LETTER

doi:10.1038/nature11011

Patterns and rates of exonic *de novo* mutations in autism spectrum disorders

Benjamin M. Neale[1,2], Yan Kou[3,4], Li Liu[5], Avi Ma'ayan[3], Kaitlin E. Samocha[1,2], Aniko Sabo[6], Chiao-Feng Lin[7], Christine Stevens[2], Li-San Wang[7], Vladimir Makarov[4,8], Paz Polak[2,9], Seungtai Yoon[4,8], Jared Maguire[2], Emily L. Crawford[10], Nicholas G. Campbell[10], Evan T. Geller[7], Otto Valladares[7], Chad Schafer[5], Han Liu[11], Tuo Zhao[11], Guiqing Cai[4,8], Jayon Lihm[4,8], Ruth Dannenfelser[3], Omar Jabado[12], Zuleyma Peralta[12], Uma Nagaswamy[6], Donna Muzny[6], Jeffrey G. Reid[6], Irene Newsham[6], Yuanqing Wu[6], Lora Lewis[6], Yi Han[6], Benjamin F. Voight[2,13], Elaine Lim[1,2], Elizabeth Rossin[1,2], Andrew Kirby[1,2], Jason Flannick[2], Menachem Fromer[1,2], Khalid Shakir[2], Tim Fennell[2], Kiran Garimella[2], Eric Banks[2], Ryan Poplin[2], Stacey Gabriel[2], Mark DePristo[2], Jack R. Wimbish[14], Braden E. Boone[14], Shawn E. Levy[14], Catalina Betancur[15], Shamil Sunyaev[2,9], Eric Boerwinkle[6,16], Joseph D. Buxbaum[4,8,12,17], Edwin H. Cook Jr[18], Bernie Devlin[19], Richard A. Gibbs[6], Kathryn Roeder[5], Gerard D. Schellenberg[7], James S. Sutcliffe[10] & Mark J. Daly[1,2]

Autism spectrum disorders (ASD) are believed to have genetic and environmental origins, yet in only a modest fraction of individuals can specific causes be identified[1,2]. To identify further genetic risk factors, here we assess the role of *de novo* mutations in ASD by sequencing the exomes of ASD cases and their parents ($n = 175$ trios). Fewer than half of the cases (46.3%) carry a missense or nonsense *de novo* variant, and the overall rate of mutation is only modestly higher than the expected rate. In contrast, the proteins encoded by genes that harboured *de novo* missense or nonsense mutations showed a higher degree of connectivity among themselves and to previous ASD genes[3] as indexed by protein-protein interaction screens. The small increase in the rate of *de novo* events, when taken together with the protein interaction results, are consistent with an important but limited role for *de novo* point mutations in ASD, similar to that documented for *de novo* copy number variants. Genetic models incorporating these data indicate that most of the observed *de novo* events are unconnected to ASD; those that do confer risk are distributed across many genes and are incompletely penetrant (that is, not necessarily sufficient for disease). Our results support polygenic models in which spontaneous coding mutations in any of a large number of genes increases risk by 5- to 20-fold. Despite the challenge posed by such models, results from *de novo* events and a large parallel case–control study provide strong evidence in favour of *CHD8* and *KATNAL2* as genuine autism risk factors.

In spite of the substantial heritability, few genetic risk factors for ASD have been identified[1,2]. Copy number variants (CNVs), in particular *de novo* and large events spanning multiple genes, have been identified as conferring risk[4,5]. Although these CNVs provide important leads to underlying biology, they rarely implicate single genes, are rarely fully penetrant, and many confer risk to a broad range of conditions including intellectual disability, epilepsy and schizophrenia[6]. There are also documented instances of rare single nucleotide variants (SNVs) that are highly penetrant for ASD[3].

Large-scale genetic studies make clear that the origins of ASD risk are multifarious, and recent estimates based on CNV data put the number of independent risk loci in the hundreds[5]. Yet knowledge regarding specific risk-determining genes and the overall genetic architecture for ASD remains incomplete. Although new sequencing technologies provide a catalogue of most variation in the genome, the profound locus heterogeneity of ASD makes it challenging to distinguish variants that confer risk from the background noise of inconsequential SNVs. *De novo* variation, being less frequent and potentially more deleterious, could offer insights into risk-determining genes. Accordingly, we sought to evaluate carefully the observed rate and consequence of *de novo* point mutations in the exomes of ASD subjects.

We performed exome sequencing of 175 ASD probands and their parents across five centres with multiple protocols and validation techniques (Supplementary Information). We used a sensitive and specific analytical pipeline based on current best practices[7–9] to analyse all data and observed no heterogeneity of mutation rate across centres.

In the entire sample, we observed 161 coding region point mutations (101 missense, 50 silent and 10 nonsense), with an additional two conserved splice site (CSS) SNVs and six frameshift insertions/deletions (indels) validated and included in pathway analyses (Supplementary Table 1).

To determine whether the rate of coding region point mutations was elevated, we estimated the mutation rate in light of coverage and base context using two parallel approaches (Supplementary Information). On the basis of both models, the exome target should have a significantly increased (~30%) mutation rate compared to the genome. Conservatively, by assuming the low end of the estimated mutation rate from recent whole-genome data (1.2×10^{-8})[10], we estimate a mutation rate of 1.5×10^{-8} for the exome sequence captured here. The observed point mutation rate of 0.92 per exome is slightly but not significantly elevated versus expectation (Table 1) and is insensitive to adjustment for lower coverage regions (Supplementary Information). Indeed our rate is similar to that of ref. 11.

Per-family events were distributed exquisitely according to the Poisson distribution (Table 1), suggesting limited variation in the underlying rate of *de novo* mutation in ASD families. The relative rates

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029, USA. [4]Seaver Autism Center for Research and Treatment, Mount Sinai School of Medicine, New York, New York 10029, USA. [5]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15232, USA. [6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [7]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [8]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [9]Division of Genetics, Department of Medicine Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [10]Vanderbilt Brain Institute, Departments of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt University, Nashville, Tennessee 37232, USA. [11]Biostatistics Department and Computer Science Department, Johns Hopkins University, Baltimore, Maryland 21205, USA. [12]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA. [13]Department of Pharmacology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA. [14]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [15]INSERM U952 and CNRS UMR 7224 and UPMC Univ Paris 06, 75005 Paris, France. [16]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [17]Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York 10029, USA. [18]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60608, USA. [19]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA.

Table 1 | Distribution of events per family

Events per family	All ASD trios		Random mut. exp.‡
	Exon DN SNVs*	Exp.†	
0	71	69.7	73.2
1	62	64.2	63.8
2	28	29.5	27.8
3	10	9.1	8.1
4	2	2.1	1.8
5	1	0.4	0.3
Mean		0.920	0.871

* Exon DN SNVs include all single nucleotide variants in coding sequence but excludes indels and intronic variants.
† The expected distribution of number of trios with a given event count as determined by the Poisson.
‡ Random mut. exp. is the expectation for 175 trios based on the sequence-context mutation rate model M1 (Supplementary Information) based on the count of the number of trios that have at least 10× coverage.

of 'functional' (missense, nonsense, CSS and read-through) versus silent changes did not deviate from expectation (Table 2). We did, however, observe ten nonsense mutations (6.2%), which exceeded expectation (3.3%) (one-tailed $P = 0.04$; Supplementary Information).

We examined missense mutations using PolyPhen-2 scores[12] to measure severity, as some missense variants can severely affect function[13]. These scores showed no deviation from random expectation. The observed PolyPhen-2 scores clearly deviate from standing variation in the parents (Table 2), but such variation, even the rarest category, has survived selective pressure and so is inappropriate for comparison to *de novo* events.

We observed three genes with two *de novo* mutations: *BRCA2* (two missense), *FAT1* (two missense) and *KCNMA1* (one missense, one silent). A gene with two or more non-synonymous *de novo* hits across a panel of trios might indicate strong candidacy. However, simulations (Supplementary Information) show that two such hits are inadequate to define a gene as a conclusive risk factor given the number of observed events in the study.

From analyses of secondary phenotypes (Supplementary Tables 2 and 3), the most striking result is that paternal and maternal age, themselves highly correlated ($r^2 = 0.679$, P-value <0.0001), each strongly predicts the number of *de novo* events per offspring (paternal age, $P = 0.0013$; maternal age, $P = 0.000365$), consistent with aggregating mutations in germ cells in the paternal line[14]. Consistent with a liability threshold model, there is an increased rate of *de novo* mutation in female versus male cases (1.214 for females versus 0.914 for males); however, the difference is not significant, owing to limited sample size. Considering phenotypic correlates, we observed no rate difference between subjects with strict autism versus those with a broader ASD classification, between positive and negative family history, or any significant effect of *de novo* mutation on verbal, non-verbal or full-scale IQ (Supplementary Table 3).

Given that hundreds of loci are apparently involved in autism[5] and *de novo* mutations therein affect ASD risk, we modelled different numbers of risk genes and penetrances (Supplementary Information) and show that a model of hundreds of genes with high penetrance mutations is excluded by our data; however, more modest contributions of *de novo* variants are not. For example, up to 20% of cases carrying a *de novo* event conferring a 10- or 20-fold increased risk is consistent with these data (Supplementary Table 4). Thus, our data are consistent with either chance mutation or a modest role for *de novo* mutations on risk. Importantly, a single deleterious event is unlikely to fully explain disease in a patient.

Table 2 | Rates of mutation annotation given variant type

Type of *de novo* mutation	*De novo* (%)*	Random *de novo* (%)	Singletons (%)†	Doubletons (%)†	≥3 (%)†
Missense	62.7	66.1	59.5	55.4	48.8
Nonsense	6.2	3.3	1.2	0.8	0.4
Synonymous	31.1	30.6	39.3	43.8	50.8
PolyPhen-2 missense classification					
Benign	35.0	35.9	46.6	51.3	63.4
Possibly damaging	21.0	18.9	18.8	17.7	15.1
Probably damaging	44.0	45.2	34.7	31.0	21.4

* All indels and failing variants were removed.
† Singletons, doubletons and ≥3 (copies) are only those variants called in 192 parents.

We therefore posed two questions of the group of genes harbouring *de novo* functional mutations: do the protein products of these genes interact with each other more than expected, and are they unusually enriched in, or connected to, previous curated lists of ASD-implicated genes? Using an *in silico* approach (DAPPLE)[15], the protein–protein connectivity defined by InWeb[16] in the set of 113 genes harbouring functional *de novo* mutations was evaluated. These analyses (Fig. 1) showed significantly greater connectivity among the *de novo* identified proteins than would be expected by chance ($P < 0.001$) (Supplementary Information).

Querying previously defined, manually curated lists of genes[3] associated with high risk for ASD with or without intellectual disability (Supplementary Table 5), and high-risk intellectual disability genes (Supplementary Table 6), we asked whether there was significant enrichment for *de novo* mutations in these genes. Five genes with functional *de novo* events were previously associated with ASD and/or intellectual disability (*STXBP1*, *MEF2C*, *KIRREL3*, *RELN* and *TUBA1A*); for four of these genes (all but *RELN*) the previous evidence indicated autosomal dominant inheritance.

We then assessed the average distance (D_i, Supplementary Fig. 2) of the *de novo* coding variants in brain-expressed genes (see supplement) to the ASD/intellectual disability list using a protein–protein interaction background network. To enhance power, data from a companion study[11] were used, including the observed silent *de novo* variants and *de novo* variants in unaffected siblings as comparators. The average distance for non-synonymous variants was significantly smaller for the case set than the comparator set (3.66 ± 0.42 versus 3.78 ± 0.59; permutation $P = 0.033$) (Supplementary Fig. 3). Much of this signal comes from 31 synaptic genes identified by three large-scale synaptic proteomic studies ($D_i = 3.47 \pm 0.46$ versus 3.57 ± 0.60; permutation $P = 0.084$) (Fig. 2; see also Supplementary Fig. 4 for the complete data). Taken in total, these independent gene set analyses, along with the modest enrichment of *de novo* variants over background rates in

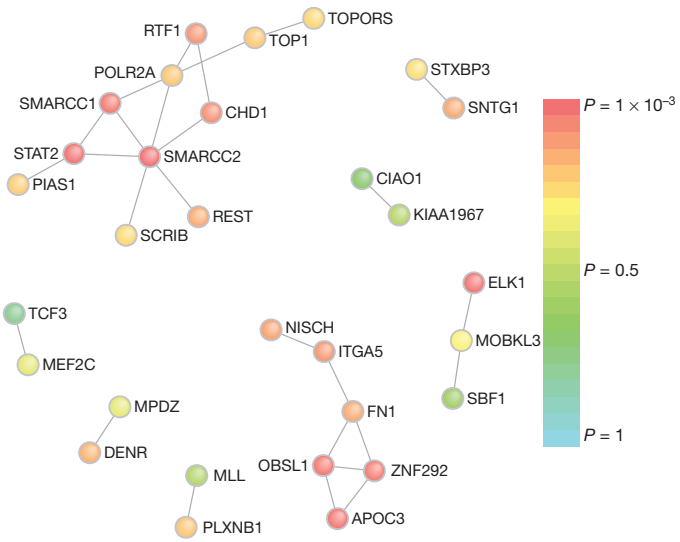

Figure 1 | Protein–protein interaction for genes with an observed functional *de novo* event. Direct protein connections from InWeb, restricting to genes harbouring *de novo* mutations for DAPPLE analysis. Two extensive networks are identified: the first is centred on SMARCC2 with 12 connections across 11 genes; the second is centred on FN1 with 7 connections across 6 genes. The P value for each gene having as many connections as those observed is indicated by node colour.

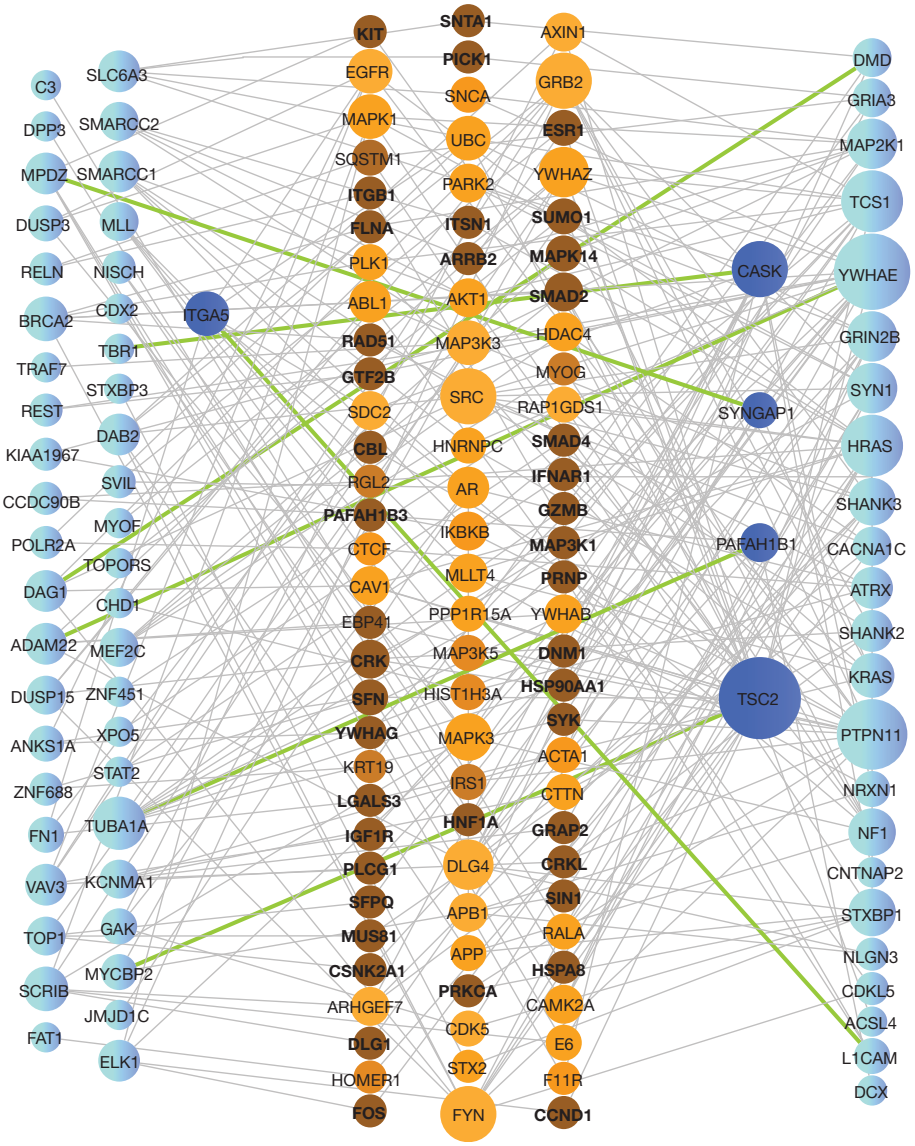

Figure 2 | Direct and indirect protein–protein interaction for genes with a functional *de novo* event and previous ASD genes. PPI network analysis for *de novo* variants and 31 previous synaptic ASD genes (see Supplementary Information). Nodes are sized based on connectivity. Genes harbouring *de novo* variants (left) and previous ASD genes (right) are coloured blue, with dark blue nodes representing genes that belong to one of these lists and are also intermediate proteins. Intermediate proteins (centre) are coloured in shades of orange based on a *P* value computed using a proportion test, where a darker colour represents a lower *P* value. Green edges represent direct connections between genes harbouring *de novo* variants (left) and previous ASD genes. All other edges, connecting to intermediate proteins, are shown in grey.

ASD, indicate that a proportion of the *de novo* events observed in this study probably contribute to autism risk.

Using whole-exome sequencing of autism trios, we demonstrate a rate, functional distribution and predicted impact of *de novo* mutation largely consistent with chance mutational processes governed by sequence context. This lack of significant deviation from random mutational processes indicates a more limited role for the contribution of *de novo* mutations to ASD pathogenesis than has previously been suggested[17], and specifically highlights the fact that observing a single *de novo* mutation, even an apparently 'severe' loss-of-function allele, is insufficient to implicate a gene as a risk factor. Yet the pathway analyses presented here assert that the overall set of genes hit with functional *de novo* mutations is not random and that these genes are biologically related to each other and to previously identified ASD/intellectual disability candidate genes. Modelling the *de novo* mutational process under a range of genetic models reveals that some models are inconsistent with the observed data—for example, 100 rare, fully penetrant Mendelian genes similar to Rett's syndrome—whereas others are not inconsistent, such as spontaneous 'functional' mutation in hundreds of genes that would increase risk by 10- or 20-fold (Supplementary Table 4). Models that fit the data are consistent with the relative risks estimated for most *de novo* CNVs[5] and suggest that *de novo* SNVs, like most CNVs, often combine with other risk factors rather than fully cause disease. Furthermore, these models indicate that *de novo* SNV events will probably explain <5% of the overall variance in autism risk (Supplementary Table 4).

Considering the two companion papers[11,18], 18 genes with two functional *de novo* mutations are observed in the complete data. Using simulations, 11.91 genes on average harbour functional mutations by chance (Supplementary Table 7). Thus, a set of 18 genes with two or more hits is not quite significant ($P = 0.063$). Matching loss-of-function variants, however, at *SCN2A*, *KATNAL2* and *CHD8* (Supplementary Table 7) are unlikely to occur by chance because of the expected very low rate of *de novo* nonsense, splice and frameshift variants. We evaluated these strong candidates further using exome sequencing on 935 cases and 870 controls, and at both *KATNAL2* and

CHD8 three additional loss-of-function mutations were observed in cases with none in controls. No additional loss-of-function mutations were seen at *SCN2A* in the case–control data, but a new splice site *de novo* event has been validated in an additional autism case while this paper was in press, strengthening the evidence for this gene as relevant to autism. Using data from more than 5,000 individuals in the NHLBI Exome Variant Server (http://evs.gs.washington.edu/EVS/) as additional controls, three loss-of-function mutations were seen in *KATNAL2* but none in *CHD8*, making the additional observation of three *CHD8* loss-of-function mutations in our cases significant evidence ($P < 0.01$) of this being a genuine autism susceptibility gene. Not all genes with double hits are nearly so promising (Supplementary Information and Supplementary Tables 8 and 9), supporting the estimate above that most of such observations are simply chance events. Overall, these data underscore the challenge of establishing individual genes as conclusive risk factors for ASD, a challenge that will require larger sample sizes and deeper analytical integration with inherited variation.

METHODS SUMMARY

We ascertained probands using the Autism Diagnostic Interview-Revised (ADI-R), the Autism Diagnostic Observation Schedule-Generic (ADOS) and the DSM-IV diagnosis of a pervasive developmental disorder. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects that were not assessed with the ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

For 175 trios, we performed exome capture and sequencing using either the Agilent 38Mb SureSelect v2 ($n = 118$), the NimbleGen Seq Cap EZ SR v2 ($n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

All sequence data were processed with Picard (http://picard.sourceforge.net/), which recalibrates quality scores and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. Putative *de novo* mutations were identified restricting to sites passing standard filters and both parents were homozygous for the reference sequence and the offspring was heterozygous, and each genotype call was made confidently (see Supplementary Information).

All putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (71 trios) or by using Sequenom MALDI-TOF (104 trios). All events were annotated using RefSeq hg19.

We modelled a Poisson process consistent with the mutation model and observed data. We varied the fraction of genes that influence risk, the probability of a functional variant, and the penetrance of said events.

We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes × 2 sites), but we also performed mega- and meta-analyses (23 genes × 2 methods).

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

Received 13 September 2011; accepted 6 March 2012.
Published online 4 April 2012.

1. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167,** 1357–1363 (2010).
2. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68,** 1095–1102 (2011).
3. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
4. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
5. 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).
6. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25,** 528–535 (2009).
7. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
8. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
10. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
11. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* http://dx.doi.org/10.1038/nature10945 (this issue).
12. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
13. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–739 (2007).
14. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1,** 40–47 (2000).
15. 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).
16. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105,** 20870–20875 (2008).
17. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
18. 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).
19. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
20. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).

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

Acknowledgements This work was directly supported by NIH grants R01MH089208 (M.J.D.), R01 MH089025 (J.D.B.), R01 MH089004 (G.D.S.), R01MH089175 (R.A.G.) and R01 MH089482 (J.S.S.), and supported in part by NIH grants P50 HD055751 (E.H.C.), RO1 MH057881 (B.D.) and R01 MH061009 (J.S.S.). Y.K., G.C. and S.Y. are Seaver Fellows, supported by the Seaver Foundation. We thank T. Lehner, A. Felsenfeld and P. Bender for their support and contribution to the project. We thank S. Sanders and M. State for discussions on the interpretation of *de novo* events. We thank D. Reich for comments on the abstract and message of the manuscript. We thank E. Lander and D. Altshuler for comments on the manuscript. We acknowledge the assistance of M. Potter, A. McGrew and G. Crockett without whom these studies would not be possible, and Center for Human Genetics Research resources: Computational Genomics Core, Genetic Studies Ascertainment Core and DNA Resources core, supported in part by NIH NCRR grant UL1 RR024975, and the Vanderbilt Kennedy Center for Research on Human Development (P30 HD015052). This work was supported in part by R01MH084676 (S.S.). We acknowledge the clinicians and organizations that contributed to samples used in this study and the particular support of the Mount Sinai School of Medicine, University of Illinois-Chicago, Vanderbilt University, the Autism Genetics Resource Exchange and the institutions of the Boston Autism Consortium. We acknowledge A. Estes and G. Dawson for patient collection/characterization. We acknowledge partial support from U54 HG003273 (R.A.G.) and U54 HG003067 (E. Lander). J.D.B., B.D., M.J.D., R.A.G., A.S., G.D.S. and J.S.S. are lead investigators in the Autism Sequencing Consortium (ASC). The ASC is comprised of groups sharing massively parallel sequencing data in autism. Finally, we are grateful to the many families, without whose participation this project would not have been possible.

Author Contributions Laboratory work: A.S., C.St., G.C., O.J., Z.P., J.D.B., D.M., I.N., Y.W., L.L., Y.H., S.G., E.L.C., N.G.C. and E.T.G. Data processing: B.M.N., K.E.S., E.L., A.K., J.F., M.F., K.S., T.F., K.G., E.Ba., R.P., M.DeP., S.G., S.Y., V.M., J.L., J.D.B., A.S., C.St., U.N., J.G.R., J.R.W., B.E.B., S.E.L., C.F.L., L.S.W. and O.V. Statistical analysis: B.M.N., L.L., K.E.S., C.Sh., B.F.V., J.M., E.R., S.S., P.P., Y.K., A.M., R.D., C.-F.L., L.-S.W., H.L., T.Z., E.Bo., R.A.G., J.D.B., C.B., E.H.C., J.S.S., G.D.S., B.D., K.R. and M.J.D. Principal Investigators/study design: E.Bo., R.A.G., E.H.C., J.D.B., K.R., B.D., G.D.S., J.S.S. and M.J.D. Y.K., L.L., A.M., K.E.S., A.S. and C.-F.L. contributed equally to this work. E.Bo., J.D.B., E.H.C., B.D., R.A.G., K.R., G.D.S., J.S.S. and M.J.D. are lead investigators of the ARRA Autism Sequencing Collaboration.

Author Information Data included in this manuscript have been deposited at dbGaP under accession number phs000298.v1.p1 and is available for download at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1. 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.J.D. (mjdaly@atgu.mgh.harvard.edu), J.D.B. (joseph.buxbaum@mssm.edu) or K.R. (kathryn.roeder@gmail.com).

METHODS

Phenotype assessment. Affected probands were assessed by research-reliable research personnel using Autism Diagnostic Interview-Revised (ADI-R), and the Autism Diagnostic Observation Schedule-Generic (ADOS) and DSM-IV diagnosis of a pervasive developmental disorder was made by a clinician. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects from AGRE that were not assessed with the ADOS. In all, 85% of probands were classified with autism on both the ADI-R and ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

Exome sequencing, variant identification and *de novo* detection. Exome capture and sequencing was performed at each site using similar methods. Exons were captured using the Agilent 38 Mb SureSelect v2 (University of Pennsylvania and Broad Institute $n = 118$), the NimbleGen Seq Cap EZ SR v2 (Mt Sinai School of Medicine, Vanderbilt University $n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

Sequence processing and variant calling was performed using a similar computational workflow at all sites. Data were processed with Picard (http://picard.sourceforge.net/), which uses base quality-score recalibration and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. The variable sites that we have considered in analysis are restricted to those that pass GATK standard filters. From this set of variants, we identified putative *de novo* mutations as sites where both parents were homozygous for the reference sequence and the offspring was heterozygous and each genotype call was made confidently (see Supplementary Information).

Validation of *de novo* events. Putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (University of Pennsylvania, Mt Sinai School of Medicine, Vanderbilt University, Baylor Medical College) or by Sequenom MALDI-TOF genotyping of trios (Broad).

Gene annotation. All identified mutations were then annotated using RefSeq hg19. The functional impact of variants was assessed for all isoforms of each gene, with the most severe annotation taking priority. Splice site variants were identified as occurring within two base pairs of any intron/exon boundary.

Expectation of *de novo* mutation calculation. To calculate the expected *de novo* rate, we assessed the mutability of all possible trinucleotide contexts in the intergenic region of the human genome for variation in two fashions: fixed genomic differences compared to chimpanzee and baboon[12] and variation identified from the 1,000 Genomes project. The overall mutation rate for the exome was then determined by summing the probability of mutation for all bases in the exome that were captured successfully. We also determined the probability of each class functional mutation by summing the annotated variants.

Pathway analyses. We applied DAPPLE[15], which uses the InWeb database[16], to determine whether there is excess protein–protein interaction across the genes hit by a functional *de novo* event. We also assessed whether these genes were more closely connected to a list of ASD genes[3].

Modelling *de novo* events. We modelled a Poisson process consistent with the expected distribution defined by the mutation model and with the observed data. We varied the fraction of genes that influence risk, the probability a variant in a gene would be functional, and the penetrance of functional *de novo* events. We also simulated a random set of *de novo* events to estimate the probability of hitting a gene multiple times.

Association analysis. We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes $\times$ 2 sites), but we also performed mega- and meta-analyses (23 genes $\times$ 2 methods).