A unified genetic theory for sporadic and inherited autism

Xiaoyue Zhao*, Anthony Leotta*, Vlad Kustanovich†, Clara Lajonchere†, Daniel H. Geschwind‡, Kiely Law§, Paul Law§, Shanping Qiu¶, Catherine Lord¶, Jonathan Sebat*, Kenny Ye‖**, and Michael Wigler*,**

*Cold Spring Harbor Laboratory, 1 Bungtown Road, P.O. Box 100, Cold Spring Harbor, NY 11724; †Autism Genetic Resource Exchange, Cure Autism Now, 5455 Wilshire Boulevard, Suite 2250, Los Angeles, CA 90036; ‡Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-1769; §Department of Medical Informatics, and Interactive Autism Network, Kennedy Krieger Institute, Baltimore, MD 21205; ¶University of Michigan Autism and Communication Disorders Center, 1111 East Catherine Street, Ann Arbor, MI 48109-2054; and ‖Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461

Contributed by Michael Wigler, June 20, 2007 (sent for review June 1, 2007)

Autism is among the most clearly genetically determined of all cognitive-developmental disorders, with males affected more often than females. We have analyzed autism risk in multiplex families from the Autism Genetic Resource Exchange (AGRE) and find strong evidence for dominant transmission to male offspring. By incorporating generally accepted rates of autism and sibling recurrence, we find good fit for a simple genetic model in which most families fall into two types: a small minority for whom the risk of autism in male offspring is near 50%, and the vast majority for whom male offspring have a low risk. We propose an explanation that links these two types of families: sporadic autism in the low-risk families is mainly caused by spontaneous mutation with high penetrance in males and relatively poor penetrance in females; and high-risk families are from those offspring, most often females, who carry a new causative mutation but are unaffected and in turn transmit the mutation in dominant fashion to their offspring.

human genetics | neurodevelopmental disorders | population genetics

Autism Spectrum Disorder (ASD) (Online Mendelian Inheritance in Man accession no. 209850) is characterized by language impairments, social deficits, and repetitive behaviors; can occur either sporadically (simplex) or in a familial (multiplex) pattern; occurs far more commonly in males; and has an overall incidence of ≈1 in 150 births (1). Monozygotic (MZ) twins show >70% concordance (2), higher with broader diagnostic criteria, and much higher than observed in dizygotic (DZ) twins, strongly suggesting that autism is genetically determined. Children with affected siblings have a higher risk than the general population, suggesting that autism can be inherited at least partially from preexisting genetic variants in parents.

Autism is likely to involve many genes. Linkage studies find no single locus of major effect but rather a very minor increase in allele sharing over the entire genome among concordant sibs (3–9). Cytogenetic studies (10), and more recently copy number analyses (9, 11, 12), support the idea that many loci may contribute to the disease.

Sibling and DZ concordance rates are perhaps one-tenth of MZ concordance rates, and this discrepancy, plus the suggestion of a large number of risk loci, has led many to expect that autism is attributable to complex multigenic interactions rather than simple dominant or recessive mutations. However, our current knowledge of genetic factors in autism suggests otherwise. Most of what we know about heritable risk factors comes from monogenic disorders, including fragile X syndrome (13–15), Rett syndrome (16), and tuberous sclerosis (17). Furthermore, cytogenetic findings and, more recently, copy number analysis point to a higher incidence of spontaneous mutation in children with sporadic autism (11), presumably occurring in a parental germ line.

An alternate to the multigenic interaction hypothesis is worth considering; most cases of autism are due to *de novo* mutation in the parental germ line, which can strike any of a number of critical loci. The large discrepancy in concordance rates between MZ twins and siblings can be explained thus: inherited or *de novo* mutation in the parental germ line affects MZ twins alike, whereas sibling and DZ concordance rates represent a mixture of modalities, sometimes inherited and sometimes *de novo* mutation.

This hypothesis makes a strong prediction: *de novo* mutations that can cause autism will also be found in resistant individuals, especially females, and these relatively asymptomatic individuals will mature to form high-risk families transmitting the mutation and hence the disorder in nearly dominant fashion to male offspring. To test one of our predictions, we analyzed autism risk in the multiplex families collected by the Autism Genetic Resource Exchange (AGRE) consortium (18) and find strong statistical evidence for families with a dominant pattern of transmission to their male offspring.

To determine the abundance of such families in the population, we extended our model beyond familial autism by incorporating two widely accepted ranges of parameters: the incidence of autism in the general population and the recurrence rate in siblings. Very simple models have a good fit to three independent databases of autism incidence; the vast majority of families have a low risk and contribute to the majority of autism, and a tiny minority of high-risk families, for whom transmission to male offspring is near 50%, contribute to the majority of the remainder. The high-risk families can be explained readily as generated by the expected number of offspring from low-risk families who have sustained *de novo* mutation but are themselves relatively asymptomatic.

We therefore propose the following unified model for sporadic and inherited autism. The majority of autisms are a result of *de novo* mutations, occurring first in the parental germ line. For reasons yet to be determined, female offspring are considerably more resistant to displaying the effects of such mutations than are males. Resistant individuals, but females in particular, carrying a mutation may marry and, with a probability of 50%, pass the mutation to their offspring, who will display the

Author contributions: X.Z., J.S., K.Y., and M.W. designed research; X.Z., K.Y., and M.W. performed research; X.Z., V.K., C. Lajonchere, D.H.G., K.L., P.L., S.Q., C. Lord, K.Y., and M.W. contributed new reagents/analytic tools; X.Z., A.L., D.H.G., S.Q., C. Lord, J.S., K.Y., and M.W. analyzed data; and X.Z., K.Y., and M.W. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: MZ, monozygotic; DZ, dizygotic; AGRE, Autism Genetic Resource Exchange; MLE, maximum likelihood estimate; IAN, Interactive Autism Network; MCMC, Markov chain Monte Carlo.

**To whom correspondence may be addressed. E-mail: kye@aecom.yu.edu or wigler@cshl.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0705803104/DC1.

GENETICS

symptoms with high probability if male. The latter process accounts for a minority of autisms, but together these two mechanisms are likely to account for the great majority of cases.

A consequence of this model is the explanation of increased autism incidence as a function of parental age (19); germ-line mutations can be expected to increase with age. The model also leads to other testable deductions and suggests practical approaches in the design of future genetic studies of autism and other disorders with similar epidemiological patterns, which will be discussed in the paper.

Results

The Autism Family Databases. We first and most thoroughly analyzed the AGRE database, comprising extensive and validated records of families with at least two affected children. Diagnoses are based on the Autism Diagnostic Interview-Revised. In most cases, the Autism Diagnostic Observation Schedule was also administered. These tools were designed to detect deficits in three behavioral domains: (*i*) reciprocal social interaction skills; (*ii*) qualitative communication skills; and (*iii*) restricted, repetitive, and stereotyped patterns of behavior (20, 21). For milder autism cases, the AGRE sample used a "broad-spectrum" category based on the presence of severe deficits in one of three domains of functioning moderate deficits in two of three or milder deficits in all three domains. We use the term "autism" or "Autism Spectrum Disorder" to refer to all three of these patterns.

We used two other independent databases. The first and much smaller from the University of Michigan was collected in part on the referral of families with at least two autistic children. The second, the Interactive Autism Network (IAN) Research Database, is a large and rapidly growing database collected by Internet registry, self referral, and self description (administered by P.L. and K.L.). The summaries of data and further information are in supporting information (SI) Tables 4 and 5 and legends.

For purposes of modeling, in all databases, we counted MZ siblings as one child, assigning the diagnosis of "affected" if even one of the MZ sibs was so diagnosed. In AGRE there was concordance of diagnosis in all 52 MZ male sibs and discordance in 3 cases of 11 female MZ sibs. In IAN, there was ≈80% concordance overall.

Transmission to Males in High-Risk Families. We look initially at families where the first two children are affected, taking these to be a subset of "high-risk" families, and consider when a third child is born. There are 165 such families in AGRE, and the third-born children can be divided by gender and affected status: 44 unaffected males, 42 affected males, 62 unaffected females, and 17 affected females. It is apparent that the risk to the third-born male child, i.e., his probability of being affected conditional upon the first two children being affected, is nearly 50%, consistent with dominant transmission with high penetrance. There is a lower risk for females (≈20%), consistent with the overall lower incidence in females and explainable as reduced penetrance in that gender.

Note that, by examining only the third-born children of families already having two children with autism, we avoid the potential bias of our risk estimate caused by stoppage. Stoppage refers to the tendency of families to cease having offspring when one of their children has autism. Stoppage can distort basic assumptions of the likelihood method used in the next section but is not operant in this data slice.

Risk estimates are affected by ascertainment bias. If the rate of recruitment for families with three children with autism was higher than the rate of recruitment for families with two, we would see an inflated risk rate. Although there is no effort from AGRE to favor families with more affected children once they meet the basic requirement of having at least two, one could argue that families with more affected children are more likely to self report to the study. However, in that case, we would also expect to see a distortion in the gender ratio of the third born, because autism is more common in males. In fact, the proportion of third-born children that are males is 86:165, or 52.1%, very close to the ratio of males to females in that age group nationally (105/205 = 51.2%) (22).

Based on the above insight, we can estimate the ascertainment bias from the gender and affected status of the third child in the 165 families whose first two children are affected, and simultaneously estimate the risks for each gender, by expressing these parameters in a likelihood equation and searching for the maximum likelihood estimate (MLE) of these parameters. Let p_f and p_m represent the risk of females and males respectively, and r represent the ascertainment bias. Using a multinomial distribution, we can express the likelihood relating gender bias and ascertainment by the following equation:

$$L(r, p_m, p_f) = \binom{n}{n_{AF}\ n_{UF}\ n_{AM}\ n_{UM}} \left(\frac{r p_f q_f}{C}\right)^{n_{AF}} \cdot \left(\frac{(1-p_f)q_f}{C}\right)^{n_{UF}} \left(\frac{r p_m q_m}{C}\right)^{n_{AM}} \left(\frac{(1-p_m)q_m}{C}\right)^{n_{UM}}, \quad [1]$$

where n_{AF} and n_{UF} represent the number of affected and unaffected females, n_{AM} and n_{UM} represent the males, and $n = n_{AF} + n_{UF} + n_{AM} + n_{UM}$, q_m and q_f are the proportion of male and female in the population, respectively, which equal 105/205 and 100/205 for this age group in the United States. In Eq. **1**, $C = rq_f p_f + q_f(1 - p_f) + rq_m p_m + q_m(1 - p_m)$ is a normalization factor.

There is a wide range of possible values for the bias r that we cannot rule out because of the small sample size; nevertheless the data do not suggest strong ascertainment bias. For details, see SI Fig. 1. The MLE for r is 1.14, a value we use for additional analysis (see below). With the consideration of ascertaining bias, the MLE for risk of males with autism is 0.46, still consistent with a pattern of dominant transmission with high penetrance.

Estimating Simple Autism Risk Models. To determine whether the observations consistent with dominant transmission in males would hold under analysis of more of the AGRE data, we sought a mathematical model fitting the data. At the same time, we could estimate the proportion of such high-risk families in the general population under the model by using accepted ranges of autism incidence, R, and sibling recurrence, S, for the general population. In the following, we restrict our analysis to male offspring only.

We assume that each male–female progenitor pair has a characteristic and time invariant risk x of producing male offspring with autism. The distribution of risk x across all progenitor pairs can be described by a density function $f(x)$ on [0, 1]. We choose to model $f(x)$ as having discrete risk components, corresponding to the intuition that there are a fixed number of genetic states for progenitor pairs. Using families of up to five offspring, there is information about the first five moments of $f(x)$, which in principle means that we can explore discrete density functions with as many as five independent parameters, that is, three-component mass functions. As it turns out, a two-component mixture model suffices to fit the available autism data very well.

Consider a family of n children. The probability that this family has m children with autism is given by:

$$P(n, m) = \int_0^1 \binom{n}{m} (1-x)^{n-m} x^m f(x)dx, \quad [2]$$

a mixture of binomial distributions of n trials. Note that Eq. **2** provides a system of equations for the moments of $f(x)$. The kth moment of $f(x)$ is defined as

$$m_k = \int_0^1 x^k f(x)dx. \quad [3]$$

Theories and methods for statistical inference of $f(x)$ in binomial mixtures are well developed (23, 24). However, our problem is unique in that families with fewer than two children with autism are not observed in AGRE. Therefore, existing methods do not directly apply to our data. In particular, we cannot infer the first two moments from the AGRE data. Therefore, in our maximum-likelihood approach, we use constraints for the first two moments, derived from R and S, and conditional probabilities for the likelihood function.

Let $F(n, m)$ denote the "shape" of a family with n children of whom m have autism, and let $f_\theta(x)$ be a density function parameterized by θ. Suppose that only families of shape $F(n, m)$ with $m \geq 2$ are observed, and here we consider only those families with at least three offspring. Under the assumption that the diagnoses of children within a family are independent, and that the families are themselves independent, a log-likelihood function of this data is given by

$$LL(\theta) = \sum_{n=3}^{N} \sum_{m=2}^{n} \text{obs}(n, m) \log(P_\theta(n, m|m \geq 2)), \quad [4]$$

where N is the size of the largest sibships, obs(n, m) is the observed number of families of shape $F(n, m)$, and $P_\theta(n, m|m \geq 2)$ denotes the conditional probability $P_\theta(n, m)/\Sigma_{k\geq 2} P_\theta(n, k)$, where P_θ is defined by Eq. **2** with $f(x) = f_\theta(x)$.

The first and second moments of f are constrained by R and S, the probability that a second child will be diagnosed with autism given that the first was, by the following two equations.

$$\int_0^1 xf(x)dx = m_1 = R \quad [5]$$

$$m_2 = S \times m_1 = S \times R. \quad [6]$$

Therefore, given the rate of autism R and sibling recurrence rate S, we have two constraints on the first two moments of $f(x)$.

We consider $f(x)$ as a k-component discrete density function

Table 1. Family shapes: Male offspring

No. of male children	No. of males with autism			
	2	3	4	5
3	85	26		
4	19	9	1	
5	2	2	1	0

Data were extracted from the AGRE database, counting the number of male offspring and of affected male offspring for each family.

$$f(x) = \sum_{i=0}^{k-1} a_i \delta(x_i), \quad [7]$$

where δ is the Dirac density function and the a_i sum to one. Here x_i is the probability of an offspring with autism, and a_i is the proportion of families with that probability. We consider two- and three-component parameterizations where the parameters θ are determined by a_i and x_i. To find the MLE for three-component models under the moment constraints, we applied a systematic grid search over all possible combinations of x_i with $0 \leq x_i < x_j \leq 1$ for $0 \leq i < j \leq 2$ using a grid with spacing 10^{-4}. For a given combination of the three x_i, the corresponding a_i are then uniquely determined by using the constraint equations on the first two moments and the constraint that the a_i sum to one. For two-component models, the three constraints allow exploration of all $0 \leq x_i \leq 1$ over a grid with the same resolution.

For our standard constraint parameters, we chose an overall autism rate R_T of 1/150 and a sibling recurrence rate S_T of 10%. Assuming an incidence ratio, males to females, of 3:1, we calculated new parameters for males: an autism rate R_M of 1%, and a sibling recurrence rate S_M of 15% (males born into a family with a single previous male child, that child having autism). In keeping with the preceding analysis, we make the conservative assumption that the ascertainment bias is 1.14, that is, a family with three or more children with autism is 1.14 times more likely to be recruited to the study than a family with only two affected (see *SI Text* for the mathematical details of the adjustment).

Table 1 describes the data from which we drew the values for obs(n, m), a total of 145 informative families, from families with up to five male children (female offspring were not counted). By considering families with up to five male children, we have enough data to examine a three-component discrete model (i.e., three risk types of families), but we examine first the parameters of simpler nested models. The MLE for parameters of one- and two-component models for males are shown in Table 2, first and second rows. x_i refers to the risk of the ith family type, and a_i refers to their proportion in all families. The one-component model $2(m, 0)$ is in actuality a two-component model with x_0, the "rate" of the first component, set to 0. That is, there are families

Table 2. Maximum-likelihood estimates for several models

Model type	Data source	a_0	x_0	a_1	x_1	a_2	x_2	LL
2(*m*,0)	AGRE	0.9333	0	0.0667	0.15			−112.70
2(*m*)	AGRE	0.9923	0.0067	0.0077	0.434			−88.211
3(*m*)	AGRE	0.7961	0.0067	0.1963	0.0068	0.0076	0.437	−88.210
2(*m*/*f*/*p*)	AGRE	0.9941	0.0072	0.0059	0.501			−783.02
2(*m*)	IAN	0.9932	0.0069	0.0068	0.462			−38.821
2(*m*/*f*/*p*)	IAN	0.9913	0.0066	0.0087	0.416			−326.49

Models 2(*m*) and 3(*m*) are the two- and three-component models for males only, respectively. Model 2(*m*,0) is for the model where x_0 is set to 0. Model 2(*m*/*f*/*p*) is for mixed gender (see *SI Text* and text below).

with zero risk, and the rest have high-risk x_1. In model 2(m), there is no constraint for x_0. That is, there are two types of families, each with uniform risk (x_0 and x_1). We achieve a much higher likelihood with the two-component model, with a P value equal to 2.59e-12 by χ^2 approximation and 0 of 1,000 by Monte Carlo simulation.

We have explored ranges of the autism rate for males, R_M, from 0.5% to 1.2% and ranges of the sibling recurrence rate for males, S_M, from 5% to 20% for model 2(m) (see SI Table 6), and it is satisfying to note that, regardless of R and S, the MLE of parameters impute low-risk families with transmission rates ranging from 0.0027 to 0.0111 (first-component rate x_0) and high-risk families with transmission rates from 0.429 to 0.519 (second-component rate x_1). Obviously, the second-rate component is always close to 0.5, yet nothing about Mendelian inheritance is embodied in our mathematical methods. Thus, we interpret families with high risk as transmitting to males in a dominant pattern with high penetrance, and this conclusion is largely independent for a wide range of incidence and sibling recurrence.

We next looked at MLE for parameters of the three-component model 3(m), males only, shown in Table 2, row 3, using the same constraint parameters, R_M and S_M. Despite providing a model with two extra parameters, presenting an opportunity to choose another component for risk, the MLE parameters show one high-rate component (x_2) that is nearly 50%, and two low-risk components that are essentially indistinguishable. Although there is no evidence for an intermediate-risk component, we examine its possibility below. The maximum likelihood of the three-component model is marginally higher than for the two-component model and not of any statistical significance (P value is 0.999 by χ^2 approximation and 0.86 by Monte Carlo simulation). Again, varying the values for rates of autism or sibling recurrence does not change this picture.

In summary, the MLE of the discrete-component models for males only imply a pattern consistent with a clear genetic interpretation; the population mixes two kinds of families: high-risk families, where the probability of a male offspring with autism is near 50%; and low-risk families. We interpret the high-risk families as transmitting the disorder in a dominant manner, with nearly complete penetrance for males.

Goodness of Fit of Simple-Risk Models to AGRE and Other Data Sets. To assess how well the statistical models fit the observations, we performed goodness-of-fit tests on the AGRE data for the two-component model, male only, listed as 2(m) in Table 2. A similar model that includes female offspring can be constructed from a two-component discrete-density function by postulating a uniform rate of penetrance in females [see 2(m/f/p) in Table 2 and *SI Text*). We used Pearson χ^2 as the test statistics, summarizing the difference between observed number of $F(n, m)$ families and expected number of $F(n, m)$ families from our model, for n = 3,4,5 and $2 \leq m \leq n$. The P values, obtained empirically from 1,000 random simulations (details described in *SI Text*), were equal to 0.98 and 0.41 for 2(m) and 2(m/f/p), respectively, suggesting our risk models fit the data well.

We performed goodness-of-fit analysis on two other data sets. For IAN, we used the estimated ascertainment bias of 1.06 (see below); otherwise, we used the same parameters (i.e., the a_i and x_i) estimated from AGRE data. For the University of Michigan data set, we obtained P values equal to 0.24 and 0.12 for 2(m) and 2(m/f/p), respectively (21 informative families for males only and 40 informative families for mixed gender). An analysis of the IAN database, an ongoing and expanding database, which, like AGRE, is based on self referral, had P values equal to 0.54 and 0.04 for 2(m) and 2(m/f/p), respectively (60 informative families for males only and 154 informative families for mixed gender).

The IAN database was large enough (3,000 families) for us to perform an ascertainment bias based on gender asymmetry, similar to what was described in Eq. **1**, and also independent maximum-likelihood estimation. The results yielded a bias of 1.26 for accruing families of size two with two affected children over families of size two with only the first child affected. For families of size three with three affected children over families of size three with only the first two children affected, the bias was only 1.06. These biases are consistent with our analysis of ascertainment bias in AGRE. The model parameters we derived from this database were very similar to those we derived from AGRE, with high-risk transmission equal to 0.46 and 0.42 for the 2(m) and 2(m/f/p) models, respectively (row 5 and row 6, Table 2). The family risk patterns inferred from these data are thus remarkably similar to that from AGRE.

The same simple models of risk fit the three databases well.

Size and Contribution of Risk Classes Using Simple and Complex Density Functions. Our simple two-component model [2(m) in Table 2] estimates that two-thirds and one-third of autism cases are contributed from low- and high-risk families, respectively. However, our biological conclusions do not depend on such a unique model. Although the data are always insufficient to determine the actual probability distribution of transmission in families, we can explore the likely functions within the same general classes of density functions and examine their properties, such as the size and contribution to autism from the various risk classes. To do this, we capped the risk at 0.5, the maximum expected risk for a dominantly inherited. model. We used our standard values for R_M and S_M and accepted for analysis those three-component mass functions with log-likelihood given the AGRE data within 1.92 ($= 0.5\chi^2_{1,0.95}$) of the maximum log-likelihood. We analyzed 10,000 three-component mass functions randomly selected from all such functions found during a grid search. We also broadened the class of allowed density functions beyond just three discrete components by considering the class of 150-component mass functions, thereby approximating the properties of continuous density functions. We used a Markov chain Monte Carlo (MCMC) method (25) to uniformly sample 10,000 150-component mass functions that satisfied the first and second moment constraints and gave a good fit to the data (with $P > 0.05$ in the Pearson χ^2 goodness-of-fit test). MCMC samples were collected every 1,000 iterations after an initial 5 million iterations. Details of the sampling method will be presented subsequently, along with some theoretical discussions relating to moment analysis and population genetics.

For each density function, we determined the size and contribution to autism from families with 0 to 0.01 risk (low risk), 0.1 to 0.3 risk (intermediate risk, which would include recessive patterns of transmission), and 0.3 to 0.5 risk (high risk, which would include dominant patterns of transmission with partial to complete penetrance). The mean and standard deviation of these values from each function class are summarized in Table 3. It is clear from our analysis that the major contribution to autism comes from low-risk families, those below the mean expected risk. Also, the contribution from high-risk families in general exceeds the overall contribution from intermediate risk families, whether we inspect 150- or 3-component mass functions.

By examining each particular reasonable density function from either class, contribution from the high-risk range exceeds the contribution from the intermediate risk range 92% of the time. Contribution from high-risk families exceeds even more broadly defined intermediate-risk families (range from 0.01 to 0.3) 65% of the time for either function class. We interpret this as evidence that, in a purely genetic model, the families with a dominant pattern contribute more cases than the families with a more complex genetic transmission pattern.

We do not know with certainty that the MCMC process (used for the 150-component mass functions) reached convergence,

Table 3. Mean (standard deviation) of sizes and contributions to autism of various risk classes

Model type	a_L	a_I	a_H	a_Lx_L/R	a_Ix_I/R	a_Hx_H/R	$a_Hx_H>a_Ix_I$	$a_Hx_H>a_{BI}x_{BI}$
3-component	0.9325 (0.0975)	0.0032 (0.0063)	0.0071 (0.0019)	0.474 (0.183)	0.051 (0.092)	0.303 (0.051)	0.921	0.653
150-component	0.9735 (0.004)	0.0089 (0.002)	0.0064 (0.0006)	0.513 (0.02)	0.172 (0.03)	0.260 (0.02)	0.949	0.731

Columns a_L, a_I, and a_H are the mean (standard deviation) of sizes of low-, intermediate-, and high-risk classes, respectively. Columns a_Lx_L/R, a_Ix_I/R, and a_Hx_H/R are the mean (standard deviation) of contributions for each class. Columns a_Hx_H/R, $a_Hx_H>a_Ix_I$, and $a_Hx_H>a_{BI}x_{BI}$ show the proportionate number of times that contribution from the high-risk range exceeds the contribution from the intermediate-risk and the broadly defined intermediate-risk ranges, respectively.

hence the sampling obtained may not be uniform from the targeted function space. However, our above conclusions are robust to the possibility of nonuniformity. This is because we deliberately chose our starting point for the MCMC to be the 150-component mass function that has the least contribution from the high-risk families. We obtained this function using mathematical programming under two linear constraints for first two moments and a quadratic constraint for the remaining moments using the Pearson χ^2 statistic.

Discussion

We must emphasize at the outset that our biological interpretation of the risk models assumes that risk is determined by genetic factors, and thus, except for their appealing simplicity, the risk models themselves should not be taken as evidence for genetic causation. We cannot rule out environmental factors, such as complications during pregnancy, contributing to the observed risk data, and the presence of these factors could impact our biological interpretation and some of our modeling assumptions. Nevertheless, in what follows, we assume a genetic basis for risk, justified largely by MZ and DZ twin studies.

Analysis of three databases shows that a simple-risk model has good fit for the incidence of autism in males, in which families fall into two major types: families for whom the risk of autism in male offspring is nearly 50%, as expected for dominant transmission with high penetrance; and families for whom male offspring have a low risk. The two-component model fits female incidence well if we simply add a penetrance factor of ≈0.3 for females. The two-component model is vastly superior to the single-component model. This latter finding is in keeping with the observation that there are at least two distinct genetic mechanisms for acquiring autism, one through *de novo* mutation in simplex families, and one through inheritance in multiplex families (11). Indeed, deleterious *de novo* mutations, which enter the gene pool for only a brief time, may be the simplest way in which to explain the discrepancy between MZ and DZ concordance rates and the origin of low- and high-risk families.

The data in our study do not rule out recessive modes of inheritance or a variety of more-complex genetic states that could give rise to families with gradations of risk. Nevertheless, if we exclude the possibility of families with greater than a dominant risk pattern, examination of simple and complex discrete density functions with plausible likelihood or χ^2 measures suggests that intermediate risk families with a recessive pattern of transmission would contribute less to autism incidence than families with a dominant pattern, and overall contribution is greatest from low-risk families.

We need to consider what factors might lead to spurious conclusions. The most obvious are overdiagnosis, stoppage, and ascertainment bias. Overdiagnosis seems to us unlikely. Standardized interviews were performed on each AGRE patient, so assessment is as objective as possible at the present time. Stoppage, the tendency to curtail procreation given affected offspring, could undermine the assumptions of the likelihood model. But stoppage is not relevant in the analysis of the third-born children, which showed evidence of families with the dominant pattern. A criticism of ascertainment bias is harder to rebut, because families with more than two affecteds might have a higher likelihood of being collected than families with just two. However, this potential bias is mitigated by two conditions: recruitment was initiated by self referral of the family, not by the encounter with a proband; and the collection was amassed over a 10-year interval. In addition, we have developed a method based on the asymmetry of gender ratios to estimate ascertainment bias and find only a minor effect in both AGRE and IAN databases.

In the following several paragraphs, we will consider the biological implications if the high-risk families are transmitting an allele in a dominant fashion and with high penetrance in males. Such families raise the question: what sort of mutation can cause autism, can be present in an apparently unaffected parent, and then can be transmitted to an offspring in a dominant manner? This pattern could be explained by X-linked mutations, but we see insufficient evidence of increased sharing of maternal X in concordant siblings in the AGRE data set (7). The most likely remaining explanation is disruptions on a single parental chromosome, with incomplete penetrance creating carrier states. Females are especially resistant and make logical carriers, but discordance between MZ twins suggests an "Autism Spectrum Disorder genotype" can exist in the absence of phenotype in either gender. To explain greater penetrance in males, we need merely consider the hypothesis that autism involves loss of cognitive abilities related to social skills, language, and repetitive behavior that may already be targets of sexual dimorphism (26–28) and hence these traits are already sensitive to perturbation.

Where are these high risk families coming from? We propose that a significant proportion of sporadic autism is caused by *de novo* mutation, deletions, duplications, other genomic rearrangements, or point mutations in the germ line of one parent that can cause loss-of-function (haploinsufficiency) or gain of function to any of a large number of target genes. This can occur in any family, regardless of genetic background, and can appear in any offspring, regardless of gender. Offspring with mild disorders or asymptomatic carriers such as females, may marry and have children who will inherit the mutation in a dominant fashion. With our standard parameters (R_T = 1/150, S_T = 10%, 3:1 males to females), the number of asymptomatic females would, by symmetry arguments, be 2 in 300 (0.0067) female newborns. This is the right magnitude needed to explain the proportion of high-risk families, which we estimate as 0.0077 (see Table 2, row 2, value a_1).

We observe *de novo* mutation in 10% of children with sporadic autism, significantly more often than observed in either healthy children or children from multiplex families (11). We believe that 10% is a gross underestimate because of the low resolution of the technique for discovering these copy-number mutations. The true rate of *de novo* copy number mutations in children with autism could well be three times that number, and we do not yet know the frequency of spontaneous point mutation. Thus the rate of spontaneous mutation of all types reasonably can be expected to account for the majority of sporadic autism. The great majority of *de novo* mutations we observe in sporadic autism are deletions (11), although other types of mutation, such

GENETICS

as duplication, do occur (29). Deletions will most likely create haploinsufficiency. There is no lack of examples of haploinsufficiency affecting neurological development, behavior, and cognition (30). The association of chromosomal abnormalities and submicroscopic deletions with autism supports the idea that gene imbalance can contribute to the disorder.

Our conjecture makes strong predictions, some testable in the short term. One parent in a high-risk family may have a *de novo* mutation. We predict this will be more commonly observed in the mother, and high risk should follow one parent when they have offspring with another partner.

We have achieved a simple model for autism in large part by focusing on the pattern of inheritance in males and using broad diagnostic criteria, while ignoring the details of the phenotype. Thus, our study does not directly address very important questions: what "modifier" genes influence the phenotypic manifestations of the disorder, the variable severity of autism in siblings, and the penetrance of autism in females. In particular, the high concordance in MZ females and the low penetrance in females suggest that additional genetic factors moderate incidence in that gender. Guided by our model, we propose that, whereas genetic association studies will fail to find causative mutations, such studies can find modifier genes.

Spontaneous mutation should be considered as a cause for any disorder clearly genetic in origin that reduces fecundity yet recurs in the population and particularly if the incidence is related to the age of the parents at conception. Besides autism, this potentially includes other severe developmental, metabolic, or neuropsychiatric disorders such as congenital heart disease, cerebellar dysfunction, morbid obesity, and severe mood and cognitive disabilities. The overall rate of *de novo* mutation in humans per generation is not a well determined figure and could be surprisingly high in light of our previous work (11). Finding *de novo* mutations by examining parent–child trios should become increasingly powerful over the next few years.

We thank Jim Simons, David Donoho, Terry Speed, Michael Zhang, Mary-Claire King, Gerry Fischbach, and Nat Heintz for helpful discussions and suggestions and Conrad Gilliam, Jonathan Pritchard, David Botstein, and Leonid Kruglyak for a critical reading of the manuscript. We gratefully acknowledge the resources provided by the AGRE Consortium and the participating AGRE families. AGRE is a program of Cure Autism Now and is supported, in part, by National Institute of Mental Health Grant MH64547 (to D.H.G.). [The AGRE Consortium: D.H.G.; Maja Bucan (University of Pennsylvania, Philadelphia, PA); W. Ted Brown (Institute for Basic Research in Developmental Disabilities, Staten Island, NY); Rita M. Cantor (University of California School of Medicine, Los Angeles, CA); John N. Constantino (Washington University School of Medicine, St. Louis, MO); T. Conrad Gilliam (University of Chicago, Chicago, IL); Martha Herbert (Harvard Medical School, Boston, MA); C. Lajonchere; David H. Ledbetter (Emory University, Atlanta, GA); Christa Lese-Martin (Emory University, Atlanta, GA); Janet Miller (Cure Autism Now); Stanley F. Nelson (University of California School of Medicine, Los Angeles, CA); Gerard D. Schellenberg (University of Washington, Seattle, WA); Carol A. Samango-Sprouse (George Washington University, Washington, DC); Sarah Spence (University of California School of Medicine, Los Angeles, CA); Matthew State (Yale University, New Haven, CT); and Rudolph E. Tanzi (Massachusetts General Hospital, Boston, MA).] This work was supported by a grant from the Simons Foundation.

1. Centers for Disease Control and Prevention. (2007) *MMWR Surveill Summ* 56:12–28.
2. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, Rutter M (1995) *Psychol Med* 25:63–77.
3. Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, Kalaydjieva L, McCague P, Dimiceli S, Pitts T, *et al.* (1999) *Am J Hum Genet* 65:493–507.
4. International Molecular genetic Study of Autism Consortium (2001) *Am J Hum Genet* 69:570–581.
5. Liu J, Nyholt DR, Magnussen P, Parano E, Pavone P, Geschwind D, Lord C, Iversen P, Hoh J, Ott J, Gilliam TC (2001) *Am J Hum Genet* 69:327–340.
6. Auranen M, Vanhala R, Varilo T, Ayers K, Kempas E, Ylisaukko-Oja T, Sinsheimer JS, Peltonen L, Jarvela I (2002) *Am J Hum Genet* 71:777–790.
7. Yonan AL, Alarcon M, Cheng R, Magnusson PK, Spence SJ, Palmer AA, Grunn A, Juo SH, Terwilliger JD, Liu J, *et al.* (2003) *Am J Hum Genet* 73:886–897.
8. Ylisaukko-oja T, Alarcon M, Cantor RM, Auranen M, Vanhala R, Kempas E, von Wendt L, Jarvela I, Geschwind DH, Peltonen L (2006) *Ann Neurol* 59:145–155.
9. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, Vincent JB, Skaug JL, Thompson AP, Senman L, *et al.* (2007) *Nat Genet* 39:319–328.
10. Vorstman JA, Staal WG, van Daalen E, van Engeland H, Hochstenbach PF, Franke L (2006) *Mol Psychiatry* 1:18–28.
11. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, *et al.* (2007) *Science* 316:445–449.
12. Jacquemont ML, Sanlaville D, Redon R, Raoul O, Cormier-Daire V, Lyonnet S, Amiel J, Le Merrer M, Heron D, De Blois MC, *et al.* (2006) *J Med Genet* 43:843–849.
13. Oberle I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boue J, Bertheas MF, Mandel JL (1991) *Science* 252:1097–1102.
14. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, *et al.* (1991) *Cell* 65:905–914.
15. Yu S, Pritchard M, Kremer E, Lynch M, Nancarrow J, Baker E, Holman K, Mulley JC, Warren ST, Schlessinger D, *et al.* (1991) *Science* 252:1179–1181.
16. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY (1999) *Nat Genet* 23:185–188.
17. The European Chromosome 16 Tuberous Sclerosis Consortium (1993) *Cell* 75:1305–1315.
18. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ (2001) *Am J Hum Genet* 69:463–466.
19. Reichenberg A, Gross R, Weiser M, Bresnahan M, Silverman J, Harlap S, Rabinowitz J, Shulman C, Malaspina D, Lubin G, *et al.* (2006) *Arch Gen Psychiatry* 63:1026–1032.
20. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E (1989) *J Autism Dev Disord* 19:185–212.
21. Lord C, Rutter M, Le Couteur A (1994) *J Autism Dev Disord* 24:659–685.
22. Mathews TJ, Hamilton BE (2005) *Natl Vital Stat Rep* 53:1–17.
23. Lindsay BG (1995) *Mixture Models: Theory, Geometry and Applications* (IMS, Hayward, CA).
24. Wood N (1999) *Ann Stat* 27:1706–1721.
25. Lovász L (1999) *Math Prog* 86:443–461.
26. Skuse DH (1999) *J Lab Clin Med* 133:23–32.
27. Skuse DH (2005) *Hum Mol Genet 14 Spec No* 1:R27–R32.
28. Baron-Cohen S, Knickmeyer RC, Belmonte MK (2005) *Science* 310:819–823.
29. Veenstra-VanderWeele J, Cook EH, Jr (2004) *Mol Psychiatry* 9:819–832.
30. Lee JA, Lupski JR (2006) *Neuron* 52:103–121.