (Genome Reference Consortium GRCh37) using GMAP[2]. The output of the alignment in SAM[3] format was subject to further analysis.

The alignments of 454 reads were analyzed using in-house codes. Only reads aligned to the genomic regions flanked by the designed primer sequences were processed. Each genomic position covered by 454 reads was annotated as either exonic or intronic depending on the quality scores of bases and gaps covering this position. The quality score of a gap was calculated as the average of the quality scores of the two flanking nucleotides. The consensus quality score (CQS) was calculated by using the following formula:

(1) $$CQS = \sum_{i=1}^{m} \frac{x_i}{i} + \sum_{j=1}^{n} \frac{y_j}{j}$$

where *m* id the number of forward reads, *n* is the number of reverse reads aligned to a genomic position, x_i is the i th highest quality score of the bases in forward reads, and y_j is the j th highest quality score of bases in reverse reads. The CQS of the gaps was calculated in similar way. “Intronic” was the default assignment for each position. A position was annotated as “exonic” only if the CQS of aligned bases was greater than CQS of aligned gaps.

The ends of the ORF clones were extended using primer information. An in house binomial model was developed and implemented to account for the ambiguity in the reads assembly. Ambiguous isoforms were removed from the analysis. After all positions within loci were annotated as “exonic” or “intronic”, isoform sequences were assembled using only exonic bases. To ensure a reliable assembly, isoforms with missing exonic regions or with low coverage region (<2 reads) were discarded. Mismatches and insertions, as well as gaps less than 30 nucleotides were annotated as SNPs or indels, and were *in silico* masked to allow high-quality splice isoform assembly and analysis. SNPs and indels were taken into consideration during translation into the protein sequences.

Full-length isoform assembly

SAMtools[3] were used to call the variance sites (SNP/indels) and the consensus nucleotides at exonic genomic positions. When multiple sets of non-reference allelic sites were called for the same genomic region, only the sites with the estimated frequency of >0.5 and with >2 reads coverage were selected for further analysis. To obtain the best combination of the variant sites, the overall quality score (OQS) of each possible combination of the sites was calculated as:

(2) $$OQS = -10 \times lg\left(1 - \prod_i \left(1 - 10^{-\frac{x_i}{10}}\right)\right)$$

where x_i is the quality score of each consensus/variant site. The combination with the highest OQS was used as the final sequence.

Additional Sanger sequencing was performed to improve the assembly quality of the isoform terminal regions. Phred[4] was used to extract sequences from Sanger sequencing raw data. Sanger reads were required to have at least 50 bases with non-zero quality scores to be further analyzed. In addition, a Sanger read was considered to be derived from the same clone if during alignment to the corresponding contig with BLAST (bl2seq) it had a minimum of 50 nucleotides aligned, with sequence identity greater than 95%. Sanger reads were integrated with corresponding 454 contigs by using CAP3[5] to generate the final isoform sequences. The variant sites previously called by SAMtools in some cases were altered due to Sanger data integration; therefore the final sequences were aligned back to the 454 contigs with variant sites masked in order to annotate SNP and Indel sites supported by the final sequences.

2. Experimental identification of protein-protein interactions using the yeast two-hybrid (Y2H) system

Yeast strains and Y2H vectors

S. cerevisiae strains Y8930 (*MATα*) and Y8800 (*MATa*) (a kind gift from X. Xin and C. Boone, U. Toronto) were generated from PJ69-4α and PJ69-4α[6]. The strain Y8930 contains three Gal4p inducible reporter genes: *GAL::HIS3*, providing growth on media lacking Histidine, *GAL7::LacZ*, for colorimetric detection of Gal4p activity, and *GAL2::ADE2*, providing growth on media lacking Adenine. The genotype of this strain is: *trp1-901leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ cyh2^R gal::HIS3@LYS2 GAL2::ADE2 GAL7::lacZ@MET2*.

AD vectors pDEST-AD-CYH2 and pDEST-DB were generated by modifying pPC86 and pPC97, respectively. Detailed information of the vectors is provided elsewhere[6, 7, 8, 9]. The entry clones of Human ORFeome v5.1 were Gateway cloned into pDEST-AD-CYH2, introduced into Y8800 (*MATa*), and pooled as small libraries (smart pools) containing 188 clones each[10]. The Iso-ORFs were transferred from entry clones into pDEST-DB and introduced into Y8930 (*MATα*).

Generating Y8930 bait strains and Y8800 prey strains for iso-ORFs

The unique set of entry clones of Iso-ORFs were cherry-picked from *E. coli* glycerol stocks and cultured in 96-well deep wells as liquid LB culture containing spectinomycin, followed by plasmid extraction using a Qiagen Biorobot 8000 (Qiagen). Iso-ORF entry clones were cloned into pDEST-DB and pDEST-AD vectors using Gateway recombination LR reaction (Invitrogen), transformed into *E. coli*, amplified, and plasmids were prepared as described above. The minipreps of baits were introduced into yeast strain Y8930 (*MATα*) to create bait

strains, and preys were introduced into Y8800 (*MATa*) using Lithium Acetate transformation method[11].

Auto-activators removal

Auto-activation of transcription of *Gal1-HIS3* reporter gene in DB-X strains was detected by growing the DB-X strains on agar plates containing SC-Leu-His +1mM 3AT at 30°C for 3 days prior to Y2H screening. *De novo* auto-activation in DB-X was identified by growth on agar plates containing SC-leu-His +1mM3AT +1µg/ml chycloheximide during screening and retests. The AD vector pDEST-AD-CYH2 with increased sensitivity to chycloheximide is lost during yeast cell division and only DB-X plasmid is kept. The diploid strains showing growth are considered to be auto-activators[6, 12]. All auto-activators were removed from the subsequent Y2H screening.

Y2H screening

To generate diploid strains, individual DB-X strains were mated with pooled AD-Y strains on YEPD plates. Known interacting pairs were added as controls. After incubation at 30°C overnight, plates were replicated into SC-Leu-Trp-His and grown at 30°C for five days. Then, yeast colonies were picked and grown in 96-well plates with SC-Leu-Trp liquid medium for two days at 30°C. Subsequently, colonies were spotted on -Leu-Trp plates, grown for two days at 30°C, and replicated to (1) SC-Leu-Trp-His + 1mM 3AT plates to test the activity of *HIS* reporter gene, and (2) Sc-Leu-His + 1mM 3AT + 1µg/ml chycloheximide plates to identify *de novo* auto-activation of DB-Xs. Phenotypes of colonies were scored after three days of growth at 30°C. Colonies that scored positive in selective medium and negative in autoactivation control plates were scored as positives and cultured at 30°C for 24 hours in SC-Leu-Trp and used in the subsequent experiments.

Stitch-Seq: 454 FLX sequencing of stitched interacting pairs

To identify interacting protein pairs, stitch-seq and 454 FLX sequencing was used[13]. Briefly, colonies scored as positive were amplified using colony-PCR of DB-X and AD-Y followed by a stitching-PCR to fuse DB-X and AD-Y through a linker region. The stitched PCR products were sequenced by massively parallel 454 FLX sequencing. The reads containing the linker region were identified from 454 GS FLX sequencing data using string match. The interacting gene pairs were identified by mapping the sequences flanking the linker region to prey sequences from human ORFeome for AD side, and to the longest ORFs of baits for DB side, using BLASTN (mismatches allowed) with E-value cutoffs of 10^{-3} and 10^{-4}, respectively. Protein-protein interactions supported by at least two sISTs were considered as confident interactions.

Y2H verification

Four rounds of pair-wise Y2H retests between isoforms of each gene as baits and all interacting partners from ORFeome 5.1 as preys were performed in matrix format as described above for the pools. Positive yeast colonies were picked, yeast colony-PCR reactions performed to amplify ORFs in DB vector and AD vector, and products were Sanger sequenced to confirm

interacting pairs to generate a high-confident isoform-level protein-protein interactome network (**Supplementary Fig. 2**).

3. Datasets used in the study

HI-II-11 dataset

A new phase of the human interactome mapping project has generated and retested the Human Interactome II-11 Prepublication data set (available for download from http://interactome.dfci.harvard.edu/H_sapiens/) containing ~14,000 novel binary interactions derived from yeast two-hybrid screens over a search space corresponding to a matrix of ~15,000 x 15,000 human ORFs. This HI-II-11 network was used as a background control for ASIN because HI-II-11 and ASIN share the same prey search space (~15,000 ORFs), and they were both generated in the same laboratory using the same experimental conditions. Furthermore, these two networks have comparable proportions of the brain-expressed preys (85% in ASIN *vs.* 87% in HI-II-11, **Supplementary Data 9**), making HI-II-11 the best currently available control PPI network for ASIN.

Binary literature-curated interactions dataset LCI

To obtain a set of protein-protein interactions from the literature comparable in quality to ASIN, we restricted the literature data to human high-quality binary physical interactions. A previously curated interaction dataset extracted from seven major databases (MIPS, BIND, DIP, MINT, IntAct, BioGRID and PDB)[14] was updated to include new PPIs reported after December 2010[1, 13, 15, 16, 17, 18] using identical filtering criteria. The final dataset was composed of 37,654 non-redundant high-quality human binary literature-curated interactions (binary-LCI) between 9,987 proteins and was frozen as of March 2012.

Genes from the autism-relevant rare *de novo* CNVs

To compile a list of genes present within *de novo* CNVs, all rare validated *de novo* CNVs identified in the autism patients were extracted from five publications[19, 20, 21, 22, 23]. CNVs were annotated as "validated" when the presence of the CNV was confirmed either using an independent approach (such as additional microarray scan, FISH, qPCR, *etc*.), or when the same CNV was called in the same patient in two different studies. A list of 198 validated rare *de novo* CNVs affecting 2,267 genes was compiled. To simplify the analyses, overlapping CNVs were merged into unique CNVs regions, spanning the union of the overlapping CNVs, resulting in 123 merged unique validated autism *de novo* CNVs.

Brain-expressed genes

The list of brain-expressed genes was obtained from the study of the human brain transcriptome through development and adulthood[24]. This dataset consisted of 1,340 brain samples and was generated by dissecting regions from 57 clinically unremarkable postmortem brains of donors ranging in age from 6 post conceptual weeks to 82 years, which were divided into 15 periods based on age. The expression levels of 17,181 protein-coding genes within each

sample were assayed using the Affymetrix GeneChip Human Exon 1.1 ST Array platform. A gene was defined as "brained-expressed" when it had a $\log_2$-transformed signal intensity ≥6 in at least one sample and a mean DABG P<0.01 in at least one brain region of at least one period. Based on these filtering criteria, 14,990 out of 17,181 genes (hg19) were defined as "brain-expressed".

Supplementary References

1. Sakai Y*, et al.* Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med* **3**, 86ra49 (2011).

2. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).

3. Li H*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

4. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185 (1998).

5. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-877 (1999).

6. Dreze M*, et al.* High-quality binary interactome mapping. *Methods Enzymol* **470**, 281-315 (2010).

7. Vidal M. The Yeast Two-Hybrid System. (ed^(eds Fields PBaS). Oxford University Press (1997).

8. Vidalain PO, Boxem M, Ge H, Li S, Vidal M. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**, 363-370 (2004).

9. Boxem M*, et al.* A protein domain-based interactome network for C. elegans early embryogenesis. *Cell* **134**, 534-545 (2008).

10. Rual JF*, et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).

11. Schiestl RH, Gietz RD. High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr Genet* **16**, 339-346 (1989).

12. Walhout AJ, Vidal M. A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res* **9**, 1128-1134 (1999).

13. Yu H*, et al.* Next-generation sequencing to generate interactome datasets. *Nat Methods* **8**, 478-480 (2011).

14. Cusick ME*, et al.* Literature-curated protein interaction datasets. *Nat Methods* **6**, 39-46 (2009).

15. Bandyopadhyay S*, et al.* A human MAP kinase interactome. *Nat Methods* **7**, 801-805 (2010).

16. Kahle JJ*, et al.* Comparison of an expanded ataxia interactome with patient medical records reveals a relationship between macular degeneration and ataxia. *Hum Mol Genet* **20**, 510-527 (2011).

17. Vinayagam A*, et al.* A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal* **4**, rs8 (2011).

18. Wang J*, et al.* Toward an understanding of the protein interaction network of the human liver. *Mol Syst Biol* **7**, 536 (2011).

19. Sebat J*, et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449 (2007).

20. Marshall CR*, et al.* Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-488 (2008).

21. Pinto D*, et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

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

23. Levy D*, et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897 (2011).

24. Kang HJ*, et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).