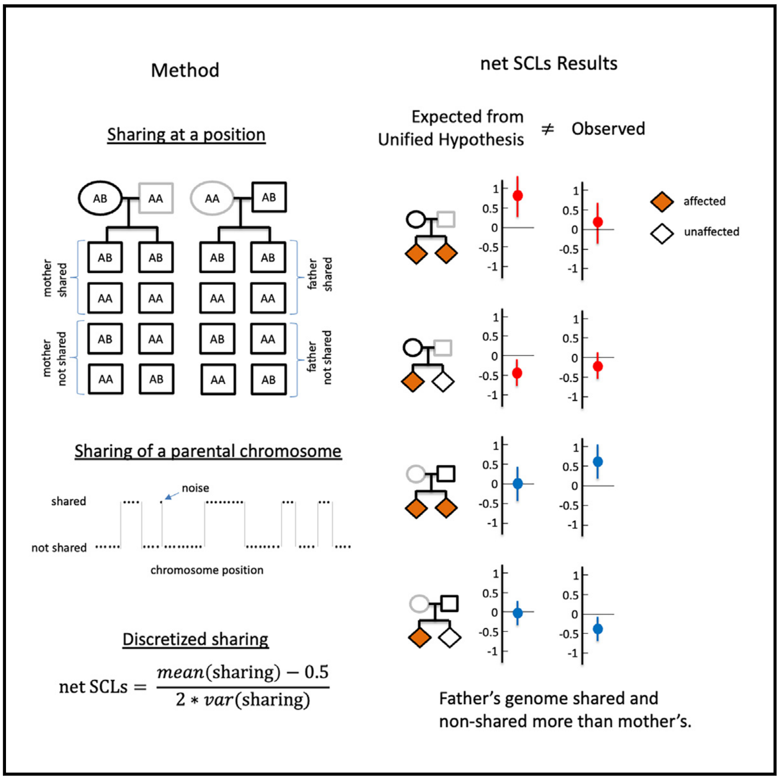
Cell Genomics

Article

Sharing parental genomes by siblings concordant or discordant for autism

Graphical abstract



Highlights

- Methods for discretized sharing of parental genomes
- Siblings with autism share parental genomes more than expected
- The excess sharing of the father's genome is significant; the mother's sharing is not
- Exploration of genetic models of populations of families with autism

Authors

Mathew Wroten, Seungtai Yoon, Peter Andrews, ..., Kenny Ye, Michael Wigler, Ivan Iossifov

Correspondence

wigler@cshl.edu (M.W.), iossifov@cshl.edu (I.I.)

In brief

Wroten et al. find that siblings concordant for autism share more parental genomes than expected; discordant siblings share less. Furthermore, the excess sharing of the father's genome is statistically significant, while the mother's sharing is not. These observations contradict certain models in which the mother carries a greater genetic predisposition.

Wroten et al., 2023, Cell Genomics *3*, 100319
June 14, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.xgen.2023.100319

Cell Genomics

CellPress
OPEN ACCESS

Article

Sharing parental genomes by siblings concordant or discordant for autism

Mathew Wroten,[1,6] Seungtai Yoon,[1,6] Peter Andrews,[1] Boris Yamrom,[1] Michael Ronemus,[1] Andreas Buja,[2,3] Abba M. Krieger,[2] Dan Levy,[1] Kenny Ye,[4] Michael Wigler,[1,5,*] and Ivan Iossifov[1,5,7,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Department of Statistics and Data Science, the Wharton School, University of Pennsylvania, Philadelphia, PA, USA
[3]Flatiron Institute, Simons Foundation, New York, NY, USA
[4]Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA
[5]New York Genome Center, New York, NY, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: wigler@cshl.edu (M.W.), iossifov@cshl.edu (I.I.)

https://doi.org/10.1016/j.xgen.2023.100319

SUMMARY

Studying thousands of families, we find siblings concordant for autism share more of their parental genomes than expected by chance, and discordant siblings share less, consistent with a role of transmission in autism incidence. The excess sharing of the father is highly significant (p value of 0.0014), with less significance for the mother (p value of 0.31). To compare parental sharing, we adjust for differences in meiotic recombination to obtain a p value of 0.15 that they are shared equally. These observations are contrary to certain models in which the mother carries a greater load than the father. Nevertheless, we present models in which greater sharing of the father is observed even though the mother carries a greater load. More generally, our observations of sharing establish quantitative constraints that any complete genetic model of autism must satisfy, and our methods may be applicable to other complex disorders.

INTRODUCTION

The genetic landscape of autism is the best understood of all the known complex cognitive-behavioral disorders.[1] Autism is at least partially genetically determined, as shown by the concordance of dizygotic twins,[2,3] involvement of known causal loci,[4] increased incidence of deleterious *de novo* mutation in affected individuals,[5–17] and biased transmission of rare and common variants to affecteds.[18–23] However, we do not know if *de novo* mutation and transmission of risk variants provide a sufficient explanation of incidence. What we might be missing from a complete genetic explanation could be extra-genetic factors. These might be relevant to other developmental and cognitive disorders.

It is sensible to ask whether any genetic model can fit all the known quantitative observations. These include the overall autism incidence, the excess incidence in males, the family risk rates after one and after two previous affected offspring, the twin concordance rates, the contribution from transmitted variants, and the absence of common loci of strong effect. Additionally, a complete genetic model would include introducing and eliminating new deleterious variants and stabilizing common risk variant frequencies. We call this the “stationary” condition.

In 2008, we proposed such a model,[24] which we called the “unified hypothesis” because it incorporated both *de novo* mutation and transmission genetics, satisfied the stationary condition, and fitted the known observations very well. In the unified hypothesis, we proposed that *de novo* mutation accounts for much of simplex autism and that, due to the resistance of females to diagnosis following a genetic insult, many deleterious mutations could be carried by females and transmitted to the next generation. This could also explain the high risk of autism to the third-born sibling of a multiplex family, a rate of nearly 50%,[24] confirmed by others.[25] In the strong form of this model, most risk would arise from recent mutation. However, in a less stringent form, additional risk is also carried by common variation, making certain individuals more vulnerable. Thus the model would easily accommodate the observations ascribed to polygenic risk.[18–20] In fact, some parents of simplex families, most often the mother, transmitted deleterious mutations preferentially to affected children.[21–23] However, these parents were not seen with the frequency needed to explain the incidence of multiplex families.

We decided to put this model to a rigorous test. We asked whose genome is shared more in concordant siblings, mother or father. We reasoned that, if the mothers were typically the carriers of strong risk variants, then concordant siblings would share the haplotype flanking her variant. This might be observable as increased genomic sharing, more so than in the father. We judged that the size of the multiplex collections was just sufficient to answer this question. We used the even larger collections of simplex families to ask the companion question about decreased genomic sharing in discordant siblings.

To conduct the analysis, we needed to work out some technical details and theory. To measure haplotype transmission, we needed to combine SNP array data and whole-genome sequencing (WGS). To adjust for the different rates of meiotic recombination in mothers and fathers, we needed to develop a method we call "discrete sharing," which adjusts genomic sharing. We present here the observations that a complete genetic model for autism must attempt to fit: the total discrete sharing in concordant and discordant siblings, partitioned into maternal and paternal components. We provide tools that may be used to test genetic models and examples of models that do or do not fit aspects of the observations. Our methods are general and applicable to modeling other complex disorders.

RESULTS

Experimental design

We examined genome data from three available family collections: the Simons Simplex Collection (SSC), about 2,000 families; multiplex from the Autism Genetic Repository Exchange (AGRE), about 800 families; and the SPARK collection, which included both simplex and multiplex families, about 2,500 and 500 each respectively. The data were high-coverage WGS from the SSC[17,26] and AGRE,[17,23] and SNP microarrays for the SPARK collection.[27,28] The set of children from these cohorts is described in Table S1. To make the genomic data compatible for all cohorts, we selected the positions of SNP arrays that were well represented in WGS. We filtered out positions that were poorly genotyped, violated Hardy-Weinberg equilibrium assumptions, or exhibited wildly distorted allele ratios in populations, indicative of abnormal copy number or mis-mapping (see STAR Methods). This resulted in a common set of ~350,000 positions (Table S2). This level of resolution is more than adequate to resolve the meiotic blocks. The genotypes for these positions for all individuals are deposited at SFARI Base (see Data and code availability).

From each family, we chose one pair of concordant siblings from the multiplex families and one discordant pair from the simplex. When there were more than two affected children, we chose one sibling pair at random. We determined sibling sharing at all SNP positions where one parent was heterozygous, and the other parent was homozygous. For example, if the father is Aa and mother is AA, then we can infer the "polarity" of the share for the father at this position: "sharing," if both sibs are Aa or both are AA, and "non-sharing" otherwise. We extended this polarity determination to blocks of consistent sharing or non-sharing, spanning all consecutive informative positions for a given parent. We then computed the length of sharing and non-sharing blocks over the entire genome.

The expectation from our present knowledge of meiotic recombination is that sharing and non-sharing would occur in long chromosomal blocks.[29] Overwhelmingly, that is what we observe (Figure S1, panels A1 and A3). However, in both WGS and array data types, we do occasionally see single-position "spikes" in which the polarity of sharing flips, interrupting an otherwise long block of sharing or non-sharing. These spikes are far more common in the array data than in WGS (compare panels A1 for WGS and A3 for SNP arrays in Figure S1). Based on an analysis of the spikes, namely their location, recurrence, and skewed allelic ratio, we decided to scrub the data to eliminate them, as shown in Figure S1, panels A2 and A4 (see STAR Methods). The distributions of numbers of chromosomal switches (share to non-share) in WGS and array data, before and after scrubbing, are shown (Figure S1, panel B); after scrubbing, the data from WGS and array are statistically indistinguishable. The distributions of the number of switches in fathers and mothers are statistically very different, as expected due to increased meiotic recombination in mothers.[29,30]

Even after scrubbing, we occasionally observed anomalous block sharing patterns: many short meiotic blocks over large chromosomal regions (Figure S1, panel C). We filtered any family (N = 67) from our analysis in which a chromosome exceeded 15 switches for mothers or 10 for fathers. On analysis, we found we could explain the patterns in each anomalous case. Sometimes there was a failure to transmit a chromosome or a sizable portion of a chromosome, or transmission of an extra chromosome or a large part of one. These events were then divisible into the imbalanced or balanced progeny, such as that caused in the latter by uniparental disomy.

After data scrubbing and family filtering, we calculated sharing over the autosomes of 4,456 discordant pairs (1,921 from SSC and 2,535 from SPARK simplex families) and 1,269 concordant pairs (766 from AGRE and 503 from SPARK multiplex families). We identified the shared or non-shared intervals for each parent and chromosome and list these intervals in Table S3. We also calculated the total length of shared or non-shared blocks in each parent for each chromosome and the total genome (Table S4). We made a similar table that was not defined by length but rather by the number of SNP positions shared (Table S5). We define the proportion of a parent shared as the sum of the lengths of shared intervals divided by the sum of the lengths of both shared and non-shared intervals of that parent. In Table 1, the mean proportion of sharing is presented by parent and by concordance over all cohorts. We calculated similarly over each individual cohort and also by the genders of the sibling pair (Table S6). We again calculated proportion based on the sharing or non-sharing of numbers of SNP positions rather than length (Table S7).

The magnitude and statistical significance of genomic sharing

We initially consider the statistical evidence for the role of transmission in autism by comparing the sharing in concordant with discordant siblings from each parent. We obtained distributions of sharing of maternal and paternal autosomes for concordant affected siblings from combined multiplex families and did likewise for discordant siblings from combined simplex families. The means and variances of each are shown in Table 1. The difference between the mean sharing of the paternal genome for concordant and discordant siblings is 0.0090 (= 0.5059–0.4969), larger in magnitude but of the same polarity as the difference for the mothers of 0.0025 (= 0.5013–0.4988). The likelihood that differences from a random sampling of the fathers could equal or exceed the observed difference can be obtained by permuting the discordant-concordant labels. This likelihood is 0.000019. The likelihood for the mothers is 0.15. Similar observations have already been reported in Risch et al.[31]

Please cite this article in press as: Wroten et al., Sharing parental genomes by siblings concordant or discordant for autism, Cell Genomics (2023), https://doi.org/10.1016/j.xgen.2023.100319

Table 1. Parental genome sharing measures

Sibling group	Number of pairs	Mean sharing	Excess (mean-null)	Variance	SEM	t test two-sided p value	Permutation two-sided p value	Net SCLs	95% CI for the net SCLs
Maternal sharing									
Discordant	4,456	0.4988	−0.0012	0.00223	0.0007	0.097	0.097	−0.26	(−0.57, 0.05)
Concordant	1,269	0.5013	0.0013	0.00222	0.0013	0.31	0.31	0.30	(−0.28, 0.87)
Paternal sharing									
Discordant	4,456	0.4969	−0.0031	0.00442	0.0010	0.0019	0.0022	−0.35	(−0.56, −0.13)
Concordant	1,269	0.5059	0.0059	0.00437	0.0019	0.0014	0.0014	0.66	(0.26, 1.07)

We show measures of maternal and paternal sharing in the concordant (from AGRE and SPARK) and discordant (from the SSC and SPARK) sibling pairs, combining the WGS and microarray data types. The first section of the table shows properties of the observed sharing distribution across the sibling pairs: the number of pairs, the mean sharing, the excess (defined as the difference of the mean and the theoretical null expectation of 0.5); and the variance in sharing. The second part of the table shows properties of the observed means. The SEM reflects the confidence in the mean sharing estimate. The table also shows the p values for two statistical tests that compare the observed mean with the theoretical null. The first test is a two-sided t test, and the second test is a non-parametric permutation test (see "results" and STAR Methods). The final section of the table shows our estimates of the net SCLs measure (and the 95% confidence of the net SCLs), which uses the excess in sharing and accounts for the differences in meiotic recombination between the mothers and fathers.

To consider the evidence for a role for transmission for separate cohorts, we initially assumed a null model of 0.5 as the expected sharing between unascertained siblings and present the mean excess sharing in Table 1. We observe an increased sharing of both parents in the siblings concordant for autism and decreased sharing of both in discordant siblings. We determined p values two ways: first, we used a t test to compare the observed mean sharing with the null expectation of 0.5, and, second, we ran 1 million simulations permuting the underlying chromosomal sharing data by randomly flipping the polarity of sharing (see STAR Methods). The two methods yielded virtually the same p values. If we use SNPs rather than length as the measure of sharing, we also obtain nearly identical values (Table S7). When we performed our analysis on individual cohorts or siblings further partitioned by gender status, we obtained values that are largely compatible with our aggregate analysis (Table S6; Figure S2).

The increase of paternal sharing above 0.5 among concordant siblings (two-sided p value = 0.0019) and the decrease of sharing below 0.5 among discordant siblings (two-sided p value = 0.0014) are highly significant. Although sharing of the maternal genome tends in the same direction as the paternal values, the magnitude is less, and the differences from the null expectation have little significance. We obtain a two-sided p value of 0.097 for discordant and 0.31 for concordant siblings (Table 1).

Although measures of parental sharing in the unascertained human population were compatible with sharing of 0.5,[32] the assumption of 0.5 sharing as the proper null might not be correct. Therefore, given this uncertainty, we consider a range of possible null expectations. The statistical significances of the observed sharing proportions are shown in Figure 1 for a range of possible nulls. We consider various null hypotheses, from 0.496 to 0.508, the range of observed sharing of the simplex to the multiplex populations. With a range of null expectations from 0.4987 (arrow D_p) to 0.5019 (arrow C_p), spanning the theoretical null, the p values of the sharing proportions of the father for discordant or concordant siblings are each below 0.05. There is no reasonable null at which the sharing of the mother in the multiplex populations achieves statistical significance. On the other hand, the sharing proportion of the mother in simplex populations becomes significant at a null of 0.5002 (arrow Dm), above the theoretical null.

The magnitude and statistical significance of discretized sharing

While the deviation of sharing from the null appears to be greater for the fathers than for the mothers, we must temper this observation with the knowledge that the mother has a higher meiotic recombination rate than the father.[29,30] Consequently, the blocks of sharing from the mother are shorter than from the father. Moreover, given the recombination rate, it is not immediately clear how one can translate excess sharing into numbers of loci shared. We, therefore, sought to normalize gender differences in meiotic recombination and to relate this number to the number of risk alleles shared. We next define "discretized" sharing and methods to compute it.

We consider that deviation in sharing from the null expectation might be driven by discrete shared causal loci (SCLs). In this case, the degree of genomic sharing would be governed not only by the number of such loci but also by the size of meiotic blocks. We begin by considering the simplest case, a single SCL necessary and sufficient for a child to be affected. One example illustrating an SCL is a single dominant allele carried by one parent. In this case, concordant affected siblings must share the allele, and discordant siblings must not. The expected amount of genomic sharing from the carrier parent will differ from the null expectation of 0.5. The expected change in sharing due to one shared SCL is

$$\delta = mean\,(X_f - 0.5) \qquad \text{(Equation 1)}$$

where X_f is the proportion of genomic sharing for the carrier parent measured over a simulated population of such families. The magnitude of δ depends on the expected size of the shared inherited block containing the SCL, which depends on the meiotic recombination rate surrounding that locus. The value of δ averaged over all possible loci would reflect the meiotic

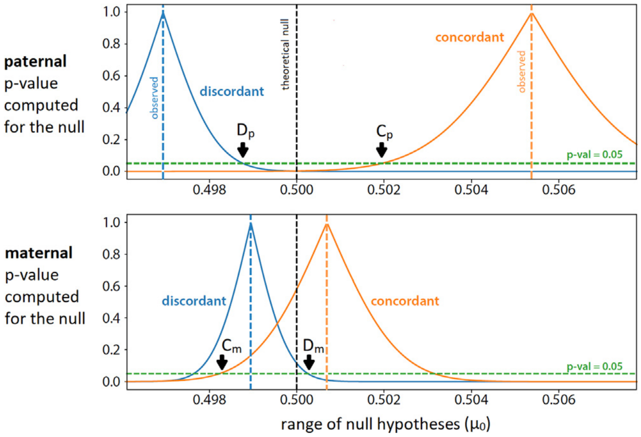


Figure 1. Statistical robustness for different null models

Statistical robustness with regard to the choice of null hypothesis about expected sharing (μ_0). The x axis shows a range of expected sharing values, μ_0, subject to hypothesis testing. The y axis shows two-sided p values of t tests as a function of μ_0. Accordingly, the p values are equal to 1 at the observed values of mean sharing, and the 0.95 confidence intervals (CIs) are the sets of expected sharing values where the tent-like function graphs exceed the 0.05 level (green horizontal line). By definition, a 0.95 CI contains exactly the values that could not be rejected by a null hypothesis test at the 0.05 level. Conversely, the values outside the CIs (hence p values <0.05) are those that can be rejected. The upper histogram shows the results for paternal sharing. For the discordant sibling pairs, the significant expected sharing values of interest are those greater than D_p and, for the concordant sibling pairs, they are those less than C_p. The interval from D_p to C_p consists of expected sharing values that can be rejected for both cohorts and contains the theoretical value μ_0 = 0.5. The lower histogram shows the results for maternal sharing. No reasonable null results in rejection of the observed maternal sharing by concordant siblings. For the null expectations μ_0 = 0.5003 and greater, the observed sharing of the mother by discordant siblings becomes significant.

recombination rate of the parent. The higher recombination rate in mothers results in a smaller δ than in fathers.

Staying with the simplest case, we estimate δ for one shared SCL as follows. We sample from our data with random polarity reversal to obtain effectively random chromosomal recombination patterns, thereby simulating patterns of "sibshares" for mothers and fathers. We then randomly chose a locus for each simulated sibshare and partitioned the sibshare patterns into those concordant and those discordant for sharing at the chosen locus. By averaging sharing over concordant sibs, we obtained δ = 0.0089 for fathers and δ = 0.0045 for mothers. These values do not depend on the cohort used to generate the random recombination patterns (Figure S3).

We note the following approximation:

$$\delta \sim 2 * \sigma^2 \qquad \text{(Equation 2)}$$

where δ is the change in sharing due to one shared SCL and σ^2 is the variance of the genomic sharing for an unascertained population. For example, we see that for fathers of simplex 2 × 0.00442 = 0.00884–0.0089, and 2 × 0.00223 = 0.00446–0.0045 for mothers. This close approximation is not an accident. We prove in Data S1 that Equation 2 is precise for the simplest case under two broad conditions: uniform random distribution of the location of the SCL, and duality of recombination outcomes; that is, the transmission of one recombinant pattern is as likely as the transmission of its complementary haplotype in the absence of ascertainment. These are precisely the conditions we assume in our simulations.

To explore the more general cases, we simulate combinations of multiple SCLs, where each SCL can be shared or non-shared. For example, we simulate sharing a parental genome under the constraint that two SCLs are shared between the two siblings while one additional SCL is non-shared, for a net of one SCL. We consider only constraint patterns where the number of constraints is small relative to the number of chromosomes to avoid substantial distortions due to linkage. In all constraint patterns we simulated, we see that sharing is nearly linear with the net number of shared SCLs (defined as the difference in the number of shared and non-shared SCLs) and with a slope δ (Figure S4). The variance in sharing decreases somewhat with the total number of constraints, but the effect is minimal (Figure S4). Thus, α (the net number of SCLs shared per parent) over simulated populations with a given constraint pattern can be approximated as

$$\alpha \sim (\overline{Y} - 0.5 / \delta) \qquad \text{(Equation 3)}$$

where $\overline{Y}$ is the mean genome sharing over the population simulated under the constraint pattern and δ is the amount of sharing change due to one shared SCL. δ is dependent mainly on the gender of the parent. As the approximations do not depend on the specific constraint pattern, it is clear that Equation 3 will hold for populations with mixed constraint patterns.

It is somewhat of a leap to assert that the approximations of Equations 2 and 3 hold in a population with an unknown causal landscape. Nevertheless, we expect that they would be valid over a large number of genetic models (see section "discussion") where the total number of specific loci that could significantly contribute to risk in any given family is small compared with the number of chromosomes, and the loci themselves are broadly distributed over the genome. The virtue of the equations is that they are simple, normalize the different meiotic recombination rates of mothers and fathers, and provide some insight into the number of shared events.

We applied Equation 3 separately to all concordant and all discordant siblings using δ = 0.0089 for fathers and 0.0045 for mothers to obtain α, the mean number of net SCLs per parent (Table 1). Using Gaussian or bootstrap analysis, we generate a distribution of α from the data (Figure 2). We obtain an estimate of 0.66 net SCLs per concordant sibling pair from fathers (95% confidence interval [CI] [0.26, 1.07]) and 0.30 from mothers (95% CI [−0.28, 0.87]). These data are consistent with more loci being shared from the fathers' genomes than the mothers' (Figure 2). However, given the present size of our concordant cohort, we cannot state with confidence that the net SCLs from fathers exceed the net from mothers (p value = 0.15, based on a one-sided test; see STAR Methods). Combining the net from mother and father we observe a total net SCLs of 0.66 + 0.30 = 0.96 (95% CI [0.26, 1.66]) per family. Repeating this process for discordant sibling pairs (Table 1; Figure 2), we find an estimate of −0.35 SCLs per sibling pair from fathers (95% CI

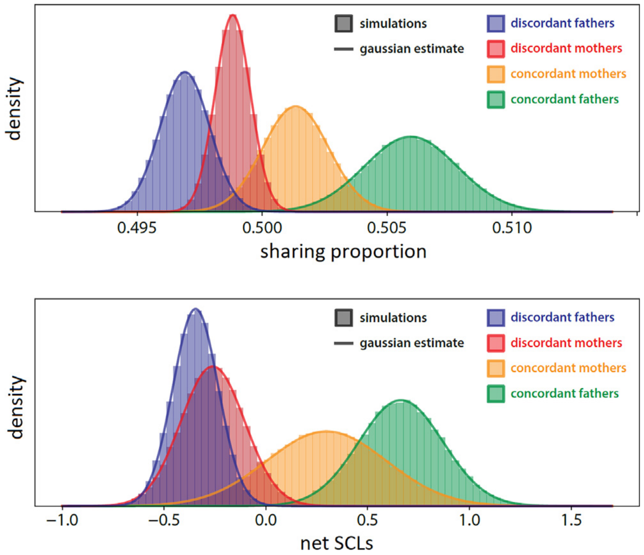


Figure 2. Confidence of observations by cohort

The top panel shows two approaches to quantify the confidence of the observed sharing of the paternal and maternal genomes in the concordant and discordant siblings (see figure legend for color coding). The first approach is to plot a normal distribution with mean equal to the observed sharing and standard deviation equal to the standard error of the mean (SEM). The second approach is to use 1 million bootstrap iterations, where we sample with replacement from original cohorts to build random cohorts of the same size. The results of the sharing across the bootstrap iteration cohorts are then plotted as histograms. The normal approximation and the bootstrap histograms overlap almost perfectly. In the lower panel, we show the confidence of the estimate of the net SCLs that uses the observed sharing proportions and accounts for differences in recombination rates of mothers and fathers. The confidence distributions for the mothers and for the fathers overlap significantly more for the net SCLs compared with the distributions for the sharing, because the former is adjusted by the respective recombination rates.

[−0.56, −0.13]) and −0.26 SCLs from mothers (95% CI [−0.57, 0.05]). Again, the magnitude of the share is greater from the father's genome than the mother's but of less significance (p value = 0.33). The total non-shared net SCLs is −0.61 (95% CI [−0.98, −0.23]).

Examples

In this section, we explore the interpretation of discrete sharing by examining simulations of specific models. The models are specified by the number of loci, their frequency, their risk weights, and the tolerance for the risk of males and females. All loci have an allele with a zero and an allele with a nonzero weight with the frequency given for the nonzero allele. Positive weights increase the risk for autism, and negative weights decrease it. Families with affected parents are excluded, as in the cohorts we observed. The models are all deterministic, additive, and unidirectional in that excess load greater than threshold tolerance results in autism. After simulating and pairing unaffected parents, multiple male offspring are generated per family by independently segregating transmission. We then simulate ascertainment of simplex and multiplex family quad collections and collect statistics in Table 2: net SCLs of each parent and risk to male and female children. In all our models, we sought conditions where male incidence was about 1.4% and female incidence about one-third of that. We describe our software tool in Supplemental Data 2 and make it available as an open-source package. Readers can readily reproduce our examples or explore different models on their own.

One might assume that, given the female protective effect,[33–35] that the mother can carry more risk load than the father, and consequently that she should be the parent more shared by concordant siblings. If all the variants confer risk (that is, they have positive weight) and if they are infrequent, then, indeed, the mean discrete sharing of the mother will typically exceed the mean for the father over the entire population of concordant male siblings. However, perhaps surprisingly, if we allow for strong protective variants, or allow the risk variants to be better tolerated in females as homozygous, then we readily find models in which the mean sharing of the father exceeds sharing of the mother over the entire population. We illustrate these results with examples.

Scenarios that favor sharing from the mother

We first consider the stringent version of the unified hypothesis, where the population carries a large-risk variant tolerated by females but not by males. In such a case, there would be no sharing of the father, only of the mother, and girls in multiplex families would have a very low-risk incidence. That solution is incompatible with the observations of a higher incidence of autism in girls from multiplex families, and also, from this report, that the fathers' genomes are very significantly shared. So, we next consider the more plausible case where the large rare risk variant is easily tolerated by both parents but close to the limit for the male. Then, given a background of weak risk alleles, the presence of the large-risk variant can push beyond the thresholds more frequently in males. The background of common weak risk variants provides a component of multi-genic risk. The compiled statistics of such a model, example 1, are summarized in Table 2. This model creates an excellent fit to all the known observations of autism epidemiology, except for one: the discrete share of the mother greatly exceeds that for the father, to an extent statistically inconsistent with the observations of this paper. The reader can reason for themself, or see example 2, that, as the strength of the rare variant diminishes, the father's share approaches the share of the mother but never exceeds it.

Scenarios that favor sharing from the father

We have found two distinct solutions to greater paternal sharing, even when females have a stronger risk tolerance than males. The first solution arises when there are strong protective variants in a population rife with low-risk variants (example 3). In this model, the ratio of paternal to maternal sharing is 1.33 to 1.18, or 1.13. The explanation is that, while a mother or father can equally likely be assigned the protective allele, a father carrying it can carry more risk variants and still be ascertained as unaffected, thus increasing the likelihood of membership in a multiplex cohort, and at the same time increasing the share at this locus by concordant affected siblings (by avoidance of the protective variant in the concordant affected). In contrast, mothers with the protective variant do not carry much more risk than they could otherwise. The protective variant reduces

Table 2. Autism risk and sharing under-five genetic models

	Model definition					All families		Multiplex				Simplex			
	Threshold		Loci			Risk		Risk		Net SCLs		Risk		Net SCLs	
Model Name	Male	Female	Weight	Frequency	Number	Male	Female	Male	Female	Pat. (0.66)	Mat. (0.30)	Male	Female	Pat. (−0.35)	Mat. (−0.26)
Example 1: strong rare positive risk	9	11	1	0.05	40	0.014	0.004	0.410	0.190	0.309	0.953	0.261	0.098	−0.136	−0.421
			8	0.01	2										
Example 2: uniform rare positive risk	8	9	1	0.05	40	0.014	0.004	0.226	0.105	1.034	1.111	0.122	0.047	−0.163	−0.175
Example 3: protective variant*	18	22	1	0.025	160	0.013	0.004	0.189	0.075	1.33	1.18	0.078	0.027	−0.128	−0.113
			2	0.025	80										
			−10	0.01	150										
Example 4: homozygous effect on sharing	1.2	2	1	0.125	1	0.013	0.004	0.292	0.076	1.000	0.832	0.266	0.074	−0.375	−0.312
			0.01	0.1	3										
			−0.01	0.1	3										
Example 5: complex frequent risk with protection*	10.7	11.24	0.15	0.9	40	0.013	0.004	0.190	0.061	1.49	1.18	0.074	0.023	−0.135	−0.107
			−15	0.03	60										

The table shows five example genetic models and the predictions of each model about the risks and the sharing of parental genomes in three types of families. The first column shows the descriptive name we gave the models, and the following five columns (or the “model definition” section of the table) define the models. All models are defined by a male and a female threshold for liability and a set of locus classes describing loci with alleles contributing weights to the liability. When an individual’s liability exceeds the individual’s gender threshold, the individual is considered affected. Each locus class describes a given number (shown in the “number” column) of population biallelic loci with one allele that contributes 0 weight to the liability and one allele that adds a given nonzero weight (shown in the “weight”’" column) to the liability. The nonzero allele has a population frequency, given in the “frequency” column. Each model occupies as many rows in the table as the number of its locus classes. For example, the model "example 1: strong rare positive risk" takes two lines because it has two locus classes, while "example 2: uniform rare positive risk" takes one line because it has one locus class. The first row of each model contains the liability thresholds and all the model predictions.

We assumed that all the loci were diploid/autosomal and examined pairs of unaffected parents. Each parent’s genotype is generated by randomly sampling two alleles for each locus and ignoring genotypes with liability larger than the gender-specific threshold. We refer to the set of all such parental pairs as "all families" in the table. We used all possible children for each parental pair to compute statistics such as the probability for a boy or a girl born in the family to be affected and the probability that an affected boy inherits the nonzero allele in each parental heterozygous locus. We also examined two subsets of families: multiplex and simplex. The multiplex are families with unaffected parents and two affected male children. The simplex are families with unaffected parents, one affected, and one unaffected male sibling. The definitions of multiplex and simplex families mimic the ascertainment criteria applied for our collections of concordant and discordant sibling pairs.

For each of the three family types, we report the average (across all families of the family type) risk for males and females to be affected in the “male” and “female” columns in the three “risk” sections of the table. For multiplex families, we also report the expected number of net SLCs from the paternal (“pat.” column) and maternal (“mat.” column) genomes between the two affected male siblings. For the simplex, we add the expected (negative) net SLCs between the affected and unaffected male siblings. For all four net SCLs measures, we show the observed values in brackets as reported in Table 1.

We computed the predictions for examples 1, 2, and 4 using an exact analytical computation by enumerating all possible family configurations. Examples 3 and 5 (labeled with *) have too many possible family configurations for the analytical computation, so we used a sampling procedure to compute estimates (see Data S2). We repeated the sampling procedure 10 times, and, in each run, we sampled 1 million families. The table reports the mean across the 10 runs of each estimated parameter up to the digit that reflects the estimate’s precision. For example, the reported value of 1.48 means that all the estimates fall in the interval of (1.48–0.05, 1.48 + 0.05) = (1.43, 1.53), and, if the reported value is 0.066, all the estimates are in the interval (0.066–0.005, 0.066 + 0.005) = (0.061, 0.071).

the risk to their offspring, making it less likely that these mothers will be recruited to a multiplex cohort.

The second solution arises when homozygosity of risk alleles is better tolerated in females, and they are sufficiently frequent to occur in that configuration. In the simplest possible case, a single locus with a risk variant is tolerated in the homozygous state by the mother but not by the father. If males tolerate the heterozygous locus, sharing will favor the father even though the mother carries a greater risk load. Autistic male offspring can come only from parents where the father has one risk variant, and the mother has either one or two at the locus. In the first case, both parents have equal sharing by concordant siblings. In the second case, sharing is only from the father because the siblings can inherit either of the mother's two risk alleles. In this simple model, girls from multiplex families do not have autism by transmission, and there would be no evidence of multi-genic contribution. So, to create a more realistic model, we add a background of low-level risk (example 4). In this particular model, the ratio of paternal to maternal sharing is 1 to 0.83, or 1.2. To create a greater variety of solutions with a large multiplicity of loci and to combine these mechanisms, we show example 5. The ratio of sharing in the multiplex families is 1.49 to 1.18, or 1.26.

We note that, in all examples except the first, the total share in the multiplex is on the order of 2-fold greater than the observed values, and the risk to the male child in multiplex families is too low. So, while these models offer solutions to the excess share of the father compared with the mother, they do not quite fit critical observations.

DISCUSSION

A complete genetic model for autism would accommodate these observations: an overall incidence of 1%–2%; a male to female ratio of about 3:1; strong concordance between monozygotic twins, on the order of 70% or higher; high sibling rates after one affected, on the order of 20%; risk to the third-born male after two prior affected siblings of nearly 50%; contribution of *de novo* mutation in at least 20% of cases; multiple causal loci with no single risk allele accounting for much more than 1% of cases; and evidence of multi-genic contribution. To this list, we can now add a mean paternal and maternal discrete sharing by concordant siblings of about 0.66 (CI [0.26, 1.07]) and 0.3 (CI [−0.28, 0.87]), respectively, and about half that amount as anti-sharing in discordant siblings (−0.35, CI [−0.56, −0.13] for the paternal and −0.26, CI [−0.57, 0.05] for the maternal).

Our previous unified hypothesis satisfied every observation in the above list except that the sharing of the father's genome in concordant siblings is not smaller than that of the mother's. However, the unified hypothesis predicted that, in multiplex families, mothers would carry recent and penetrant rare risk alleles, transmitted to their offspring. In the strong form of this hypothesis, unaffected fathers could not carry the variant and would play no role in transmission. The present work establishes paternal transmission in the incidence of autism, both in multiplex (p value = 0.0014) and simplex (p value = 0.0022) families, in conflict with the strong version of the unified hypothesis. The unified hypothesis has a weaker and more plausible form, in which mothers and fathers can both carry strong and rare risk alleles, but, in the presence of background risk, such as caused by multi-genic contribution, the mother is the more likely carrier. This weaker form still predicts greater sharing of the maternal genome by concordant siblings, as she would often be the parent to carry a single large-risk variant. We examined such models, and while we do not provide a mathematical bound on them, we demonstrate with example 1 how a typical version would fail.

Not only did we observe significant genome sharing of the father by concordant siblings but its significance and magnitude also exceeded that of the mother. To explore this further, we needed to adjust genomic sharing for differences in meiotic recombination rates, higher in mothers than fathers. We first developed the mathematics for the case in which sharing was driven by a single discrete loci and then generalized it, showing by simulation that the formulation is a good approximation for the net excess of SCL when they number only a few. We call this discrete sharing, measured in units of net SCLs, and this statistic allows us to adjust genomic sharing by the meiotic recombination rate and to compare mothers and fathers on equal footing (Figure 2). We thus see that there are more net SCLs from the father for concordant and discordant children (0.66 and −0.35 SCL) than from the mother (0.30 and −0.26 SCL). The hypothesis that the discrete sharing of fathers relative to mothers is equal has a one-sided p value of 0.15.

As this p value is shy of nominal statistical significance, we could ignore it. Rather, we chose to determine whether increased discrete sharing of the father is even possible when the mother has a greater risk tolerance, and we can answer in the affirmative. In the event of only rare positive risk variants, we find no plausible models in which the concordant siblings share more net SCLs from father than mother (see example 2). However, if we drop that constraint on risk frequency, or consider negative risk (that is, protective variants), we can construct examples in which the father is more often shared than the mother even though she carries more genetic load. The examples occur either when there are strong protective variants (see example 3) or when the mother is better able to carry homozygous risk variants than the father (see example 4). Finally, we present a model with both features that provides an even better fit to the sharing ratio (see example 5).

Although these examples demonstrate that an increased share from the father is possible even though the mother has more risk tolerance, they do not fit the data for two related parameters. The total discrete share is too high, by a factor of two, and the incidence of autism in a male child is too low by a factor of two when two previous children from the family have been diagnosed. We have tried strenuously but have not yet found any solution within the constraints of an additive deterministic model that satisfies overall autism incidence, ratio of maternal to paternal discrete sharing, total discrete sharing, and risk to the third-born male child in multiplex families. We cannot yet satisfy all these parameters with additive genetic models in the absence of strong and rare risk alleles.[36] We are also considering several extensions of these models, including gender-specific weights and assortative mating. However, identifying and exploring the complex sub-space of models consistent with all the observations is far from trivial and is the subject of future investigations.

In the absence of a classical model that fits all the data, or even in its presence, we should still consider non-classical models that do. We mention two because they form an important class of testable possibilities. First, epigenetic events, akin to an error in imprinting, might occur early in gametogenesis. For example, if one allele was incorrectly silenced in the germ cell precursors, the improperly imprinted allele might be shared or avoided by concordant siblings. Second, the affected siblings might share paternal antigens that in themselves do not confer risk but, due to prior sensitization in a given mother, might cause the fetus to encounter a maternal immune response; for example, an immunoglobulin (Ig) G that passes the placental barrier, resulting in a developmental abnormality. This possibility is consistent with persistent reports of increased autism incidence given certain immunological preconditions.[37–41] These non-classical mechanisms, if a driving force shaping the phenotypic landscape, would not be restricted to autism and, therefore, might be observable as excess paternal share in other disorders of development.

Limitations of the study

We have not yet extended our studies to the role of the X chromosome. We have yet to identify a genetic model consistent with the major known incidence and sharing parameters of autism. This work is an ongoing effort, and the hope is that this publication will motivate others to join the search.

The discretized sharing method we describe is general and requires only array genotyping. It can be applied to other disorders when a sufficient (on the order of 1,000) number of quad families are available. However, such large collections are generally not available.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Probe selection
  - ○ Selection of sibling pairs
  - ○ Sharing at informative SNPs
  - ○ Spikes and spike removal
  - ○ Filtering for chromosomal abnormalities and large copy number variants
  - ○ Measures of sharing
  - ○ Permutation test comparing two groups of sibling pairs for their mean sharing of a parent
  - ○ Permutation test for comparing the mean sharing of a parent for a group of sibling pairs with the theoretical null expectation
  - ○ Simulation of sharing under constraints
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100319.

ACKNOWLEDGMENTS

We thank Tobias Janowitz for his helpful suggestions. We also thank all the families at the participating SSC sites, as well as the principal investigators (A.L. Beaudet, R. Bernier, J. Constantino, E.H. Cook, Jr., E. Fombonne, D. Geschwind, D.E. Grice, A. Klin, D.H. Ledbetter, C. Lord, C.L. Martin, D.M. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M.W. State, W. Stone, J.S. Sutcliffe, C.A. Walsh, and E. Wijsman) and the coordinators and staff at the SSC sites for the recruitment and comprehensive assessment of simplex families, and the Simons Foundation Autism Research Initiative (SFARI) staff for facilitating access to the SSC. We appreciate useful discussions with scientists at Decode Genetics in Iceland on the subject of genomic sharing by unascertained sibling pairs. This work was supported by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory, SFARI grants SF497800 (to M.W. and I.I.); SF529232, SF677963, SF666590 (to I.I.); and the Centers for Common Disease Genomics grant (UM1 HG008901). The Centers for Common Disease Genomics are funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute.

AUTHOR CONTRIBUTIONS

Conceptualization, M. Wigler, I.I., K.Y., and D.L.; methodology, M. Wigler, I.I., M. Wroten, D.L., A.M.K., and A.B.; investigation, M. Wroten, S.Y., I.I., B.Y., and P.A.; writing – original draft, M. Wigler, M. Wroten, and I.I; writing – review & editing, M. Wigler, I.I., A.M.K., A.B., and M.R.; funding acquisition, I.I. and M. Wigler.; resources, I.I., P.A, S.Y., and B.Y.; supervision, M. Wigler and I.I.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 23, 2022
Revised: April 30, 2022
Accepted: April 12, 2023
Published: May 8, 2023

REFERENCES

1. Dias, C.M., and Walsh, C.A. (2020). Recent advances in understanding the genetic architecture of autism. Annu. Rev. Genomics Hum. Genet. *21*, 289–304. https://doi.org/10.1146/annurev-genom-121219-082309.

2. Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., and Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. Psychol. Med. *25*, 63–77. https://doi.org/10.1017/s0033291700028099.

3. Tick, B., Bolton, P., Happé, F., Rutter, M., and Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. J. Child Psychol. Psychiatry *57*, 585–595. https://doi.org/10.1111/jcpp.12499.

4. Sztainberg, Y., and Zoghbi, H.Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. Nat. Neurosci. *19*, 1408–1417. https://doi.org/10.1038/nn.4420.

5. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449. https://doi.org/10.1126/science.1138659.

6. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221. https://doi.org/10.1038/nature13908.

7. Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886–897. https://doi.org/10.1016/j.neuron.2011.05.015.

8. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885. https://doi.org/10.1016/j.neuron.2011.05.002.

9. Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. *94*, 677–694. https://doi.org/10.1016/j.ajhg.2014.03.018.

10. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into autism spectrum disorder genomic architecture and Biology from 71 risk loci. Neuron *87*, 1215–1233. https://doi.org/10.1016/j.neuron.2015.09.016.

11. Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M., and Geschwind, D.H. (2016). Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. Am. J. Hum. Genet. *99*, 540–554. https://doi.org/10.1016/j.ajhg.2016.06.036.

12. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250. https://doi.org/10.1038/nature10989.

13. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241. https://doi.org/10.1038/nature10945.

14. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299. https://doi.org/10.1016/j.neuron.2012.04.009.

15. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215. https://doi.org/10.1038/nature13772.

16. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245. https://doi.org/10.1038/nature11011.

17. Yoon, S., Munoz, A., Yamrom, B., Lee, Y.H., Andrews, P., Marks, S., Wang, Z., Reeves, C., Winterkorn, L., Krieger, A.M., et al. (2021). Rates of contributory de novo mutation in high and low-risk autism families. Commun. Biol. *4*, 1026. https://doi.org/10.1038/s42003-021-02533-z.

18. Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat. Genet. *46*, 881–885. https://doi.org/10.1038/ng.3039.

19. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444. https://doi.org/10.1038/s41588-019-0344-8.

20. Ye, K., Iossifov, I., Levy, D., Yamrom, B., Buja, A., Krieger, A.M., and Wigler, M. (2017). Measuring shared variants in cohorts of discordant siblings with applications to autism. Proc. Natl. Acad. Sci. USA *114*, 7073–7076. https://doi.org/10.1073/pnas.1700439114.

21. Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y.H., Yamrom, B., and Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proc. Natl. Acad. Sci. USA *112*, E5600–E5607. https://doi.org/10.1073/pnas.1516376112.

22. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat. Genet. *47*, 582–588. https://doi.org/10.1038/ng.3303.

23. Ruzzo, E.K., Pérez-Cano, L., Jung, J.Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and de novo genetic risk for autism impacts shared networks. Cell *178*, 850–866.e26. https://doi.org/10.1016/j.cell.2019.07.015.

24. Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D.H., Law, K., Law, P., Qiu, S., Lord, C., Sebat, J., et al. (2007). A unified genetic theory for sporadic and inherited autism. Proc. Natl. Acad. Sci. USA *104*, 12831–12836. https://doi.org/10.1073/pnas.0705803104.

25. Ozonoff, S., Young, G.S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., Bryson, S., Carver, L.J., Constantino, J.N., Dobkins, K., et al. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. Pediatrics *128*, e488–e495. https://doi.org/10.1542/peds.2010-2825.

26. An, J.Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576. https://doi.org/10.1126/science.aat6576.

27. pfeliciano@simonsfoundation.org, SPARK Consortium Electronic address pfeliciano@simonsfoundationorg, and SPARK Consortium (2018). SPARK: a US cohort of 50,000 families to accelerate autism research. Neuron 97, 488-493. https://doi.org/10.1016/j.neuron.2018.01.015.

28. Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J.B., Obiajulu, J.U., Wright, J.R., Murali, S.C., Xu, S.X., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat. Genet. *54*, 1305–1319. https://doi.org/10.1038/s41588-022-01148-2.

29. Zelkowski, M., Olson, M.A., Wang, M., and Pawlowski, W. (2019). Diversity and determinants of meiotic recombination landscapes. Trends Genet. *35*, 359–370. https://doi.org/10.1016/j.tig.2019.02.002.

30. Lenormand, T., and Dutheil, J. (2005). Recombination difference between sexes: a role for haploid selection. PLoS Biol. *3*, e63. https://doi.org/10.1371/journal.pbio.0030063.

31. Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D., Hallmayer, J., Kalaydjieva, L., McCague, P., Dimiceli, S., Pitts, T., et al. (1999). A genomic screen of autism: evidence for a multilocus etiology. Am. J. Hum. Genet. *65*, 493–507. https://doi.org/10.1086/302497.

32. Gagnon, A., Beise, J., and Vaupel, J.W. (2005). Genome-wide identity-by-descent sharing among CEPH siblings. Genet. Epidemiol. *29*, 215–224. https://doi.org/10.1002/gepi.20090.

33. Zhang, Y., Li, N., Li, C., Zhang, Z., Teng, H., Wang, Y., Zhao, T., Shi, L., Zhang, K., Xia, K., et al. (2020). Genetic evidence of gender difference in autism spectrum disorder supports the female-protective effect. Transl. Psychiatry *10*, 4. https://doi.org/10.1038/s41398-020-0699-8.

34. Dougherty, J.D., Marrus, N., Maloney, S.E., Yip, B., Sandin, S., Turner, T.N., Selmanovic, D., Kroll, K.L., Gutmann, D.H., Constantino, J.N., and Weiss, L.A. (2022). Can the "female protective effect" liability threshold model explain sex differences in autism spectrum disorder? Neuron *110*, 3243–3262. https://doi.org/10.1016/j.neuron.2022.06.020.

35. Wigdor, E.M., Weiner, D.J., Grove, J., Fu, J.M., Thompson, W.K., Carey, C.E., Baya, N., van der Merwe, C., Walters, R.K., Satterstrom, F.K., et al. (2021). The female protective effect against autism spectrum disorder. Preprint at medRxiv. https://doi.org/10.1101/2021.03.29.21253866.

Please cite this article in press as: Wroten et al., Sharing parental genomes by siblings concordant or discordant for autism, Cell Genomics (2023), https://doi.org/10.1016/j.xgen.2023.100319

36. Mukhopadhyay, S., Wigler, M., and Levy, D. (2015). Simple genetic models for autism spectrum disorder. Preprint at bioRxiv. https://doi.org/10.1101/017301.

37. Vinet, É., Pineau, C.A., Clarke, A.E., Scott, S., Fombonne, É., Joseph, L., Platt, R.W., and Bernatsky, S. (2015). Increased risk of autism spectrum disorders in children born to women with systemic lupus erythematosus: results from a large population-based cohort. Arthritis Rheumatol. *67*, 3201–3208. https://doi.org/10.1002/art.39320.

38. Tioleco, N., Silberman, A.E., Stratigos, K., Banerjee-Basu, S., Spann, M.N., Whitaker, A.H., and Turner, J.B. (2021). Prenatal maternal infection and risk for autism in offspring: a meta-analysis. Autism Res. *14*, 1296–1316. https://doi.org/10.1002/aur.2499.

39. Agrawal, S., Rao, S.C., Bulsara, M.K., and Patole, S.K. (2018). Prevalence of autism spectrum disorder in preterm infants: a meta-analysis. Pediatrics *142*, e20180134. https://doi.org/10.1542/peds.2018-0134.

40. Xie, S., Heuvelman, H., Magnusson, C., Rai, D., Lyall, K., Newschaffer, C.J., Dalman, C., Lee, B.K., and Abel, K. (2017). Prevalence of autism spectrum disorders with and without intellectual disability by gestational age at birth in the stockholm youth cohort: a register linkage study. Paediatr. Perinat. Epidemiol. *31*, 586–594. https://doi.org/10.1111/ppe.12413.

41. Meldrum, S.J., Strunk, T., Currie, A., Prescott, S.L., Simmer, K., and Whitehouse, A.J.O. (2013). Autism spectrum disorder in children born preterm-role of exposure to perinatal inflammation. Front. Neurosci. *7*, 123. https://doi.org/10.3389/fnins.2013.00123.

Please cite this article in press as: Wroten et al., Sharing parental genomes by siblings concordant or discordant for autism, Cell Genomics (2023), https://doi.org/10.1016/j.xgen.2023.100319

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
SSC whole-genome data	An et al.[26]; Yoon et al.[17]	SFARI Base (https://base.sfari.org/): ID: SFARI_SSC_WGS_2
AGRE whole-genome data	Ruzzo et al.[23]; Yoon et al.[17]	iHART; http://www.ihart.org
SPARK microarray genotype data	SPARK et al.[27]; Zhou et al.[28]	SFARI Base (https://base.sfari.org/): ID: SFARI_SPARK_WES_1
Genotypes at the selected ~300,000 positions for the SSC, AGRE, and SPARK individuals.	This paper	SFARI Base (https://base.sfari.org/): ID: SFARI_DS229125
Software and algorithms		
autopop tool	This paper; Zenodo	https://doi.org/10.5281/zenodo.7779998; https://github.com/iossifovlab/autpop

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ivan Iossifov (iossifov@cshl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All original code is available as an open-source project on GitHub (https://github.com/iossifovlab/autpop). The version that was used in the preparation of the manuscript has been deposited at Zenodo and is publicly available as of the date of publication (https://doi.org/10.5281/zenodo.7779998).

We deposited the genotypes for the common set of ~350,000 filtered genomic positions for all individuals in our cohorts as VCF files at SFARI Base under the dataset id: SFARI_DS229125. Access to this resource is subject to approval by the SFARI.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We started with a set of 2,380 (1,939 quads and 441 trios) families from the Simons Simplex Collection[17,26] (SFARI_SSC_WGS_2 at SFARI Base) and 859 nuclear families from the AGRE collection[17,23] of multiplex families with whole genome sequencing data. The New York Genome Center generated the whole genome sequencing at coverage of ≥30X. We generated genotypes for the families using our family-based multinomial genotyper.[14] We added 546 multiplex families and 2,576 simplex quad families from the first release of SPARK[27,28] (SFARI_SPARK_WES_1 at SFARI Base) with available microarray data based on a design developed by Regeneron that comprised ~635,000 probes. We verified that no individuals were included in more than one of these collections. Table S1 describes the families that entered our analysis pipeline.

METHOD DETAILS

Probe selection

We selected autosomal genomic positions with biallelic SNPs targeted by the Regeneron chip that 1) had both alleles observed in all four of our cohorts (SSC, AGRE, SPARK QUADS, and SPARK MULTIPLEX); 2) were genotyped in at least 95% of the individuals in each of the four cohorts; and 3) had no other alleles observed in the whole genome data from SSC and AGRE. We further filtered out positions that violated Hardy-Weinberg Equilibrium (p value <0.00001) or displayed transmission bias (binomial p value <0.001) in any of the four cohorts or exhibited abnormal distribution of the alternative allele ratios for the heterozygous genotypes in the WGS data. We list the 370,000 positions that passed all the filters in Table S2.

Please cite this article in press as: Wroten et al., Sharing parental genomes by siblings concordant or discordant for autism, Cell Genomics (2023), https://doi.org/10.1016/j.xgen.2023.100319

Selection of sibling pairs

We selected one non-monozygotic twin pair from the families of the four cohorts: discordant (one affected and one unaffected) from the SSC and SPARK QUADS and concordant (two affected children) from AGRE and SPARK MULTIPLEX. Like the SSC trios or the twin families from AGRE, some of the families did not contribute a pair. We also excluded multi-generation families from AGRE. We detected the twins as children with 90% identical genotypes in our selected positions. When there were multiple available sibling pairs from a family, we randomly selected one.

Sharing at informative SNPs

We identified all positions informative for sharing each parent's genomes for the selected sibling pairs. A position is informative for sharing a parental genome if that parent has a heterozygous genotype while the other parent has a homozygous genotype at the position. The two children share the parent's genome at informative positions if they have identical homozygous or heterozygous genotypes. They do not share the parental genome if they have different genotypes.

Spikes and spike removal

When positions informative for sharing a parental chromosome for a pair of siblings are placed in genome order, they form consecutive large blocks of positions with the same sharing status (either shared or non-shared). But there are occasional positions that break these blocks (see Figure S1A). We define spikes as informative positions with a different sharing status than their immediate upstream and downstream informative neighbors. We remove the spikes from the list of informative positions for all chromosomes and all parents of the selected sibling pairs leading to a substantial decrease in the number of shared and non-shared blocks per chromosome (see Figures S1A and S1B).

Filtering for chromosomal abnormalities and large copy number variants

After we removed spikes, we filtered out 67 sibling pairs with a maternal chromosome with more than 15 switches (a transition from a shared to a non-shared block or vice versa) or with a paternal chromosome with more than ten switches. A large number of switches per chromosome indicates chromosomal abnormalities or large copy number variants (see Figure S1C for an example of trisomy 21). This is the last filtering step in our pipeline, and it yields 4,456 discordant pairs (1,921 from SSC and 2,535 from SPARK QUADS) and 1,269 concordant pairs (766 from AGRE and 503 from SPARK MULTIPLEX). For these sibling pairs, we recorded the first and the last positions and the number of informative positions for each shared and non-shared block for both the paternal and maternal genomes in all 22 autosomes in Table S3.

Measures of sharing

We measure sharing of a parental genome for a pair of siblings in two alternative ways. In the genomic length-based measure (Table S4), we assign each shared and unshared block a length equal to the genomic size of the blocks, defined as the difference of the last informative position and the first informative position for each block. In the SNP number-based measure (Table S5), we assign the length of the block to be equal to the number of informative positions within the block. Once we have set a length value for each of the blocks, we sum the lengths of all shared blocks and all non-shared blocks for each of the 22 autosomes separately: S_{klc} and N_{klc} stand for the total length of the shared and not shared blocks, respectively, for chromosome c of parent k in sibling pair l. We can then compute the sharing proportion (or "sharing" for short) $p_{kl} = \sum_c S_{klc} / (\sum_c S_{klc} + \sum_c N_{klc})$. Note that there is a gap between any pair of neighboring blocks (between the last positions of the first block and the first position of the second block), and sometimes these gaps are large (e.g. if a centromere falls in one of the gaps). Thus in the genomic-length-based measure, the $S_{klc} + N_{klc}$ is smaller than the length of chromosome c.

For a set of sibling pairs, we defined our discretized measure of sharing of a parental genome as:

$$\text{net SCLs} = \frac{\text{mean(sharing)} - 0.5}{2 * \text{var(sharing)}}$$

where the mean(sharing) and var(sharing) represent the mean and the variance of the sharing measurements (p_{kl}) for the parent across the sibling pairs. The definition of the net SCLs results from the theorem proven in Data S1, which shows that, under the assumptions of uniformity and duality, the expected sharing of a parental genome for siblings forced to share one genomic locus is equal to 0.5 plus twice the variance in sharing in an unascertained population. Thus, the value of the net SCLs measure is interpretable. For example, a net SCLs of zero means the sharing is at the expected level; a net SCLs of one means siblings share more than expected, and the extra sharing is equal to the sharing in siblings forced to share one locus; and a net SCLs of minus one means siblings share less than expected, and the decrease in sharing is equal to sharing seen in siblings forced to be different at one locus. The net SCLs measure also allows us to compare sharing of the paternal and maternal genomes because it accounts for the differences in maternal and paternal meiotic recombination through variance-based normalization.

Please cite this article in press as: Wroten et al., Sharing parental genomes by siblings concordant or discordant for autism, Cell Genomics (2023), https://doi.org/10.1016/j.xgen.2023.100319

Permutation test comparing two groups of sibling pairs for their mean sharing of a parent

We use label swapping to test if the parental sharing of two separate groups of sibling pairs are significantly different. For example, we use this permutation test to compare the mean sharing of the paternal genomes in concordant vs. discordant siblings (see "magnitude and statistical significance of sharing" in the Results). We use the difference of the mean sharing proportions between the two groups as a test statistic. We create an empirical null distribution by permuting one million times the label assignment for the sibling pairs from the two groups, making sure that the sizes of the groups are kept constant, and by recording the difference in means in every permutation. We can then use the empirical distribution to assign one- or two-sided p values.

Permutation test for comparing the mean sharing of a parent for a group of sibling pairs with the theoretical null expectation

We designed a permutation method to test if the mean sharing $p_k = \sum_{l \in L} p_{kl} / |L|$ of a parent k for a group of sibling pairs L is incompatible with the theoretical expectation of 0.5 sharing. We used the mean sharing as a test statistic. We generated an empirical null distribution by simulating one million chromosome sharing datasets (S^i_{klc} and N^i_{klc}) of the same size as the observed one. To generate each of the random datasets, we independently flip the observed S_{klc} and N_{klc} lengths (either keep them as observed or exchange them with probability 0.5) for each chromosome c and each in sibling pair l. For each simulated sharing dataset, we then compute the mean sharing $p^i_k = \sum_{l \in L} p^i_{kl} / |L|$ where $p^i_{kl} = \sum_c S^i_{klc} / \left(\sum_c S^i_{klc} + N^i_{klc} \right)$ and use the million p^i_k as empirical null to assign p value.

Note that the method described above is only applicable for the theoretical null of 0.5 sharing. For tests against different null models (as done in Figure 1), we use a t-test. T-tests results for the 0.5 null are virtually identical to the permutation method.

Simulation of sharing under constraints

We developed a simulation procedure to estimate the expected sharing for a parent for pairs of siblings that are forced to share a given number of random loci, U, while at the same time are forced not to share a given number, V, of additional random loci. We assume that the loci are independently and uniformly selected across the genome for each simulated sibling pair. We used this procedure to estimate δ, that is, the change from the unascertained background sharing to sharing under the constraint of one shared locus (see results), and evaluate sharing under more general constraints presented in Figure S4. The procedure has two steps: (1) generating sharing data for a random unascertained population of sibling pairs and (2) selecting (or ascertaining) the sibling pairs that obey all the requested sharing constraints.

The first step depends on an observed sharing dataset for the given parent, k, for a set of sibling pairs, L, with chromosomal sharing and non-sharing lengths S_{klc} and N_{klc}. We simulate sharing data for one million unascertained sibling pairs. To simulate a sibling pair, we select the sharing data from a random pair, l_i, from the observed sharing dataset and randomly flip the pairs of numbers S_{kl_ic} and N_{kl_ic}. We then compute the sharing proportion for the generated family p^i_k. At the second step, the ascertainment, we retain a simulated pair with probability $\left(p^i_k\right)^U \left(1 - p^i_k\right)^V$, or the probability that all the U random shared loci fall in a shared area while all the V random non-shared loci fall in a non-shared area of the simulated pair. In the end, we use the retained p^i_k numbers to estimate the expected sharing under the given forced constraints (as the mean of the retained p^i_k) and quantify the confidence of the simulated estimates.

The estimated sharing computed by the procedure is not sensitive to the cohort we use as an observed sharing dataset (Figure S3). In our experiments, we used all concordant and discordant sibling pairs.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantitative and statistical analyses are described in the relevant sections of the method details or the table and figure legends.