Journal of Human Genetics (2009) 54, 721–726

www.nature.com/jhg

ORIGINAL ARTICLE

Major contribution of dominant inheritance to autism spectrum disorders (ASDs) in population-based families

Takeshi Nishiyama[1,2], Morihiro Notohara[3], Satoshi Sumi[4], Satoshi Takami[5] and Hirohisa Kishino[5]

Results of twin studies have shown that autism spectrum disorders (ASDs) are attributable to complex multigenic interactions rather than to a single susceptibility gene. However, the growing number of distinct, individually rare genetic causes of ASDs, mostly copy number variations (CNVs), favors an alternative to the polygenic hypothesis, the two-component model, which suggests that ASDs are caused either by *de novo* mutation or by dominant inheritance from asymptomatic carriers of such a mutation. To verify this hypothesis, we estimated the distribution of ASD-risk among both catchment area-based families and multiplex families. Our results suggest that the models with more than three risk components are preferable to the two-component model. Our results also suggest that the largest proportion of ASD cases is caused by dominant inheritance. We additionally show that Supplementary information regarding prevalence has a crucial role in analyzing proband-ascertained data.
Journal of Human Genetics (2009) **54,** 721–726; doi:10.1038/jhg.2009.105; published online 6 November 2009

Keywords: autism spectrum disorders (ASDs); mixture distribution; prevalence; sibling recurrence risk

INTRODUCTION

The DSM-IV category of Pervasive Developmental Disorders, also termed Autism Spectrum Disorders (ASDs) (MIM 209850), includes autism as well as Pervasive Developmental Disorders Not Otherwise Specified and Asperger's disorder. There is evidence that ASDs are highly heritable,[1–3] although ASDs show wide clinical variability and a heterogeneous genetic architecture.[4,5] Dizygotic and sibling concordance rates are about one-tenth of monozygotic concordance rates, suggesting that ASDs are attributable to complex multigenic interactions rather than to a single susceptibility gene.

The great majority of identified ASD genes, mostly copy number variations (CNVs), show an unexpectedly high frequency of *de novo* mutation.[5–10] The increasing number of distinct, individually rare genetic causes of ASDs suggest an alternative to the polygenic hypothesis; most cases of ASDs are due to *de novo* mutations in the parental germ line, which can cause ASDs in most individuals. However, resistant individuals, mostly female, can be relatively asymptomatic carriers yet transmit the mutation and the resulting disorder, in a nearly dominant fashion. This hypothesis, which hereafter we refer to as the two-component model (in contrast to the classical polygenic threshold model[11]), was proposed by Zhao *et al.* (2007).[12] These authors analyzed the susceptibility risk for ASDs in multiplex families using data collected mainly by the Autism Genetic Resource Exchange (AGRE) consortium[13] and found only two types of ASD families. One type of families, which comprised approximately 99% of the sample, had a low risk (slightly less than 0.01) of producing a child with ASD. The second type of families, comprising approximately 1% of the sample, had a high risk (about 0.5) of producing a child with ASD. Although these results support a two-component model, the sample analyzed was originally designed for linkage analyses and was not systematically ascertained. This ascertainment scheme could bias the genetic composition of the ASD families.

Previously, we conducted a cohort study of siblings thoroughly ascertained through at least one ASD proband in the catchment area.[14] Census data of children in the same area has also become available. Therefore, we undertook this study to estimate the distribution of ASD risk within families in this area based on both our sample and the census data. We hoped to verify the two-component model with this study. In addition, to compare the estimates from our analysis and those from Zhao *et al.* (2007)[12] we also analyzed the same dataset used in their study, which was collected by the AGRE.

MATERIALS AND METHODS

Subjects

The sample used for this study, with informed consent and Institutional Review Board approval, has been described in detail elsewhere.[14] Briefly, subjects in this

[1]Clinical Trial Management Center, Nagoya City University Hospital, Nagoya, Japan; [2]Doctor of Public Health Program in Biostatistics, National Institute of Public Health, Wako, Japan; [3]Department of Information and Biological Sciences, Graduate School of Natural Sciences, Nagoya City University, Nagoya, Japan; [4]Department of Pediatrics, Nagoya Western Rehabilitation Center for Children with Disabilities, Nagoya, Japan and [5]Department of Agricultural and Environmental Biology, Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan
Correspondence: Dr T Nishiyama, Department of Information and Biological Sciences, Graduate School of Natural Sciences, Nagoya City University, Yamanohata 1, Mizuho-cho, Mizuho-ku, Nagoya 467-8601, Japan.
E-mail: nishiyama@minos.ocn.ne.jp
Received 21 July 2009; revised 17 September 2009; accepted 2 October 2009; published online 6 November 2009

sample were siblings born between 1993 and 2004, who were ascertained through at least one proband affected by ASD, and living in the western region of Nagoya city (this region is administered by the West District Care Center for Disabled Children). In this catchment area, all children with ASDs were ascertained through the regional screening system, which consists of a three-stage system of health check-ups and also captures missed cases through referrals from kindergartens, nursery schools, clinics and hospitals. Given that, during the study, the average participation rates for health check-ups were 95.3% for 18-month-old children and 86.5% for 3-year-old children, and given that 99.7% of the infants in the catchment area attended kindergartens or nursery schools, it is likely that most infants with developmental problems were identified. Thus, the screening for ASDs could be considered thorough.

A consensus diagnosis of ASD, based on the DSM-IV criteria, was made on the basis of all available information prepared in a semi-structured case vignette. This information included medical examination, psychological assessment and a clinical report based on repeated observations by psychologists and pediatric psychiatrists at ages 4 years or above. Inter-rater reliability was assessed by comparing diagnoses made by two raters based on data from 27 subjects with names and ages removed. The kappa coefficient between the two diagnosticians was 0.70 for both ASD and non-ASD cases.

To compare the estimates from our analysis and those from Zhao *et al.* (2007)[12] we also analyzed the same dataset used in their study, which was collected by the AGRE.

A summary of the dataset, including the total number of children, number of affected children and the estimated sibling recurrence risk, is shown in Table 1. Here, sibling recurrence risk was estimated using the proband method.[15]

Statistical models

The likelihood of sample-only data. We assumed that a mother–father pair in each family had a characteristic and time-invariant risk, x, of producing a male offspring with ASD. We set the risk of producing a female offspring with ASD equal to $p\times x$, where p (female penetrance) represents the factor by which the risk for female offspring is greater than for male offspring. As ASDs are approximately four times more common in males than females, female penetrance, p, ranges from zero to one. Let q_m *and* q_f represent the proportions of males and females in the population, respectively, the values of which were 6949/13 568 and 6619/13 568 in the catchment area. Then the probabilities of producing a male child with ASD and a male child without ASD are given by $q_m\times x$ and $q_m\times(1-x)$, respectively. Similarly, the probabilities of producing a female child with ASD and a female child without ASD are given by $q_f\times(px)$ and $q_f\times(1-px)$, respectively. Note that these probabilities sum to one, that is, $q_m\times x+q_m\times(1-x)+q_f\times(px)+q_f\times(1-px)=1$ because $q_m+q_f=1$. Let $\mathbf{n}=(n_{\mathrm{AM}}, n_{\mathrm{UM}}, n_{\mathrm{AF}}, n_{\mathrm{UF}})$ represent the number of affected males (n_{AM}), unaffected males (n_{UM}), affected females (n_{AF}) and unaffected females (n_{UF}) in a family. If we assume only one risk of producing an affected child, x, the probability that a family with $\mathbf{n}$ children is given by multinomial distributions, $f(\mathbf{n}|x, p)$ as:

$$f(\mathbf{n}|x,p)=\binom{n}{n_{\mathrm{AM}}\ n_{\mathrm{UM}}\ n_{\mathrm{AF}}\ n_{\mathrm{UF}}}(q_m x)^{n_{\mathrm{AM}}}(q_m(1-x))^{n_{\mathrm{UM}}}\times(q_f px)^{n_{\mathrm{AF}}}(q_f(1-px))^{n_{\mathrm{UF}}}$$

Table 1 Sample characteristics

	Our sample	*AGRE sample*
Number of families	269	382
Number of children	510	1340
Male	337	887
Female	173	453
Number of affected children	293	864
Male	232	676
Female	61	188
Sibling recurrence risk[a]	0.183	0.536

[a]Sibling recurrence risk was estimated using the proband method.

where the first term on the right side represents $n!/n_{\mathrm{AM}}!\ n_{\mathrm{UM}}!\ n_{\mathrm{AF}}!\ n_{\mathrm{UF}}!$, and n is the number of siblings in a family, thus $n=n_{\mathrm{AM}}+n_{\mathrm{UM}}+n_{\mathrm{AF}}+n_{\mathrm{UF}}$.

As suggested by Zhao *et al.* (2007),[12] the family population may consist of more than one family subpopulation, each of which may have a different risk of producing an affected child. In this case, it is desirable to model the distribution of $\boldsymbol{n}$ as a mixture of K components. For $i=1, \ldots, K$, the parameter denoting the proportion of the family subpopulation, i, is a_i, with $\sum_{i=1}^{K} a_i = 1$. Therefore, the distribution of n is given by:

$$P(\mathbf{n}|x_i, a_i, p)=\sum_{i=1}^{K} a_i f(\mathbf{n}|x_i, p) \qquad (1\leqslant i\leqslant K)$$

For the two-risk component example, when the first type of family with risk x_1 has the proportion of a_1 in the population and the second type of family with risk of x_2 has the proportion of a_2 in the population,

$$\begin{aligned}P(\mathbf{n}|x_1, x_2, a_1, a_2, p) = {} & a_1\binom{n}{n_{\mathrm{AM}}\ n_{\mathrm{UM}}\ n_{\mathrm{AF}}\ n_{\mathrm{UF}}}(q_m x_1)^{n_{\mathrm{AM}}}(q_m(1-x_1))^{n_{\mathrm{UM}}}\\ & \times(q_f px_1)^{n_{\mathrm{AF}}}(q_f(1-px_1))^{n_{\mathrm{UF}}}\\ & +a_2\binom{n}{n_{\mathrm{AM}}\ n_{\mathrm{UM}}\ n_{\mathrm{AF}}\ n_{\mathrm{UF}}}(q_m x_2)^{n_{\mathrm{AM}}}(q_m(1-x_2))^{n_{\mathrm{UM}}}\\ & \times(q_f px_2)^{n_{\mathrm{AF}}}(q_f(1-px_2))^{n_{\mathrm{UF}}}\end{aligned}$$

Based on the assumption that the individual families are independent and the diagnoses of children within a family are also independent, a log-likelihood function $\mathrm{LL}_{\mathrm{sample}}(\theta)$ of our sample, ascertained from at least one case, is given by:

$$\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})=\sum_{i=1,\ m\geqslant 1}^{N}\mathrm{obs}(\mathbf{n}_i)\log P(\mathbf{n}_i|m\geqslant 1) \qquad (1)$$

where $\mathrm{obs}(\mathbf{n}_i)$ is the observed number of families with $\mathbf{n}_i$ children, m is the number of affected children in a family ($m=n_{\mathrm{AM}}+n_{\mathrm{AF}}$), N is the number of families in the sample and the parameters are $\boldsymbol{\theta}=(x_i, a_i, p)$ $(1\leqslant i\leqslant K)$. For the AGRE sample, we replaced the conditional in equation (1) with $m\geqslant 2$ to reflect the ascertainment procedure of this sample.

Under this model, the prevalence of ASDs, R, is given by:

$$R=(q_m+q_f p)\sum_{i=1}^{K} a_i x_i \qquad (2)$$

and the sibling recurrence risk of ASDs, S, is given by:

$$S=(q_m+q_f p)\sum_{i=1}^{K} a_i x_i^2 \Big/ \sum_{i=1}^{K} a_i x_i \qquad (3)$$

The notation and modeling of ASD-risk described above is identical to that used in the previous study.[12] Details are described in the Supplementary information.

Including the Supplementary information on prevalence

Although Zhao *et al.* (2007)[12] used these equations (2) and (3) as constraints to estimate the parameters, this method does not take into account the statistical uncertainty accompanied by the estimation of R and S. Therefore, instead of using these as constraints, we incorporated the prevalence information into the log-likelihood function $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$, based on a binomial model as:

$$\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})=\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})+\log\binom{N_1}{M_1}R^{M_1}(1-R)^{N_1-M_1}$$

where N_1 and M_1 refer to the number of age-matched children (including affected children) and the number of affected children in the catchment area, respectively, the values of which were 13 568 and 281. This equation for $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ implies that the first term, $\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})$, contains information regarding the sampled population, and the second term contains information regarding the prevalence of ASDs, which can be supplied from the out-of-sample dataset. For the AGRE sample, the equation of $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ is also justified. Details are described in the Supplementary information section.

Identifiability on the number of risk components and model selection

We employed a Bayesian framework to estimate parameters. For the finite mixtures of multinomial distributions, a restriction on the number of components, K, and the maximum number of siblings in a family, n, was imposed because of the identifiability (that is, when the mixture has exactly one representation). This restriction is given by $n \geqslant 2K-1$, where $n=5$ in our sample and $n=8$ in the AGRE sample.[16] Therefore, we considered the model up to three components ($K=3$) in our sample and up to four components ($K=4$) in the AGRE sample.

Finally, as the limit of the discrete distribution with increasing number of components, we examined a continuous risk model by using a beta distribution, $Be(\alpha, \beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$, as:

$$f(\mathbf{n}|\alpha, \beta, p) = \int_0^1 \binom{n}{n_{\mathrm{AM}}\ n_{\mathrm{UM}}\ n_{\mathrm{AF}}\ n_{\mathrm{UF}}} (q_m x)^{n_{\mathrm{AM}}} (q_m(1-x))^{n_{\mathrm{UM}}} \times (q_f p x)^{n_{\mathrm{AF}}} (q_f(1-px))^{n_{\mathrm{UF}}} \times Be(\alpha, \beta)\mathrm{d}x$$

where $B(\alpha, \beta)$ is the beta function.

The prior distributions for x_i, a_i $(1 \leqslant i \leqslant k)$ and p were all assumed to be uniform on [0, 1], where $\sum_{i=1}^{k} a_i = 1$ and $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_k$ for the identifiability of mixture models. We used a Markov Chain Monte Carlo method for estimation (details are described in the Supplementary information). To show how well a statistical model fits the observations, the deviance information criteria (DIC) was calculated from Markov Chain Monte Carlo samples.[17] DIC is defined as the posterior mean deviance, $\bar{D}$, plus the 'effective number of parameters', p_D, where D is the deviance of the model, $-2\times$log-likelihood. Usually, p_D is computed from the difference between $\bar{D}$ and the deviance at the posterior mean parameter estimates, $D(\hat{\theta})$. However, in the finite mixture model, p_D can often be negative because the overdispersion in mixture models leads to $D(\hat{\theta}) > \bar{D}$. Therefore, as proposed by Gelman *et al.* (2004),[18] we computed p_D from half the posterior variance of the deviance.

RESULTS

Risk components

We first examined the parameters in our sample and the AGRE sample, based on the log-likelihood, $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$. Table 2 shows the posterior means (s.d.) of the parameters. The first column refers to the model type by the number of components, x_i refers to the risk in the ith family type, and a_i refers to the proportion of families with risk x_i. The second column refers to DIC as a model choice criterion. For model fitting, the two-component model resulted in a substantially poor fit to the data from both samples compared with the other models. Irrespective of the model, the female penetrance, p, had posterior means between 0.28 and 0.31, with a narrow s.d. (approximately 0.03) in both samples.

The columns $a_i x_i/R$ refer to the means (s.d.) of the proportions of ASD cases contributed from the risk x_i, where R is calculated from equation (4). Note that this is the proportion of ASD cases, not the proportion of families. Based on each model, the higher risk component showed larger proportions of ASD cases. For example, based on the three-component model, the highest risk component (x_3) had the largest proportion of ASD cases, followed by the intermediate risk component (x_2), then by the lowest risk component (x_1).

Next, we examined the continuous risk model by using a beta distribution with parameters α and β. In Table 2, the bottom rows show the posterior means (s.d.) of the parameters and the DIC values. This continuous model clearly shows higher DIC values than any discrete model.

Lastly, we examined the discrete risk models, which constrained the highest risk to 0.5, forcing the assumption that the highest risk corresponds to a dominant risk. This model had slightly poor fit compared with the model without the constraint. The results of the three-component model are shown in Figure 1 and the results based on $\mathrm{LL}_{\mathrm{sample}}$ are shown in the Supplementary Table A.

Information content of the sample and the supplementary data on prevalence

Next, we examined the ascertainment bias by comparing the estimates of prevalence, $\hat{R}$, and sibling recurrence risk, $\hat{S}$, with the expected values from the census data. The values of $\hat{R}$ and $\hat{S}$ obtained from each model are shown in Table 3. The $\hat{R}$ values estimated from the log-likelihood $\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})$, which did not include information on prevalence, were far from the MLE of 2.2%. Of course, the estimates of prevalence, $\hat{R}$, based on the log-likelihood $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ were essentially identical to 2.2%, as indicated by the shaded cells in Table 3 because this likelihood included information on prevalence and constrained $\hat{R}$ to be equal to the MLE. In contrast, the estimates of sibling recurrence risk, $\hat{S}$, based on $\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})$ and $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ were both close to the estimate of 18.3%, which was based on the proband method,[15] although these likelihoods did not accommodate the information on sibling recurrence risk. This finding indicates that our sample contains sufficient information regarding sibling recurrence risk but not prevalence.

To compare the estimates from our sample, we analyzed the data used by Zhao *et al.* (2007) (AGRE sample), and the results are also shown in Tables 2 and 3. The $\hat{R}$ and $\hat{S}$ values obtained from the AGRE sample based on $\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})$ were also far from 2.2 and 18.3%, respectively. However, the discrepancies between these figures and the estimates based on $\mathrm{LL}_{\mathrm{sample}}(\boldsymbol{\theta})$ were much larger than the differences from our sample. As in our sample, the estimates of prevalence, $\hat{R}$, based on the log-likelihood $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ were essentially identical to the MLE of 2.2%, as indicated by the shaded cells in Table 3. However, incorporation of the prevalence information based on $\mathrm{LL}_{\mathrm{sample+supp}}(\boldsymbol{\theta})$ resulted in a marked change in the estimates of sibling recurrence risk, $\hat{S}$, in the AGRE sample. This finding may indicate that the AGRE sample, ascertained by more than two probands, does not contain sufficient information regarding prevalence or sibling recurrence risk.

DISCUSSION

Inspired by a previous study suggesting only two-risk components of ASDs, here we verified this finding using both the dataset previously analyzed by Zhao *et al.* (2007)[12] and our independently collected dataset. The conclusion of Zhao *et al.* (2007)[12] is mainly based on the finding that the risk estimates of male children show one high-risk component, which is near 50%, and two low-risk components, which are both below 1% and are essentially indistinguishable. Because their analysis was based on the MLE method, under some constraints (equations (2) and (3) in this paper), their analysis of male children alone is restricted up to the three-component model, and the analysis of both male and female children is restricted up to only the two-component model, because of limited degrees of freedom. Instead of the MLE framework, we employed a Bayesian framework, allowing us to consider the models for both male and female children up to three components in our sample and up to four components in the AGRE sample. Our results show that the estimates of ASD risks are not divided into two parts, one near 50% and one below 1%. Using models with more than three components, we find intermediate risks ranging from 5 to 30% in both samples. Furthermore, our results also demonstrated that the two-component model resulted in a substantially poor fit to the data compared with the other models. Therefore, we can conclude that the models with more than three risk components are preferable to the two-component model.

Table 2 Risks, family proportions and case proportions contributed from each risk

(a) Our sample

Model[a]	*DIC*[b]	x_1	a_1	X_2	a_2	x_3	a_3	a_1x_1/R	a_2x_2/R	a_3x_3/R
2	696.098	0.013 (0.007)	0.946 (0.038)	0.477 (0.187)	0.054 (0.038)			0.392 (0.209)	0.608 (0.209)	
3	597.654	0.010 (0.005)	0.730 (0.248)	0.114 (0.130)	0.237 (0.246)	0.561 (0.189)	0.033 (0.026)	0.231 (0.176)	0.323 (0.204)	0.446 (0.198)
3 dominant	627.146	0.010 (0.005)	0.740 (0.258)	0.123 (0.127)	0.231 (0.255)	0.5	0.029 (0.001)	0.240 (0.159)	0.313 (0.186)	0.447 (0.142)

Model	*DIC*	α	β
Continuous	723.790	0.102 (0.029)	3.067 (0.852)

(b) AGRE sample

Model	*DIC*	x_1	a_1	X_2	a_2	x_3	a_3	x_4	a_4	a_1x_1/R	a_2x_2/R	a_3x_3/R	a_4x_4/R
2	1694.762	0.016 (0.008)	0.968 (0.017)	0.504 (0.052)	0.032 (0.017)					0.499 (0.238)	0.501 (0.238)		
3	1595.250	0.011 (0.007)	0.754 (0.243)	0.138 (0.136)	0.223 (0.241)	0.565 (0.092)	0.023 (0.014)			0.260 (0.206)	0.357 (0.202)	0.384 (0.197)	
3 dominant	1599.467	0.011 (0.006)	0.813 (0.232)	0.174 (0.158)	0.157 (0.232)	0.5	0.030 (0.011)			0.289 (0.194)	0.238 (0.176)	0.473 (0.172)	
4	1608.127	0.008 (0.006)	0.664 (0.272)	0.071 (0.082)	0.247 (0.240)	0.241 (0.158)	0.072 (0.102)	0.607 (0.097)	0.017 (0.011)	0.177 (0.164)	0.241 (0.171)	0.275 (0.181)	0.307 (0.165)

Model	*DIC*	α	β
Continuous	1822.240	0.073 (0.010)	2.163 (0.269)

The model represents each discrete model by the number of risk components and represents the continuous model as 'continuous'. The model, '3 dominant' means the three -component model that constrains the highest risk to a dominant risk, 0.5. DIC represents the deviance information criteria. x_i refers to the risk in the ith family type, and a_i refers to the proportion of families with risk x_i. a_ix_i/R refers to the proportion of the autism spectrum disorders (ASDs) contributed by risk x_i. α and β represent the parameters of the beta distribution used in the continuous model. The posterior mean of each term (x_i, a_i, a_ix_i/R, α and β) is shown with the s.d. given in parentheses. Shaded cells indicate that the highest risk (x_3) is constrained to 0.5.

[a]The estimates based on our sample.

[b]The estimates based on AGRE sample.

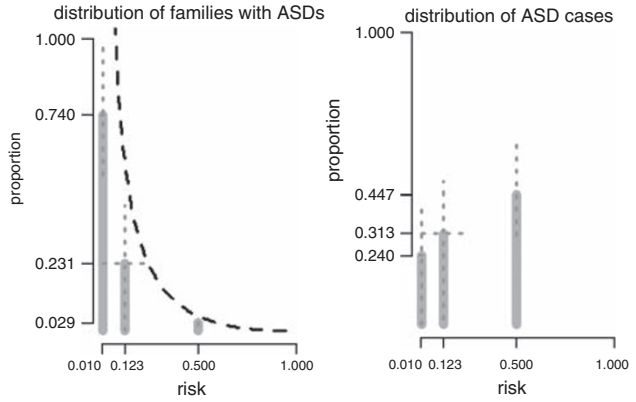

Figure 1 Highlighted estimates for our sample based on the three-component model, which constrained the highest risk to 0.5. (left) The proportions of families with each risk (0.010, 0.123 and 0.500) are represented by gray lines. Highest density regions (HPDs (80%)) are represented by dotted lines. The continuous risk estimated using the beta distribution is represented by a dashed line. (right) The proportions of ASD cases contributed from each risk (0.010, 0.123 and 0.500) are represented by gray lines. HDRs (80%) are represented by dotted lines.

Table 3 Estimates of prevalence and sibling recurrence risk based on different models

	LIsample		*LLsample+supp*	
Model	*R*	*S*	*R*	*S*
(a) Our sample				
2	0.137 (0.048)	0.181 (0.031)	0.021 (0.001)	0.172 (0.030)
3	0.131 (0.044)	0.184 (0.029)	0.021 (0.001)	0.173 (0.030)
Continuous	NA		0.021 (0.001)	0.176 (0.030)
(b) AGRE sample				
2	0.237 (0.070)	0.280 (0.040)	0.021 (0.001)	0.166 (0.066)
3	0.223 (0.064)	0.277 (0.037)	0.021 (0.001)	0.171 (0.056)
4	0.216 (0.059)	0.276 (0.034)	0.021 (0.001)	0.172 (0.052)
Continuous	NA		0.021 (0.001)	0.203 (0.016)

The model represents each discrete model by the number of risk components and represents a continuous model as 'continuous'. LLsample represents estimates based on the information content of only the sample data (LL is log-likelihood). Similarly, LLsample+supp represents estimates based on the sample and the supplementary information about the prevalence. *R* refers to the prevalence, and *S* refers to the sibling recurrence risk. The posterior means of *R* and *S* are shown for each model with the s.d. given in parentheses. The Markov Chain Monte Carlo method is not feasible for LLsample of the beta model, because the detail balance condition is not satisfied, as indicated by 'NA'. Shaded cells indicate that the estimates of prevalence are essentially identical to the MLE of 2.2% because $LL_{sample+supp}$ constrained the estimates to be equal to the MLE.
[a]The result from our sample.
[b]The result from AGRE sample.

Although the estimated risks themselves cannot be assumed to be genetic, based on twin studies it is reasonable to assume that the risks are mostly genetic in origin.[1–3] Thus, from a genetic point of view, we can interpret families with the highest risk, close to 50%, as transmitting ASDs in a dominant pattern. The risk estimates for our sample under the model, which constrained the highest risk to 0.5 are shown in Figure 1. Here, it is worth noting that the highest risk component, which can be regarded as a dominant risk, was associated with the largest proportion of ASD cases in any component model. With regard to the substantial contribution from a dominant risk of ASDs, our finding is in agreement with Zhao *et al.* (2007).[12]

Previous twin studies have suggested that the polygenic factors responsible for ASDs may also be responsible for more common social impairments in the general population, the severity of which falls below the threshold for categorical ASD diagnosis.[19,20] Moreover, autosomal recessive genes responsible for autism have been identified by homozygosity mapping in consanguineous pedigrees.[21] The risks from polygenic or recessive genetic factors is included in an intermediate risk class. Therefore, it is likely that more than a three-component model with intermediate risks is preferable to the two-component model, which does not contain intermediate risks.

In addition, our results show the importance of adding information regarding prevalence to the analysis of ASD risk among families selected through affected probands. Any differences between the results obtained with and without incorporating the information on prevalence reflect the characteristics of the sample determined by the particular ascertainment procedure. Incorporation of the prevalence information resulted in a marked change in the estimates of sibling recurrence risk in the AGRE sample. However, the estimates of sibling recurrence risk showed strong stability to the incorporation of prevalence information in our sample. This finding may indicate that our sample, which was thoroughly ascertained by more than one proband, contains sufficient information regarding sibling recurrence risk, whereas the AGRE sample, which was ascertained by the presence of more than two probands, does not. This conclusion is supported by the extremely high sibling recurrence risk estimate of 0.536 in the AGRE sample, which suggests that the sample overly contains numerous carrier parents that transmit ASDs in a nearly dominant pattern. Therefore, for the analysis of the proband-ascertained sample, we can empirically justify the incorporation of the prevalence information.

The results of this report should be interpreted in the context of several potential limitations. First, we assume that the distribution of the number of offspring among nuclear families is independent of the risk for ASDs. However, this assumption may be violated because parents may choose to stop having children after the birth of a child with ASD. This situation, referred to as stoppage, can severely bias the estimates of the ASD-risk distribution. Without knowledge of the distribution of the number of offspring among families, a correction for stoppage appears to be almost intractable.[22] Therefore, we tested for the existence of stoppage by using the Mann–Whitney *U*-test, according to a previous study.[23] In brief, if U is the number of times a normal child precedes an affected child in all k sibships, a_i is the number of affected children in sibship i, and n_i is the number of normal children in sibship i, then,

$$z = \frac{U - \sum_{i=1}^{k} a_i n_i/2}{\sqrt{\sum_{i=1}^{k} a_i n_i (a_i + n_i + 1)/12}}$$

is a unit normal deviate. Mann–Whitney *U*-tests on our sample and the AGRE sample resulted in z-values equal to 1.70 (P-value=0.058) and 1.65 (P-value=0.099), respectively. This test suggests that this potential bias seems unlikely.

Second, our findings largely depend on the approach of adjusting for ascertainment, which uses the conditional distribution of the phenotype of non-probands, given the phenotype of probands. This method is attractive because it does not necessitate correctly modeling the ascertainment process. For singly ascertained data, conditioning on probands should provide an asymptotically unbiased estimator. However, this is not true for multiple ascertainment, that is, ascertaining through multiple probands in each family.[24] Thus, serious asymptotic bias can occur when adjusting for at least two affected children in the AGRE sample.

Third, as in previous work,[12] the analysis of the AGRE sample is based on an extrapolation of the prevalence information. In contrast,

the analysis of our sample uses prevalence information derived from the same data source. Thus, our sample has higher internal validity than the AGRE sample. That is, our sample should be well specified in terms of the geography and time covered, as well as ethnic group (all subjects in our sample are Japanese), as compared with the AGRE sample. For sensitivity analysis of the AGRE sample, we have explored ranges of prevalence, R, from 0.5 to 2.2% and found that the DIC values were insensitive to changes in R (Supplementary Table B).

Forth, we assume that the same female penetrance, p, even for different risk components is based on the previous study.[12] Although this assumption allows us to avoid the increase in the number of parameters to be estimated, this assumption may be too restrictive to be justified.

The final limiting issue concerns the restriction of the number of risk components incorporated into the statistical model. For the finite mixtures of multinomial distributions, the maximum number of siblings, n, constrains the number of components up to $(n+1)/2$, due to identifiability. Under normal conditions, in which the maximum number of siblings in a sample is limited (for example, five to eight), the number of components is capped at three to four. Therefore, even if a 'true' model comprises of components beyond the capped number, we cannot evaluate the model. Even in that case, it is inferred from the estimated results below the capped number that dominant genes substantially contribute to the development of ASDs. On the other hand, we can consider many more components (for example, 30 or more) by approximating discrete functions by the continuous function, as previously shown. In our results, this model was not supported in terms of DIC. Therefore, we conclude that a limited number of risks are involved in producing children with ASDs.

Despite these limitations, which are largely related to the AGRE data, the estimates from our sample are in remarkable agreement with the estimates from the AGRE sample, suggesting that our results are robust. From our results, the largest risk (dominant risk) can cause the largest proportion of ASD cases in any component model. Recent studies have revealed that submicroscopic CNVs can have a role in ASDs, and the frequencies of 7–10% are observed in simplex families.[6,10] However, CNVs in asymptotic carriers have not yet been fully identified.[10] From our results, we predict that the frequency of any kind of mutations being transmitted from carrier parents will increase significantly, once higher resolution genome-scanning methods become available. The identification of *de novo* CNVs associated with ASDs has progressed considerably in recent years, but detection of mutations transmitted from parents, through examining parent–offspring trios, should become increasingly critical.

ACKNOWLEDGEMENTS

We thank Dr Toshiro Tango and Dr Kunihiko Takahashi for helpful discussions and suggestions. We gratefully acknowledge the resources provided by the AGRE Consortium, the participating AGRE families and the participating families in the Nagoya city sample.

1 Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol. Med.* **25,** 63–77 (1995).
2 Le Couteur, A., Bailey, A., Goode, S., Pickles, A., Robertson, S., Gottesman, I. *et al.* A broader phenotype of autism: the clinical spectrum in twins. *J. Child Psychol. Psychiatry* **37,** 785–801 (1996).
3 Taniai, H., Nishiyama, T., Miyachi, T., Imaeda, M., Sumi, S. Genetic influences on the broad spectrum of autism: study of proband-ascertained twins. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B,** 844–849 (2008).
4 Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D., Hallmayer, J. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65,** 493–507 (1999).
5 Szatmari, P., Paterson, A. D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X. Q. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39,** 319–328 (2007).
6 Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445–449 (2007).
7 Jacquemont, M. L., Sanlaville, D., Redon, R., Raoul, O., Cormier-Daire, V., Lyonnet, S. *et al.* Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J. Med. Genet.* **43,** 843–849 (2006).
8 Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
9 Kumar, R. A., KaraMohamed, S., Sudi, J., Conrad, D. F., Brune, C., Badner, J. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17,** 628–638 (2008).
10 Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82,** 477–488 (2008).
11 Falconer, D. S. *Introduction to Quantitative Genetics* (Oliver & Boyd, Edinburgh, 1960).
12 Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D. H., Law, K. *et al.* A unified genetic theory for sporadic and inherited autism. *Proc. Natl. Acad. Sci. USA* **104,** 12831–12836 (2007).
13 Geschwind, D. H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P. *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69,** 463–466 (2001).
14 Sumi, S., Taniai, H., Miyachi, T. & Tanemura, M. Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in Nagoya, Japan. *J. Hum. Genet.* **51,** 518–522 (2006).
15 Sham, P. C. *Statistics in Human Genetics* (Arnold Publications, London, 1998).
16 Elmore, R. & Wang, S. *Identifiability and Estimation in Finite Mixture Models With Multinomial Components* (Technical Report, Department of Statistics, Pennsylvania State University: PA, USA, 2003).
17 Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. Bayesian measures of model complexity and fit. *J. Roy Statist. Soc. B* **64,** 583–639 (2002).
18 Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. *Bayesian Data Analysis,* 2nd ed. (Chapman & Hall, London, 2004).
19 Constantino, J. N., Lajonchere, C., Lutz, M., Gray, T., Abbacchi, A. & McKenna, K. *et al.* Autistic social impairment in the siblings of children with pervasive developmental disorders. *Am. J. Psychiatry* **163,** 294–296 (2006).
20 Ronald, A., Happe, F., Price, T. S., Baron-Cohen, S. & Plomin, R. Phenotypic and genetic overlap between autistic traits at the extremes of the general population. *J. Am. Acad. Child. Adolesc. Psychiatry* **45,** 1206–1214 (2006).
21 Morrow, E. M., Yoo, S. Y., Flavell, S. W., Kim, T. K., Lin, Y., Hill, R. S. *et al.* Identifying autism loci and genes by tracing recent shared ancestry. *Science* **321,** 218–223 (2008).
22 Slager, S. L., Foroud, T., Haghighi, F., Spence, M. A. & Hodge, S. E. Stoppage: an issue for segregation analysis. *Genet. Epidemiol.* **20,** 328–339 (2001).
23 Jones, M. B. & Szatmari, P. Stoppage rules and genetic studies of autism. *J. Autism. Dev. Disord.* **18,** 31–40 (1988).
24 Boehnke, M. & Greenberg, D. A. The effects of conditioning on probands to correct for multiple ascertainment. *Am. J. Hum. Genet.* **36,** 1298–1308 (1984).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)