ARTICLES
https://doi.org/10.1038/s41588-018-0130-z
nature genetics

An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders

Siwei Chen[1,2,3,7], Robert Fragoza[1,2,3,7], Lambertus Klei[4], Yuan Liu[1,2], Jiebiao Wang[5], Kathryn Roeder[5,6*], Bernie Devlin[4*] and Haiyuan Yu[1,2*]

Identifying disease-associated missense mutations remains a challenge, especially in large-scale sequencing studies. Here we establish an experimentally and computationally integrated approach to investigate the functional impact of missense mutations in the context of the human interactome network and test our approach by analyzing ~2,000 de novo missense mutations found in autism subjects and their unaffected siblings. Interaction-disrupting de novo missense mutations are more common in autism probands, principally affect hub proteins, and disrupt a significantly higher fraction of hub interactions than in unaffected siblings. Moreover, they tend to disrupt interactions involving genes previously implicated in autism, providing complementary evidence that strengthens previously identified associations and enhances the discovery of new ones. Importantly, by analyzing de novo missense mutation data from six disorders, we demonstrate that our interactome perturbation approach offers a generalizable framework for identifying and prioritizing missense mutations that contribute to the risk of human disease.

Mutations disrupting the function of proteins are recognized as an important source of risk for developmental disorders, such as intellectual disability[1,2], autism spectrum disorder (ASD)[3] and congenital heart defects[4]. Whole-exome sequencing (WES) has produced a boon of findings linking de novo mutations to risk for developmental disorders[5–19]. Not all mutations are simple to interpret as causing a loss of gene function. Missense mutations are especially difficult; although there are bioinformatics tools to predict the level of damage[20–22], these annotators are far from perfect. This is a critical deficiency because the majority of coding mutations are missense. Here we show that one key feature in evaluating the disruptiveness of mutations is whether they fall in known or predicted protein–protein interaction interfaces and their likelihood to disrupt these interactions.

Large-scale studies of known disease-associated mutations have already reported a strong association with binding interfaces of protein interactions[23,24]. The major bottleneck for wide application of this feature is limited knowledge about the set of interactions and the binding interfaces of all interactions. To experimentally evaluate the impact of mutations on protein interactions, we establish a high-throughput mutagenesis and interactome-scanning pipeline for generating site-specific mutant clones and testing corresponding mutant protein interactions. However, such a pipeline cannot evaluate the impact of missense mutations on many interactions, because high-throughput interaction assays are limited in their coverage[25–27]. For this reason, we also explore a computational approach for systematically examining the functional impact of missense mutations on protein interactions. This approach builds on our newly established full-interactome interface predictions[28] to computationally predict the impact of all missense mutations on all associated interactions. Here we apply our experimental and computational approaches in tandem, which can be applied to any WES study.

To evaluate the effectiveness of our integrated experimental–computational approach, we focus on 2,821 de novo missense (dnMis) mutations identified from WES of ~2,500 families from the Simons Simplex Collection (SSC)[29] (Supplementary Table 1). The SSC targets the study of ASD through a cohort of parent–offspring trios or quads with two unaffected parents, an ASD proband and, for most families, an unaffected sibling[30]. Previous analyses of the SSC data have reported significantly higher de novo mutation rates in ASD probands versus unaffected siblings across various mutation types, from copy-number variants[29,31,32], to frameshift indels[33], to missense mutations[12–14,16]. While a number of risk de novo copy-number[29,31,32,34,35] and protein-truncating[12–14,16,33] variants have been identified, exactly which dnMis mutations play a role and to what extent are open questions. We applied our integrated framework to evaluate the effect of 1,733 dnMis mutations within a protein interactome framework, aiming to identify potentially disease-contributing dnMis mutations. Although there are many ways by which a missense mutation can impact a protein's function, such as by destabilizing protein folding, we evaluate the disruptiveness of a mutation within our framework exclusively on its capacity to disrupt protein interactions, measured experimentally or through prediction. We further compare the network properties of proteins impacted by interaction-disrupting and non-disrupting dnMis mutations, using unaffected siblings as negative controls throughout. While our analyses focus on dnMis mutations in ASD, the integrated experimental–computational approach provides a generalizable framework for investigating the impact of missense mutations uncovered by WES for human diseases.

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [4]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [5]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA. [7]These authors contributed equally: Siwei Chen, Robert Fragoza. *e-mail: roeder@andrew.cmu.edu; devlinbj@upmc.edu; haiyuan.yu@cornell.edu

Results

Proband dnMis mutations are enriched on interaction interfaces. We previously reported that inherited in-frame disease-associated mutations are significantly enriched on protein interaction interfaces and demonstrated that alteration of specific protein interactions is crucial in the pathogenesis of many disease-associated genes[36]. To explore the relationship between non-inherited dnMis mutations and autism, we used a structurally resolved three-dimensional human interactome network[36,37] to examine where dnMis mutations reside with respect to interaction interfaces. We found that in probands, dnMis mutations are significantly enriched on interaction interfaces. While interaction interfaces cover 30.1% of the proteins harboring these mutations, 38.2% of the mutations fall in interaction interfaces (1.27-fold, $P = 2.9 \times 10^{-3}$ by a two-tailed exact binomial test). In contrast, dnMis mutations in siblings fall in interaction interfaces on corresponding proteins at an expected rate (observed 37.6% versus expected 36.5%, 1.03-fold, $P = 0.76$). Thus, disruption of specific interactions could contribute to ASD etiology for dnMis mutations in probands (Fig. 1a), underscoring the functional significance of dnMis mutations on protein interaction interfaces.

Proband dnMis mutations have elevated disruption rates. We next explored the impact of dnMis mutations on protein interactions by intersecting all 2,821 dnMis mutations with 59,073 human protein interactions, a comprehensive set of high-quality physical interactions compiled in HINT[38] from eight widely used interaction databases, including BioGRID[39], MINT[40], iRefWeb[41], DIP[42], IntAct[43], HPRD[44], MIPS[45] and the PDB[46]. Of these mutations, 1,733 are on proteins with at least 1 known interaction within the current human interactome data set. To experimentally assess the impact of a subset of these mutations, 208 individual clones were generated carrying dnMis mutations—corresponding to 109 in probands and 99 in siblings, respectively—using Clone-seq, a massively parallel site-directed mutagenesis pipeline[24]. Protein interactions amenable to yeast two-hybrid (Y2H) were then tested, yielding 667 total protein interactions corresponding to 151 of our cloned dnMis mutations (Fig. 1b; Methods).

To explore the remaining dnMis mutations and interactions untested by Y2H, we applied a two-tiered computational approach that first predicts whether a particular residue is an interface residue using Interactome INSIDER[28], a unified machine-learning framework comprising the first full-interactome map of human interaction interfaces. To determine whether a particular mutation is deleterious, we used PolyPhen-2 (PPH2)[20] predictions: if a particular residue is predicted to be an interface residue and its mutation is scored as 'probably damaging' by PPH2, that mutation was predicted as interaction-disrupting; if a mutation is unlikely to occur at an interface residue and is scored as 'benign' by PPH2, it was predicted as interaction non-disrupting (Fig. 1b; Methods).

To evaluate the performance of our computational predictions, we applied this two-tiered prediction approach to our 667 experimentally tested protein interactions and obtained an accuracy of 80.8% (sensitivity: 65.0%, specificity: 82.5%). Moreover, when our approach was applied to a previously published, independent data set of 204 disease-associated mutations and their impact on protein–protein interactions[24], we obtained a similar prediction performance (accuracy: 77.4%, sensitivity: 81.0%, specificity: 75.0%).

We then analyzed the distribution of disrupted interactions across ASD probands and unaffected siblings. Examining our experimental data showed that 74/361 (20.5%) tested interactions were disrupted in probands. In contrast, only 21 out of 208 (10.1%) interactions were disrupted in unaffected siblings. Modeling the count of disruptions per subject with a negative binomial model, using case status as the predictor, yielded a 2.54-fold

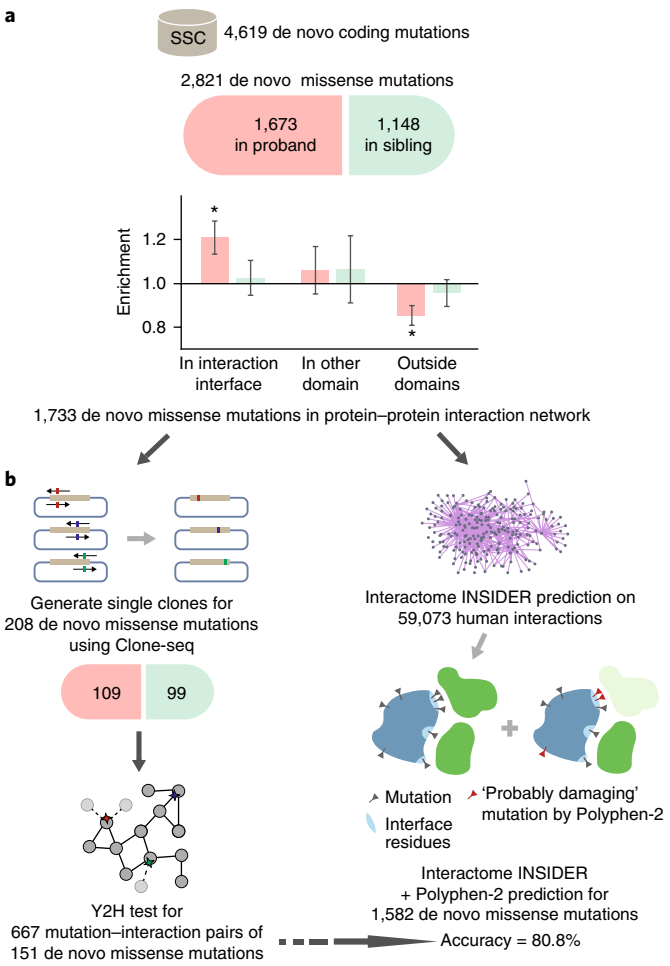

Fig. 1 | Workflow of our integrated experimental–computational interactome perturbation framework. a, Distribution of dnMis mutations from the SSC across different protein locations. Enrichment was calculated as the ratio of the observed fraction of dnMis mutations that occur on interaction interfaces over the fraction of interface residues on corresponding proteins (expected fraction). *P* values were calculated using a two-tailed exact binomial test (*$P < 0.05$; Methods). The error bars indicate standard error. **b**, Experimental (left) and computational (right) pipelines for assessing the functional impact of dnMis mutations on protein–protein interactions.

higher rate of disruptions in probands ($P = 0.012$, Fig. 2a and Supplementary Table 2; Methods). This sharp contrast in interaction disruption rate suggests that disruption of the interactome network by dnMis mutations contributes to autism etiology in probands. Combining the experimental data with predictions for all remaining dnMis mutations and interactions, there was again a significant, 2.34-fold higher disruption rate for probands (25.3%) versus unaffected siblings (10.8%, $P = 5.6 \times 10^{-4}$, Fig. 2b). Furthermore, the predicted disruptions alone showed significantly higher rate of disruption in probands than siblings (2.15-fold, $P = 0.013$, Fig. 2c). These observations suggest that dnMis mutations in ASD probands are of higher functional consequence than those in unaffected siblings. Therefore, interaction-disrupting mutations identified by our integrated experimental–computational framework could serve as a viable approach for identifying candidate risk variants, which may go undetected by other methodologies. Hereinafter, we shall present results using the combined data. Results using only the Y2H data or predictions are provided in Supplementary Fig. 1.

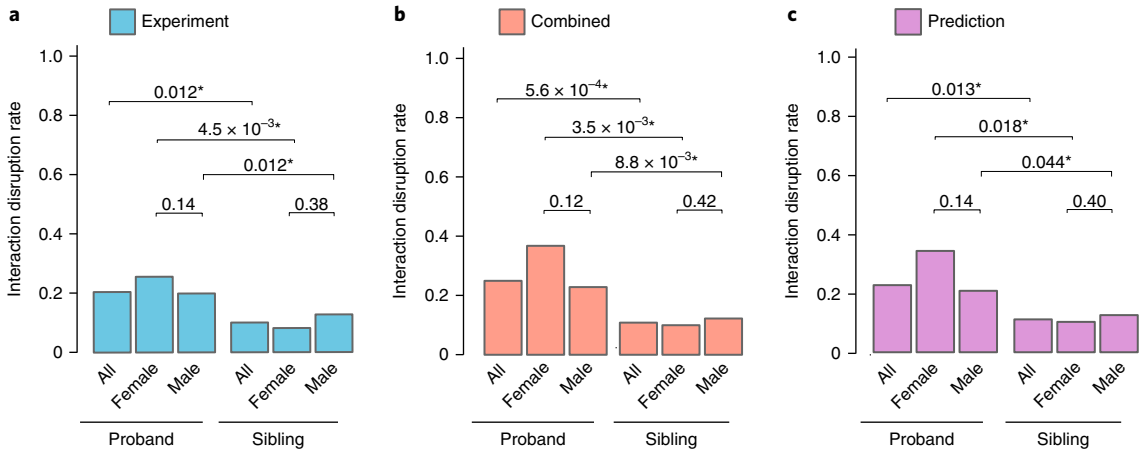

Fig. 2 | dnMis mutations are more disruptive in ASD probands than in unaffected siblings. a–c, Interaction disruption rates of dnMis mutations tested experimentally (**a**), by combining experimental results and predictions (**b**) and predicted computationally (**c**). Probands and unaffected siblings are divided by sex. The count of disruptions per subject was modeled with a negative binomial model (*$P < 0.05$). Combined: 1,080 out of 4,275 measured interactions were disrupted in ASD probands (25.3%) and 322 out of 2,973 were disrupted in unaffected siblings (10.8%). The interaction disruption rate is significantly higher in ASD probands than that in unaffected siblings (FC = 2.34 (1.44–3.79, 95% CI), $P = 5.6\times10^{-4}$ by a two-tailed negative binomial test). The trend persists in male and female subgroups: 23.1% disruption rate in male probands versus 12.3% in male siblings (FC = 2.21 (1.15–4.25, 95% CI), $P = 8.8\times10^{-3}$ by a one-tailed negative binomial test); 37.3% disruption rate in female probands versus 9.9% in female siblings (FC = 3.50 (1.41–8.72, 95% CI), $P = 3.5\times10^{-3}$). Comparing disruption rates between males and females showed a higher rate, although not significant, in females than males in ASD probands (FC = 1.71 (0.71–4.09, 95% CI), $P = 0.12$ by a one-tailed negative binomial test), whereas similar rates were observed in female and male siblings (FC = 1.08 (0.52–2.22, 95% CI), $P = 0.42$).

The female protective effect postulates that females require a larger genetic burden before being diagnosed with ASD[7,47]. Accordingly, we anticipate dnMis mutations in female probands to be more disruptive than those in male probands, although the 6.5:1 male/female ratio of probands could obscure true differences by limiting power. Indeed, we observed a higher disruption rate in females than in males among ASD probands, fold = 1.71, but the difference is not significant ($P = 0.12$). In contrast, the disruption rate in female versus male siblings is 1.08-fold and does not approach significance ($P = 0.42$, Fig. 2b).

Disruptive dnMis mutations in probands impact network hubs. Previous research has shown that genes harboring known disease-associated mutations differ strongly in their network properties in comparison to non-disease-associated genes[48,49]. Early studies reported that disease-associated genes often encode for protein hubs that mediate a greater number of protein interactions than their non-disease-associated counterparts as a whole[50,51]. However, researchers later argued that the observed hub–disease gene correlation might be entirely driven by a handful of hub-encoding essential genes classified within the disease-associated gene class[48]. Here we investigated whether proteins harboring disruptive dnMis mutations in ASD probands exhibit distinguishable network properties in the human interactome.

We first compared the degree of all proteins harboring interaction-disrupting dnMis mutations to those harboring non-disrupting dnMis mutations. We found that in ASD probands, proteins with interaction-disrupting dnMis mutations on average have a significantly higher degree than proteins with non-disrupting dnMis mutations (mean ± s.e.m.: 18.4 ± 2.8 versus 9.3 ± 1.0, fold change (FC) = 1.98, $P = 2.0\times10^{-6}$ by a two-tailed *U*-test, Fig. 3a), whereas no significant difference was observed in unaffected siblings (mean ± s.e.m.: 7.9 ± 1.0 versus 11.4 ± 1.3, FC = 0.69, $P = 0.60$). This suggests that interaction-disrupting dnMis mutations in ASD probands preferentially impact hub proteins, which play a central role in maintaining the integrity of the human interactome[52]. Importantly, when we excluded essential human genes[53] from our analysis, the correlation between interaction-disrupting dnMis mutations and protein hubs persisted in probands (mean ± s.e.m.: 17.6 ± 2.9 versus 9.2 ± 1.0, FC = 1.91, $P = 7.2\times10^{-6}$, Fig. 3a). Similarly, no such correlation was observed in unaffected siblings (mean ± s.e.m.: 8.0 ± 1.0 versus 11.0 ± 1.3, FC = 0.73, $P = 0.50$). Likewise, when we analyzed betweenness, another measure of network centrality based on shortest paths, proteins harboring interaction-disrupting dnMis mutations have a significantly higher betweenness value than proteins harboring non-disrupting dnMis mutations in ASD probands, regardless of whether essential genes were included (Fig. 3b).

To further assess whether disruptive dnMis mutations tend to be on essential genes, we analyzed gene essentiality measured using CRISPR (clustered regularly interspaced short palindromic repeat) gene knockout screens[54]. Using this CRISPR score, we observed no significant difference in essentiality between genes with interaction-disrupting and non-disrupting dnMis mutations for probands (mean ± s.e.m.: −0.43 ± 0.08 versus −0.33 ± 0.04, FC = 1.30, $P = 0.28$ by a two-tailed *U*-test) or for unaffected siblings (mean ± s.e.m.: −0.37 ± 0.09 versus −0.41 ± 0.05, FC = 0.90, $P = 0.39$). This confirms that disruptive dnMis mutations have no tendency to be on essential genes while preferentially affecting topologically central positions in the interactome network in ASD probands.

We then investigated whether proteins with dnMis mutations tend to form inter-connected modules within the interactome network. We found that in ASD probands, proteins with interaction-disrupting dnMis mutations on average have a significantly smaller shortest path length to each other than proteins harboring non-disrupting dnMis mutations (mean ± s.e.m.: 3.48 ± 0.04 versus 3.94 ± 0.03, FC = 0.88, $P = 3.9\times10^{-18}$ by a two-tailed *U*-test, Fig. 3c). This result indicates that proteins with disruptive dnMis mutations in probands tend to be closely connected to each other in the network and may therefore function as modules with specific roles in ASD etiology. In contrast, no such trend was observed for proteins with disruptive dnMis mutations in unaffected siblings (mean ± s.e.m.: 3.77 ± 0.05 versus 3.79 ± 0.03, FC = 0.99, $P = 0.96$), underscoring the functional significance of modules derived from interaction-disrupting dnMis mutations in ASD probands.

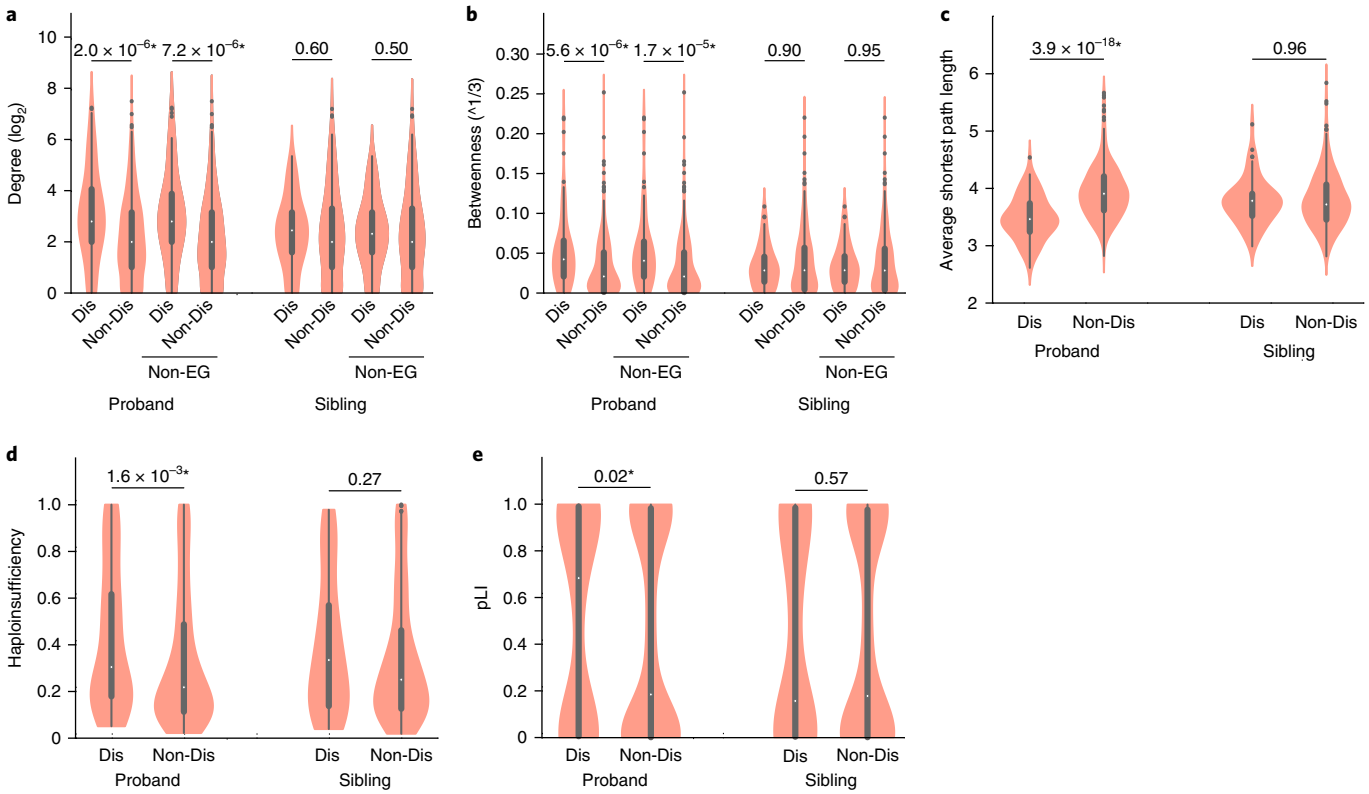

Fig. 3 | Disruptive proband dnMis mutations exhibit characteristic network and haploinsufficiency properties. a,**b**, Degree (**a**) and betweenness (**b**) distributions of proteins with interaction-disrupting (Dis, $n=109$ in probands and $n=68$ in siblings) or non-disrupting (Non-Dis, $n=342$ in probands and $n=241$ in siblings) dnMis mutations across all proteins and across non-essential gene-encoded proteins (Non-EG) in ASD probands (Dis: $n=106$; Non-Dis: $n=338$) and unaffected siblings (Dis: $n=66$; Non-Dis: $n=238$). Degree and betweenness values are transformed by $\log_2$ and cube root (^1/3) for presentation purposes, respectively. **c**, Average shortest path length distributions of proteins with dnMis mutations (in probands, Dis: $n=109$, Non-Dis: $n=342$; in siblings, Dis: $n=68$, Non-Dis: $n=241$). **d**,**e**, Haploinsufficiency (**d**) and pLI (**e**) distributions of genes with dnMis mutations. Genes with available haploinsufficiency or pLI scores were included in corresponding analyses (haploinsufficiency: in probands, Dis: $n=95$, Non-Dis: $n=304$; in siblings, Dis: $n=63$, Non-Dis: $n=217$; pLI: in probands, Dis: $n=106$, Non-Dis: $n=338$; in siblings, Dis: $n=63$, Non-Dis: $n=237$). Genes carrying dnPTVs in SSC data were excluded from all analyses. Violin plots: thick black bar, interquartile range; white dot, median; whiskers, upper and lower limits; points, outliers; the width of each 'violin' is proportional to element abundance. *P* values were calculated using a two-tailed *U*-test (*$P<0.05$).

Overall, our results indicate that network topology should be considered when interpreting the impact of dnMis mutations. By analyzing network properties, we can investigate how individual disruptive missense mutations locally alter protein complexes and functional modules and how multiple mutations work together to rewire the whole cellular network which can, as a result, lead to autism or other disease-associated phenotypes.

Disruptive dnMis mutations and haploinsufficient genes. Disruptive dnMis mutations typically occur only on one copy of the gene. To affect risk, they should occur more frequently on haploinsufficient genes, where a single copy of the wild-type gene is insufficient to carry out its normal function. In probands, genes harboring interaction-disrupting dnMis mutations have a significantly higher probability of being haploinsufficient[55] than genes harboring non-disrupting dnMis mutations (mean ± s.e.m.: 0.42 ± 0.03 versus 0.33 ± 0.02, FC = 1.27, $P=1.6\times10^{-3}$ by a two-tailed *U*-test, Fig. 3d). In contrast, no significant difference was observed in unaffected siblings (mean ± s.e.m.: 0.39 ± 0.04 versus 0.34 ± 0.02, FC = 1.15, $P=0.27$). Reinforcing these findings, we also found that genes with interaction-disrupting dnMis mutations in probands are less tolerant to genetic variation, as indicated by their higher average pLI[56] scores in comparison to genes with non-disrupting dnMis mutations (mean ± s.e.m.: 0.52 ± 0.04 versus 0.43 ± 0.02, FC = 1.21, $P=0.02$, Fig. 3e). No such contrast was found in unaffected siblings (mean ± s.e.m.: 0.44 ± 0.06 versus 0.44 ± 0.03, FC = 1.00, $P=0.57$). Collectively, these results demonstrate that interaction-disrupting dnMis mutations in ASD probands tend to affect haploinsufficient genes, for which heterozygous variations are not tolerated, and they may therefore contribute to ASD outcomes through dosage effect[57].

Disruptive dnMis mutations in probands cluster to ASD genes. To evaluate whether interaction-disrupting dnMis mutations are associated with ASD risk, we first investigated whether such mutations are enriched in previously reported ASD-associated genes. Using a curated list of 881 genes implicated in ASD in the SFARI database[58], we observed a significant enrichment in probands for genes with interaction-disrupting dnMis mutations compared to genes with non-disrupting dnMis mutations (21/109 versus 32/342, odds ratio (OR) = 2.3, $P=5.7\times10^{-3}$ by a one-tailed Fisher's exact test, Supplementary Table 3). In contrast, no significant enrichment was observed in unaffected siblings (6/68 versus 17/241, OR = 1.3, $P=0.39$). Thus, characterizing interaction perturbation captures new evidence to establish associations of genes with ASD.

Previous studies have reported functional clustering in genes with de novo protein-truncating variants (dnPTVs) in ASD individuals[13,14,16,59]. Here we assessed the network distance within the human interactome between genes harboring interaction-disrupting dnMis mutations (excluding genes with dnPTVs)

and seven classes of known ASD-associated genes. These genes (Supplementary Table 4) include: fragile X mental retardation protein (FMRP) target genes, with transcripts bound by FMRP; genes encoding chromatin modifiers; genes expressed preferentially in embryos; genes encoding postsynaptic density proteins; 881 genes in the SFARI database; a high-quality SFARI subset (141 genes scored as syndromic, high confidence or strong candidate[58]); and the latest set of 65 ASD genes discovered by de novo mutations[29]. We found that in probands, proteins harboring interaction-disrupting dnMis mutations are significantly closer to proteins from all seven classes in comparison to proteins with non-disrupting dnMis mutations (Table 1 and Supplementary Note; Methods). In contrast, no significant differences were observed among unaffected siblings in any category. These findings demonstrate that disruptive dnMis mutations identified by our study in ASD probands are indeed closely related to known ASD genes and functional classes and that they may contribute to ASD etiology by disrupting common pathways shared with dnPTVs.

Identification of candidate ASD genes and mutations. Towards the identification of new candidate ASD-associated genes, we examined mutations on the protein RARA (Supplementary Note). RARA binds with RXRB to form the retinoic acid receptor complex. When bound to retinoic acid, the retinoic acid receptor can then bind retinoic acid receptor elements to co-activate transcription of downstream genes. In agreement with our Y2H experiments, our computational approach predicted that a proband mutation p.Pro375Leu on RARA (NP_000955.1) is disruptive, while an unaffected sibling mutation, p.Arg83His, is not (Fig. 4a,b). We note that PPH2 predicts both mutations to be probably damaging and cannot distinguish the two. We further confirmed by co-immunoprecipitation in human cells that the proband mutation p.Pro375Leu disrupts the RARA–RXRB interaction while the sibling mutation p.Arg83His does not (Fig. 4c; Methods).

While there is insufficient evidence to directly link mutations on RARA to ASD, there is compelling evidence that mutated RARA does induce ASD risk by affecting retinoic acid signaling. Specifically, we would expect the p.Pro375Leu mutation to diminish retinoic acid signaling by disrupting its binding to RXRB. Notably, one of the most common genetic risk factors for ASD is maternal duplication of 15q11-q13 and isodicentric chromosome 15[60], both of which increase transcription of *UBE3A*, among other genes. It has recently been shown that UBE3A negatively regulates ALDH1A proteins[61], which act as rate-limiting enzymes in retinoic acid synthesis. Increased dosage of UBE3A diminishes retinoic acid synthesis and retinoic acid signaling, altering neuronal development and features such as homeostatic synaptic plasticity[61]. Moreover, in mice, ASD-like phenotypes are induced by overexpression of UBE3A or by an ALDH1A antagonist, while the wild-type phenotype can be rescued by retinoic acid supplementation[61]. Thus, together with published results regarding the role of UBE3A in retinoic acid signaling and autism risk[61], our results implicate *RARA* as an ASD-associated gene, and our experimentally validated interaction-disrupting prediction for RARA p.Pro375Leu demonstrates how our methodology can be used to identify functional dnMis mutations.

The occurrence of a predicted disruptive mutation near other closely related disease-associated dnMis mutations across interacting proteins can lend strong evidence towards the postulated functionality and shared phenotypic impact of the mutation in question. In this regard, our computational approach predicted an ASD proband mutation, p.Lys1431Met, on the guanine nucleotide exchange factor TRIO (NP_651960.2) that disrupts its interaction with the GTPase RAC1 (Fig. 4d). Of note, two neurodevelopmental disorder dnMis mutations[18,62] on TRIO, p.Arg1428Gln and p.Pro1461Leu, occur in structural proximity to the ASD proband interface mutation, p.Lys1431Met, as do three dnMis mutations on RAC1 (NP_008839.2) interface residues, p.Asn39Ser, p.Tyr64Asp and p.Pro73Leu, which all result in mild to severe intellectual disability[63] (Fig. 4d). Moreover, p.Lys1431Met has recently been reported to functionally inhibit synaptic function in human cell lines and statistically postulated to reside within a hotspot for ASD-related de novo mutations in the GEF1 domain of TRIO[64]. As sequencing data from developmental disorder studies become more readily available, we anticipate the use of predicted interaction-disrupting mutations to uncover shared molecular pathways between related developmental disorders.

dnMis mutations in other developmental disorders. To demonstrate the generalizability of our interactome perturbation approach towards studying the impact of missense mutations in human disease, we investigated how ~10,000 dnMis mutations previously detected in developmental disorders correspond with protein interaction interfaces. The mutation data comprise a collection of 4,565 dnMis mutations from the Deciphering Developmental Disorders project and five lists of dnMis mutations curated from studies of autism, congenital heart disease, intellectual

Table 1 | Distance of genes with interaction-disrupting and non-disrupting dnMis mutations to seven classes of known ASD-associated genes in a protein interactome network background

	Proband					Sibling				
	Dis (109)		Non-Dis (342)		*P* value	Dis (68)		Non-Dis (241)		*P* value
	Mean	s.d.	Mean	s.d.		Mean	s.d.	Mean	s.d.	
FMRP (794)	2.61	0.38	2.85	0.46	$\mathbf{1.5\times10^{-6}}$	2.77	0.37	2.77	0.48	0.57
CHM (408)	2.55	0.38	2.79	0.47	$\mathbf{1.3\times10^{-6}}$	2.70	0.40	2.72	0.50	0.44
EMB (1,865)	2.65	0.38	2.88	0.46	$\mathbf{2.9\times10^{-6}}$	2.79	0.38	2.81	0.48	0.45
PSD (1,395)	2.61	0.37	2.84	0.46	$\mathbf{2.4\times10^{-6}}$	2.76	0.37	2.77	0.47	0.47
SFARI (881)	2.69	0.38	2.92	0.46	$\mathbf{1.8\times10^{-6}}$	2.83	0.37	2.85	0.48	0.52
SFARI hq (141)	2.62	0.38	2.86	0.46	$\mathbf{1.1\times10^{-6}}$	2.77	0.37	2.77	0.49	0.58
DN65 (65)	2.70	0.39	2.94	0.46	$\mathbf{1.0\times10^{-6}}$	2.85	0.37	2.86	0.48	0.52

The number of genes in each class is indicated in parentheses. Genes carrying dnPTVs in SSC data were excluded from the analyses. *P* values were calculated using a one-tailed *U*-test (*P* < 0.05 in bold). Dis, interaction-disrupting; Non-Dis, non-disrupting. FMRP, fragile X mental retardation protein target genes; CHM, genes encoding chromatin modifiers; EMB, genes expressed preferentially in embryos; PSD, genes encoding postsynaptic density proteins; SFARI, 881 genes in the SFARI database; SFARI hq, a high-quality SFARI subset (141 genes scored as syndromic, high confidence or strong candidate); DN65, the latest set of 65 ASD genes discovered by de novo mutations.

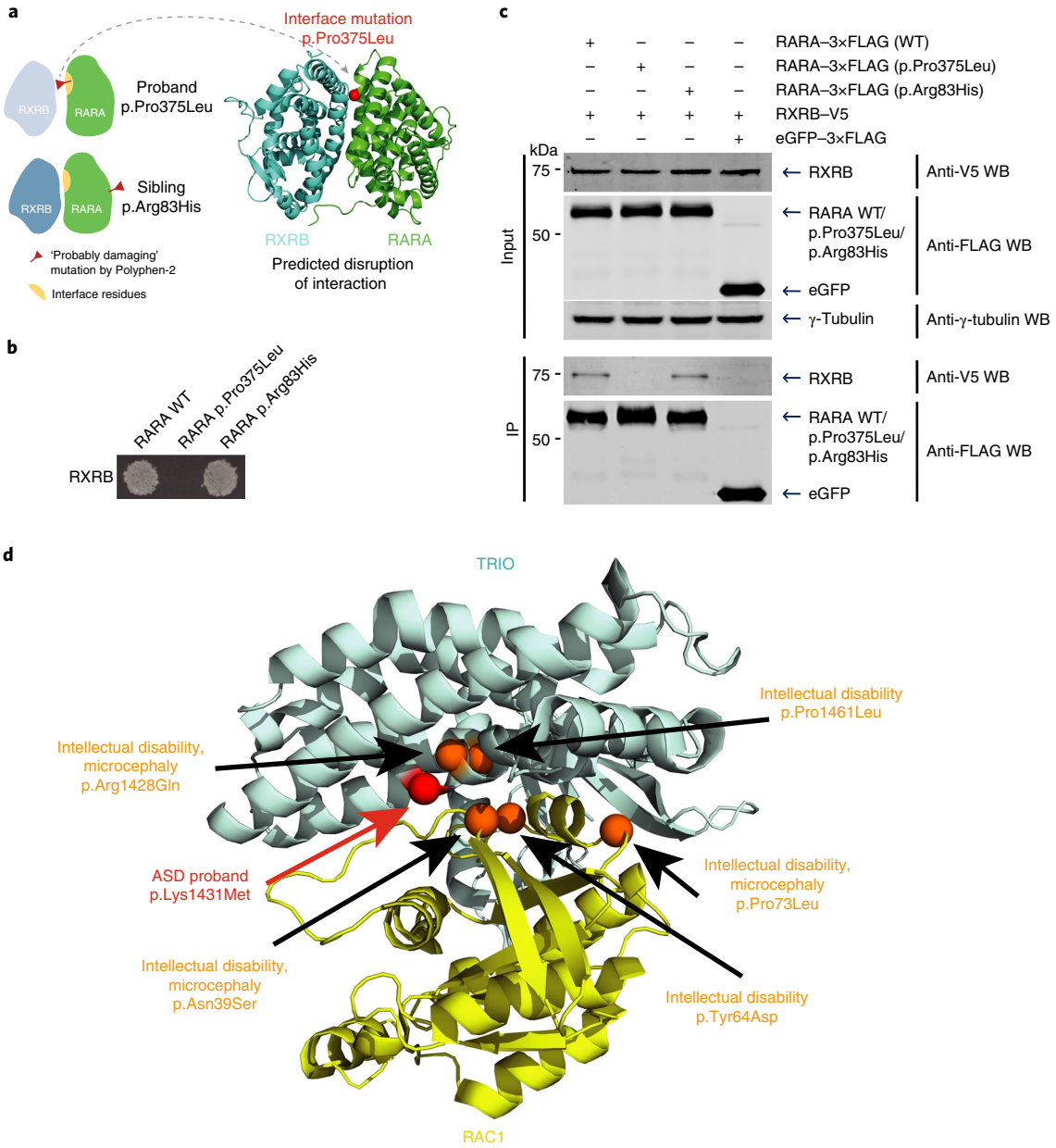

Fig. 4 | Identification of candidate ASD-associated genes and mutations through our interactome perturbation framework. a, Computational prediction of the effects of RARA p.Pro375Leu and RARA p.Arg83His on the RARA–RXRB interaction. A homology model highlighting the RARA p.Pro375Leu interface mutation is shown. **b**, RARA p.Pro375Leu disruption and RARA p.Arg83His non-disruption of the RARA–RXRB interaction by Y2H. **c**, Co-immunoprecipitation confirming RARA p.Pro375Leu disruption and RARA p.Arg83His non-disruption of the RARA–RXRB interaction in HEK 293T cells. See Supplementary Fig. 10 for uncropped gel images. **d**, Co-crystal structure of TRIO-RAC1 (PDB ID: 2NZ8) displaying the structural locations of proband ASD (red) and intellectual disability and/or microcephaly (orange) dnMis mutations across the interaction interfaces.

disability, schizophrenia and epilepsy (denovo-db v.1.5)[65]. We found that in all six data sets, dnMis mutations occur significantly more frequently on protein interaction interfaces than expected (Fig. 5), indicating that dnMis mutations in developmental disorders can contribute to disease risk by impacting protein interactions. In particular, the strongest signal was observed in intellectual disability: 23.5% of the dnMis mutations occurred on interaction interfaces, resulting in an enrichment of 2.09 (1.77–2.44, 95% confidence interval (CI)) in comparison to the fraction of interface residues on corresponding proteins (11.2%, $P = 1.4\times10^{-14}$ by a two-tailed exact binomial test). In contrast, dnMis mutations in schizophrenia had the weakest significance (enrichment = 1.61 (1.22–2.06, 95% CI), $P = 8.7\times10^{-4}$), which agrees with previous findings that schizophrenia has a much weaker de novo signal than other developmental disorders[66].

Forty dnMis hotspots implicated in neurodevelopmental disorder pathogenesis have recently been reported[67]. When we examined the 31 corresponding hotspots within the interactome network, we found that they occur on protein interaction interfaces at a very high rate of 48.4% (enrichment = 4.03 (2.51–5.58, 95% CI), $P = 6.9\times10^{-7}$, Fig. 5). This suggests that interactome perturbations play an important role in the pathogenesis linked with these recurrent events. Taken together, these findings reinforce that our integrated experimental–computational interactome perturbation approach offers a scalable and generalizable framework to identify risk dnMis mutations in human disease.

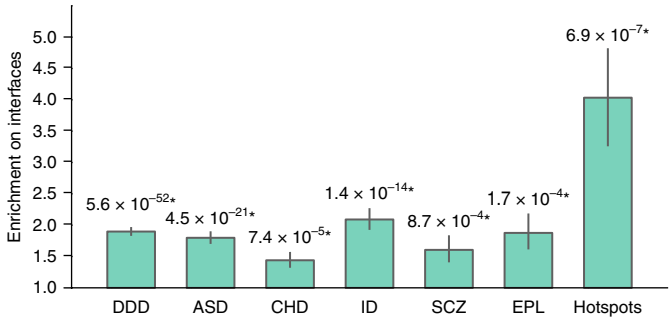

Fig. 5 | dnMis mutations are enriched on protein interaction interfaces in developmental disorders. Enrichment was calculated as the ratio of the observed fraction of dnMis mutations that occur on interaction interfaces over the fraction of interface residues on corresponding proteins (expected fraction). The error bars indicate standard error. *P* values were calculated using a two-tailed exact binomial test (*$P < 0.05$). DDD (Deciphering Developmental Disorders project, $n = 2{,}914$ dnMis mutations): enrichment = 1.90 (1.76–2.04, 95% CI); ASD (autism spectrum disorder, $n = 1{,}512$): enrichment = 1.80 (1.61–2.00, 95% CI); CHD (congenital heart disease, $n = 759$): enrichment = 1.44 (1.21–1.70, 95% CI); ID (intellectual disability, $n = 498$): enrichment = 2.09 (1.77–2.44, 95% CI); SCZ (schizophrenia, $n = 312$): enrichment = 1.61 (1.22–2.06, 95% CI); EPL (epilepsy, $n = 181$): enrichment = 1.88 (1.36–2.48, 95% CI); hotspots ($n = 31$): enrichment = 4.03 (2.51–5.58, 95% CI).

Discussion

Here we demonstrated that dnMis mutations can contribute to ASD risk by disrupting protein–protein interactions and that our interactome perturbation framework offers a novel and effective way to identify ASD risk dnMis mutations. As only a small fraction of dnMis mutations found in ASD subjects are believed to be functional[12], this framework helps overcome a significant challenge in identifying risk dnMis mutations. Our analyses focused on dnMis mutations from the SSC families because the information on unaffected siblings in the data set provides robust negative controls. Our results demonstrated that interaction-disrupting dnMis mutations in ASD probands preferentially impact proteins that have many interaction partners in the interactome network (that is, hubs) and disrupt these interactions at a significantly higher rate than those in unaffected siblings. Our results also lend evidence to previously reported ASD-associated genes and pathways by showing that interaction-disrupting dnMis mutations are closely clustered to proteins in ASD-associated functional classes in the interactome network. Thus, characterizing interactome perturbation provides additional and potentially orthogonal information to strengthen previously identified genetic associations and helps in the discovery of new genes that contribute to ASD risk.

Integration of computational predictions with experimental data imbued more meaning onto missense mutations found in ASD probands and their siblings. Thus, the prediction model alone can enhance researchers' ability to prioritize damaging missense mutations and can be applied across a wide range of human disease studies. We emphasize that the strength of this prediction model is rooted in its integration of PPH2 scores and Interactome INSIDER interface predictions. To demonstrate this, we repeated all analyses using PPH2 and Interactome INSIDER separately. The results show that neither method individually is sufficient to reproduce most signals towards identifying disease-contributing dnMis mutations in ASD (Supplementary Fig. 2 and Supplementary Fig. 3). This confirms that our two-tiered predictor, which evaluates the disruptiveness of a variant on protein interactions, greatly improves the effectiveness of predicting functional missense mutations. We also note that our predictor is robust at different PPH2 score cutoffs (Supplementary Fig. 4). Taken together, we demonstrate that our computational prediction approach can serve as an effective and robust method to identify disease-contributing missense mutations.

Stronger associations between ASD proband mutations and clinical data can be established by filtering out ASD dnMis mutations that are also identified as population variants in the Exome Aggregation Consortium (ExAC)[56] since these variants represent standing variations in the population and are less likely to be deleterious as a result[68]. Adopting this same principle, we found that dnMis mutations in developmental disorders are significantly more enriched on interaction interfaces when mutations coinciding with ExAC are filtered out in comparison to the non-filtered set (Supplementary Fig. 5 and Supplementary Note). Importantly, our results show that the characteristic network and haploinsufficiency properties of disruptive proband dnMis mutations are not unique to ASD but are shared features across different developmental disorders (Supplementary Fig. 6), indicating that our interactome perturbation framework is generalizable to prioritize dnMis mutations across a wide range of developmental disorders (Supplementary Note).

Our analyses indicate that network properties are important in interpreting the functional impact of dnMis mutations and their relevance towards disease etiology. However, we recognize that the human interactome with which these analyses are performed is currently incomplete. As a result, certain classes of protein interactions, for example interactions mediated by membrane-bound proteins, may be under-represented in the current interactome, limiting potential insights from such proteins. Moreover, literature-derived segments of the human interactome are subject to sampling bias present in small-scale studies[38,69]. Therefore, we re-examined the network topology analyses across a chronologically ordered series of unbiased high-throughput-derived human interactomes (Supplementary Note). Not only do we show that our results are robust across all high-throughput-derived interactomes, more importantly, we also demonstrate that the topological differences between interaction-disrupting and non-disrupting dnMis mutations in probands become more significant as the interactome coverage increases (Supplementary Fig. 7). Moreover, we show that all of our results remain the same when we expanded our disruptiveness predictions to include all dnMis mutations in the current human interactome (Supplementary Fig. 8 and Supplementary Note). Taking these results together, we fully expect that as increasingly more human protein–protein interactions and mutations are uncovered, our interactome perturbation framework can be applied to these new interactions and mutations to identify new or currently under-characterized disease-associated mutations and genes.

As large-scale WES studies continue to produce mutation data at ever-increasing scales, our interaction-disruption prediction approach can greatly extend the reach of interactome perturbation studies to investigate complex genotype–phenotype relationships and improve our understanding of how genetic variation affects disease risk through the alteration of topological and community structures of networks.

URLs. Interactome INSIDER predictions, http://interactomeinsider.yulab.org; SFARI database (downloaded on June 6, 2017), https://gene.sfari.org/database/human-gene/; denovo-db (v.1.5), http://denovo-db.gs.washington.edu/denovo-db/Download.jsp.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0130-z.

Received: 4 October 2017; Accepted: 6 April 2018;
Published online: 11 June 2018

References

1. Ropers, H. H. Genetics of early onset cognitive impairment. *Annu. Rev. Genomics Hum. Genet.* **11**, 161–187 (2010).
2. Mefford, H. C., Batshaw, M. L. & Hoffman, E. P. Genomics, intellectual disability, and autism. *N. Engl. J. Med.* **366**, 733–743 (2012).
3. Devlin, B. & Scherer, S. W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).
4. Bruneau, B. G. The developmental genetics of congenital heart disease. *Nature* **451**, 943–948 (2008).
5. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
6. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
7. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
8. Epi, K. C. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
9. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project & Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
10. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
11. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
12. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
13. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
14. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
15. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
16. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
17. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
18. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
19. de Ligt, J., Veltman, J. A. & Vissers, L. E. Point mutations as a source of de novo genetic disease. *Curr. Opin. Genet. Dev.* **23**, 257–263 (2013).
20. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
21. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
22. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
23. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
24. Wei, X. et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
25. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
26. Braun, P. et al. An experimentally derived confidence score for binary protein–protein interactions. *Nat. Methods* **6**, 91–97 (2009).
27. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
28. Meyer, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 1–8 (2018).
29. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
30. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
31. Levy, D. et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897 (2011).
32. Sanders, S. J. 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).
33. Dong, S. et al. De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9**, 16–23 (2014).
34. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
35. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
36. Wang, X. et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
37. Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29**, 1577–1579 (2013).
38. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
39. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–D478 (2015).
40. Stelzl, U. et al. A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
41. Turner, B. et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023 (2010).
42. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
43. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
44. Keshava Prasad, T. S. et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
45. Mewes, H. W. et al. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220–D224 (2011).
46. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
47. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nat. Neurosci.* **18**, 191–198 (2015).
48. Goh, K. I. et al. The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
49. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA* **105**, 4323–4328 (2008).
50. Xu, J. & Li, Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* **22**, 2800–2805 (2006).
51. Jonsson, P. F. & Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297 (2006).
52. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
53. Chen, W. H., Lu, G., Chen, X., Zhao, X. M. & Bork, P. OGEEv2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.* **45**, D940–D944 (2017).
54. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
55. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
56. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
57. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.* **15**, 133–141 (2014).
58. Basu, S. N., Kollu, R. & Banerjee-Basu, S. AutDB: a gene reference resource for autism research. *Nucleic Acids Res.* **37**, D832–D836 (2009).
59. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
60. Schanen, N. C. Epigenetics of autism spectrum disorders. *Hum. Mol. Genet.* **15**, R138–R150 (2006).
61. Xu, X. et al. Excessive UBE3A dosage impairs retinoic acid signaling and synaptic plasticity in autism spectrum disorders. *Cell Res.* **28**, 48–68 (2017).
62. Pengelly, R. J. et al. Mutations specific to the Rac-GEF domain of TRIO cause intellectual disability and microcephaly. *J. Med. Genet.* **53**, 735–742 (2016).
63. Reijnders, M. R. F. et al. RAC1 missense mutations in developmental disorders with diverse phenotypes. *Am. J. Human Genet.* **101**, 466–477 (2017).
64. Sadybekov, A., Tian, C., Arnesano, C., Katritch, V. & Herring, B. E. An autism spectrum disorder-related de novo mutation hotspot discovered in the GEF1 domain of Trio. *Nat. Commun.* **8**, 601 (2017).
65. Turner, T. N. et al. denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* **45**, D804–D811 (2017).
66. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
67. Geisheker, M. R. et al. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* **20**, 1043–1051 (2017).
68. Robinson, E. B. et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552 (2016).

69. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

Acknowledgements
We would like to thank J. F. Beltrán, J. Liang, S. D. Wierbowski and other Yu laboratory members for constructive discussions. This work was supported by National Institute of General Medical Sciences grants (R01 GM104424, R01 GM124559, R01 GM125639); a National Cancer Institute grant (R01 CA167824); a Eunice Kennedy Shriver National Institute of Child Health and Human Development grant (R01 HD082568); a National Human Genome Research Institute grant (UM1 HG009393); a National Science Foundation grant (DBI-1661380) to H.Y.; a National Institute of Mental Health grant (R37MH057881) to B.D. and K.R.; and Simons Foundation Autism Research Initiative grants (SF367561 to H.Y., B.D. and K.R. and SF402281 to B.D. and K.R.). We would like to thank the SSC principal investigators (A. L. Beaudet, R. Bernier, J. Constantino, E. H. Cook, Jr, E. Fombonne, D. Geschwind, D. E. Grice, A. Klin, D. H. Ledbetter, C. Lord, C. L. Martin, D. M. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M. W. State, W. Stone, J. S. Sutcliffe, C. A. Walsh and E. Wijsman) and the coordinators and staff at the SSC clinical sites; the SFARI staff, in particular N. Volfovsky; D. B. Goldstein for contributing to the experimental design; and the Rutgers University Cell and DNA repository for accessing biomaterials.

Author contributions
S.C., R.F., K.R., B.D. and H.Y. conceived the study. H.Y. oversaw all aspects of the study. S.C. and L.K. performed computational analyses with extensive input from K.R., B.D., and H.Y. R.F. and Y.L. performed laboratory experiments. S.C. and R.F. wrote the manuscript with input from J.W., K.R., B.D. and H.Y. All authors edited and approved of the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0130-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.R. or B.D. or H.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Enrichment of dnMis mutations on interaction interfaces. The set of 412 proteins with dnMis mutations and containing at least 1 interaction interface and 1 known domain was included for calculating dnMis mutation distribution. The sequences were divided into three regions: 'in interaction interface', 'in other domain' and 'outside domains'. Interaction interfaces were determined by our previously developed human structural interaction network (hSIN[36], comprising 4,222 structurally resolved interactions between 2,816 proteins). 'Other domains' referred to protein domains (obtained from Pfam[70] database) that exclude interacting interfaces in hSIN. The rest of the residues then were categorized as 'outside domains'. If the locations of mutations were not influenced by the domain architecture of the protein, then their relative lengths should determine the frequency of mutations in these three regions. The fraction of mutations expected by chance in each region was calculated by adding the total sequence length of each region in all proteins, and dividing it by the length of all proteins combined; call the probability of falling in an interaction interface p. The number of observed mutations in each region over all proteins was also computed; call the number falling in the interaction interfaces S, and let N be the total number of dnMis missense mutations. An exact binomial test was then computed from p, S and N. CIs are based on the 95% CI for an exact binomial, and then transformed to the risk ratio (enrichment) using the expectation in the denominator and the lower/upper bound in the numerator.

In ASD probands, the total length of 248 proteins is 377,421, which compromises 113,449 residues on interaction interfaces, 69,870 residues on other domains and 194,102 residues outside domains. The probabilities for a mutation to fall in these regions were computed to be 30.1%, 18.5% and 51.4%, respectively. The observed distribution of the 296 dnMis mutations on these proteins was 113 on interaction interfaces, 59 on other domains and 124 outside domains, showing that dnMis mutations in ASD probands are significantly enriched on protein interaction interfaces (enrichment = 1.27 (1.09–1.46, 95% CI), $P = 2.9 \times 10^{-3}$) while they occur on other domains with an expected rate (enrichment = 1.08 (0.84–1.35, 95% CI), $P = 0.55$) and are depleted from regions outside domains (enrichment = 0.81 (0.70–0.93, 95% CI), $P = 1.1 \times 10^{-3}$). In contrast, the observed 186 dnMis mutations in unaffected siblings occur on all three regions with expected rates: 70/186 fall on interaction interfaces (37.6% versus expected 36.5%, enrichment = 1.03 (0.84–1.23, 95% CI), $P = 0.76$), 32/186 fall on other domains (17.2% versus expected 15.9%, enrichment = 1.08 (0.76–1.47, 95% CI), $P = 0.62$) and 84/186 fall outside domains (45.2% versus expected 47.6%, enrichment = 0.95 (0.79–1.10, 95% CI), $P = 0.51$).

Cloning of 208 dnMis mutations using our massively parallel Clone-seq pipeline. Single-colony-derived mutant clones were constructed using a high-throughput mutagenesis and next-generation sequencing pipeline called Clone-seq[24] (Supplementary Fig. 9). Wild-type clones were picked from hORFeome v8.1[71] to serve as templates for site-directed mutagenesis (Eurofins). Mutagenesis was performed at 96-well scales using site-specific mutagenesis primers and full-length human ORF templates. PCR product was digested overnight using DpnI (NEB) without a ligation step to maximize throughput and then transformed directly into competent cells to isolate single colonies. Then, 4 colonies per mutagenesis reaction were hand-picked and arrayed into 96-well plates. After 21 h incubation at 37 °C, glycerol stocks were generated and then clones were pooled into 4 respective bacterial pools. Maxiprepped DNAs from each of the 4 pools were then combined through multiplexing (NEBNext) and then sequenced in a single 1 × 100 single-end Illumina HiSeq run. Properly mutated clones were then identified by next-generation sequencing analysis and recovered from single-colony glycerol stocks. In total, we generated individual clones for 208 dnMis mutations comprising 109 from ASD probands and 99 from unaffected siblings.

Experimental examination of 667 protein–protein interactions using our high-throughput Y2H assay. To perform Y2H, pDEST-AD and pDEST-DB plasmid vectors corresponding to the GAL4-activating domain (AD) and DNA-binding (DB) domain, respectively, were used. Full-length Clone-seq-identified mutant clones were transferred into Y2H-amenable pDEST-DB and pDEST-AD vectors by Gateway LR reactions and then transformed into *MATα* Y8930 and *MAT**a*** Y8800, respectively. All DB-ORF *MATα* transformants, including wild-type ORFs, were then mated against corresponding wild-type and mutant AD-ORF *MAT**a*** transformants in a pairwise orientation on YEPD agar plates. After mating, yeast was replica-plated onto selective SC–Leu–Trp–His+1 mM of 3-amino-1,2,4-triazole (3AT) as well as SC–Leu–Trp–Adenine plates. Interactions were scored after three days of incubation and five days of incubation for SC–Leu–Trp+3AT and SC–Leu–Trp–Ade plates, respectively. To screen out autoactivating DB-ORFs, all DB-ORF *MATα* transformants were also mated pairwise against empty pDEST-AD *MAT**a*** transformants and scored for growth on SC–Leu–Trp+3AT and SC–Leu–Trp–Ade plates. DB-ORFs that trigger reporter activity under this setup were removed from further experiments. We finally examined 667 interactions, of which the wild-type proteins could be detected with strong Y2H-positive phenotypes in our experiments, for 151 out of the 208 total dnMis mutations that we have successfully generated. The other 57 dnMis mutations corresponded to proteins with no testable interaction partners by Y2H; therefore, they were excluded from Y2H experiments. While on average each of the 151 mutations was tested against 4 or 5 interaction partners, 2 proband mutations (Q8TBB1 p.Glu295Lys and Q8TD31 p.Trp337Arg) had >40 interaction partners tested and disrupted >30 of their corresponding interactions. Thus, we excluded these two outliers when comparing the disruption rates of dnMis mutations in ASD probands and unaffected siblings (Fig. 2a).

Computational prediction for protein–protein interaction disruption. For the remaining 1,582 dnMis mutations, we assessed their probabilities to disrupt an interaction based on whether they are likely to be on protein interaction interfaces and whether they tend to have damaging functional effects on the protein. We first applied an ensemble machine-learning algorithm to predict interface residues (Interactome INSIDER). For each of these dnMis mutations, on each of its interactions with an interaction-specific partner, we considered a mutation to be an interaction interface residue for this specific interaction if it has a probability score of very high, high or medium in Interactome INSIDER prediction. We next evaluated its deleteriousness using PolyPhen-2 (PPH2). If a mutation predicted as an interface residue also has a 'probably damaging' PPH2 score (Interface+ and PPH2+), we considered this mutation to disrupt the interaction. On the other hand, we called a mutation non-disrupting if it was predicted as unlikely to be an interaction interface residue (probability below 'medium' by Interactome INSIDER) and to be 'benign' to the protein by PPH2 (Interface− and PPH2−). Considering that using individual measurements (PPH2 alone or Interactome INSIDER alone) do not provide sufficient signal towards whether a mutation is damaging or not (Supplementary Fig. 2 and Supplementary Fig. 3), mutations that only meet one of these two criteria (Interface+ and PPH2−; Interface− and PPH2+) were excluded from the analyses. Importantly, when we included all of the Interface+PPH2− and Interface−PPH2+ mutations as non-disrupting to our analyses, we found that all our results remain the same (Supplementary Fig. 8 and Supplementary Note).

Modeling the number of disrupted interactions as a function of case-control status. Some missense mutations fail to disrupt any interactions, $D = 0$ disruptions. However, other mutations can disrupt $D = 1, 2, \ldots, I$ interactions. To account for the dispersion in D, and to determine whether D is stochastically greater for missense mutations found in ASD probands versus unaffected siblings, we modeled D as a negative binomial distribution and fit it to case-control status. We also evaluated other models for goodness-of-fit, specifically Poisson and zero-inflated versions of Poisson and negative binomial. After accounting for degrees of freedom, none of these models fit the data as well as the negative binomial by the Akaike information criterion.

Construction of plasmids for western blot and co-immunoprecipitation. Wild-type RARA and RXRB entry clones were obtained from the hORFeome v8.1[71] collection. Gateway LR reactions were used to transfer bait RARA wild-type, p.Pro375Leu and p.Arg83His into a pQXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a carboxy-terminal 3×FLAG. Prey RXRB was transferred into pcDNA-DEST40 that includes a V5 tag (Invitrogen, 12274-015) also using Gateway LR reactions.

Cell culture, co-immunoprecipitation and western blotting. HEK 293T cells were maintained in complete DMEM medium supplemented with 10% FBS. Cells were seeded onto 6-well dishes and incubated until 70–80% confluency. Cells were then transfected with a mixed solution of 1 µg bait construct, 1 µg prey construct, 10 µl of 1 mg ml^{-1} PEI (Polysciences, 23966), and 150 µl OptiMEM (Gibco, 31985-062). After 24 h incubation, cells were gently washed three times in 1 × PBS and then resuspended in 200 µl cell lysis buffer (10 mM Tris-Cl pH 8.0, 137 mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA and 1 × EDTA-free Complete Protease Inhibitor tablet (Roche)) and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 min at 16,000*g* at 4 °C. For co-immunoprecipitation, 100 µl cell lysate per sample was incubated with 5 µl EZ view Red Anti-FLAG M2 Affinity Gel (Sigma, F2426) for 2 h at 4 °C under gentle rotation. After incubation, bound proteins were washed three times in cell lysis buffer and then eluted in 50 µl elution buffer (10 mM Tris-Cl pH 8.0, 1% SDS) at 65 °C for 10 min. Cell lysates and co-immunoprecipitated samples were then treated in 6 × SDS protein loading buffer (10% SDS, 1 M Tris-Cl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% bromophenol blue) and subjected to SDS–PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804), anti-V5 (Invitrogen, R960-25) and anti-γ-tubulin (Sigma, T5192) at 1:5,000, 1:3,000 and 1:3,000 dilutions, respectively, were used for immunoblotting analysis. Full scans of all blots are supplied in Supplementary Fig. 10.

Evaluation of the distance between gene sets in the interactome network. We evaluated the distance between two gene sets using a previously published method[59]: in an interactome background, the distance between two gene sets (L_1 and L_2) is the average distance of each gene i in L_1 to L_2, where the distance

of a specific gene i in L_1 to L_2 is the average distance of gene i to each gene j in L_2. Let n_1 and n_2 be the number of genes in L_1 and L_2,

$$\text{Distance}(L_1, L_2) = \frac{1}{n_1}\sum_i \text{Distance}(i, L_2) \text{ where}$$

$$\text{Distance}(i, L_2) = \frac{1}{n_2}\sum_j \text{Distance}(i, j)$$

Then consider i and j as two nodes in the interactome network; the distance between these two nodes Distance(i, j) here is defined as the minimum number of intermediate nodes that connect i and j in the shortest path.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. dnMis mutations in ASD subjects and their unaffected siblings came from published data in ref. [12] and are available in Supplementary Table 1. Interaction disruption results from Y2H experiments are available in Supplementary Table 2.

References

70. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
71. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).

Corresponding author(s): Kathryn Roeder
Bernie Devlin
Haiyuan Yu

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Statistical tests were performed using models built in R (3.3.2) and Python (2.7).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

dnMis mutations in ASD subjects and their unaffected siblings came from published data in ref. 12 and are available in Supplementary Table 1. Interaction disruption results from Y2H experiments are available in Supplementary Table 2.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by data availability. No statistical method was used to predetermine sample size.
Data exclusions	No data was excluded from analyses.
Replication	NA -- only bulk experiments performed, with trends derived from many individuals.
Randomization	Protein interactions were selected in an unbiased manner for experimental test using yeast two-hybrid assay.
Blinding	Data blinding present through all experimental and computational measurements in terms of whether mutations were from affected or unaffected samples.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☐	☒ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Unique materials

Obtaining unique materials	There are no restrictions on experimental materials. ORF clones used in yeast two-hybrid assays and vectors used for co-IP are available upon request.

Antibodies

Antibodies used	Anti-FLAG (Sigma, F1804), anti-V5 (Invitrogen, R960-25), and anti-gamma-Tubulin (Sigma, T5192) at 1:5000, 1:3000, and 1:3000 dilutions, respectively, were used for immunoblotting analysis.
Validation	Species validation of all primary antibodies used in this study can be found in the corresponding manufacturer´s websites.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells were obtained from ATCC.
Authentication	Cell lines have been thoroughly tested and authenticated by ATCC.
Mycoplasma contamination	HEK293T cells were tested negative for Mycoplasma contaminations.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging