LETTER

doi:10.1038/nature10945

De novo mutations revealed by whole-exome sequencing are strongly associated with autism

Stephan J. Sanders[1], Michael T. Murtha[1], Abha R. Gupta[2]*, John D. Murdoch[1]*, Melanie J. Raubeson[1]*, A. Jeremy Willsey[1]*, A. Gulhan Ercan-Sencicek[1]*, Nicholas M. DiLullo[1]*, Neelroop N. Parikshak[3], Jason L. Stein[3], Michael F. Walker[1], Gordon T. Ober[1], Nicole A. Teran[1], Youeun Song[1], Paul El-Fishawy[1], Ryan C. Murtha[1], Murim Choi[4], John D. Overton[4], Robert D. Bjornson[5], Nicholas J. Carriero[5], Kyle A. Meyer[6], Kaya Bilguvar[7], Shrikant M. Mane[8], Nenad Šestan[6], Richard P. Lifton[4], Murat Günel[7], Kathryn Roeder[9], Daniel H. Geschwind[3], Bernie Devlin[10] & Matthew W. State[1]

Multiple studies have confirmed the contribution of rare *de novo* copy number variations to the risk for autism spectrum disorders[1–3]. But whereas *de novo* single nucleotide variants have been identified in affected individuals[4], their contribution to risk has yet to be clarified. Specifically, the frequency and distribution of these mutations have not been well characterized in matched unaffected controls, and such data are vital to the interpretation of *de novo* coding mutations observed in probands. Here we show, using whole-exome sequencing of 928 individuals, including 200 phenotypically discordant sibling pairs, that highly disruptive (nonsense and splice-site) *de novo* mutations in brain-expressed genes are associated with autism spectrum disorders and carry large effects. On the basis of mutation rates in unaffected individuals, we demonstrate that multiple independent *de novo* single nucleotide variants in the same gene among unrelated probands reliably identifies risk alleles, providing a clear path forward for gene discovery. Among a total of 279 identified *de novo* coding mutations, there is a single instance in probands, and none in siblings, in which two independent nonsense variants disrupt the same gene, *SCN2A* (sodium channel, voltage-gated, type II, α subunit), a result that is highly unlikely by chance.

We completed whole-exome sequencing in 238 families from the Simons Simplex Collection (SSC), a comprehensively phenotyped autism spectrum disorders (ASD) cohort consisting of pedigrees with two unaffected parents, an affected proband, and, in 200 families, an unaffected sibling[5]. Exome sequences were captured with NimbleGen oligonucleotide libraries, subjected to DNA sequencing on the Illumina platform, and genotype calls were made at targeted bases (Supplementary Information)[6,7]. On average, 95% of the targeted bases in each individual were assessed by ≥8 independent sequence reads; only those bases showing ≥20 independent reads in all family members were considered for *de novo* mutation detection. This allowed for analysis of *de novo* events in 83% of all targeted bases and 73% of all exons and splice sites in the RefSeq hg18 database (http://www.ncbi.nlm.nih.gov/RefSeq/; Supplementary Table 1; Supplementary Data 1). Given uncertainties regarding the sensitivity of detection of insertion-deletions, case-control comparisons reported here consider only single base substitutions (Supplementary Information). Validation was attempted for all predicted *de novo* single nucleotide variants (SNVs) via Sanger sequencing of all family members, with sequence readers blinded to affected status; 96% were successfully validated. We determined there was no evidence of systematic bias in variant detection between affected and unaffected siblings through comparisons of silent *de novo*, non-coding *de novo*, and novel transmitted variants (Fig. 1a; Supplementary Figs 1–5; Supplementary Information).

Among 200 quartets (Table 1), 125 non-synonymous *de novo* SNVs were present in probands and 87 in siblings: 15 of these were nonsense (10 in probands; 5 in siblings) and 5 altered a canonical splice site (5 in probands; 0 in siblings). There were 2 instances in which *de novo* SNVs were present in the same gene in two unrelated probands; one of these involved two independent nonsense variants (Table 2). Overall, the total number of non-synonymous *de novo* SNVs was significantly greater in probands compared to their unaffected siblings ($P = 0.01$, two-tailed binomial exact test; Fig. 1a; Table 1) as was the odds ratio (OR) of non-synonymous to silent mutations in probands versus siblings (OR = 1.93; 95% confidence interval (CI), 1.11–3.36; $P = 0.02$, asymptotic test; Table 1). Restricting the analysis to nonsense and splice site mutations in brain-expressed genes resulted in substantially increased estimates of effect size and demonstrated a significant difference in cases versus controls based either on an analysis of mutation burden ($N = 13$ versus 3; $P = 0.02$, two-tailed binomial exact test; Fig. 1a; Table 1) or an evaluation of the odds ratio of nonsense and splice site to silent SNVs (OR = 5.65; 95% CI, 1.44–22.2; $P = 0.01$, asymptotic test; Fig. 1b; Table 1).

To determine whether factors other than diagnosis of ASD could explain our findings, we examined a variety of potential covariates, including parental age, IQ and sex. We found that the rate of *de novo* SNVs indeed increases with paternal age ($P = 0.008$, two-tailed Poisson regression) and that paternal and maternal ages are highly correlated ($P < 0.0001$, two-tailed linear regression). However, although the mean paternal age of probands in our sample was 1.1 years higher than their unaffected siblings, re-analysis accounting for age did not substantively alter any of the significant results reported here (Supplementary Information). Similarly, no significant relationship was observed between the rate of *de novo* SNVs and proband IQ ($P \geq 0.19$, two-tailed linear regression, Supplementary Information) or proband sex ($P \geq 0.12$, two-tailed Poisson regression; Supplementary Fig. 6; Supplementary Information).

Overall, these data demonstrate that non-synonymous *de novo* SNVs, and particularly highly disruptive nonsense and splice-site *de novo* mutations, are associated with ASD. On the basis of the conservative assumption that *de novo* single-base coding mutations observed in siblings confer no autism liability, we estimate that at least 14% of

[1]Program on Neurogenetics, Child Study Center, Department of Psychiatry, Department of Genetics, Yale University School of Medicine, 230 South Frontage Road, New Haven, Connecticut 06520, USA. [2]Child Study Center, Department of Pediatrics, Yale University School of Medicine, 230 South Frontage Road, New Haven, Connecticut 06520, USA. [3]Neurogenetics Program, UCLA, 695 Charles E. Young Dr. South, Los Angeles, California 90095, USA. [4]Department of Genetics, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06510, USA. [5]Department of Computer Science, Yale Center for Genome Analysis, Yale University, 51 Prospect Street, New Haven, Connecticut 06511, USA. [6]Department of Neurobiology, Kavli Institute for Neuroscience, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA. [7]Department of Neurosurgery, Center for Human Genetics and Genomics, Program on Neurogenetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA. [8]Yale Center for Genome Analysis, 300 Heffernan Drive, West Haven, Connecticut 06516, USA. [9]Department of Statistics, Carnegie Mellon University, 130 DeSoto Street, Pittsburgh, Pennsylvania 15213, USA. [10]Department of Psychiatry and Human Genetics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. *These authors contributed equally to this work.

©2012 Macmillan Publishers Limited. All rights reserved

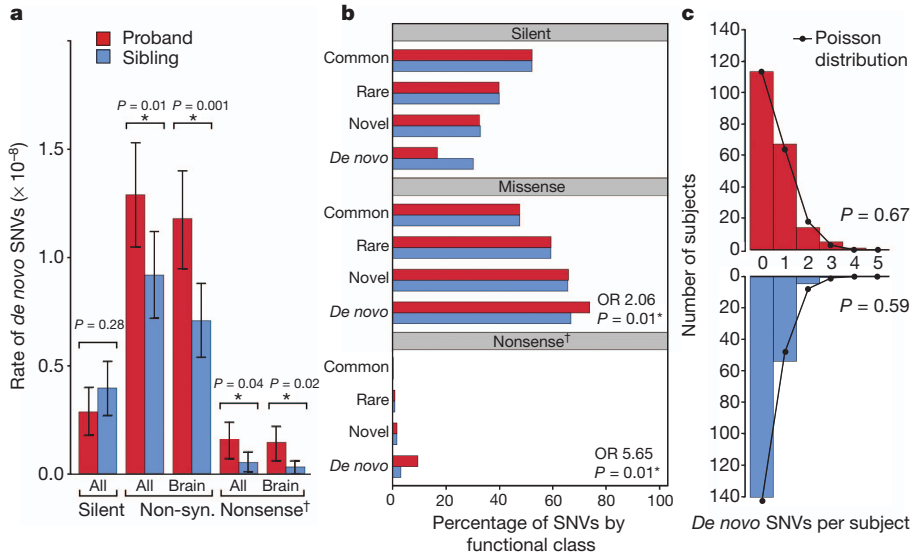

Figure 1 | Enrichment of non-synonymous *de novo* variants in probands relative to sibling controls. **a**, The rate of *de novo* variants is shown for 200 probands (red) and matched unaffected siblings (blue). 'All' refers to all RefSeq genes in hg18, 'Brain' refers to the subset of genes that are brain-expressed[24] and 'Non-syn' to non-synonymous SNVs (including missense, nonsense and splice site SNVs). Error bars represent the 95% confidence intervals and *P* values are calculated with a two-tailed binomial exact test. **b**, The proportion of transmitted variants in brain-expressed genes is equal between 200 probands (red) and matched unaffected siblings (blue) for all mutation types and allele frequencies, including common (≥1%), rare (<1%) and novel (single allele in one of the 400 parents); in contrast, both non-synonymous and nonsense *de novo* variants show significant enrichment in probands compared to unaffected siblings (73.7% versus 66.7%, $P = 0.01$, asymptotic test and 9.5% versus 3.1%, $P = 0.01$ respectively). **c**, The frequency distribution of brain-expressed non-synonymous *de novo* SNVs is shown per sample for probands (red) and siblings (blue). Neither distribution differs from the Poisson distribution (black line), suggesting that multiple *de novo* SNVs within a single individual do not confirm ASD risk. Nonsense† represents the combination of nonsense and splice site SNVs.

affected individuals in the SSC carry *de novo* SNV risk events (Supplementary Information). Moreover, among probands and considering brain-expressed genes, an estimated 41% of non-synonymous *de novo* SNVs (95% CI, 21–58%) and 77% of nonsense and splice site *de novo* SNVs (95% CI, 33–100%) point to *bona fide* ASD-risk loci (Supplementary Information).

We next set out to evaluate which of the particular *de novo* SNVs identified in our study confer this risk. On the basis of our prior work[3],

Table 1 | Distribution of SNVs between probands and siblings

Category	Total number of SNVs*		SNVs per subject		Per base SNV rate (x10^{-8})		*P*†	Odds ratio (95% CI)‡
	Pro *N* = 200	Sib *N* = 200	Pro *N* = 200	Sib *N* = 200	Pro *N* = 200	Sib *N* = 200		
De novo								
				All genes				
All	154	125 §	0.77	0.63	1.58	1.31	0.09	NA
Silent	29	39	0.15	0.20	0.29	0.40	0.28	NA
All non-synonymous	125	87	0.63	0.44	1.29	0.92	**0.01**	1.93 (1.11–3.36)
Missense	110	82	0.55	0.41	1.13	0.86	0.05	1.80 (1.03–3.16)
Nonsense/splice site	15	5	0.08	0.03	0.16	0.05	**0.04**	4.03 (1.32-12.4)
				Brain-expressed genes				
All	137	96	0.69	0.48	1.41	1.01	**0.01**	NA
Silent	23	30	0.12	0.15	0.24	0.31	0.41	NA
All non-synonymous	114	67	0.57	0.34	1.18	0.71	**0.001**	2.22 (1.19–4.13)
Missense	101	64	0.51	0.32	1.04	0.68	**0.005**	2.06 (1.10–3.85)
Nonsense/splice site	13	3	0.07	0.02	0.14	0.03	**0.02**	5.65 (1.44–22.2)
Novel transmitted								
				All genes				
All	26,565	26,542	133	133	277	277	0.92	NA
Silent	8,567	8,642	43	43	90	91	0.57	NA
All non-synonymous	17,998	17,900	90	90	188	187	0.61	1.01 (0.98–1.05)
Missense	17,348	17,250	87	86	181	180	0.60	1.01 (0.98–1.05)
Nonsense/splice site	650	650	3.3	3.3	7	7	1.00	1.01 (0.90–1.13)
				Brain-expressed genes				
All	20,942	20,982	105	105	219	220	0.85	NA
Silent	6,884	6,981	34	35	72	74	0.42	NA
All non-synonymous	14,058	14,001	70	70	147	146	0.74	1.02 (0.98–1.06)
Missense	13,588	13,525	68	68	142	141	0.71	1.02 (0.98–1.06)
Nonsense/splice site	470	476	2.3	2.4	5	5	0.87	1.00 (0.88–1.14)

* An additional 15 *de novo* variants were seen in the probands of 25 trio families; all were missense and 14 were brain-expressed.
† The *P* values compare the number of variants between probands and siblings using a two-tailed binomial exact test (Supplementary Information); *P* values below 0.05 are highlighted in bold.
‡ The odds ratio calculates the proportion of variants in a specific category to silent variants and then compares these ratios in probands versus siblings. NA, not applicable.
§ The sum of silent and non-synonymous variants is 126, however one nonsense and two silent *de novo* variants were indentified in *KANK1* in a single sibling, suggesting a single gene conversion event. This event contributed a maximum count of one to any analysis.

©2012 Macmillan Publishers Limited. All rights reserved

Table 2 | Loss of function mutations in probands

Gene symbol	Gene name	Mutation type
ADAM33	ADAM metallopeptidase domain 33	Nonsense
CSDE1	cold shock domain containing E1, RNA-binding	Nonsense
EPHB2	EPH receptor B2	Nonsense
FAM8A1	family with sequence similarity 8, member A1	Nonsense
FREM3	FRAS1 related extracellular matrix 3	Nonsense
MPHOSPH8	M-phase phosphoprotein 8	Nonsense
PPM1D	protein phosphatase, Mg^{2+}/Mn^{2+} dependent 1D	Nonsense
RAB2A	RAB2A, member RAS oncogene family	Nonsense
SCN2A	sodium channel, voltage-gated, type II, α subunit	Nonsense
SCN2A	sodium channel, voltage-gated, type II, α subunit	Nonsense
BTN1A1	butyrophilin, subfamily 1, member A1	Splice site
FCRL6	Fc receptor-like 6	Splice site
KATNAL2	katanin p60 subunit A-like 2	Splice site
NAPRT1	nicotinate phosphoribosyltransferase domain containing 1	Splice site
RNF38	ring finger protein 38	Splice site
SCP2	sterol carrier protein 2	Frameshift*
SHANK2	SH3 and multiple ankyrin repeat domains 2	Frameshift*

* Frameshift *de novo* variants are not included in any of the reported case-control comparisons (Supplementary Information).

we hypothesized that estimating the probability of observing multiple independent *de novo* SNVs in the same gene in unrelated individuals would provide a more powerful statistical approach to identifying ASD-risk genes than the alternative of comparing mutation counts in affected versus unaffected individuals. Consequently, we conducted simulation experiments focusing on *de novo* SNVs in brain-expressed genes, using the empirical data for per-base mutation rates and taking into account the actual distribution of gene sizes and GC content across the genome (Supplementary Information). We calculated probabilities (P) and the false discovery rate (Q) based on a wide range of assumptions regarding the number of genes conferring ASD risk (Supplementary Fig. 7; Fig. 2). On the basis of 150,000 iterations, we determined that under all models, two or more nonsense and/or splice site *de novo* mutations were highly unlikely to occur by chance ($P = 0.008$; $Q = 0.005$; Supplementary Information; Fig. 2a). Importantly, these thresholds were robust both to sample size, and to variation in our estimates of locus heterogeneity. Similarly, in our sample, two or more nonsense or splice site *de novo* mutations remained statistically significant when the simulation was performed using the lower bound of the 95% confidence interval for the estimate of *de novo* mutation rates in probands (Supplementary Fig. 7).

Only a single gene in our cohort, *SCN2A*, met these thresholds ($P = 0.008$; Fig. 2a), with two probands each carrying a nonsense *de novo* SNV (Table 2). This finding is consistent with a wealth of data showing overlap of genetic risks for ASD and seizure[8]. Gain of function mutations in *SCN2A* are associated with a range of epilepsy phenotypes; a nonsense *de novo* mutation has been described in a patient with infantile epileptic encephalopathy and intellectual decline[9], *de novo* missense mutations with variable electrophysiological effects have been found in cases of intractable epilepsy[10], and transmitted rare missense mutations have been described in families with idiopathic ASD[11]. Of note, the individuals in the SSC carrying the nonsense *de novo* SNVs have no history of seizure.

We then considered whether alternative approaches described in the recent literature[4,12], including identifying multiple *de novo* events in a single individual or predicting the functional consequences of missense mutations, might help identify additional ASD-risk genes. However, we found no differences in the distribution or frequency of multiple *de novo* events within individuals in the case versus the control groups (Fig. 1c). In addition, when we examined patients carrying large *de novo* ASD-risk CNVs, we found a trend towards fewer non-synonymous *de novo* SNVs (Supplementary Fig. 11; Supplementary Information). Consequently, neither finding supported a 'two *de novo* hit' hypothesis. Similarly, we found no evidence that widely used measures of conservation or predictors of protein disruption, such as PolyPhen2[13], SIFT[14], GERP[15], PhyloP[16] or Grantham Score[17],

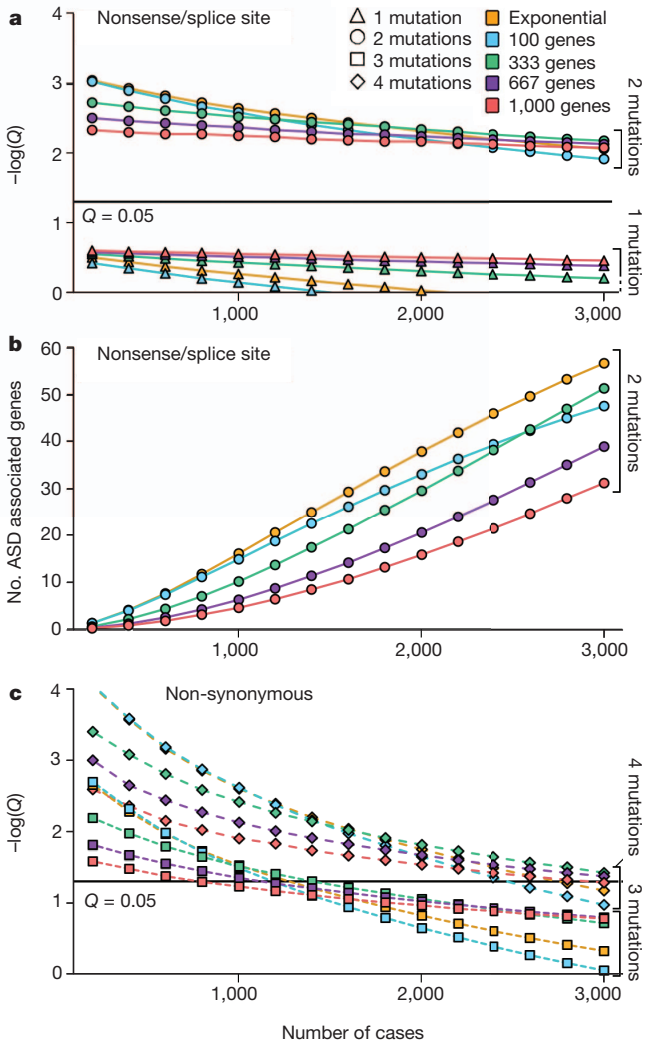

Figure 2 | Identification of multiple *de novo* mutations in the same gene reliably distinguishes risk-associated mutations. **a**, Results of a simulation experiment modelling the likelihood of observing two independent nonsense/splice site *de novo* mutations in the same brain-expressed gene among unrelated probands. We modelled the observed rate of *de novo* brain-expressed mutations in probands and siblings, gene size, GC content and varying degrees of locus heterogeneity, including 100, 333, 667 or 1,000 ASD-contributing genes, as well as using the top 1% of genes derived from a model of exponential distribution of risk (indicated by colour). A total of 150,000 iterations were run. The rate of occurrences of two or more *de novo* variants in non-ASD genes was used to estimate the *P*-value (Supplementary Fig. 7) while the ratio of occurrences of two or more *de novo* variants in non-ASD genes to similar occurrences in ASD genes was used to estimate the false discovery rate (Q). The identification of two independent nonsense/splice site *de novo* variants in a brain-expressed gene in this sample provides significant evidence for ASD association ($P = 0.008$; $Q = 0.005$) for all models. This observation remained statistically significant when the simulation was repeated using the lower bound of the 95% confidence interval for the estimate of the *de novo* mutation rate in probands (Supplementary Fig. 7). **b**, The simulation described in **a** was used to predict the number of genes that will be found to carry two or more nonsense/splice site *de novo* mutations for a sample of a given size (specified on the *x* axis). **c**, The simulation was repeated for non-synonymous *de novo* mutations. The identification of three or more independent non-synonymous *de novo* mutations in a brain-expressed gene provides significant evidence for ASD association ($P < 0.05$; $Q < 0.05$) in the sample reported here, however these thresholds are sensitive both to sample size and heterogeneity models.

either alone or in combination differentiated *de novo* non-synonymous SNVs in probands compared to siblings (Supplementary Fig. 9; Supplementary Information). Additionally, among probands, the *de*

©2012 Macmillan Publishers Limited. All rights reserved

novo SNVs in our study were not significantly over-represented in previously established lists of synaptic genes[18–20], genes on chromosome X, autism-implicated genes[2], intellectual disability genes[2], genes within ASD-risk associated CNVs[3] or *de novo* non-synonymous SNVs identified in schizophrenia probands[12,21]. Finally we conducted pathway and protein–protein interaction analyses[22] for all non-synonymous *de novo* SNVs, all brain-expressed non-synonymous *de novo* SNVs and all nonsense and splice site *de novo* SNVs (Supplementary Fig. 9, 10; Supplementary Information) and did not find a significant enrichment among cases versus controls that survived correction for multiple comparisons, though these studies were of limited power.

These analyses demonstrate that neither the type nor the number of *de novo* mutations observed solely in a single individual provides significant evidence for association with ASD. Moreover, we determined that in the SSC cohort at least three, and most often four or more, brain-expressed non-synonymous *de novo* SNVs in the same gene would be necessary to show a significant association (Fig. 2c; Supplementary Figs 7, 8). Unlike the case of disruptive nonsense and splice site mutations, these simulations were highly sensitive to both sample size and heterogeneity models (Fig. 2c; Supplementary Figs 7, 8; Supplementary Information).

Finally, at the completion of our study, we had the opportunity to combine all *de novo* events in our sample with those identified in an independent whole-exome analysis of non-overlapping Simons Simplex families that focused predominantly on trios[23]. From a total of 414 probands, two additional genes were found to carry two highly disruptive mutations each, *KATNAL2* (katanin p60 subunit A-like 2) (our results and ref. 23) and *CHD8* (chromodomain helicase DNA binding protein 8) (ref. 23), thereby showing association with the ASD phenotype.

Overall, our results substantially clarify the genomic architecture of ASD, demonstrate significant association of three genes—*SCN2A*, *KATNAL2* and *CHD8*—and predict that approximately 25–50 additional ASD-risk genes will be identified as sequencing of the 2,648 SSC families is completed (Fig. 2b). Rare non-synonymous *de novo* SNVs are associated with risk, with odds ratios for nonsense and splice-site mutations in the range previously described for large multigenic *de novo* CNVs[3]. It is important to note that these estimates reflect a mix of risk and neutral mutations in probands. We anticipate that the true effect size for specific SNVs and mutation classes will be further clarified as more data accumulate. From the distribution of large multi-genic *de novo* CNVs in probands versus siblings, we previously estimated the number of ASD-risk loci at 234 (ref. 3). Using the same approach, the current data result in a point estimate of 1,034 genes, however the confidence intervals are large and the distribution of this risk among these loci is unknown (Supplementary Information). What is clear is that our results strongly support a high degree of locus heterogeneity in the SSC cohort, involving hundreds of genes or more. Finally, via examination of mutation rates in well-matched controls, we have determined that the observation of highly disruptive *de novo* SNVs clustering within genes can robustly identify risk-conferring alleles.

The focus on recurrent rare *de novo* mutation described here provided sufficient statistical power to identify associated genes in a relatively small cohort—despite both a high degree of locus heterogeneity and the contribution of intermediate genetic risks. This approach promises to be valuable for future high-throughput sequencing efforts in ASD and other common neuropsychiatric disorders.

METHODS SUMMARY

Sample selection. In total 238 families (928 individuals) were selected from the SSC[5]. Thirteen families (6%) did not pass quality control, leaving 225 families (200 quartets, 25 trios) for analysis (Supplementary Data 1). Of the 200 quartets, 194 (97%) probands had a diagnosis of autism and 6 (3%) were diagnosed with ASD; the median non-verbal IQ was 84.

Exome capture, sequencing and variant prediction. Whole-blood DNA was enriched for exonic sequences through hybridization with a NimbleGen custom array ($N = 210$) or EZExomeV2.0 ($N = 718$). Captured DNA was sequenced using an Illumina GAIIx ($N = 592$) or HiSeq 2000 ($N = 336$). Short read sequences were aligned to hg18 with BWA[6], duplicate reads were removed and variants were predicted using SAMtools[7]. Data were normalized within families by only analysing bases with at least 20 unique reads in all family members. *De novo* predictions were made blinded to affected status using experimentally verified thresholds (Supplementary Information). All *de novo* variants were confirmed using Sanger sequencing blinded to affected status.

Gene annotation. Variants were analysed against RefSeq hg18 gene definitions; in genes with multiple isoforms the most severe outcome was chosen. All nonsense and canonical splice site variants were present in all RefSeq isoforms. A variant was listed as altering the splice site only if it disrupted canonical 2-base-pair acceptor (AG) or donor (GT) sites. Brain-expressed genes were identified from expression array analysis across 57 post-mortem brains (age 6 weeks post conception to 82 years) and multiple brain regions; 80% of RefSeq genes were included in this subset[24].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 September 2011; accepted 14 February 2012.
Published online 4 April 2012.

1. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
2. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
3. 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).
4. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genet.* **43,** 585–589 (2011).
5. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
8. Meisler, M. H., O'Brien, J. E. & Sharkey, L. M. Sodium channel gene family: epilepsy mutations, gene interactions and modifier effects. *J. Physiol. (Lond.)* **588,** 1841–1848 (2010).
9. Kamiya, K. *et al.* A nonsense mutation of the sodium channel gene *SCN2A* in a patient with intractable epilepsy and mental decline. *J. Neurosci.* **24,** 2690–2698 (2004).
10. Ogiwara, I. *et al. De novo* mutations of voltage-gated sodium channel alphall gene *SCN2A* in intractable epilepsies. *Neurology* **73,** 1046–1053 (2009).
11. Weiss, L. A. *et al.* Sodium channels *SCN1A*, *SCN2A* and *SCN3A* in familial autism. *Mol. Psychiatry* **8,** 186–194 (2003).
12. Xu, B. *et al.* Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nature Genet.* **43,** 864–868 (2011).
13. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
14. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4,** 1073–1081 (2009).
15. Cooper, G. M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods* **7,** 250–251 (2010).
16. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
17. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).
18. Abul-Husn, N. S. *et al.* Systems approach to explore components and interactions in the presynapse. *Proteomics* **9,** 3303–3315 (2009).
19. Bayés, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature Neurosci.* **14,** 19–21 (2011).
20. Collins, M. O. *et al.* Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J. Neurochem.* **97** (suppl. 1), 16–23 (2006).
21. Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genet.* **43,** 860–863 (2011).
22. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
23. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* http://dx.doi.org/10.1038/nature10989 (this issue).
24. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to all of the families participating in the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (SSC). This work was supported by a grant from the Simons Foundation. R.P.L. is an Investigator of the

©2012 Macmillan Publishers Limited. All rights reserved

Howard Hughes Medical Institute. We 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 sites for the recruitment and comprehensive assessment of simplex families; the SFARI staff, in particular M. Benedetti, for facilitating access to the SSC; Prometheus Research for phenotypic data management and Prometheus Research and the Rutgers University Cell and DNA repository for accessing biomaterials; the Yale Center of Genomic Analysis, in particular M. Mahajan, S. Umlauf, I. Tikhonova and A. Lopez, for generating sequencing data; T. Brooks-Boone, N. Wright-Davis and M. Wojciechowski for their help in administering the project at Yale; I. Hart for support; G. D. Fischbach, A. Packer, J. Spiro, M. Benedetti and M. Carlson for their suggestions throughout; and B. Neale and M. Daly for discussions regarding *de novo* variation. We also acknowledge T. Lehner and the Autism Sequencing Consortium for providing an opportunity for pre-publication data exchange among the participating groups.

Author Contributions S.J.S., M.T.M., R.P.L., M.G., D.H.G. and M.W.S. designed the study; M.T.M., A.R.G., J.M., M.R., A.G.E.-S., N.M.D., S.M., M.W., G.O., Y.S., P.E., R.M. and J.O. designed and performed high-throughput sequencing experiments and variant confirmations; S.J.S., M.C., K.B., R.B. and N.C. designed the exome-analysis bioinformatics pipeline; S.J.S., A.J.W., N.N.P., J.L.S., N.T., K.A.M., N.Š., K.R., D.H.G., B.D. and M.W.S. analysed the data; S.J.S., A.J.W., K.R., B.D. and M.W.S. wrote the paper; J.M., M.R., A.J.W., A.R.G., A.G.E.-S. and N.M.D. contributed equally to the study. All authors discussed the results and contributed to editing the manuscript.

Author Information Sequence data from this study is available through the NCBI Sequence Read Archive (accession number SRP010920.1). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.W.S. (matthew.state@yale.edu), B.D. (devlinbj@upmc.edu) or D.H.G. (dhg@mednet.ucla.edu).

©2012 Macmillan Publishers Limited. All rights reserved

METHODS

Sample selection. In total 238 families (928 individuals) were selected from the SSC on the basis of: male probands with autism, low non-verbal IQ (NVIQ), and discordant Social Responsiveness Scale (SRS) with sibling and parents ($N = 40$); female probands ($N = 46$); multiple unaffected siblings ($N = 28$); probands with known multigenic CNVs ($N = 15$); and random selection ($N = 109$). Thirteen families (6%) did not pass quality control (Supplementary Information) leaving 225 families (200 quartets, 25 trios) for analysis (Supplementary Data 1). Of the 200 quartets, 194 (97%) probands had a diagnosis of autism and 6 (3%) were diagnosed with ASD; the median NVIQ was 84. Three of these quartets have previously been reported as trios[4]; there is no overlap between the current sample and those presented in the companion article[23].

Exome capture, sequencing and variant prediction. Whole-blood DNA was enriched for exonic sequences (exome capture) through hybridization with a NimbleGen custom array ($N = 210$) or EZExomeV2.0 ($N = 718$). The captured DNA was sequenced using an Illumina GAIIx ($N = 592$) or HiSeq 2000 ($N = 336$). Short read sequences were aligned to hg18 with BWA[6], duplicate reads were removed and variants were predicted using SAMtools[7]. The data were normalized across each family by only analysing bases with at least 20 unique reads in all family members (Supplementary Information). *De novo* predictions were made blinded to affected status using experimentally verified thresholds (Supplementary Information). All *de novo* variants were confirmed using Sanger sequencing blinded to affected status.

Variant frequency. The allele frequency of a given variant in the offspring was determined by comparison with dbSNPv132 and 1,637 whole-exome controls including 400 parents. Variants were classified as: 'novel', if only a single allele was present in a parent and none were seen in dbSNP or the other control exomes; 'rare', if they did not meet the criteria for novel and were present in $<1\%$ of controls; and 'common', if they were present in $\geq 1\%$ of controls.

Gene annotation. Variants were analysed against the RefSeq hg18 gene definitions, a list that includes 18,933 genes. Where multiple isoforms gave varying results the most severe outcome was chosen. All nonsense and canonical splice site variants were checked manually and were present in all RefSeq isoforms. A variant was listed as altering the splice site only if it disrupted canonical 2-base-pair acceptor (AG) or donor (GT) sites.

Brain-expressed genes. A list of brain-expressed genes was obtained from expression array analysis across 57 post-mortem brains (age 6 weeks post conception to 82 years) and multiple brain regions[24]. Using these data, 14,363 (80%) of genes were classified as brain-expressed (Supplementary Information).

Rate of *de novo* SNVs. To allow an accurate comparison between the *de novo* burden in probands and siblings, the number of *de novo* SNVs found in each sample was divided by the number of bases analysed (that is, bases with ≥ 20 unique reads in all family members) to calculate a per-base rate of *de novo* SNVs. Rates are given in Table 1.

Simulation model. The likelihood of observing multiple independent *de novo* events of a given type for a given sample size in an ASD risk-conferring gene was modelled using gene size and GC content (derived from the full set of brain-expressed RefSeq genes) and the observed rate of brain-expressed *de novo* variants in probands and siblings. These values were then used to evaluate the number of genes contributing to ASD showing two or more variants of the specified type (Fig. 2); comparing this to the number of genes with similar events not carrying ASD risk gave the likelihood of the specified pattern demonstrating association with ASD. The simulation was run through 150,000 iterations across a range of samples sizes and multiple models of locus heterogeneity (Supplementary Information).

Severity scores. Severity scores were calculated for missense variants using web-based interfaces for PolyPhen2[13], SIFT[14] and GERP[15], using the default settings (Supplementary Information). PhyloP[16] and Grantham Score[17] were determined using an in-house annotated script. For nonsense/splice site variants the maximum score was assigned for Grantham, SIFT and PolyPhen2; for GERP and PhyloP, every possible coding base for the specific protein was scored and the highest value selected.

Pathway analysis. The list of brain-expressed genes with non-synonymous *de novo* SNVs was submitted to KEGG using the complete set of 14,363 brain-expressed genes as the background to prevent bias. For IPA the analysis was based on human nervous system pathways only, again to prevent bias. Otherwise default settings were used for both tools.

Protein–protein interactions. Genes with brain-expressed non-synonymous *de novo* variants in probands were submitted to the Disease Association Protein–protein Link Evaluator (DAPPLE)[22] using the default settings.

Comparing *de novo* SNV counts to gene lists. To assess whether non-synonymous *de novo* SNVs were enriched in particular gene sets, the chance of seeing a *de novo* variant in each gene on a given list was estimated based on the size and GC content of the gene. The observed number of *de novo* events was then assessed using the binomial distribution probability based on the total number of non-synonymous *de novo* variants in probands and the sum of probabilities for *de novo* events within these genes.

©2012 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature10945

Table of Contents

1. Supplementary Figures and Legends

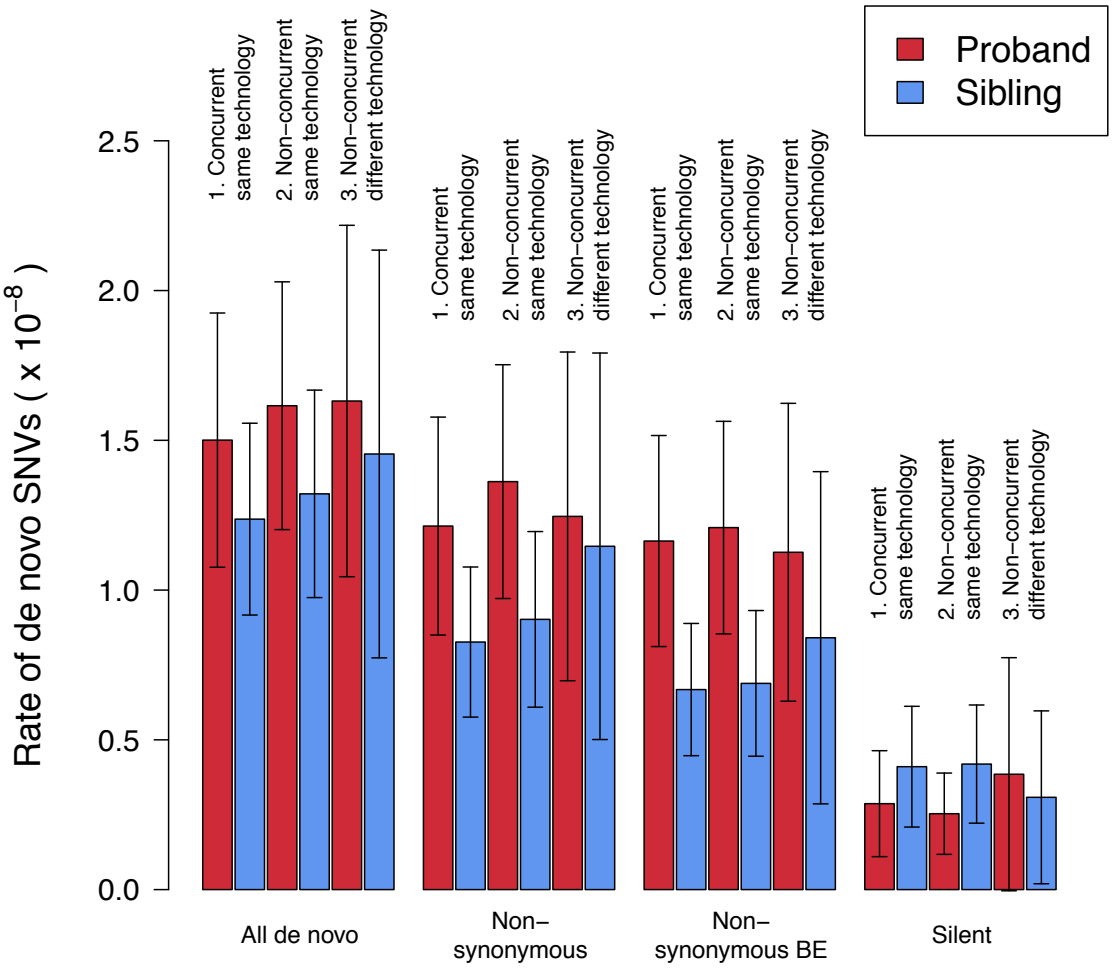

Figure S1. *De novo* rate of mutation between probands and siblings split by consistency of data generation between probands and siblings.

In 38% of families the proband and sibling were hybridized using the same version of the NimbleGen capture array and sequenced on the same flowcell (group 1 – Concurrent, same technology); in 44.5% of families the proband and sibling were hybridized using the same version of the NimbleGen capture array and analyzed on the same sequencing instrument (Illumina GAIIx or HiSeq2000) however, they were not run concurrently on the same flowcell (group 2 – non-concurrent, same technology); finally, 17.5% of families were hybridized on different versions of the NimbleGen capture array and/or different sequencing instruments (group 3 – non-concurrently, different technology). The greatest difference in *de novo* rate between probands and siblings was seen for the samples in which probands and siblings were analyzed in the most consistent manner supporting the conclusion that the difference in *de novo* mutation rate seen in the experiment (Fig1A, main manuscript) was not the result of different technology but a true biological signal present in the samples analyzed. Error bars represent the 95% confidence interval; N=200.

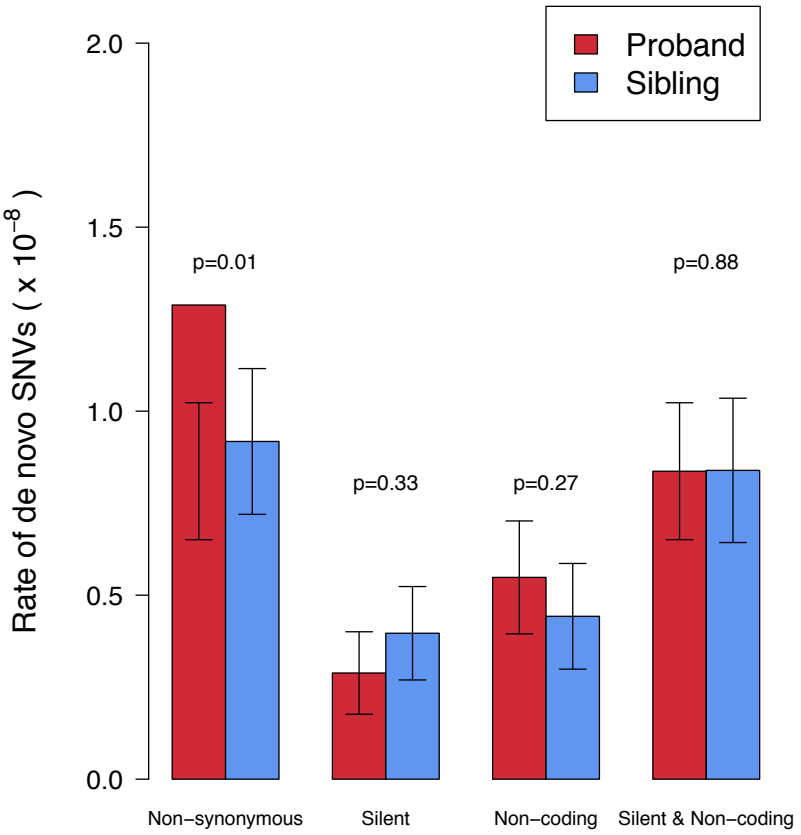

Figure S2. *De novo* rate in probands and siblings comparing non-synonymous variants with silent and non-coding variants.

The rate of *de novo* variants in probands and siblings is shown for different categories of predicted variants. Of note, non-coding variants were not confirmed by PCR (these are the only *de novo* data that are not confirmed in the entire manuscript). However, other categories of *de novo* mutation predicted using identical thresholds showed a 96% confirmation rate. No systematic increase in *de novo* detection is observed for probands compared to siblings. Error bars represent the 95% confidence interval and p-values are calculated with a two-tailed Wilcoxon test; N=200.

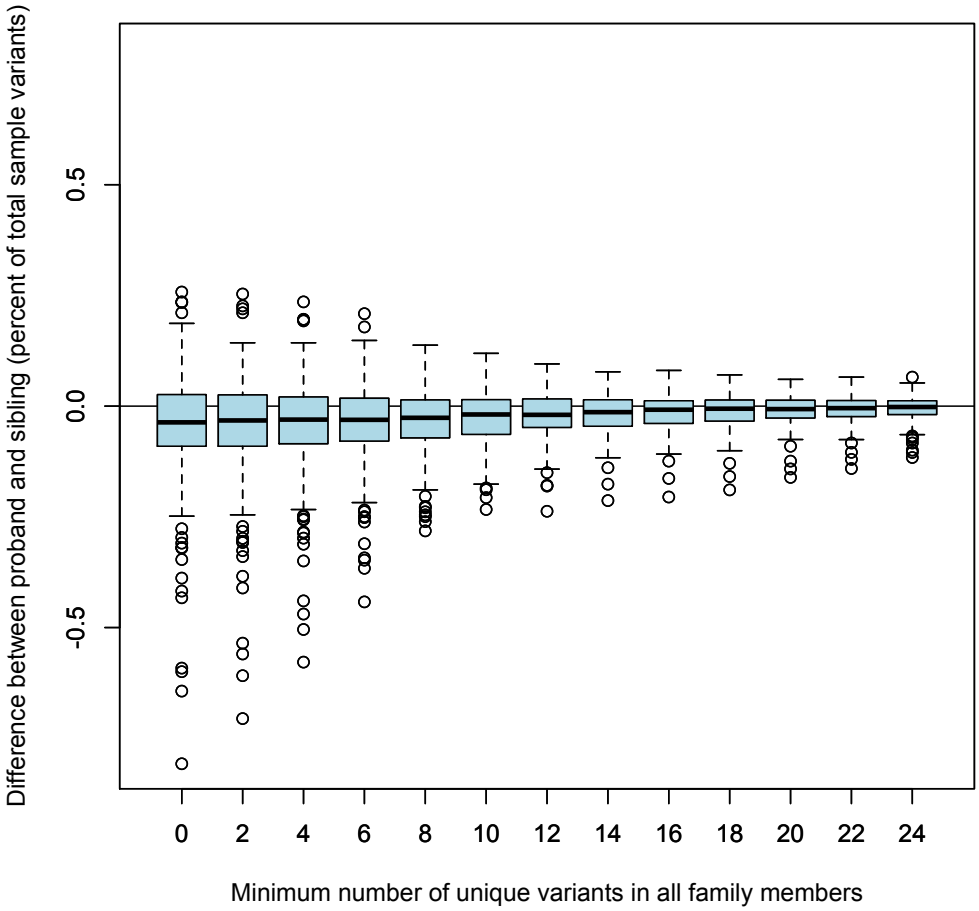

Figure S3. Minimal difference in detection accuracy for novel variants at a threshold of 20 unique reads.

To demonstrate that the detection accuracy for *de novo* variants was equal between probands and siblings, the rate of novel variants, those in which a single allele was present in a parent and not seen in dbSNP or the other control exomes, is plotted against the threshold used for minimum number of unique sequencing reads in all family members at that position. If a variant was at a position in which any one family member did not meet this threshold the variant would not be included. These variants were detected using the same detection criteria as for *de novo* variants. For each family the difference in the number of variants in probands and siblings was calculated (probands variants – sibling variants). To allow for a straightforward comparison, this difference is then expressed as a percentage of the average number of variants detected in all samples at this threshold: (probands variants – sibling variants) / average variants per sample. The 'percentage difference' for each family is shown as a boxplot. At a minimum number of unique reads of 20 in all family members the percentage difference is almost 0 with a slight bias toward siblings. Error bars represent the last data point 1.5 times the IQR from the median; outliers from this range are shown as points; N=200.

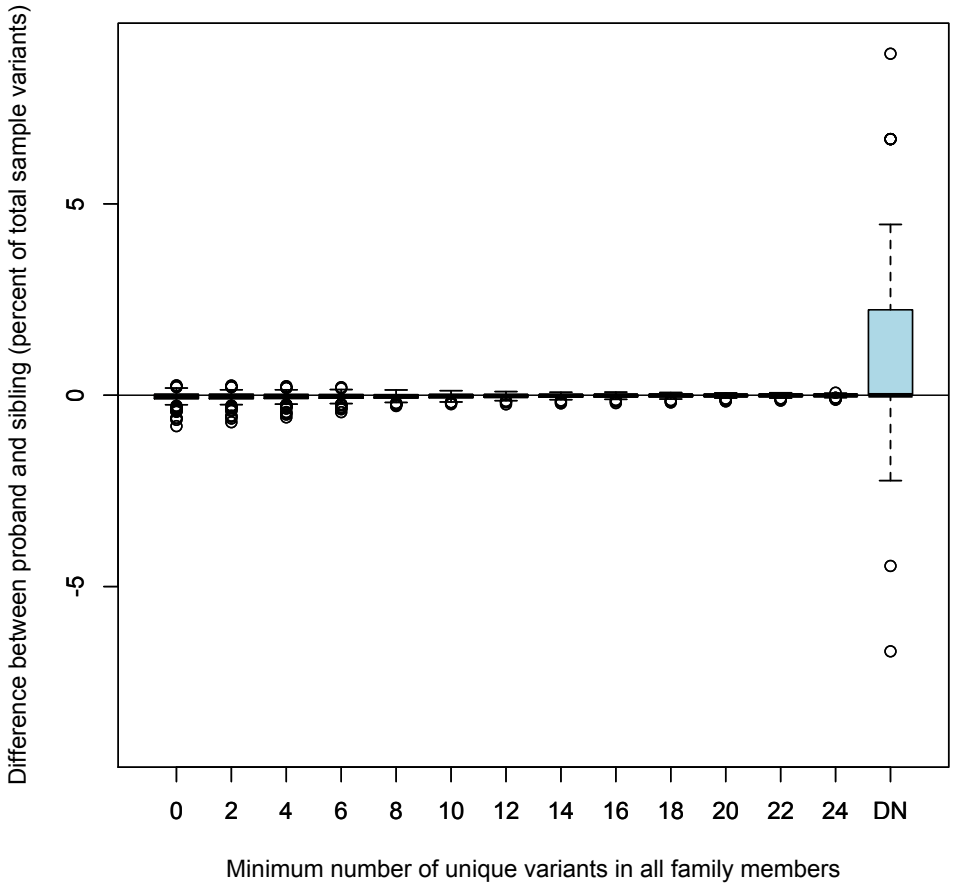

Figure S4. Minimal difference in detection accuracy for novel variants at a threshold of 20 unique reads with the corresponding value for *de novo* variant detection included.

This plot shows the same data as for FigS3, however the corresponding plot for *de novo* variants has been added on the far right and the y-axis has been rescaled. The mean 'percentage difference' for *de novo* variants at ≥20 unique reads in all family members is 66-fold higher than the corresponding value for novel transmitted variants. Error bars represent the last data point 1.5 times the IQR from the median; outliers from this range are shown as points; N=200.

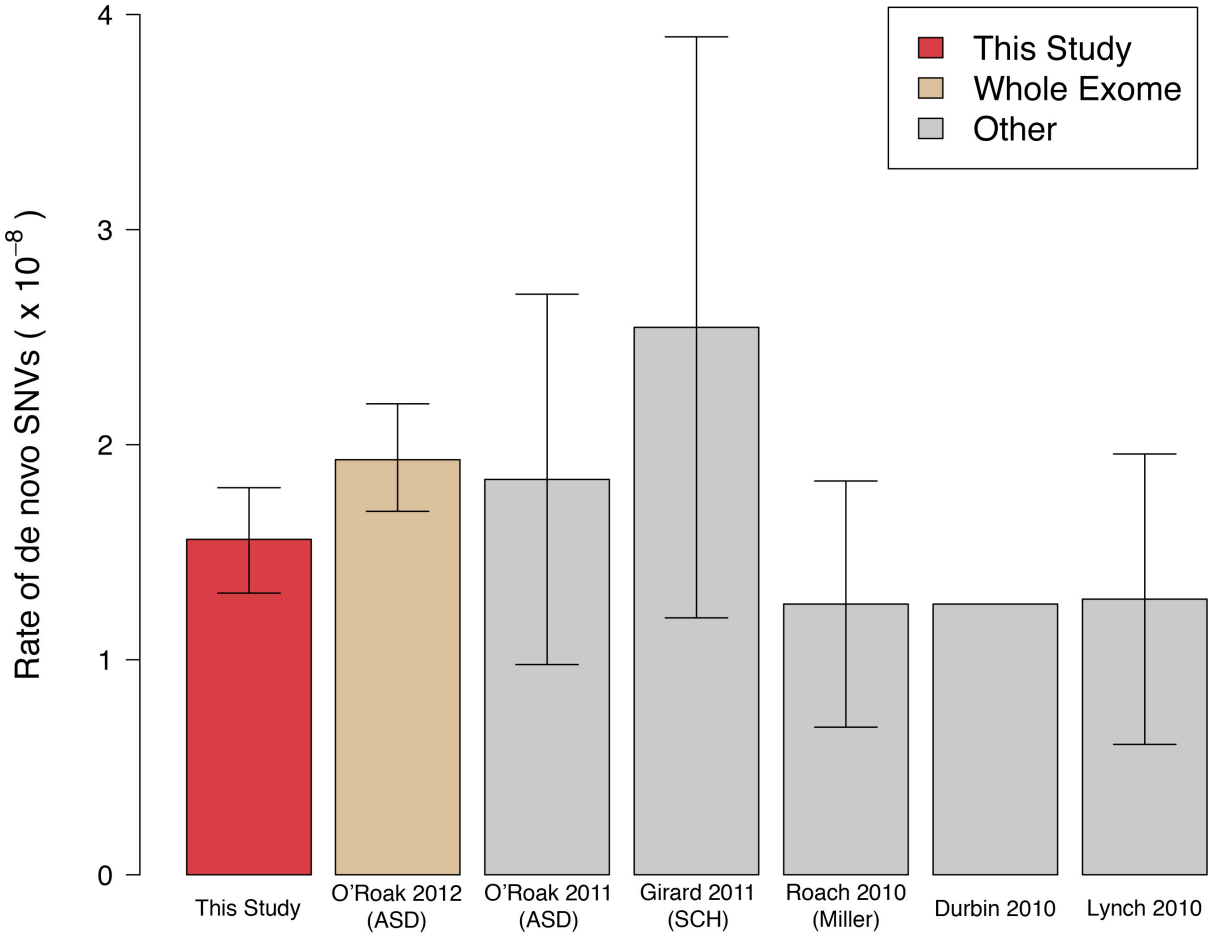

Figure S5. Rate of *de novo* variation across multiple studies.

Our observed rate of *de novo* variants in coding and splice sites is compared to the published haploid *de novo* rate estimated in seven studies of humans, including our two companion papers. Four of these are based on whole-exome sequence in affected probands with ASD[1] and Schizophrenia.[2] Roach et al. use whole-genome sequencing in a single family with Miller syndrome;[3] Durbin et al.[4] show the rate estimated by the 1000Genomes Consortium; Lynch[5] used dominant patterns of human disease to estimate the rate based on population prevalence. Estimates based on whole-genome data (Roach and Durbin) have been adjusted for GC content to be comparable with coding regions by increasing the rate by 1.15 estimated by assessing GC content in RefSeq exons and introns (excluding 50bp nearest exon boundaries). Error bars represent the 95% confidence interval.

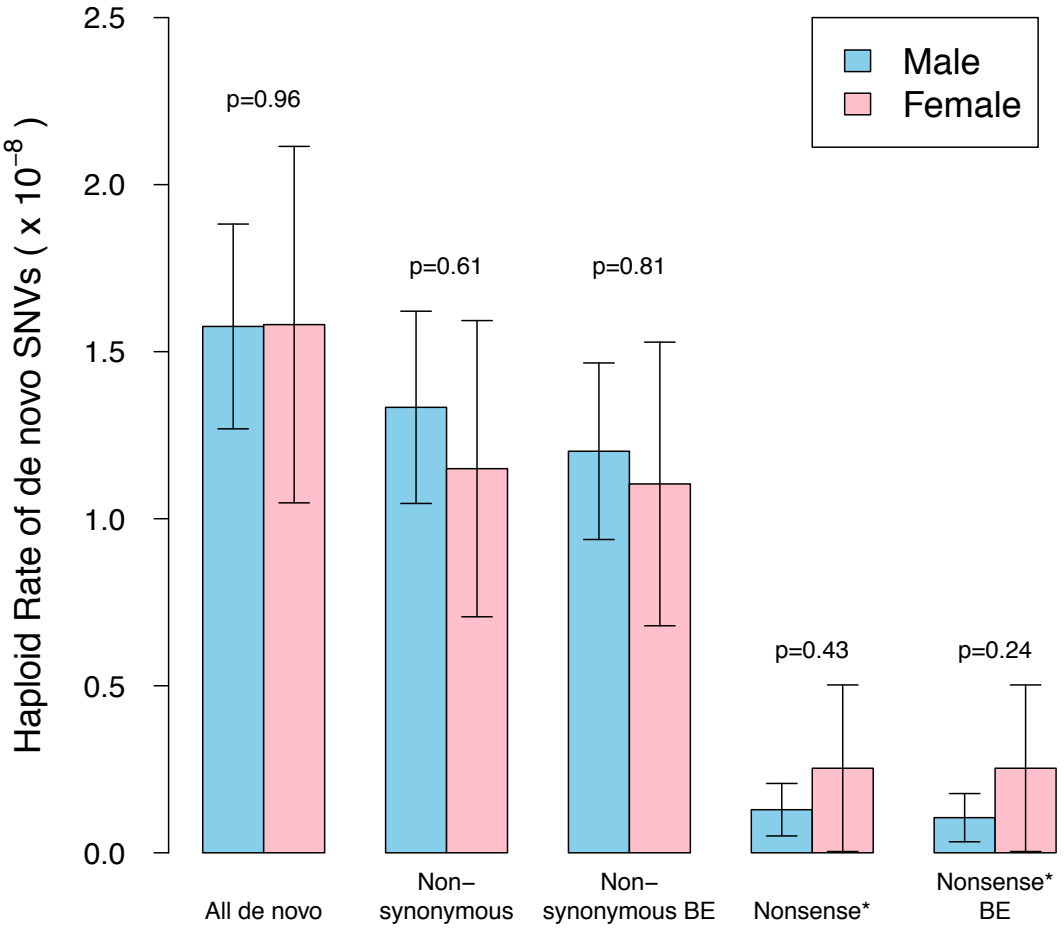

Figure S6. Comparison of the rate of *de novo* point mutations in male and female probands. No difference in the rate of *de novo* SNVs was observed between 151 male probands and 49 female probands. The rate in siblings was approximately equal between the sexes too (data not shown). Error bars represent the 95% confidence interval and the p-values shown are calculated by comparing the rate of *de novo* SNVs in all samples using a two-tailed Wilcoxon test. No significant relationship was seen when a Poisson regression taking parental age into account was used.

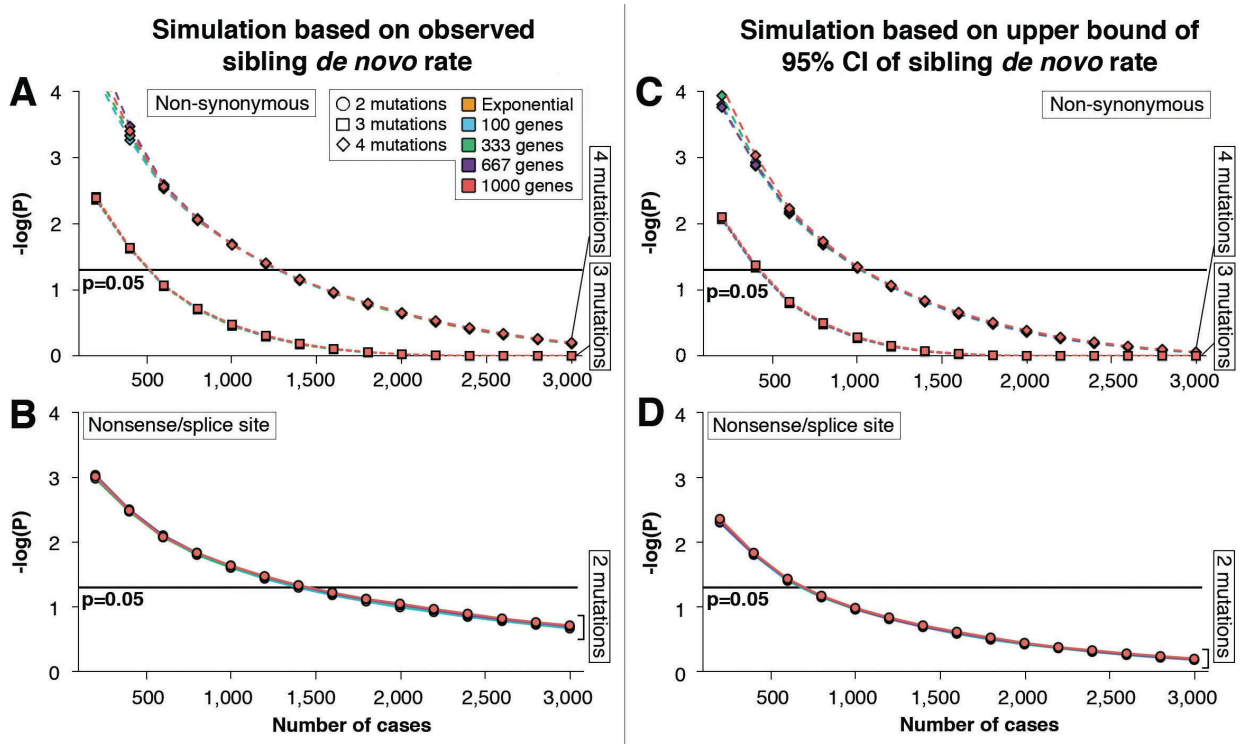

Figure S7: Probability of seeing multiple *de novo* variants in the same non-ASD gene by chance (p-value).

A) This plot shows the p-value (P) of seeing at least one gene with multiple independent non-synonymous *de novo* variants for a given sample size. The probability of a gene with multiple independent non-synonymous *de novo* variants contributing ASD risk is shown in figure S8. The p-value (P) was estimated from a simulation experiment based on: the observed rate of non-synonymous *de novo* brain-expressed mutations in siblings (0.71×10^{-8}; Table 1); gene size; GC content; and an estimate of locus heterogeneity (we evaluated various models including 100, 333, 667, or 1,000 contributing genes, as well as using the top 1% of genes derived from a model of exponential distribution of risk). A total of 150,000 iterations were run. The p-value is calculated as the number of iterations in which a sibling had ≥3 or ≥4 mutations in the same non-ASD gene divided by the total number of iterations. The observation of ≥3 *de novo* non-synonymous mutations present in the same gene in different probands is significant ($p<0.05$) evidence for ASD association for 225 families. **B)** Shows the same approach to a simulation experiment as 'A', but estimates the p-value (P) of observing ≥2 independent nonsense/splice site *de novo* variants in the same brain-expressed gene by chance. The observed rate of nonsense/splice site *de novo* mutation in siblings was 0.03×10^{-8} (Table 1). The identification of two or more independent nonsense/splice site *de novo* variants in a brain-expressed gene provides significant evidence for ASD association ($p=0.008$) for 225 families. **C)** The simulation shown in 'A' is repeated using the upper bound of the 95% confidence interval for the rate of non-synonymous brain-expressed *de novo* variants in siblings (0.88×10^{-8}) as a conservative estimate. Three *de novo* non-synonymous variants in the same gene provide evidence of association ($p<0.05$) for all models of locus heterogeneity at 225 families. **D)** The simulation shown in 'B' is repeated using the upper bound of the 95% confidence interval for the rate of nonsense/splice site brain-expressed *de novo* variants in siblings (0.06×10^{-8}). The threshold of

significance ($p<0.05$) remains at ≥2 *de novo* variants in the same gene for up to 700 families.

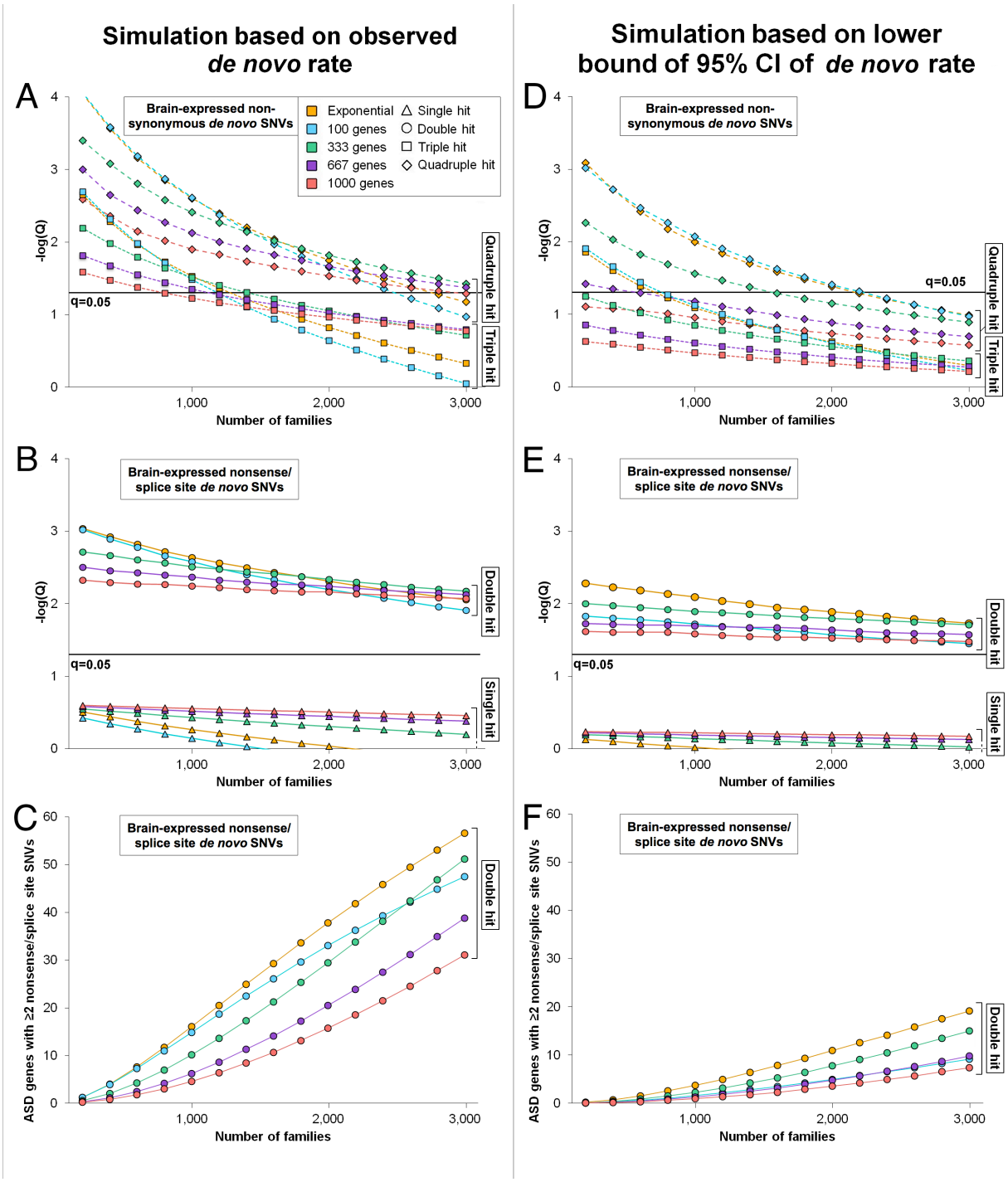

Figure S8: Probability of a gene with multiple *de novo* variants contributing ASD risk (q-value).

A) This plot shows the false discovery rate (Q) of using multiple independent non-synonymous *de novo* variants to detect genes with ASD risk for a given sample size. The probability of observing multiple independent non-synonymous *de novo* variants in a gene that does not contribute ASD risk at least once is shown in figure S7. The false discovery rate (Q) was estimated from a simulation experiment based on: the observed rate of non-synonymous *de novo*

brain-expressed mutations in probands and siblings (1.18×10^{-8} and 0.71×10^{-8} respectively; Table 1); gene size; GC content; and an estimate of locus heterogeneity (we evaluated various models including 100, 333, 667, or 1,000 contributing genes, as well as using the top 1% of genes derived from a model of exponential distribution of risk). A total of 150,000 iterations were run. The false discovery rate is calculated as the number of observations of non-ASD risk genes with ≥3 or ≥4 non-synonymous mutations in probands over the corresponding number of observations in ASD risk genes. The observation of ≥3 *de novo* non-synonymous mutations present in the same gene in different probands is significant ($q<0.05$) evidence for ASD association for 225 families. **B)** Shows the same approach to a simulation experiment as 'A', but estimates the false discovery rate (Q) of ascribing ASD association to a brain-expressed gene with ≥2 independent nonsense/splice site *de novo* variants. The observed rate of *de novo* nonsense/splice site mutations in probands and siblings was 0.14×10^{-8} and 0.03×10^{-8} respectively (Table 1). The identification of ≥2 independent nonsense/splice site *de novo* variants in a brain-expressed gene provides significant evidence for ASD association ($q=0.005$) for 225 families. **C)** The simulation described in 'B' was used to predict the number of genes conferring ASD risk that will be identified by the observation of ≥2 independent nonsense/splice site *de novo* mutations for a sample of a given size (specified on the x-axis). Predictions are given for the specified models of locus heterogeneity. **D)** The simulation shown in 'A' is repeated using the lower bound of the 95% confidence interval for the rate of non-synonymous brain-expressed *de novo* variants in probands (0.95×10^{-8}). The threshold of significance ($q<0.05$) remains at ≥3 *de novo* variants in the same gene for 225 families. **E)** The simulation shown in 'B' is repeated using the lower bound of the 95% confidence interval for the rate of nonsense/splice site brain-expressed *de novo* variants in probands (0.06×10^{-8}). The threshold of significance ($q<0.05$) remains at ≥2 *de novo* variants in the same gene for all samples sizes shown. **F)** The simulation described in 'E' was used to predict a conservative estimate for the number of genes conferring ASD risk that will be identified by the observation of ≥2 independent nonsense/splice site *de novo* mutations for a sample of a given size (specified on the x-axis). Predictions are given for the specified models of locus heterogeneity.

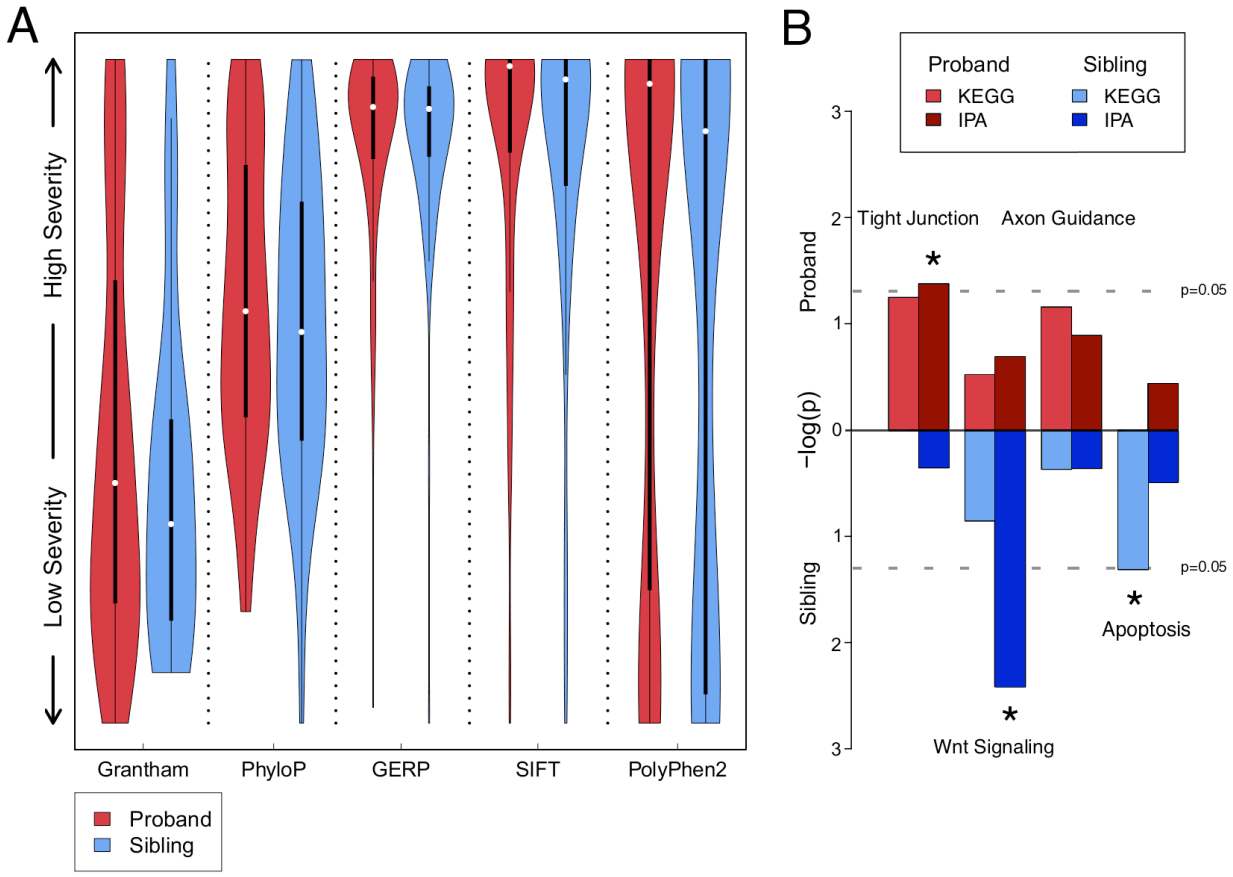

Figure S9: Metrics of functional severity and pathway analysis fail to differentiate risk-associated from neutral *de novo* variants.

A) The distribution of various metrics of functional severity for brain-expressed non-synonymous *de novo* variants, as predicted by the specific bioinformatics tool, is shown by violin plots of 200 probands and their unaffected siblings. The median is indicated by the white dot and interquartile range by the thick black bars; the colored area shows the kernel distribution of the data. All metrics have been rescaled to allow side-by-side comparison; no significant differences in scores are seen between probands and siblings. B) Gene ontology analysis of brain-expressed non-synonymous *de novo* variants in 200 probands and their unaffected siblings using KEGG and IPA tools. Equal numbers of pathways are enriched in probands (red) and siblings (blue). The p-values shown are uncorrected for multiple comparisons and none survive correction.

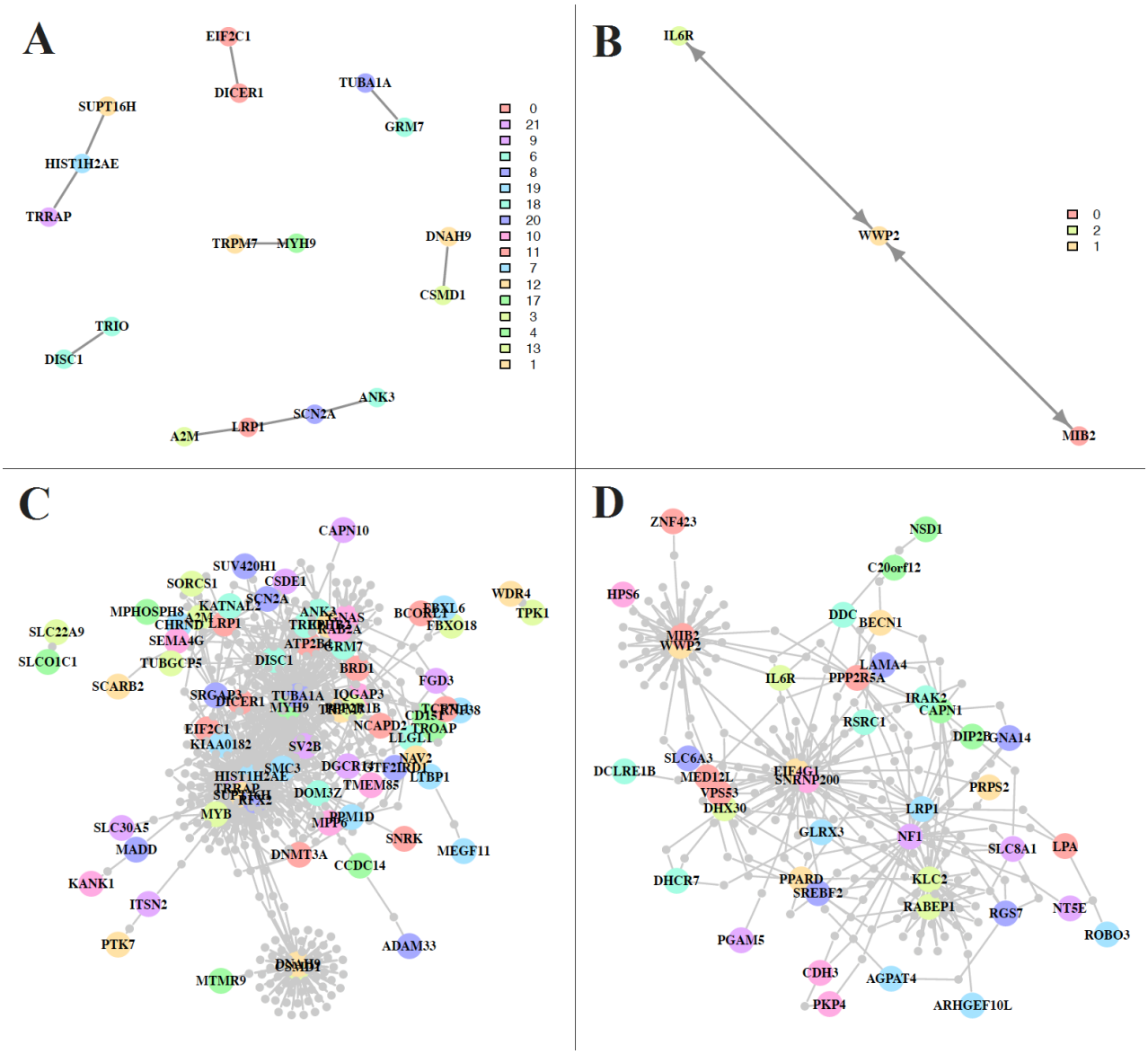

Figure S10. Disease Associated Protein-Protein Interaction for genes with brain-expressed non-synonymous *de novo* SNVs.

A) Direct interactions in probands showing 10 direct protein-protein interactions out of the 114 genes with brain-expressed non-synonymous *de novo* variants that were submitted for analysis. This degree of connectivity is not greater than expected (p=0.77). B) Direct network for siblings showing two direct protein-protein interactions out of the 67 genes submitted for analysis. This is also not more than expected by chance (p=0.12). C) Indirect interactions (i.e. protein-protein interactions though intermediate genes not on the submitted list) between the 114 proband genes. This level of connectivity is not greater than expected by chance (p=0.34). D) Indirect interactions between the 67 sibling genes; again this is not different from expectation (p=0.70).

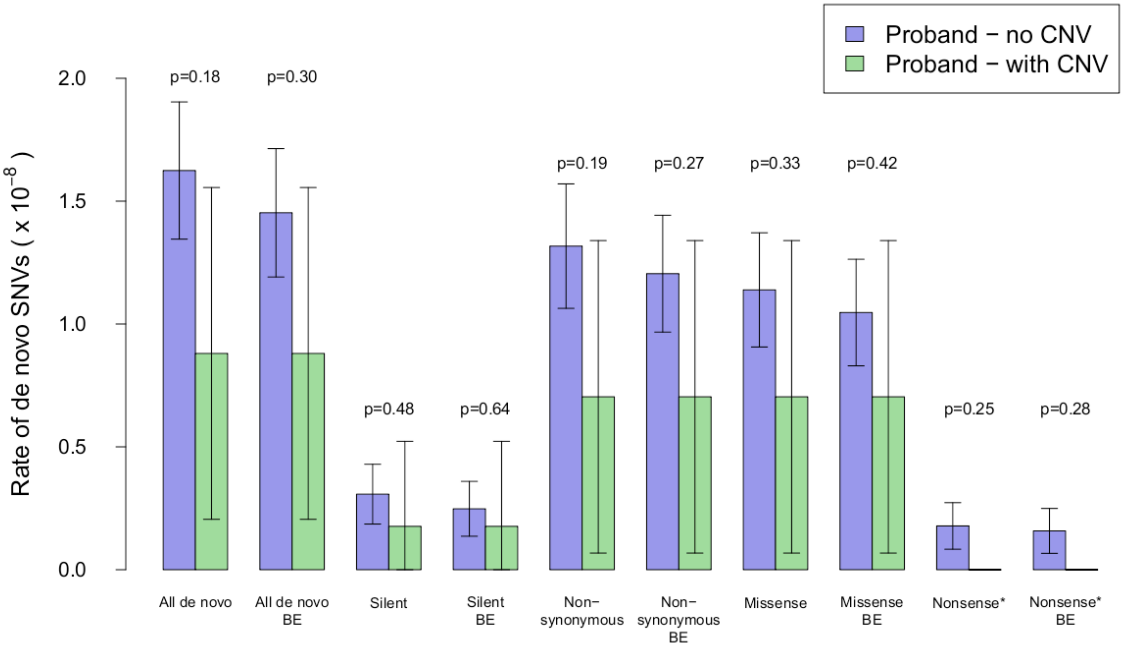

Figure S11. *De novo* rate of SNVs in samples with and without large multigenic CNVs (data from this study only).

The rate of *de novo* SNVs is shown in 15 probands with large multigenic CNVs (≥16 RefSeq genes) and 179 probands without such CNVs (demonstrated previously by genotype).[6] In the probands with large multigenic CNVs a trend toward less non-synonymous *de novo* SNVs are seen compared to probands without. Error bars represent the 95% confidence interval and the p-values shown are calculated by comparing the rate of *de novo* SNVs in all samples using a two-tailed Wilcoxon test.

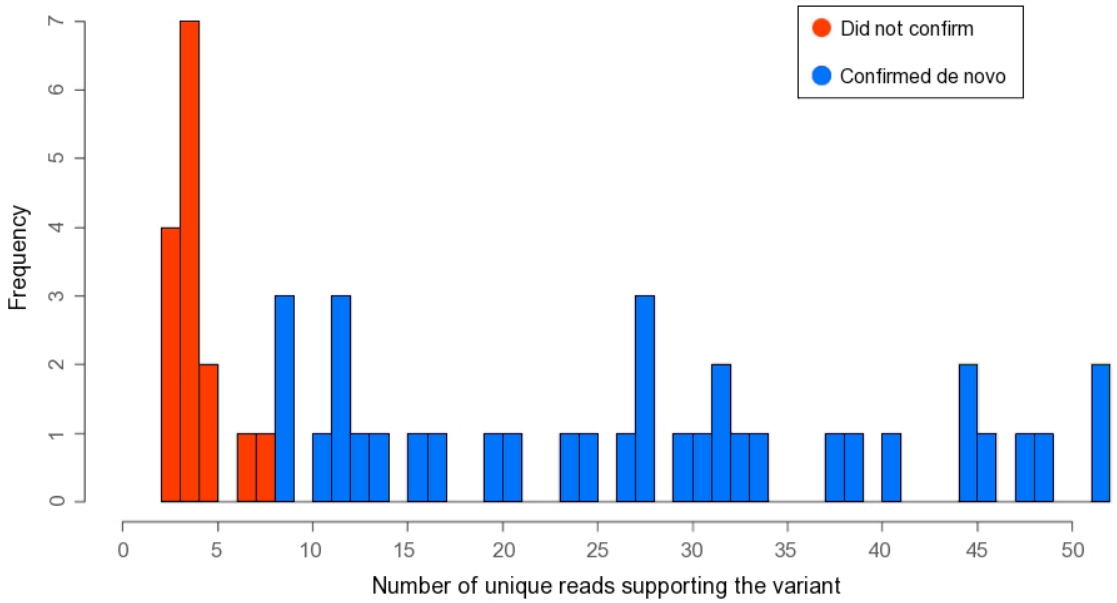

Figure S12. Validation of *de novo* predictions with Sanger sequencing.

A total of 50 *de novo* variant predictions were validated blind to affected status using PCR and Sanger sequencing from blood-derived DNA. There were 15 *de novo* variants that had less than 8 unique reads supporting the variant, and all of these failed to confirm due to false positives in the proband. Of the 35 variants that had at least 8 unique supporting reads all 35 confirmed as true *de novo* events.

2. Supplementary Methods

Sample selection

928 samples from 238 families of the Simons Simplex Collection (SSC)[7] were selected for whole-exome analysis; 214 families were quartets with an affected proband, unaffected sibling and unaffected father and mother; the other 24 were trios without the unaffected sibling. The samples were selected on the basis of:

- Most severely affected female probands in SSC (based on low NVIQ) with highest degree of discordance with siblings (based on SRS) (34 quartets, 12 trios)
- Most severely affected male probands in SSC (based on low NVIQ) with highest degree of discordance with siblings (based on SRS) (40 quartets)
- Random selection from the SSC (102 quartets, 3 trios);
- Known rare sequence variants in contactin genes (4 trios);
- Multiple unaffected siblings (27 quartets, 1 trio);
- Known large multigenic copy number variants (CNVs) (11 quartets; 4 trios)
    - *de novo* deletions in 16p11.2 (6 quartets);
    - *de novo* 16p11.2 duplications (3 quartets, 1 trio);
    - inherited 16p11.2 duplication (1 trio);
    - *de novo* 7q11.23 duplication (1 quartet, 2 trios);
    - *de novo* 17q12 deletion (1 quartet);

A complete list of samples with the reason they were selected is given in Supplementary_data_S1.

Capture and sequence

The samples were enriched for exonic DNA using two versions of NimbleGen hybridization arrays: custom whole-exome array (35Mbp target, 210 samples); and SeqCap EZ Exome v2 (45Mbp target, 718 samples). Cell-line derived genomic DNA was used for the first 12 samples (4 trios with known rare contactin variants), followed by whole-blood-derived genomic DNA for the remaining 916 samples. The exon enriched DNA was sequenced using the Illumina GAIIx (592 samples) or the Illumina HiSeq 2000 (336 samples). Sequencing data were generated using: single lane of 74bp paired-end reads (657 samples); two samples per HiSeq lane with barcoding and 74bp paired-end reads (224 samples); single lane of 99bp paired-end reads (11 samples); or 1-4 lanes of 74bp single-end reads (36 samples). A complete list of samples with the conditions under which they were run is given in Supplementary_data_S1.

Variant detection overview

Sequencing data were run through Illumina's Cassava pipeline and then aligned to the entire human genome reference (hg18/NCBI 36) using the Burrows-Wheeler Aligner (BWA)[8]. Reads that did not align, or aligned outside of the target region, were discarded. Duplicate reads were filtered out using SAMtools,[9] which was also used to identify single nucleotide variants (SNVs). All SNVs were analyzed using SysCall[10] to remove systematic errors in Illumina data.

Alignment and SAMtools conversion

Rescaled FASTQ format data were aligned to unmasked human genome build 18 (NCBI 36) using the Burrows-Wheeler Aligner (BWA) with the default settings using the following command: bwa aln -t 8 'BWA_reference' 'Fastq_input' > 'Output.sai'.

Aligned reads were converted to SAMtools format using the following commands: Single-end: bwa samse 'BWA_reference' 'Output.sai' 'Fastq_input' > 'Output.sam'; Paired-end: bwa sampe 'BWA_reference' 'Output_pair1.sai' 'Output_pair2.sai' 'Fastq_input_pair1' 'Fastq_input_pair2' > 'Output.sam'.

If different family members had single-end and paired-end data, then the single-end settings were used for all family members.

Trimming to target

The aligned reads were trimmed to the exome target using an in-house script. If any read overlapped at least 1bp of a probe on the NimbleGen array, then the read was considered 'on-target'. All reads that did not meet this definition, including unaligned reads, were discarded. If members of the same family had been analyzed using different targets then trimming was performed using the consensus of the 35Mb custom array and the 45Mb EZ v2 array. This consensus target included 32,159,763 nucleotides.

Duplicate removal and pileup conversion

The trimmed aligned data were converted to a sorted binary format (BAM), and then duplicates were removed using SAMtools on the default settings. The following commands were used: samtools view -bSt 'SAM_reference' 'Input.sam' | samtools sort – 'Output.bam', followed by: samtools rmdup -u 'Input.bam' - | samtools view - -o 'Output.sam'.

The aligned, trimmed, and duplicate-free SAM file was then converted to pileup format using SAMtools with the default settings: samtools pileup -cAf 'Reference' -t 'SAM_reference' 'Input.sam' > 'Output.pileup'.

Quality control

Thirteen families and one sibling were excluded due to poor quality data. Exome enrichment failed on one proband (12418.p1) and one father (11031.fa); both families were removed. Four families processed using single-end reads had very low coverage after removing duplicate reads (11028, 11135, 11334, 12219). In two families mismatch was seen between exome data and blood DNA showing that the proband and sibling had been mislabeled within the family. We will return to the original DNA stocks to resolve this issue, but both families were excluded from analysis (11347, 11372). Finally excessive *de novo* predictions (over 1,000 compared with less than 10 in all other offspring) were seen in both probands and siblings in five families (11714, 11998, 11999,14019,14025) and the sibling only in one family (11630.s1). All of these offspring had been processed on the same three flowcells; while the error rate and coverage appeared normal we believe some subtle recurrent error resulted in excess noise. Family 11630 was converted to a trio family as a result. Following quality control there were 225 families (200 quartets and 25 trios).

Variant detection and data cleaning

Variants from the reference genome were filtered from the pileup file using the SAMtools variant filter script: samtools.pl varFilter -d 4 -D 10000000 'Input.pileup' > 'Output.var'.

To remove systematic errors from the Illumina data, the variants were assessed using SysCall.[10] This algorithm was trained using discrepant variants detected in overlapping paired-end data and uses a combination of read-direction, quality scores, and surrounding sequence to identify systematic errors. The algorithm was used with default settings: SysCall.pl 'Input.var' 'Input.sam' 'Output' 'Path'. All genome positions that were present in the 'Error' file in one or more samples were removed from the dataset.

Consistency within quartets

Of the 200 quartets that passed quality control 167 (84%) were analyzed using the same target array, and 197 (99%) were analyzed using the same sequencing instrument. 165 (83%) were analyzed with the same target array and the same sequencing instrument. Of the 165 families analyzed with the same target array and sequencing instrument, there were 76 quartets (38%) in which the probands and sibling were run concurrently (for both capture analysis and on the same flowcell for sequencing).

One of the major findings in this paper was that there is a significant increase in the rate of *de novo* missense/nonsense/splice site (non-synonymous) mutations in probands compared to their unaffected matched siblings. The key question is whether batch effects could be responsible for this difference rather than a true biological effect. If batch effects were responsible we would expect to see the greatest difference in *de novo* rate in the samples run under different conditions, while there should be less of a difference in the 38% of families in which samples were run concurrently with the same technology. In fact the reverse is seen supporting that the signal seen reflects the underlying biology (FigS1).

Separating the probands into three groups: 1) Run concurrently with same technology (38% of quartets); 2) Not run concurrently, but with same technology (44.5%); and 3) Not run concurrently and using different technology (capture array version and/or sequencing instrument, 17.5%), the greatest difference in *de novo* rate between probands and siblings is seen for group 1 – the samples run under the most consistent conditions (FigS1). The result in group 2 is very similar to that in group 1, while the least dramatic difference is seen in group 3 – those run using differing technologies and not run concurrently (FigS1). If there is any underlying bias in detection accuracy due to non-concurrent analysis it favors the detection of *de novo* events in siblings rather than probands.

The less dramatic results in the third group (those run separately with different technology) could be a result of improvements in technology leading to enhanced detection accuracy in siblings (who were run after the probands). Alternatively, it may reflect the presence of multigenic CNVs in 13 out of the 34 (38%) probands in this group since a lower rate of *de novo* point mutations was detected in samples with such CNVs.

De novo rate in non-coding and silent regions

To further demonstrate that the increased rate of non-synonymous *de novo* events was a true biological signal, the rate of non-coding (variants that are predicted to be in UTRs, introns, intergenic, and non-coding genes) and silent *de novo* variants was calculated in probands and siblings. The comparison of confirmed silent variants in probands and siblings is shown in the

main manuscript Fig1A. Non-coding variants were not confirmed by PCR; however, from the coding regions we demonstrated a 96% confirmation rate, and non-coding regions were treated in a similar manner to coding regions. This is the only analysis in the entire manuscript or supplements that is based on unconfirmed *de novo* results.

FigS2 shows the rate of non-coding, silent, and non-coding plus silent *de novo* variants; non-synonymous *de novo* variants are included for comparison. There is no difference between probands and siblings for any of these categories.

Normalization

To allow for a fair comparison between probands and siblings in quartets, only those bases with 20x diploid coverage of unique reads in all four family members were considered. A highly conservative definition of unique reads was used – counting only the number of positive strand starting positions, thereby allowing a maximum count of 74 reads per sequence base (due to the 74bp length reads). This normalization procedure gave extremely consistent numbers of variants in probands vs. siblings across all variant frequencies (FigS3-4).

The category most sensitive to small variations in detection accuracy (and therefore most informative for *de novo* detection) is novel heterozygous variants. These are variants that are observed only once in a parent in the entire dataset and never seen in dbSNP or 1,289 control whole-exome samples. Considering only the variants in coding and splice sites detected using the exact settings used for *de novo* variant detection (except that variants present in parents were not excluded) there are 22,996 novel variants in probands vs. 22,992 novel variants in siblings (main manuscript – Table1), a difference of just 4 variants representing 0.02% (4/22,994) of novel variants. This strongly supports the conclusion that we are detecting variants equally in probands and siblings.

Changes in normalization settings

To demonstrate the effectiveness of the normalization strategy and the appropriateness of the threshold chosen (20 unique reads per base in all four family members) the difference in the number of variants detected in probands and siblings is represented as the threshold is varied from 0 to 24 (FigS3). The difference in variant count was calculated for each family separately and expressed as a percentage of the average number of variants detected in all samples (to allow comparability as the counts decrease with the rising threshold). At a threshold of 20 the mean difference between probands and siblings is just 0.009% of all novel variants with a slight bias towards siblings. These variants were detected using the exact settings used for *de novo* variant detection except that variants present in parents were not excluded.

To give a context to the scale of difference seen in the rate of *de novo* mutations in probands and siblings compared with the detection of novel variants, FigS4 shows the same plot with the data for *de novo* variants included at the end. At a minimum number of unique reads for family member of 20 (the threshold used for *de novo* detection) the mean difference for *de novo* variants is 0.58% in favor of probands, a 66-fold higher value than seen with novel variants.

Transmission Disequilibrium

The detection accuracy can also be assessed by considering novel variants in the parents (defined as variants in which only one allele was seen throughout all 400 parents and never seen in dbSNPv132 or in 1,289 whole-exome controls analyzed locally, by definition such a variant would be heterozygous on an autosomal chromosome).

By considering only novel silent (presumed neutral) autosomal variants detected using the same thresholds as *de novo* variant detection (except for excluding variants with evidence of a variant in a parent) and seen in either the proband or sibling (but not both), we can test the data against the assumption that 50% of such variants should be present in probands and 50% in siblings. The data show a result of 49.6:50.4 for probands:siblings (13,503 variants vs. 13,395 variants). This does not differ from expectation (chi-square = 0.42, p=0.51); moreover, any resulting bias that is present would actually favor detection of variants in the sibling.

Variant detection sensitivity

Considering the novel autosomal silent variants present in parents (as described in the previous section) also allows an estimate of the sensitivity of variant detection within the regions analyzed. By testing the assumption that 50% of such variants present in the parent should be present in either child, we can estimate how many of these variants are missed in the children. There are 14,709 variants seen in the parents, therefore 29,418 transmission events with 50% chance are to be considered. The children have 14,413 variants (representing 49.0% of transmissions); though the difference is small, it is different from the expected rate (p=0.0003, binomial distribution). Based on these results, the sensitivity of detection of a variant in a child is 98.0% (97.5% in probands and 98.4% in siblings, though as described in the previous section the difference between probands and siblings is not significant, p=0.09). An estimated sensitivity of 98% is consistent with the observation of two *de novo* predictions that Sanger sequencing revealed to be inherited events (2/304 variants = 0.7%).

This calculation is based on regions in which a parental variant was detected; the sensitivity of variant detection in a proband or sibling base, given that a variant has previously been detected there (in a parent), is likely to be slightly higher than for a base in which no variant has been detected. However, the rate of *de novo* variants observed is highly consistent with previous estimates (FigS5) leading us to believe that our sensitivity for *de novo* variants in the bases assessed is high.

Defining unique variants

The model of expected erroneous *de novo* predictions assumes that every read is an independent observation. This is not the case in sequencing data since PCR duplicates introduced during the amplification steps of the Illumina sequencing protocol lead to multiple identical reads which may all contain a specific variant.

All data were run through the SAMtools duplicate removal pipeline, however PCR duplicates may remain due to random sequencing errors. To ensure that each read was a truly independent observation, the number of positive-strand read starting positions supporting the variant and reference were calculated for all variants predictions. This scale would give a maximal value of 74 unique observations for 74bp paired-end data. The number of unique reads in all family members was used to normalize the data (above).

Blinding and randomization

When more than one sample was run per flowcell lane on HiSeq2000 machines, care was taken to randomize barcodes assigned to probands and siblings. Furthermore, probands and siblings were run concurrently on the same lane.

Throughout the entire alignment and variant prediction pipeline, probands and siblings were treated in an equal and unbiased manner. Determination of variants for confirmation was by

preset thresholds determined through theoretical calculations (see section "5. Supplementary Equations") and confirmation experiments (FigS12).

All confirmation by Sanger PCR was performed in all family members (including the both children in quartets); interpretation of chromatograms and determination of variant status was performed blinded to affected status. The overall confirmation rate was 96% (95% in probands and 96% in siblings).

De novo variant detection

Sequence data for each variant in a proband or sibling were compared with all other family members at the same position. A variant was predicted to be *de novo* if it was not predicted in either parent (single parent for chrX and chrY in male offspring); there were at least 20 unique reads in all family members, at least 8 unique reads supporting the variant in the offspring (supplementary methods, FigS12), at least 90% of reads supporting the reference in both parents (single parent for chrX and chrY in male offspring), and a mean PHRED-like quality score of at least 15 for reads supporting the variant. These thresholds were determined through calculations of the chance of seeing recurrent error in sequencing data (see equations section at the end of the SOM) and confirmed experimentally (FigS12).

All such predictions were validated experimentally using PCR to amplify the region from whole-blood derived DNA in all family members and Sanger dideoxynucleotide sequencing to confirm the variant was present in the offspring only.

De novo confirmations in cell-line DNA

The whole-exome data for the initial 4 trios were generated from transformed lymphoblast cell-line derived DNA. The *de novo* predictions made were confirmed using Sanger sequencing initially in DNA from the same source where all 7 variants were present in the proband and absent in the parent. When confirmation with Sanger sequencing was repeated using whole-blood derived DNA, only 3 of the variants were present in the proband and none were present in parents. The 4 unconfirmed variants were subsequently excluded.

De novo confirmations

The *de novo* variant prediction pipeline yielded 297 potentially *de novo* variants in 200 quartets (162 in probands, 135 in siblings) with an additional 18 potential *de novo* variants in 25 trios (18 in probands) to give a total of 315 variants.

Oligonucleotides were designed around the variant in question using Primer3. The oligonucleotides were synthesized by Integrated DNA Technologies (IDT, http://www.idtdna.com) and the region was amplified using standard PCR in both parents and children. The amplified DNA was analyzed using Sanger dideoxynucleotide sequencing at Keck, the Yale core facility (http://medicine.yale.edu/keck). For each amplicon, sequencing was performed with both forward and reverse oligos.

The chromatograms were analyzed in all 4 family members using Sequencher and classified as: 'Confirmed *de novo*,' 'Inherited,' 'No variant,' or 'Inconclusive.' For inconclusive results new sets of primers were designed and synthesized, then the region was reassessed with PCR and Sanger sequencing. Recurrent inconclusive results were reassessed through four iterations only. If a definitive result was not obtained at this stage, no further attempts at confirmation were made and the variant was removed from the analysis.

Of the 304 variants with definitive results, 291 (96%) were confirmed as true *de novo* events. The remaining 13 variants were either not detected in the child (11 variants, 7 in probands and 4 in siblings) or present in a parent (2 variants in 1 proband and 1 sibling).

False positive rate for *de novo* detection

The successful confirmation of 291 out of 304 variants allows estimation of the specificity for *de novo* variant detection. Since the exomes of 425 individuals (200 probands and 200 siblings in quartets plus 25 probands in trios) were examined with an average of 48,561,202 nucleotides analyzed per sample, a total of 20,589,949,648 nucleotides were analyzed. There were 13 false positives, giving an estimate of the false positive rate of 6.3×10^{-10} per base.

Gene conversion by mismatch repair

Sample 11382.s1 was found to have 3 *de novo* variants (1 nonsense, 2 silent) in the same exon of *KANK1* but separated by over 200bp (so that they would appear on different reads). Sanger sequencing confirmed all 3 variants. The observation of 3 independent events in such close proximity is extremely improbable and we feel this likely represents an example of gene conversion by mismatch repair. Accordingly we counted this as a single *de novo* event in each category that it was examined. Since there is only one non-synonymous event, the method of counting has no effect on the significant results shown in the main manuscript.

Variant frequency

The population frequency of predicted variants was predicted by comparison with three datasets: dbSNPv132, 1,289 whole-exomes from a Swedish blood-pressure study, and 400 whole-exomes from parents in this study; both whole-exome data sets were run locally on the same sequencing machines and under the same conditions as the data presented. The allele frequency was determined as the higher of the estimates from the Swedish controls and parents (not counting single incidence alleles in the parents). A variant was considered 'novel' if it was not present in dbSNPv132, not present in 1,289 control exomes, and not seen on more than one allele in all the parents. A variant that did not meet these criteria, but that was seen at <1% population frequency was considered rare and all variants with a population frequency of ≥1% were considered common.

Gene annotation

Variants were annotated against the RefSeq gene definitions to determine the effect on the resulting amino acid sequence. Where multiple isoforms were present, the most-deleterious interpretation was selected.

Canonical splice site

The phrase 'splice site' used throughout the manuscript and supplements refers to the 2bp donor and acceptor canonical splice site found on either side of 98.5% of exons in the human genome. These are among the most highly conserved base pairs in the human genome and have the potential to cause highly disruptive events. For this reason splice site variants are considered in the same category as nonsense variants and frameshift indels. When a variant is labeled as 'splice site' this means that the reference sequence (hg18) shows the expected two base pair donor or acceptor sequence and that this is altered by the presence of a variant (e.g. a deletion

that resulted in the same two base pairs adjacent at the end of an exon would not be labeled as a splice site variant).

Brain-expressed genes

The list of brain-expressed genes was obtained from a recent study of the human brain transcriptome throughout development and adulthood.[11] Their 1,340-sample dataset was generated by dissecting regions from 57 clinically unremarkable postmortem brains of donors ranging from 6 post conceptual weeks to 82 years, which were divided into 15 periods based on age. The expression levels of 17,565 protein-coding genes within each sample were assayed using the Affymetrix GeneChip Human Exon 1.0 ST Array platform. A "brain-expressed" gene was defined as having 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. Using these criteria, 15,132 of 17,565 genes were expressed in at least one brain region during at least one period.

Using this list of 15,132 genes, the 18,933 RefSeq genes in hg18 were defined as being 'brain-expressed' (14,363), not brain-expressed (1,833), or unknown (2,737). Genes not within the RefSeq gene list accounted for the difference between 15,132 and 14,363. In text and figures where the results are described as being 'brain-expressed' this means they were on the list of 14,363 genes rather than being unknown or not brain-expressed.

Synaptic genes

Genes were defined as being 'synaptic' if they were implicated in proteomic analysis completed in 3 prior publications.[12-14]

Insertion and deletion detection

Small insertions and deletions (<40bp) can be detected in short-read sequencing data using gapped aligners such as BWA. While we were able to successfully confirm 3 such *de novo* events in the first 51 probands analyzed (Supplementary_data_S2), we were unable to get a confirmation rate above 10%. Such a low confirmation rate could amplify any small biases present in the underlying data and would risk generating a non-interpretable result between matched probands and siblings.

While the potential for such events to disrupt protein function is plain, especially if they alter the reading frame, we elected not to pursue these events further in the present study. We hope to further refine detection methods and reevaluate this decision in the future.

No indel events are included in the analyses presented in this paper except in consideration of the combination of *de novo* events from probands only across this study and O'Roak et al..

Five scores commonly used as measures of functional severity were calculated to try and distinguish pathogenic *de novo* variants from neutral, non-risk associated variants. PhyloP[15] and GERP[16] are based on measuring conservation between species; Grantham Scores[17] assess the chemical differences between amino acids; PolyPhen2[18] and SIFT[19] use a range of sequence and structural features to estimate variant severity. PhyloP scores were downloaded from UCSC Genome Browser[20] and annotated using an in-house script. GERP scores were calculated using SeattleSeq Annotation on default settings (http://gvs.gs.washington.edu/SeattleSeqAnnotation/index.jsp). Grantham scores were calculated directly based on the table in the stated reference. Polyphen2 results were obtained through batch query web interface on default hg18 settings (http://genetics.bwh.harvard.edu/pph2/). SIFT

results were obtained through batch submission on default hg18 settings to the web interface (http://sift.jcvi.org). Missense variants that could not be annotated were converted to hg19 and reannotated using Polyphen2 or SIFT.

In view of their predicted deleterious nature nonsense and splice site variants were given the highest possible scores for Grantham (215), SIFT (0) and PolyPhen2 (1). For GERP and PhyloP, when analyzing nonsense and splice site variants, every possible coding base for the specific protein was scored and the highest value selected.

To present the results together in FigS9B the results were rescaled to range from 0 to 1. Unlike the other metrics, SIFT assigns a low score to severe variants; these scores were inverted to allow easy comparison.

Multivariate severity score analysis.

Multivariate analyses were restricted to genes with brain-expressed *de novo* variants. Genes were categorized as "case" and "control" based on the following rule: proband genes with missense, nonsense, or splice variants were labeled as case; sibling genes and proband genes with silent variants were labeled as control. Using the 5 severity scores (see previous section) as response variables, MANOVA analyses did not detect a significant association with case/control status. From these analyses we concluded that the fraction of genes designated as cases included a mixture of damaging and non-damaging mutations from which a clear signal could not be extracted.

Chance of observing a *de novo* event

Since genes differ in coding size and GC content, the chance of a *de novo* variant occurring by chance varies across genes. To allow this difference to be taken into account when considering the distribution of *de novo* variants, the chance of a *de novo* variant occurring by chance in each gene was calculated.

All RefSeq genes were analyzed to determine the size of the coding region and splice site along with GC content. The size of each gene was compared to the total RefSeq coding and splice site region (33,102,852bp) to obtain a probability. This probability was modified in light of GC content using previously published estimates of the ratio of *de novo* variation by base in humans (GC bases mutate at a 1.76-fold greater rate than AT bases).[5] The resulting probability (expressed as a fraction of 1 for that specific gene so that the combined probability across all genes is 1) gave the likelihood of a *de novo* variant hitting each specific gene by chance. Size accounted for the vast majority of risk compared to GC content.

Parental age and *de novo* burden covariate analysis

To determine if the rate of *de novo* events differed between probands and their siblings, we used Poisson regression with proband status as a predictor and the count of relevant *de novo* events as the outcome. Because it is well known that *de novo* SNV rates increase with paternal age,[21] the model also included paternal age as a predictor. Maternal age is also potentially related to *de novo* events, but it is highly correlated in our data with paternal age ($r=0.69$, p-value < 0.0001). To include independent information into the model we included, as predictors, paternal age and the difference between paternal and maternal ages. In essence the model is an analysis of covariance using Poisson regression. In addition, to account for the relatedness of proband and sibling, we fit the data using GEE methods.[22]

When all *de novo* events are fitted against paternal age and proband status, the effect of paternal age on *de novo* rate is significant (estimated slope $b_A = 0.033 \pm 0.012$ [standard error] p = 0.008). To express this effect in numerical terms, it represents a 1.39-fold increase in *de novo* events per increased decade of paternal age. The difference in parental ages is not quite significant as judged by two-sided criteria, but the estimated slope is negative and thus consistent with expectation ($b_{DA} = -0.030 \pm 0.016$, p = 0.061). Proband status does not have a significant effect on *de novo* rate (estimated increase in rate for probands $b_P = 0.174 \pm 0.125$, p = 0.165). One might choose to use a one-sided test for these hypotheses because the signs of the effects can be assigned *a priori*; only the interpretation of the difference in parental ages would be altered. All p-values reported herein are two-sided.

When only non-synonymous *de novo* events are fitted against paternal age, difference in parental ages, and proband status, the effects of parental ages are no longer significant ($b_A = 0.020 \pm 0.014$, p = 0.161; $b_{DA} = -0.015 \pm 0.018$, p = 0.404), but the effect of proband status is significant ($b_P = 0.341 \pm 0.144$, p = 0.018). Thus b_P suggests that probands will carry 1.41 times more *de novo* non-synonymous events than their siblings. Likewise, for the subset of *de novo* events hitting brain-expressed genes, parental age is not significant ($b_A = 0.009 \pm 0.007$, p = 0.227; $b_{DA} = -0.011 \pm 0.009$, p = 0.224), but the effect of proband status is ($b_P = 0.175 \pm 0.060$, p = 0.004). From this model, b_P suggests that probands will carry 1.19 times more *de novo* events hitting brain-expressed genes than their siblings. Finally, for the subset of non-synonymous *de novo* variants falling in brain-expressed genes, the effect of proband status is highly significant ($b_P = 0.332 \pm 0.085$, $p = 4.4 \times 10^{-5}$); ages have no predictive value and the slopes make little sense ($b_A = -0.005 \pm 0.011$, p = 0.627; $b_{DA} = 0.004 \pm 0.012$, p = 0.747). In this instance b_P predicts that probands will carry 1.39 times more non-synonymous *de novo* events falling in brain-expressed genes than their siblings.

Because the significant effect of parental age on the count of *de novo* events disappears for subsets of genes in which proband status is a critical predictor, we wondered how well it would predict the counts of synonymous events. In this analysis, both paternal age and the difference between paternal and maternal ages are significant predictors of count of *de novo* events ($b_A = 0.077 \pm 0.026$, p = 0.003; $b_{DA} = -0.076 \pm 0.032$, p = 0.016) whereas proband status was not ($b_P = -0.350 \pm 0.245$, p-value p = 0.154). The slopes $b_A = 0.077$ and $b_{DA} = -0.076$ translate into similar effects on rates by paternal and maternal age on synonymous events, equivalent to a 2.15-fold increase in rate per increased decade of parental age.

When nonsense and splice site variants were considered independently, proband status is a critical predictor of *de novo* count (p=0.02) while both paternal age and the difference between paternal and maternal ages are not significant predictors of count of *de novo* events. A similar pattern is seen for brain-expressed nonsense and splice site variants with probands status remaining significant (p=0.04) while parental age is not. These results should be treated with caution due to the low number of observations; however, they are in accordance with results of the binomial test used in the main manuscript.

These results demonstrate several important points. Both paternal and maternal ages potentially contribute to the observed *de novo* rate, and this contribution is greatest towards the rate of synonymous variants. By contrast, the greatest effect of proband status is on the rate of non-synonymous *de novo* events falling in brain-expressed genes. Still, probands have a significantly higher rate of *de novo* events than siblings for 3 subsets: non-synonymous *de novo* variants; *de novo* variants hitting genes that are brain-expressed; and the intersection of those sets. Moreover, this difference in *de novo* rates between probands and siblings stands even after

accounting for parental age, which for functional variation explains only a small portion of the variance.

IQ and *de novo* burden covariate analysis

To determine if there is a relationship between the number of *de novo* events in probands and their measured IQ, we used linear regression with count of relevant *de novo* events as the predictor and one of three measures of IQ as the outcome variable (i.e., full-scale IQ, non-verbal IQ and verbal IQ). An alternative approach would be to treat the IQ measure as a predictor of the count of *de novo* events in a Poisson regression. Either modeling approach yielded similar results, although on a different scale. Moreover, for either modeling approach and for any measure of IQ, there was never a significant relationship between IQ and count of *de novo* events. The slope of the linear model was negative, as would be expected, when IQ was predicted as function of count of *de novo* events, and the tightest relationship was between verbal IQ and total *de novo* count ($b_A = -3.18 \pm 2.451$, $p = 0.196$). Analyzing the subset of non-synonymous *de novo* events does not improve the fit, although again the tightest relationship is with verbal IQ ($b_A = -1.99 \pm 2.743$, $p = 0.469$). If only the subset of *de novo* events falling in brain-expressed genes is analyzed, the relationship weakens further and the sign of the slope is not always consistent with expectation.

Sex and *de novo* burden covariate analysis

Given the observation of a higher rate of *de novo* CNVs in female probands compared with male probands,[6,23] and the lower prevalence of ASD in females,[24] we expected to see a higher rate of *de novo* SNVs in female probands than male probands. To determine if there is a relationship between the number of *de novo* events in probands and sex, we used Poisson regression accounting for paternal age, paternal age minus maternal age, and sex. We fit three outcomes, the count of all *de novo* events, the count of non-synonymous events, and the count of synonymous events. The sexes did not differ significantly for any of these outcomes (all $p > 0.12$) (FigS6).

Recurrent *de novo* CNVs

Since specific recurrent *de novo* CNVs have been strongly associated with ASD,[6,25] specifically: 16p11.2, 7q11.23, 22q11.2, 15q11-13.2, and *NRXN1*, we aimed to identify whether *de novo* point mutations would cluster within specific genes to help characterize the risk associated with these CNVs. Two such *de novo* SNVs were identified, both were missense variants in probands in the genes *GTF2IRD1* in 7q11.23 and *DGCR14* in 22q11.2. The observation of two *de novo* variants within these 143 genes is not greater than would be expected by chance when taking gene size and GC content into account ($p=0.16$, binomial). No further *de novo* variants were detected within these genes when considering the dataset from O'Roak et al..

Estimation of percentage of *bona fide* risk-associated variants

Since 154 *de novo* SNVs were detected in probands and the 125 were detected in siblings (Table1) the percentage of *de novo* variants in probands associated with ASD risk is (154-125)/154 = 19%. A similar calculation with non-synonymous *de novo* variants gives an estimate of (125-87)/125 = 30% and in brain-expressed non-synonymous *de novo* variants this value is (114-67)/114 = 41%. Finally in brain-expressed nonsense and splice site *de novo* variants the estimate is (13-3)/13 = 77%. To obtain 95% confidence intervals for each of these ratios we

simulated Poisson counts for probands and siblings. We then obtained empirical confidence intervals based on the distribution of the resulting ratios. Empirical values matched asymptotic-based estimates quite well. For each scenario we used the observed *de novo* counts in probands and siblings to estimate of the expected rates.

Estimation of percentage of individuals with *bona fide* risk-associated variants

Since 87 probands have at least one non-synonymous *de novo* variant in a brain-expressed gene compared to only 60 siblings we estimate that at least 14% ((87 – 60) / 200) of probands carry a risk-associated *de novo* SNV.

Estimation of the number of genes contributing ASD risk

To estimate the number of ASD risk loci from brain-expressed non-synonymous *de novo* SNVs we used the "unseen species problem". This approach uses the observed frequency and number of risk-associated genes (or species) to infer the total number of risk-associated genes in the population, including those yet to be observed. To estimate the number of risk-associated genes we assumed that 47 proband brain-expressed non-synonymous *de novo* SNVs carried risk (114 in proband – 67 in sibling) and that the 2 genes with recurrent *de novo* SNVs (*SCN2A* and *SUV420H1*) represent risk-associated events. Therefore in this sample there are 47 risk-associated events, 45 risk-associated genes, and 43 single occurrence risk-associated genes (though we do not know which 43). We then apply a formula for calculating the number of species (C):

- $C = c/u + g^2*d*(1-u)/u$

In which: c = the total number of distinct species observed (45); c_1 = the number of singleton species (43); d = total number of CNVs observed (47); g = the coefficient of variation of the fractions of CNVs of each type, and $u = 1 – c_1/d$.[26] In the calculations presented in this manuscript we assume that g equals 1 due to the small number of observations. Applying these figures the total number of risk-associated genes (C) is 1,034.

To estimate the confidence interval of this observation we used the upper and lower confidence intervals for the difference in the rate of brain-expressed non-synonymous *de novo* SNVs in probands and siblings to estimate the number of risk-associated *de novo* SNVs in this sample:

- (Upper 95% difference in rate) * 47,663,807 mean number of RefSeq bases assessed * 200 samples = $(0.76 \times 10^{-8}) * 47{,}663{,}807 * 200 = 73$
- (Lower 95% difference in rate) * 47,663,807 mean number of RefSeq bases assessed * 200 samples = $(0.18 \times 10^{-8}) * 47{,}663{,}807 * 200 = 17$

Using the same logic to identify the number of genes with recurrent and single occurrence *de novo* SNVs, an approximate 95% confidence interval is 119 to 2,555 genes.

Simulation model

Consideration of the size and GC content of a gene is essential when estimating the likelihood of observing a *de novo* SNV in a specific gene. To estimate the likelihood of seeing multiple *de novo* SNVs in the same gene, we conducted a simulation experiment to determine the likelihood of seeing such events by chance. Genes were modeled using 14,363 brain-expressed RefSeq genes to obtain the likelihood of observing a *de novo* event based on size and GC content (see 'Chance of observing a *de novo* event').

We next used the rate of non-synonymous *de novo* variants in the sibling data to estimate the rate of non-synonymous *de novo* variants in brain-expressed genes (0.70770×10^{-8}). To allow the simulation to replicate our results, it was necessary to estimate the mean penetrance for non-synonymous *de novo* variants in the subset of genes randomly assigned as carrying risk to ASD. The penetrance was set so that the rate of non-synonymous *de novo* variants estimated in the probands matched that observed in the experiment (1.17802×10^{-8}, TableS3).

Based on the model for the number of genes contributing ASD risk (100, 333, 667, or 1,000 genes) the corresponding number of RefSeq modeled genes were assigned as being 'ASD genes' or not. This was randomized between iterations, but constant within a given iteration.

The simulation generated samples and assigned *de novo* mutations at the rate observed in siblings. Using the list of RefSeq genes and probabilities determined by size and GC content, the mutation was assigned to a specific gene. The percentage of non-synonymous *de novo* mutations in siblings was used to randomly assign mutations as being non-synonymous or silent. If the mutation was in a previously defined 'ASD gene' then the estimated penetrance was used to determine whether the *de novo* variant caused ASD. In addition, samples were randomly assigned a diagnosis of ASD regardless of *de novo* variants to match the expected prevalence of ASD in the population (0.21%).[24]

The simulation was run until the desired number of ASD cases was reached (whether from background risk or because of *de novo* variants in ASD genes); a matching number of controls (samples without ASD due to background risk or *de novo* variants in ASD genes) were selected at random and the population incidence recorded. The set of matched cases and controls were then compared and the rate of *de novo* variants was also calculated. All of these variables yielded similar average values to those observed in the real experiment when multiple iterations were run, though individual iterations were seen to vary markedly for all measures (TableS3).

Two outcomes were recorded: firstly, whether a non-ASD gene was observed to contain multiple non-synonymous *de novo* variants in a sibling. The p-value (P) of observing multiple *de novo* variants was estimated by dividing the number of iterations in which such an event was detected by the total number of iterations. Secondly, the number of randomly assigned 'non-ASD genes' with multiple *de novo* variants in a proband and the number of randomly assigned 'ASD genes' with multiple *de novo* variants in a proband. The non-ASD gene count was divided by the ASD gene count to estimate the false discovery rate (Q). These values are shown in FigS7 and FigS8 respectively.

The values P and Q reflect two different methods of assigning significance to an observation. The p-value (P) estimates the probability of seeing at least one instance of multiple independent *de novo* events in a non-ASD gene by chance, irrespective of how many times such an event is observed in ASD genes. The false discovery rate (Q) evaluates the probability of the observation of multiple independent *de novo* events in a gene that is not associated with ASD risk.

Since the entire genome was being considered this result does not need to be adjusted for multiple comparisons and because it was modeled with RefSeq genes, gene size and GC content are accounted for. The observation that double hits are more common in larger genes in our data was also seen in the model for both ASD genes and non-ASD genes.

One further model was considered in which a degree of ASD risk was assigned to all genes with the penetrance varying in an exponential distribution. As for the other models the distribution of penetrance across genes was obtained through trial and error with the aim of replicating the observed rate of *de novo* SNVs in cases and controls. In this model the objective

was to identify the genes with the top 1% of risk (effectively the top 144 genes) and denote these as 'ASD genes.'

To obtain accurate estimates the simulation was run through 150,000 iterations for each model (varying the number of ASD genes and the size of the population being considered). The results are shown in Fig2C of the main manuscript, FigS7 and FigS8.

Conservative simulation

To ensure that the results derived from the simulation experiment were robust across a range of estimates for the rate of *de novo* variants, the simulation was rerun using the upper bound of the 95% confidence interval of non-synonymous *de novo* rate in siblings (0.87576×10^{-8}) to estimate the p-value (P). The lower bound of the 95% confidence interval of non-synonymous *de novo* rate in probands (0.95349×10^{-8}) was used to estimate the false discovery rate (Q). The results are shown in FigS7 and FigS8. Varying the estimate for ASD prevalence from 0.21% to 0.93%[27] did not substantively alter the results.

Simulation with nonsense/splice site variants

The simulation was modified to consider only nonsense/splice site *de novo* mutations. The sibling nonsense/splice site *de novo* was 0.03020×10^{-8} while the rate used in probands was 0.14142×10^{-8}. All other values were unchanged. The results are shown in FigS7 and FigS8.

To obtain a conservative estimate the simulation was rerun using the upper bound of the 95% confidence interval of nonsense/splice site *de novo* rate in siblings (0.06492×10^{-8}) to estimate the p-value (P). The lower bound of the 95% confidence interval of nonsense/splice site *de novo* rate in probands (0.05948×10^{-8}) was used to estimate the false discovery rate (Q). The results are shown in FigS7 and FigS8.

Pathway analysis

To determine whether the *de novo* variants identified in probands and siblings showed enrichment for specific pathways they were run through two pathway analysis tools: Kyoto Encyclopedia of Genes and Genomes[28] (KEGG accessed January, 21, 2010 through the WebGestalt tool)[28] and Ingenuity Pathway Analysis (IPA, Ingenuity Systems, www.ingenuity.com).

Firstly a list of genes with brain-expressed non-synonymous *de novo* variants was submitted to KEGG using the list of 14,363 RefSeq brain-expressed genes as a background; otherwise default settings were used. All pathways in probands or siblings with uncorrected p-values ≤ 0.10 for enrichment were noted. If the corresponding pathway was present in the other group it was also noted regardless of p-value.

The same list of genes with non-synonymous brain-expressed genes was also submitted to IPA to confirm overlapping pathways. It is not possible to submit a list of background genes with this tool, however the background can be changed to select for human nervous system pathways only; otherwise the default setting were used. Any pathway that was noted from the prior KEGG analysis and also present in the IPA results was included regardless of the uncorrected p-value. Two pathways were enriched in probands using IPA but not seen in KEGG: GABA Receptor Signaling (p=0.03) and Sphingolipid metabolism (p=0.05). Two pathways were unique to IPA for the siblings: Dopamine Receptor Signaling (p=0.004) and Histidine metabolism (p=0.009). These additional pathways were not included in the analysis since they were absent in the primary analysis tool (KEGG).

A pathway was considered to be nominally enriched if it was present in both analysis tools and significant in at least one. The results are shown in TableS5 and FigS9C.

Protein-Protein interaction analysis

To test whether the genes detected by *de novo* variants had a greater degree of connectivity than expected by chance, we submitted the list of brain-expressed non-synonymous *de novo* variants found in probands to the Disease Association Protein-Protein Link Evaluator (DAPPLE)[29] using the InWeb database of protein-protein interactions.[30] The corresponding list of sibling variants was also submitted for analysis.

While more connectivity was seen for the proband results than sibling results, neither group displayed more connectivity than expected by chance when considering direct interactions (direct protein-protein interaction between proteins with *de novo* variants; FigS10A-B) or indirect interactions (interactions between proteins with *de novo* variants via an intermediate protein; FigS10C-D).

Rare homozygous variants

A slight over-representation of homozygous variants was seen in probands compared with siblings, however a small number of families with large differences between siblings was responsible for this result. Families were filtered to remove samples with known large CNVs (17 samples), an African American family, and a family with high consanguinity. For the remaining 155 quartet families, autosomal missense/nonsense/splice site homozygous variants were filtered to remove variants in which the population frequency was over 1%, or the variant was present as a homozygote in both siblings, or as a homozygote in any parent. Of the remaining 814 variants, 439 were seen in probands compared with 375 in siblings. This difference is significant ($p = 0.01$, binomial distribution); however, this difference is generated entirely from the 22 families with the most discordant number of filtered homozygous variants between the two siblings. In the remaining 133 families the number of variants is almost equal (239 in probands vs. 241 in siblings). A similar pattern was seen after restricting to brain-expressed variants.
The difference in rare homozygous variants could represent an association with ASD; however, in the 22 families contributing to the observed difference, half the variants demonstrated synteny, suggesting undetected deletion CNVs or blocks of homozygosity.

Rare compound heterozygous variants

No evidence of an increased burden of rare compound heterozygotes in probands was seen. Restricting missense/nonsense/splice site compound heterozygotes to those in which neither allele was present at over 1% in the population showed 328 events in probands, which is slightly less than the 343 seen in siblings. Restricting to brain-expressed genes gave a similar distribution with 232 in probands compared with 239 in siblings. Filtering to only variants with at least one nonsense or splice site variant gave 7 in probands compared with 10 in siblings.

***De novo* compound heterozygous variants**

15 *de novo* missense/nonsense/splice variants in brain-expressed genes formed compound heterozygotes with transmitted missense/nonsense/splice variants: 12 seen in probands and 3 seen in siblings. Using the rate per *de novo* in siblings ($3/55 = 0.055$) this shows that the probands have more than would be expected by chance ($p=0.009$, binomial distribution). The gene *KANK1* was seen in both the proband and sibling lists, removing this gives a more

significant result (p=0.001). All the compound heterozygotes in probands were formed by a common missense and *de novo* missense with the exception of *DOM3Z* (DOM-3 homolog Z, clears mRNAs with aberrant 5-prime-end caps) and *LLGL1* (Lethal giant larvae homolog 1, regulates of cell polarity, within the Smith-Magenis region), both with rare (<1% population frequency) instead of common missense variants, and *FCRL6* (Fc receptor-like 6, may play a role in immune function) with a *de novo* splice site and common nonsense variant.

Nonsense variants

No evidence of an increased burden of rare nonsense heterozygotes in probands was seen. To identify the most likely damaging nonsense variants, the rare autosomal heterozygous variants were filtered to remove variants in genes with common nonsense or splice site variants identified in other subjects. The variants were further restricted to those with no homozygous variants in any parents and which were only seen in one of the two siblings. 376 such variants were detected in probands compared with 370 in siblings. Restricting to brain-expressed variants gave a similar pattern with 254 in probands and 246 in siblings.

101 genes had recurrent rare nonsense variants. Of these 56 were seen in both probands and siblings, 30 were seen in only probands and 15 were seen in only siblings. One gene had three rare nonsense variants in probands: *RNASEL* (Ribonuclease L, involved in removing viral RNA from cells).

Splice site variants

No evidence of an increased burden of rare canonical splice site heterozygotes in probands was seen. To identify the most likely damaging canonical splice site variants, the rare autosomal heterozygous variants were filtered to remove variants in genes with common nonsense or splice site variants identified in other subjects. The variants were further restricted to those with no homozygous variants in any parents and which were only seen in one of the two siblings. 167 such variants were detected in probands compared with 168 in siblings. Restricting to brain-expressed variants gave a similar pattern with 117 in probands and 126 in siblings.

33 genes had recurrent rare canonical splice site variants. Of these 19 were seen in both probands and siblings, 6 were seen in only probands and 8 were seen in only siblings.

Inherited variants within high-risk CNV samples

16 families had large multigenic CNVs (6 16p11.2 deletions, 5 16p11.2 duplications, 3 7q11.23 duplications, 1 17q12 deletion, and 1 3q29 deletion). While several samples had common missense variants overlying the CNVs, only two genes within overlying CNVs had novel missense variants. Both were in the same sample with a 7q11.23 duplication. These variants were in the genes *GTF2I repeat domain-containing 1* (GTF2IRD1), a genetic determinant of mammalian craniofacial and cognitive development, and *Frizzled 9* (*FZD9*), part of the Wnt signaling pathway. Of note the *FZD9* variant formed a compound heterozygote with a common missense variant in the same gene.

- **16p11.2** 9 of the 11 samples with 16p11.2 CNVs (6 deletions and 5 duplications) had missense variants within genes in the 16p11.2 interval; no nonsense variants or canonical splice site variants were detected. The variants were all common (population allele frequency of 2-79%) and located in the genes *QPRT*, *SEZ6L2* and *DOC2A*. All of the variants in samples with deletion CNVs appeared to be hemizygous on the sequencing

data. Three of the samples had the same 3bp insertion in the gene ASPHD1; no other indels were detected.

- **7q11.23** All 3 of the samples with 7q11.23 duplications had missense variants within the 7q11.23 interval; no nonsense variants or canonical splice site variants were detected. There were two novel missense variants in the same individual in the genes *FZD9* (missense on non-duplicated paternal allele) and *GTF2IRD1* (missense on the duplicated maternal allele). Common variants were identified in the genes *TRIM50*, *FZD9*, *MLXIPL*. The novel and common variant in *FZD9* were in the same sample and from different parents giving this sample a compound heterozygote missense variant overlying a *de novo* duplication. No indels were detected.

- **17q12** No missense, nonsense or canonical splice site variants or coding indels were identified within the 17q12 region within the single sample with a 17q12 deletion.

- **3q29** A single common missense variant with an allelic frequency of 12% was present within the gene *LRRC33* within the deletion region of the single individual with this deletion. No coding indels were identified.

3. Supplementary Tables

Table S1. Overview of exome sequencing data in all quartet samples (n=800) passing quality control.

Measure	Mean (±95% CI)
Total reads (million)	115.5(±4.1)
% reads aligned	97.8% (±0.2%)
% reads on target	72.5% (±0.7%)
% Duplicate reads	11.8% (±0.6%)
Median coverage	87.1x (±3.0x)
% target at 4x	96.8% (±0.2%)
% target at 8x	94.5% (±0.3%)
% target at 20x	87.0% (±0.6%)
Base pair error rate	1.2% (±0.1%)
Coding bases analysed (million)	24.3 (±0.3)
% RefSeq bases analysed	73.4% (±1.2)
% target coding bases analysed	83.2% (±1.3)
Transition/Transversion ratio	2.71 (±0.004)
% variants not in dbSNP132	3.13% (±0.08%)

Table S2. Genes with multiple hits in this study and O'Roak et al..

Gene	Brain expressed	Count	Affected status	Variant type	p-value*
SCN2A	Yes	2	Both probands	Double nonsense	0.005
KATNAL2	Yes	2	Both probands	Double splice site	0.005
CHD8	Unknown	2	Both probands	Nonsense and frameshift	0.005
DNAH5	Yes	2	Both probands	Frameshift and missense	0.29
KIAA0100	Yes	2	Both probands	Nonsense and missense	0.29
KIAA0182	Yes	2	Both probands	Double missense	0.29
MEGF11	Yes	2	Both probands	Double missense	0.29
MYO7B	No	2	Both probands	Double missense	0.29
NTNG1	Yes	2	Both probands	Double missense	0.29
RFX8	Unknown	2	Both probands	Double missense	0.29
SLCO1C1	Yes	2	Both probands	Double missense	0.29
SUV420H1	Yes	2	Both probands	Double missense	0.29
TRIO	Yes	2	Both probands	Double missense	0.29
NAV2	Yes	2	Both probands	Missense and silent	1.00
SLC22A9	Yes	2	Both probands	Missense and silent	1.00
KANK1	Yes	2	Proband and sibling	Nonsense and missense	1.00
ARHGEF10L	Yes	2	Proband and sibling	Double missense	1.00
EIF4G1	Yes	2	Proband and sibling	Double missense	1.00
MUC16	Unknown	2	Proband and sibling	Double missense	1.00
NF1	Yes	2	Proband and sibling	Double missense	1.00
LRP1	Yes	3	Sibling, proband, proband	Missense, missense and silent	1.00
RGS7	Yes	2	Both siblings	Missense and silent	1.00
SNRNP200	Yes	2	Both siblings	Missense and silent	1.00

*Estimated using the simulation shown in Fig2, main manuscript.

Table S3. Comparison of non-synonymous simulation metrics with observed values.

Model	Observed	100 genes	333 genes	667 genes	1000 genes	Exponential
Mean estimated penetrance	NA	18.20%	5.46%	2.73%	1.82%	0.51%
% ASD cases due to *de novo* SNVs	NA	25.9%	25.9%	25.9%	25.9%	26.4%
Rate in probands (x 10^{-8})	1.41	1.41	1.41	1.40	1.41	1.41
Rate in siblings (x 10^{-8})	1.01	1.01	1.01	1.01	1.01	1.01
% Non-synonymous SNVs in probands	83.2%	78.0%	78.0%	78.0%	78.0%	78.0%
% Non-synonymous SNVs in siblings	69.8%	69.5%	69.5%	69.6%	69.6%	69.5%
Odds ratio	2.22	1.56	1.56	1.56	1.56	1.56

Table S4. Comparison of nonsense/splice site simulation metrics with observed values.

Model	Observed	100 genes	333 genes	667 genes	1000 genes	Exponential
Mean estimated penetrance	NA	100%	34.5%	17.2%	11.5%	8.92%
% ASD cases due to *de novo* SNVs	NA	6.6%	7.5%	7.5%	7.5%	7.7%
Rate in probands (x 10^{-8})	1.12	1.13	1.13	1.13	1.13	1.13
Rate in siblings (x 10^{-8})	1.01	1.01	1.01	1.01	1.01	1.01
% Nonsense and splice site SNVs in probands	9.5%	12.8%[1]	12.8%	12.8%	12.8%	12.8%
% Nonsense and splice site SNVs in siblings	3.1%	3.0%	3.0%	3.0%	3.0%	3.0%
Odds ratio	5.65	4.86[1]	4.86	4.85	4.86	4.86

[1] Note that the comparative increase in the % of nonsense variants and lower odds ratio in the simulation is because no risk was attributed to missense variants in this simulation. This does not affect the determination of significance threshold calculated by the simulation.

Table S5. Pathway analysis results based on probands and siblings in this data set.

Pathway	Probands				Siblings			
	KEGG		IPA		KEGG		IPA	
	p-value*	Genes	p-value	Genes	p-value	Genes	p-value	Genes
Thiamine metabolism	0.001	MTMR2, TPK1	NA	NA	NA	NA	NA	NA
Hematopoietic cell lineage	NA	NA	NA	NA	0.03	IL6R, ANPEP	NA	NA
Taste Transduction	0.04	GNAS, TAS2R3	NA	NA	NA	NA	NA	NA
Tight junction	0.06	LLGL1, MYH9, PPP2R1B	0.04	LLGL1, MYH9, PPP2R1B	NA	NA	0.45	PPP2R5A
Axon guidance	0.07	EPHB2, SRGAP3, SEMA4G	0.13	SRGAP3, TUBA1A, EPHB2, SEMA4G	0.44	ROBO3	0.44	GNA14, ROBO3
Wnt signaling pathway	0.30	TCF7L1, PPP2R1B	0.20	PPP2R1B, LRP1	0.14	PPARD, PPP2R5A	0.004	CDH3, PPARD, LRP1, PPP2R5A
ECM-receptor interaction	0.43	SV2B	NA	NA	0.04	LAMA4, COL11A1	NA	NA
Apoptosis	NA	NA	0.37	CAPN10	0.05	CAPN1, IRAK2	0.30	CAPN1

*p-values shown are uncorrected for multiple comparisons.

Table S6. The PPV and specificity of variant detection as the prior probability is varied.

The accuracy column shows the change in prediction accuracy as the specificity is held at 99.99% while the 'Specificity required' column shows the specificity required to maintain a 90% PPV.

Type of heterozygous event	Prior Probability	Accuracy (Specificity of 99.99%)	Specificity required (>90% accuracy)
Variant	1 in 1,000	91%	99.99%
Rare variant	1 in 20,000	33%	99.9995%
Rare missense variant	1 in 40,000	20%	99.9998%
Rare nonsense variant	1 in 2,000,000	0.5%	99.999995%
De novo variant	1 in 50,000,000	0.02%	99.9999998%

Table S7. Expected numbers of false positive *de novo* events.

Expected number of erroneous *de novo* predictions per exome and specificity assuming a target of 32 million giving 30,000 variants with a per sequenced base error rate of 2%. The variant allele sampling frequency is assumed to be 50%.

Reads supporting the variant	Total reads	Number of false positives per exome	Specificity
1	2	1,800,000	94%
2	4	60,000	99.8%
3	6	2,620	99.99%
4	8	201	99.999%
5	10	33	99.9998%
6	12	8	99.99995%
7	14	2	99.99998%
8	16	0.5	99.999997%
9	18	0.1	99.9999992%
10	20	0	99.9999998%

4. Supplementary Equations

Determining the required specificity for *de novo* prediction

De novo variants are difficult to detect due to the low prior probability of detection, estimated at $1x10^{-8}$.[5] Using Bayes theorem the specificity required to accurately predict a *de novo* event can be calculated:

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+|D)\,P(D) + P(+|N)\,P(N)}$$

- $P(D|+)$ Probability of result being true given that it is positive (PPV)
- $P(D)$ Probability of event (*prior probability* of D)
- $P(N)$ Probability of alternative ($1 - P(D)$)
- $P(+|D)$ Probability of a positive result given that the result is true (TPR or sensitivity)
- $P(+|N)$ Probability of a positive result given that the result is false (FDR or 1 - specificity)

This can be rearranged to calculate the specificity required to predict variants at 90% accuracy for a specific prior probability:

$$P(+|N) = \frac{P(+|D)\,P(D) - P(+|D)\,P(D)\,P(D|+)}{P(D|+) - P(D|+)\,P(D)}$$

Assuming a sensitivity [$P(+|D)$] of 95%, and a desired positive predictive value [$P(D|+)$] of 90% the specificity required for a given prior probability can be calculated. The results are shown in table S3 and shows that a specificity of 99.99999998% is required for accurate *de novo* prediction.

Modeling specificity of variant prediction using unique reads

A *de novo* variant can be wrongly predicted for two reasons:

1) The variant is not present in the child (false positive)
2) The variant is present in a parent (false negative)

The contribution of both errors to false *de novo* predictions can be modeled by considering the chance of recurrent errors in sequence or the chance of missing a variant allele through random sampling. The former (false positives in children) has a larger contribution to erroneous *de novo* predictions because of the large number of bases being considered (32-45 million) while the parent errors are only considered in bases predicted to be variant (20-30 thousand).

False positive predictions in children

False positives can arise through either systematic or random errors in sequencing. Two processes remove systematic errors:

1) Cleaning the data with SysCall[10]
2) Concentrating on novel variants (systematic errors are likely to be present in multiple samples, including the parents)

The contribution of random errors can be assessed using the calculated per base error rate for the sequencing. This value is estimated by assessing the number of non-reference bases in every aligned read. The calculated error varies from 0.3-6.8% with a mean of 1.3%. Some of the non-reference bases are true variants, however these should account for 0.1-0.2% only.

Assuming the errors are distributed evenly between the three non-reference bases the probability of seeing each base can be calculated as 0.43% per base and 98.70% for the reference base. Since the errors need to consistently give the same base the probability can be considered as 0.43% for the predicted variant and 99.57% for all other bases.

The chance of seeing a specific outcome from a given number of reads is therefore:

$$\left[\frac{\text{Error rate}}{3}\right]^{\text{Variant reads}} \times \left[1 - \frac{\text{Error rate}}{3}\right]^{\text{Non-variants reads}}$$

To complete the model two further considerations must be added:

1) The number of combinations of reads that could give this specific result (the equation above only considers all variant reads being the first to be read, followed by all non-variant reads; in practice the variant reads can be in any combination). The number of combinations for a specific number of variant and non-variant reads can be calculated using Pascal's triangle.

2) That any one of the three bases could cause the false positive variant, therefore the probability calculated must be multiplied by three.

$$\left[\frac{\text{Error rate}}{3}\right]^{\text{Variant reads}} \times \left[1 - \frac{\text{Error rate}}{3}\right]^{\text{Non-variant reads}} \times \left[\frac{n!}{k!(n-k)!}\right] \times 3$$

Where :

n = row number (total reads – 1)

k = element in the row (number of variant reads)

This equation estimates the per base probability of seeing a false positive through random errors given the error rate, number of variant reads and number of non-variant reads.

False negative predictions in the parents

Missing a variant in the parents occurs because the variant allele is not present in the sequenced DNA through random sampling error. Since there are usually two alleles the chance of seeing the variant allele is 0.5 per read (assuming 100% sequence detection and no hybridization bias against the variant allele; both assumptions are unlikely to have a significant effect). The per base probability of a false negative result through random sampling is simply 0.5 to the power of the number of reads.

Exome-wide *de novo* detection

The equations described above model per base false positive and per base false negative rates in sequencing data. To estimate the specificity corresponding to these probabilities the per base false positive rate is multiplied by the target size (32-45 million) while the per base false negative is multiplied by the number of variants (20-30 thousand) since the parental data is only considered if a variant is present.

TableS7 shows the expected number of erroneous *de novo* predictions per exome and specificity assuming a target of 32 million giving 30,000 variants with a per sequenced base error rate of 2% (a high estimate was used in view of the non-stochastic nature of errors). It assumes a variant allele frequency of exactly 50%. The number of reads supporting the variant are the major determinant in this equation while total reads has a less dramatic effect.

Experimental validation of *de novo* predictions

De novo variant predictions were tested by PCR and Sanger sequencing in blood derived DNA. The number of unique reads supporting the variant was calculated and FigS12 shows the validation results for different values of unique reads supporting the variant. All variants with less than 8 unique reads supporting the variant failed to confirm, while all variants with at least 8 unique reads supporting the variant were confirmed. This matches the prediction of the model for expected erroneous *de novo* variants shown in TableS7.

5. Supplementary Data

Supplementary_Data_S1 Quality metrics and sample IDs
This excel file details the key quality metrics for all 928 samples obtained during sequencing, including the number of RefSeq coding and splice site bases per family meeting the 20 unique reads in all family members threshold for *de novo* detection. Gender, IQ, trio vs. quartet, and quality control result are listed.

Supplementary_Data_S2 List of *de novo* variants
This excel file details the confirmed *de novo* variants in probands and siblings including genomic co-ordinates, gene annotation, and severity scores.

6. References for supplementary materials

1. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat Genet* **43**, 585-9 (2011).
2. Girard, S.L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat Genet* (2011).
3. Roach, J. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
4. Consortium, G.P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
5. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).
6. 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**, 863-85 (2011).
7. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-5 (2010).
8. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
10. Singer, M., Meacham, F. & Pachter, L. SysCall. (2011).
11. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-9 (2011).
12. Bayés, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci* **14**, 19-21 (2011).
13. Collins, M.O. *et al.* Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J Neurochem* **97 Suppl 1**, 16-23 (2006).
14. Abul-Husn, N.S. *et al.* Systems approach to explore components and interactions in the presynapse. *Proteomics* **9**, 3303-15 (2009).
15. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).
16. Cooper, G. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* **7**, 250-1 (2010).
17. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).
18. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
19. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
20. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

21. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* **1**, 40-7 (2000).
22. Zeger, S.L. & Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-30 (1986).
23. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-97 (2011).
24. Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatr Res* **65**, 591-8 (2009).
25. Moreno-De-Luca, D. *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* **87**, 618-30 (2010).
26. Bunge, J. & Fitzpatrick, M. Estimating the Number of Species: A Review. *Journal of the American Statistical Association* **88**, 364-373 (1993).
27. Kim, Y.S. *et al.* Prevalence of autism spectrum disorders in a total population sample. *Am J Psychiatry* **168**, 904-12 (2011).
28. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
29. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
30. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309-16 (2007).