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.

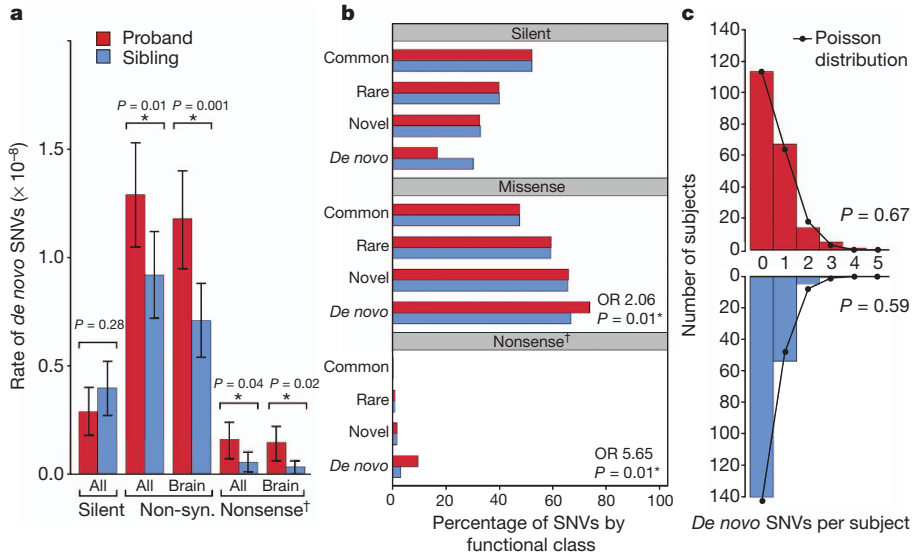

Figure 1 | Enrichment of non-synonymous *de novo* variants in probands relative to sibling controls. **a**, The rate of *de novo* variants is shown for 200 probands (red) and matched unaffected siblings (blue). 'All' refers to all RefSeq genes in hg18, 'Brain' refers to the subset of genes that are brain-expressed[24] 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.

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],

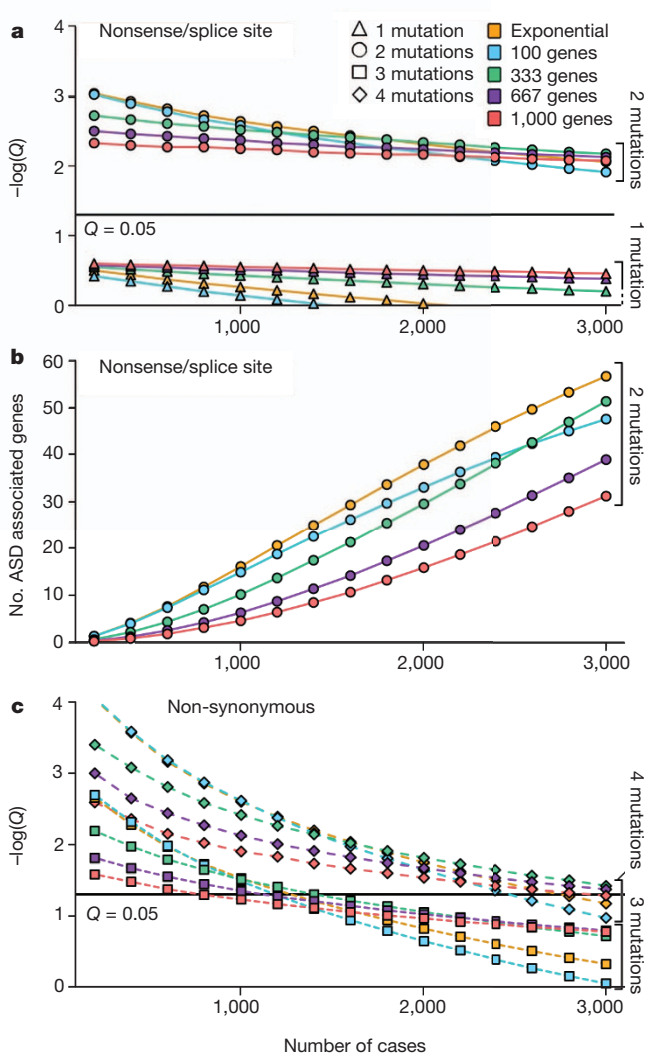

Figure 2 | Identification of multiple *de novo* mutations in the same gene reliably distinguishes risk-associated mutations. **a**, Results of a simulation experiment modelling the likelihood of observing two independent nonsense/splice site *de novo* mutations in the same brain-expressed gene among 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*

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

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).

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.