nature genetics

Excess of rare, inherited truncating mutations in autism

Niklas Krumm[1,5], Tychele N Turner[1,5], Carl Baker[1], Laura Vives[1], Kiana Mohajeri[1], Kali Witherspoon[1], Archana Raja[1,2], Bradley P Coe[1], Holly A Stessman[1], Zong-Xiao He[3], Suzanne M Leal[3], Raphael Bernier[4] & Evan E Eichler[1,2]

To assess the relative impact of inherited and *de novo* variants on autism risk, we generated a comprehensive set of exonic single-nucleotide variants (SNVs) and copy number variants (CNVs) from 2,377 families with autism. We find that private, inherited truncating SNVs in conserved genes are enriched in probands (odds ratio = 1.14, $P = 0.0002$) in comparison to unaffected siblings, an effect involving significant maternal transmission bias to sons. We also observe a bias for inherited CNVs, specifically for small (<100 kb), maternally inherited events ($P = 0.01$) that are enriched in CHD8 target genes ($P = 7.4 \times 10^{-3}$). Using a logistic regression model, we show that private truncating SNVs and rare, inherited CNVs are statistically independent risk factors for autism, with odds ratios of 1.11 ($P = 0.0002$) and 1.23 ($P = 0.01$), respectively. This analysis identifies a second class of candidate genes (for example, *RIMS1*, *CUL7* and *LZTR1*) where transmitted mutations may create a sensitized background but are unlikely to be completely penetrant.

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder diagnosed in approximately 1 of 88 children[1] and manifests as deficits in social behavior and language development, as well as restricted or stereotyped interests. ASD is highly heritable, with consensus estimates suggesting that ~50–60% of ASD etiologies are genetic in origin[2,3]. In particular, *de novo* mutations have been implicated as an underlying genetic cause in autism, and these mutations have provided a rich source for understanding pathogenic genes and neurobiological mechanisms in ASD[4–10]. However, *de novo* mutations are rare, and previous work suggests that they could account for the development of ASD in only 25–30% of cases[9], a fraction of the cases likely to have a genetic basis. This suggests that other genetic factors contribute to ASD, including both rare and common inherited genetic variation[2,11].

Previous reports have put forward genetic models for ASD in which rare, inherited CNVs or disruptive SNVs are disproportionally inherited by affected probands when compared to their unaffected siblings[10–15]. Specifically, it has been posited that risk factors for autism must exist that are essentially non-penetrant in females but are preferentially transmitted to affected sons. Although CNVs show some evidence of this transmission pattern[12,16], conclusive evidence from SNVs has been lacking[17]. We sought to test this hypothesis by reanalyzing exome sequence data from a family-based study design, where there are sequence data from a single autism proband, an unaffected sibling and both parents. Our goals were to assess and quantify this SNV transmission disequilibrium, to identify potential candidate genes associated with ASD risk, and to integrate both inherited and *de novo* factors to create a unified ASD risk model for rare disruptive SNV and CNV alterations.

RESULTS

SNV discovery and quality control

To generate a standard call set of inherited variants for analysis, we reprocessed 8,917 exomes sequenced at 3 different genome centers[4,5,7–9]. The set included 2,377 families from the Simons Simplex Collection (SSC)—of which 1,786 consisted of exome sequence data from both parents, an affected child and an unaffected sibling (referred to here as 'quad' families). In combination, we identified a total of 1,303,385 transmitted variants called by both the Genome Analysis Toolkit (GATK) HaplotypeCaller and FreeBayes and passing our quality filters (**Table 1** and Online Methods). Of these variants, 31% were not observed in dbSNP (v137). As a quality control check, we generated a principal-component analysis (PCA) graph for the transmitted variants and compared the results to the self-identified ancestry of the samples (**Supplementary Fig. 1**). As expected, the numbers of rare variant alleles in probands and siblings were highly correlated ($r^2 = 0.99$; **Fig. 1a**), with no significant difference in heterozygosity observed within proband-sibling pairs (**Fig. 1b**). Using the FreeBayes and GATK intersection set, we found a median of 23,055 transmitted variants per exome for probands and siblings (95% confidence interval (CI) = 15,885–27,845 variants; **Fig. 1c**). A median of 377 (95% CI = 154–692) sites per family were new and not observed in dbSNP (v137); conversely, a median of 98.6% of sites were in dbSNP (**Fig. 1e**), and 99.7% of those were in agreement with respect to the alternate allele (**Fig. 1f**). The intersection set of variants had a median transition-to-transversion (Ti/Tv) ratio of 2.94 (95% CI = 2.79–3.03) for all sites, a ratio of 2.95 (95% CI = 2.83–3.04) for dbSNP sites and a ratio of 1.94 (95% CI = 1.05–2.75) for new sites.

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [3]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 14 November 2014; accepted 20 April 2015; published online 11 May 2015; doi:10.1038/ng.3303

Table 1 SNV and CNV discovery

Variants	Quads (n = 1,786)	Trios (n = 591)	All (n = 2,377)[a]
All	1,123,040	614,190	1,737,230
SNVs	1,060,422	581,154	1,641,576
Indels	56,008	31,501	87,509
Private SNVs or indels	52,279	12,634	64,913
CNVs	6,610	1,535	8,145
Deletions	2,289	492	2,781
Duplications	4,321	1,043	5,364
<500 kb	6,369	1,480	7,849
>500 kb	241	55	296

Summary of SNVs, indels and CNVs from exome sequence data for 2,377 families (1,786 quads and 591 trios) from the Simons Simplex Collection (SSC), including transmitted SNV and indel calls from the intersection of GATK HaplotypeCaller and FreeBayes lists and all CNVs with orthogonal validation.
[a]Events in each category are given as the sum of quad and trio numbers, resulting in ~1.6 million SNVs and indels (the number of unique independent sites is ~1.3 million).

In addition, we compared SNPs from exome calls with SNP calls from existing Illumina SNP microarray data[18] (S.J. Sanders, personal communication) and found the median genotype-level concordance to be 99.4% (**Fig. 1d**) (for a median of 17,731 overlapping SNPs in 3,052 offspring in 1,796 families for which microarray data were available).

Although discovery of *de novo* events was not the primary goal of this study, our use of independent SNV callers allowed us to identify additional *de novo* mutations (**Table 2**). Our reanalysis pipeline predicted 1,544 *de novo* SNVs not previously reported (**Supplementary Table 1**). We selected a subset of 141 events for Sanger-based validation because they represented either new recurrences or likely gene-disruptive (LGD) events. Of these new sites, 55% (77) were confirmed to be *de novo* as well as an additional 132 events that had been called but not confirmed in previous studies (**Supplementary Table 2**). *Post-hoc* analysis using three different classifiers (support vector machine (SVM), decision tree and random forest) suggested that the allele balance (defined as the number of reads supporting the alternate allele divided by the total number of reads covering that site) in each proband was the best individual predictor of validation for a *de novo* variant and that classification models could accurately predict which events were most likely to be validated (Online Methods and **Supplementary Fig. 2**). Extrapolating the allele balance in probands across all untested candidate variants in probands (n = 771) suggests that there are 463 (60%) additional true *de novo* variants in probands (at an allele balance cutoff of >0.3); similarly, the predictions generated by the random forest model suggest that 445 (58%) additional *de novo* variants in probands would be validated.

After validation, we identified 21 new recurrently mutated genes (**Table 2**). Notably, these validated mutations established recurrent *de novo* mutations for *GIGYF2* and *SSPO* (encoding a brain-secreted protein involved in axon growth), as well as added a new LGD mutation to *GIGYF1* and to *ASH1L* for a total of three LGD *de novo* mutations each.

SNV transmission disequilibrium

We tested for transmission disequilibrium between probands and siblings using Fisher's exact and Mann-Whitney U tests and by logistic regression (where the dependent variable was the presence of a variant found in a proband or sibling). We considered only transmitted variants reported using both FreeBayes and GATK and defined private events as those unique to a single family. When considering all rare or private protein-altering mutations (LGD and missense) together, we observed no statistically significant difference in the overall burden within proband-sibling pairs. Under the assumption that LGD mutations in genes intolerant to deleterious mutations would be more likely to be pathogenic, we repeated the analysis using residual variation intolerance score (RVIS) values[19,20]. Restricting our analysis to private LGD mutations in genes with the lower 50% of RVIS values, we observed a significant enrichment for transmission in probands when compared to siblings (odds ratio (OR) = 1.14,

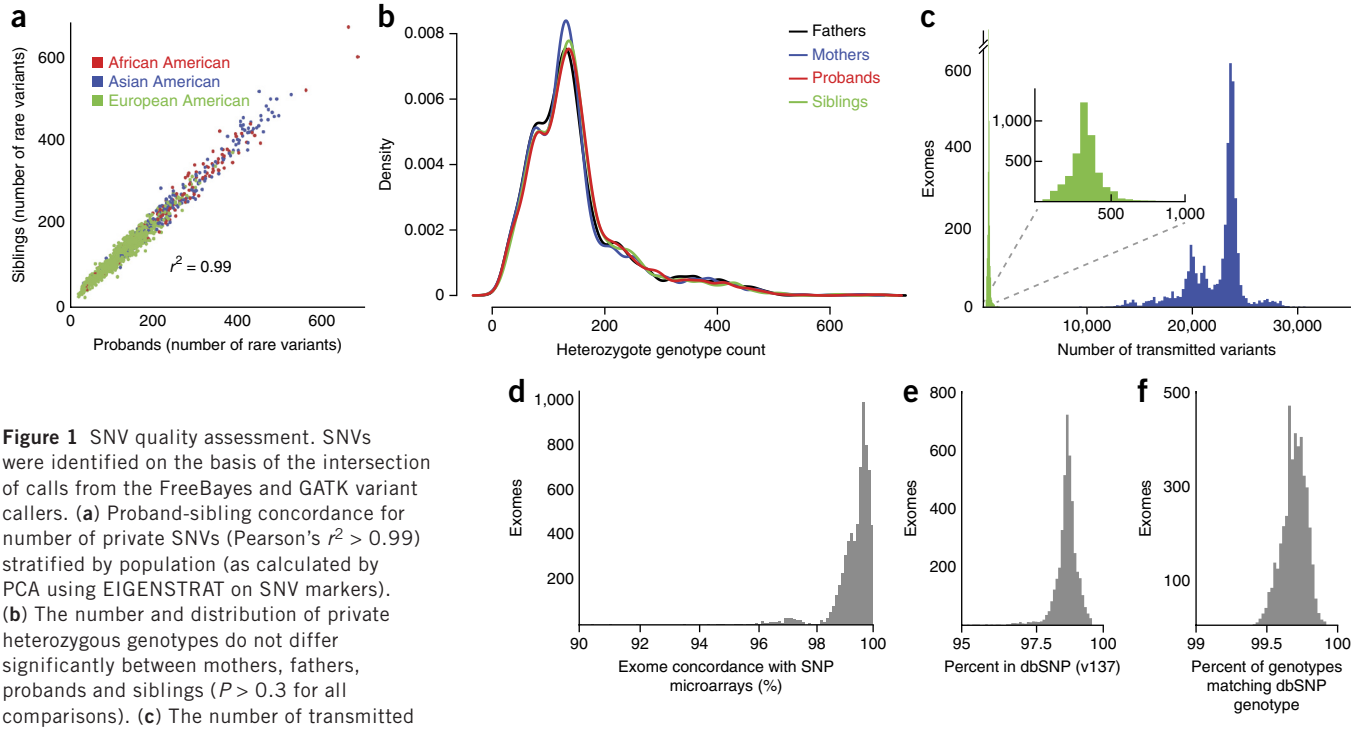


Figure 1 SNV quality assessment. SNVs were identified on the basis of the intersection of calls from the FreeBayes and GATK variant callers. (**a**) Proband-sibling concordance for number of private SNVs (Pearson's $r^2 > 0.99$) stratified by population (as calculated by PCA using EIGENSTRAT on SNV markers). (**b**) The number and distribution of private heterozygous genotypes do not differ significantly between mothers, fathers, probands and siblings ($P > 0.3$ for all comparisons). (**c**) The number of transmitted SNVs per exome for SNVs in dbSNP (blue) and new SNVs (green). (**d**) Concordance between exome and SNP microarray calls[18] (S.J. Sanders, personal communication). (**e**) Fraction of events per exome found in dbSNP. (**f**) Genotype concordance of SNVs found in dbSNP, per exome.

Table 2 Additional genes with recurrent *de novo* mutations

Gene	Krumm proband count	Iossifov proband count	De Rubeis proband count	Mutation type	Number of Sanger-validated mutations	Number of sibling mutations	*P*
ASH1L	1	1	1	2 N, 1 FS	3	0	5.70×10^{-6}
CCDC88C	1	1	0	2 M	2	0	0.07
CDC42BPB	1	1	1	1 N, 2 M	3	0	8.45×10^{-4}
CGNL1	1	1	0	2 M	2	0	0.04
CUL7	1	1	0	2 M	2	0	0.04
DMXL2	1	1	0	2 M	2	0	0.07
FAM92B	1	1	0	2 M	2	0	7.96×10^{-3}
GIGYF2	1	1	1	1 N, 2 M	3	0	3.40×10^{-4}
GRIK5	1	1	1	3 M	3	0	0.01
HECW2	1	1	0	2 M	2	0	0.05
P4HA2	1	1	1	3 M	3	0	1.21×10^{-4}
PHRF1	1	1	1	3 M	3	0	0.20
*PYHIN1**	1	1	1	1 N, 2 M	3	0	5.53×10^{-4}
RAB43	1	1	0	2 M	2	0	1.10×10^{-3}
RBM27	1	1	0	2 M	2	0	4.63×10^{-3}
*SCN4A**	1	1	0	2 M	2	0	0.17
TBC1D31	1	1	0	2 M	2	0	7.29×10^{-3}
TET2	1	1	0	2 M	2	0	0.02
*XIRP1**	1	1	0	2 M	2	0	0.07
ZNF462	1	1	1	1 FS, 2 M	3	0	4.03×10^{-3}
SSPO	2	0	0	1 S, 1 M	2	0	0.91

New validated *de novo* events (Krumm, this study) are compared to previously discovered events (Iossifov *et al.*[9], SSC; De Rubeis *et al.*[10], Autism Sequencing Consortium (non-SSC probands)). The total number of events in probands ($n = 2,377$) is contrasted to the total number of *de novo* events in siblings ($n = 1,786$). All genes except those with an asterisk are brain expressed according to the Gene-Tissue Expression (GTEx) portal. *P* values are based on O'Roak *et al.*[6]; recurrence in genes with marginal or non-significant *P* values is potentially by chance. Mutation type: N, nonsense, FS, frameshift; M, missense; S, splice site.

$P = 0.0002$, Fisher's exact test) and at a family level ($P < 0.0001$, two-tailed paired *t* test; **Fig. 2a**).

This signal persisted for all LGD mutations in genes (regardless of frequency) with RVIS values below the 50th percentile (OR = 1.06; $P = 0.03$, Fisher's exact test; $P = 0.02$, two-tailed paired *t* test). Furthermore, RVIS was a significant predictor of proband or sibling inheritance in a logistic regression model built on all LGD mutations ($P = 0.028$, OR = 1.01 per RVIS percentage point). As suggested by this model, the burden of private LGD mutations in genes with progressively lower RVIS values continued to increase (**Fig. 2b**). At the extreme, the burden between probands and siblings in genes with the lowest 1% of all RVIS values reached an OR of 1.4 (although this comparison was not statistically significant, owing to the small number of mutations present at this threshold in the current data set). When we examined the fractions of probands and siblings who inherited LGD SNVs in highly conserved genes (RVIS values below the 10th percentile), we found that 50.6% of probands (903/1,786 quads) and 47.9% of siblings (855/1,786 quads) harbored such events, a difference of 2.7%. Finally, we performed extensions of the rare variant–transmission disequilibrium test (RV-TDT)[21] at the individual gene level comparing transmission of rare variants to probands and siblings within the SSC families. Several promising candidate genes emerged (**Supplementary Table 3**), although none survived a multiple-testing correction (Online Methods).

We considered the relationship between the set of private LGD mutations in RVIS-restricted genes and phenotypic features of the SSC families (**Fig. 3**). First, we examined how inherited burden correlated with the overall clinical diagnosis. For the 1,575 probands with a diagnosis of 'autism' or 'pervasive developmental disorder' (PDD), the OR values were 1.15 and 1.18 ($P = 0.001$ and 0.05), respectively. In contrast, probands with a diagnosis of 'Asperger's syndrome' ($n = 205$) showed a lower OR of 1.04 ($P > 0.7$; **Fig. 3a**) for inherited gene-disruptive mutations. Consistent with this observation, we found that probands with full-scale IQ between 70 and 100 had an OR of 1.18 ($P = 0.002$), whereas those with an IQ above 100 had a lower, non-significant OR of 1.06 (**Fig. 3b**). For probands in the SSC, IQ and clinical diagnosis were weakly correlated ($r^2 = 0.18$, $P < 1 \times 10^{-10}$; **Supplementary Fig. 3**), but we note that burden of private LGD mutations in RVIS-restricted genes in probands depends on both IQ and clinical diagnosis: for probands diagnosed with autism or PDD and a full-scale IQ above the median for the SSC probands at large (IQ = 84), the OR was 1.1, whereas the burden for probands with Asperger's syndrome of similar IQ was 1.03. Similarly, the OR of this burden for probands with autism and IQ above 100 was 1.19, whereas that for probands with PDD or Asperger's syndrome at this IQ threshold was less than 1 (**Supplementary Table 4**).

Our previous work with CNVs suggested that simplex families could be distinguished into two groups on the basis of their overall social responsiveness scale (SRS) T-scores[22]. Probands and siblings with very different SRS scores ('SRS-discordant' sibling pairs) should show stronger transmission disequilibrium when compared to sibling pairs in which unaffected siblings show elevated ASD symptomatology ('SRS-concordant' sibling pairs; Online Methods). Using our previous threshold definitions[12], we observed a stronger proband-sibling

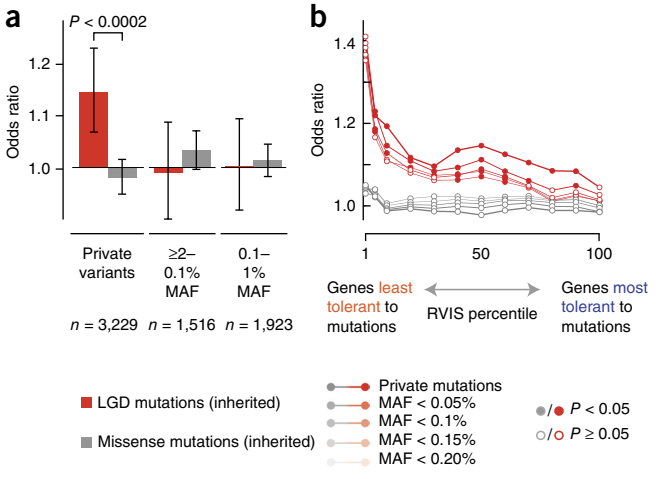


Figure 2 Transmission disequilibrium of SNVs in ASD. (**a**) Private, inherited LGD SNVs (red bars) in genes not tolerant to functional variation were significantly enriched in probands. The analysis examined only SNVs in genes with an RVIS value in the bottom 50%. Non-private, rare variants and inherited missense SNVs (gray bars) are not enriched in probands. Bar heights are Fisher's exact test OR values, and whiskers represent 95% confidence interval (CI) bounds. (**b**) RVIS is a critical determinant for enrichment in probands. Burden was highest (reaching OR = 1.4) for private, inherited LGD SNVs among the genes with the lowest RVIS values (within the first percentile). MAF, minor allele frequency.

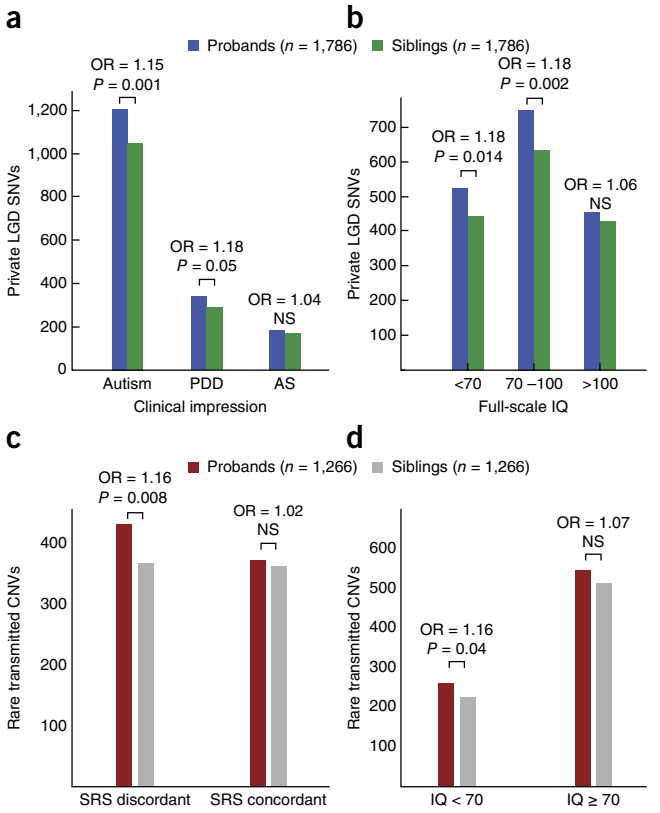


Figure 3 Transmitted mutations and their effect on phenotype. (**a**) Private, inherited LGD SNVs are enriched in probands with autism or PDD diagnosis but not Asperger's syndrome (AS) diagnosis. (**b**) Private, inherited LGD SNVs are primarily enriched in probands with lower IQ than average (<100). (**c**) We observe transmission disequilibrium of rare, inherited CNVs in SRS-discordant families (proband SRS score > 75, sibling SRS score < 50) but not in families where the SRS score was mild or more balanced between the proband and sibling. (**d**) Rare, inherited CNVs are enriched in probands (versus their siblings) with IQ <70, but the effect is not significant in probands with IQ >70. All tests and reported *P* values are paired *t* tests based on proband-sibling pairs. All analyses were restricted to genes with RVIS values below the 50th percentile. NS, not significant.

differential of 3.7% for private LGD SNVs in conserved genes (RVIS below the 10th percentile) for SRS-discordant quads only (probands, 484/923; siblings, 450/923), whereas SRS-concordant quads had only a 1.6% differential (probands, 419/863; siblings, 405/863).

CNV discovery and validation

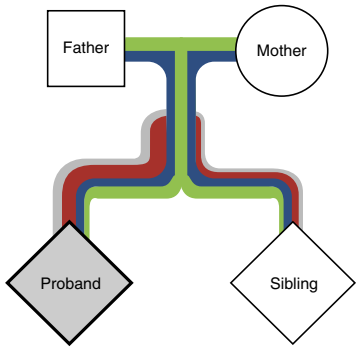
Because exome and SNP microarray data provide the opportunity to accurately detect a subset of smaller CNVs within the exonic regions of genes[12], we also revisited the burden of both inherited and *de novo* CNVs with respect to autism. We characterized CNVs from 1,266 quads with available SNP microarray data (validation shown in **Supplementary Fig. 4**) and tested an additional 50 samples with CNVs of interest identified by array comparative genomic hybridization (aCGH). We focused in particular on validating smaller CNV events that affected genes recurrently affected by *de novo* SNVs, such as *DSCAM*, *CHD2*, *ARID1B* and *TNRC6B* (**Supplementary Table 5**). We identified a total of 2,891 CNVs with an excess of autosomal events in probands when compared to siblings (854 versus 743; OR = 1.25, $P = 0.006$, binomial two-sided test). The overall ratio of duplications to deletions was 1.6, consistent with previous results for a smaller SSC data set[12]. Restricting the analysis to *de novo* CNVs, we identified, as expected, a more significant 2.4-fold excess ($P = 6.7 \times 10^{-5}$, paired *t* test) in probands ($n = 79$) when compared to siblings ($n = 33$), driven primarily by deletions ($P = 4.2 \times 10^{-5}$, paired *t* test) and not duplications ($P = 0.18$, paired *t* test) (**Table 3**). Overall, *de novo* CNVs were larger in probands than in siblings ($P = 0.03$, Wilcoxon) and carried genes with significantly lower total RVIS values ($P = 0.02$, Wilcoxon). Both FMRP and CHD8 target genes were enriched in *de novo* CNVs (OR = 3.1 and 2.7; $P = 6.6 \times 10^{-4}$ and 1.7×10^{-3}, Fisher's exact test and $P = 1.4 \times 10^{-4}$ and 2.6×10^{-4}, paired *t* test, respectively), and this is likely owing, in part, to the larger size of the *de novo* events among probands.

The validated inherited CNV data set (frequency < 0.8%) consisted of a total of 1,485 events ($n = 775$ in probands and $n = 710$ in siblings) from 1,266 quads. We replicated the previously reported[12] preferential transmission of CNVs to probands when compared to siblings

Table 3 CNV burden and transmission

Data set	Inheritance	OR[a]	*P* (*t* test)	*t*-test mean of the differences[a]	Number of CNV events	
					Probands	Siblings
All	*De novo*	1.90 (1.32, Inf)	6.7×10^{-5}	0.46 (0.28, Inf)	79	33
	All inheritance	1.10 (0.95, Inf)	0.03	0.08 (0.02, Inf)	775	710
	Maternal	1.15 (1.00, Inf)	0.01	0.11 (0.04, Inf)	411	357
	Paternal	1.02 (0.89, Inf)	0.59	0.02 (–0.05, Inf)	364	353
Deletions	*De novo*	3.07 (1.79, Inf)	4.2×10^{-5}	0.68 (0.43, Inf)	49	13
	All inheritance	1.11 (0.96, Inf)	0.05	0.09 (0.01, Inf)	297	262
	Maternal	1.08 (0.90, Inf)	0.20	0.08 (–0.02, Inf)	156	139
	Paternal	1.09 (0.90, Inf)	0.14	0.09 (–0.01, Inf)	141	123
Duplications	*De novo*	1.20 (0.74, Inf)	0.18	0.21 (–0.05, Inf)	30	20
	All inheritance	1.12 (0.98, Inf)	0.20	0.05 (–0.02, Inf)	478	448
	Maternal	1.16 (0.99, Inf)	0.03	0.11 (0.03, Inf)	255	218
	Paternal	1.00 (0.86, Inf)	0.67	–0.02 (–0.11, Inf)	223	230
Duplications <100 kb	*De novo*	0.60 (0.27, Inf)	0.55	–0.15 (–0.57, Inf)	9	12
	All inheritance	1.12 (0.97, Inf)	0.38	0.04 (–0.04, Inf)	315	298
	Maternal	1.19 (0.99, Inf)	0.01	0.15 (0.05, Inf)	177	143
	Paternal	0.98 (0.81, Inf)	0.22	–0.08 (–0.19, Inf)	138	155

Test results (one-sided paired *t* test) are shown for all CNV events, deletions, duplications and small (<100 kb) duplications.
[a]Numbers in parentheses indicate 95% confidence intervals.



$P(\text{ASD}) \sim \beta_0 + \beta_1$ (*de novo* SNVs) + β_2 (*de novo* CNVs) + β_3 (inherited CNVs) + β_4 (inherited SNVs) + strata (family)

	Odds ratio	*P* value
De novo SNVs (disruptive)	1.72	$<1 \times 10^{-5}$
De novo CNVs	2.05	0.0004
Inherited CNVs (rare)	1.23	0.01
Inherited SNVs (rare)	1.11	0.0002

Figure 4 Combined risk model for SNVs and CNVs (inherited and *de novo*). Integrative risk model for ASD, based on *de novo* and inherited events and covering both SNVs and CNVs. The model used is a stratified logistic regression model, which uses proband-sibling pairs to estimate the OR (i.e., risk) of ASD for each type of event (**Supplementary Table 7**).

($P = 0.03$, paired *t* test). This effect was driven almost exclusively by smaller (<100 kb) maternally inherited events ($P = 0.01$, paired *t* test). In contrast to *de novo* events, for inherited CNVs, there was no difference in the size of the CNVs transmitted in probands relative to those transmitted in siblings ($P = 0.59$, Wilcoxon). Similar to our observations of SNV mutations in conserved genes, we found that genes within CNV intervals in probands had a lower average RVIS than those in siblings, with the difference borderline statistically significant ($P = 0.05$, Wilcoxon).

To more fully understand the potential biology of these inherited CNV events, we tested whether the CNVs were enriched in either FMRP[23] or CHD8 (ref. 24) targets. Although no overall enrichment of FMRP ($P = 0.22$) or CHD8 ($P = 0.19$) target genes was observed among inherited CNVs, when we restricted the analysis to maternally inherited duplications, we observed a significant enrichment for CHD8 targets (OR = 1.5; $P = 0.02$, Fisher's exact test and $P = 3.9 \times 10^{-3}$, paired *t* test). In particular, this enrichment was strongest for small duplications (<100 kb) (OR = 1.5; $P = 0.05$, Fisher's exact test and $P = 7.4 \times 10^{-3}$, paired *t* test). As truncating mutations of *CHD8* have been associated with a subtype of autism characterized by macrocephaly[25], we tested whether patients carrying CNVs that intersected CHD8 target genes showed any deviation in head circumference. We specifically stratified the patient population into two groups: those containing a maternally inherited CNV with a CHD8 target and those that had a maternally inherited CNV without a CHD8 target. We then tested whether there was an enrichment of macrocephalic or microcephalic patients in carriers of CNVs with CHD8 targets. Interestingly, we observed a modest enrichment for macrocephaly in patients with maternally inherited autosomal deletions containing CHD8 targets (OR = 2.9, $P = 0.03$, Fisher's exact test), including for smaller deletions (<100 kb) (OR = 3.5, $P = 0.04$, Fisher's exact test). The reciprocal was also observed, with enrichment for microcephaly with borderline significance in maternally inherited autosomal duplications containing CHD8 targets (OR = infinity, $P = 0.04$, Fisher's exact test) (**Supplementary Fig. 5**). As a control, we repeated the same analysis for inherited CNVs carrying FMRP targets, which we did not expect to have any relevance for head circumference, and we found no statistically significant enrichment for increased head size among carriers of CNVs with these targets.

SNV and CNV integration and sex bias

We jointly examined SNVs and CNVs at a gene level to identify potentially new ASD candidate genes (**Supplementary Table 6**). On the basis of our findings, we considered all *de novo* CNVs, private, inherited SNVs in genes with an RVIS below the 50th percentile and rare, inherited CNVs where at least one gene had an RVIS below the 50th percentile; we then created a combined gene-level table identifying several candidate genes. In particular, the three highest-ranked genes—*RIMS1*, *CUL7* and *CSMD1*—each display brain-specific expression or have identified neural functions (**Supplementary Fig. 6**). The highest-ranked gene, *RIMS1*, had two *de novo* LGD mutations and two private, inherited LGD mutations in probands. Additionally, there were six rare, inherited LGD mutations in probands (two were shared with siblings and one was found in a trio) and one mutation in a sibling alone. *CUL7* had two *de novo* and two inherited LGD mutations in probands (none in siblings). Finally, *CSMD1* had three *de novo* missense mutations in probands, four LGD SNVs (one was shared with a sibling and one was found in a trio) and five rare, inherited CNVs (one was shared with a sibling and one was found in a trio). (See **Supplementary Fig. 6** for details and the locations of these variants.)

We quantified the risk for ASD by examining *de novo* and inherited CNVs and SNVs using a conditional logistic regression model (**Fig. 4**; see Online Methods). In this model, the binary outcome of an ASD proband or unaffected sibling is predicted by four independent counts: (i) the number of *de novo* CNVs; (ii) the number of *de novo* LGD SNVs; (iii) the set of rare, inherited CNVs; and (iv) the set of private, inherited LGD SNVs in genes with an RVIS below the 50th percentile (**Supplementary Table 6**). Additionally, we accounted for familial stratification effects by adding a family-level stratum to the model. Using data from the 1,786 quads, we found robust effects for *de novo* events (**Supplementary Table 7**): each *de novo* CNV increased the risk for ASD by 2.05-fold, whereas each *de novo* SNV increased risk by 1.72-fold ($P = 0.0004$ and $P < 1 \times 10^{-7}$, respectively). In addition, the results from this analysis show a significant role for inherited mutations in ASD risk: rare, inherited CNVs contribute an increased risk of 1.23 ($P = 0.01$), and private LGD SNVs have an OR of 1.11 ($P = 0.0002$). These results suggest that each of the four types of mutations modeled additively contribute to the risk of ASD and that they do so in a statistically independent manner.

Finally, by calculating the attributable fraction in the population (**Supplementary Fig. 7**), we were able to identify the contribution of each variant type as follows: *de novo* LGD SNVs, 6.62% (4.18%, 8.99%); private, inherited LGD SNVs, 8.54% (−24.23%, 32.66%); *de novo* CNVs, 2.92% (1.37%, 4.44%); rare, inherited CNVs, 3.18% (−3.71%, 9.6%). Whereas these values give high confidence for *de novo* events, the contributions of the inherited events are less clear. When stratifying by inheritance and RVIS for private, inherited LGD SNV events, the results became much tighter and showed a clear contribution for maternal events (7.15% (−0.25, 14.01%)) and not for paternal events (1.01% (−6.56%, 8.04%)). The same was found for maternal duplications (2.99% (−0.45%, 6.31%)), especially those under 100 kb in size (2.65% (−0.16%, 5.38%)).

Specifically, we extended the work to investigate the roles of maternally transmitted events to males and females. First, we assessed the attributable fractions in all quad families and subsequently in quads with male probands and female probands separately. For LGD SNVs, we were able to identify private, maternally inherited LGD SNVs in genes with RVIS values below the 50th percentile as the category

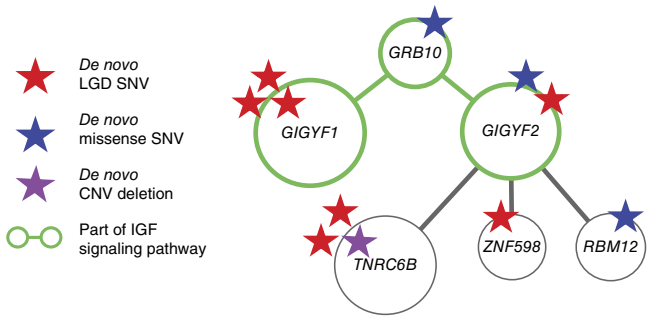


Figure 5 Networks and pathways. A highly interconnected network was identified on the basis of new *de novo* mutations identified in this study (note: one additional *de novo* missense mutation was recently identified in an independent study[10]). Gene Ontology (GO) annotation of the genes in this network suggests involvement of the insulin-like growth factor (IGF) signaling pathway (*GIGYF1*, *GIGYF2* and *GRB10*; accession GO:0048009), which has been previously implicated in the development of ASD[28]. Furthermore, *GIGYF2* and *ZNF598* form part of the m4EHP mRNA-binding complex and have widespread roles in translational repression, especially in the brain and lungs[38]. Red stars, *de novo* LGD mutations (frameshift, stop gain, splice site); blue stars, *de novo* missense mutations; purple star, CNV deletion.

with the highest attributable fraction (estimated) in the population (8.32% (0.56%, 15.48%)) in the families with male probands, whereas the families with female probands had a value of −2.33% (−29.06%, 18.87%) in this same category. No effect was observed for paternally inherited LGD events. This is in stark contrast to *de novo* LGD events, which contribute 5.7% (−2.26%, 13.04%) of the attributable fraction in females (**Supplementary Fig. 7**). Although larger sample sizes will be required, these findings are consistent with the maternal bias observed for large and small CNVs and now extend the observation to maternally inherited SNVs. To further explore this difference, we examined all four possible quad types based on sex: male proband/male sibling, male proband/female sibling, female proband/male sibling and female proband/female sibling. These observations for maternally inherited LGD SNVs held true regardless of the sex of the sibling (**Supplementary Fig. 7** and **Supplementary Table 8**)[26].

DISCUSSION

In this study, we have explored the effect of rare, inherited variation on the risk of autism. Our results provide some of the first genetic evidence that private, inherited SNVs that truncate proteins are enriched in autism probands. Remarkably, this effect is only observed for truncating SNVs that disrupt genes intolerant to functional variation and shows bias in transmission from mothers to their sons. The effect becomes more pronounced the more intolerant the gene is to mutation, consistent with the notion that such genes are subject to strong selective pressure. While the effect is strongest for individuals with a diagnosis of autism, it is most significant for SRS-discordant quads and probands with an IQ between 70 and 100. Extending previous work[12–14] on the role of rare inherited CNVs, we report that smaller maternally inherited duplications show the largest bias toward transmission to probands, and these duplications are enriched for gene targets of CHD8. The reciprocal shift in macrocephaly and microcephaly when comparing CHD8 target gene duplications and deletions, respectively, is intriguing but warrants further investigation. In addition, the application of two SNV callers identified 77 additional *de novo* SNVs that were previously missed[9]. The recurrent hits highlight potential new pathways such as the insulin-like growth factor protein-interaction network (**Fig. 5**). This is interesting because variable levels of IGF1 are considered a biomarker of autism[27] and are of potential therapeutic relevance[28].

In some cases, inherited and *de novo* mutations of both SNVs and CNVs converge on the same gene (**Supplementary Table 6**). *RIMS1* has been previously suggested as an ASD candidate as a result of recurrent *de novo* truncating mutations[4,29]. In this analysis, we also find a nominally significant transmission disequilibrium of private, disruptive events of *RIMS1* to probands ($P = 0.013$, TDT-Combined Multivariate and Collapsing (CMC) analytical[21]) but not siblings ($P = 0.841$) (**Supplementary Table 3**). The gene displays brain-specific expression, and disruption of the gene in mice leads to increased postsynaptic density and impaired learning. *CUL7* has two *de novo* and two LGD-inherited mutations in probands (none in siblings); functionally, it is an E3 ligase with high cerebellar brain expression and a selective role in neural dendrite patterning and growth[30]. For the highly conserved gene *CSMD1*, there are three *de novo* missense mutations, one shared inherited LGD SNV, and four rare inherited focal CNVs (one shared with siblings). Overall, there are eight events in ASD probands and two in siblings clustered at the exon-dense 3′ end of *CSMD1*, a region nearly devoid of exonic CNVs in the Database of Genomic Variants (DGV; **Supplementary Fig. 8**). Functionally, *CSMD1* exhibits strong and specific brain expression; this gene has been associated with schizophrenia[31], and damaging variants of the gene segregated in two ASD families with distantly related probands[32].

We also identified candidate genes for which no *de novo* events have yet been reported despite evidence of over-transmission of private LGD events to affected probands within the SSC families (**Supplementary Tables 3** and **6**). Using the RV-TDT on rare inherited events, we identified candidates that have not yet reached locus-specific significance, including *LZTR1* (commonly deleted in DiGeorge syndrome[33]) and *CENPJ* (a gene with autosomal recessive mutations known to cause microcephaly and intellectual disability[34]). While these genes and genes like *RIMS1* may represent important risk factors for ASD, the fact that gene-disruptive events are inherited from normal parents and/or occasionally transmitted to unaffected siblings argues that they are neither necessary nor sufficient to cause autism. This stands in contrast to other genes, such as *ADNP*, *CHD8* or *DYRK1A*, where *de novo* LGD mutations have been observed almost exclusively in probands. In fact, genes enriched for *de novo* LGD mutations have significantly fewer inherited LGD mutations than expected from randomly sampled gene sets (empirical $P < 1 \times 10^{-4}$, see Online Methods and **Supplementary Fig. 9**), suggesting that inherited and *de novo* mutation risk factors may often target different genes.

We hypothesize the second class of inherited-LGD genes simply predisposes an individual to ASD, requiring additional genetic or non-genetic factors to reach a disease state. Notably, the largest effect appears to be for maternal transmission to sons, consistent with other recent findings[16] and models of autism[11]. Such oligogenic models have been proposed previously for CNVs[35] as well as other forms of severe mutation associated other human diseases[36,37]. The availability of CNVs and SNVs from exome sequence data is the first step toward obtaining a more complete genetic picture at an individual level in the context of autism. In this light, it is interesting that our analysis uncovered a paternally inherited two-exon intragenic deletion of *NRXN3* and a *de novo* missense mutation of *NLGN2* in proband 13367.p1 (**Supplementary Fig. 10**). Both of these genes have been identified as ASD risk factors, but crucially, they are also protein-protein interacting partners. The neuroligin-neurexin interaction has long been hypothesized to be a key underlying pathway in ASD

pathology, but to our knowledge, this is the first identification of a case with mutations in both binding partners. As the genetic profile of the SSC becomes more complete through full genome sequencing, it is likely that examples supporting an oligogenic model for ASD will become more prevalent and informative to our understanding of the genetic etiology.

URLs. epiR, http://cran.r-project.org/web/packages/epiR/index.html.

METHODS

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

Accession codes. Raw data (including BAM files, VCF files, additional data files and the software pipelines) have been deposited in the National Database for Autism Research (NDAR) under study 353 (doi:10.15154/1151812).

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

ACKNOWLEDGMENTS

We thank D. Obenshain, D. Hall, B. Koser and S. Novikova for providing support for usage of the Amazon Cloud and for assistance in the deposition of SNV and CNV call sets into the National Database for Autism Research (NDAR). We are grateful to the laboratories of M. Wigler and M. State for providing early access to exome sequencing data as well as access to SNP microarray data. We also thank T. Brown for assistance in editing this manuscript. Funding for this study was provided, in part, by the US National Institutes of Health (1U01MH100233 to E.E.E.), by the National Institute for Mental Health (R01MH101221 to E.E.E. and R01MH100047 to R.B.) and by the Simons Foundation (SFARI 89368 to R.B. and SFARI 137578 to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren and E. Wijsman). We appreciate obtaining access to phenotypic data on Simons Foundation Autism Research Initiative (SFARI) Base. Approved researchers can obtain the SSC population data set described in this study by applying at https://base.sfari.org/.

AUTHOR CONTRIBUTIONS

N.K., T.N.T. and E.E.E. designed experiments and wrote and edited the manuscript. N.K. performed sequence data reanalysis and created and analyzed the SNV call set. T.N.T. created and analyzed the CNV call set, analyzed SNP microarray data, performed statistical analyses for SNV and CNV quality control, and examined epidemiological features for the full data set. C.B., L.V., K.M., K.W. and H.A.S. performed validation experiments and sample handling. A.R. and B.P.C. provided additional computational support. Z.-X.H. and S.M.L. performed the TDT tests and statistical analyses. R.B. provided phenotype data and additional SSC variables where needed.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators and Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders—Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR Surveill. Summ.* **61**, 1–19 (2012).
2. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
3. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095–1102 (2011).
4. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
5. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
6. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
7. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
8. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
9. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
10. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
11. Zhao, X. *et al.* A unified genetic theory for sporadic and inherited autism. *Proc. Natl. Acad. Sci. USA* **104**, 12831–12836 (2007).
12. Krumm, N. *et al.* Transmission disequilibrium of small CNVs in simplex autism. *Am. J. Hum. Genet.* **93**, 595–606 (2013).
13. Poultney, C.S. *et al.* Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.* **93**, 607–619 (2013).
14. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
15. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
16. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
17. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.* **15**, 133–141 (2014).
18. 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).
19. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
20. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
21. He, Z. *et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.* **94**, 33–46 (2014).
22. Constantino, J.N. *et al.* Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *J. Autism Dev. Disord.* **33**, 427–433 (2003).
23. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
24. Subtil-Rodríguez, A. *et al.* The chromatin remodeller CHD8 is required for E2F-dependent transcription activation of S-phase genes. *Nucleic Acids Res.* **42**, 2185–2196 (2014).
25. Bernier, R. *et al.* Disruptive *CHD8* mutations define a subtype of autism early in development. *Cell* **158**, 263–276 (2014).
26. Taylor, J.W. Simple estimation of population attributable risk from case-control studies. *Am. J. Epidemiol.* **106**, 260 (1977).
27. Steinman, G. & Mankuta, D. Insulin-like growth factor and the etiology of autism. *Med. Hypotheses* **80**, 475–480 (2013).
28. Bozdagi, O., Tavassoli, T. & Buxbaum, J.D. Insulin-like growth factor-1 rescues synaptic and motor deficits in a mouse model of autism and developmental delay. *Mol. Autism* **4**, 9 (2013).
29. Dong, S. *et al. De novo* insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9**, 16–23 (2014).
30. Litterman, N. *et al.* An OBSL1-Cul7Fbxw8 ubiquitin ligase signaling mechanism regulates Golgi morphology and dendrite patterning. *PLoS Biol.* **9**, e1001060 (2011).
31. Håvik, B. *et al.* The complement control–related genes *CSMD1* and *CSMD2* associate to schizophrenia. *Biol. Psychiatry* **70**, 35–42 (2011).
32. Cukier, H.N. *et al.* Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol. Autism* **5**, 1 (2014).
33. Kurahashi, H. *et al.* Isolation and characterization of a novel gene deleted in DiGeorge syndrome. *Hum. Mol. Genet.* **4**, 541–549 (1995).
34. Kaindl, A.M. *et al.* Many roads lead to primary autosomal recessive microcephaly. *Prog. Neurobiol.* **90**, 363–383 (2010).
35. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
36. Béna, F. *et al.* Molecular and clinical characterization of 25 individuals with exonic deletions of *NRXN1* and comprehensive review of the literature. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **162B**, 388–403 (2013).
37. Lupski, J.R. Digenic inheritance and Mendelian disease. *Nat. Genet.* **44**, 1291–1292 (2012).
38. Morita, M. *et al.* A novel 4EHP-GIGYF2 translational repressor complex is essential for mammalian development. *Mol. Cell. Biol.* **32**, 3585–3593 (2012).

ONLINE METHODS

Data sets. We analyzed exome data from 2,377 families with ASD (2,391 before quality control) from the SSC[39], including 1,786 quads and 591 trios (total of 8,917 exomes). These exomes were recently analyzed for *de novo* variants[4,5,8,9] but were reanalyzed here to increase sensitivity and to create a unified call set for private variants (**Table 1**). The raw sequence data for these exomes are available in the National Database for Autism Research (NDAR), and the reanalyzed data, including the complete variant call format (VCF) files from the SNV call set and bioinformatics pipelines for this study, are available (see URLs). We used Illumina 1M, 1MDuo or Omni2.5 SNP microarray data for 1,266 complete quads for CNV validation[18] (S.J. Sanders, personal communication). Relevant phenotype scores were extracted for both SRS (parent-assessed T-scores[22]) and full-scale IQ (as in Krumm *et al.*[12]) from the SSC Simons Foundation Autism Research Initiative (SFARI) Base. Normalized head circumference scores were determined as previously described[40]. Published databases of FMRP[23] and CHD8 (ref. 24) target genes were used to assess enrichment of targets within CNVs. The institutional review board (IRB) of the University of Washington approved this study (IRB 46179).

Sequence data processing. Reads from all 8,917 exomes were realigned using BWA-MEM[41] (v0.7.5a, options -k 17) to the 1000 Genomes Project Phase 1 reference genome (hg19/GRCh37). We mapped all available libraries for samples, including single-end and paired-end reads where appropriate. Mapped BAM files were processed according to GATK[42] best practices, including duplicate marking and mate fixing. We applied GATK (v. 2.7-4) IndelRealigner in a family-aware manner, ensuring that each member of a family was realigned at the same positions across the family. Base qualities were recalibrated using GATK. We next used QPLOT[43] and computed 24 read- and exome-level statistics (**Supplementary Table 9**) for quality control assessment. Finally, to ensure that we did not have any sample, family or data mix-ups, we used a custom-developed tool (S.J. Sanders, personal communication) to identify and match 287 polymorphic SNPs in each exome to an existing database of 'SNP fingerprints' derived from Illumina SNP microarray data[18] and 96 SNP fingerprints collected by the Rutgers sample distribution center. We excluded 14 families for issues with sample identity, and concordance by center is shown in **Supplementary Table 10**.

SNV discovery. To identify SNVs, we batched families into groups of 16–20 families, or approximately 70 exomes, to ensure better sensitivity for events. We called SNVs and indels with both the GATK HaplotypeCaller (v 2.7-4) and FreeBayes[44] (v0.99) to within 20 bp of the NimbleGen EZ-SeqCap v2.0 targets. Family-level VCF files from FreeBayes and GATK were merged into a union set. Merged VCF files were annotated using SnpEff[45] (v 3.4i), dbNSFP[46] (v2.1), CADD score[47], dbSNP (v137), tandem repeats and segmental duplications. Allele frequency was estimated by counting non-reference alleles across all parents (n = 4,754).

For *de novo* events, we applied a minimum read depth of 6 alternate alleles in offspring and a read depth of >10 reference reads in parents and allowed for no more than 2 low-quality bases of the *de novo* variant. Because the FreeBayes and GATK SNV calling routines report only the number of high-quality reads supporting the alternate or reference allele, we queried the original BAM files at each site to include the count of low-quality bases in these filters. To exclude common artifacts, we only considered *de novo* sites that were private to a family. Inherited events were derived from the intersection set of both algorithms, with a minimum depth filter (DP > 20) and quality filter (QUAL > 50) for all events (**Fig. 1** and **Supplementary Fig. 11**). In addition, we applied a batch exclusion filter, which filtered out variants found at high frequency exclusively in one batch (3 or more times among 16–20 families). Using the FreeBayes and GATK intersection set, we found a median of 23,055 transmitted variants per exome for probands and siblings (95% CI = 15,885–27,845; **Fig. 1c**) and a median of 26,920 transmitted variants per family (95% CI = 23,394–31,401). A median of 377 (95% CI = 154–692) sites per family were new and not observed in dbSNP (v137); conversely, a median of 98.6% of sites were in dbSNP, and 99.7% of these were in agreement with respect to the alternate allele. Overall, 81% of all transmitted variants were found by both FreeBayes and GATK, 12% were found by FreeBayes alone and 7% were found by GATK alone. The intersection set of variants had a median Ti/Tv ratio of 2.94 (95% CI = 2.79–3.03) for all sites, a ratio of 2.95 (95% CI = 2.83–3.04) for dbSNP sites and a ratio of 1.94 (95% CI = 1.05–2.75) for new sites. Of all the inherited mutations in the intersection set, an average of 341 (95% CI = 133–632) sites were new and not observed in dbSNP (v137); 98.6% of sites were in dbSNP with a concordance rate of 99.7% (for all transmitted variants, 93.4% of variants were found in dbSNP and 99.5% were concordant). In addition, we compared SNPs from exome calls with SNP calls from existing Illumina SNP microarray data[18] (S.J. Sanders, personal communication) and found the median genotype-level concordance to be 99.4% (for a median of 17,731 overlapping SNPs in 3,052 offspring in 1,796 families for which microarray data were available).

Modeling *de novo* SNV validation efficiency. We used the Sanger sequencing validation results from our 141 tested *de novo* SNV events to better understand which SNV calls would be the most likely to validate. In this *post-hoc* analysis, we constructed a feature matrix of 77 validated events (truly *de novo*) and 63 'invalidated' events (which turned out to be inherited or otherwise not present), along with event- or site-level quality data emitted by GATK and/or FreeBayes (**Supplementary Table 1**). This quality information included data such as QUAL (Phred-scale quality score for the assertion made for the alternate allele), BaseQRankSum, MQ (mapping quality), MQ0 (number of reads with mapping quality equal to 0 covering the variant) and MQRankSum (z score from Wilcoxon rank-sum testing of alternative versus reference read mapping qualities), as well as sample-specific data for allele depth, allele quality and GATK-specific fields such as PL (Phred likelihood). As many of the invalidated events were found to be present in one of the parents (i.e., they were inherited heterozygous SNVs), we also included the maximum (or minimum, when appropriate) values of both parents for PL and GQ (genotype quality). For values that were not outputted by both FreeBayes and GATK, we imputed values on the basis of the mean of all values not missing.

Using this feature matrix, we investigated three types of classifiers present in the Python Scikit-Learn package[48]: an SVM (linear kernel; 'svm.SVC' module), a decision tree ('tree.DecisionTreeClassifier') and a random forest model ('ensemble.RandomForestClassifier', with 'n_estimators' = 200). We estimated all accuracy statistics by cross-validation (scores are the reported average from cross-validation; **Supplementary Table 11**). Using these data, the random forest method had the best overall performance across most performance metrics, although the SVM method had slightly better recall. For the decision tree and random forest methods, we were able to compute matrices corresponding to individual feature importance with respect to classification (**Supplementary Table 12**; note that this is not possible using the SVM implementation). For both classifiers, we found that the most important feature was the proband's allele balance, and this finding was recapitulated when observing the allele balance values directly (**Supplementary Fig. 2**).

CNV discovery. We used the CoNIFER[49] and XHMM[50] algorithms to discover copy number variation from exome data at a single-exon resolution. Identification with CoNIFER was carried out as described[12]. Briefly, we split reads into 36-mers and aligned them using mrsFAST[51] to NimbleGen EZ-SeqCap v.2 targets and flanking sequence. Using CoNIFER, we processed all samples with the specific setting of –components-removed equal to 40. CNV calls were made using the CoNIFER tools package, which implements DNAcopy[52]. In parallel, XHMM was applied using best-practice guidelines. GATK was used to calculate depth of coverage (from BWA-MEM alignments) for each individual, and all individuals were then combined into one composite file. The XHMM-specific steps included hard filtering of samples and targets, PCA on the data, filtering on the basis of the PCA results and discovery of CNVs. Post-discovery CNVs were genotyped by family, and a score cutoff of 10 was ultimately used to determine inheritance in families on the basis of SQ and NQ values[50].

Using the union of the XHMM and CoNIFER call sets, we first genotyped all loci across family members to recover false negative calls and then identified transmitted and *de novo* CNVs. CNVs were clustered into copy number–variable regions, or CNVRs (as previously described[12]), and then annotated with family frequency across the entire cohort. To focus our analysis on the CNVs most likely to be relevant to ASD pathogenesis, we restricted our analysis to rare CNVs found at frequency <0.8% (<10 events/1,266 families) mapping outside of repetitive genomic elements (**Supplementary Fig. 12**).

Validation experiments. Our reanalysis pipeline identified 1,544 new candidate *de novo* SNVs not detected by previous analyses of the same data set (**Supplementary Table 1**). Using Sanger sequencing, we attempted validation of 141 (of the 1,544) previously unidentified *de novo* variants, including all new LGD (stop-gain, frameshift and splice-site) events, as well as recurrent missense mutations in autism candidate genes[53]. We were able to validate 77 new sites (55%). These SNVs included various functional classes: 1 codon change plus codon deletion, 11 frameshift mutations, 51 missense mutations, 3 splice-site acceptor mutations, 4 splice-site donor mutations, 6 stop-gain mutations and 1 stop-loss mutation. In addition, we also validated by Sanger sequencing another 132 previously called[9] but not confirmed events, resulting in a total of 209 validated *de novo* SNVs and indels (**Supplementary Table 2**). This new analysis identified 21 new recurrently hit genes not identified in previous studies (**Table 2**). We did not attempt validation of any inherited SNVs but rather used the intersection of the FreeBayes and GATK sets to obtain the highest-quality variant events (all rare SNVs are shown in the **Supplementary Data Set**).

We used SNP microarray data (available for 1,266 quads) for validation of CNV events discovered using CoNIFER and XHMM. Probe-level copy number estimates were generated for each array sample using Corrected Robust Linear Model with Maximum-Likelihood Classification (CRLMM) software[54,55]. A permutation-based method examined the mean copy number of all probes in each CNVR versus random sampling of the same number of probes from the genome ($n = 10{,}000$) to assess event confidence. Events with permutation $P < 0.01$ and with percentile ranking of <30 or >70 were considered to be validated deletions and duplications, respectively (a full list of validated events for all quads and trios is provided in **Supplementary Table 13**). To further validate and genotype *de novo* events, we employed the CRLMM method to recall genotypes in trios and were able to recover events that were truly *de novo* as well as events inherited from a parent but missed in the exome analysis. Our final CNV data set for all statistical testing consisted of validated events in the 1,266 quads for which SNP microarray data were available. We further tested an additional 50 *de novo* CNVs in individuals lacking SNP microarray data by aCGH[12] using a customized Agilent Technologies microarray. In this design, we targeted events and flanking genomic regions (up to 5 kb or three exons) where probe density ranged from a spacing of 150 bp to 5 kb, depending on the size of the event. Of these CNV events, 26 were validated, of which 21 were confirmed to be *de novo* whereas 5 were transmitted events (**Supplementary Table 5**).

Statistical analyses. We tested for transmission disequilibrium between probands and siblings in aggregate with a Fisher's exact test (by comparing summed proband and sibling variant counts for LGD versus non-LGD events) and at the level of each proband-sibling pair using the Mann-Whitney *U* test (by comparing the variant counts in each proband-sibling pair). In addition, we used a logistic regression model in which the dependent variable was the presence of a variant in a proband (true or false), and the independent variables were characteristics of the variant (such as its frequency or conservation score). Note that we applied a different conditional logistic regression to assess the risk by variant class to affected and unaffected individuals within the families. We used RVIS[19] to identify genes that were not tolerant of functional or deleterious mutation in control populations (defined here by an RVIS value below the 50th percentile) and hypothesized that the score may have similar relevance to ASD genes (see also ref. 20). We examined the RVIS profile of genes in a protein-protein interaction network based on published *de novo* mutations in ASD[53] and found that these ASD-related genes had an average RVIS percentile of 26.3. This average was significantly lower than those for randomly picked sets of genes, suggesting that RVIS percentile is a relevant predictor of ASD genes ($P < 1 \times 10^{-6}$, permutation testing; **Supplementary Fig. 13**). To integrate both CNV and SNV data for specific genes, events were tabulated on the basis of variation type (SNV or CNV) and inheritance class as presented throughout this manuscript. In particular, we counted all *de novo* CNVs and LGD or missense SNV events, private inherited LGD SNVs in genes with an RVIS below the 50th percentile and rare, inherited CNVs in which at least one gene had an RVIS below the 50th percentile. From these values, we calculated *P* values for *de novo* SNVs[6] and inherited SNVs and CNVs (binomial test). Genes were ranked on the basis of a Fisher's combined *P*-value test (**Supplementary Table 6**; family-based aggregation shown in **Supplementary Table 14**). We also applied RV-TDT[21] using trio data (parents and either the affected proband or the unaffected sibling) to test for an association between rare, inherited LGD events in conserved genes (at two different RVIS percentile cutoffs, <50 and <20) and ASD.

Combined gene-level ranking. For each gene (**Supplementary Table 6**), we calculated a 'delta score', defined as the difference in counts between proband and sibling in terms of the number of *de novo* LGD and missense SNVs. The delta score was adjusted for gene size and gene-specific mutation rate as described previously[6]. Because of the rarity of *de novo* CNVs and the difficulty in assigning gene-specific *P* values to large CNVs, we did not include *de novo* CNVs in this calculation. For inherited SNVs, we also calculated the proband-sibling delta score on the basis of private LGD SNVs and used a simple binomial test to rank genes. The *P* values specific to *de novo* or inherited variants were integrated using Fisher's combined *P*-value test.

Conditional logistic regression. We estimated the contribution of genetic risk to ASD for both inherited and *de novo* CNVs and SNVs using an additive conditional logistic regression model and adding strata for families (or proband-sibling pairs). This model took the form

$$\text{logit}[P(\text{ASD})] \sim \textit{de novo}\ \text{CNVs} + \textit{de novo}\ \text{SNVs} + \text{inherited CNVs} + \text{inherited SNVs} + \text{strata(family)}$$

Each term is composed of the total number of events in each individual. We included all *de novo* CNVs, all *de novo* LGD SNVs, the set of private, inherited LGD SNV mutations in genes with RVIS values below the 50th percentile and the set of rare, inherited CNVs with a minimum RVIS at or below the 50th percentile. We counted only autosomal events for all domains. The model was run with the survival.clogit function in the R language.

To test for nonlinear—or exponential—effects, we contrasted two simplified logistic regression models. In the first, we predicted proband (ASD) or sibling (unaffected) status simply on the basis of the summed number of mutations defined above (and again including family-level strata). The OR for each mutation (regardless of type) in this model was 1.17 ($P < 1 \times 10^{-8}$). In the second model, we added a term consisting of the total number of mutations squared. In this model, the simple sum was again significant (OR = 1.20, $P = 0.002$), but the squared sum term was not (OR = 1.00, $P = 0.59$).

Overlap between genes enriched for *de novo* and inherited events. We examined whether genes enriched for *de novo* mutations were also enriched for the class of inherited, private LGD mutations. Using data from **Supplementary Table 6**, we ranked all genes by their enrichment for *de novo* mutations (via the "de.novo.SNV.p.value" column). We took the top 100 genes in this sorted list and compared the summed gene counts for all inherited CNVs and SNVs in this group against 10,000 iterations of 100 randomly selected genes (without replacement) from the list. Observation of the resulting histogram and observed values suggests that genes enriched for *de novo* mutations do not overlap with genes enriched for inherited LGD mutations or rare disruptive CNVs (**Supplementary Fig. 9**).

Population attributable risk. We assessed the contribution of different variant types to risk in the population. Included in the variant types were SNVs of the following classes: inheritance (*de novo*, private inherited), RVIS (no cutoff, 50, 20) and transmission (all, maternal, paternal). CNV classes tested included inheritance (*de novo*, rare (<0.8%) inherited), type (deletion, duplication) and size (no cutoff, <100 kb). To assess the attributable fraction (estimated) in exposed (ASD probands) and attributable fraction (estimated) in the population, we used the epi.2by2 function in epiR. We calculated population attributable risk using the method detailed in Taylor *et al.*[26]. For a given variant type, the attributable fraction in the exposed gives the fraction of cases with the variant type that have autism because of that variant type; the attributable fraction in the population is the number of cases with the variant type that have autism because of the variant type, and the population attributable risk is the proportion of autism relevant to the variant type[26]. Complete results for all categories are listed in **Supplementary Table 8**.

doi:10.1038/ng.3303

39. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
40. Stessman, H.A., Bernier, R. & Eichler, E.E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).
41. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
42. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
43. Li, B. *et al.* QPLOT: a quality assessment tool for next generation sequencing data. *Biomed. Res. Int.* **2013**, 865181 (2013).
44. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* http://arxiv.org/abs/1207.3907 (2012).
45. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
46. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
47. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
48. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22**, 1525–1532 (2012).
50. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
51. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010).
52. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
53. Krumm, N., O'Roak, B.J., Shendure, J. & Eichler, E.E. A *de novo* convergence of autism genetics and molecular neuroscience. *Trends Neurosci.* **37**, 95–105 (2014).
54. Ritchie, M.E., Carvalho, B.S., Hetrick, K.N., Tavare, S. & Irizarry, R.A. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics* **25**, 2621–2623 (2009).
55. Scharpf, R.B., Irizarry, R.A., Ritchie, M.E., Carvalho, B. & Ruczinski, I. Using the R package crlmm for genotyping and copy number estimation. *J. Stat. Softw.* **40**, 1–32 (2011).