LETTER

doi:10.1038/nature11011

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

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

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

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

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

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

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

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

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

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

Table 1 | Distribution of events per family

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

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

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

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

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

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

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

Table 2 | Rates of mutation annotation given variant type

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

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

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

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

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

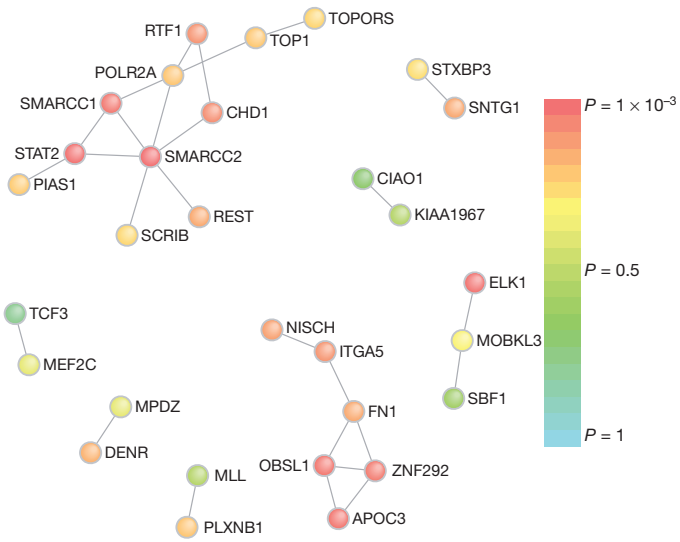


Figure 1 | Protein–protein interaction for genes with an observed functional *de novo* event. Direct protein connections from InWeb, restricting to genes harbouring *de novo* mutations for DAPPLE analysis. Two extensive networks are identified: the first is centred on SMARCC2 with 12 connections across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

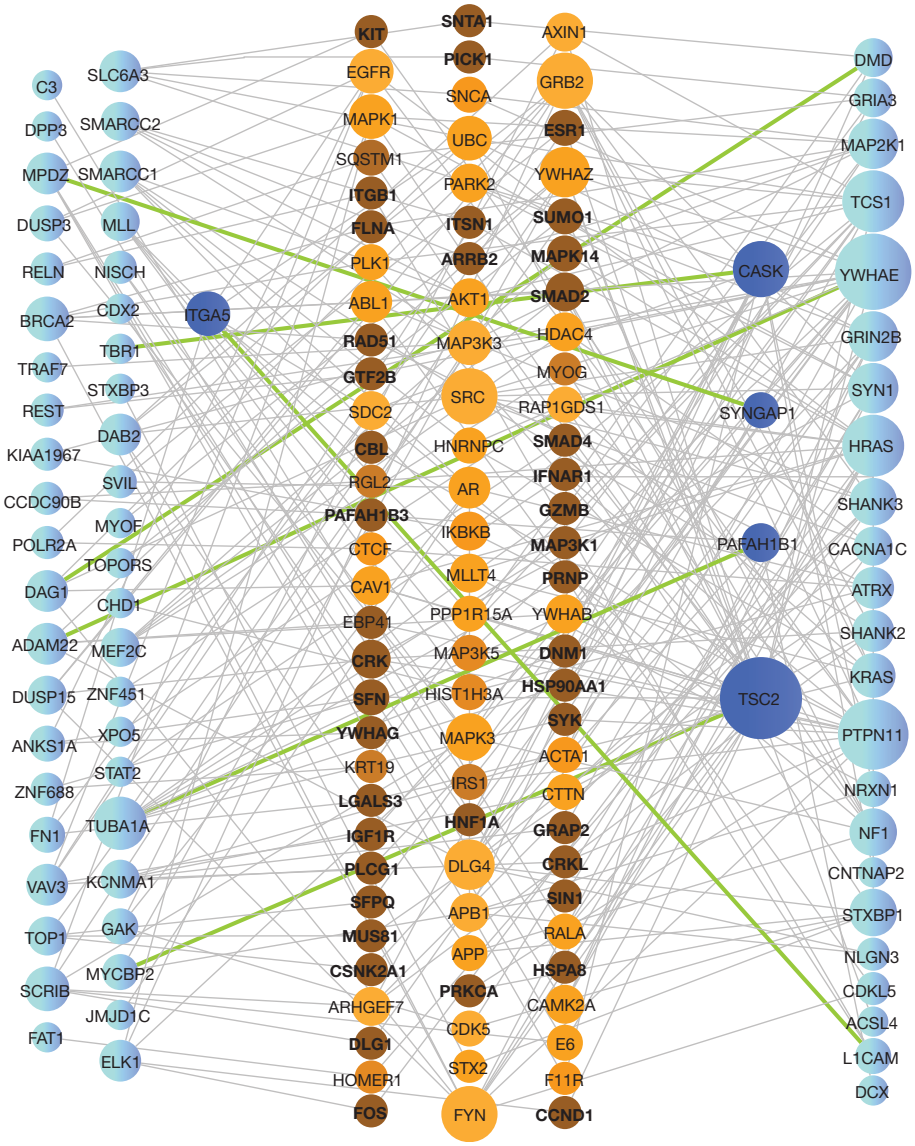


Figure 2 | Direct and indirect protein–protein interaction for genes with a functional *de novo* event and previous ASD genes. PPI network analysis for *de novo* variants and 31 previous synaptic ASD genes (see Supplementary Information). Nodes are sized based on connectivity. Genes harbouring *de novo*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

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

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

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

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

METHODS SUMMARY

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

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

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

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

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

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

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

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

1. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167,** 1357–1363 (2010).
2. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68,** 1095–1102 (2011).
3. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
4. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
5. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
6. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25,** 528–535 (2009).
7. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
8. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
10. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
11. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* http://dx.doi.org/10.1038/nature10945 (this issue).
12. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
13. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–739 (2007).
14. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1,** 40–47 (2000).
15. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
16. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105,** 20870–20875 (2008).
17. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
18. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* http://dx.doi.org/10.1038/nature10989 (this issue).
19. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
20. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).

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

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

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

Author Information Data included in this manuscript have been deposited at dbGaP under accession number phs000298.v1.p1 and is available for download at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu), J.D.B. (joseph.buxbaum@mssm.edu) or K.R. (kathryn.roeder@gmail.com).

METHODS

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

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

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

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

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

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

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

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

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

doi:10.1038/nature11011

Data generation, processing and identifying de novo mutations

Data generation

Exome capture and sequencing was performed at each site using similar methods. Genomic DNA (~3 ug) was sheared to 200-300 bp using a Covaris Acoustic Adaptor, and (Vanderbilt) DNA purified using Agencourt's AMPure XP Solid Phase Reversible Immobilization paramagnetic (SPRI) bead. Fragments were end-repaired, dA-tailed, and sequencing adaptor oligonucleotides ligated using reagents from New England BioLabs. Libraries were barcoded using the Illumina index read strategy, which uses six-base sequences within the adapter that are sequenced separately from the genomic DNA insert. Ligated products were size selected with gel electrophoresis (Mt Sinai School of Medicine) or purified using SPRI beads (Vanderbilt). The DNA library was subsequently enriched for sequences with 5′ and 3′ adapters by PCR amplification using with primers complementary to the adapter sequences (ligation-mediated PCR, LM-PCR). Exons were captured using either the Agilent 38Mb SureSelect v2 (University of Pennsylvania and Broad Institute), the NimbleGen Seq Cap EZ SR v2 (Mt Sinai School of Medicine, Vanderbilt University), or NimbleGen VCRome 2.1 (Baylor). In some cases, barcoded libraries from 2-4 subjects were mixed prior to hybridization with the capture reagent. After capture, another round of LM-PCR was performed to generate enough DNA to sequence. Libraries were sequenced using an IlluminaHiSeq2000.

Sequence processing and variant calling was performed using a similar computational workflow at all sites. Data was processed with Picard (http://picard.sourceforge.net/), which utilizes base quality-score recalibration and local realignment at known indels[1] and BWA [2] for mapping reads to hg19. SNPs were called using GATK [1,3] for all trios jointly. The variable sites that we have considered in analysis are restricted to those that pass GATK standard filters to eliminate SNPs with strand-bias, low quality for the depth of sequencing achieved, homopolymer runs, and SNPs near indels. The same thresholds on allelic depth and likelihood were used to identify the likely *de novo* mutations. Putative *de novo* mutations were identified as sites where both parents were homozygous for the reference sequence and the offspring was heterozygous.

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

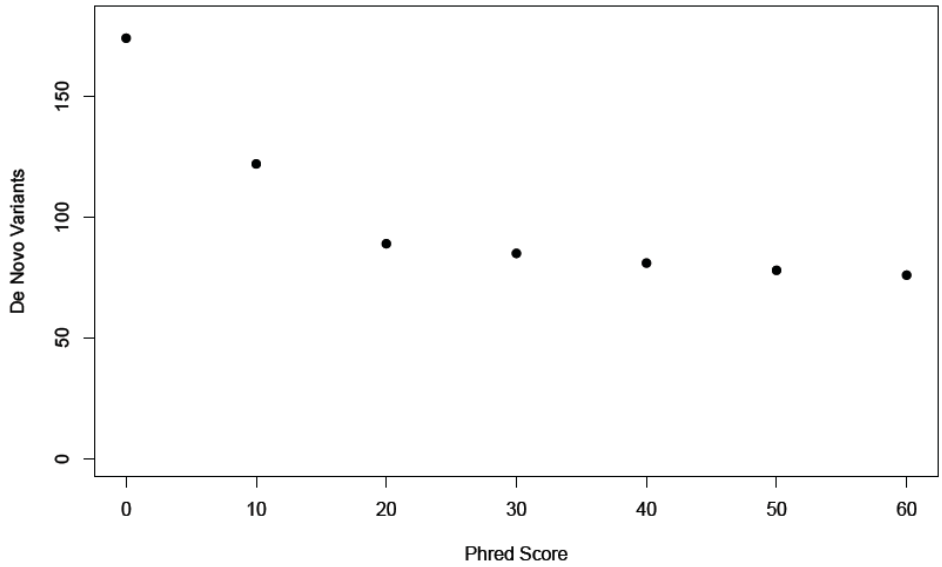
Identifying de novo mutations

We identified potential de novo mutations if we observed a heterozygous genotype in the offspring and observed reference homozygote genotypes in both parents and did not

observe any other copy of the alternate allele in the trio sample.

We further cleaned the genotypes beyond the standard GATK filters by imposing a threshold on the observed allele balances for the family. Conceptually, we aimed to remove instances where the child was likely miscalled for some proportion of sequencing error by removing heterozygote calls in the child when more than 70% of reads were reference, as well as cases where a parent was likely miscalled reference homozygous as indicated by more than 5% non-reference reads that matched the child's heterozygous call. Such data configurations as these do not conform to expected proportions (and may arise from erroneous read mapping, duplicated segments or biases in data generation) but at high sequence depth can generate genotype calls with high confidence despite being clear outliers. Specifically, the construction of the genotype likelihood in GATK is such that we are estimating the probability of the data given an assumed genotype. Thus, in the presence of many copies of the non-reference allele at a position even if this only represents 10% of reads, for example, the likelihood of the reference homozygote decreases more rapidly than that of the heterozygote. In addition we removed sites where the child read depth was <10% of the total parental read depth – eliminating occasional spurious calls occurring in cases where the child may have been homozygously deleted or an exon failed to capture.

To further refine the list of possible *de novo* events, sites were filtered based on the likelihood of the data, given the genotype, represented in the 'phred-scale' (-10log10(p)). p is the likelihood ratio defined such that the denominator is the most likely genotype for the designated individual and site. Hence PL is represented by three numbers PL{AA, AB, BB} and the most likelihood genotype has PL score 0. If this is = {30, 0, 40}, then the most likely genotype is heterozygous with L(data|AB) being 1000 times more likely than L(data|AA) and 10000 times more likely then L(data|BB). To insulate against including positions for which the genotype calls are uncertain, we explored the rate of *de novo* mutation as a function of threshold on the PL. We define *de novo* events at a PL threshold of T to be those sites where the child's PL(AA) score exceeds T and for both parents scores of PL(AB) and PL(BB) exceed T. As expected, many sites contain low confidence genotype calls (largely due to low coverage), where the most likely genotypes would suggest a de novo event, but a consistent Mendelian arrangement of alleles is nearly as likely. We enumerated the effect of the threshold on the number of de novo events as a function of T in the 96 trios. By the time we reach higher confidences of 20 (100:1 odds) and 30 (1000:1 odds), we see a plateau identifying the mixture of false events (rapidly declining distribution from T=0) with true events (a relatively flat distribution governed by the depth of coverage in practice flat through PLs of 100-200) (Supplementary Figure 1).



Supplementary Figure 1. Number of de novo variants by Phred score threshold.

In these trios, we observe 87 events, when T = 30 (i.e., the next most likely genotype has a PL of >=30 for the child and both parents). Of these, 60 are missense, 4 are nonsense, 22 are silent and 1 is at an intronic conserved splice site. A first batch of Sequenom genotyping confirmed 74 of 75 events for which a genotyping assay was successfully run validated as true de novo events, suggesting high specificity as intended. To insure sensitivity of the selected threshold we also advanced all events at a lower confidence T=20 – there were only 4 events added by lowering the threshold (and none were validated). The fact that few events, and none which validated, are included when this threshold is lowered, when combined with the careful evaluation of coverage described below that suggests more than 90% of the targeted exome is covered at a level that would reliably provide this statistical support, suggests the depth of sequencing coverage and analytic approach provides high sensitivity as well. All of the filtering and likelihood analysis performed is incorporated in a Python script to identify de novo sites from a GATK generated VCF-formatted file.

Wave Two of this experiment comprised 78 trios, sequenced at Baylor College of Medicine, the Broad Institute, Mount Sinai School of Medicine, University of Pennsylvania, and Vanderbilt University. These trios were then processed using precisely the same approach as defined for the Wave 1 trios. Specifically, the same quality thresholds for genotype likelihood, allele balance and the other filtering criteria were applied to these data. Validation of all events has been performed via Sanger sequencing at all sites except the Broad Institute and has also demonstrated high confirmation rates with these analytic approaches.

Computing the expected de novo mutation rate in the exome

Our goal is to estimate the de novo rate over the exome, v_e. We could take the *de novo* rate of the entire genome, $v_g = 1.2 \times 10^{-8}$, as an estimate. Because the base pair composition of the genome is somewhat different from that of the exome, and because the *de novo* mutation rate is known to depend on sequence context, the genomewide rate is not sufficiently accurate. Notably the exome has an average GC content of approximately 50% while the genome is approximately 40%, and the most common sites for mutation are C to T transitions at CpG sites. Thus we should do better by considering the context-specific mutation rates.

To obtain v_e we, following Kryukov[4] and Krawczak[5], assume that the trinucleotide context is sufficient, specifically the probability a base mutates in the context of its adjacent bases. What we require is the *de novo* probability for each of the 64 possible (4 x 4 x 4) trinucleotides to mutate directionally to any of the other 3 possible bases in the middle position (that is, $XY_1Z \rightarrow XY_2Z$). The data on *de novo* events in the human (or any mammalian) genome are too sparse to estimate these 64x3 probabilities directly. Nonetheless it is reasonable to assume that these context-specific mutation rates, which have given rise to single-nucleotide variation within and between mammals, are reflected precisely by the relative rates of standing variants in each context in the intergenic genome. For Model M1, we utilize the context-dependent mutation-rate matrix of Kryukov[4] which utilizes fixed differences in human, chimp and baboon to empirically calculate the "directed" 64x3 mutation matrix. The proportionality constant, λ, follows from some algebra and the assumption that $v_g = 1.2 \times 10^{-8}$.

We additionally calculated a second version of the 64x3 context-dependent rate matrix for human polymorphism alone using emerging 1000 Genomes data and restricting our sequence analysis to intergenic regions that are orthologous between human and chimp. For the entire orthologous sequence we tallied every instance of each of the 64 trinucleotide sequences. We identified all instances of mutations in the 1,000 Genomes data by assuming that the chimp allele represents the ancestral of the two alleles at a polymorphic site in humans and that the alternate base at the SNP is the derived allele. Thus, we have counts for all 64 configurations, as well as the 64x3 possible context-specific directional mutations.

Let T denote a trinucleotide background and V denote whether the central base is variable (i.e., a SNP). Let P(V) be the marginal probability a nucleotide is variable. We estimate P(V) with the sum of all mutations divided by the length of the all sequence. Likewise, let P(V|T=t) be the conditional probability a nucleotide in the background T=t is variable. We estimate P(V|T=t) using the sum of mutations in that trinucleotide divided by the occurrence of the trinucleotide.

The key to converting standing variation rates to trinucleotide specific mutation rates is the following. We need to calibrate P(V) to be representative of the expected mutation rate in a single generation. We assume a mutation rate of $v_g = 1.2 \times 10^{-8}$, consistent with recent work from 1,000 Genomes[6]. And so, $\lambda = v_g / P(V)$.

To estimate the exome specific mutation rate we take a weighted average of λP(V|T), weighted by the probability of each trinucleotide T=t in the exome. If the distribution of T were equal in the exome to the genome as a whole, this simplifies to 1.2×10^{-8}, as desired. However, the distributions are not equal, and the relative rate of mutation, P(V|T)/P(V), is higher for trinucleotides that occur more frequently in the exome.

Successful Target Identification

To estimate accurately the expected *de novo* mutation count in our experiment, we must also have an estimate of how much sequence is truly captured with adequate depth and quality by the sequencing experiment. To determine what parts of the targeted exome were captured, we considered each trio jointly and deemed a target to be covered for a family if all three members of the trio exceeded 10x coverage – given high very correlation in segmental coverage across families, we considered an exon assayed in the experiment if more than half of the families passed this threshold. We empirically determined that 10x depth of coverage was a sensitive threshold for accurate variant identification in the wave 1 data set.

As described in the sensitivity analysis above, we explored all sites where the likelihood of the offspring and parental genotypes were each supported by 100 to1 odds or better (i.e., PL >= 20). Conditioning on 10x coverage, we see that 98.5% of calls have a PL of 20 or better for all individuals and focusing on individual genotype calls at singleton sites in the data set, more than 99.9% of individual genotype likelihoods exceeded the threshold of PL >= 20. By applying this definition of successfully covered sequence, the total territory included in the experiment is 30.23 Mb of sequence. Considering only coding exon targets under analysis in this manuscript, this translates to 27.86 Mb of sequence.

Supplementary Table 1

Supplementary Table 1 is a stand-alone Excel spreadsheet providing all validated de novo events and annotation.

Expectation for Exome and Experiment

Based on the trinucleotide specific mutation rate, for the entire Consensus Coding DNA Sequence (CCDS) we calculate the effective mutation rate of the entire exome to be approximately $1.65e^{-8}$. Consequently the expected de novo mutation rate is approximately 1.032 per family. For our experiment, more than 90% of the exome is well captured by our experiment, but there is a strong bias against coverage of particularly high GC% regions that are more mutable. We calculate the effective mutation rate of the covered coding sequence at approximately 1.54e-8 for the bases covered for each trio at 10x. Thus we estimate the de novo rate per family to be approximately 0.87 per family. In this target region we observed 85 de novo mutations in 96 trios (the 86th was an intronic

conserved splice site not part of the coding target tally) – an observation of 0.885 per trio consistent with expectation.

Proband diagnosis and familial characterization

Affected probands were assessed using standard diagnostic instruments by research-reliable research personnel, the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule-Generic (ADOS), DSM-IV diagnosis of a pervasive developmental disorder by a clinician and received a medical screen. Intellectual ability and/or adaptive function were assessed for all probands and some parents. Additional measures, largely for probands, included the Vineland Adaptive Behavior Scales, Peabody Picture Vocabulary Test, and physical measures (Supplementary Table 2).

All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the 3 subjects from AGRE that were not assessed with the ADOS. In all 85% of probands were classified with autism on both the ADI-R and ADOS.

Clinical recruitment was largely uniform across recruitment sites. The largest set of families assessed here (Supplementary Table 2) was recruited by the Autism Consortium of Boston, who recruited families only if the index proband received a diagnosis of autism or ASD. This was also the procedure for Vanderbilt, University of Illinois (Chicago) for its Autism Center of Excellence (ACE), and for Mt. Sinai. For the University of Pennsylvania, families were drawn from the ACE center at the University of Washington and were strictly families with at least a proband and sibling both diagnosed with ASD (i.e., multiplex families). For the Baylor Sequencing Center, families were drawn from the Autism Genetic Resource Exchange (AGRE), and they were multiplex families.

All families were characterized for family history of ASD and other psychiatric conditions, however, specific Family History forms varied by site. Results will be summarized for three groups, siblings, parents and extended relatives. For our sample, almost all siblings were screened for affection status and most found unaffected – specifically, siblings in just 33 families from our collection of 174 were diagnosed with ASD. Parents were assessed by the interviewing clinicians who spent extensive time during interviews to determine whether they might exhibit symptoms of autism spectrum disorder and/or mild/moderate intellectual disability. Moreover, 76% of parents in our study were formally screened by at least one of these instruments, Social Reciprocity Scale-Adult Research Version (SRS-ARV), Broader Autism Phenotype Questionnaire (BAP-Q), Broader Autism Phenotype Symptom Scale (BPASS). Three parents across the combined collection were determined to have ASD symptoms noted by clinicians – these we considered to have positive family history. For all families, parental report was used to determine any further family history in parents, siblings, grandparents, uncles/aunts and first cousins. After these formal assessments of first-degree relatives and parental reporting of first through third degree relatives, 69% of probands have no first to third degree relative family history of ASD.

Supplementary Table 2. Characterization of trios by recruitment sites. If all subjects were assessed on the instrument, the entry in the cell is Y=yes, if it was not assessed it is N=no, if a subset were assessed the cell is set as X%.

A. **Proband characterization**

Characterization	**Instruments**	**Recruitment Sites**					
		Autism Consortium[1] N = 104	UIC N = 18	Mt. Sinai N = 18	VU N = 15	UW (UPenn) N = 14	AGRE (Baylor) N = 6
Diagnosis	Medical Screen	Y	Y	Y	Y	Y	P=17%
	ADI-R	Y	Y	Y	Y	Y	Y
	ADOS	Y	Y	Y	Y	Y	P=50%
	DSM-IV Clinician Diagnosis	Y	Y	Y	Y	Y	P=50%
Communication	Peabody Picture Vocabulary Test	Y	Y	Y	Y	N	P=17%
Intellectual Ability/Adaptive Function	Standardized IQ	Y	Y	Y	Y	Y	P=17%
	Vineland Adaptive Behaviour Scales	Y	Y	Y	Y	Y	P=33%
Physical Attributes	Height	Y	Y	Y	Y	Y	P=17%
	Weight	Y	Y	Y	Y	N	P=17%
	Head Circumference	Y	Y	Y	Y	Y	P=17%

B. **Parent characterization**

Broader Phenotype	Social Reciprocity Scale Adult Research Version	Y	N	22%	47%	N	N
	BAPQ or BPASS	Y	Y	17%	80%	93%	N

Secondary analyses of trio data

We conducted a range of secondary analyses on the data (Table 2; Supplementary Table 3). We find no significant differences in the number of *de novo* mutations based on diagnostic severity, sex, or family history. Similarly, the number of *de novo* mutations is not a significant predictor of IQ. Paternal and maternal ages are significant predictors of the number of *de novo* mutations. However, paternal and maternal age are highly correlated in the dataset (r^2=0.679), meaning that we cannot effectively disentangle whether the effect is a consequence of maternal, paternal, or both ages without assigning the origin of the mutation

Supplementary Table 3

	N	Mean de novo	P-value
Family History no	120	0.975	0.518
Family History yes	55	0.872	
Male	146	0.910	0.2345
Female	29	1.172	
Broad Criteria	26	0.962	0.723
Strict Criteria	149	0.946	
		Beta	P-value
Verbal IQ	74	-1.639	0.472
Nonverbal IQ	60	-1.267	0.552
Full Scale IQ	77	-1.627	0.468
Paternal Age	139	0.0038	0.000239
Maternal Age	139	0.0061	3.80E-05

Supplementary Table 3. Presented are the additional analyses of clinical data for the autism probands in the trios. For the dichotomous traits (Family history, sex, and broad vs. strict (ADI-R autism and ADOS autism) criteria) the means of the subgroups are presented and P-value as defined by Poisson regression. For the three IQ measures, we predicted IQ as a function of the number of *de novo* mutations. For paternal and maternal age, we predicted the number of *de novo* mutations based on the age of the parents.

Expected distribution of de novo events under a range of genetic models for ASD

To explore models in which *de novo* mutations contribute to ASD risk and which are consistent with observed data, we modeled a Poisson process with the observed and expected mean of 1 *de novo* point mutations per exome as a function of the following unknown parameters:

- G = fraction of genes (or exome territory) contributing to autism risk
- P = probability that a novel missense variant is highly deleterious
- γ = genotype relative risk (GRR) contributed to an individual with one or more such hits

For the model we require the following notation:

- A = autism
- X = number of de novo SNVs
- B = bad hit with potentially deleterious effect
- H = number of bad hits

For the purpose of the model, we consider any nonsense, splice or deleterious missense variant to be in the 'risk mutation' category.

P has been estimated at 20% [4] and 35% [13]; however given the significant fraction of de novo events that are silent or predicted neutral, varying this number has little impact on the model because it equates to a range of 20-30% of de novo events overall having significant likelihood of deleterious impact. We set this value at 30%. We assume that 66.7% of exonic *de novo* events are missense and 3.3% are nonsense[4] with the latter assumed to produce loss of function. Consequently, the probability a coding mutation is deleterious (bad) is

$$P(B) = (.667\,P + .033)\,G.$$

Based on our empirical findings and those of others, we assume that X is distributed as a Poisson with rate parameter equal to 1. Our objective is to compute this distribution for a population with autism for varying choices of γ.

Clearly $P(X=x|A) = P(A|X=x)\,P(X=x)/\,P(A)$. To compute $P(A|X=x)$ we partition the outcomes into those with and without at least one bad mutation.

$$P(A \mid X = x) = P(H = 0 \mid X = x)P(A \mid H = 0) + P(H > 0 \mid X = x)P(A \mid H > 0)$$
$$= P(H = 0 \mid X = x)[\,P(A \mid H = 0) + P(H > 0 \mid X = x)\,\gamma],$$

in which $P(H = 0 \mid X = x) = [1 - P(B)]^x$, $P(H > 0 \mid X = x) = 1 - [1 - P(B)]^x$ and

$$\gamma = \frac{P(A \mid H > 0)}{P(A \mid H = 0)}.$$

Consequently

$$P(X = x \mid A) = \frac{P(X = x)[\,P(A \mid H = 0) + P(H > 0 \mid X = x)\gamma]}{\sum_x P(X = x)[\,P(A \mid H = 0) + P(H > 0 \mid X = x)\gamma]}$$

In these calculations, we verify that the model parameters satisfy the following constraint imposed by the prevalence of autism. Specifically, because

$P(A) = P(B)P(A \mid H = 0)\gamma + (1 - P(B))P(A \mid H = 0)$. For a given choice of γ, we solve for P(A|H=0) and ensure that it is bounded above by 1.

We have estimated that in practice 13% of all mutations are missed, based on the amount and composition of targeted coding sequence. Thus, in the simulations below, we set the baseline Poisson rate at 0.87.

Recent CNV studies suggest hundreds of loci underly autism risk. Such polygenicity is both routinely inferred for complex disease in general and can explain much of the difficulty in gene discovery in autism to date. We simulated six scenarios using 200, 500 or 1000 genes (or specifically 1, 2.5 or 5% of coding exon territory) to be involved in autism risk, and two choices of γ, the genotype relative risk associated with acquisition of a *de novo* risk mutation (Supplementary Table 6). γ=20 corresponds to a value routinely observed in large meta-analyses for the relative risks associated with de novo CNVs [14]; $\gamma = 200$ represents a much higher penetrance models that are essentially Mendelian (closer to, for example, of the properties of *MECP2* and *CDKL5* mutations in Rett syndrome).

As shown in the table below, the more modest value of γ introduces little deviation from the expected 0.87 events per exome and would not introduce a highly significant distortion in the current study. By contrast the larger value of γ is starkly inconsistent with the distribution observed here. Most notably, with a large number of high penetrance genes, the models predict very few cases would carry no functional events. To reinforce this point, we drew 10,000 sets of 174 trios and tallied how often by chance under each model we would observe as few or fewer de novo mutations than the 161 observed. These results demonstrate that models postulating large numbers of genes where de novo mutations are highly penetrant can be rejected by these data. Thus, while there may exist a few hidden Mendelian forms of autism exposed by high penetrance de novo mutations, the majority of de novo mutations must confer a more modest risk. In addition, we also tally the observed number of cases with no functional events and calculate the binomial probability of observing as large or larger number according to each model with similar results.

Finally, the implication of each model is presented in two ways: first - the overall proportion of cases that harbor a disease-relevant de novo event is reported – for all consistent models, this number is far less than 50%; second – the implied variance explained by *de novo* protein-coding point mutations is reported. This second quantity is calculated assuming a liability threshold model and additive contributions from the many genes contributing to autism risk. Because the events are incompletely penetrant, the percent variance explained is considerably lower than the proportion of cases carrying a relevant event.

Supplementary Table 4 Expected Patterns of Mutations.

Given 13% events missed				**NULL**	**GRR=20**			**GRR=200**		
		Observed No.	**Observed Proportion**	**Expected**	**200**	**500**	**1000**	**200**	**500**	**1000**
	0	71	0.408	0.419	0.403	0.383	0.353	0.303	0.223	0.165
	1	62	0.356	0.364	0.366	0.369	0.373	0.379	0.390	0.398
	2	28	0.161	0.158	0.166	0.176	0.191	0.215	0.254	0.283
	3	10	0.057	0.046	0.050	0.056	0.064	0.077	0.098	0.114
	4	2	0.011	0.010	0.011	0.013	0.016	0.020	0.027	0.032
	≥5	1	0.006	0.002	0.002	0.003	0.004	0.005	0.007	0.008
	mean		0.925	0.87	0.908	0.958	1.028	1.147	1.338	1.476
P(as few events as observed)					ns	ns	0.1	0.002	1E-04	1E-04
Proportion of cases no functional events)		94	0.54	0.543	0.523	0.496	0.456	0.392	0.285	0.205
P(as few cases as observed)					ns	ns	0.016	0.001	<1E-10	<1E-10
Model implication										
Proportion of cases without de novo coding event					0.96	0.90	0.81	0.68	0.46	0.30
Percent variance explained					1.0	2.4	4.6	7.8	8.4	11.7

Legend: In this table, we show the expected number of *de novo* mutations that would be observed in the trios, conditional on the specified genetic models. The GRR=20 or 200 reflects scenarios where some *de novo* mutations have a relative risk of 20 or 200. We present the probability of observing as few events in total given these models, the percentage of cases with no functional events and probability of observing as many cases without events as we do given these models. Based on the number of cases in particular, we can rule out scenarios of 1,000 genes with a GRR of 20 and 200 genes with a GRR of 200. A GRR of 20 is consistent with the effect sizes estimated from CNVs and a GRR of 200 would be nearly fully penetrant. Observed number and observed proportion represent the observed number of *de novo* mutations and observed proportion of *de novo* mutations in the actual experiment.

Evaluation of Enrichment for ASD Gene

To obtain a reference list related to ASD or ASD with intellectual disability (ID), we used a list from a recent publication [7], further updated based on recent reports, for a total of 112 genes ("ASD112") (Supplementary Table 4). We also developed a similar list for ID genes that had not yet been described in ASD (see Supplementary Table 5, modified from Pinto[8]; contact CB for further details). To explore the overlap of ASD/ID genes with those in synaptic compartments, we took advantage of large-scale proteome studies for synaptic proteins. Eight lists, including synaptic and postsynaptic lists were derived from a single study in mouse [9] and two presynaptic lists were derived from a second study in mouse [10]. For the human postsynaptic density, we made use of a recent list derived from purified postsynaptic densities (PSD) from human neocortex [11]. As the ASD gene list was for human proteins, we mapped murine proteins to their human orthologs. The intersection of these synaptic protein lists with the ASD112 list yielded a subset of 31 high risk ASD genes found in the synapse (ASD31). Genes in ASD31 are: *ACSL4, ALDH5A1, ALDH7A1, ATRX, CACNA1C, CASK, CDKL5, CNTNAP2, DCX, DMD, GRIA3, GRIN2B, HRAS, IL1RAPL1, IQSEC2, KRAS, L1CAM, MAP2K1, NF1, NLGN3, NRXN1, PAFAH1B1, PTPN11, SHANK2, SHANK3, STXBP1, SYN1, SYNGAP1, TSC1, TSC2,* and *YWHAE.* Enrichment P values were calculated using lists2networks[12], making use of the Fishers exact (hypergeometric) test.

Supplementary Table 5. ASD genes

	Gene symbol	Chr	Locus	Start	End	Disorder	Inheritance pattern
1	*POMGNT1*	1	1p34.1	46 426 940	46 436 708	Muscle-eye-brain disease	AR
2	*RPE65*	1	1p31.3	68 667 095	68 688 230	Leber congenital amaurosis	AR
3	*DPYD*	1	1p21.3	97 315 888	98 159 203	Dihydropyrimidine dehydrogenase deficiency	AR
4	*NRXN1*	2	2p16.3	50 000 992	51 113 178	disrupted in ASD, MR, schizophrenia (dominant?); Pitt-Hopkins-like syndrome-2 (recessive)	AD?/AR
5	*NPHP1*	2	2q13	110 238 203	110 319 928	Joubert syndrome type 4, nephronophthisis	AR
6	*MBD5*	2	2q23.1	148 932 242	148 987 514	autosomal dominant MR, responsible for the 2q23.1 microdeletion syndrome	AD
7	*SCN1A*	2	2q24.3	166 553 916	166 638 395	severe myoclonic epilepsy of infancy (Dravet syndrome)	AD
8	*SATB2*	2	2q33.1	199 842 469	200 033 500	Haploinsufficiency of SATB2 causes some of the clinical features of the 2q33.1 microdeletion syndrome	AD
9	*BTD*	3	3p24.3	15 618 259	15 662 329	Biotinidase deficiency	AR
10	*FOXP1*	3	3p14.1	71 087 426	71 715 830	non-syndromic MR and autism	AD
11	*PRSS12*	4	4q26	119 421 865	119 493 370	autosomal recessive non-syndromic MR	AR
12	*NIPBL*	5	5p13.2	36 912 618	37 101 678	Cornelia de Lange syndrome	AD
13	*MEF2C*	5	5q14.3	88 051 922	88 214 780	syndromic MR, responsible for the 5q14.3 microdeletion syndrome	AD
14	*ALDH7A1*	5	5q23.2	125 906 817	125 958 981	Pyridoxine-dependent epilepsy	AR
15	*NSD1*	5	5q35.2-q35.3	176 493 439	176 659 820	Sotos syndrome	AD
16	*ALDH5A1*	6	6p22.2	24 603 176	24 645 414	Succinic semialdehyde dehydrogenase deficiency (gamma-hydroxybutyric aciduria)	AR
17	*SYNGAP1*	6	6p21.32	33 495 825	33 529 444	non-syndromic MR	AD
18	*AHI1*	6	6q23.3	135 646 817	135 860 576	Joubert syndrome 3	AR
19	*HOXA1*	7	7p15.2	27 099 139	27 102 150	HOXA1 syndrome, Bosley-Salih-Alorainy variant	AR
20	*BRAF*	7	7q34	140 080 282	140 271 033	Cardio-facio-cutaneous syndrome	AD
21	*CNTNAP2*	7	7q35-q36.1	145 444 386	147 749 019	Cortical dysplasia-focal epilepsy syndrome, Pitt-Hopkins-like syndrome-1 (recessive); the clinical significance of the disruption of 1 allele is unknown	AR
22	*HGSNAT*	8	8p11.21	43 114 749	43 177 127	Mucopolysaccharidosis type IIIC (Sanfilippo syndrome C)	AR
23	*CHD7*	8	8q12.2	61 753 893	61 942 021	CHARGE syndrome	AD
24	*VPS13B*	8	8q22.2	100 094 670	100 958 984	Cohen syndrome	AR
25	*STXBP1*	9	9q34.11	129 414 389	129 494 816	autosomal dominant non-syndromic epilepsy, MR and autism	AD
26	*POMT1*	9	9q34.13	133 368 110	133 389 014	Limb-girdle muscular dystrophy with MR; Walker-Warburg syndrome	AR
27	*TSC1*	9	9q34.13	134 756 557	134 809 841	Tuberous sclerosis	AD
28	*EHMT1*	9	9q34.3	139 725 240	139 850 399	9q subtelomeric deletion syndrome (Kleefstra syndrome)	AD
29	*PTEN*	10	10q23.31	89 613 175	89 718 512	PTEN hamartoma-tumor syndrome, MR and ASD with macrocephaly	AD
30	*FGFR2*	10	10q26.13	123 227 834	123 347 962	Apert syndrome	AD
31	*HRAS*	11	11p15.5	522 242	525 550	Costello syndrome	AD
32	*IGF2*	11	11p15.5	2 106 923	2 118 917	Aberrant imprinting of *IGF2* is associated with Beckwith–Wiedemann syndrome and Silver–Russell syndrome	AD
33	*KCNJ11*	11	11p15.1	17 363 372	17 366 782	DEND syndrome (developmental delay, epilepsy, and neonatal diabetes)	AD
34	*SHANK2*	11	11q13.3	69 990 609	70 186 520	non-syndromic MR and ASD	AD
35	*DHCR7*	11	11q13.4	70 823 105	70 837 125	Smith-Lemli-Opitz syndrome	AR
36	*FOLR1*	11	11q13.4	71 578 250	71 585 014	Cerebral folate transport deficiency	AR
37	*HEPACAM*	11	11q24.2	124 294 356	124 311 518	Megalencephalic leukoencephalopathy with subcortical cysts (recessive); leukodystrophy and macrocephaly (dominant)	AR/AD
38	*CACNA1C*	12	12p13.33	2 032 677	2 677 376	Timothy syndrome	AD
39	*GRIN2B*	12	12p13.1	13 605 677	14 024 289	autosomal dominant MR	AD
40	*KRAS*	12	12p12.1	25 249 447	25 295 121	Cardio-facio-cutaneous syndrome	AD
41	*GNS*	12	12q14.3	63 393 489	63 439 493	Mucopolysaccharidosis type IIID (Sanfilippo disease D)	AR
42	*CEP290*	12	12q21.32	86 966 921	87 060 124	Joubert syndrome 5, Leber congenital amaurosis, Bardet-Biedl syndrome 14, Meckel syndrome 4	AR
43	*PAH*	12	12q23.2	101 756 234	101 835 511	Phenylketonuria	AR
44	*PTPN11*	12	12q24.13	111 340 919	111 432 100	Noonan syndrome	AD
45	*FOXG1*	14	14q12	28 306 038	28 308 622	congenital variant of Rett syndrome	AD
46	*L2HGDH*	14	14q22.1	49 778 902	49 848 697	L-2-hydroxyglutaric aciduria	AR

47	*UBE3A*	15	15q11.2	23 133 489	23 204 888	Angelman syndrome	AD
48	*GATM*	15	15q21.1	43 440 614	43 458 272	Arginine:glycine amidinotransferase (AGAT) deficiency	AR
49	*MAP2K1*	15	15q22.31	64 466 265	64 570 936	Cardio-facio-cutaneous syndrome	AD
50	*TSC2*	16	16p13.3	2 037 991	2 078 714	Tuberous sclerosis	AD
51	*CREBBP*	16	16p13.3	3 715 057	3 870 122	Rubinstein-Taybi syndrome	AD
52	*RPGRIP1L*	16	16q12.2	52 191 319	52 295 272	Joubert syndrome 7, Meckel syndrome, COACH syndrome	AR
53	*YWHAE*	17	17p13.3	1 194 593	1 250 267	Miller-Dieker syndrome	AD
54	*PAFAH1B1*	17	17p13.3	2 443 673	2 535 659	isolated lissencephaly, Miller-Dieker syndrome	AD
55	*GUCY2D*	17	17p13.1	7 846 713	7 864 383	Leber congenital amaurosis	AR
56	*RAI1*	17	17p11.2	17 525 512	17 655 490	Smith-Magenis syndrome (deletion, mutation), Potocki-Lupski syndrome (duplication)	AD
57	*RNF135*	17	17q11.2	26 322 082	26 351 053	overgrowth syndrome; haploinsufficiency of *RNF135* contributes to the phenotype of the NF1 microdeletion syndrome	
58	*NF1*	17	17q11.2	26 446 121	26 728 821	Neurofibromatosis type 1	AD
59	*NAGLU*	17	17q21.31	37 941 477	37 949 992	Mucopolysaccharidosis type IIIB (Sanfilippo syndrome B)	AR
60	*SGSH*	17	17q25.3	75 797 674	75 808 794	Sanfilippo syndrome A (mucopolysaccharidosis III A)	AR
61	*GAMT*	19	19p13.3	1 349 606	1 352 552	guanidine acetate methyltransferase (GAMT) deficiency	AR
62	*NFIX*	19	19p13.13	12 967 584	13 070 610	Sotos-like overgrowth syndrome, Marshall-Smith syndrome	AD
63	*DMPK*	19	19q13.32	50 964 816	50 977 655	Myotonic dystrophy type 1 (Steinert disease)	AD
64	*MKKS*	20	20p12.2	10 333 833	10 362 866	Bardet-Biedl syndrome	AR
65	*TBX1*	22	22q11.21	18 124 226	18 134 855	responsible for some of the phenotypic features of the 22q11 deletion syndrome (velocardiofacial/DiGeorge syndrome)	AD
66	*ADSL*	22	22q13.1	39 072 450	39 092 521	adenylosuccinate lyase deficiency	AR
67	*SHANK3*	22	22q13.33	49 459 936	49 518 507	22q13 deletion syndrome (Phelan-McDermid syndrome)	AD
68	*NLGN4X*	X	Xp22.31-p22.32	5 818 083	6 156 706	non-syndromic X-linked MR and ASD	XL
69	*MID1*	X	Xp22.2	10 373 596	10 761 730	Opitz syndrome (Opitz/BBB syndrome)	XL
70	*AP1S2*	X	Xp22.2	15 753 850	15 783 021	syndromic and non-syndromic X-linked MR	XL
71	*NHS*	X	Xp22.13	17 303 464	17 664 034	Nance-Horan syndrome	XL
72	*CDKL5*	X	Xp22.13	18 353 646	18 581 670	Rett-like syndrome with infantile spasms	XL
73	*PTCHD1*	X	Xp22.11	23 262 906	23 324 839	non-syndromic X-linked MR and ASD	XL
74	*ARX*	X	Xp21.3	24 931 732	24 943 986	X-linked lissencephaly and abnormal genitalia, West syndrome, Partington syndrome, non-syndromic X-linked MR	XL
75	*IL1RAPL1*	X	Xp21.2-p21.3	28 515 602	29 883 938	non-syndromic X-linked MR and ASD	XL
76	*DMD*	X	Xp21.1-21.2	31 047 266	33 139 594	Muscular dystrophy, Duchenne and Becker types	XL
77	*OTC*	X	Xp11.4	38 096 680	38 165 647	Ornithine transcarbamylase deficiency	XL
78	*CASK*	X	Xp11.4	41 259 133	41 667 231	syndromic and non-syndromic X-linked MR	XL
79	*NDP*	X	Xp11.3	43 692 968	43 717 788	Norrie disease	XL
80	*ZNF674*	X	Xp11.3	46 243 490	46 289 820	non-syndromic X-linked MR	XL
81	*SYN1*	X	Xp11.23	47 316 244	47 364 200	X-linked epilepsy and MR	XL
82	*ZNF81*	X	Xp11.23	47 581 245	47 666 554	non-syndromic X-linked MR	XL
83	*FTSJ1*	X	Xp11.23	48 219 493	48 229 696	non-syndromic X-linked MR	XL
84	*PQBP1*	X	Xq11.23	48 640 139	48 645 364	Renpenning syndrome, non-syndromic MR	XL
85	*CACNA1F*	X	Xp11.23	48 948 467	48 976 777	X-linked incomplete congenital stationary night blindness, severe form	XL
86	*JARID1C*	X	Xp11.22	53 237 378	53 271 329	syndromic and non-syndromic X-linked MR	XL
87	*IQSEC2*	X	Xp11.22	53 278 783	53 367 247	non-syndromic X-linked MR	XL
88	*SMC1A*	X	Xp11.22	53 417 795	53 466 343	Cornelia de Lange syndrome	XL
89	*PHF8*	X	Xp11.22	53 979 838	54 087 036	Siderius-Hamel syndrome	*XL*
90	*FGD1*	X	Xp11.22	54 488 612	54 539 324	Aarskog-Scott syndrome, non-syndromic X-linked MR	XL
91	*OPHN1*	X	Xq12	67 178 911	67 570 024	MR with cerebellar and vermis hypoplasia	XL
92	*MED12*	X	Xq13.1	70 255 131	70 279 029	Lujan-Fryns syndrome	XL
93	*NLGN3*	X	Xq13.1	70 280 436	70 308 776	non-syndromic X-linked MR and ASD	XL
94	*KIAA2022*	X	Xq13.3	73 870 137	74 061 709	syndromic X-linked MR	XL
95	*ATRX*	X	Xq21.1	76 647 012	76 928 375	ATRX syndrome, non-syndromic X-linked MR	XL
96	*PCDH19*	X	Xq22.1	99 433 298	99 551 927	X-linked female-limited epilepsy and MR	XL
97	*ACSL4*	X	Xq22.3	108 771 220	108 863 277	non-syndromic X-linked MR	XL
98	*DCX*	X	Xq22.3	110 423 663	110 541 030	Type 1 lissencephaly	XL
99	*AGTR2*	X	Xq23	115 215 986	115 220 253	non-syndromic X-linked MR	XL
100	*UPF3B*	X	Xq24	118 852 017	118 870 996	non-syndromic X-linked MR	XL
101	*LAMP2*	X	Xq24	119 444 031	119 487 232	Danon disease	XL
102	*GRIA3*	X	Xq25	122 145 777	122 452 447	non-syndromic X-linked MR	XL

103	*OCRL*	X	Xq25	128 501 933	128 554 211	Lowe syndrome	XL
104	*PHF6*	X	Xq26.2	133 335 008	133 390 488	Borjeson-Forssman-Lehmann syndrome	XL
105	*SLC9A6*	X	Xq26.3	134 895 252	134 957 094	syndromic X-linked MR, Christianson type	XL
106	*ARHGEF6*	X	Xq26.3	135 575 377	135 691 169	non-syndromic X-linked MR	XL
107	*FMR1*	X	Xq27.3	146 801 201	146 840 333	Fragile X syndrome	XL
108	*AFF2*	X	Xq28	147 389 831	147 889 899	Fragile X mental retardation 2	XL
109	*SLC6A8*	X	Xq28	152 606 946	152 615 240	Creatine deficiency syndrome, non-syndromic X-linked MR	XL
110	*L1CAM*	X	Xq28	152 780 581	152 794 593	MASA (mental retardation, aphasia, shuffling gait, and adducted thumbs) syndrome	XL
111	*MECP2*	X	Xq28	152 940 458	153 016 382	Rett syndrome, non-syndromic X-linked MR (mutation, deletion); MECP2 duplication syndrome	XL
112	*RAB39B*	X	Xq28	154 140 720	154 147 046	non-syndromic X-linked MR	XL

Genomic coordinates correspond to the hg18 genome assembly (Build 36). Abbreviations: AD, autosomal dominant; ASD, autism spectrum disorder; AR, autosomal recessive; MR, mental retardation; XL, X-linked

Supplementary Table 6. ID genes

	Gene symbol	Chr	Locus	Start	End	Disorder/Phenotype	Inheritance pattern
1	*FUCA1*	1	1p36.11	24 044 159	24 067 408	Fucosidosis	AR
2	*SLC2A1*	1	1p34.2	43 163 633	43 197 434	Glucose transport defect	AD
3	*STIL*	1	1p33	47 488 398	47 552 406	Primary microcephaly	AR
4	*ALG6*	1	1p31.3	63 605 886	63 675 466	Congenital disorder of glycosylation, type Ic	AR
5	*DBT*	1	1p21.2	100 425 066	100 487 997	Maple syrup urine disease, type II	AR
6	*NRAS*	1	1p13.2	115 048 601	115 061 038	Noonan Syndrome	AD
7	*KCNJ10*	1	1q23.2	158 274 657	158 306 585	SESAME syndrome (seizures, sensorineural deafness, ataxia, MR, and electrolyte imbalance)	AR
8	*ASPM*	1	1q31	195 319 880	195 382 447	Microcephaly and MR	AR
9	*CRB1*	1	1q31.3	195 504 031	195 714 208	Leber congenital amaurosis 8	AR
10	*RD3*	1	1q32.3	209 717 404	209 732 882	Leber congenital amaurosis 12	AR
11	*TBCE*	1	1q42.3	233 597 351	233 678 903	Hypoparathyroidism-retardation-dysmorphism syndrome	AR
12	*FH*	1	1q43	239 727 527	239 749 677	Fumarase deficiency	AR
13	*MYCN*	2	2p24.3	15 998 134	16 004 580	Feingold syndrome (microcephaly-oculo-digito-esophageal-duodenal syndrome), Microcephaly and digital abnormalities with normal intelligence	AD
14	*SOS1*	2	2p22.1	39 062 194	39 201 108	Noonan Syndrome	AD
15	*ERCC3*	2	2q14.3	127 731 336	127 768 222	Trichothiodystrophy	AR
16	*RAB3GAP1*	2	2q21.3	135 526 323	135 644 016	Warburg Micro syndrome 1	AR
17	*ZEB2*	2	2q22.3	144 862 053	144 994 386	Mowat-Wilson syndrome (Hirschsprung disease-mental retardation syndrome)	AD
18	*BBS5*	2	2q31.1	170 044 252	170 071 411	Bardet-Biedl syndrome 5	AR
19	*GAD1*	2	2q31.1	171 381 446	171 425 905	Cerebral palsy, spastic, symmetric, autosomal recessive	AR
20	*HDAC4*	2	2q37.3	239 634 801	239 987 580	Brachydactyly mental retardation syndrome (2q37 deletion syndrome)	AD
21	*CRBN*	3	3p26.2	3 166 696	3 196 390	autosomal recessive non-syndromic MR	AR
22	*SUMF1*	3	3p26.2	4 377 830	4 483 954	Multiple sulfatase deficiency	AR
23	*TSEN2*	3	3p25.1	12 501 028	12 549 812	Pontocerebellar hypoplasia type 2B	AR
24	*RAF1*	3	3p25.1	12 600 100	12 680 700	Noonan Syndrome	AD
25	*TGFBR2*	3	3p24.1	30 622 998	30 710 637	Loeys–Dietz syndrome	AD
26	*GLB1*	3	3p22.3	33 013 104	33 113 698	GM1-gangliosidosis, Mucopolysaccharidosis IVB	AR
27	*ARL13B*	3	3q11.2	95 181 672	95 256 813	Joubert syndrome 8	AR
28	*ARL6*	3	3q11.2	98 966 285	99 000 063	Bardet-Biedl syndrome 3	AR
29	*ATR*	3	3q23	143 650 767	143 780 358	Seckel syndrome	AR
30	*ALG3*	3	3q27.1	185 442 811	185 449 440	Congenital disorder of glycosylation, type Id	AR
31	*KIAA0226*	3	3q29	198 882 656	198 948 170	syndromic MR with ataxia, dysarthria and epilepsy	AR
32	*IDUA*	4	4p16.3	970 785	988 317	Mucopolysaccharidosis Ih (Hurler syndrome)	AR
33	*CC2D2A*	4	4p15.3	15 080 587	15 212 278	Joubert syndrome 9, Meckel syndrome 6, COACH syndrome	AR
34	*QDPR*	4	4p15.32	17 097 121	17 122 811	Hyperphenylalaninemia due to dihydropteridine reductase deficiency	AR
35	*SRD5A3*	4	4q12	55 907 166	55 932 235	Kahrizi syndrome, type 1 congenital disorder of glycosylation	AR
36	*SLC4A4*	4	4q13.3	72 271 867	72 656 663	Renal tubular acidosis, proximal, with ocular abnormalities	AR
37	*BBS7*	4	4q27	122 965 085	123 011 092	Bardet-Biedl syndrome 7	AR
38	*BBS12*	4	4q27	123 873 307	123 885 548	Bardet-Biedl syndrome 12	AR
39	*LRAT*	4	4q32.1	155 884 613	155 893 720	Leber congenital amaurosis 14	AR
40	*AGA*	4	4q34.3	178 588 918	178 600 585	Aspartylglucosaminuria	AR
41	*ANKH*	5	5p15.2	14 757 909	14 924 887	Chondrocalcinosis 2, Craniometaphyseal dysplasia	AR
42	*MOCS2*	5	5q11.2	52 429 652	52 441 082	Molybdenum cofactor deficiency, type B	AR
43	*ERCC8*	5	5q12.1	60 205 415	60 276 662	Cockayne syndrome type A	AR
44	*TUBB2B*	6	6p25.2	3 169 514	3 172 870	Asymmetric polymicrogyria	AD
45	*NEU1*	6	6p21.3	31 934 808	31 938 688	Sialidosis type I and type II	AR
46	*MOCS1*	6	6p21.2	39 980 024	40 003 433	Molybdenum cofactor deficiency, type A	AR
47	*SLC17A5*	6	6q13	74 359 823	74 420 458	Salla disease, Sialic acid storage disorder, infantile	AR
48	*LCA5*	6	6q14.1	80 251 427	80 303 844	Leber congenital amaurosis 5	AR

49	*BCKDHB*	6	6q14.1	80 873 063	81 112 706	Maple syrup urine disease, type Ib	AR
50	*GRIK2*	6	6q16.3	101 953 626	102 624 651	autosomal recessive non-syndromic MR	AR
51	*SOBP*	6	6q21	107 918 010	108 089 206	autosomal recessive syndromic and nonsyndromic MR	AR
52	*LAMA2*	6	6q22.33	129 245 979	129 879 403	Merosin-deficient congenital muscular dystrophy type 1A	AR
53	*ARG1*	6	6q23.2	131 936 058	131 947 161	Argininemia	AR
54	*PEX7*	6	6q23.3	137 185 416	137 276 752	Refsum disease, Rhizomelic chondrodysplasia punctata, type 1	AR
55	*GTF2H5*	6	6q25.3	158 511 490	158 533 364	Trichothiodystrophy	AR
56	*BBS9*	7	7p14.3	33 135 677	33 612 205	Bardet-Biedl syndrome 9	AR
57	*C7orf11*	7	7p14.1	40 138 867	40 140 783	Trichothiodystrophy	AR
58	*GUSB*	7	7q11.21	65 063 108	65 084 681	Mucopolysaccharidosis VII	AR
59	*AP4M1*	7	7q22.1	99 537 066	99 542 739	autosomal recessive tetraplegic cerebral palsy with MR	AR
60	*RELN*	7	7q22	102 899 473	103 417 198	Lissencephaly	AR
61	*DLD*	7	7q31.1	107 318 822	107 348 879	Maple syrup urine disease, type III	AR
62	*IMPDH1*	7	7q32.1	127 819 567	127 837 272	Leber congenital amaurosis 11	AD
63	*MCPH1*	8	8p23	6 251 529	6 493 434	Microcephaly and MR	AR
64	*TUSC3*	8	8p22	15 442 101	15 666 366	autosomal recessive non-syndromic MR	AR
65	*TMEM67*	8	8q21	94 836 269	94 899 523	Joubert syndrome 6, Meckel-Gruber syndrome	AR
66	*KCNK9*	8	8q24.3	140 692 762	140 784 481	Birk-Barel mental retardation dysmorphism syndrome, genomic-imprinting syndrome	AD
67	*TRAPPC9*	8	8q24.3	140 811 770	141 537 860	autosomal recessive non-syndromic MR	AR
68	*VLDLR*	9	9p24.2	2 611 793	2 644 485	Cerebellar ataxia and MR	AR
69	*TGFBR1*	9	9q22.33	100 907 233	100 956 294	Loeys–Dietz syndrome	AD
70	*FKTN*	9	9q31.2	107 360 232	107 443 220	Fukuyama congenital muscular dystrophy with type 2 lissencephaly, Walker-Warburg syndrome	AR
71	*TRIM32*	9	9q33.1	118 489 402	118 503 400	Bardet-Biedl syndrome 11	AR
72	*CDK5RAP2*	9	9q33.2	122 190 968	122 382 258	Microcephaly vera	AR
73	*SPTAN1*	9	9q34.11	130 354 687	130 435 761	West syndrome with severe cerebral hypomyelination, spastic quadriplegia and MR	AD
74	*INPP5E*	9	9q34.3	138 442 893	138 454 077	Joubert syndrome 1	AR
75	*ERCC6*	10	10q11.23	50 334 497	50 417 153	Cockayne syndrome type B, Cerebro-oculo-facio-skeletal syndrome	AR
76	*KIAA1279*	10	10q21.3	70 418 499	70 446 742	Goldberg-Shprintzen megacolon syndrome	AR
77	*SMC3*	10	10q25.2	112 317 439	112 354 382	Cornelia de Lange syndrome	AD
78	*SHOC2*	10	10q25.2	112 713 873	112 763 413	Noonan Syndrome	AD
79	*SLC25A22*	11	11p15.5	780 475	788 235	autosomal recessive neonatal epileptic encephalopathy	AR
80	*PAX6*	11	11p13	31 762 916	31 796 085	isolated and syndromic aniridia, including Gillespie syndrome (aniridia, cerebellar ataxia and MR)	AD
81	*SLC35C1*	11	11p11.2	45 783 912	45 791 143	Congenital disorder of glycosylation, type IIc	AR
82	*TMEM216*	11	11q12.2	60 916 441	60 922 899	Joubert syndrome 2	AR
83	*BBS1*	11	11q13.1	66 034 695	66 057 660	Bardet-Biedl syndrome 1	AR
84	*ALG8*	11	11q14.1	77 489 636	77 528 347	Congenital disorder of glycosylation type Ih	AR
85	*MED17*	11	11q21	93 157 053	93 186 144	Primary microcephaly of postnatal onset, spasticity, epilepsy, and profound MR	AR
86	*ALG9*	11	11q23.1	111 158 129	111 247 515	Congenital disorder of glycosylation, type Il	AR
87	*CBL*	11	11q23.3	118 582 200	118 684 069	Noonan syndrome-like phenotype	AD
88	*PVRL1*	11	11q23.3	119 036 913	119 104 645	Cleft lip/palate ectodermal dysplasia syndrome	AR
89	*KIRREL3*	11	11q24.2	125 799 613	126 375 976	autosomal dominant non-syndromic MR	AD
90	*MLL2*	12	12q13.12	47 699 025	47 735 374	Kabuki syndrome	AD
91	*TUBA1A*	12	12q13.12	47 864 850	47 869 128	Lissencephaly	AD
92	*DIP2B*	12	12q13.13	49 185 035	49 428 717	Mental retardation, FRA12A type	XL
93	*SUOX*	12	12q13.2	54 677 310	54 685 576	Sulfite oxidase deficiency	AR
94	*BBS10*	12	12q21.2	75 262 397	75 266 353	Bardet-Biedl syndrome 10	AR
95	*GNPTAB*	12	12q23.2	100 663 408	100 748 763	Mucolipidosis III alpha/beta	AR
96	*IGF1*	12	12q23.2	101 335 584	101 398 508	Growth retardation with deafness and MR due to IGF1 deficiency	AR
97	*ATP6V0A2*	12	12q24.31	122 762 818	122 810 393	Cutis laxa with epilepsy and mental retardation	AR
98	*CENPJ*	13	13q12.12	24 354 412	24 395 085	Microcephaly vera, Seckel syndrome	AR
99	*SLC25A15*	13	13q14.11	40 261 597	40 282 246	Hyperornithinemia-hyperammonemia-homocitrullinemia syndrome	AR

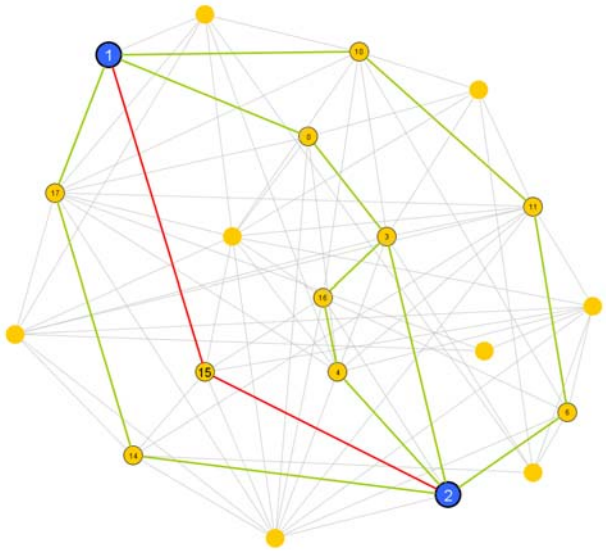
100	*ERCC5*	13	13q33.1	102 257 497	102 322 749	Cockayne syndrome, Cerebro-oculo-facio-skeletal syndrome	AR
101	*COL4A1*	13	13q34	109 599 311	109 757 497	Porencephaly	AD
102	*RPGRIP1*	14	14q11.2	20 825 976	20 889 300	Leber congenital amaurosis 6	AR
103	*MGAT2*	14	14q22.1	49 157 239	49 159 949	Congenital disorder of glycosylation, type IIa	AR
104	*RDH12*	14	14q24.1	67 238 356	67 270 921	Leber congenital amaurosis 13	AR
105	*POMT2*	14	14q24.3	76 811 054	76 856 978	Walker-Warburg syndrome	AR
106	*GALC*	14	14q31.3	87 469 111	87 529 660	Krabbe disease	AR
107	*SPATA7*	14	14q31.3	87 921 495	87 974 557	Leber congenital amaurosis 3	AR
108	*TTC8*	14	14q31.3	88 360 671	88 414 088	Bardet-Biedl syndrome 8	AR
109	*VRK1*	14	14q32.2	96 333 437	96 417 704	Pontocerebellar hypoplasia type 1	AR
110	*SPRED1*	15	15q14	36 332 344	36 436 742	Neurofibromatosis type 1-like syndrome /Legius syndrome	AD
111	*CEP152*	15	15q21.1	46 817 640	46 890 476	Primary microcephaly	AR
112	*AP4E1*	15	15q21.2	48 988 238	49 085 389	novel autosomal recessive cerebral palsy syndrome with microcephaly and MR	AR
113	*BBS4*	15	15q24.1	70 765 588	70 817 869	Bardet-Biedl syndrome 4	AR
114	*GNPTG*	16	16p13.3	1 341 933	1 353 353	Mucolipidosis III gamma	AR
115	*TBC1D24*	16	16p13.3	2 465 148	2 493 489	autosomal recessive syndrome of focal epilepsy, dysarthria, and MR	AR
116	*PMM2*	16	16p13.2	8 799 171	8 850 695	Congenital disorder of glycosylation, type Ia	AR
117	*BBS2*	16	16q13	55 075 799	55 111 696	Bardet-Biedl syndrome 2	AR
118	*GPR56*	16	16q13	56 220 023	56 256 445	autosomal recessive bilateral frontoparietal polymicrogyria	AR
119	*COG8*	16	16q22.1	67 920 025	67 931 027	Congenital disorder of glycosylation, type IIh	AR
120	*CDH15*	16	16q24.3	87 765 664	87 789 401	autosomal dominant non-syndromic MR	AD
121	*AIPL1*	17	17p13.2	6 267 783	6 279 243	Leber congenital amaurosis 4	AR
122	*MPDU1*	17	17p13.1	7 427 854	7 432 247	Congenital disorder of glycosylation, type If	AR
123	*SLC46A1*	17	17q11.2	23 745 788	23 757 355	Folate malabsorption	AR
124	*GFAP*	17	17q21.31	40 338 519	40 348 394	Alexander disease	AD
125	*MKS1*	17	17q22	53 637 797	53 651 665	Bardet-Biedl syndrome 13, Meckel syndrome 1	AR
126	*COG1*	17	17q25.1	68 700 768	68 716 240	Congenital disorder of glycosylation, type IIg	AR
127	*TSEN54*	17	17q25.1	71 024 204	71 032 415	Pontocerebellar hypoplasia type 2A	AR
128	*SETBP1*	18	18q12.3	40 535 138	40 898 771	Schinzel-Giedion syndrome	AD
129	*TCF4*	18	18q21.2	51 040 560	51 406 858	Pitt-Hopkins syndrome	AD
130	*MAP2K2*	19	19p13.3	4 041 320	4 075 126	Cardio-facio-cutaneous syndrome	AD
131	*MCOLN1*	19	19p13.2	7 493 512	7 504 863	Mucolipidosis IV	AR
132	*CC2D1A*	19	19p13.12	13 878 052	13 902 692	autosomal recessive non-syndromic MR	AR
133	*WDR62*	19	19q13.12	41 237 623	41 287 852	severe brain malformations, including microcephaly, pachygyria and hypoplasia of the corpus callosum	AR
134	*BCKDHA*	19	19q13.2	46 595 544	46 622 750	Maple syrup urine disease, type Ia	AR
135	*ERCC2*	19	19q13.32	50 546 686	50 565 669	Cockayne syndrome, Trichothiodystrophy, Cerebro-oculo-facio-skeletal syndrome	AR
136	*ERCC1*	19	19q13.32	50 608 532	50 618 642	Cerebro-oculo-facio-skeletal syndrome	AR
137	*FKRP*	19	19q13.32	51 941 143	51 953 582	Congenital muscular dystrophy 1C, limb-girdle muscular dystrophy type 2I, muscle-eye-brain disease, Walker-Warburg syndrome	AR
138	*CRX*	19	19q13.32	53 016 911	53 038 398	Leber congenital amaurosis 7	AD
139	*PNKP*	19	19q13.33	55 056 273	55 062 630	Microcephaly, seizures and defects in DNA repair	AR
140	*DNMT3B*	20	20q11.2	30 813 852	30 860 823	ICF syndrome (immune deficiency, centromeric instability, facial dysmorphy and MR)	AR
141	*CTSA*	20	20q13.12	43 952 998	43 960 865	Galactosialidosis	AR
142	*ARFGEF2*	20	20q13.13	46 971 682	47 086 637	autosomal recessive periventricular heterotopia with microcephaly	AR
143	*DPM1*	20	20q13.13	48 984 812	49 008 467	Congenital disorder of glycosylation, type Ie	AR
144	*CBS*	21	21q22.3	43 346 370	43 369 493	Homocystinuria	AR
145	*PCNT*	21	21q22.3	46 568 464	46 690 110	Seckel syndrome, Majewski osteodysplastic primordial dwarfism type II	AR
146	*SNAP29*	22	22q11.21	19 543 292	19 574 109	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome	AR
147	*LARGE*	22	22q12.3	31 999 062	32 646 416	Congenital muscular dystrophy	AR
148	*EP300*	22	22q13.2	39 818 560	39 906 027	Rubinstein-Taybi syndrome	AD
149	*ALG12*	22	22q13.33	48 682 857	48 698 110	Congenital disorder of glycosylation, type Ig	AR

150	*HCCS*	X	Xp22.2	11 039 336	11 051 122	Microphthalmia with linear skin defects syndrome	XL
151	*OFD1*	X	Xp22.2	13 662 753	13 697 401	Oral-facial-digital syndrome type I, Simpson-Golabi-Behmel syndrome, type 2, Joubert syndrome 10	XL
152	*FANCB*	X	Xp22.2	14 771 450	14 801 105	VACTERL with hydrocephalus, Fanconi anemia of complementation group B	XL
153	*PDHA1*	X	Xp22.12	19 271 932	19 289 723	Pyruvate decarboxylase deficiency	XL
154	*RPS6KA3*	X	Xp22.12	20 077 950	20 194 671	Coffin-Lowry syndrome, non-syndromic MR	XL
155	*SMS*	X	Xp22.11	21 868 763	21 922 876	Snyder-Robinson syndrome	XL
156	*GK*	X	Xp21.2	30 581 397	30 658 646	Glycerol kinase deficiency	XL
157	*TSPAN7*	X	Xp11.4	38 305 675	38 433 116	non-syndromic X-linked MR	XL
158	*BCOR*	X	Xp11.4	39 795 443	39 841 663	syndromic Lenz microphthalmia-2, oculofaciocardiodental syndrome	XL
159	*ATP6AP2*	X	Xp11.4	40 325 160	40 350 832	X-linked MR with epilepsy	XL
160	*MAOA*	X	Xp11.3	43 400 353	43 491 012	Brunner syndrome (monoamine oxidase A deficiency)	XL
161	*ZNF41*	X	Xp11.3	47 190 505	47 227 289	non-syndromic X-linked MR	XL
162	*PORCN*	X	Xp11.23	48 252 315	48 264 146	Focal dermal hypoplasia	XL
163	*SYP*	X	Xp11.23	48 931 209	48 943 605	non-syndromic X-linked MR	XL
164	*SHROOM4*	X	Xp11.22	50 351 387	50 573 784	Stocco dos Santos X-linked MR syndrome, non-syndromic XLMR	XL
165	*HSD17B10*	X	Xp11.22	53 474 931	53 478 048	2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency	XL
166	*HUWE1*	X	Xp11.22	53 575 797	53 730 398	non-syndromic and syndromic X-linked MR	XL
167	*KLF8*	X	Xp11.21	56 275 632	56 328 254	non-syndromic X-linked MR	XL
168	*ARHGEF9*	X	Xq11.1	62 771 573	62 891 756	syndromic X-linked MR, hyperekplexia and epilepsy	XL
169	*IGBP1*	X	Xq13.1	69 270 043	69 302 898	syndromic X-linked MR, agenesis of the corpus callosum, ocular coloboma, and micrognathia	XL
170	*DLG3*	X	Xq13.1	69 581 449	69 642 062	non-syndromic X-linked MR	XL
171	*SLC16A2*	X	Xq13.2	73 557 810	73 670 477	T3 transporter deficiency; syndromic and non-syndromic MR	XL
172	*MAGT1*	X	Xq21.1	76 968 520	77 037 721	non-syndromic X-linked MR	XL
173	*ATP7A*	X	Xq21.1	77 052 850	77 192 548	Menkes disease, occipital horn syndrome	XL
174	*PGK1*	X	Xq21.1	77 246 322	77 268 980	Phosphoglycerate kinase-1 deficiency	XL
175	*BRWD3*	X	Xq21.1	79 818 339	79 951 889	non-syndromic X-linked MR	XL
176	*ZNF711*	X	Xq21.1	84 385 653	84 415 025	non-syndromic X-linked MR	XL
177	*SRPX2*	X	Xq22.1	99 785 819	99 812 952	X-linked Rolandic epilepsy, speech dyspraxia and MR	XL
178	*TIMM8A*	X	Xq22.1	100 487 306	100 490 343	Mohr-Tranebjaerg syndrome, Jensen syndrome	XL
179	*PLP1*	X	Xq22.2	102 918 095	102 934 203	Pelizaeus-Merzbacher disease	XL
180	*PRPS1*	X	Xq22.3	106 758 310	106 780 912	Phosphoribosylpyrophosphate synthetase I superactivity	XL
181	*PAK3*	X	Xq22.3	110 252 961	110 350 829	non-syndromic X-linked MR	XL
182	*UBE2A*	X	Xq24	118 592 527	118 602 407	syndromic X-linked MR, seizures, speech impairment, and hirsutism	XL
183	*NDUFA1*	X	Xq24	118 889 762	118 894 657	Mitochondrial complex I deficiency (syndromic X-linked MR)	XL
184	*CUL4B*	X	Xq24	119 542 474	119 593 712	non-syndromic X-linked MR	XL
185	*ZDHHC9*	X	Xq25	128 766 594	128 805 554	non-syndromic X-linked MR (Marfanoid habitus)	XL
186	*GPC3*	X	Xq26.2	132 497 442	132 947 332	Simpson-Golabi-Behmel syndrome type 1	XL
187	*HPRT1*	X	Xq26.2	133 421 841	133 462 364	Lesch-Nyhan syndrome	XL
188	*SOX3*	X	Xq27.1	139 412 818	139 414 891	Isolated GH deficiency, short stature and MR	XL
189	*IDS*	X	Xq28	148 368 203	148 394 769	Mucopolysaccharidosis II (Hunter syndrome)	XL
190	*NSDHL*	X	Xq28	151 750 167	151 788 563	CK syndrome	XL
191	*ABCD1*	X	Xq28	152 643 517	152 663 410	Adrenoleukodystrophy	XL
192	*AVPR2*	X	Xq28	152 823 622	152 825 814	X-linked nephrogenic diabetes insipidus	XL
193	*FLNA*	X	Xq28	153 230 094	153 256 200	Bilateral periventricular nodular heterotopia, otopalatodigital syndrome, frontometaphyseal dysplasia	XL
194	*GDI1*	X	Xq28	153 318 453	153 325 009	non-syndromic X-linked MR	XL
195	*IKBKG*	X	Xq28	153 423 653	153 446 455	Incontinentia pigmenti	XL
196	*DKC1*	X	Xq28	153 644 225	153 659 157	Dyskeratosis congenita	XL

Genomic coordinates correspond to the hg18 genome assembly (Build 36). Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MR, mental retardation; XL, X-linked

Evaluating distance between gene lists

In a given background network, for example the human protein-protein interactome, the distance between two nodes can be defined as the shortest path between the nodes. The distance is the minimum number of intermediates connecting the two nodes in the shortest path (Supplementary Fig. 2).



Supplementary Figure 2. Shortest path between node 1 and node 2. Red edges highlight the shortest path between node 1 and node 2, which goes through one intermediate node, node 15. The distance is defined as $D_{12}=1$. The green edges show other possible paths in the given background network.

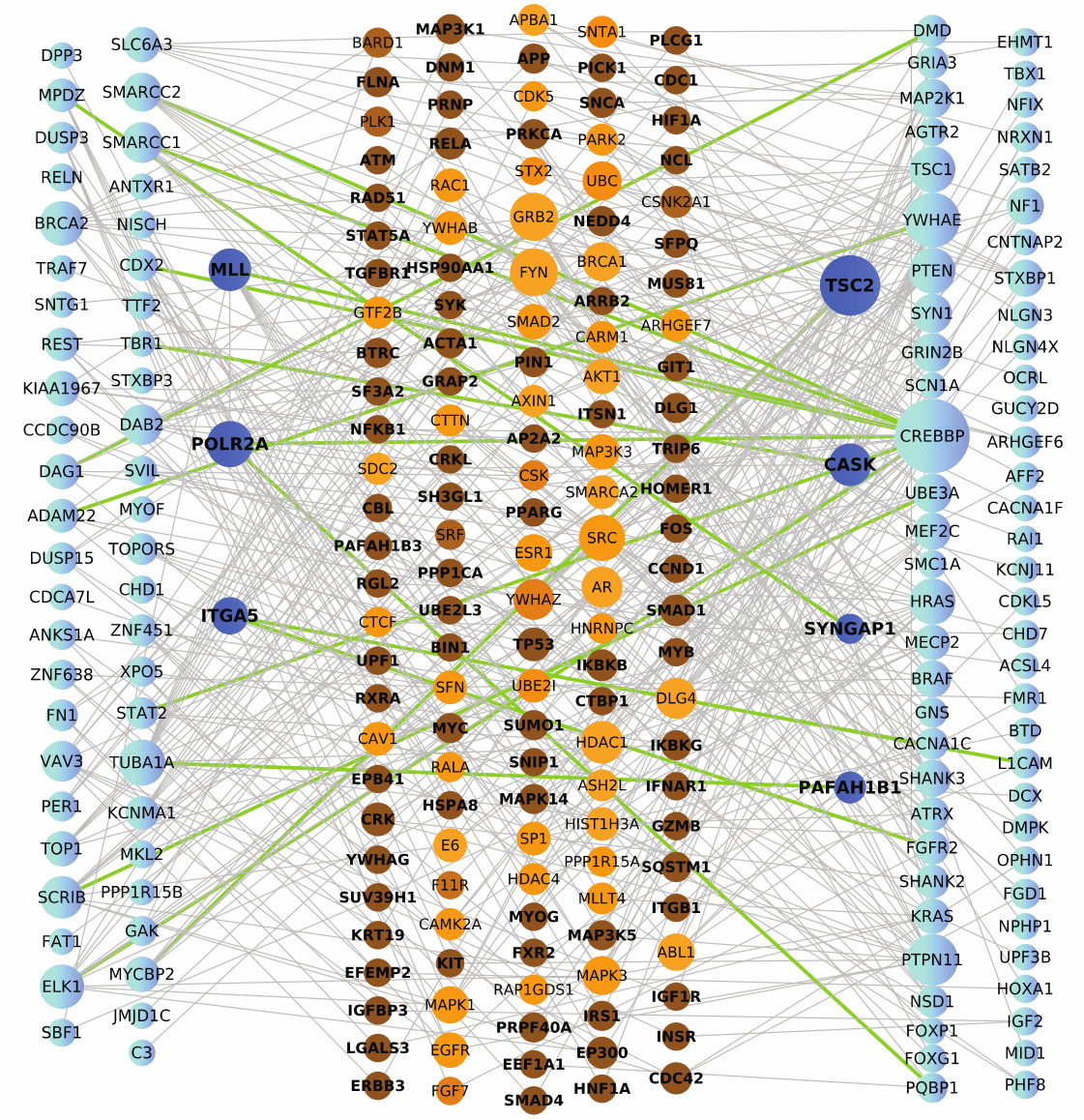
Now consider the average distance between a gene list and a reference list. Let n_1 and n_2 be the number of genes in list 1 (L_1) and the reference list (L_R), respectively. For a given background network, let D_{ij} be the shortest distance between the i'th gene in L_1 and the j'th gene in L_R. The distance between gene i in L_1 and the set of genes in L_R is defined as the mean distance between this gene and each gene in L_R : $D_i = 1/n_2 \Sigma_j D_{ij}$. Define the average distance between list L_1 and list L_R as $D(L_1) = 1/n_1 \Sigma_i D_i$. The standard error of $D(L_1)$ is defined as $SE(L_1) = s/(n_1)^{1/2}$, where s is the standard deviation of the set of distances $\{D_i\}$.

To address the question whether either the genes in the list defined by the cases (L_1) or the controls (L_2) is closer to a reference list (L_R, e.g., a list of synapse genes), a two-sample t test can be used. Under the null hypothesis, the average distance from L_1 to L_R is similar to the distance from L_2 to L_R versus the alternative, which says the average distance is shorter (i.e., a one-sided hypothesis). The t statistic (for unequal variances) can be computed as follows: $t = [D(L_1) - D(L_2)]/[SE(L_1)^2 + SE(L_2)^2]^{1/2}$. To allow for modest violations of the assumption of normality, one-sided p-values are obtained via a permutation test, creating an appropriate empirical distribution of test statistics.

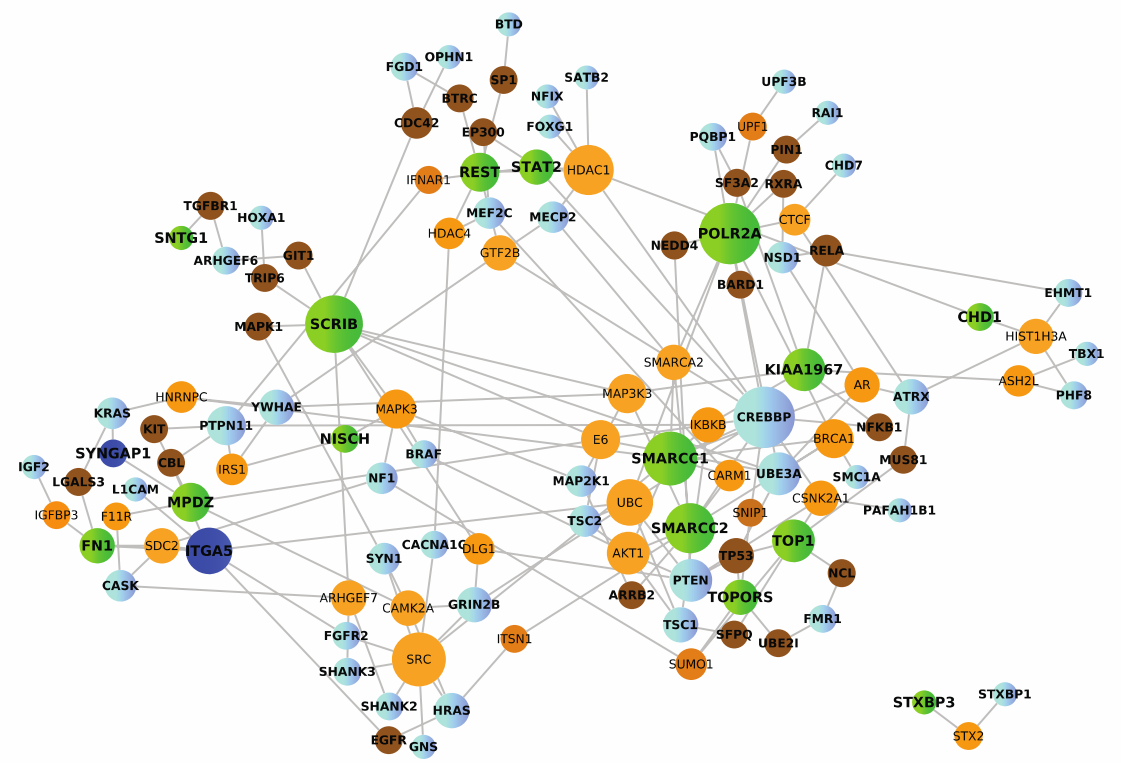
For the background network we used a highly connected PPI network composed from the following databases: BioGrid (PMID: 21071413), MINT (PMID: 19897547), KEGG (PMID: 18428742), PPID (PMID: 14755292), HPRD (PMID: 18988627), DIP (PMID: 11752321, BIND (PMID: 21233089), IntAct (PMID: 19850723), InnateDB (PMID: 18766178), and SNAVI (PMID: 19154595). Interactions from those online resources provided direct physical PPIs identified experimentally. The consolidated dataset from these PPI databases was filtered to include publications with only a maximum of 10 interactions from a single publication. Because the shortest path between two genes can only be found when they are connected, we restricted our analysis to genes in this network. Thus, the actual sample sizes (i.e. n_1 and n_2) in the t test is the number of genes in the lists found in the network. In addition, for any comparison of lists, overlapping genes were removed.

D_i was calculated for de novo variants using ASD112 and ASD31 (defined above). The ASD112 and even more so ASD31 lists are enriched for brain expressed genes. As brain expressed genes are more connected to other brain expressed genes, we consider that we can reduce bias with a focus on brain-expressed *de novo* variants, so we restricted the analyses of novel variants to those in brain expressed genes. Results for ASD112 are presented below (Supplementary Figure 3) and for ASD31 in the main text. Direct interactions between genes with de novo mutations and reference ASD genes as well as immediate intermediates (first neighbors) were included for subnetwork reconstruction in the figures.

We also developed subnetworks with genes identified in DAPPLE analysis and genes in the ASD112 list (Supplementary Figure 4).



Supplementary Figure 3. PPI network analysis for de novo variants and prior ASD genes. Nodes are sized based on connectivity. Genes harboring de novo variants (left) and prior ASD genes (ASD112, right) are colored blue with dark blue nodes represent genes that belong to one of these lists and are also intermediate proteins. Intermediate proteins (center) are colored in shades of orange based on a p-value computed using a proportion test where darker color represents a lower p-value. Green edges represent direct connections between genes harboring de novo variants (left) and prior ASD genes.



Supplementary Figure 4. Relationship between DAPPLE genes and prior ASD genes (ASD112). Subnetworks were created by connecting the genes identified in DAPPLE (see Figure 1 in the main text) with previous ASD genes (Supplementary Table 1) using known high confidence protein-protein interactions (PPI). Clusters within the topology of the subnetwork were identified automatically using the organic option implemented within the yEd software. Nodes are sized based on connectivity. Genes identified in DAPPLE are colored in green, previously known ASD genes are colored in light blue, dark blue nodes represent genes that belong to either the DAPPLE genes (*ITGA5*) or ASD genes (*SYNGAP1*) and are also intermediate proteins. Intermediate proteins are colored in shades of orange and are shaded based on a p-value computed using a proportion test where darker color represents more significance.

Estimation of probability of hitting a gene multiple times

Based on the evolutionary model, we generated a set of ~60,000 random mutations. From this set, we drew random subsets of mutations according to the number of observed events. For each set of mutations, we then counted the number of instances for which the same gene was hit more than once. The frequency of the multiple hits is shown in Supplementary Table 7.

Supplementary Table 7: Simulated *de novos* and number of genes hit

No. Events	2 hits	3 hits	4 hits	≥5 hits
100	0.58	0.008	0.001	0
150	1.32	0.0279	0.0017	0.0002
200	2.2756	0.0573	0.0037	0.0003
350	6.6854	0.2485	0.0274	0.0072
439	11.9156	0.4198	0.0188	0.0186
500	12.9861	0.598	0.0799	0.0259

Legend. The number of simulated *de novo* events is shown in the first column. The subsequent columns, 2, 3, 4, and 5+ hits refer to the expected number of genes across with that number of *de novos* across 10,000 draws from the random mutations. We determined the significance of 18 double hit genes by simulating a set of 439 mutations (the total number of missense mutations observed across all three datasets). Across 10,000 draws, we determined that that the p-value is 0.0632 by counting the number of instances that 18 or more genes were hit by two mutations.

Analysis of Case-Control Data

Based on the ARRA exome sequencing of unrelated cases and controls (Baylor cases: 440 males, 65 females; Baylor controls: 240 males, 251 females: Broad cases: 344 males, 86 females; Broad controls: 177 males, 202 females), we have 935 cases and 870 controls available for analysis of association with rare variants. We performed association tests using SKAT [15] with gender as a covariate. Analysis was restricted to non-synonymous variants with minor allele frequency less than 0.01. Functional singleton variants were pooled to create a single additional variant. Analysis was initially performed on data from each sequencing center separately due to differences in the sequencing and variant calling routines. Results of the full study were obtained by combining these statistics (meta analysis) and combining the full set of data across sequencing centers (mega analysis). In the mega analysis we removed any variants with minor allele frequency greater than 0.01 in either individual data set. We restricted our analysis to the 18 genes with double non-synonymous hits across all three sources of data (Supplementary Table 8A,B,C). We included gender in the model to control for the strong unbalance in the gender distribution among cases. For example, in TUBA1A only two singleton variants were observed, but both were in female cases, enhancing the significance of this observation.

Supplementary Table 8A. Rare variant distribution in Baylor case-control sample for double hit de novo genes.

Gene	Non-Singleton Variant No.	Singleton Count	Non Singleton Count	P-value
BRCA2	21	41:23	50:38	0.027
CHD8	5	18:13	6:7	0.195
DNAH5	23	47:44	44:51	0.913
FAT1	27	52:55	55:58	0.925
KATNAL2	1	6:5	0:2	0.727
KIAA0100	11	16:11	45:39	0.565
KIAA0182	6	8:14	13:9	0.181
MEGF11	2	8:5	5:5	0.514
MYO7B	10	22:25	14:16	0.389
NTNG1	0	8:3	0:0	0.250
RFX8	8	7:5	31:25	0.804
SBF1	2	11:5	1:3	0.292
SCN2A	1	19:7	2:1	0.013
SLCO1C1	3	10:2	6:11	0.015
SUV420H1	4	4:2	8:8	0.777
TBR1	2	1:2	4:0	0.489
TRIO	2	8:14	2:2	0.142
TUBA1A	1	0:0	2:0	0.005

Supplementary Table 8B. Rare variant distribution in Broad case-control sample for double hit de novo genes.				
Gene	Non-Singleton Variant No.	Singleton Count	Non Singleton Count	P-value
BRCA2	18	18:16	48:57	0.658
CHD8	3	4:9	2:7	0.092
DNAH5	23	37:27	49:47	0.089
FAT1	33	44:31	104:85	0.710
KATNAL2	4	3:4	6:5	0.798
KIAA0100	8	7:10	31:27	0.694
KIAA0182	19	22:14	59:37	0.234
MEGF11	4	11:4	8:10	0.121
MYO7B	21	24:24	49:33	0.969
NTNG1	2	5:1	3:1	0.219
RFX8	1	3:1	1:3	0.113
SBF1	22	25:14	39:41	0.161
SCN2A	6	8:9	19:15	0.965
SLCO1C1	1	2:3	6:3	0.581
SUV420H1	5	3:4	9:6	0.673
TBR1	1	6:5	1:5	0.383
TRIO	5	23:19	10:10	0.283
TUBA1A	3	1:0	3:6	0.360

Supplementary Table 8C. Rare variant distribution in combined case-control sample for double hit de novo genes.

Gene	Damaging	Non Singleton Variant No.	Non Singleton Count	Singleton Count	Nonsense (Stop/Gain) Count	Mega P-value	Meta P-value
BRCA2	yes	34	49:35	108:99	15:17	0.212	0.108
CHD8	yes	7	22:18	8:18	0:0	0.400	0.066
DNAH5	yes	40	74:64	103:105	1:3	0.479	0.582
FAT1	yes	51	80:71	168:146	1:0	0.785	0.929
KATNAL2	yes	5	8:8	7:8	1:0	0.827	0.840
KIAA0100	yes	15	20:17	79:70	2:0	0.951	0.672
KIAA0182	yes	25	25:27	79:47	1:2	0.712	0.122
MEGF11	yes	6	16:7	16:17	1:0	0.099	0.239
MYO7B	yes	29	42:45	67:53	8:6	0.598	0.848
NTNG1	yes	2	13:3	3:2	0:0	0.040	0.156
RFX8	yes	8	10:6	24:20	1:0	0.440	0.461
SBF1	yes	24	36:18	40:45	3:3	0.047	0.147
SCN2A	yes	8	26:14	22:18	1:2	0.048	0.281
SLCO1C1	yes	3	12:5	12:14	1:1	0.138	0.060
SUV420H1	yes	7	7:5	17:15	0:0	0.917	0.809
TBR1	yes	3	6:6	6:6	0:0	0.531	0.417
TRIO	yes	11	26:27	17:18	0:4	0.847	0.117
TUBA1A	yes	4	1:0	5:6	0:0	0.161	0.014

Supplementary Table 8. Summary of rare variant distribution in ARRA case control samples for genes with double hit non-synonymous *de novo* variants. “Non-Singleton Variant No” is the total number of non-singleton coding rare variant sites recorded; "Singleton Count" is the total number of non-synonymous singleton realizations added over all the cases and all the controls, reported as case:control; "Non-Singleton Count" is the total number of non-synonymous non-singleton realizations added over all cases and and all controls; “Nonsense (stop gain) Count” is total number of nonsense and splice site realizations added over all cases and all controls. P-values are derived from the SKAT statistic. The data summaries provided are not utilized directly in this test statistic.

Supplementary Table 9. Count of transmitted (T) or untransmitted (U) rare variants in genes hit with two, functional de *novo* mutations.

Gene	Nonsense		Missense Singletons		Missense Singletons	
	Combined T	Combined U	Combined T	Combined U	Combined T	Combined U
RFX8	1	2	1	0	7	9
KIAA0182	0	0	0	1	1	5
CHD8	0	0	2	1	6	5
BRCA2	2	2	10	7	116	112
TUBA1A	0	0	0	0	0	0
TBR1	0	0	1	1	2	1
SBF1	1	0	1	7	11	16
SLCO1C1	0	0	1	0	91	81
KATNAL2	0	0	1	2	76	71
DNAH5	0	0	9	12	543	562
KIAA0100	0	0	7	3	40	26
MEGF11	0	0	1	1	24	26
SCN2A	0	0	2	1	21	20
TRIO	0	0	4	3	6	6
MYO7B	0	0	5	5	64	55
NTNG1	0	0	0	0	1	0
SUV420H1	0	0	0	0	2	3
FAT1	0	0	9	19	615	716

Supplementary Table 10: Case control counts for all genes with *de novo* loss of function alleles

Gene	Cases Lof	Controls Lof	Gene	Cases LoF	LoF Controls
ADAM33	41	33	NAPRT1	0	0
ADNP	1	0	PDCD1	0	0
APH1A	0	0	PLXNB1	1	0
ARID1B	0	0	POGZ	0	0
ASAH2	11	12	POLRMT	0	0
BRSK2	0	1	PPM1D	1	0
BRWD1	0	1	PPP1R15B	0	0
BTN1A1	1	2	PRPF39	0	0
C20orf111	0	1	RAB2A	0	1
CDC42BPB	1	0	RELN	0	1
CHD8	3	0	RNF38	0	0
CNOT3	0	0	RPS6KA3	0	0
COL25A1	3	0	SCN2A	0	0
CSDE1	1	0	SCP2	0	0
CUBN	1	1	SETBP1	0	0
CUL3	2	0	SETD2	0	0
DHRS4L1	1	0	SHANK2	0	0
DNAH5	3	3	SLC7A7	1	0
DYRK1A	0	0	SMARCC2	0	0
EPHB2	0	0	SPAST	0	0
ETFB	1	0	SPP2	1	1
FAM8A1	0	0	ST3GAL6	1	0
FCRL6	3	3	SUCLA2	0	0
FREM3	0	0	SVIL	0	0
IQGAP2	3	3	TBR1	0	0
ITGA5	0	0	TBX18	0	0
KATNAL2	3	0	TCF3	0	0
KIAA0100	2	0	TRPM5	1	0
LTN1	0	2	TSPAN17	0	0
MBD5	0	0	UBR3	0	0
MLL3	0	1	USP15	1	1
MPHOSPH8	0	0	ZNF292	0	0
MTMR12	0	0			

Table 10 Legend: Included in case and control tallies here are 935 cases and 870 controls as well as 104 trios where transmitted alleles are considered case counts and nontransmitted alleles are considered control counts

Supplemental References

1. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-8 (2011).
2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-95 (2010).
3. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-303 (2010).
4. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80, 727-39 (2007).
5. Krawczak, M., Ball, E. V. & Cooper, D. N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63, 474-488 (1998).
6. Conrad, D. F. et al. Variation in genome-wide mutation rates within and between human families. Nat Genet 43, 712-714 (2011).
7 Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res 1380, 42-77 (2011).
8 Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466, 368-372 (2010).
9. Collins, M. O. et al. Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. J Neurochem 97 Suppl 1, 16-23 (2006).
10. Abul-Husn, N. S. et al. Systems approach to explore components and interactions in the presynapse. Proteomics 9, 3303-15 (2009).
11. Bayes, A. et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat Neurosci 14, 19-21 (2011).
12. Lachmann A, Ma'ayan A. Lists2Networks: integrated analysis of gene/protein lists. BMC Bioinformatics. Feb 12, 11-87 (2010)
13. Boyko, A. R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4, e1000083 (2008).
14. McCarthy, S. E. et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41, 1223-7 (2009).
15. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93 (2011).