RESOURCE

nature medicine

Whole-genome sequencing of quartet families with autism spectrum disorder

Ryan K C Yuen[1], Bhooma Thiruvahindrapuram[1], Daniele Merico[1], Susan Walker[1], Kristiina Tammimies[1,2], Ny Hoang[3], Christina Chrysler[4], Thomas Nalpathamkalam[1], Giovanna Pellecchia[1], Yi Liu[1,5], Matthew J Gazzellone[1], Lia D'Abate[1], Eric Deneault[1], Jennifer L Howe[1], Richard S C Liu[1], Ann Thompson[4], Mehdi Zarrei[1], Mohammed Uddin[1], Christian R Marshall[1,6], Robert H Ring[7], Lonnie Zwaigenbaum[8], Peter N Ray[6], Rosanna Weksberg[3,9], Melissa T Carter[3,10], Bridget A Fernandez[11,12], Wendy Roberts[10], Peter Szatmari[10,13,14] & Stephen W Scherer[1,15]

Autism spectrum disorder (ASD) is genetically heterogeneous, with evidence for hundreds of susceptibility loci. Previous microarray and exome-sequencing studies have examined portions of the genome in simplex families (parents and one ASD-affected child) having presumed sporadic forms of the disorder. We used whole-genome sequencing (WGS) of 85 quartet families (parents and two ASD-affected siblings), consisting of 170 individuals with ASD, to generate a comprehensive data resource encompassing all classes of genetic variation (including noncoding variants) and accompanying phenotypes, in apparently familial forms of ASD. By examining *de novo* and rare inherited single-nucleotide and structural variations in genes previously reported to be associated with ASD or other neurodevelopmental disorders, we found that some (69.4%) of the affected siblings carried different ASD-relevant mutations. These siblings with discordant mutations tended to demonstrate more clinical variability than those who shared a risk variant. Our study emphasizes that substantial genetic heterogeneity exists in ASD, necessitating the use of WGS to delineate all genic and non-genic susceptibility variants in research and in clinical diagnostics.

© 2015 Nature America, Inc. All rights reserved.

Evidence from twin and family studies suggests substantial heritability in ASD[1]. Risk of recurrence in families is high, ranging from 12.9 (ref. 2) to 18.7% (ref. 3), suggesting that in quartet families with two (or more) affected siblings with ASD (generally referred to as multiplex families) affected children might carry the same susceptibility allele(s)[4–6]. Families with a single affected offspring with ASD (simplex families) have been characterized and used to discover highly penetrant *de novo* mutations, because sporadic genetic changes are believed to be more prevalent in ASD-simplex families than in the general population[7,8]. Earlier studies of copy-number variations (CNVs) in ASD yielded mixed results, with some studies finding differences in the *de novo* mutation rate in ASD-affected individuals between simplex and multiplex (MPX) families[9,10], while others (based on larger cohort sizes) did not[11,12].

To date, >100 ASD-susceptibility genes have been identified, mainly by microarray and exome-sequencing approaches[13,14]. Most of these investigations have been conducted in simplex cases and have focused on attributing ASD risk to the protein-coding regions of the genome. Apparent highly penetrant mutations in nongenic[15,16], noncoding RNA[17–19] and large CNV regions[10–12,20] are also known, suggesting that WGS, which detects all classes and sizes of mutations, be considered as the preferred genomic platform in studies of ASD. We and others have also previously demonstrated that WGS provides more uniform coverage in the coding regions of the genome than does exome sequencing[21,22], thereby increasing the detection rate of rare variants with clinical utility for diagnosis yields and management.

To expand the knowledge of the genetic characteristics in familial forms of ASD, we applied WGS and bioinformatics analyses to investigate an extensively phenotyped cohort of multiplex families. Such families are favored in genetic analysis studies, as it is thought that inherited genetic factors are likely to be involved in the etiology of diseases[20,23]. Our project, which yields a multitude of high-quality data, all publicly accessible for further analysis, further supports the observation that substantial genetic and clinical heterogeneity exists

[1]The Centre for Applied Genomics, Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. [2]Department of Women's and Children's Health, Pediatric Neuropsychiatry Unit, Center of Neurodevelopmental Disorders at Karolinska Institutet (KIND), Karolinska Institutet, Stockholm, Sweden. [3]Department of Pediatrics, Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [4]Department of Psychiatry and Behavioural Neurosciences, Offord Centre for Child Studies, McMaster University, Hamilton, Ontario, Canada. [5]Jinan Pediatric Research Institute, Qilu Children's Hospital of Shandong University, Shandong, China. [6]Department of Molecular Genetics, Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada. [7]Autism Speaks, Princeton, New Jersey, USA. [8]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada. [9]Department of Paediatrics and Genetics and Genome Biology Program, The Hospital for Sick Children and Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [10]Autism Research Unit, The Hospital for Sick Children, Toronto, Ontario, Canada. [11]Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada. [12]Provincial Medical Genetic Program, Eastern Health, St. John's, Newfoundland, Canada. [13]Child Youth and Family Services, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [14]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [15]Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada. Correspondence should be addressed to S.W.S. (stephen.scherer@sickkids.ca).

Received 27 October 2014; accepted 22 December 2014; published online 26 January 2015; doi:10.1038/nm.3792

in ASD and that much larger surveys will need to be completed to completely understand the genetic architecture of ASD.

RESULTS

Characteristics of subjects and WGS

We used the Complete Genomics technology to perform WGS of both parents and two affected siblings in 85 ASD multiplex families. Of these 85 families, 59 (69%) had two male children with ASD, 23 (27%) had one male and one female child with ASD, and three (4%) had two female children with ASD (**Supplementary Table 1**). These 170 children with ASD included 139 males and 31 females (4.5: 1 male-to-female ratio). All ASD subjects were diagnosed using the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule–Generic (ADOS) protocols plus clinical evaluation. We also assessed the ASD subjects with standardized measures of intelligence, language, and adaptive functioning, and we collected information on developmental, medical and family history, as well as physical measures (**Fig. 1**).

Most of the samples were karyotyped and screened for fragile X mutations[20]. We excluded families from the study if either of the affected siblings had chromosomal abnormalities or a fragile X mutation. DNA from whole blood (n = 300) or from cell lines (derived from lymphoblast) (n = 40) were used, and the biosamples are available at the National Institute for Mental Health (NIMH) Center for Collaborative Genetic Studies, the European Collection of Cell Cultures and the Autism Genetics Research Exchange.

We sequenced the genomes from 340 subjects with an average of 96.8% coverage, relative to the hg19 human genome reference sequence (**Fig. 2a** and **Supplementary Table 1**) and with an average of 56× sequence depth (**Fig. 2b** and **Supplementary Table 1**). In particular, we achieved high sequence coverage in the coding (exome) regions of the genome. On average, 99.6% of the exome was covered with at least 5× sequence depth. Likewise, 95.6% and 74.8% of the exome was covered with at least 20× and 40× sequence depth, respectively (**Fig. 2c** and **Supplementary Table 1**). The high exome coverage of WGS allows optimal variant calling and potentially captures many clinically relevant mutations, which could have been missed by other exome-sequencing technologies[21,22]. We also performed high-resolution microarray experiments to test for CNV detection accuracy.

Complete Genomics called variants as previously described using version 2.2 and 2.4 of the Complete Genomics software[24]. As we demonstrated in our previous study[25], the Complete Genomics technology detects a substantial portion of genetic variants across the complete size spectrum (**Supplementary Fig. 1**). Annotated variants are in known genes, including a description of their putative effect on the protein (frameshift, nonsynonymous, and so on). The Complete Genomics approach can identify small and mid-size insertions and deletions (indels), CNVs and structural variants (SVs) (**Supplementary Table 2**). For each nucleotide position, the accuracy of variant calling was assessed with a confidence score, which was calculated by taking into account the read depth, base-call quality values, and mapping probabilities estimated under an equal allele fraction (EAF) model. A given variant was considered fully called at a minimum varScoreEAF (confidence score) threshold of 20dB (dB is a likelihood ratio unit used by Complete Genomics) for homozygotes and 40 dB for heterozygotes. We then annotated known polymorphisms in the database SNP (dbSNP) for each variant detected, as well as for the allelic frequency from the 1000 Genomes Project[26], the Exome Sequencing Project[27], the 69 Complete Genomics public genomes[28] and the Wellderly Project[29] using the ANNOVAR software tool[30].

Genomic annotation strategy

De novo variants are generally considered more deleterious than inherited variants because they have been subjected to less-stringent evolutionary selection[31]. However, heritability data in ASD support a major role for inherited factors[1]. Genome-wide studies of inherited autosomal mutations in ASD have been mainly limited to the model of recessive inheritance, but given the increasing number of new rare variants recognized in the human population[32,33], we considered the possible impact of accumulated deleterious variants under a haploinsufficiency model (**Fig. 1**). Given the ~4:1 male-to-female sex bias in ASD and the recent progress in finding X-linked forms of ASD and intellectual disability[34,35], we also considered X-linked mutations in males, inherited from their unaffected mothers (**Fig. 1**; details of the annotation strategy and pipeline can be found in the Online Methods).

De novo mutations in multiplex families with ASD

Initially, using a *de novo* mutation detection strategy, we identified an average of 72.6 apparently spontaneous events per genome: 59.3 single-nucleotide variants (SNVs), 13.2 *de novo* small insertions and deletions (indels: <100bp) and 0.09 CNVs per genome

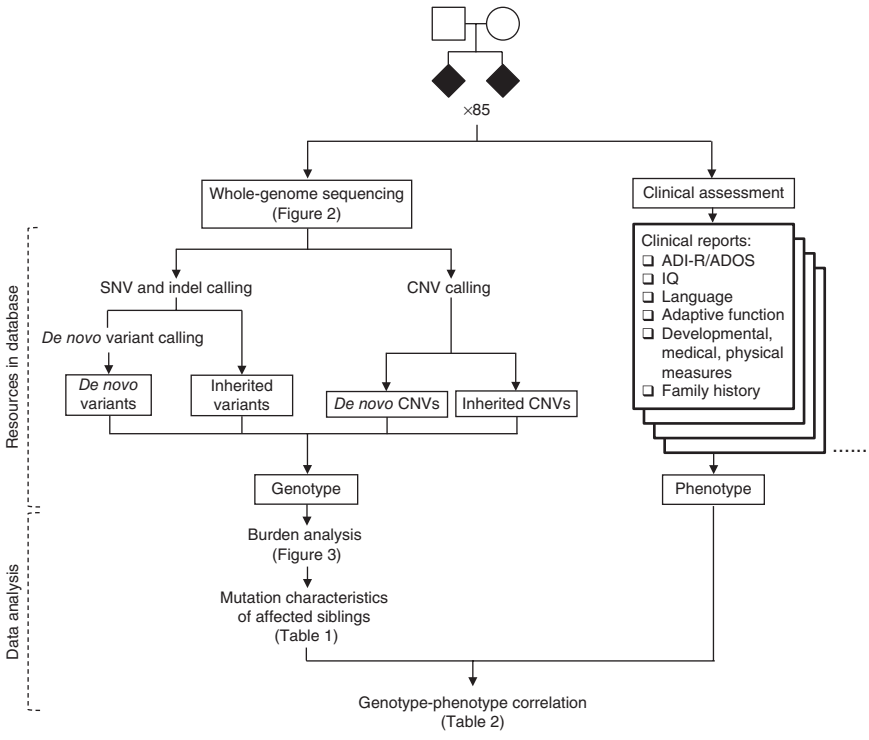

Figure 1 Schematic of genetic and phenotypic data processing in 85 multiplex families with ASD. Detailed genetic data processing can be found in **Supplementary Figure 4**.

© 2015 Nature America, Inc. All rights reserved.

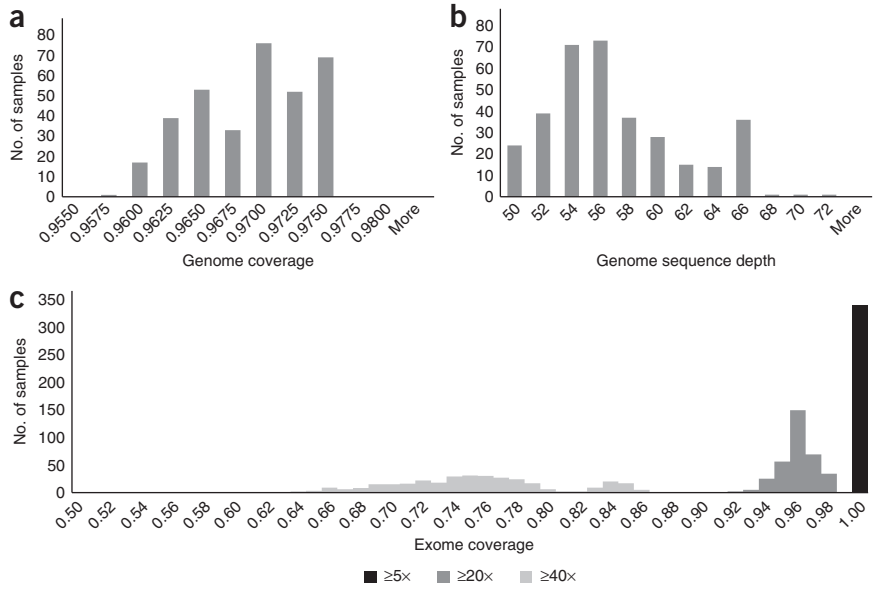

Figure 2 Quality of whole genome sequencing for all samples. (**a**) Genome sequence coverage. (**b**) Genome sequence depth. (**c**) Exome sequence coverage. Details of quality for individual samples can be found in **Supplementary Table 1**.

(**Supplementary Table 2**). We found an average of 1.02 *de novo* exonic variants per genome, similar to data previously reported in simplex families[7,8,21]. Experimental validation rates for selected *de novo* SNVs, indels and CNVs by Sanger sequencing were 90.2% (174 out of 193), 64.3% (9 out of 14) and 66.7% (10 out of 15), respectively (**Supplementary Tables 2** and **3**). Therefore, we estimated an average of 62 spontaneous events per genome. The validated *de novo* CNVs include variants with different size spectra (1kb–1Mb) (**Supplementary Table 4**). We estimated that 95.6% of the small CNVs (1kb–10kb) detected by WGS could have been missed by microarrays (**Supplementary Fig. 1**). Pairwise comparison of the number of *de novo* mutations found between two affected siblings within families showed no significant difference at the whole-genome ($P = 0.94$) or exome ($P = 0.26$) level (**Supplementary Fig. 2**), even though the number of *de novo* mutations (both SNVs and indels) correlated with paternal age ($P < 0.0001$ for SNVs and $P = 0.0084$ for indels) (**Supplementary Fig. 3**). We also found that 76% of the *de novo* mutations originated from the father, which is consistent with other reports on *de novo* mutations using simplex families[36]. Therefore, we observed no significant difference in the whole-genome *de novo* mutation rate between siblings with ASD in our multiplex families, and the rate per individual is similar to that observed in simplex families[7,8,21].

Mutation characteristics of individuals with ASD

Rare genetic variants, as well as common variants[37], are thought to contribute to ASD, but rare variants exert a larger effect than common variants on genes, which facilitates evaluation. Among all the rare (minor allele frequency or MAF≤0.01 in the 1000 Genomes Project[26] or the Exome Sequencing Project[27]) *de novo* and inherited loss-of-function (LoF) and missense mutations detected, we found a significantly higher burden in the offspring compared with parents in two out of 27 brain-related and constrained gene sets tested ($P = 0.0003$, Wilcoxon test, one-sided, FDR = 0.8% for the PhHs_MindFun_All gene set; $P = 0.004$, Wilcoxon test, one-sided, FDR = 5.7% for the NeuronalCellBody gene set) (**Fig. 3**; **Supplementary Tables 5** and **6**), whereas no significant difference of burden in common variants (MAF≥0.01; all with FDR>25% after multiple test correction) was found between offspring and parents. Many of the genes involved in the significant genesets were previously described as ASD-risk genes (such as *UBE3A* (encoding ubiquitin protein ligase E3A) and *STXBP1* (encoding syntaxin binding protein 1))[14]. The mutation burden in the first-born child with ASD was not the same as in the later-born child when compared with the parents (**Fig. 3a**,**b**). Only 29.5% of the variants within these two genesets ($n = 687$ genes for PhHs_MindFun_All and $n = 309$ genes for NeuronalCellBody) are shared between two siblings, which is significantly different from assumed complete segregation ($P < 0.0001$ for PhHs_MindFun_All and $P < 0.0001$ for NeuronalCellBody).

The observation that variants within brain-related genesets are not shared between two affected siblings from the same family suggested that either the genetic causes in our samples of siblings were heterogeneous or the gene sets analyzed were not ASD relevant. We therefore used a comprehensive medical annotation strategy[21], which included following mutation classification guidelines from the literature[38] to identify the mutations that were most likely to be relevant to ASD in the two siblings (**Supplementary Fig. 4**). For smaller variants (point mutations and indels), we considered only LoF or damaging missense mutations (as predicted by at least two of the five functional prediction algorithms) if they were *de novo* in origin, and only LoF alterations if they were inherited (**Supplementary Fig. 4**). Regarding CNVs, we considered both *de novo* and inherited exonic deletions or duplications. After filtering out all the common variants (MAF>0.01), we grouped the genes carrying putative mutations in ASD-risk genes

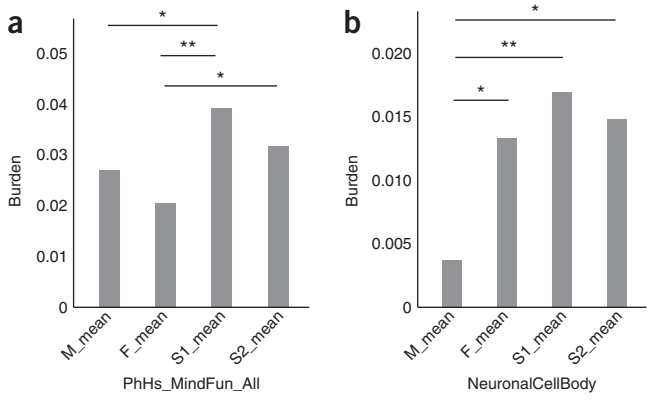

Figure 3 Mutation burden in families with two siblings with ASD. (**a**,**b**) Mutation burden between family members for the PhHs_MindFun_All (**a**) and NeuronalCellBody (**b**) gene sets. Each bar represents the exact burden value of each family member ($n = 85$ per family member). PhHs_MindFun_All comprises genes implicated in human disorders with abnormality of higher mental function in all mode of inheritance. NeuronalCellBody comprises neuronal and brain function–related genes as derived from Gene Ontology (GO:0043025 neuronal cell body; **Supplementary Table 6**). M, mother; F, father; S1, first-born child; S2, second-born child. $^*P < 0.05$, $^{**}P < 0.005$ (paired, one-sided, Wilcoxon test).

© 2015 Nature America, Inc. All rights reserved.

(termed class I, genes known to be involved in ASD[7,14]), candidate ASD-risk genes (class II, genes that have been functionally implicated in ASD), and putative ASD-risk genes (class III, novel ASD-risk genes identified by a large-scale exome-sequencing study and meta-analysis from the Autism Sequencing Consortium[7]) (**Supplementary Fig. 4**). Most of the genes involved (except class III, in which the functions of the genes are mostly unknown) are known to be X-linked or autosomal dominant (AD). Because we found no obvious biallelic rare deleterious variants, we classified the remaining mutations as being associated with genes that are involved in known AD neurodevelopmental disorders (class IV). Mutations were classified as ASD-relevant (variants potentially contributing to ASD risk) when they fell in any of the above four categories (**Table 1**).

We identified ASD-relevant mutations in 36 of 85 (42.4%) families, which is similar to the diagnostic yield reported in intellectual disability using the same sequencing platform (42%)[22]. However, we found that in only 14 of these 36 families did both affected siblings carry ASD-relevant mutations (**Table 1**). Among these 14 'genetically resolved' families, 11 (of 36, or 31%) had the same, rare, presumed-penetrant mutation in two affected siblings (**Table 1**). Ten of these shared mutations were inherited, but both ASD-affected siblings in family 2–1408 had a 1,743-bp deletion encompassing exon 18 in *SCN2A* (which probably leads to premature protein truncation) that was not found in either parent (**Fig. 4a**,**b** and **Supplementary Table 7**). Mutations in *SCN2A*, which encodes the type II α subunit of a voltage-gated sodium channel, have been described in several other ASD sequencing studies[7,8]. Another *de novo* missense mutation was found in *RELN* (which encodes reelin) in one of the affected offspring from this family (2–1408) (**Table 1**); this gene seems to act in an autosomal recessive manner and we did not find additional deleterious mutations in the other allele.

Table 1 Summary of families with clinically relevant mutations

No.	Family ID[a]	Sib1	Sib2	Gene	Changes	Type[b]	Effect	Class[c]
1	2–1408*	–04	–03	*SCN2A*	1.7kb del	DN	Deletion	I
			–03	*RELN*	p.S1985Y	DN	Missense	I
2	2–0006	–03		*STXBP1*	p.A527fs	DN	Frameshift	I
			–04	*UBE3A*	p.A619fs	MI	Frameshift	I
3	3–0107	–00		*SHANK3*	p.309_309del	DN	Frameshift	I
4	1–0234		–04	*ANK2*	p.E3429V	DN	Missense	I
5	2–0018*	–03	–04	*RAB39B*	p.E1887X	MI	Nonsense	I
6	1–0232	–03		*DMD*	31kb del	MI	Deletion	I
7	1–0273		–04	*KATNAL2*	p.Q53X	MI	Nonsense	I
		–03		*THRA*	p.R384C	DN	Missense	IV
8	2–1362*	–03	–04	*MIB1*	p.R906fs	PI	Frameshift	I
9	2–0323		–03	*MAP2K2*	c.581–2A>G	MI	Splice site	II
			–03	*KCNQ4*	c.E509X	PI	Nonsense	IV
10	2–0197*		–04	*CHRNA2*	p.Y331X	MI	Nonsense	II
		–03	–04	*RNF213*	6kb dup	MI	Duplication	IV
11	2–0309		–05	*KMT2D*	p.Q1035K	DN	Missense	II
12	2–1335		–04	*KAT6B*	88kb dup	PI	Duplication	II
13	1–0130*	–03	–04	*SCN9A*	p.V205fs	PI	Frameshift	II
14	2–1169	–04		*NLGN1*	p.V320I	DN	Missense	II
15	2–1341	–03		*LRRC7*	p.114_115del	DN	Frameshift	II
16	2–0122		–03	*KIF11*	p.R297fs	DN	Frameshift	II
17	2–0210		–04	*TENM1*	p.W1882X	MI	Nonsense	II
18	2–0223*	–03		*CACNB2*	p.V2D	DN	Missense	II
		–03	–04	*BIRC6*	p.Q1166X	PI	Nonsense	III
		–03	–04	*MYH14*	p.F446fs	MI	Frameshift	IV
19	2–0299	–03		*CD163L1*	c.2686+1G>A	MI	Splice site	III
		–03		*UTP6*	p.C333fs	MI	Frameshift	III
		–03		*RAD21*	p.F114L	DN	Missense	IV
20	2–0143*	–04	–05	*SLCO1B3*	p.598_599del	MI	Frameshift	III
21	3–0027		–01	*DNAH10*	p.L3068M	DN	Missense	III
22	2–0102	–03		*DNAH10*	p.R3963C	DN	Missense	III
23	2–0303*	–04	–03	*DNAH10*	p.R1888X	MI	Nonsense	III
24	2–0003		–04	*SLCO1B3*	p.L68fs	MI	Frameshift	III
25	2–0319*	–03	–04	*ZNF774*	p.433_438del	MI	Frameshift	III
26	1–0171		–05	*KRT24*	p.R44X	PI	Nonsense	III
27	2–0256		–04	*ZNF559*	p.359_362del	MI	Frameshift	III
28	2–0081	–03		*PCOLCE*	p.P297fs	PI	Frameshift	III
29	2–1086		–04	*ANO3*	4.5kb del	PI	Deletion	IV
30	1–0458	–04		*GARS*	p.R464fs	MI	Frameshift	IV
31	1–0339		–04	*RRM1/STIM1*	8.1kb dup	MI	Duplication	IV
		–03		*PER2*	p.A731T	DN	Missense	IV
32	2–0295	–03		*GJB6*	p.N230fs	MI	Frameshift	IV
33	1–0433	–03		*COL11A1*	p.R779X	PI	Nonsense	IV
34	1–0366*	–03	–06	*RTN2*	p.P313fs	PI	Frameshift	IV
35	1–0389*	–03	–04	*TRPV4*	p.K192fs	PI	Frameshift	IV
36	1–0160		–04	*KIF2A*	p.R723H	DN	Missense	IV

[a]Asterisks indicate families with at least one mutation shared between two affected siblings. [b]Sib, sibling; DN, *de novo*; MI, maternal inherited; PI, paternal inherited. [c]I, known ASD-risk genes; II, candidate ASD-risk genes; III, putative ASD-risk genes; IV, genes known to be associated with AD neurodevelopmental disorders.
Numbers are assigned to each sibling for identification in the family.

Apparent *de novo* events shared between siblings are not uncommon in ASD-affected families, and mechanistically they can be attributed to gonadal or germline mosaicism[20] or parental somatic mosaicism[39]. To date, most of the presumed germline mosaic events have been identified through CNV studies, but point mutations are also known[21], suggesting that they may have a role in the genetic etiology of families with multiple ASD-affected offspring. We identified and validated 21 *de novo* SNV events shared between two siblings in 16 families; none of these SNV events was in the exonic regions (**Supplementary Table 8**). Therefore, we found identical *de novo* mutations in 18.8% (16 out of 85) of the sibling pairs we investigated (**Supplementary Table 8**). Among these shared *de novo* mutations, we observed that only 55.6% of them originated in the father, compared with the 76% for the total *de novo* mutations (**Supplementary Table 8**). Although the difference is not statistically significant ($P = 0.15$) owing to the small number of shared *de novo* mutations detected, it may support the idea that the maternal germline is more mosaic[39].

Clinical features in affected individuals with ASD-relevant mutations

Affected siblings carry the same mutation in only 11 of the 36 multiplex families (31%) in which an ASD-relevant mutation was identified (**Table 1**), highlighting the genetic heterogeneity underlying

© 2015 Nature America, Inc. All rights reserved.

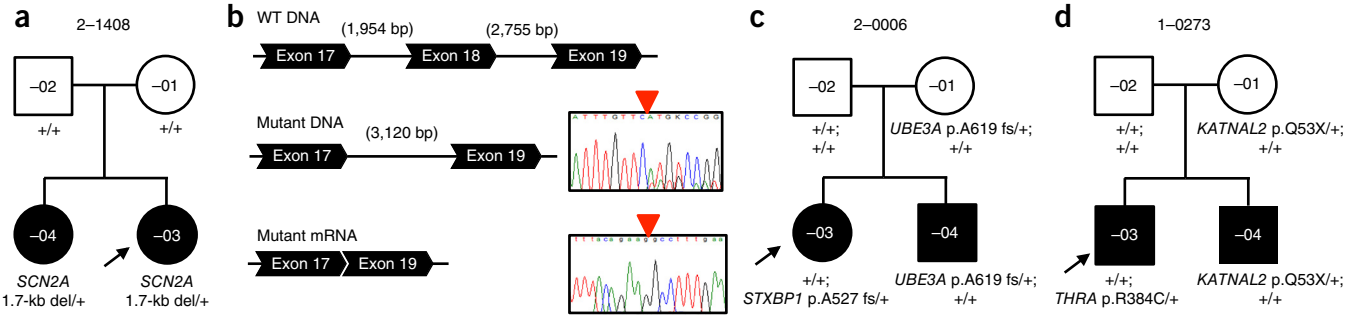

Figure 4 Families with underlying genetic etiology resolved. (**a**) Putative *de novo* 1.7-kb exonic deletion at *SCN2A* in both affected offspring. (**b**) Deletion of exon 18 in *SCN2A* and Sanger sequencing confirmation of DNA and cDNA for the deletion. Red triangles indicate the breakpoints in DNA and mRNA. (**c**) A *de novo* frameshift mutation at *STXBP1* in sib–03 and a maternally inherited frameshift (fs) mutation affecting *UBE3A* in sib–04. (**d**) A *de novo* missense mutation at *THRA* in sib–03 and a maternally inherited nonsense mutation at *KATNAL2* in sib –04. Arrows indicate the index cases in the families. WT, wild type.

ASD[11,14,20]. Indeed, we found significant differences in autism symptoms related to social and communication domains of the ADOS in siblings that probably carry different ASD-relevant mutations ($P < 0.05$), but not between siblings that share the same mutations ($P > 0.05$) (**Table 2**). This finding did not apply to repetitive stereotyped behavior (the other ASD phenotypic dimension; $P = 0.23$) or IQ ($P = 0.06$). Such heterogeneity can be further illustrated in two of the families we studied. In family 2–0006, there is a *de novo* frameshift mutation identified in *STXBP1*, a gene which is reported to be involved in both autism and epilepsy[40], in the first-born individual with ASD. However, the other sibling with ASD inherited a frameshift mutation in *UBE3A* from his unaffected mother (**Fig. 4c**). *UBE3A* is a paternally imprinted gene involved in Angelman syndrome, which may have ASD among its clinical features[41]. In family 1–0273 (**Fig. 4d**), one of the children inherited a nonsense mutation in a recently identified ASD-risk gene, *KATNAL2* (ref. 7; encodingkatanin p60 subunit A–like 2), from his mother. The other affected child had a *de novo* damaging missense mutation in *THRA*, which encodes a thyroid hormone receptor that is highly expressed in the central nervous system. *De novo* LoF mutations of *THRA* have been reported in individuals with hypothyroidism, many of whom have cognitive deficits[42,43]. Although the functional consequences remain to be confirmed, it is possible that the *de novo* missense mutation in *THRA* contributed to the pathogenesis of ASD in this individual in a dominant-negative fashion, as was recently reported for another missense mutation in individuals with developmental delay[44]. In four families, we found LoF mutations in at least one affected individual in the *KCNQ4*, *MYH14*, *GJB6* and *COL11A1* autosomal dominant hereditary deafness-related genes[45] (**Table 1**). The involvement of deafness-associated genes is further supported by the higher burden of LoF mutations in known hearing loss-related genes (derived from knockout mice studies) in children than in parents ($P = 0.009$; FDR = 2.8%; Online Methods). Given that hearing loss can be observed in individuals with ASD[46], pleiotropic effects of ASD-risk genes may add to the clinical heterogeneity seen in affected families.

DISCUSSION

This study represents the largest published WGS data set in ASD. While still small compared with what is anticipated to be required to more fully resolve all of the genetic factors involved in autism[47], several observations can be made. For example, in more than half (69.4%; 25 out of 36) of the multiplex ASD families studied here, the two affected siblings did not share the same rare penetrant ASD risk variant(s). This is partly due to the fact that 16 out of the 46 (35%) ASD-relevant mutations we identified were derived sporadically. Although it is possible that other undetected or uncharacterized combinations of rare variants not considered in this study or other common variants may be contributing, many ASD-risk genes implicated in these families are considered on their own to be sufficiently penetrant to cause ASD-relevant phenotypes[14]. We found a similar discordant rate between siblings (57.1%; 4 out of 7) when we restricted our analysis to LoF mutations in those we considered 'high–confidence' ASD-risk genes (class I genes; for example, *SHANK3* in case 3–0107–00, *DMD* in case 1–0232–03, *STXBP1* in case 2–0006–03)[14]. Similarly, analyzing a larger cohort of 2,446 ASD-affected families from our previous CNV studies[11,12,20] revealed that 50% (10 out of 20) of the pathogenic CNVs in one affected individual were not present in the other affected sibling(s) (**Supplementary Table 9**).

As we and others have shown previously[21,22], WGS technologies allow detection of all classes and sizes of mutations, many of which could have been missed by other technologies such as microarray and exome sequencing. We have further shown here that 95.6% of the small CNVs detected by WGS were not detected by high-resolution microarrays. In particular, the 1.7-kb deletion in *SCN2A*, with its small size and its breakpoints located within the introns, is difficult to detect by microarray or exome sequencing. We have also shown that WGS covered, on average, 74.8% of the coding regions with at least 40× read depth, which is higher than the 48% coverage reported in a recent exome-sequencing study[8]. With the

Table 2 Phenotype comparison of ADOS scores and IQ between siblings (gender-matched) with shared mutations and non-shared mutations in ASD-risk genes

	Shared mutations				Non-shared mutations		
Category	Sibling 1	Sibling 2	*P*	Category	Sibling 1	Sibling 2	*P*
Comm ($n = 11$)	5.00 ± 2.6	5.36 ± 2.0	0.33	Comm ($n = 15$)	4.67 ± 2.1	5.67 ± 1.8	**0.01**
Social ($n = 11$)	9.36 ± 2.3	9.27 ± 3.2	0.47	Social ($n = 15$	8.20 ± 2.5	9.67 ± 3.2	**0.04**
Soccom ($n = 11$)	14.36 ± 4.2	14.64 ± 4.9	0.44	Soccom ($n = 15$)	12.87 ± 3.9	15.33 ± 4.3	**0.02**
Play ($n = 8$)	1.88 ± 1.2	2.63 ± 1.3	0.14	Play ($n = 13$)	1.46 ± 1.1	1.92 ± 1.6	0.15
Behav ($n = 11$)	4.20 ± 1.5	3.80 ± 2.2	0.28	Behav ($n = 15$)	3.80 ± 1.9	3.33 ± 1.9	0.20
IQ ($n = 7$)	67 ± 24.4	64 ± 29.2	0.41	IQ ($n = 9$)	92 ± 12.7	110 ± 27.3	0.06

Comm, Communication; social, Reciprocal Social Interaction; soccom, Communication and Social Interaction; behav, Stereotyped Behaviors and Restricted Interests. Values are means ± sd (paired, one-sided *t*-test).

© 2015 Nature America, Inc. All rights reserved.

availability of such a high-quality WGS resource, other analyses such as integrating genomic sequences with transcriptomic[48] and epigenomic[49] data, and additional annotation using new tools or databases to characterize the noncoding genome[15] will be possible in the future.

Several approaches for molecular diagnostic testing in ASD have been contemplated, including using microarrays, targeted gene sequencing and exome sequencing[13]. We and others have previously reported that some highly penetrant CNVs, such as those affecting the *PTCHD1* (encoding patched domain containing 1) and 16p11.2 loci, did not necessarily segregate between siblings with ASD[10,17,50]. By using WGS, we can now often detect other ASD–relevant mutations in multiplex families, thereby providing a more complete description of the mutational characteristics involved. Although our current study is limited by the number of families examined, our data indicate that the substantial genetic heterogeneity of ASD (both between and within families) necessitates that a full assessment of each individual's genome be performed when determining the role of genetic factors in risk- or health-management strategies. These WGS data represent an important first step in a much larger initiative to sequence the genomes of the members of thousands of other families with ASD.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data has been deposited at the European Genome–phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001001023. The data, as part of a larger autism whole-genome sequencing project, will also be hosted in the MSSNG database on Google Genomics (for access see http://www.mss.ng/researchers). No computer source code is provided.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S.S. Gross, M. Bookman and D. Glazer from Google for the development of the MSSNG database. We thank the families for their participation in the study and we thank The Centre for Applied Genomics for informatics support. We would also like to acknowledge the Autism Sequencing Consortium (AASC) principal investigators for the use of data. This work was supported by Autism Speaks (S.W.S., R.K.C.Y.), Autism Speaks Canada (S.W.S.), Neurodevelopment Network (NeuroDevNet) (S.W.S.), the Canadian Institutes for Advanced Research (S.W.S.), the University of Toronto McLaughlin Centre (S.W.S.), Genome Canada and Ontario Genomics Institute (S.W.S.), the government of Ontario (S.W.S., P.S.), the Canadian Institutes of Health Research (S.W.S., P.S.), and The Hospital for Sick Children Foundation (S.W.S.). R.K.C.Y. holds the Autism Speaks Meixner Postdoctoral Fellowship in Translational Research and a NARSAD Young Investigator award. K.T. holds a fellowship from the Swedish Research Council. E.D. holds the Banting Postdoctoral Fellowship. P.S. holds the Patsy and Jamie Anderson Chair in Child and Youth Mental Health. S.W.S. holds the GlaxoSmithKline–Canadian Institutes of Health Research (CIHR) Chair in Genome Sciences at the University of Toronto and The Hospital for Sick Children.

AUTHOR CONTRIBUTIONS

R.K.C.Y. and S.W.S. conceived and designed the experiments. R.K.C.Y., B.T., D.M. and T.N. processed and analyzed the whole genome sequencing data. S.W. and K.T. designed and performed experiments for variants characterization and validation. N.H., C.C., J.L.H. and A.T. collected phenotypic information from the participants. G.P., Y.L., M.J.G., L.D., E.D., R.S.C.L., M.Z., M.U. and C.R.M. helped perform different components of whole genome sequencing analysis and validation experiments. R.K.C.Y., R.H.R. and S.W.S. conceived and coordinated the project. L.Z., P.N.R., R.W., M.T.C., B.A.F., W.R. and P.S. recruited, diagnosed and examined the recruited participants. R.K.C.Y. and S.W.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ronald, A. & Hoekstra, R.A. Autism spectrum disorders and autistic traits: a decade of new twin studies. *Am. Med. Genet. B Neuropsychiatr. Genet.* **156B**, 255–274 (2011).
2. Sandin, S. *et al.* The familial risk of autism. *J. Am. Med. Assoc.* **311**, 1770–1777 (2014).
3. Ozonoff, S. *et al.* Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics* **128**, e488–e495 (2011).
4. Toma, C. *et al.* Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Mol. Psychiatry* **19**, 784–790 (2014).
5. Zhao, X. *et al.* A unified genetic theory for sporadic and inherited autism. *Proc. Natl. Acad. Sci. USA* **104**, 12831–12836 (2007).
6. Constantino, J.N. *et al.* Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. *Mol. Psychiatry* **18**, 137–138 (2013).
7. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
8. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
9. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
10. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
11. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
12. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
13. Devlin, B. & Scherer, S.W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).
14. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42–77 (2011).
15. Xiong, H.Y. *et al.* The splicing code reveals new insights into the genetic determinants of disease. *Science* (in press) (2014).
16. Bae, B.I. *et al.* Evolutionarily dynamic alternative splicing of *GPR56* regulates regional cerebral cortical patterning. *Science* **343**, 764–768 (2014).
17. Noor, A. *et al.* Disruption at the *PTCHD1*locus on Xp22.11 in autism spectrum disorder and intellectual disability. *Sci. Transl. Med.* **2**, 49ra68 (2010).
18. Kerin, T. *et al.* A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci. Transl. Med.* **4**, 128ra140 (2012).
19. Ghahramani Seno, M.M. *et al.* Gene and miRNA expression profiles in autism spectrum disorders. *Brain Res.* **1380**, 85–97 (2011).
20. Autism Genome Project Consortium. *et al.* Mapping autism-risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319–328 (2007).
21. Jiang, Y.H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
22. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
23. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493–507 (1999).
24. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19**, 279–292 (2012).
25. Pang, A.W., Macdonald, J.R., Yuen, R.K., Hayes, V.M. & Scherer, S.W. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3* **4**, 63–65 (2014).
26. 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
27. Exome Variant Server. (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA) ⟨http://evs.gs.washington.edu/EVS/⟩.
28. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
29. *Wellderly Study.* ftp://stsi-ftp.sdsc.edu/pub/wellderly/.
30. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
31. Veltman, J.A. & Brunner, H.G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
32. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
33. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
34. Piton, A. *et al.* Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia. *Mol. Psychiatry* **16**, 867–880 (2011).
35. Lim, E.T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235–242 (2013).

© 2015 Nature America, Inc. All rights reserved.

36. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
37. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum. Mol. Genet.* **21**, 4781–4792 (2012).
38. Richards, C.S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008).
39. Campbell, I.M. *et al.* Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am. J. Hum. Genet.* **95**, 173–182 (2014).
40. Campbell, I.M. *et al.* Novel 9q34.11 gene deletions encompassing combinations of four Mendelian disease genes: *STXBP1*, *SPTAN1*, *ENG*, and *TOR1A*. *Genet. Med.* **14**, 868–876 (2012).
41. Kishino, T., Lalande, M. & Wagstaff, J. *UBE3A/E6-AP* mutations cause Angelman syndrome. *Nat. Genet.* **15**, 70–73 (1997).
42. Bochukova, E. *et al.* A mutation in the thyroid hormone receptor alpha gene. *N. Engl. J. Med.* **366**, 243–249 (2012).
43. van Mullem, A. *et al.* Clinical phenotype and mutant TRα1. *N. Engl. J. Med.* **366**, 1451–1453 (2012).
44. Moran, C. *et al.* Resistance to thyroid hormone caused by a mutation in thyroid hormone receptor (TR)α1 and TRα2: clinical, biochemical, and genetic analyses of three related patients. *Lancet Diabetes Endocrinol.* **2**, 619–626 (2014).
45. Smith, R.J.H., Shearer, A.E., Hildebrand, M.S. & Van Camp, G. Deafness and Hereditary Hearing Loss Overview. in *GeneReviews* (eds. Pagon, R.A., *et al.*) (University of Washington, 1993).
46. Rosenhall, U., Nordin, V., Sandstrom, M., Ahlsen, G. & Gillberg, C. Autism and hearing loss. *J. Autism Dev. Disord.* **29**, 349–357 (1999).
47. Buxbaum, J.D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2012).
48. Uddin, M. *et al.* Brain-expressed exons under purifying selection are enriched for *de novo* mutations in autism spectrum disorder. *Nat. Genet.* **46**, 742–747 (2014).
49. Numata, S. *et al.* DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.* **90**, 260–272 (2012).
50. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).

© 2015 Nature America, Inc. All rights reserved.

ONLINE METHODS

Samples for whole-genome sequencing. From a cohort of Canadian ASD families, we selected 85 unrelated families with at least two children with ASD. Both parents and two children with ASD were recruited, with selection based on availability of genomic DNA from whole blood and completeness of phenotype information. We recruited additional siblings and members of the extended family across generations whenever possible. We obtained informed consent from all participants, as approved by the Research Ethics Boards at The Hospital for Sick Children, McMaster University and Memorial Hospital. We genotyped all samples on high-resolution microarray platforms for the detection of copy number variants (CNVs).

Whole-genome sequencing (WGS). We sequenced family quartets (two parents and two ASD-affected children). Genomic DNA extracted from blood- or lymphoblast-derived cell lines (LCLs) was assessed for quality by PicoGreen and gel electrophoresis, and then sequenced by Complete Genomics (Mountain View, CA) as previously described[28]. At least 10 μg of non-degraded DNA was provided for WGS. Complete Genomics performed additional quality controls, including DNA quality assessment, sex check and comparison of samples with results from 96-SNP genotyping assay to avoid sample mix-up. The Complete Genomics Analysis Platform employs high-density DNA nanoarrays that are populated with 35-base, mate-paired reads, generated from the ends of approximately 500-bp genomic fragments biochemically engineered into DNA nanoballs[28]. Base identification was performed using a non-sequential, unchained read technology known as combinatorial probe–anchor ligation[28].

Detection of *de novo* SNVs and indels. We extracted SNVs and indels reported in 'var' format using cgatools testvariants (http://cgatools.sourceforge.net/) and compared each variant from the children to the sequence in the same position in the parents. A variant inconsistent with Mendelian inheritance (present in the offspring but not in either parent or the sibling), was considered to be a potential *de novo* mutation for that child. For an autosomal variant, we considered it to be a potential *de novo* mutation if there was a heterozygous alternative genotype in the offspring, and homozygous reference genotypes in both parents. For an X-linked variant, we considered male and female offspring with different criteria: in non-pseudoautosomal regions in male offspring, we considered a variant to be a *de novo* if there was a hemizygous alternative genotype in the offspring and a homozygous reference genotype in the mother; X-linked variants in female offspring and X-linked variants in pseudoautosomal regions in male offspring were considered as for autosomal regions. We considered a Y-linked variant to be *de novo* when a hemizygous alternative variant was present in the male offspring but absent at the same position in the father.

To the list of all apparent *de novo* variants, we applied a further systematic filter matrix to remove potential false positive calls, as previously described[21]. We applied the following quality filters for each variant: (i) varQuality of allele1 and allele2 is either VQHIGH (for v2.2) or PASS (for v2.4); (ii) ploidy of the child = 2 (or = 1 for X- and Y-linked variants in male subjects) and the ploidy of both parents is not 'N'; (iii) the ratio of sequence reads supporting the alternative call to that of the reference call is 0.3–0.7 (or ≥0.7 for X- and Y-linked variants in male subjects); (iv) the variant call does not overlap with known regions of segmental duplication; (v) the refscore (likelihood of the region being the same as the reference sequence) in both parents is >40 or '–'; (vi) the variant call does not overlap with any variants found in Complete Genomics public genomes; (vii) the variant call has frequency <0.01 in the 1000 Genomes Project; (viii) the SNV call in the child does not overlap with any variant call (SNV or indel) in either parent; and (ix) variants clustered within a distance of 100 bp have been eliminated.

Lastly, we selected only the *de novo* mutations above an optimal varScoreEAF, determined by applying the above filter matrix to an in-house control trio sequenced by Complete Genomics. We sequenced the offspring in this control trio twice at the same time (i.e. a sample replicate control) and considered variant calls that were concordant between the two samples to be true positive calls, and those that were discordant to be false positive calls. We found an optimal varScoreEAF at 175, which gave false positive rates of 3–7% and 4% for *de novo* SNVs and *de novo* indels, respectively; and false negative rates of 5–11% and 61% for *de novo* SNVs and *de novo* indels, respectively (**Supplementary Fig. 5**). Therefore, we eliminated all the variant calls with varScoreEAF less than 175 at allele1 (or at both allele1 and allele2 for insertions in regions with ploidy = 2).

Among the 170 offspring DNA samples, we detected 13,238 putative *de novo* SNVs and 2,273 putative *de novo* indels (**Supplementary Table 2**). From the distribution of number of *de novo* SNVs, we found that the majority of the DNA samples had fewer than 100 *de novo* SNVs (**Supplementary Fig. 6a**); nine samples had more than 100 *de novo* SNVs identified (**Supplementary Fig. 6a**), one of which had more than 1,500 apparent *de novo* SNVs (**Supplementary Table 2**). The nine samples with high numbers of *de novo* SNVs were all coming from LCL samples (**Supplementary Fig. 6b**,**c**). Deviation of allelic ratio from 50% between alternative and reference calls (proportion reference) can be applied to filter out *de novo* SNVs induced in somatic cell lines given that they tend to be mosaic[51]. Because we applied similar criteria in our filters for *de novo* variant detection (variants with allelic ratios less than 30% or more than 70% were removed), yet were unable to eliminate such induced *de novo* SNVs, we decided to exclude all *de novo* variant calls from the 30 LCL samples for subsequent statistical analyses. When clinically relevant mutations were detected in these LCL samples, we validated with DNA from wholeblood of the same individual.

Inheritance state determination and phasing. We applied the inheritance state analysis algorithm developed by Roach *et al.*[52]. The algorithm resolves contiguous blocks of SNVs into one of the four Mendelian inheritance states using a Hidden Markov Model: paternal identical, maternal identical, identical and non-identical. We then resolved the phase of heterozygous variant calls in the children by the assigned inheritance state of the contiguous blocks of SNVs. For phasing of *de novo* mutations, we extracted all the surrounding heterozygous variants that form a haplotype with a particular *de novo* mutation (i.e. with the same Haplink ID that derived from local *de novo* assembly). The parent of origin for the *de novo* mutation with a Haplink ID will then be assigned according to the inheritance states of the surrounding phased heterozygous variants. For all the non–cell line samples, 839 out of 8,300 *de novo* mutations were phased.

Validation of SNVs and indels. We used Primer 3 to design primers to span at least 100 bp upstream and downstream of a putative variant, avoiding regions of repetitive elements, segmental duplication or known SNPs. We randomly selected putative *de novo* mutations from the whole genome of one offspring (2–1292–003) in a quartet family, and validated all the exonic *de novo* and ASD-relevant variants by Sanger sequencing, using DNA from whole blood if available. Candidate regions were amplified by PCR for all quartets, as well as for other family members (if available) and assayed by Sanger sequencing.

Manual curation of potential *de novo* exonic variants. Although we considered the X-linked variants in male offspring for *de novo* mutations from our pipeline, the estimated false negative rates for detection of *de novo* SNV (5–11%) and indels (61%) suggested that our algorithm could have missed some true *de novo* variants. We therefore manually curated potential *de novo* exonic variants, as previously described[21]. Considering all the putative exonic variants, we identified and validated three additional *de novo* mutations (*LRRC7*p.114_115del, *RAD21*p.F114L and *SHANK3*p.309_309del) (**Supplementary Table 3**).

Assessment of CNVs by genomic array and WGS. To assess the quality of the CNVs called by Complete Genomics, we compared the concordance of CNV calls between the two sample replicates. We found a high concordance of calls the by read depth method (94%) but not by the paired-end method (73.2%). We found that the CNV calls made by the paired-end method achieved optimal false positive (19~30%) and false negative (27~31%) rates when the variant calls were supported by 20 or more mate-pair reads (**Supplementary Fig. 7**); therefore, we eliminated all the CNV calls supported by fewer than 20 mate-pair reads. We then combined the CNVs recognized by the two methods, in order to list non-redundant CNVs that cover most of the size spectrum. For any CNV with overlap between the two methods of at least 1bp, we calculated the percentage of reciprocal overlap and used the breakpoints given by the paired-end method.

The CNVs from Complete Genomics were then compared with the CNVs detected by the CytoScan HD Array. For the 40 quartet families assayed by both WGS and microarray, we found that 61.1% of the CNVs called by microarray were detected by WGS, whereas only 5.9% of those called by WGS were

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/nm.3792

detected by microarray (**Supplementary Fig. 1**). The overlap between CNVs from Complete Genomics WGS and microarray was consistent with our previous findings[25]. We also found that 93.3% of the all the summarized CNVs from Complete Genomics overlapped (50% reciprocal) with the gold-standard set of CNVs from DGV[53], further suggesting that the CNVs included in the present study are of high quality. One sample, 2–0142–003, had an abnormally high number of CNV calls (five times higher than the average) and was removed from the subsequent analysis.

Detection of CNVs by genomic array and WGS. DNA samples from 40 of the 85 quartet families were run in parallel using the CytoScan HD Array (Affymetrix, Santa Clara, CA). The microarray consists of about 2.67 million probes from across the whole genome. We called CNVs using a combination of four algorithms: Chromosome Analysis Suite software (Affymetrix, Santa Clara, CA) iPattern[12], Nexus (BioDiscovery Inc., CA), and Partek (Partek Inc., St. Louis, MO). A call was considered to be confident when two of the four algorithms made the same CNV call with 50% reciprocal overlap.

We detected CNVs from WGS using two different approaches: read depth and paired-end, as previously described[25]. We extracted variants detected by the paired-end method for deletion, distal and tandem duplications from the SV/highConfidenceSVEventsBeta file. We also extracted the variants detected by read depth method for gain and loss from the CNV/cnvSegmentsDiploidBeta file. The read depth method by Complete Genomics was based on deviation from expectation of the sequence depth in a diploid baseline reference genome using 2-kb, GC-corrected windows with a hidden Markov model. Therefore, the CNVs called were mostly larger than 2kb. In contrast, the paired-end method by Complete Genomics employed junction detections on uniquely mapped discordant mate-pairs, Therefore, it can detect CNVs below 2kb and compensate for the size limitation from the read depth method.

Detection of *de novo* CNVs. We assessed transmission of a CNV by determining whether the same CNV was present in the summarized CNV list from both the offspring and his/her parents. 'Putative *de novo* CNV' was assigned when the CNV detected in the offspring was not present in either of the parents.

To filter all the deletions and duplications to improve the accuracy of *de novo* CNVs detected, we: (i) removed all the CNVs present (with 50% reciprocal overlap) in the Complete Genomics baseline calls, (ii) eliminated the CNV calls with a Complete Genomics structural variant (SV) event frequency >0, (iii) removed all the CNVs with ref Score <20 at the same region in either parents, (iv) required all the CNVs to have gcCorrected Coverage ≥10 at the same region in both parents, (v) required the region covered by the CNV calls in the offspring to have gcCorrectedCoverage less than 0.75 times for deletions and more than 1.25 times for duplication relative to that of both parents, (vi) required the ploidy to be 1 for deletions and more than 2 for duplications, (vii) removed all the CNV calls that had more than 50% of the size covered by known segmental duplication, (viii) removed any CNV calls present in another individual from an unrelated family, (ix) removed the calls where the same CNV was called in the sibling, but in low quality (filtered by the above criteria), and (x) removed the calls for which the region covered by the CNV had a common CNV in DGV[53].

Burden analysis of rare variants for brain-related gene sets. Variants were filtered to retain only the high quality (as described above) and rare ones that overlap coding regions or essential splice sites (2 intronic bp of intron–exon boundaries) of the genes. Rare variants were defined as not exceeding 1% allele frequency based on the 1000 Genomes, NHLBI–ESP exomes and two private Complete Genomics control datasets provided by CGI; we used both the global and ethnic group specific allele frequencies (1000 Genomes: Caucasian, Easter Asian, Latin American, African; NHLBI–ESP: Caucasian, African-American). Variants were further categorized as LoF (i.e. stop-gain, frameshift or essential splice site alterations) and missense damaging (missense predicted damaging by at least two of these six criteria: SIFT[54] ≤ 0.05, PolyPhen2 (ref. 55) ≥ 0.95, Mutation Assessor[56] ≥ 2, placental mammal PhyloP[57] ≥ 2.4, vertebrate PhyloP[57] ≥ 4, CADD[58] Phred score ≥ 15).

We curated 27 gene sets representing brain expression, brain function, neurodevelopmental phenotypes and evolutionarily conserved genes (**Supplementary Table 5**). These gene sets were tested for higher burden of rare damaging variants in siblings compared with parents (paired, one-sided, Wilcoxon and Student's *t*-test), (ii) in siblings compared with the mother (paired, one-sided, Wilcoxon and Student's *t*-test), (iii) in siblings compared with the father (paired, one-sided, Wilcoxon and Student's *t*-test), (iv) in the elder sibling compared with the younger one (paired, two-sided, Wilcoxon and Student's *t*-test).

In particular, we tested the ratio of damaging variants in the gene set compared with all coding genes; when testing siblings and parents, we used the average ratio for the two subjects; subjects were paired by family membership, and all families included one father, one mother and two siblings. Wilcoxon and *t*-test results tended to agree, and we finally filtered results using the Wilcoxon *P* value of the 'siblings compared to parents' test. Multiple test correction was performed using the Benjamini–Hochberg FDR. Mean, median and percentiles (10%, 25%, 75%, 90%) of the variant ratios for each of the tested subject groups (mother, father, proband and sibling) were also reported (**Supplementary Table 6**). We have also performed the same comparisons using additional three gene sets that are related to hearing loss (1. Hs_earPheno_ADX: human abnormality of the ear phenotype with autosomal dominant/X-linked inheritance HP:0000598, 271 genes; 2. Hs_earPheno_all: human abnormality of the ear phenotype with all modes of inheritance HP:0000598, 664 genes; 3. Mm_earPheno: mouse hearing/vestibular/ear phenotype MP:0005377, all inheritance modes multigenic, 427 genes). Hearing loss genes known from knockout mice were significantly enriched in the children with ASD compare to their parents ($P = 0.009$; FDR = 2.8%), but not in the known human hearing loss genes ($P = 0.21$; FDR = 21% for Hs_earPheno_ADX and $P = 0.18$; FDR = 21% for Hs_earPheno_all).

Characterization of rare variants. To assess the rare inherited SNVs and indels, we used the population frequency of the variants as a filter to differentiate putative deleterious alterations (which would be rare in frequency) from the probable benign events (more common in frequency) (**Supplementary Fig. 4**). We define a rare variant as one that is present in <1% of the population from each of the databases used. For rare inherited CNVs, we computed the population frequency of each CNV with at least 50% reciprocal overlap from among the 54 unrelated Complete Genomics public genomes. We removed all the WGS-defined CNVs that overlapped those in public genomes and those with a SV event baseline frequency from the public genomes. We removed all CNVs that encompassed more than 50% of a known segmental duplication. We also removed CNVs that were inherited from parents (likely to be common CNVs), and recurrent CNVs that presented in more than two unrelated families within our WGS cohort.

We prioritized deleterious variants for further characterization. We defined rare deleterious variants as: all rare (≤1% minor allele frequency) LoF mutations (nonsense, splice site and frameshift mutations), and damaging missense mutations that are *de novo*. We define mutations as damaging *de novo* missense if they fulfilled two of the following five conditions (less-stringent thresholds were applied than that used in burden analysis in order to increase sensitivity, given that *de novo* variants are generally considered to be more deleterious): (i) SIFT[54] score for predicted deleterious effect with values ≤0.05; (ii) Polyphen2 (ref. 55)(HDIV) score for predicted possibly damaging effect with values ≥0.453; (iii) Mutation Assessor[56] score for predicted damaging effect with values ≥2; (iv) Phred-transformed CADD[58] score for top 1% of predicted deleterious effect with value ≥20; (v) PhyloP[57] nucleotide–level conservation (inferred from 100 vertebrate genomes) ≥2. We also included the rare CNVs (deletion and duplication) that overlapped with the coding regions of the human genome.

Medical annotation and family analysis. To evaluate the medical relevance of selected genetic variants to ASD, we developed the assessment protocol outlined in **Supplementary Figure 4**: (i) to predict whether they are likely to have a deleterious effect on splicing of the gene or functional properties of the protein product; (ii) to assess the frequency of predicted deleterious variants in a population database; (iii) to compare genes affected to known ASD candidate genes; (iv) to assess segregation in families and comparing to the Online Mendelian Inheritance in Man (OMIM) database[43] and (v) to compare to the Mouse Genome Informatics (MGI) phenotype database[59]. Details of each step have been described previously[21].

We examined the function of genes with rare deleterious mutations in the documented neurodevelopmental/behavioral phenotype of human or mouse from the human phenotype ontology (HPO)[60] and the MGIdatabases[59].

© 2015 Nature America, Inc. All rights reserved.

The mode of inheritance for the genes can also be assessed by the information provided in available online resources. For human phenotypes, we imported the information from HPO for mode of inheritance (AD, autosomal dominant; XL, X-linked; AR, autosomal recessive). For mouse phenotypes, the mode of inheritance can be inferred on the basis of the genotype associated with the phenotype of interest: (i) autosomal heterozygous genotypes were inferred to be dominant, (ii) autosomal homozygous genotypes were inferred to be recessive and (iii) genes with human orthologs on the X chromosome outside of the pseudoautosomal regions were inferred to be X linked. The confirmation of mode of inheritance was further evaluated by comparing the information from Clinical Genomics Database[61].

After thorough evaluation using the algorithm of analysis outlined above, the genes carrying deleterious mutations were classified as: (i) known, (ii) candidate, (iii) putative ASD-risk genes; or (iv) genes for autosomal dominant (AD) diseases (**Table 1** and **Supplementary Fig. 4**). The known ASD-risk genes were determined from to a list of 120 expert-annotated ASD-risk genes that is regularly updated in our analyses[11,12,14], we also included the high-confidence (top 33 ASD-risk genes with FDR < 0.1) ASD-risk genes identified from the latest results from Autism Sequencing Consortium (ASC) exome-sequencing study[7]. The candidate genes were those known to be associated with other related neuropsychiatric conditions, such as intellectual disability, schizophrenia and epilepsy. Their relevance to ASD was also evaluated on the basis of the phenotypes possessed in the knockout mice experiments (information from MGI or reported in the literature). The putative ASD-risk genes were those found as significantly enriched in the ASC exome–sequencing study (ASD-risk gene with $0.1 < \text{FDR} < 0.3$)[7]. For the genes responsible for autosomal dominant diseases, the mode of inheritance was determined as described above. We assessed the relevance of the associated autosomal dominant disease to ASD by whether it has any known neurodevelopmental dysfunction.

51. Schafer, C.M. *et al.* Whole-exome sequencing reveals minimal differences between cell line– and whole blood–derived DNA. *Genomics* **102**, 270–277 (2013).
52. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
53. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. & Scherer, S.W. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
54. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
55. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
56. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8**, R232 (2007).
57. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
58. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
59. Blake, J.A. *et al.* The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.* **42**, D810–D817 (2014).
60. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
61. Solomon, B.D., Nguyen, A.D., Bear, K.A. & Wolfsberg, T.G. Clinical genomic database. *Proc. Natl. Acad. Sci. USA* **110**, 9851–9855 (2013).

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/nm.3792