DISEASE MECHANISMS — OPINION

The role of *de novo* mutations in the genetics of autism spectrum disorders

Michael Ronemus, Ivan Iossifov, Dan Levy and Michael Wigler

Abstract | The identification of the genetic components of autism spectrum disorders (ASDs) has advanced rapidly in recent years, particularly with the demonstration of *de novo* mutations as an important source of causality. We review these developments in light of genetic models for ASDs. We consider the number of genetic loci that underlie ASDs and the relative contributions from different mutational classes, and we discuss possible mechanisms by which these mutations might lead to dysfunction. We update the two-class risk genetic model for autism, especially in regard to children with high intelligence quotients.

The terms autism and autism spectrum disorders (ASDs) refer to severe developmental defects in social response and communication that are accompanied by inappropriate repetitive behaviour. ASDs afflict ~1% of the human population[1,2]. There is a strong male bias among those diagnosed, especially among those who are less severely affected[3]. The disabilities that are associated with ASDs are such that the affected children do not generally become parents[4]. Because of this, the population models and the genetic mechanisms for ASDs can be expected to share elements with other paediatric or juvenile disorders that reduce fecundity[5].

Concordance for ASDs between identical twins is estimated to be between 70% and 90%, which is higher than that for any other known cognitive and/or behavioural disorders[6,7]. The risk for a newborn child is more than tenfold higher if a previous sibling has an ASD[8]. Strong monogenic risk factors are known, such as the mutations that underlie fragile X syndrome and Rett syndrome[9–11]. For all of these reasons, genetic variation is recognized as a major aetiological factor. Nevertheless, studies that have been designed to detect signals from common variants have yet to find evidence that these variants confer risk for ASDs. Moreover, findings from tests of total contribution of common variants to ASDs have ranged widely[12,13] and do not provide convincing arguments in support of a role for this type of variation. Clearly, some contribution from common variants cannot be dismissed[14], but given the current state of knowledge it seems likely that their overall impact is minor[13].

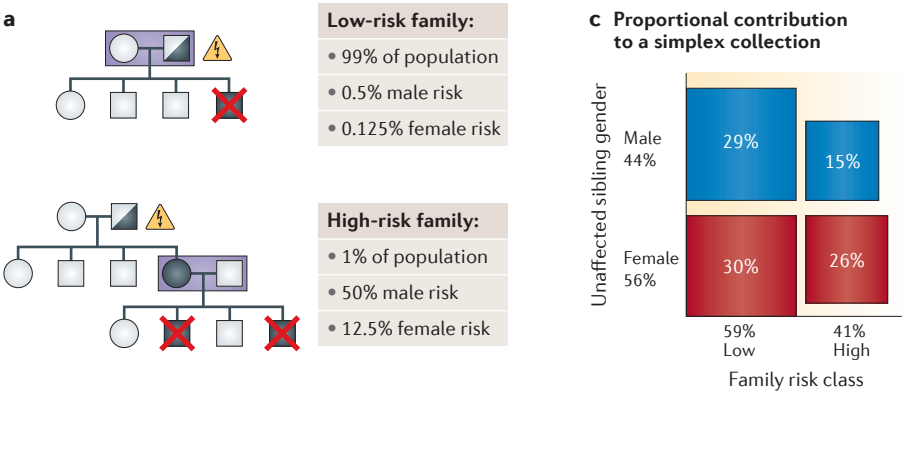
> “some contribution from common variants cannot be dismissed, but given the current state of knowledge it seems likely that their overall impact is minor”

Since 2007, studies have shown a strong source of causality for ASDs, namely *de novo* mutations (that is, new mutations) that originate in the parental germ line[15,16]. These studies allow an estimate of the magnitude of the contribution of *de novo* mutations and provide a clear path to the discovery of candidate genes[17]. In this Opinion article, our purpose is to review these developments, to estimate the contribution of *de novo* mutations to ASDs and to determine the number of genetic loci that contribute in this way. We investigate mechanisms by which *de novo* mutations may result in dominant phenotypes. We also update our unified two-class risk genetic model in light of current findings from studies of simplex families and speculate on the contribution of inherited mutations in light of gender bias. We find good support for our model in females and in males with low intelligence quotients (IQs), but this model requires revision to explain the smaller and even more gender-biased set of affected individuals with high IQs.

Population models of autism risk

It is necessary, from the start, to have some quantitative understanding of the different types of family risk for ASDs. In the case of autism, it has long been known that having one affected child increases the likelihood that subsequent children will also be affected. A recent estimate is a ~20% risk for males if a previous sibling has been affected[18]. But are there distinct classes of risk? It was the failure to ask this question that initially led to premature conclusions about the number of loci that are involved in each family at risk. These earlier studies applied conventional genetic models to fairly small ASD cohorts and concluded that the observed signal was most compatible with the involvement of multiple loci and that each locus conferred only low or moderate risk[19,20]. More recently, we recognized that, for families with two previously affected siblings, the chance that a third-born boy is affected is nearly 50%[18,21], which suggests that there is a single dominant transmitted trait for most high-risk families.

From knowledge of overall incidence and the observed family recurrence rates, one can make constrained models of the family risk distribution with no assumptions about genetics. Six observations are needed: overall risk to a newborn child (that is, incidence), risk if there is a previous sibling with an ASD and risk if there are two previous siblings with ASDs, all three of which need to be broken down by the gender of the newborn child. It is mathematically demonstrable that there are at least two positive risk classes among families with affected children[21].



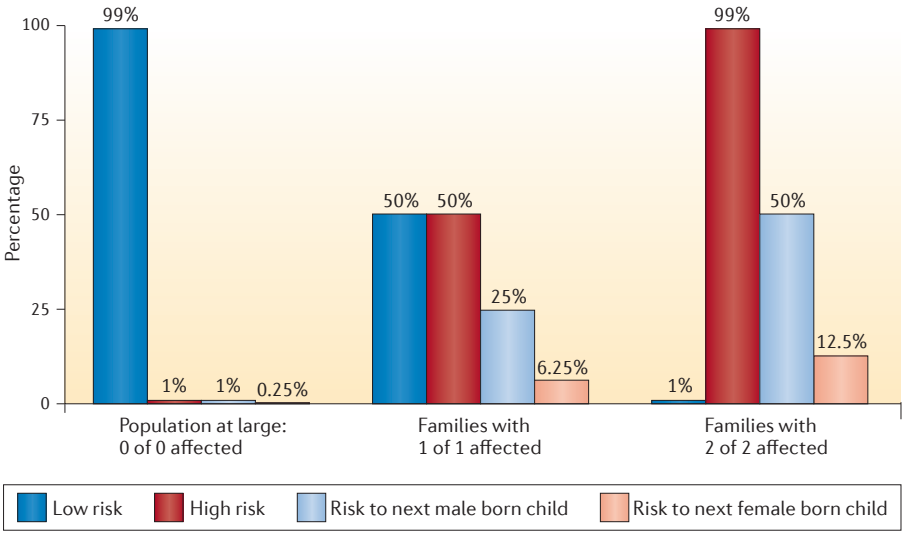


Figure 1 | **A unified two-class risk model and its consequences for the composition of a simplex collection.** **a** | The Zhao model with two positive risk classes[21] is shown. Yellow triangles indicate the incidence of *de novo* mutations; individuals with potentially causal germline *de novo* mutations are represented by partially shaded pedigree symbols, whereas individuals who inherit such alleles are fully shaded. Affected status is indicated by the red 'x'. Males are indicated by squares and females by circles. In this simplest of models, there are no families with zero risk. The children of low-risk families (upper panel) are susceptible to *de novo* mutations that occur in a parental germ line, and their risk is 0.5% for boys. These families comprise the great majority (99%) of the population. In the remaining 1% of families (lower panel), one parent carries a highly penetrant allele, which puts male children at a high risk of 50%. In this example, the mother is shown as the carrier. For both risk classes, and for subsequent estimations, we set the penetrance of mutations in girls at 0.25 to match the male:female incidence ratio. **b** | Restating the assumptions of the Zhao model[21] in the absence of further information, the high-risk families constitute 1% of all families. It follows that, for families with one child who has been ascertained as being affected with an autism spectrum disorder (ASD) — and with no other information about the other children of these families — the proportions of low- and high-risk families are equal. Among families with two children who have been ascertained as being affected by an ASD, nearly all families are at high risk. With knowledge of whether children from previous births are affected by an ASD, the risk to the next born child is the sum of the family risks that are weighted by the proportion of the family risk class. These risks approximately match the observed risks. **c** | The composition by gender and risk class for a simplex collection is shown. Under the assumptions of the two-class risk model, there is little exclusion of high-risk families because both children need not be affected, especially when one is female. With these assumptions, we estimate that only ~60% of the families in a simplex collection are actually at low risk. Moreover, as the female children of high-risk families have lower risk than males, such a collection should have a bias in the gender of the unaffected children. The predictions of this simple model match the observed gender bias of the Simons Simplex Collection: of unaffected siblings, 44% are male and 56% are female[22].

Whereas there are infinite models with any number of risk classes that satisfy the observations, all models share two properties: a substantial proportion of ASDs come from high-risk families and a substantial proportion come from low-risk families (which have low but non-zero risk). At a minimum, any successful genetic model must account for the source of ASDs in low- and high-risk families, and a 'unified' model must relate the two.

To illustrate these principles, we consider the two-class risk model, which is the simplest model that can satisfy the observations of sibling risk (FIG. 1a,b). The particular two-class model presented here is unified because it incorporates the connection between *de novo* and transmitted mutational events[21]. However, the consequences of the risk model are independent of the genetic mechanism. In this model, high-risk families constitute a small proportion of the population but account for a large proportion of autism incidence (FIG. 1b).

Below, we describe published results on studies of simplex families. These families only have a single affected child (for example, those in the Simons Simplex Collection[22] (SSC)) and are in contrast with multiplex families in which two or more children are affected (for example, those in the Autism Genetic Repository Exchange (AGRE) collection[23]). As most families have few children and as females are resistant to the disorder[24], simplex collections are made up of all family risk classes. On the basis of the two-class model, we suggest that there is little enrichment for low-risk families in simplex collections and that the other classes could make up nearly 50% of the collections (FIG. 1c).

Contribution of *de novo* mutations

In 2004, two groups used comparative genomic hybridization (CGH) microarrays to discover a rich source of genetic variability in the human gene pool in the form of large-scale copy number variants (CNVs)[25,26]. As these variants were often found to overlap with genes, it was suggested that they would frequently alter gene dosage and thus have functional consequences. Such a rich source of variation would have an intrinsic rate of origination through *de novo* mutations, and the tools for detecting CNVs were the first to be used to test the idea that *de novo* mutations widely contribute to ASD incidence. The initial studies used 'trios' and searched for CNVs in the child that were not present in either of the

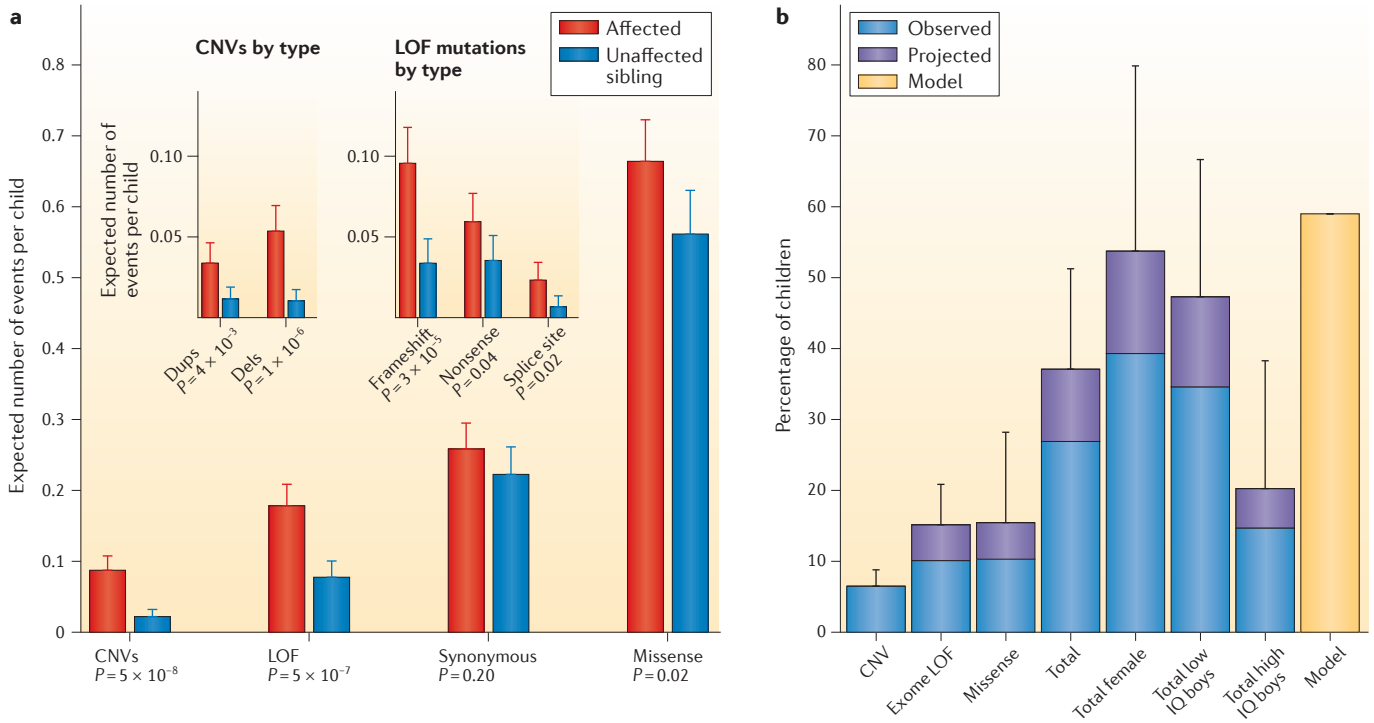


Figure 2 | **Differential signal of *de novo* mutations in affected and unaffected siblings. a** | The observed incidence per child of various types of *de novo* mutations is shown; these mutations include copy number variants (CNVs) such as duplications (dups) and deletions (dels), as well as loss-of-function (LOF) mutations such as frameshift, nonsense and splice-site mutations. The *P* values of the differentials are listed below each designated type. The copy number data are based on REF. 27; the data from exome sequencing of families in the Simons Simplex Collection (SSC) — which include LOF, synonymous and missense mutations — are taken from three published data sets (REFS 34,36,37) with our own reprocessing of data from REF. 37 to include insertions and deletions. The error bars represent 95% confidence intervals. **b** | A summary of the known and potential contributions to autism spectrum disorders (ASDs) of different mutational classes is shown. Under the assumption that the observed differential in the rate of *de novo* mutations is due to events that contribute to ASD phenotypes, the proportion of affected children with such events is equal to the differential in rates. Owing to limitations in exome capture and in sequencing coverage, we estimate that the SSC exome sequencing studies capture only two-thirds of exonic variants. On the basis of these numbers, we project an additional 50% differential for missense and LOF variants. All observed, adjusted and projected differentials are summed under 'total'. We make a similar projection for females alone (shown as 'total female'). Given the calculations of the composition by gender and risk class for a simplex collection (FIG. 1c), we expect that 60% of the children in the SSC arise from low-risk families, which is labelled 'model'. The total contribution of the mutational load for all affected individuals is 50% less than that needed to explain the observations for all children of low-risk families, but is close to that expected for female children. The error bars represent 95% confidence intervals. IQ, intelligence quotient.

parents. In 2007 and 2008, two groups published reports showing that *de novo* CNVs are more abundant in children with ASDs than in controls[15,16].

In 2010 and 2011, three studies confirmed these results in much larger sets of samples[27–29]. The two studies in 2011 were more definitive: one made use of extensive correction for system noise[30] and the highest-density microarrays that were available at the time[27], and both analysed families from the SSC[27,29]. The 'quad' structure of the SSC included an unaffected child in each family, which provided a robust control, and the SSC probands were ascertained using highly consistent clinical criteria. Each of the three studies used CGH platforms that could effectively detect CNVs of 20 kb or larger. The two groups that studied the SSC concurred that in children with autism, *de novo* CNVs were more frequent (8%) and richer in genes than in their unaffected siblings, with a 6% differential in frequency (FIG. 2).

With the advent of inexpensive exome sequencing, large studies of the contribution of *de novo* point mutations and small insertions and deletions (indels) became affordable[31–33]. Four groups published similar findings on different cohorts[34–37], which were predominantly from the SSC[34,36,37]. Heterozygous *de novo* loss-of-function mutations (LOF mutations) — those that create stop codons, cause frameshifts or alter splice sites — occur in ~20% of the probands but in only ~10% of the unaffected siblings[34,37]. Thus, *de novo* LOF mutations contribute to at least 10% of simplex cases, although this is undoubtedly an underestimate as discussed below. Neither transmitted nor *de novo* LOF mutations were present in the remaining allele.

One study reported a differential signal for *de novo* missense mutation between probands and unaffected siblings[37], but this was not found in another study[34]. As a combined study, the net differential is significant. *De novo* missense mutations are far more abundant than LOF mutations, and most have no effect; hence, there is a loss of statistical power in the difference. Taking the two studies together, *de novo* missense mutations may contribute to ASD in up to 10% of affected children (FIG. 2).

De novo LOF mutations, including both CNVs and those detected by exome sequencing, are more frequent in autistic females than in autistic males[27,34]. Consistent with this, large CNVs, which encompass many genes and are thus

presumably more damaging, are especially enriched in affected females[27,29]. This enrichment suggests to us that other causative factors that have not yet been identified make a considerable contribution in males but have lower effects in females.

The risk of some disorders, including ASDs, is dependent on paternal age[38–40]. In fact, the majority (~75%) of *de novo* point mutations originate from the father, and the rate of germline mutation increases with paternal age[34–37,41]. On the basis of our published work[34], we see a correlation between paternal age and the number of *de novo* single-nucleotide variants (SNVs) that arise from the paternal germ line, from which we derive a 1.3-fold increase in the number of *de novo* SNVs for every ten years of paternal age. The rate of increase in autism incidence is similar to that in the number of germline mutations, which is consistent with the idea that new mutation governs the increased risk with paternal age, although age-related risk might derive from other phenomena, such as epigenetic changes within gametes.

Hits	Published data (948 probands)	Published and updated data (948 probands)
1	123	146
2	5	5
3	0	1

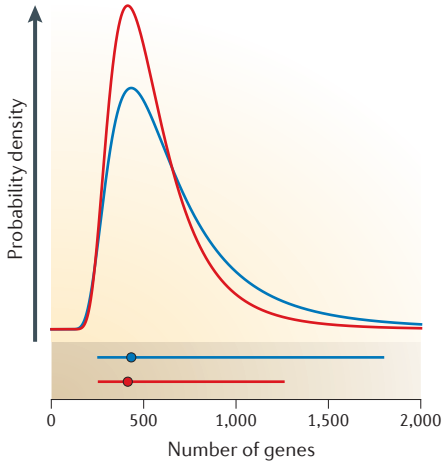


Figure 3 | **Estimates of ASD gene target sizes.** Assuming that all target genes have a uniform probability of mutation and detection, we can apply a likelihood method to estimate the number of target genes. Using all published data on 948 probands, we observe that there are 133 loss-of-function (LOF) mutations, including five genes that are hit recurrently. Assuming that half of the LOF mutations contribute to autism spectrum disorder (ASD) phenotypes and occur in target genes, we compute a posterior probability distribution for the target size (blue curve) with a maximum likelihood solution (blue dot) of 430 genes and a 95% confidence interval (blue line) of 250–1,800 genes. When we add our calls for insertions and deletions (which is indicated as published and updated data) on the data from REF. 37, the probability distribution (red curve) narrows, with a maximum likelihood solution (red dot) of 415 genes and a 95% confidence interval (red line) of 260–1,250 genes.

Target genes and individual vulnerability

Here, we define a target gene or locus as a transcription unit that, when mutated, can significantly contribute to the incidence of ASDs. The identities and the types of such genes are of great importance but, in this section, we focus on their target size (that is, the number of causal genes), which is an essential parameter in any genetic model. Our estimate of target size comes from exome sequence analyses. There are two methods for the estimation of target size.

The first method uses recurrence analysis (FIG. 3). The fundamental idea is that if all genes are assumed to have an equal probability of being mutated and detected, then the number of recurrently hit genes that is found in a given number of hits has a predictable distribution that is only dependent on the number of target genes (N). Conversely, the likelihood of a given target size (N) can be computed from the observed numbers of recurrences and hits. Calculating the number of hits requires an important adjustment: it is not the total number of genes hit by *de novo* mutations in affected individuals that determines target size, but the number of those observed mutations that are likely to have functional consequences. As discussed above, only half of the observed *de novo* mutations are likely to be causal. Additionally, an adjustment should be made for coincident mutations among the remainder of non-causal mutations, but this is a small correction and is therefore omitted. The numbers from published data suggest that the most likely target size is ~450–500 genes. Work in progress sharpens and confirms this estimate (I.I., M.R., D.V. and M.W., unpublished observations). The assumption that all target genes have an equal likelihood of being mutated and detected is obviously wrong, and this has two consequences. On the one hand, any computation of target size that is based on these assumptions is an underestimate. On the other hand, most contribution to ASDs from *de novo* mutations would come from fewer targets than we estimate from recurrence.

The second method computes the mean individual genetic vulnerability[34]. We first define the pre-mutant zygote (PMZ) as the putative zygote of the child without any of the new parental germline mutations that actually contributed to the zygote. The individual genetic vulnerability of the child is the number of genes that are likely to cause the development of the disorder if they were disrupted in the PMZ. An estimate of the mean of this value can be determined from R_{aut} and R_{pop}, which are the rates of *de novo* LOF mutations in individuals with ASDs and in the general population, respectively. If the incidence of ASDs in the population is 1 in 100, then in a population of size P, there are $P \times 0.01 \times (R_{aut} - R_{pop})$ *de novo* LOF mutations that contribute to ASDs. However, the total number of LOF mutations in such a population is $\sim P \times R_{pop}$. Therefore, the proportion of *de novo* LOF mutations that contribute to ASDs is $0.01 \times (R_{aut} - R_{pop})/R_{pop}$. From the SSC studies[33,35,36], we can take $R_{aut} = 0.2$ and, using the unaffected siblings, we estimate R_{pop} to be 0.1. Applying our formula, 1 in 100 *de novo* LOF mutations contribute to ASDs. If all genes have an equal mutation rate and if all LOF mutations are completely penetrant, then 250 target genes of a total of 25,000 genes would be predicted as the target size. If penetrance is 50% in the population — for example, if not all children were vulnerable to the effects of mutation or if children had different target gene vulnerabilities — then this method of estimation and the previous method would agree. Despite making different assumptions, the two methods of estimation are in fairly good agreement, which suggests that the underlying assumption of high penetrance is not far off. Both estimates predict that there are hundreds of target genes of LOF mutations that contribute to ASDs.

Insights into target gene function

On the basis of the calculations above, the study of new mutations has yielded a rich source of candidate genes. Scanning the exome for *de novo* LOF mutations in affected children from the ~2,500 families of the SSC alone should result in the generation of ~500 candidate genes, each with a ~50% chance of being contributory. The recurrently hit genes among these will, of course, be far better candidates. Summing over published work, there are currently ~150 candidates, and the false-positive rate

is 50% on the basis of the rate of *de novo* LOF mutations in unaffected siblings. Nine are recurrently hit genes (TABLE 1). A network analysis suggests a signal for genes that are involved in the neuroskeleton and the synapse[42], and for genes that encode chromatin modulators[43]. In 2011, one study[44] hypothesized that genes with transcripts that bind to FMRP — the protein encoded by the fragile X mental retardation locus — are enriched for targets of mutations in ASDs. This hypothesis now has strong statistical support from published exome sequencing studies[34]. Relative to other gene classes, LOF mutations in FMRP-associated genes are underrepresented in the parents of affected children in the SSC (TABLE 2), and *de novo* LOF mutations are not seen at a higher rate than the background mutation rate in unaffected siblings (TABLE 3). But in affected children, perhaps 30% of ASD-target gene transcripts bind to FMRP (TABLE 3), assuming a 50% false-positive rate among target genes. Such a strong signal of overlap is not seen between candidate target genes for ASDs and the sets of genes that are expressed in the brain, or even in postsynaptic densities. Accumulating data continue to provide strong support for disruption of FMRP-associated genes in the pathogenesis of ASDs (I.I., M.R., D.V. and M.W., unpublished observations). FMRP itself is involved in neuroplasticity through its role in modulating long-term potentiation and depression[45,46]. Thus, a sensible general hypothesis is that disruption of neuroplasticity contributes to ASDs.

Gene dosage and dominance

Several lines of evidence point to altered gene dosage as the major effect of new mutations that contribute to ASDs. This is a direct inference from the observation of increased incidence of CNVs, including both deletions and duplications, in affected individuals. This idea receives additional support from exome sequencing, which shows that the new mutations in children with ASDs rarely occur opposite an allele that is already defective[34].

The mechanism of dosage sensitivity is far from established. Sensitivity could occur in three different ways. First, decreased expression of critical genes, even by a half, could cause partial dysfunction. The genes that are required for the recently acquired human traits of speech and complex social behaviours may be particularly vulnerable. Second, if expression of a target gene locus is monoallelic, then a loss of expression

Table 1 | **Genes with recurrent *de novo* loss-of-function mutations***

Gene	Total hits	WES hits	Additional TGS hits‡
CHD8 (chromodomain helicase DNA-binding protein 8)	8	2	6
DYRK1A (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A)	3	3	0
GRIN2B (glutamate receptor, ionotropic, *N*-methyl D-aspartate 2B)	3	1	2
KATNAL2 (katanin p60 subunit A-like 2)	2	2	–
RIMS1 (regulating synaptic membrane exocytosis 1)	2	2	–
SCN2A (sodium channel, voltage-gated, type II, alpha subunit)	2	2	–
POGZ (pogo transposable element with ZNF domain)	2	2	–
ADNP (activity-dependent neuroprotector homeobox)	2	1	1
ARID1B (AT-rich interactive domain 1B (SWI1-like))	2	1	1
TBR1 (T-box, brain, 1)	2	1	1

TGS, targeted sequencing study; WES, whole-exome sequencing. *Most *de novo* events in recurrently hit genes were discovered by WES of 948 affected children. The data shown in this table are taken from published Simons Simplex Collection (SSC) exome sequencing results, and we have added our calls for insertions and deletions by reprocessing the data from REF. 37. ‡Additional *de novo* mutations were discovered in a TGS of 44 candidate genes from 2,446 SSC probands[43]. Genes that were not included in the TGS are indicated with a dash.

Glossary

Coincident mutations
Mutations in both alleles at a given locus.

Comparative genomic hybridization
(CGH). A microarray-based technique for identifying large deletions or duplications in the genome.

Concordance
The probability that multiple siblings are affected given that one of them is already known to be affected.

Copy number variants
(CNVs). Large deletions or duplications that either alter the number of copies of genes or disrupt the function of genes.

***De novo* mutations**
New mutations that arise either in the parental germ line or somatically.

Dosage sensitivity
A defining feature of phenotypes that result from heterozygous mutation.

Gender bias
The phenomenon whereby four times as many males are affected by autism spectrum disorders compared with females, with a male:female ratio of nearly 6:1 among individuals who are diagnosed as being high functioning.

High-risk families
Families that contain a highly penetrant segregating risk allele for autism spectrum disorders.

Insertions and deletions
(Indels). Small insertions or deletions in the genome that are generally < 10 bp.

Loss-of-function mutations
In the context of this article, events that result in a nonsense allele or that change the reading frame.

Low-risk families
Families that do not contain a segregating risk allele for autism spectrum disorders (ASDs) and that are only at risk of ASDs in cases of *de novo* mutation.

Monoallelic
Pertaining to the expression of only one allele at a given locus.

Multiplex families
Families with multiple affected children.

Neuroplasticity
The dynamic state of the brain, which enables it to respond to changes in environment and development.

Penetrant
Pertaining to the probability that an individual with a given mutation will be affected by the corresponding condition.

Recurrence
Independent mutational 'hits' within a given gene in unrelated individuals.

Sibling risk
The probability that a sibling of an affected child will also be affected.

Simplex families
Families with only one affected child; all other children (if any) of these families are unaffected.

Transmitted
Inheritance of a mutant allele from a parent, who may be phenotypically normal owing to gender bias.

Trios
Family units that consist of both parents and one child in each unit.

Table 2 | **Protection from LOF mutations in parents of affected children from the SSC***

Gene class‡	Number of genes	Proportion of exome	Overlap with ultra-rare synonymous variants in parents (44,701 variants)		Overlap with ultra-rare LOF mutations in parents (4,824 mutations)§		
			Observed	Proportion of variants	Observed	Expected	Ratio
Mendelian disease genes	256	0.02	980	0.02	91	106	0.86
Chromatin modifiers	428	0.03	1,470	0.03	65	159	0.41
FMRP targets	842	0.10	4,852	0.11	111	524	0.21
PSD genes	1,445	0.09	4,450	0.10	228	480	0.48
Essential genes	1,750	0.12	5,732	0.13	260	619	0.42
Brain-expressed genes	14,727	0.77	35,344	0.79	3,372	3,814	0.88

FMRP, fragile X mental retardation protein; LOF, loss-of-function; PSD, postsynaptic density; SSC, Simons Simplex Collection. *For each gene class, we report the number of genes in the class and the proportion that they contribute to the total coding capacity in the genome. Data were obtained from the parents of the SSC sequencing studies[34,36,37], as well as from additional analysis that we carried out using methods described in REF. 34 on one published data set for which calls for insertions and deletions were not initially made[37]. There are a total of 44,701 ultra-rare synonymous variants, which are variants that are seen in only one parent. We compute the proportion of these variants in each gene class and use this to correct for uneven distribution of gene number, gene length and sequence coverage. ‡Mendelian disease genes consist of positionally cloned human disease genes[68], whereas essential genes are the human orthologues of mouse genes that have been associated with lethality in the Mouse Genome Database[69]. FMRP targets are genes with transcripts that are observed to bind to FMRP[44], and PSD genes encode proteins that are identified in postsynaptic densities[70]. Brain-expressed genes come from the expression profiles of post-mortem human brains[71], and the classification of chromatin modifiers is derived from the Gene Ontology[72]. §The last set of columns lists observed and expected overlaps of the 4,824 ultra-rare LOF mutations with the gene classes. Expectation within a class is based on the total number of mutations as predicted by the proportion of synonymous variants that fall in that class. The ratio is obtained by dividing the observed count by the expected count. All of the gene classes have a significantly skewed ratio; however, FMRP is exceptionally depleted for LOF mutations, with fivefold fewer variants than expected.

from a defective allele could result in a net homozygous deficit in terms of function[47]. We do not know the proportion of monoallelically expressed genes in the brain, but it could be as high as 5%[47,48], which greatly exceeds current expectations of the amount of genomic imprinting[49]. Last, some LOF mutations in coding regions might also create dominant alleles by altering gene products. For example, prematurely terminated proteins can either interfere with the function of the remaining wild-type proteins (that is, dominant-negative LOF mutations) or lose inhibitory regulatory control (that is, gain-of-function mutations).

Is there an identifiable class of genes that are enriched in the property that mutation in a single allele is dominant? Our data on the incidence of LOF mutations in FMRP-associated genes suggest that this is the case (TABLES 2,3). For various gene classes, we counted the incidence of rare LOF mutations in the human gene pool and normalized that count on the basis of the incidence of rare synonymous mutations within the gene class. The classes we examined included the FMRP-associated genes, as well as known targets of Mendelian inherited disabilities, essential genes that were discovered in mouse modelling, genes that are expressed in the brain and in postsynaptic densities, and genes that encode chromatin modifiers. In humans, the FMRP-associated genes show a marked 'protection' from the accumulation of LOF mutations (TABLE 2). Thus, LOF mutations in the FMRP-associated genes may generally be deleterious and dominant.

The missing burden of transmission

The first large-scale attempt to determine genetic mechanisms for ASDs was mounted by the AGRE[23], which is a collection of pedigrees and biological samples from nuclear and extended families. A transmitted genetic signal was expected to be observed most easily in multiplex families, which was a major part of the rationale for putting together the collection. However, the linkage studies that were carried out using the AGRE collection did not reproducibly identify loci of major effect[50–53]. In parallel, genome-wide association studies that were carried out using large case-control cohorts[54–56] identified few significantly replicated signals, which is in violation of the common disease–common variant hypothesis[57]. Nevertheless, many in the field believed — and many still do[12] — that the transmitted genetic component of autism is mostly the result of the chance combination of many loci of small effect, which we refer to as the 'complex interaction' mechanism[54]. Such a mechanism is plausibly invoked to explain quantitative traits such as height and IQ[58,59]. However, it is hard to see how such a mechanism can explain more than a minor proportion of ASDs[12] because it fails to explain the large contribution of autism from high-risk families[18,21], ignores the unification of genetic mechanisms and fails to account for individuals with highly penetrant LOF mutations who, nonetheless, have mild phenotypes and continue to produce offspring.

Evidence for transmission has been sought in the families in the SSC[27,34,60]. As discussed above, we estimate that up to half of the simplex collection is comprised of high-risk families[27]. Moreover, transmission might have a role in low-risk families, in which new mutations merely push children who are already at risk over the threshold. If this is the case, then some signal from transmission would be observed. In fact, transmission of rare CNVs was observed to occur at higher rates to the affected than to the unaffected child in the SSC, but the significance was marginal[27]. In initial exome sequencing studies of the SSC, there was a weak signal from compound heterozygous rare variants, but this was also of low significance[34].

A recent report finds a small signal in the form of compound heterozygous and rare homozygous LOF variants in ASD, which accounts for perhaps a 5% differential between affected and unaffected children[60]. This study used samples from both multiplex and simplex families. However, the overall signal from transmission is so far underwhelming.

The unified two-class risk model

There are excellent reasons for seeking a quantitative and more accurate genetic model for ASDs. Even an incomplete genetic model that does not fully account

Table 3 | **Enrichment for *de novo* mutations in affected children from the SSC***

Gene class	Overlap with *de novo* LOF mutations in affected children (157 mutations)			Overlap with *de novo* missense mutations in affected children (615 mutations)			Overlap with *de novo* LOF mutations in unaffected siblings (46 mutations)			Overlap with *de novo* missense mutations in unaffected siblings (333 mutations)		
	Observed	Expected	P‡	Observed	Expected	P‡	Observed	Expected	P‡	Observed	Expected	P‡
Mendelian disease genes	2	3	0.78	16	13	0.49	1	1	0.99	12	7	0.09
Chromatin modifiers	18	5	4×10^{-6}	32	20	0.01	3	2	0.19	15	11	0.22
FMRP targets	37	17	7×10^{-6}	71	67	0.56	2	5	0.23	47	36	0.06
PSD genes	20	16	0.23	64	61	0.69	9	5	0.04	41	33	0.17
Essential genes	31	20	0.02	83	79	0.63	7	6	0.66	52	43	0.14
Brain-expressed genes	134	124	0.06	502	486	0.12	34	36	0.37	260	263	0.64

FMRP, fragile X mental retardation protein; LOF, loss-of-function; PSD, postsynaptic density; SSC, Simons Simplex Collection. *The proportions of synonymous mutations determine the expected number of hits within each of the six gene classes (TABLE 2). ‡*P* values are computed from a two-sided binomial test and give the significance of the deviation between observed and expected. By this measure, chromatin modifiers and FMRP targets are disproportionately burdened with *de novo* LOF mutations in the affected children relative to their unaffected siblings. The FMRP hypothesis was made prior to, and is hence blind to, the acquisition of exome sequencing data[44].

for all of the incidence of ASDs would be of great aid in early diagnosis, treatment and counselling, as well as in the design of future studies. Any complete model for ASDs would need to explain many puzzles, including overall incidence, risk classes, gender bias, gene dosage sensitivity, partial penetrance and the observations that some genetic lesions that are firmly associated with ASDs are also associated with other cognitive impairments[13,61]. By that standard, we still have a long way to go. For now, we consider the performance of the simplest model that is consistent with the observed parameters of overall and sibling risk — the unified two-class risk model.

The most conspicuous gap in our understanding is the nature and the origin of concordant sibling risk. Explaining concordance between monozygotic twins is intrinsic to any genetic model. The observed discordance between monozygotic twins is more problematic and leaves room for the 'environment' — broadly defined to include stochastic processes, DNA methylation, chromatin modification, early somatic mutation, and *in utero* and postnatal events — to play a part. Some increase in concordance between dizygotic twins over that observed in non-twin siblings has been reported, but only in small samples[7,18,62]. What is truly needed is the calculation of accurate monozygotic, dizygotic and non-twin sibling concordance rates, sorted by gender, a sibling concordance rate that is properly defined and adjusted for stoppage (that is, the parental choice not to bear further offspring after having an affected child) and phenotypes that are uniformly assessed. Such data are not found in the literature. An increase in dizygotic concordance over sibling concordance, if it does exist, might arise either from the environmental causes listed above or from a higher likelihood of shared germline *de novo* events in dizygotic twins (which has yet to be demonstrated). As stated above, future studies are needed to clarify this important issue.

Ultimately, the most important observation to explain is sibling risk. The overall sharing of *de novo* mutations between siblings is too infrequent[34] to explain high sibling risk. To account for such risk, the simplest hypothesis is that high-risk families transmit one or more highly penetrant alleles, which arose in recent generations as new mutations but are carried in parents who have 'overcome' the defective allele. Considering the skewed gender ratio of ASD diagnoses, females are a likely source of carrier individuals, but conclusive evidence in support of this idea is still lacking.

We do have a good explanation for incidence in low-risk families, which is the acquisition of *de novo* gene-disrupting mutations. A full account of new mutations is currently lacking: we have good data so far only for CNVs that are larger than 20 kb, as well as small indels and SNVs in the exome. Moreover, the data for small indels and SNVs are incomplete owing to insufficient coverage. However, the number of missing small mutations in the exome can be accurately estimated (FIG. 2b). There are other sources of *de novo* variation that cannot be readily identified with exome sequencing, which include large-scale copy number-neutral rearrangements[63], medium-scale copy number changes[64,65], highly variable long repetitive regions, mutations that affect gene expression but that fall outside the exome[66] and somatic mutations[67]. We can only guess the relative contributions from these other sources, and they are therefore omitted from the 'total' in FIG. 2b.

From the simplest two-class risk model, we expect ~60% of children in simplex collections to come from low-risk families[27]. The model predicts that these children will carry *de novo* mutations of strong effect. Tallying the contribution of *de novo* mutations, including both the observed and projected contribution (FIG. 2b), we project that only 35% of SSC probands actually have causal *de novo* small exomic or large copy number mutations, which represents a large shortfall. However, the incidence of *de novo* mutations in females is projected to be 55%, which matches well with the expectation of the simple two-class risk model (FIG. 2b). The entire shortfall from the model is in the affected males and, as discussed below, only in males with high IQs.

There may be a separate class of affected males in the SSC. The collection began as a study of high-functioning individuals with

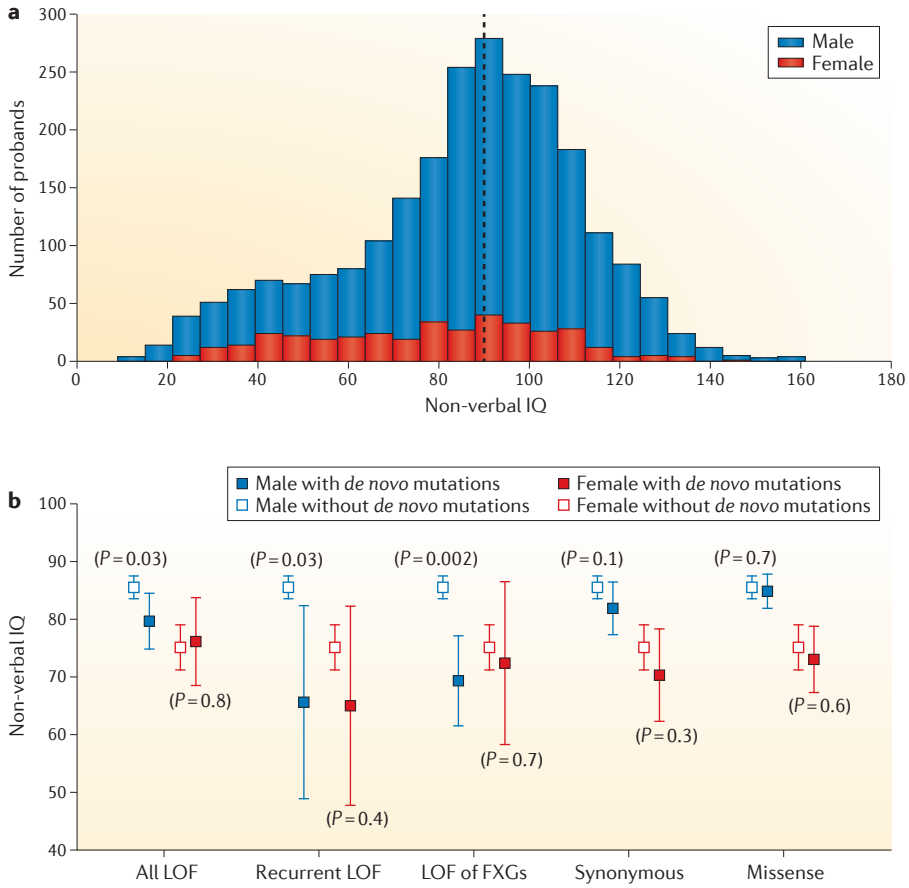


Figure 4 | **Non-verbal IQ in SSC studies by gender and mutational type.** **a** | The distribution of non-verbal intelligence quotient (IQ) in the Simons Simplex Collection (SSC) by gender is shown. Affected females account for 13.5% of the collection and have a mean IQ of 78, whereas affected males have a mean IQ of 86. The *P* value, from which the means are drawn by sampling from the same distribution, is 10^{-7} as determined by Student's *t*-test. The vertical dashed line indicates a non-verbal IQ of 90; below this threshold, affected males have a rate of *de novo* loss-of-function (LOF) mutations that is nearly equal to that observed in affected females (FIG. 2b). **b** | Effect of *de novo* mutations on non-verbal IQ by gender is shown. Affected children were drawn from the three published SSC exome studies with additional processing of the data from REF. 37 for small insertions and deletions. We consider five classes of *de novo* mutations, including all LOF mutations, LOF mutations in recurrently hit genes, LOF mutations in FMRP-associated genes (FXGs), synonymous mutations and missense mutations. For each class we divide the affected children into two groups by the presence or the absence of a *de novo* mutation of the given class and count only children with sequenced exomes. We then divide each group by gender, which yields four groups per mutational class. We tested whether the means of the non-verbal IQ for children within the gender-matched groups were statistically different. The means and their 95% confidence intervals, as well as the *P* values (computed with Student's *t*-test), are shown for each class. Notably, affected boys with *de novo* LOF mutations have significantly lower non-verbal IQs than affected boys who do not carry these mutations. This trend is even more marked in recurrently hit genes and in FXGs. By contrast, there are no significant differences for synonymous and missense mutations in affected males, or for any mutational class in affected females.

ASDs, and this group is known to have an even more marked male bias than ASDs overall[3]. Affected males in the SSC have statistically higher IQs than affected females (FIG. 4a). No significant difference in IQ distribution is seen between males that bear *de novo* synonymous mutations and those with missense mutations. However, males with *de novo* LOF mutations have an IQ distribution that is statistically indistinguishable from all affected females (FIGS 2b,4b). IQ is statistically lower in males with *de novo* LOF mutations than in males without detectable mutations ($P = 0.03$). This trend is even stronger in the group of males with mutations in gene sets that are enriched for high-confidence candidates (that is, recurrent LOF mutations and/or LOF mutations in FMRP-associated genes). It is worth noting that the effects of mutations on IQ might be better estimated by considering the IQs of parents, but parental and sibling IQ data are absent from the SSC.

Future perspectives

The results of exome sequencing of the entire SSC, which consists of ~2,500 families, will be published in the near future. Extrapolating from the earlier published work on ~700 families[34,36,37], we expect that ~50 genetic loci with recurrent LOF mutations will be found, and that nearly 100 loci will be found either in affected females or overlapping with the FMRP-associated gene class. Overall, we predict that there will be ~300 candidate genetic loci, with a net false discovery rate of ~50%.

Obviously, more genome sequencing and candidate resequencing studies are needed to get a more complete list of strong candidates. However, we expect that the large number of candidates from the SSC will soon greatly enhance the prospects of addressing many unresolved problems. The mechanisms of gene dosage can be explored. We can determine whether monoallelic expression is a mechanism for each candidate gene by examining allele-specific expression in human brains[47]. We can explore a statistically unexpected overlap between candidate genes and the list of genes that are reported to be monoallelically expressed in mice. We can determine mechanisms for dominant interfering alleles by comparing the precise mutations that are found in humans with heterozygous gene deletions in animal models, brain slices or cell-based systems, such as human induced pluripotent stem cell cultures.

Aspects of the two-class model can be investigated in several ways. Most importantly, with a large list of good candidates we can study parents of multiplex families, such as those found in the AGRE collection, for transmission of disruptive mutations in these candidates to affected children. We can also look at the correlation between the severity of mutations and IQs in children with ASDs. This allows testing of the possibility that affected individuals with high IQs carry mutations in the same genes that are mutated in more severely affected children but that the mutations in the high-IQ individuals have more moderate effects.

Finally, the basis of gender bias can be explored. We can ask whether mothers are indeed the more frequent carriers in multiplex families and can more generally investigate subclinical phenotypes in parents who are carriers. By targeted resequencing of a larger collection of individuals with ASDs, we can determine the genetic loci that show significant gender specificity, which might shed light on mechanisms for gender susceptibility.

All authors are at the Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.

Correspondence to M.W.
e-mail: wigler@cshl.edu

doi:10.1038/nrg3585
Published online 16 January 2014

1. Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatr. Res.* **65**, 591–598 (2009).
2. Muhle, R., Trentacoste, S. V. & Rapin, I. The genetics of autism. *Pediatrics* **113**, e472–e486 (2004).
3. Newschaffer, C. J. *et al.* The epidemiology of autism spectrum disorders. *Annu. Rev. Publ. Health* **28**, 235–258 (2007).
4. Power, R. A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse versus their unaffected siblings. *JAMA Psychiatry* **70**, 22–30 (2013)
5. Ku, C. S. *et al.* A new paradigm emerges from the study of *de novo* mutations in the context of neurodevelopmental disease. *Mol. Psychiatry* **18**, 141–153 (2013).
6. Bailey, A. *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psych Med.* **25**, 63–77 (1995).
7. Rosenberg, R. E. *et al.* Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch. Pediatr. Adolesc. Med.* **163**, 907–914 (2009).
8. Constantino, J. N., Zhang, Y., Frazier, T., Abbacchi, A. M. & Law, P. Sibling recurrence and the genetic epidemiology of autism. *Am. J. Psychiatry* **167**, 1349–1356 (2010).
9. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nature Genet.* **23**, 185–188 (1999).
10. Fu, Y. H. *et al.* Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**, 1047–1058 (1991).
11. Verkerk, A. J. *et al.* Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
12. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 9 (2012).
13. Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genet.* **45**, 984–994 (2013).
14. Devlin, B. & Scherer, S. W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).
15. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
16. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
17. Veltman, J. A. & Brunner, H. G. *De novo* mutations in human genetic disease. *Nature Rev. Genet.* **13**, 565–575 (2012).
18. Ozonoff, S. *et al.* Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics* **128**, e488–e495 (2011).
19. Jorde, L. B. *et al.* Complex segregation analysis of autism. *Am. J. Hum. Genet.* **49**, 932–938 (1991).
20. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493–507 (1999).
21. Zhao, X. *et al.* A unified genetic theory for sporadic and inherited autism. *Proc. Natl Acad. Sci. USA* **104**, 12831–12836 (2007).
22. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
23. Geschwind, D. H. *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69**, 463–466 (2001).
24. Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl Acad. Sci. USA* **110**, 5258–5262 (2013).
25. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
26. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
27. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897 (2011).
28. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
29. 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).
30. Lee, Y. H. *et al.* Reducing system noise in copy number data using principal components of self–self hybridizations. *Proc. Natl Acad. Sci. USA* **109**, E103–E110 (2012).
31. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Rev. Genet.* **12**, 745–755 (2011).
32. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
33. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
34. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
35. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
36. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
37. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
38. Lundstrom, S. *et al.* Trajectories leading to autism spectrum disorders are affected by paternal age: findings from two nationally representative twin studies. *J. Child Psychol. Psychiatry* **51**, 850–856 (2010).
39. Waller, D. K. *et al.* The population-based prevalence of achondroplasia and thanatophoric dysplasia in selected regions of the US. *Am. J. Med. Genet. A* **146A**, 2385–2389 (2008).
40. Zammit, S. *et al.* Paternal age and risk for schizophrenia. *Br. J. Psychiatry* **183**, 405–408 (2003).
41. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
42. Gilman, S. R. *et al.* Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898–907 (2011).
43. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
44. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
45. Auerbach, B. D., Osterweil, E. K. & Bear, M. F. Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature* **480**, 63–68 (2011).
46. Bear, M. F., Huber, K. M. & Warren, S. T. The mGluR theory of fragile X mental retardation. *Trends Neurol.* **27**, 370–377 (2004).
47. Gregg, C., Zhang, J., Butler, J. E., Haig, D. & Dulac, C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**, 682–685 (2010).
48. DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA-seq: a new perspective. *PLoS Genet.* **8**, e1002600 (2012).
49. Barlow, D. P. Genomic imprinting: a mammalian epigenetic discovery model. *Ann. Rev. Genet.* **45**, 379–403 (2011).
50. Cantor, R. M. *et al.* Replication of autism linkage: fine-mapping peak at 17q21. *Am. J. Hum. Genet.* **76**, 1050–1056 (2005).
51. Chen, G. K., Kono, N., Geschwind, D. H. & Cantor, R. M. Quantitative trait locus analysis of nonverbal communication in autism spectrum disorder. *Mol. Psych* **11**, 214–220 (2006).
52. McCauley, J. L. *et al.* Genome-wide and ordered-subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates. *BMC Med. Genet.* **6**, 1 (2005).
53. Ylisaukko-oja, T. *et al.* Search for autism loci by combined analysis of Autism Genetic Resource Exchange and Finnish families. *Ann. Neurol.* **59**, 145–155 (2006).
54. Anney, R. *et al.* A genomewide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072–4082 (2010).
55. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
56. Weiss, L. A., Arking, D. E., Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802–808 (2009).
57. Murdoch, J. D. & State, M. W. Recent developments in the genetics of autism spectrum disorders. *Curr. Opin. Genet. Dev.* **23**, 310–315 (2013).
58. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
59. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
60. Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235–242 (2013).
61. Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum. Mol. Genet.* **19**, R176–R187 (2010).
62. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095–1102 (2011).
63. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525–537 (2012).
64. Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
65. Hancks, D. C. & Kazazian, H. H. Jr. Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* **22**, 191–203 (2012).
66. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
67. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *New Engl. J. Med.* **365**, 611–619 (2011).
68. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA* **105**, 4323–4328 (2008).
69. Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E. & Eppig, J. T. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39**, D842–D848 (2011).
70. Bayes, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature Neurosci.* **14**, 19–21 (2011).
71. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
72. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).

Competing interests statement
The authors declare no competing interests.