

Supplementary Note

Ascertainment of subjects.

We developed an epidemiological sample of autistic disorder taking advantage of the detailed birth and medical registries and universal access to health care. Cases with autistic disorder (Swedish ICD9-codes: 299A or ICD-10 codes: F84.0–F84.1) were identified from the Swedish National Patient Register. As of May 2011 there were 7,290 cases who were 10 years or older. Note that there are no major differences in the population and samples underlying the Swedish family study (described in the Main Text) and PAGES. Both access children born in Sweden and use the same case ascertainment scheme to identify autistic individuals. The family study includes all subjects born in Sweden between 1982 and 2007, the PAGES sampling frame includes these subjects and also allows recruitment of subjects born as early as 1960, although there are currently few of these older subjects in the sample. Methodologically the family study adjusted for differences in prevalence over time to calculate heritability over different birth cohorts, but this time component had little impact on relative recurrence risk. At the present time PAGES samples from most counties in Sweden, whereas the family study encompassed all of them. There is no reason to expect that this difference would bias the sources of risk explaining autism diagnosis.

To validate the registry-based diagnoses, we conducted a medical records review in 88 cases with autism diagnosis (representing about 1% of cases) randomly selected from all autism cases listed in the register. Medical records contain assessments from treating doctors including psychiatrists, psychologists and neurologists, as well as from nurses for inpatient and outpatient treatment. Medical record review substantiated the presence of DSM-IV autism in 94.3% of cases (83 of 88), and there was no indication of Autism Spectrum Disorder (ASD) in the medical records of the 5 cases that did not meet DSM-IV autism or ASD criteria. Note that 20% of these cases had an ADI and ADOS in their records, but the rate of validation for both groups (with and without ADI and ADOS) was indistinguishable.

We then identified subjects in the registry for recruitment, with exclusion and inclusion criteria as described in Online Methods. Five hundred thirty six (536) autistic disorder subjects were recruited from 12 counties in Sweden.

Genetic characterization.

All samples were genotyped on the Illumina HumanOmniExpressExome BeadChip. Here we analyzed only the OmniExpress content, which attempts to genotype more than 715,000 SNPs across the genome.

Quality control: Data from all subjects were processed through the PGC pipeline¹, which imputed autosomal SNPs with CEU (Utah residents of Northern and Western European ancestry) and TSI (Toscani in Italia) HapMap Phase 3 data as the reference panel. Only samples of Swedish descent, as judged by PCA analysis, were targeted for imputation. A SNP was kept only if it met the following criteria: had minor allele frequency $MAF > 0.01$; genotyping completion rate $> 95\%$; minor allele frequency > 0.05 ; Hardy-Weinberg equilibrium $p\text{-value} > 0.005$; imputation $R^2 > 0.6$. Of the original set of 711,027 autosomal SNP probes genotyped on the OmniExpress, 531,906 were retained for analysis. (We focus the analysis on the SNPs genotyped in the case sample, which ensures very high quality genotypes enter the analysis. In fact, because imputation can alter genotypes of SNPs that were “physically genotyped” by the chip, we checked to ensure the vast majority of imputed genotypes were identical to the physical genotypes, and they were; i.e., heritability should not be inflated or deflated by imputation error.)

Individuals were omitted if they were a genetic duplicate of another sample (n=11) or had a genotyping completion rate < 98% (n=1). This results in a final sample of 3046 individuals, of which 466 are cases and 2580 are controls.

Population structure: We further accounted for more subtle population structure by running an additional ancestry analysis on the data (n=3044, omitting one of each set of twins) with the GemTools software². A sample of 50,000 SNPs that had passed QC was used to determine seven significant dimensions of ancestry. These seven ancestry dimensions were then used as covariates in all subsequent analyses to control for the population structure.

Prevalence.

For the purpose of our heritability calculation we estimated autistic disorder prevalence in Sweden. Individuals were grouped into four-year birth cohorts for each of Sweden's 21 counties, starting with the 1980-1984 cohort and ending with the 2005-2009 cohort. An estimate of the prevalence within each county and birth cohort combination is obtained as a function of the age of the subjects. The estimate of prevalence is necessarily a monotone non-decreasing function of age because over time more people with the disorder are diagnosed, hence the prevalence increases as a function of age (**Supplementary Figure 1**). **Supplementary Figure 1** demonstrates that, when aggregating across all birth cohorts and counties, the population prevalence of 'strict autism' (i.e., autistic disorder) in Sweden, as a function of age, stabilizes after subjects reach 23 years of age, and converges on the value of .3%. Data are separated into birth cohorts to account for changes in prevalence over time. As demonstrated in **Supplementary Figure 1**, indeed, the county level prevalence estimates generally increase over time. Note that this figure excludes birth cohorts 2000-2004 and 2005-2009, as these groups had insufficient time to yield reliable estimates of prevalence. However cohorts spanning 1990-1999 show an average prevalence of .3%, hence we use this value of K in the current study when estimating the heritability of autistic disorder from our Swedish sample.

Heritability.

Early twin studies: There have been several twin studies in ASD (**Supplementary Table 1**). In every case, concordance for monozygotic twins was higher than that for dizygotic twins. In some of these studies, heritability was estimated. In three studies, heritability estimates were between 73-93%, while in 1, heritability was estimated at 38-39%.

Study	Diagnosis	Number of twins	MZ:DZ concordance estimates	Reported heritability estimates
Folstein and Rutter, 1977 ³	Kanner criteria	11 MZ, 10 DZ	Autism: MZ, 36%; DZ, 0% BAP: MZ, 82%; DZ, 10%	
Ritvo <i>et al</i>	DSM-III	23 MZ, 17 DZ	Autism: MZ,	

1985⁴			96%; DZ, 24%	
Steffenburg et al 1989⁵	DSM-III-R	11 MZ, 10 DZ, plus one set of identical triplets	Autism: MZ, 91%; DZ, 0% BAP*: MZ, 91%; DZ, 30%	
Bailey et al 1995⁶	ICD-10	25 MZ, 20 DZ, plus 2 triplet sets	Autism: MZ, 60%; DZ, 0% BAP: MZ, 92%, DZ, 10%	91-93%
Taniai et al 2008⁷	CARS	19 MZ, 26 DZ	ASD: MZ, 95%; DZ, 31%	73% for males and 87% for females
Rosenberg et al 2009⁸	IAN; diagnosis supplied by families	67 MZ, 210 DZ	ASD: MZ, 88%; DZ, 31%	
Lichtenstein et al 2010⁹	A-TAC	29 MZ, 88 DZ	ASD: MZ, 39%; DZ, 15%	80%
Hallmayer et al 2011¹⁰	ADI-R and ADOS	54 MZ, 138 DZ	Autism: MZ, 58-60%; DZ, 21-27% ASD: MZ, 50-77%; DZ, 31-36%	37% for autism and 38% for ASD

Supplementary Table 1 Twin studies in autism spectrum disorders. ADI-R, Autism Diagnostic Interview-Revised; ADOS, Autism Diagnostic Observation Schedule; A-TAC, the autism, tics, ADHD and other comorbidities inventory; BAP, Broader Autism Phenotype; CARS, Childhood Autism Rating Scale; DSM-III-R, Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised; DZ, dizygotic; IAN, Interactive Autism Network; ICD-10, 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD); MZ, monozygotic.

Heritability estimates from the Swedish family study: The Swedish family study¹¹ discussed in Main Text and Online Methods relied on an extended sibling design, an extension of the classical twin design, to estimate heritability¹². To increase power and to make results more generalizable, this design included full siblings and maternal and

paternal half siblings in addition to twins. As noted in Online Methods, the approach also facilitated estimation of additive and non-additive genetic sources of variance, as well as shared and non-shared environmental sources of variance. To choose a model, as a first step, all potential sources of variance were included. Next, parameters were excluded sequentially until the most parsimonious model that fitted the data was discovered. The chosen model incorporated covariance due to additive genetic and non-shared environment. Other terms in the full model did not have sufficient explanatory power to warrant retention in the final model.

Heritability estimated from GCTA: To estimate the genetic relationship matrix (GRM) we used common SNP markers only. To obtain estimates of the heritability, we follow the steps outlined in Lee *et al*¹³. In brief, we model the case-control status via the mixed linear model $y = X\beta + g + e$, where y is the vector of case-control status, β is the vector of coefficients for the fixed effects (7 ancestry dimensions returned by GemTools) with associated design matrix X , g is the vector of random additive genetic effects associated with SNPs, and e is a vector of random errors, which are assumed to be independent. The variance of the phenotypes can be written as $Var(y) = A\sigma_g^2 + I\sigma_e^2$, where A is the genetic relationship matrix (GRM) and I is an identity matrix, while σ_g^2 and σ_e^2 partition the total phenotypic variation into pieces attributable to additive genetic effects and random error, respectively. The heritability is then simply calculated as $h^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \sigma_e^2)}$ on the observed scale, which is transformed to the liability scale as a function of the population prevalence (K) as outlined previously¹³.

To estimate heritability due to common SNPs we use GCTA, applied to a GRM calculated based on a sample of essentially unrelated individuals ($A < .025$). To estimate total additive heritability we include all sampled individuals, compute the GRM, smooth this matrix using TCS (see below), and then compute heritability from the GCTA package.

Heritability by treelet covariance smoothing (TCS): In principle, given A , total heritability can be estimated from any sample that includes relatives. Thus one option is to estimate A based on the GRM from the full sample, which includes distant relatives; however this could produce a negatively biased estimate of heritability¹⁴. An alternative that works well is a thresholded version of the GRM, zeroing any entries in the GRM less than t (GRM_t)¹⁴. For estimating heritability of quantitative traits in the Icelandic population, Zaitlin *et al*¹⁴ used $t = .05$. For our sample from Sweden, however, such an approach is not feasible because there are too few close relatives (**Supplementary Figure 2**) and we did not perform this estimate.

Instead we use a more refined estimator that estimates degrees of relatedness and does a good job of thresholding unrelated individuals. Starting with GRM we applied a smoothing/dimension reduction technique known as treelets to obtain the treelet covariance smoothing matrix (TCS)¹⁵. Treelets cluster related individuals, even very distantly related individuals to smooth and denoise the GRM. The fundamental assumption is that A is a sparse matrix (mostly 0 entries) because most pairs of individuals are unrelated. If the sample has underlying structure in the form of small groups of related individuals, with little or no relatedness across groups, then transforming the data into a treelet representation and thresholding out noise to obtain a sparse covariance matrix will produce a better estimate¹⁵. To do so, treelets introduces novel construction (via a series of Jacobi rotations) of orthonormal multi-scale bases supported on nested clusters in a hierarchical tree. In particular, if \mathbf{z} is a random vector with $Var(\mathbf{z}) = A$, then the orthonormal basis $\{v_i\}_{i=1}^n$ can represent \mathbf{z} as $\mathbf{z} = \sum_{i=1}^n c_i v_i$,

where $c_i = \langle z, v_i \rangle$, and thus $A = \sum_{i=1}^n \gamma_{ii} v_i (v_i)^t + \sum_{i \neq j} \gamma_{ij} v_i (v_j)^t$, where $\gamma_{ij} = \text{Cov}(c_i, c_j)$. By applying a hard threshold to the γ_{ij} coefficients, some noisy relationships are filtered out producing a smoothed version of the GRM. Implementation requires first estimating A, then using an R function for treelets and then plugging this estimate into GCTA to obtain an estimate of h^2 . The treelet algorithm, as well as its implementation, is available in R on CRAN as the treelet library (<http://cran.r-project.org/web/packages/treelet/index.html>).

In published simulations of samples containing a mix of distant relatives and unrelated individuals, TCS estimated the genetic relationship matrix accurately. Indeed the bulk of the estimates were within one degree of the true relationship¹⁵. Moreover, TCS estimated genetic relationships with less error (in terms of root mean squared error) than GRM and GRM_t for relatives of degree 7 and greater; for closer relatives (degree < 7) GRM provided the best estimate while GRM_t and TCS produced slight under-estimates of the true relationships. For the current analysis only .03% of the relative pairs were estimated to be 6th degree or closer (**Supplementary Figure 2**), hence accurate estimation of the distant relative pairs was critical.

Carrying out these analyses based on the estimate of GRM produced by GCTA¹⁶, with people who are unrelated in the classical sense ($A < .025$), produces an estimate of heritability attributable only to common variants, in the sense of references^{6,7,13 16-18}. Including closer relatives, produces estimates of total additive heritability, including the variance attributable to both common and rare variation.

Consistent with the results in Zaitlen *et al*¹⁴, by using the complete GRM for A without any smoothing, the estimated heritability was 51.9%, lower than that using TCS (52.4%) but slightly higher than that from common variants (49.4%). We also attempted to estimate heritability using thresholded versions of the GRM, but the procedure failed to converge due to the small number of closely related subjects. To evaluate the stability of the TCS estimate of total additive heritability we perturbed entries in the TCS matrix based on a simple model for the sampling variability to evaluate the effect of estimation error and found that the results were very stable (standard deviation = 1.7%).

Simulations of pedigrees for heritability estimation: Simulations were conducted following the procedure used in Crossett *et al*¹⁵. To produce realistic simulations, we started with phased genomes (haplotypes) of individuals from the HapMap 3 database, selecting two populations with European ancestry (CEU and TSI). Utilizing the small sample of available haplotypes, we generated a large sample of haplotypes, representative of those that might be sampled from unrelated founders of a population. The challenge was to keep intact the realistic haplotype structure for a human population, including linkage disequilibrium (LD), without generating unusual sharing between the founders. To accomplish this goal, we took the HapMap data on CEU and Tuscan samples, which were phased quite accurately into haplotypes, as the initial sample of chromosomes from which to generate founders. Now each founder haplotype was created by sampling pieces of chromosomes (or haplotypes) from the initial sample. To do so, the number of recombination spots per chromosome was determined using an overall recombination of $\theta = 10^{-6}$ per Mb, which is 100 times the normal rate of recombination for humans. The actual location of the recombination spots were then determined using the recombination map provided by HapMap, a procedure that successfully keeps intact the LD structure of the chromosome. From this pool of generated haplotype pairs, chromosomes were randomly assigned to each of the founders in each of 100 families. These founder chromosomes were then dropped

through a five generation pedigree. At each generation the chromosomes underwent recombination with an overall rate of $\theta = 10^{-8}$ at locations determined by HapMap's recombination map. One hundred sets of independent pedigrees including 20 sampled individuals were combined to generate the full genotype sample of size 2000. For the given set of genotypes, 50 independent vectors of phenotypes were simulated. For each simulation a random set of causal variants were chosen: 1000 rare (MAF < .01) and 1000 common variants. These two classes of SNPs generated 25% and 50% of the heritability, respectively, for a total of $h^2 = 75\%$.

GRMs are generated using 300,000 markers with MAF > 0.05 and $r^2 < 0.64$. Data were then analyzed using the full GRM (2000 individuals) and the sub-GRM (n=1390) in which no relationship is > 0.025.

Impact of clinical features on estimates of heritability, exemplified by diagnosis and intellectual function.

Autistic disorder versus ASD and intellectual function in Sweden: Autistic disorder has a prevalence of 0.003 (= .3%) in the Swedish population. Using the same sampling frame we estimate a prevalence of 0.01 (=1%) for ASD in the population. Although recent CDC estimates show a somewhat higher rate in the United States¹⁹, both rates from Sweden are comparable to rates obtained from other studies, including the CDC.

By linking registry data from Sweden we were able to estimate the fraction of subjects with autistic disorder who also have intellectual disability or ID (IQ < 70). In Swedish subjects with autistic disorder, 43.57% also have ID. This rate is similar to other recent epidemiological studies. In a recent analysis from the Center for Disease Control Autism and Developmental Disabilities Monitoring Network¹⁹, seven sites had sufficient data on rates of ID. In this cohort 31% (range 18-37%) of children with the broader ASD diagnosis were classified as having ID (IQ \leq 70). The rate of comorbidity is expected to increase with more stringent levels of diagnosis. For example, in a UK study the rate of ID in subjects with autistic disorder, narrowly defined, was 73% whereas in broadly defined ASD it was 55%²⁰. In their "consensus autism" diagnosis, which by frequency seems closest to autistic disorder as defined here, 53% of subjects have ID. In a total population from Korea and when autistic disorder was analyzed separately from other ASD diagnoses, 48.1% of individuals with autistic disorder had ID²¹.

Heritability as assessed on data from the Autism Genome Project: Consistent with quantitative genetics theory it has already been shown that families who are multiplex for ASD carry a larger load of liability alleles, presumably common variants, relative to simplex families (defined as families with only one affected subject within the set of first and second degree relatives)¹⁷. Clinical phenotypes could also affect heritability/genetic load, although how much impact they might have is an open question. To evaluate this question we evaluate two phenotypes thought to have major impact on the genetics of ASD, namely diagnosis per se and high versus low functioning, as measured by IQ. For diagnosis we follow the AGP by analyzing strict autism, as defined by meeting criteria for autism on the ADI-R and ADOS²², versus ASD, which includes strict autism and subjects who meet looser criteria for a spectrum diagnosis. For IQ we again follow the AGP by targeting subjects with IQ \geq 80, well beyond the bound for intellectual disability.

For data we use the AGP sample of 2121 ASD probands from European descent and 1663 controls from European descent from the HABC collection. The AGP data are described in Anney *et al*^{23,24}, the HABC data are described in Klei¹⁷ and are the same data used therein. The number of probands for AGP samples was limited to one per family. There was an overlap of 837,454 markers between the two datasets. Conducting

a series of quality control steps for SNP, similar to those from Klei, SNPs are retained for analysis if they meet these criteria:

- 1) Non call rate per SNP < 0.02 (20,542 SNPs failed)
- 2) Minor allele frequency MAF > 0.01 (317 failed)
- 3) Exact Hardy Weinberg in the HABC samples 0.005 (5366 failed)

Samples are removed if they had a non-call rate for SNPs > 2.5%, and this removed 24 cases from the analysis. After QC there were 2097 AGP cases and 1663 HABC controls genotyped for 828,352 markers. After analysis using GCTA we obtain the following results:

Phenotype	N	Heritability (%)	
		Estimate	Standard error
ASD	2097	51.1	4.8
Strict autism	1242	52.3	6.2
IQ ≥ 80	871	59.3	7.8

For all analyses the prevalence for autism is set at 0.01 to make the estimates of heritability comparable. Note that clinical phenotypes do not have a substantial impact on heritability. This conclusion is consistent with that reached by analysis of a larger spectrum of clinical phenotypes in the Simons Simplex Collection (Chaste et al., unpublished data).

Meta analysis of heritability.

This is described fully in Online Methods.

Estimating the contribution of *de novo* mutations and heritable variation to liability and variation in liability to ASD.

Motivation: Both *de novo* copy number variants (CNVs) and *de novo* loss of function (LoF) mutations have been associated with ASD by multiple independent researchers²⁵⁻³¹. To estimate the contribution of these *de novo* mutations to the overall risk of ASD, we assume ASD risk is a direct function of the excess proportion of cases carrying *de novo* mutations, relative to control subjects. In other words because control subjects do not have an ASD diagnosis and because *de novo* mutations mediating ASD risk are rare events, a reasonable assumption is that *de novo* mutations found in controls do not mediate ASD risk. Rather, these events are incidental to an ASD diagnosis and provide a good approximation to the background rate of mutation in the general population.

In subjects with ASD (cases), on the other hand, we expect a mix of *de novo* mutations that do, and that do not, impart ASD risk. The *de novo* mutations that do not impart ASD risk should be present at roughly the same rate in cases as in controls because they are an incidental finding, unrelated to ASD, so should not be affected by the ascertainment of ASD samples. The remaining *de novo* mutations that impart ASD risk in cases will be present as a consequence of ascertainment bias for ASD. We will estimate this differential rate by the odds ratio, which, in this case, is an excellent approximation to the relative risk (RR), and then use the estimated RR to estimate the liability accounted for by *de novo* mutations, in the same manner as heritability on the liability scale (see Falconer³² for development).

While using unobservable liability to model the heritability of the binary trait has a long history, the challenge of this parameterization is its interpretation. Take, for example, a disease with prevalence $K = 0.01$ (like ASD), a risk allele with a population frequency (p) of 0.001 for allele d , versus allele D occurring at a population frequency of $q=1-p$. For genotypes Dd and DD , the genotype relative risk (G) can be defined as the ratio of two penetrance probabilities (f_0 and f_1) so that $G = \frac{P(\text{disease}|Dd)}{P(\text{disease}|DD)} = \frac{f_0}{f_1}$. Let $G = 20$. Then, using equations developed below we can estimate the contribution of Dd towards disease heritability (h^2) as 0.4%. Still the penetrance for genotype Dd is 0.193 so an individual carrying the d variant is at substantial risk for the disease and 19.3% of subjects carrying this variant will have a disease diagnosis. Furthermore the variant d will be present in 1.9% of individuals with the disease, the majority of whom would not have the disease in the absence of the d variant (>99%, see calculations below). Yet, because d is rare and K is small, the locus explains a small fraction of the variance in liability at the population level: there is the gist of it, the mutation substantially affects liability when present, but at the population level it is found in only 2 in a thousand subjects and therefore cannot, by itself, explain much of the population's variability in liability.

Calculation of liability: Risch³³, in his analysis of heritability for a single risk locus, lays out an equation relating key quantities as a quasi-additive model:

$$K = q^2 f_0 + 2pqf_1 + p^2 f_2 = q^2 f_0 + 2pq(f_0 + a) + p^2(f_0 + 2a) = f_0 + 2pa$$

Other quantities flow from this simple parameterization. Given K , p and G , $f_0 = K/(1 + 2p(G - 1))$ and $a = (G - 1)f_0$. Because *de novo* events observed in ASD subjects appear to be acting dominantly, i.e., they generate a single loss-of-function in one (for sequence) or more (some CNVs) haplo-insufficient genes, we make a slight adjustment to the formulas, setting penetrance $f_1 = Gf_0$ for Dd and dd ; and $f_0 = K/(q^2 + 2pqG + p^2G)$.

Our interest here is to relate *de novo* variation – generation of allele 'd' in our example above – to heritable variation by setting them on the same liability scale. This we will do by using the model described in Falconer³².

De novo CNVs: Fifty-four *de novo* CNVs were observed in 872 ASD cases from the Simons Simplex Collection (SSC) compared with 16 in 872 matched sibling controls²⁷. A similar result was obtained from a complementary analysis, which used the same sample but with a different CNV detection protocol, with 75 *de novo* CNVs in 858 probands and 19 *de novo* CNVs in 863 sibling controls²⁶. Using the latter, more comprehensive estimates, we compute the relative risk $RR=4.25$. Assuming an 'exposure rate' of 0.022, the rate of *de novo* CNVs in unaffected siblings of ASD probands, variability on the liability scale accounted for by *de novo* CNVs is 1.46%.

De novo LoF mutations: Seventy-two of 599 ASD probands from the SSC have a *de novo* LoF mutation, compared with 32 of the 599 matched sibling controls^{28,31,34}. These observations lead to an estimate of $RR = 2.42$. Assuming an exposure rate of 0.053, variability on the liability scale accounted for by *de novo* LoF mutations is 1.11%.

De novo missense mutations: Reports of an association between *de novo* missense mutations and ASD differ between publications^{28,31} and the combined data show only a trend towards enrichment in probands ($p=0.10$)³⁴; however, because a proportion of missense mutations are known to have a highly deleterious effect on protein function, and because of the reproducible association between *de novo* LoF mutations and ASD, it is highly likely that a subset of *de novo* missense mutations do mediate ASD risk. In the combined dataset, 253 out of 599 probands had at least one *de novo* missense

mutation compared with 238 of the 599 matched sibling controls. These observations lead to an estimate of RR = 1.11. Assuming an exposure rate of 0.397, variability on the liability scale accounted for by *de novo* missense mutations is negligible (0.04%).

Risk from *de novo* mutations and generalizability: By assuming *de novo* mutations act independently, a reasonable estimate of the contribution to the variance in liability is 2.6%. Relative to heritable variation, it is roughly 5% (i.e., .026/.5). On the other hand, these *de novo* events are often meaningful for liability when they occur. In our simple example ($p=.001$ and $G=20$ for the additive model) we asserted two facts: 1.9% of individuals with the disease would carry the variant and 99% would not have the disease in the absence of the risk variant. Empirically, relative to sibling controls, we find an excess of 13% of *de novo* copy number or *de novo* LoF mutations in ASD subjects who participated in the Simons Simplex Collection study. The subjects are largely idiopathic, as opposed to syndromic (see below for discussion of syndromic subjects and **Supplementary Table 1**), which will reduce the rate of *de novo* mutations per subject. Therefore *de novo* events are inarguably an important source of risk. For *de novo* CNV and LoF events we can also estimate what fraction of subjects would not be affected if they did not carry the *de novo* mutation. For *de novo* CNVs, roughly 80% of subjects would not be affected and for *de novo* LoF mutations, 57%.

Further details of calculations: The low contribution of *de novo* events to variability in liability could seem counterintuitive. There is a way to verify these estimates are sensible, by determining if the observed rate of *de novo* CNVs in ASD subjects for each class of variants is captured by the estimated parameters of the model. To do so we estimate $P(\text{mutation}|\text{ASD})$ by the conditional probability model:

$$\frac{P(\text{ASD}|\text{mutation})P(\text{mutation})}{P(\text{ASD}|\text{mutation})P(\text{mutation}) + P(\text{ASD}|\text{no mutation})(1 - P(\text{mutation}))}$$

= $pf_1/(pf_1 + (1 - p)f_0)$. The rate calculated for each class of mutations matches perfectly with the observed rate of *de novo* events in probands.

To compute the probability of carrying a risk mutation, one can use this conditional probability argument: $P(Dd|\text{disease}) = \frac{P(Dd)P(\text{disease}|Dd)}{P(\text{disease})}$. To compute the probability that someone would not be affected were they not a carrier, we first need the impact of the mutation on liability measured from f_1 and K . Call it 'd' for displacement. Then, using the normal distribution describing the distribution of liability in the population, calculate $1 - \frac{1-\Phi(d)}{1-\Phi(K)}$, where Φ is the normal distribution function.

Caveats: We note that *de novo* missense mutations are observed in over 40% of cases and have only a modest impact on risk, 1.11. It is possible that we over-value exposure for this class of *de novo* variants, but they account for almost nothing of the variability of liability. We assume *de novo* events are independent sources of risk, which comports with the data. We simplify the modeling of *de novo* missense variants by ignoring their multiplicity in some individuals, but these missense variants, as a class, have little impact in any case.

We also note that all of the estimates of *de novo* impact are derived from the Simons Simplex Collection, a cohort composed solely of families with sporadic ASD that was designed to enrich for *de novo* mutations. Therefore the estimate of the contribution of *de novo* mutations to explanation of ASD liability is likely to be inflated somewhat compared with the general population; however the degree of inflation is likely to be small because the majority of ASD families are simplex³⁵ and other studies have shown

that the rates of *de novo* CNVs and SNVs are similar when simplex and multiplex families are contrasted^{22,29}. Conversely, it is likely that technological advances will demonstrate ASD association for categories of *de novo* mutation that are missed by current studies, including small CNVs and non-coding mutations, leading to an underestimate of the total contribution from *de novo* mutations.

Syndromic variants: A proportion of ASD is described as being syndromic, with ASD being one of a constellation of features, and a substantial proportion of these syndromic causes are also *de novo*. Comorbidity of ASD and Fragile X syndrome accounts for about 1% of ASD diagnoses³⁶, making it the most common single-gene syndrome associated with ASD. While it is often cited as an inherited “cause of ASD”, the CGG triplet repeat frequently expands from a pre-mutation in the mother so that the full mutation in the child is ‘*de novo*’³⁷ and whether the pre-mutation is sufficient in itself to cause ASD is debatable. Considering four syndromes in which ASD comorbidity is frequently observed (after Fragile X), over 60% of cases arise from *de novo* mutations in autosomal genes and 31% for X-linked Duchenne muscular dystrophy (**Supplementary Table 2**). Similarly for CNV syndromes associated with ASD *de novo* mutations frequently account for over 50% of the observed variants (**Supplementary Table 3**).

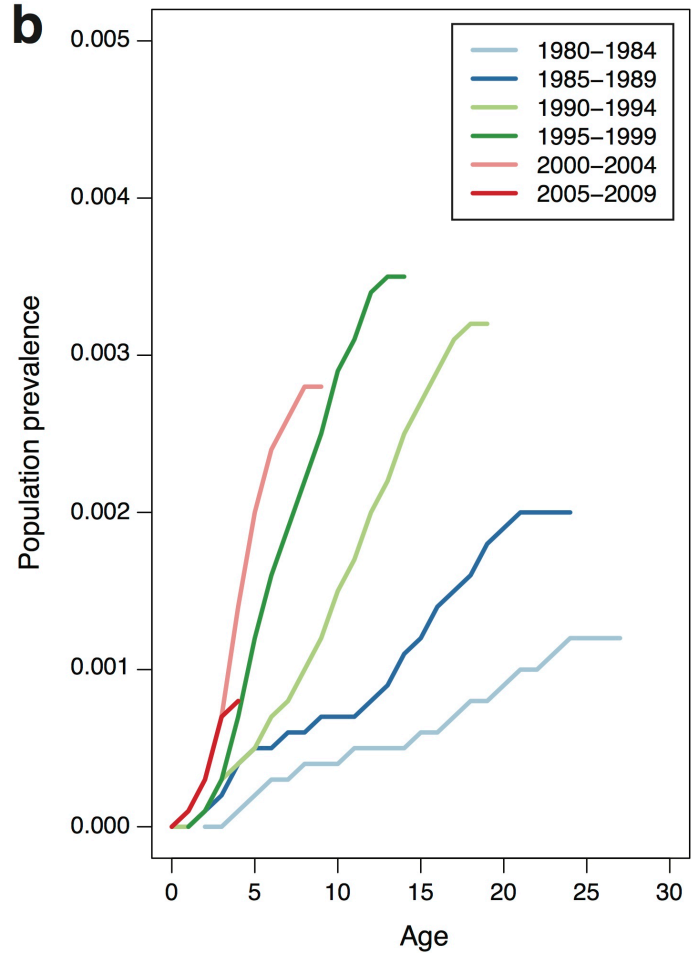
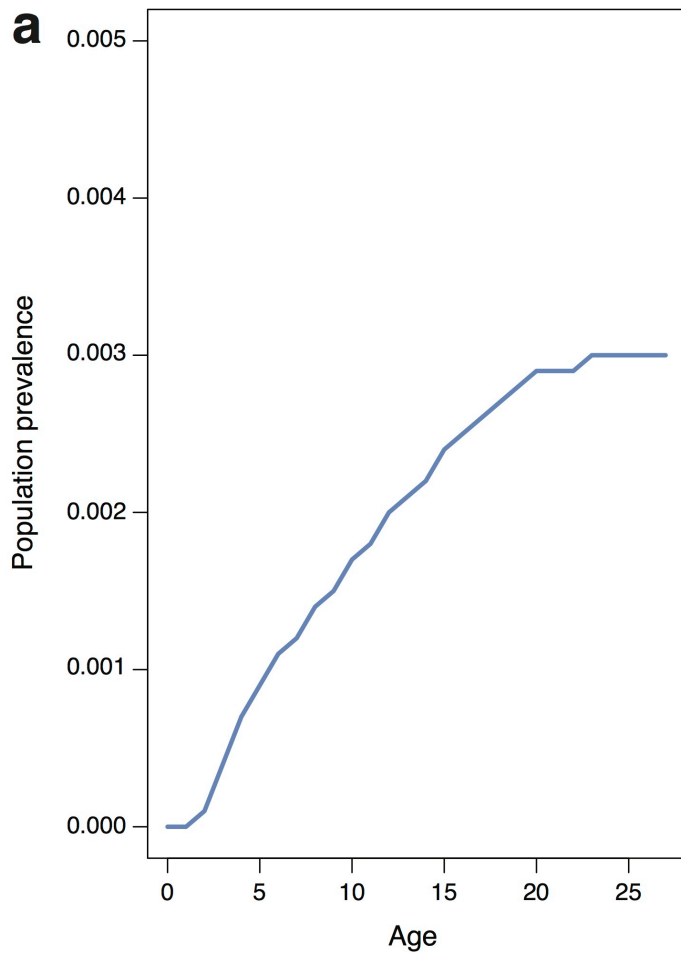
Syndrome	Gene	Syndrome prevalence	Percentage <i>de novo</i>	ASD prevalence in syndrome
Duchenne muscular dystrophy	DMD	0.029	31%	0.04
Neurofibromatosis, type 1	NF1	0.02	60%	0.04
Tuberous sclerosis	TSC2	0.003	64%	0.48
Tuberous sclerosis	TSC1	0.001	64%	0.48

Supplementary Table 2 Single gene syndromes in which comorbidity with ASD is relatively common.

Syndrome	Type	Percentage <i>de novo</i>
16p11.2	Del	92%
16p11.2	Dup	50%
15q11.2-q13 (BP2-BP3)	Dup	50%
1q21.1	Dup	17%
22q11.2	Dup	33%
7q11.23	Dup	33%
1q21.1	Del	100%
3q29	Del	50%
22q11.2	Del	100%

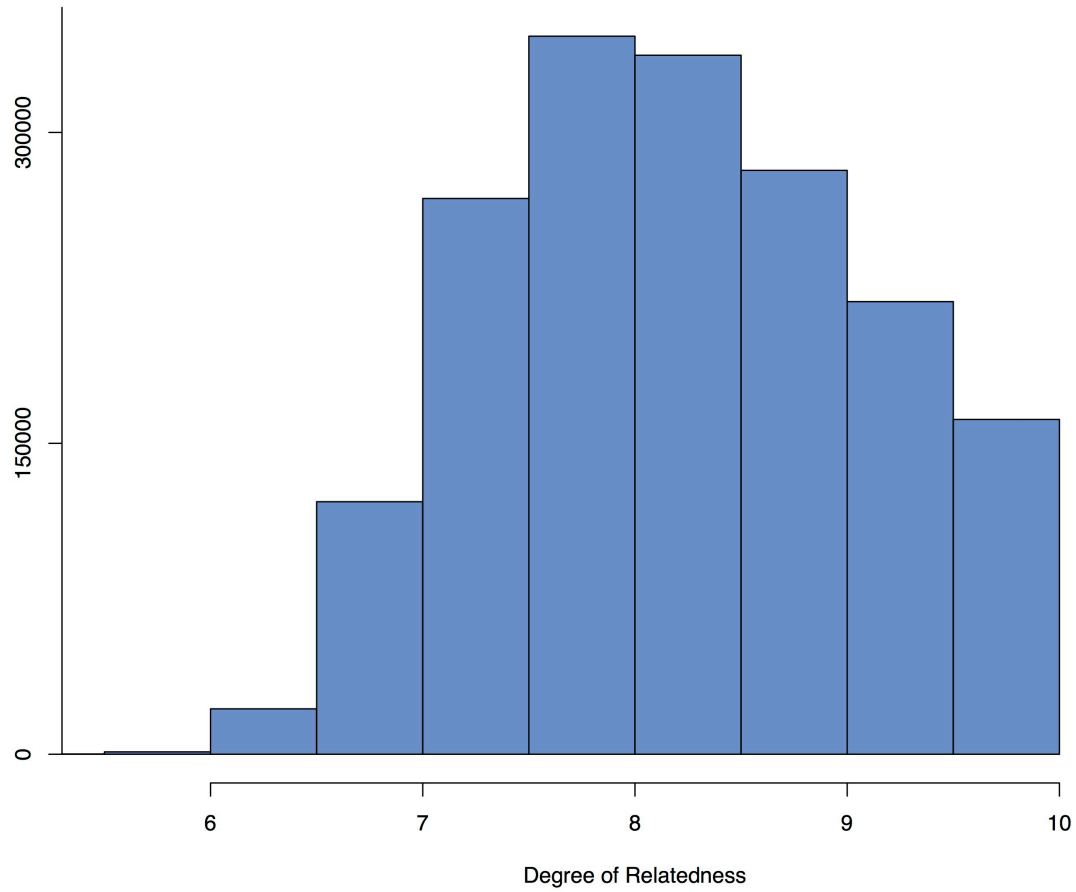
Supplementary Table 3 CNV syndromes in which comorbidity with ASD is relatively common. The percentage of CNVs that are *de novo* and the overall contribution of each CNV to PAR for ASD is shown for the CNVs showing the strongest association with ASD. All data are based on Moreno de Luca *et al*³⁸.

While this list of syndromes is not exhaustive, it does include syndromes frequently observed in ASD cohorts. The contribution from other syndromes to overall ASD liability is small because such syndromes are either very rare, such as Timothy syndrome³⁹, or they have very weak association with ASD, such as Sotos syndrome⁴⁰. Because the syndromes are relatively rare or ASD is rarely comorbid with the syndrome, we do not include these in our calculation for the contribution of *de novo* variation to liability for ASD.



Supplementary Figure 1 Estimated prevalence of strict autism for the entire Swedish population as a function of individuals' ages.

Prevalence was calculated by Kaplan-Meier function utilizing year of birth, age at autistic disorder diagnosis and censoring due to death, emigration or end of study period, whichever came first. (a) Prevalence pooled across cohorts and counties; (b) Prevalence by cohort.



Supplementary Figure 2 Distribution of relative pairs by degree of relatedness estimated using TCS.

Based on 3,044 subjects, there are 4,631,446 subject pairs, and 63% of the pairs are estimated to be unrelated (degree > 10). 1,346 of these were estimated to be degree 6 or closer with 12, 37, 104 and 1,193 of these falling in the intervals 2-3, 3-4, 4-5 and 5-6.

Supplemental References

1. Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**, 969-76 (2011).
2. Lee, A.B., Luca, D., Klei, L., Devlin, B. & Roeder, K. Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol* **34**, 51-9 (2010).
3. Folstein, S. & Rutter, M. Infantile autism: a genetic study of 21 twin pairs. *Journal of child psychology and psychiatry, and allied disciplines* **18**, 297-321 (1977).
4. Ritvo, E.R., Freeman, B.J., Mason-Brothers, A., Mo, A. & Ritvo, A.M. Concordance for the syndrome of autism in 40 pairs of afflicted twins. *The American journal of psychiatry* **142**, 74-7 (1985).
5. Steffenburg, S. *et al.* A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *Journal of child psychology and psychiatry, and allied disciplines* **30**, 405-16 (1989).
6. Bailey, A. *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological medicine* **25**, 63-77 (1995).
7. Tani, H., Nishiyama, T., Miyachi, T., Imaeda, M. & Sumi, S. Genetic influences on the broad spectrum of autism: study of proband-ascertained twins. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **147B**, 844-9 (2008).
8. Rosenberg, R.E. *et al.* Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of pediatrics & adolescent medicine* **163**, 907-14 (2009).
9. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *The American journal of psychiatry* **167**, 1357-63 (2010).
10. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry* **68**, 1095-102 (2011).
11. Sandin, S. *et al.* The familial risk of autism. *J Am Med Assoc*, in press. (2014).
12. Neale, M.C., Cardon, L.R. & North Atlantic Treaty Organization. Scientific Affairs Division. *Methodology for genetic studies of twins and families*, xxv, 496 p. (Kluwer Academic Publishers, Dordrecht ; Boston, 1992).
13. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
14. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
15. Crossett, A., Lee, A.B., Klei, L., Devlin, B. & Roeder, K. Refining genetically inferred relationships using treelet covariance smoothing. *Annals of Applied Statistics* **7**, 669-690 (2013).
16. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
17. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism* **3**, 9 (2012).
18. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).
19. Developmental Disabilities Monitoring Network Surveillance Year Principal, I. Prevalence of autism spectrum disorder among children aged 8 years - autism

- and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill Summ* **63 Suppl 2**, 1-21 (2014).
20. Baird, G. *et al.* Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *Lancet* **368**, 210-5 (2006).
 21. Kim, Y.S. *et al.* Prevalence of autism spectrum disorders in a total population sample. *Am J Psychiatry* **168**, 904-12 (2011).
 22. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
 23. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet* **21**, 4781-92 (2012).
 24. Anney, R. *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* **19**, 4072-82 (2010).
 25. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).
 26. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-97 (2011).
 27. 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**, 863-85 (2011).
 28. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
 29. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
 30. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
 31. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
 32. Falconer, D.S. *Introduction to quantitative genetics*, viii, 340 p. (Longman, London ; New York, 1981).
 33. Risch, N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol Biomarkers Prev* **10**, 733-41 (2001).
 34. Willsey, A.J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).
 35. Constantino, J.N., Zhang, Y., Frazier, T., Abbacchi, A.M. & Law, P. Sibling recurrence and the genetic epidemiology of autism. *Am J Psychiatry* **167**, 1349-56 (2010).
 36. Reddy, K.S. Cytogenetic abnormalities and fragile-X syndrome in Autism Spectrum Disorder. *BMC Med Genet* **6**, 3 (2005).
 37. Nolin, S.L. *et al.* Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. *Am J Hum Genet* **72**, 454-64 (2003).
 38. Moreno-De-Luca, D. *et al.* Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol Psychiatry* **18**, 1090-5 (2013).
 39. Zafeiriou, D.I., Ververi, A., Dafoulis, V., Kalyva, E. & Vargiami, E. Autism spectrum disorders: the quest for genetic syndromes. *Am J Med Genet B Neuropsychiatr Genet* **162B**, 327-66 (2013).
 40. Buxbaum, J.D. *et al.* Mutation analysis of the NSD1 gene in patients with autism spectrum disorders and macrocephaly. *BMC Med Genet* **8**, 68 (2007).