Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations

Brian J. O'Roak¹, Laura Vives¹, Santhosh Girirajan¹, Emre Karakoc¹, Niklas Krumm¹, Bradley P. Cox¹, Roie Levy¹, Arthur Ko¹, Choli Lee¹, Joshua D. Smith¹, Emily H. Turner¹, Ian B. Stanaway¹, Benjamin Vernot¹, Maïka Malig¹, Carl Baker¹, Beau Reilly², Joshua M. Akey¹, Elhanan Borenstein¹,³,⁴, Mark J. Rieder¹, Deborah A. Nickerson¹, Raphael Bernier², Jay Shendure¹ & Evan E. Eichler¹,⁵

It is well established that autism spectrum disorders (ASD) have a strong genetic component; however, for at least 70% of cases, the underlying genetic cause is unknown. Under the hypothesis that de novo mutations underlie a substantial fraction of the risk for developing ASD in families with no previous history of ASD or related phenotypes—so-called sporadic or simplex families—we sequenced all coding regions of the genome (the exome) for parent–child trios exhibiting sporadic ASD, including 189 new trios and 20 that were previously reported. Additionally, we also sequenced the exomes of 50 unaffected siblings corresponding to these new (n = 31) and previously reported trios (n = 19), for a total of 677 individual exomes from 209 families. Here we show that de novo point mutations are overwhelmingly paternal in origin (4:1 bias) and positively correlated with paternal age, consistent with the modest increased risk for children of older fathers to develop ASD. Moreover, 39% (49 of 126) of the most severe or disruptive de novo mutations map to a highly interconnected β-catenin/chromatin remodelling protein network ranked significantly for autism candidate genes. In proband exomes, recurrent protein-altering mutations were observed in two genes: CHD8 and NTNG1. Mutation screening of six candidate genes in 1,703 ASD trios and 20 that were previously reported4. Additionally, we also observed complex classes of de novo mutation including: five cases of multiple mutations in close proximity; two events consistent with paternal germline mosaicism (that is, where both siblings contained a de novo event observed in neither parent); and nine events showing a weak minor allele profile consistent with somatic mosaicism (Supplementary Table 3 and Supplementary Figs 2 and 3).

Of the severe de novo events, 28% (33 of 120) are predicted to truncate the protein. The distribution of synonymous, missense and nonsense changes corresponds well with a random mutation model1 (Supplementary Fig. 4 and Supplementary Table 2). However, the difference in nonsense rates between de novo and rare singleton events (not present in 1,779 other exomes) is striking (4:1) and suggests strong selection against new nonsense events (Fisher’s exact test, P < 0.0001). In contrast with a recent report4, we find no significant difference in mutation rate between affected and unaffected individuals; however, we do observe a trend towards increased non-synonymous rates in probands, consistent with the findings of ref. 9 (Supplementary Tables 1 and 2).

Given the association of ASD with increased paternal age and our previous observations4, we used molecular cloning, read-pair information, and obligate carrier status to identify informative markers linked to 51 de novo events and observed a marked paternal bias (41:10; binomial P < 1.4 × 10⁻³; Fig. 1a and Supplementary Tables 3 and 5). This provides strong direct evidence that the germline mutation rate in protein-coding regions is, on average, substantially higher in males. A similar finding was recently reported for de novo CNVs20. In addition, we observe that the number of de novo events is positively correlated with increasing paternal age (Spearman’s rank correlation = 0.19; P < 0.008; Fig. 1b). Together, these observations are consistent with the hypothesis that the modest increased risk for children of older fathers to develop ASD is the result of an increased mutation rate.

Using sequence read-depth methods in 122 of the 189 families, we scanned ASD probands for either de novo CNVs or rare (<1% of controls), inherited CNVs. Individual events were validated by either array CGH or genotyping array (see Methods). We identified 76 events in 53 individuals, including six de novo (median size 467 kilobases (kb)) and 70 inherited (median size 155 kb) CNVs (Supplementary Table 6). These include disruptions of EHMT1 (Kleefstra’s syndrome, Online Mendelian Inheritance in Man (OMIM) accession 610253), CNTNAP4 (reported in children with developmental delay and autism19) and the 16p11.2 duplication (OMIM 611913) associated with developmental delay, bipolar disorder and schizophrenia. We performed a multivariate analysis on non-variant IQ (NVIQ), verbal IQ (VIQ) and the load of extreme de novo mutations—where extreme is defined as point mutations that truncate proteins, intersect

1Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. 2Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington 98195, USA. 3Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA. 4Santa Fe Institute, Santa Fe, New Mexico 87501, USA. 5Howard Hughes Medical Institute, Seattle, Washington 98195, USA.

©2012 Macmillan Publishers Limited. All rights reserved
Mendelian or ASD loci (n = 57), or de novo CNVs that intersect genes (n = 5) (Fig. 1c and Supplementary Discussion). NVIQ, but not VIQ, decreased significantly (P < 0.01) with increased number of events. Covariant analysis of the samples with CNV data showed that this finding was strengthened, but not exclusively driven, by the presence of either de novo or rare CNVs (Supplementary Fig. 5).

Among the de novo events, we identified 62 top ASD risk contributing mutations based on the deleteriousness of the mutations, functional evidence, or previous studies (Table 1). Probands with these mutations spanned the range of IQ scores, with only a modest non-significant trend towards individual’s co-morbid with intellectual disability (Supplementary Figs 1 and 6). We observed recurrent, protein-disruptive mutations in two genes: NTNG1 (netrin G1) and CHD8 (chromodomain helicase DNA binding protein 8). Given their locus-specific mutation rates, the probability of identifying two independent mutations in our sample set is low (uncorrected, NTNG1: P < 1.2 × 10⁻⁴⁶; CHD8: P < 6.9 × 10⁻⁵) (Supplementary Fig. 7, Supplementary Table 8 and Methods). NTNG1 is a strong biological candidate given its role in laminar organization of dendrites and axonal guidance and was also reported as being disrupted by a de novo translocation in a child with Ret’s syndrome, without MECP2 mutation. Both de novo mutations identified here are missense (p.Tyr23Cys and p.Thr135Ile) at highly conserved positions predicted to disrupt protein function, although there is evidence of mosaicism for the former mutation (Supplementary Table 3).

CHD8 has not previously been associated with ASD and codes for an ATP-dependent chromatin-remodelling factor that has a significant role in the regulation of both β-catenin and p53 signalling. We also identified de novo missense variants in CHD3 as well as CHD7 (CHARGE syndrome, OMIM 214800), a known binding partner of CHD8 (ref. 16). ASD has been found in as many as two-thirds of children with CHARGE, indicating that CHD7 may contribute to an ASD syndromic subtype.

We identified 30 protein-altering de novo events intersecting with Mendelian disease loci (Supplementary Table 3) as well as inherited hemizygous mutations of clinical significance (Supplementary Table 9). The de novo mutations included truncating events in syndromic intellectual disability genes (MBD5 (mental retardation, autosomal dominant 1, OMIM 156200), RPS6KA3 (Coffin–Lowry syndrome, OMIM 303600) and DYRK1A (the Down’s syndrome candidate gene, OMIM 600855)), and missense variants in loci associated with syndromic ASD, including CHD7, PTEN (macrocephaly/autism syndrome, OMIM 605309) and TSC2 (tuberous sclerosis complex, OMIM 613254). Notably, DYRK1A is a highly conserved gene mapping to the Down’s syndrome critical region (Supplementary Fig. 8). The proband here (13890) is severely cognitively impaired and microcephalic, consistent with previous studies of DYRK1A haploinsufficiency in both patients and mouse models.

Twenty-one of the non-synonymous de novo mutations map to CNV regions recurrently identified in children with developmental delay and ASD (Supplementary Table 10), such as MBD5 (2q23.1 deletion syndrome), SYNRG (17q12 deletion syndrome) and POLRMT (19p13.3 deletion). There is also considerable overlap with genes disrupted by single de novo CNVs in children with ASD (for example, NLGN1 and ARID1B; Supplementary Table 11). Given the prior probability that these loci underlie genomic disorders, the disruptive de novo SNVs and small indels may be pinpointing the possible major effect locus for ASD-related features. For example, we identified a complex de novo mutation resulting in truncation of SETBP1 (SET binding protein 1), one of five genes in the critical region for del(18)(q12.2q21.1) syndrome (Fig. 1d), which is characterized by hypotonia, expressive language delay, short stature and behavioural problems. Recurrent de novo missense mutations at SETBP1 were recently reported to be causative for a distinct phenotype, Schinzel–Giedion syndrome, probably through a gain-of-function mechanism, indicating diverse phenotypic outcomes at this locus depending on mutation mechanism.

Several of the mutated genes encode proteins that directly interact, suggesting a common biological pathway. From our full list of genes carrying truncating or severe missense mutations (126 events from all 209 families), we generated a protein–protein interaction (PPI) network based on a database of physical interactions (Supplementary Table 12). We found 39% (49 of 126) of the genes mapped to a highly significant trend towards individual’s co-morbid with intellectual disability (Supplementary Figs 1 and 6). We observed recurrent, protein-disruptive mutations in two genes: NTNG1 (netrin G1) and CHD8 (chromodomain helicase DNA binding protein 8). Given their locus-specific mutation rates, the probability of identifying two independent mutations in our sample set is low (uncorrected, NTNG1: P < 1.2 × 10⁻⁴⁶; CHD8: P < 6.9 × 10⁻⁵) (Supplementary Fig. 7, Supplementary Table 8 and Methods). NTNG1 is a strong biological candidate given its role in laminar organization of dendrites and axonal guidance and was also reported as being disrupted by a de novo translocation in a child with Ret’s syndrome, without MECP2 mutation. Both de novo mutations identified here are missense (p.Tyr23Cys and p.Thr135Ile) at highly conserved positions predicted to disrupt protein function, although there is evidence of mosaicism for the former mutation (Supplementary Table 3).

CHD8 has not previously been associated with ASD and codes for an ATP-dependent chromatin-remodelling factor that has a significant role in the regulation of both β-catenin and p53 signalling. We also identified de novo missense variants in CHD3 as well as CHD7 (CHARGE syndrome, OMIM 214800), a known binding partner of CHD8 (ref. 16). ASD has been found in as many as two-thirds of children with CHARGE, indicating that CHD7 may contribute to an ASD syndromic subtype.

We identified 30 protein-altering de novo events intersecting with Mendelian disease loci (Supplementary Table 3) as well as inherited hemizygous mutations of clinical significance (Supplementary Table 9). The de novo mutations included truncating events in syndromic intellectual disability genes (MBD5 (mental retardation, autosomal dominant 1, OMIM 156200), RPS6KA3 (Coffin–Lowry syndrome, OMIM 303600) and DYRK1A (the Down’s syndrome candidate gene, OMIM 600855)), and missense variants in loci associated with syndromic ASD, including CHD7, PTEN (macrocephaly/autism syndrome, OMIM 605309) and TSC2 (tuberous sclerosis complex, OMIM 613254). Notably, DYRK1A is a highly conserved gene mapping to the Down’s syndrome critical region (Supplementary Fig. 8). The proband here (13890) is severely cognitively impaired and microcephalic, consistent with previous studies of DYRK1A haploinsufficiency in both patients and mouse models.
The observed number of de novo events was compared with expectations based on the mutation rates estimated for each gene (Methods and Supplementary Table 8), with GRIN2B showing the highest significance (uncorrected \( P \) value <0.0002). Notably, the three de novo events observed in GRIN2B are all predicted to be protein truncating, whereas no events truncating GRIN2B were found in more than 3,000 controls (Methods).

Our analysis predicts extreme locus heterogeneity underlying the genetic aetiology of autism. Under a strict sporadic disorder–de novo mutation model, if 20–30% of our de novo point mutations are considered to be pathogenic, we can estimate between 384 and 821 loci (Methods and Supplementary Fig. 13). We reach a similar estimate if we consider recurrences from ref. 9. It is clear from phenotype and genotype data that there are many ‘autisms’ represented under the current umbrella of ASD and other genetic models are more likely in different contexts (for example, families with multiple affected
There is marked convergence on genes previously implicated in intellectual disability and developmental delay. As has been noted for CNVs, this indicates that nosological divisions may not readily translate into differences at the molecular level. We believe that there is value in comparing mutation patterns in children with developmental delay (without features of autism) to those in children with ASD.

Although there is no one major genetic lesion responsible for ASD, it is still largely unknown whether there are subsets of individuals with a common or strongly related molecular aetiology and how large these subsets are likely to be. Using gene expression, protein–protein interactions, and CNV pathway analysis, recent reports have highlighted the role of synapse formation and maintenance. We find it intriguing that 49 proteins found to be mutated here have critical roles in fundamental developmental pathways, including β-catenin and p53 signalling, and that patients have been identified with multiple disruptive de novo mutations in interconnected pathways. The latter observations are consistent with an oligogenic model of autism where both de novo and extremely rare inherited SNV and CNV mutations contribute in conjunction to the overall genetic risk. Recent work has supported a role for these interconnected pathways in neuronal stem-cell fate-determination, differentiation and synaptic formation in humans and animal models. Given that fundamental developmental processes have previously been found to underlie syndromic forms of autism, a wider role of these pathways in idiopathic ASD would not be entirely surprising and would help explain the extreme genetic heterogeneity observed in this study.

METHODS SUMMARY

Exome capture, alignments and base-calling. Genomic DNA was derived directly from whole blood. Exomes were considered to be completed when ~90% of the capture target exceeded 8-fold coverage and ~80% exceeded 20-fold coverage. Exomes for the 189 trios and 31 unaffected siblings were captured with NimbleGen EZ Exome V2.0. Reads were mapped as in ref. 4 to a custom reference genome assembly (GRC build37). Genotypes were generated with GATK unified genotyper and parallel SAMtools pipeline. Exomes for the unaffected siblings matching the pilot trios were captured and analysed as in ref. 4. Predicted de novo events were called as in ref. 4 and confirmed by capillary sequencing in all family members (for 176 of the 189 trios, this also included one unaffected sibling). Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid.

Exome read-depth CNV analysis. Reads were mapped using mrFAST and normalized reads per kilobase of exon per million mapped reads (RPKM) values normalized reads per kilobase of exon per million mapped reads (RPKM) values.

Figure 2 | Mutations identified in protein–protein interaction (PPI) networks. a, The 49-gene connected component of the PPI network formed from 126 genes with severe de novo mutations among the 209 probands. b, Proband 13844 inherits three rare gene-disruptive CNVs and carries two de novo mutations. There is marked convergence on genes previously implicated in intellectual disability and developmental delay. As has been noted for CNVs, this indicates that nosological divisions may not readily translate into differences at the molecular level. We believe that there is value in comparing mutation patterns in children with developmental delay (without features of autism) to those in children with ASD.

Although there is no one major genetic lesion responsible for ASD, it is still largely unknown whether there are subsets of individuals with a common or strongly related molecular aetiology and how large these subsets are likely to be. Using gene expression, protein–protein interactions, and CNV pathway analysis, recent reports have highlighted the role of synapse formation and maintenance. We find it intriguing that 49 proteins found to be mutated here have critical roles in fundamental developmental pathways, including β-catenin and p53 signalling, and that patients have been identified with multiple disruptive de novo mutations in interconnected pathways. The latter observations are consistent with an oligogenic model of autism where both de novo and extremely rare inherited SNV and CNV mutations contribute in conjunction to the overall genetic risk. Recent work has supported a role for these interconnected pathways in neuronal stem-cell fate-determination, differentiation and synaptic formation in humans and animal models. Given that fundamental developmental processes have previously been found to underlie syndromic forms of autism, a wider role of these pathways in idiopathic ASD would not be entirely surprising and would help explain the extreme genetic heterogeneity observed in this study.

METHODS SUMMARY

Exome capture, alignments and base-calling. Genomic DNA was derived directly from whole blood. Exomes were considered to be completed when ~90% of the capture target exceeded 8-fold coverage and ~80% exceeded 20-fold coverage. Exomes for the 189 trios and 31 unaffected siblings were captured with NimbleGen EZ Exome V2.0. Reads were mapped as in ref. 4 to a custom reference genome assembly (GRC build37). Genotypes were generated with GATK unified genotyper and parallel SAMtools pipeline. Exomes for the unaffected siblings matching the pilot trios were captured and analysed as in ref. 4. Predicted de novo events were called as in ref. 4 and confirmed by capillary sequencing in all family members (for 176 of the 189 trios, this also included one unaffected sibling). Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid.

Exome read-depth CNV analysis. Reads were mapped using mrFAST and normalized reads per kilobase of exon per million mapped reads (RPKM) values normalized reads per kilobase of exon per million mapped reads (RPKM) values.

Figure 2 | Mutations identified in protein–protein interaction (PPI) networks. a, The 49-gene connected component of the PPI network formed from 126 genes with severe de novo mutations among the 209 probands. b, Proband 13844 inherits three rare gene-disruptive CNVs and carries two de novo mutations. There is marked convergence on genes previously implicated in intellectual disability and developmental delay. As has been noted for CNVs, this indicates that nosological divisions may not readily translate into differences at the molecular level. We believe that there is value in comparing mutation patterns in children with developmental delay (without features of autism) to those in children with ASD.

Although there is no one major genetic lesion responsible for ASD, it is still largely unknown whether there are subsets of individuals with a common or strongly related molecular aetiology and how large these subsets are likely to be. Using gene expression, protein–protein interactions, and CNV pathway analysis, recent reports have highlighted the role of synapse formation and maintenance. We find it intriguing that 49 proteins found to be mutated here have critical roles in fundamental developmental pathways, including β-catenin and p53 signalling, and that patients have been identified with multiple disruptive de novo mutations in interconnected pathways. The latter observations are consistent with an oligogenic model of autism where both de novo and extremely rare inherited SNV and CNV mutations contribute in conjunction to the overall genetic risk. Recent work has supported a role for these interconnected pathways in neuronal stem-cell fate-determination, differentiation and synaptic formation in humans and animal models. Given that fundamental developmental processes have previously been found to underlie syndromic forms of autism, a wider role of these pathways in idiopathic ASD would not be entirely surprising and would help explain the extreme genetic heterogeneity observed in this study.

METHODS SUMMARY

Exome capture, alignments and base-calling. Genomic DNA was derived directly from whole blood. Exomes were considered to be completed when ~90% of the capture target exceeded 8-fold coverage and ~80% exceeded 20-fold coverage. Exomes for the 189 trios (and 31 unaffected siblings) were captured with NimbleGen EZ Exome V2.0. Reads were mapped as in ref. 4 to a custom reference genome assembly (GRC build37). Genotypes were generated with GATK unified genotyper and parallel SAMtools pipeline. Exomes for the unaffected siblings matching the pilot trios were captured and analysed as in ref. 4. Predicted de novo events were called as in ref. 4 and confirmed by capillary sequencing in all family members (for 176 of the 189 trios, this also included one unaffected sibling). Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid.

Exome read-depth CNV analysis. Reads were mapped using mrFAST and normalized reads per kilobase of exon per million mapped reads (RPKM) values normalized reads per kilobase of exon per million mapped reads (RPKM) values.
12. Nishimura-Akiyoshi, S., Niimi, K., Nakashiba, T. & Itohara, S. Axonal netrin-Gs transneuronally determine lamina-specific subdendritic segments. Proc. Natl Acad. Sci. USA 104, 14801–14806 (2007).
13. Borg, I. et al. Disruption of Netrin G1 by a balanced chromosome translocation in a girl with Rett syndrome. Eur. J. Hum. Genet. 13, 921–927 (2005).
14. Nishiyama, M. et al. CHD8 suppresses p53-mediated apoptosis through histone H1 recruitment during early embryogenesis. Nature Cell Biol. 11, 172–182 (2009).
15. Thompson, B. A., Tremblay, V., Lin, G. & Bochar, D. A. CHD8 is an ATP-dependent chromatin remodeling factor that regulates β-catenin target genes. Mol. Cell. Biol. 28, 3894–3904 (2008).
16. Batsuihi, T. et al. CHD8 interacts with CHD7, a protein which is mutated in CHARGE syndrome. Hum. Mol. Genet. 19, 2859–2866 (2010).
17. Bedogni, F. et al. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res. 1380, 42–77 (2011).
18. Moller, R. S. et al. Truncation of the Down syndrome candidate gene DYRK1A in two unrelated patients with microcephaly. Am. J. Hum. Genet. 82, 1165–1170 (2008).
19. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. Nature Genet. 43, 838–846 (2011).
20. De Ferrari, G. V. & Moon, R. T. The ups and downs of Wnt signaling in prevalent neurological disorders. Nature Genet. 42, 483–485 (2010).
21. Borg, I. et al. Delineation of a critical region on chromosome 18 for the del(18)(q12.2q21.1) syndrome. Am. J. Med. Genet. A. 146A, 1330–1334 (2008).
22. Thompson, B. A., Tremblay, V., Lin, G. & Bochar, D. A. CHD8 is an ATP-dependent chromatin remodeling factor that regulates β-catenin target genes. Mol. Cell. Biol. 28, 3894–3904 (2008).
23. Warde-Farley, D. et al. Delineation of a critical region on chromosome 18 for the del(18)(q12.2q21.1) syndrome. Am. J. Med. Genet. A. 146A, 1330–1334 (2008).
24. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. Nature Genet. 43, 838–846 (2011).
25. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. Nature Genet. 43, 838–846 (2011).
26. Bedogni, F. et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nature Genet. 42, 483–485 (2010).
27. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38, W214–W220 (2010).
28. Enright, S., Sebek, G., Ewing, R. & Koytûrük, M. DADA: Degree-aware algorithms for network-based disease gene prioritization. BioData Mining 4, 19 (2011).
29. De Ferrari, G. V. & Moon, R. T. The ups and downs of Wnt signaling in prevalent neurological disorders. Oncogene 25, 7545–7553 (2006).
30. Bedogni, F. et al. Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. Proc. Natl Acad. Sci. USA 107, 13129–13134 (2010).
31. Ilie, F. & Sommer, L. Wnt signaling: multiple functions in neural development. Cell. Mol. Life Sci. 62, 1100–1108 (2005).
32. Tedeschi, A. & Di Giovanni, S. The non-apoptotic role of p53 in neuronal biology: enlightening the dark side of the moon. EMBO Rep. 10, 576–583 (2009).

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

Acknowledgements We would like to thank and recognize the following ongoing studies that produced and provided exome variant calls for comparison: NHLBI Lung Cohort Sequencing Project (HL 1029230), NHLBI WHI Sequencing Project (HL 102924), NHGRI SNPs (HHSN273200800010C), NHLBI/NHGRI SeattleSeq (HL 094976), and the Northwest Genomics Center (HL 102926). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantinou, E. Cook, Fombonne, D. Geschwind, E. Hanson, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelsphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren and E. Wijsman). We also acknowledge Dr. A. Beaudet and the Simons Simplex Collection Genetics Consortium for providing Illumina genotyping data, T. Lehner and the Autism Sequencing Consortium for providing an opportunity for pre-publication data exchange among the participating groups. We appreciate obtaining access to phenotypic data on SAFARI Base. This work was supported by the Simons Foundation Autism Research Initiative (SFARI 137578 and 191889; E.E.E., J.S. and R.B.) and NIH HD065285 (E.E.E. and J.S.). E.B. is an Alfred P. Sloan Research Fellow. E.E.E. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions E.E.E., J.S. and B.J.O. designed the study and drafted the manuscript. E.E.E. and J.S. supervised the study. R.B., B.R. and B.J.O. analysed the clinical information. R.B., R.V., S.G., E.K., N.K. and B.P.C. contributed to the manuscript. S.G., N.K., B.P.C., A.K., C.B., M.M. and L.V. generated and analysed CNV data. B.J.O. and L.V. performed MIP resequencing and mutation validations. I.B.S., E.H.T., B.J.O. and J.S. developed MIP protocol and analysis. B.V. and R.B., and J.M.A. generated loci-specific mutation rate estimates. R.L. and E.B. performed PPI network analysis and simulations. E.K. performed DADA analysis. C.L. performed Illumina sequencing. J.D.S., I.B.S., E.H.T. and C.L. analysed sequence data. B.P.C. performed IPA analysis. B.J.O., E.K. and N.K. developed the de novo analysis pipelines and analysed sequence data. D.A.N., M.J.R., J.D.S. and E.H.T. supervised exome sequencing and primary analysis.

Author Information Access to the raw sequence reads can be found at the NCBI database of Genotypes and Phenotypes (dbGaP) and National Database for Autism Research under accession numbers phs000482.v1.p1 and NDARCOL0001878, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. 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 E.E.E. (eee@gs.washington.edu) or J.S. (shendure@uw.edu).

©2012 Macmillan Publishers Limited. All rights reserved
METHODS
Exome capture, alignments and base-calling. Exomes for the 189 trios (and 31 unaffected siblings) were captured with NimbleGen EZ Exome V2.0. Final libraries were then sequenced on either an Illumina GAIIx (paired- or single-end 76-bp reads) or HiSeq2000 (paired- or single-end 50-bp reads). Reads were mapped to a custom GRCh37/hg19 build using BWA 0.5.6 (ref. 32). Read qualities were recalibrated using GATK Table Recalibration 1.2.9025 (ref. 33). Picard-tools 1.14 was used to flag duplicate reads (http://picard.sourceforge.net/). GATK IndelRealigner 1.2.9025 was used to realign reads around insertion/deletion (indel) sites. Genotypes were generated with GATK Unified Genotyper42 with FILTER = “QUAL >= 50.0 | AB >= 0.75 | HR [ Run > 3 ] | QD < 5.0” and in parallel with the SAMTools pipeline as described previously41. Only positions with at least eightfold coverage were considered. All pilot sibling exomes were captured and analysed as described previously41. Predicted de novo events were called and compared against a set of 946 other exomes to remove recurrent artefacts and likely undecalled sites. Indels were also called with the GATK Unified Genotyper and SAMTools and filtered to those with at least 25% of reads showing a variant at a minimum depth of 8x. Mutations were phased using molecular cloning of PCR fragments, read-pair information, linked informative SNPs, and obligate carrier status. To identify rare private variants (singleton), the full variant list was compared against a larger set of 1,779 other exomes. Predicted de novo indels were also filtered against this larger set.

Sanger validations. All reported de novo events (exome or MIP capture) were validated by designing primers with BatchPrimer3 followed by PCR amplification and Sanger sequencing. We performed PCR reactions using 10 ng of DNA from father, mother, unaffected sibling (when available), and proband and performed Sanger capillary sequencing of the PCR product using forward and reverse primers. In some cases, one direction could not be assessed due to the presence of repeat elements or indels in close proximity to the mutation event.

Mutation candidate gene analysis. We examined whether each non-synonymous or CNV de novo event may be contributing to the aetiology of ASD by evaluating the likelihood deleteriousness of the change (GERP, Grantham score) and intersecting with known syndromic and non-syndromic candidate genes, CNV morbidity maps, and information in OMIM and PubMed. Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid. For genes that had not previously been implicated in ASD, we gave priority to those with structural similarities to known candidate or strong evidence of neural function or development.

Exome read-depth CNV discovery. To find CNVs using exome read-depth data, we first mapped sequenced reads to the hg19 exome using the mrsFAST aligner41. Next, we applied a novel method (N.K. et al., manuscript in preparation), which uses normalized RPKM values35 of the ∼194,000 captured exons/sequences, subsequent population normalization using 36 exomes from the Exome Sequencing Project and singular value decomposition to remove systematic bias present within exome capture reactions. Rare CNVs were detected using a threshold cutoff of the normalized RPKM values, and we required at least three exons above our threshold in order to make a call. We made a total of 1,077 deletion or duplication calls in 366 individuals (range 0–14, median = 3, mean = 2.94).

CNV detection using array CGH. A custom-targeted 2×400K Agilent chip with median probe spacing of 500bp in the genomic hotspots flanked by segmental duplications or Alu repeats and probe spacing of 14kb in the genomic backbone was designed. All experiments were performed according to the manufacturer’s instructions using NAI12878 as the female reference and NA18507 as the male reference (Coriell). Data analysis was performed following feature extraction using DNA analytics with ADM-2 setting. All CNV calls were visually inspected in the UCSC Genome Browser. CNV calls from probands were then intersected with those from parents and also with 377 controls recruited through NIMH Genetics Initiative39–41 and ClinSeq cohort42 analysed on the same microarray platform. The NIMH set of controls were ascertained by the NIMH Genetics Initiative36 through the Wellcome Trust Case Control Consortium (WTCCC) National Blood Services Cohort43–45 and filtered all CNVs present in 1% (20) of WTCC2 controls or 1% (16) of parents by 50% reciprocal overlap with matching copy number status. In addition, similar to the filtering criteria used for array CGH detection, we selected only CNVs that contained less than 50% segmental duplication and intersected with RefSeq coding sequence. To select putative de novo CNVs, we further refined the CNV not to be present in family-matched parents and siblings. Additionally, we filtered CNVs present in >0.1% (2) of the full 1,651 parent set. For total potential, rare inherited events, we required the CNV be detected in a matched parent or sibling. Finally, we filtered the genes inside each CNV under the same criteria (to account for smaller or larger CNPs) and removed CNVs with no remaining genes.

CNV cross validation. High-confidence, cross-validated de novo and inherited CNVs were selected by identifying events detected by at least two of three methodologies. To account for the variable breakpoint definitions in array CGH, SNP arrays, and exome copy number profiles, we aligned the CNVs by at least one overlapping gene ID and reported each CNV region by its maximal outer boundaries. This identified six de novo and 70 rare inherited events for further study (Supplementary Table 6).

Ingenuity pathway analysis. Ingenuity pathway analysis (IPA) was performed to identify potential functional enrichments within both our PPI (49 genes) and overall set of 126 genes. ReSeq reference gene list was used as a background list for all analysis. To confirm our results pertaining to CTNNB1 upstream enrichment, we simulated 10,000 random populations of 209 individuals using Poisson priors for each gene based on their estimated mutation rates (see below), with a global correction factor resulting in selecting a mean of 126 genes per population.

We then used this simulation data to calculate the probability of observing eight direct upstream interactors of CTNNB1 and determined that our data set is enriched for these genes with P = 0.0030.

Estimating locus-specific mutation rates. Human–chimpanzee alignments were downloaded from the UCSC Genome Browser (reference versions GRCh37 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/panTro2/syntenicNet/). The more conservative syntenicNet alignments were used (details in http://hgdownload.cse.ucsc.edu/goldenPath/hg19/panTro2/README.txt). Gene definitions were downloaded from the UCSC Table Browser, from the ReSeq Genes track, and the refFlat table. Exons were extended by 2bp, and overlapping exons were merged using BEDTools. Non-exonic sequence was not considered. For each gene, we extracted: (1) d = the number of differences between chimpanzee and human; and (2) n = the number of bases aligned. We assumed a divergence time between human and chimpanzee of 12 million years (Myr) and an average generation time of 25 years. We then calculated gene-specific mutation rates per site per generation: $r = (d/n)/(12\text{Myr}/25\text{years})$. We calculated the probability of observing X+ events using the Poisson distribution defined by the number of mutations observed on the exomes screened and the size of the coding region, including actual splice bases.

Network simulation and null model estimation. To generate a null distribution of gene mutations, de novo mutation rates were estimated from human–chimpanzee mutation rates. A pseudocount of 2.0833×10−8 (the smallest calculated in the gene set) was applied to any exon with a mutation rate of zero. To create null gene sets, genes were drawn uniformly from this background distribution. Human protein–protein interaction data were collected from GeneMANIA39–40 on 29 August 2011. Only direct physical interactions from the Homo sapiens database were considered. The list comprises approximately 1.5 million physical interactions retrieved from 150 studies. A protein interaction network was created from each experimental and null gene set by drawing edges between genes with physical interactions reported in the GeneMANIA database. Qualitatively similar results were achieved by including only interactions supported by multiple independent data sources. For each network, clustering coefficient, centralization, average shortest path length, density, and heterogeneity were determined using Cytoscape46 and Network Analyzer44. Duplicate- and self-interactions were not considered in calculating network statistics.

Disease gene prioritization based on PPI networks. We applied degree-aware algorithms to rank a set of candidate genes with respect to a set of products of genes expressed in the ASD using Protein–Protein Interaction (PPI) networks. We used the integrated human PPI network data collected from GeneMANIA39–40 on 29 August 2011. The PPI network contains 12,007 proteins with ~1.5 million direct physical interactions associated with a reliability score. We obtain the seed proteins for the ASD from the list of ref. 17. For the candidate set we used 126 gene products from the severe families selected for CNV comparisons in this study, calls were generated for 107 probands. Of these, both parents were profiled for 101 families and one parent was profiled for the remaining six families. In addition, at least one sibling was profiled for 99 of these families.

Independent of array CGH detection, to identify putatively pathogenic CNVs, we first compared our data to 2,090 control samples derived from the Wellcome Trust Case Control Consortium (WTCCC) National Blood Services Cohort43–45 and filtered all CNVs present in 1% (20) of WTCC2 controls or 1% (16) of parents by 50% reciprocal overlap with matching copy number status. In addition, similar to the filtering criteria used for array CGH detection, we selected only CNVs that contained less than 50% segmental duplication and intersected with RefSeq coding sequence. To select putative de novo CNVs, we further refined the CNV not to be present in family-matched parents and siblings. Additionally, we filtered CNVs present in >0.1% (2) of the full 1,651 parent set. For total potential, rare inherited events, we required the CNV be detected in a matched parent or sibling. Finally, we filtered the genes inside each CNV under the same criteria (to account for smaller or larger CNPs) and removed CNVs with no remaining genes.
disruptive de novo events from the pilot autism project and the current study. Given the GeneMANIA PPI network and Betancur seed gene product list, we used DADA for ranking the candidate genes. We emphasize that this ranking is not implying causality but rather relatedness to genes previously and independently associated with ASD. For testing the significance of this ranking, we rank all the gene products except the seed set using the same algorithm. On the basis of the ranking result, we applied a Mann–Whitney U rank sum test (one-tailed) on the candidate set compared to all the other genes.

**MIP protocol.** Each of 1,703 autism probands from the SSC collection and 744 controls from the NIMH collection was subjected to MIP-based multiplex capture of the six genes: SCN1A, GRIN2B, GRIN2A, LAMC3, FOXP1 and FOXP2. For each library, 50 ng of DNA was used. Individually synthesized 70 mer MIPs (n = 355) were pooled and 5’ phosphorylated with T4 PNK (NEB). Hybridization with MIPs, gap filling and ligation were performed in one step for 45–48 h at 60 °C, followed by an exonuclease treatment of 30 min at 37 °C, similar to ref. 45, with modifications for reduced MIP number (B.J.O. et al., manuscript in preparation). Amplification of the library was performed by PCR using different barcoded primers for each library. Then barcoded libraries were pooled, purified using Agencourt AMPure XP and one lane of 101-bp paired-end reads was generated for each mega-pool (~384) on an Illumina HiSeq 2000 according to manufacturer’s instructions. Raw reads were mapped to the genome as in ref. 4. MIP targeting arms were then removed and variants called using SAMtools. A 25-fold coverage, with AB allele ration <0.7, and quality 30 threshold was used for high-confident variant calling. Private (possible de novo) variants were identified by filtering against 1,779 other exomes. The parents of children with disruptive rare variants were then captured. Variants not seen or with low coverage in the parents were validated by Sanger capillary-based fluorescent sequencing. No truncating variants of

[**Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.** Bioinformatics 25, 1754–1760 (2009).]

[**Bunge, J. & Fitzpatrick, M. Estimating the number of species - a Review.** Bioinformatics 25, 1754–1760 (2009).]

[**Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq.** Nature Methods 5, 621–628 (2008).]

[**Hach, F. et al. mrsFAST: a cache-oblivious algorithm for short-read mapping.** Nature Methods 7, 576–577 (2010).]

[**Kessler, R. C. & Ustun, T. B. The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI).** Int. J. Methods Psychiatr. Res. 13, 93–121 (2004).]

[**Biesecker, L. G. et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine.** Genome Res. 19, 1665–1674 (2009).]

[**Talati, A., Fyer, A. J. & Weissman, M. M. A comparison between screened NIMH and clinically interviewed control samples on neuroticism and extraversion.** Mol. Psychiatry 13, 122–130 (2008).]

[**Baum, A. E. et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder.** Mol. Psychiatry 13, 197–207 (2008).]

[**Itsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease.** Am. J. Hum. Genet. 84, 148–161 (2009).]

[**Garrido, R. & Heyes, M. Target-enrichment strategies for next-generation sequencing.** Bioinformatics 28, 282–284 (2012).]

[**Mamanova, L. et al. Target enrichment strategies for next-generation sequencing.** Nature Methods 7, 111–118 (2010).]

[**Bunge, J. & Fitzpatrick, M. Estimating the number of species - a Review.** J. Am. Stat. Assoc. 88, 364–373 (1993).]

[**Chao, A. & Lee, S. M. Estimating the number of classes via sample coverage.** J. Am. Stat. Assoc. 87, 210–217 (1992).]