

# A Genome-Wide Association Study of Social and Non-Social Autistic-Like Traits in the General Population Using Pooled DNA, 500 K SNP Microarrays and Both Community and Diagnosed Autism Replication Samples

Angelica Ronald · Lee M. Butcher · Sophia Docherty ·  
Oliver S. P. Davis · Leonard C. Schalkwyk ·  
Ian W. Craig · Robert Plomin

Received: 21 September 2009 / Accepted: 14 October 2009 / Published online: 13 December 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** Two separate genome-wide association studies were conducted to identify single nucleotide polymorphisms (SNPs) associated with social and nonsocial autistic-like traits. We predicted that we would find SNPs associated with social and non-social autistic-like traits and that different SNPs would be associated with social and nonsocial. In Stage 1, each study screened for allele frequency differences in ~430,000 autosomal SNPs using pooled DNA on microarrays in high-scoring versus low-scoring boys from a general population sample ( $N = \sim 400/\text{group}$ ). In Stage 2, 22 and 20 SNPs in the social and non-social studies, respectively, were tested for QTL association by individually genotyping an independent community sample of 1,400 boys. One SNP (rs11894053) was nominally associated ( $P < .05$ , uncorrected for multiple testing) with social autistic-like traits. When the sample was increased by adding females, 2 additional SNPs were nominally significant ( $P < .05$ ). These 3 SNPs, however, showed no significant association in transmission disequilibrium analyses of diagnosed ASD families.

**Keywords** Autistic traits · Genome-wide association · Autism · Microarrays · Heritability · Pooling

Social interaction problems and ‘nonsocial’ behaviors, such as restricted repetitive behaviors, are two core symptoms that define autistic spectrum disorders (ASD). Recent population studies show that autistic-like traits vary dimensionally in the general population (Baron-Cohen et al. 2001; Constantino et al. 2003; Posserud et al. 2006; Ronald et al. 2005; Skuse et al. 2005).

Twin studies have reported that autistic-like traits measured dimensionally in the general population are highly heritable (Constantino and Todd 2000, 2003; Hoekstra et al. 2007; Ronald et al. 2006a, 2005; Scourfield et al. 1999; Skuse et al. 2005). Furthermore, liability threshold models and extreme group twin regression analyses using DeFries–Fulker analysis have demonstrated that ‘extreme’ autistic traits (i.e., the most severely affected 15, 10, 5 or 2% of the population) are also highly heritable and show a similar heritability to autism (Ronald et al. 2006a). This suggests that dimensional measures of autistic-like traits might be genetically related to autistic behaviors at the high (impaired) extreme.

Some of these recent twin analyses of autistic-like traits have also explored the nature of the relationship between the different autistic-like traits that together form the diagnostic criteria. In data collected from parents and teachers on over 3,000 7-year-old pairs in a community twin sample, social and nonsocial autistic-like traits were both found to be highly heritable, but showed only modest genetic overlap (Ronald et al. 2005). The genetic correlation was estimated at 0.2, which suggests that only a small proportion of the genes influencing variation in social and nonsocial traits in the general population were overlapping, with the majority of genetic influences acting specifically on each trait. This finding has since been replicated using a different measure (Ronald et al. 2006a). Modest genetic overlap between social and nonsocial autistic-like traits has

A. Ronald · L. M. Butcher · S. Docherty ·  
O. S. P. Davis · L. C. Schalkwyk · I. W. Craig · R. Plomin  
Social Genetic and Developmental Psychiatry Centre, Institute  
of Psychiatry, De Crespigny Park, London SE5 8AF, UK

A. Ronald (✉)  
Centre for Brain and Cognitive Development, Birkbeck College,  
Malet Street, London WC1E 7HX, UK  
e-mail: a.ronald@bbk.ac.uk

also been reported for autistic-like traits at the impaired extreme (Ronald et al. 2006b).

Family studies have demonstrated that undiagnosed relatives of individuals with autism show sub-threshold traits characteristic of autism (the ‘broader autism phenotype’), suggesting that these behaviors are familial and supporting the notion that these behaviors lie on a continuum of impairment (Bailey et al. 1998). Furthermore, it has been noted that there is some segregation of the phenotype among relatives, that is, often relatives show some but not all autistic-like traits, for example, social difficulties without any nonsocial behaviors or communication problems (Bailey et al. 1998; Bishop et al. 2004; Pickles et al. 2000; Szatmari et al. 2000). These findings from family and twin studies suggest that different causal influences might affect quantitative variation in social and nonsocial autistic-like traits (Happé et al. 2006).

There is also some indirect evidence from linkage studies using diagnosed autism samples that different genetic influences may play a role in different autistic behaviors. For example, in several studies, linkage signals have been shown to increase when families were selected based on particular nonsocial features such as having high scores on insistence on sameness (Shao et al. 2003), obsessive compulsive behaviors (Buxbaum et al. 2004), savant skills (Nurmi et al. 2003), high scores on the restricted repetitive behaviors and interests (RRBI) domain (Sutcliffe et al. 2005), severe compulsive behaviors and rigidity (McCauley et al. 2004) and repetitive behaviors (Alarcon et al. 2002).

Many linkage studies have been carried out for diagnosed autism and nearly every chromosome has been implicated (Abrahams and Geschwind 2008; Sykes and Lamb 2007; Yang and Gill 2007). A previous linkage study that used a quantitative assessment of autistic traits with the Social Reciprocity Scale in 100 multiplex ASD families found linkage signals on chromosomes 11 and 17 (Duvall et al. 2007). The first study to directly test for different linkage regions for social and non-social autistic behaviors has recently been carried out (Liu et al. 2009). In a sample of 2,025 individuals with an ASD, the ADI-R social interaction and the non-social behavior domains correlated .28. Genome-wide linkage analyses were performed separately for these two domains—reciprocal social interaction and restricted repetitive and stereotyped behaviors—but no genome-wide significant linkage signals were found. For complex traits though, linkage is limited to detecting large effects that may reflect a summary of effects over vast genetic distances. For this reason, allelic association, which is more powerful than linkage for detecting quantitative trait loci (QTLs) of small effect size (Risch 2000; Sham et al. 2000), has become the latest hope to unearth causal variants underlying complex traits and disease.

There have been many candidate genes proposed for autism (Abrahams and Geschwind 2008), and, like the linkage studies mentioned above, candidate gene studies have begun to explore the possibility of symptom-specific genetic associations in autism. A good example is the set of studies on the serotonin transporter gene (*SLC6A4*). A recent study reported that subjects with the short version of the serotonin transporter gene promoter polymorphism (5-HTTLPR) (S/L or S/S genotypes) were rated as more severe on a social subdomain “failure to use nonverbal communication to regulate social interaction,” whereas subjects with the long version (L/L genotype) were more severe on a nonsocial subdomain “stereotyped and repetitive motor mannerisms” and on an aggression measure (Brune et al. 2006). Increased severity on social/communication domains in individuals with the short version was also found in an earlier study (Tordjman et al. 2001), and other variants within this gene have also been found to be specifically associated with increased severity on nonsocial domains (Mulder et al. 2005; Sutcliffe et al. 2005).

However a problem with candidate gene studies is their unsystematic nature. Genome-wide association studies (GWAS) provide a solution to this problem: they are highly systematic and are now possible using SNP microarrays (Hirschhorn and Daly 2005). One economical strategy for screening large samples is to pool DNA for groups such as low and high groups on a quantitative trait, which averages allele frequencies biologically for the comparison groups rather than obtaining individual genotypes and averaging them statistically (Darvasi and Soller 1994; Knight and Sham 2006; Norton et al. 2004; Sham et al. 2002).

In the present study we have combined the strengths of microarrays and pooled DNA in a method we call *SNP Microarrays and Pooling* (SNP-MaP). Pooled DNA can be genotyped reliably on microarrays (Butcher et al. 2004; Docherty et al. 2007; Kirov et al. 2006; Meaburn et al. 2005, 2006; Pearson et al. 2007). To our knowledge four SNP GWAS for autism have so far been published, three with positive findings (SNPs identified on chromosome 15p; Ma et al. 2009; Wang et al. 2009; Weiss et al. 2009) and one which identified no genome-wide significant SNPs (Arking et al. 2008).

The purpose of the present study was to undertake the first GWAS of autistic-like traits in the general population, using a dimensional measure of autistic-like behaviors. Our *first* hypothesis, based on the high heritability of both social and nonsocial autistic-like traits, was that SNP associations would be found for both social and nonsocial autistic-like traits. A quantitative trait model has several advantages beyond the practical advantages of using community rather than clinical samples. First, quantitative information about the degree of autistic-like traits may be more informative than categorical information about presence or absence of a

disorder (Abrahams and Geschwind 2008; Duvall et al. 2007). Second, and most pertinent to the current study, a trait approach allows the relationship between the different symptoms within a disorder to be studied independently, for example, the social and nonsocial behaviors that are both key features in the autism diagnosis. Importantly, we conducted two separate GWAS, one for social and one for nonsocial autistic-like traits. The Affymetrix GeneChip Human Mapping 500 K Array Set was employed, and the findings were followed up in a second stage using an independent community sample. Our *second* hypothesis, again based on the findings from twin studies, was that most markers associated with social autistic-like traits would be different from those associated with nonsocial autistic-like traits. Finally, we took advantage of the availability of the Autism Genetic Resource Exchange (AGRE) dataset, and tested, in a third replication stage, whether SNPs found to be associated with autistic-like traits in the general population were also associated with diagnosed ASD. Our *third* hypothesis was that SNPs associated with variation in autistic-like traits would also be associated with diagnosed ASD.

## Materials and methods

### Participants

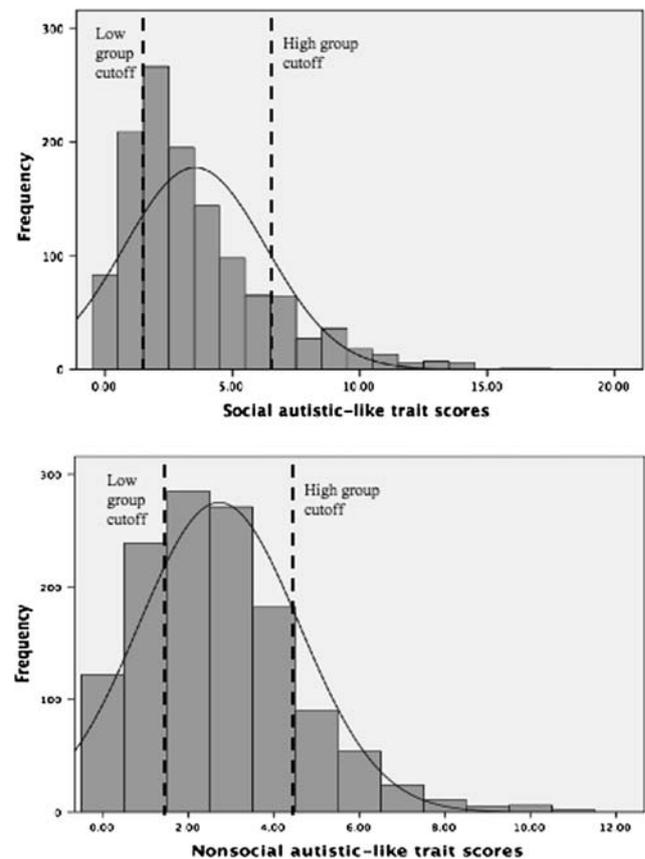
The general population sample came from the Twins Early Development Study (TEDS), a UK-based sample born in England and Wales in 1994–1996 (Oliver and Plomin 2007). Children were excluded who did not have ethnicity information or DNA available. Other exclusion criteria were extreme medical conditions (other than ASDs), severe perinatal difficulties, or non-Caucasian ethnicity. Males only were selected in order to avoid sex differences in the high and low groups in Stage 1. The TEDS sample is reasonably representative of the UK population (Oliver and Plomin 2007). Comparing the TEDS sample that provided data when the twins were age 7 to the General Household Survey (Office for National Statistics 2002), 94 versus 93% were white, 48 versus 50% were male, and 37 versus 32% of mothers had one or more A-level (UK advanced educational qualification). 4% of children in the TEDS 7-year sample had a statement of special educational needs versus 3% of children in England (Department for education and skills 2002).

### Stage 1: SNP-MaP high and low groups

For Stage 1 of the two separate studies of social and nonsocial autistic-like traits, boys were selected at the low and high extremes of the quantitative trait distribution on a measure of social and nonsocial autistic-like traits (Ronald

et al. 2005)—see “Measures” section. Figure 1 shows the distributions of the social and nonsocial autistic-like trait measures and the high and low group cutoff criteria according to raw scores. For the social study, boys were selected if they scored in the most severe 10.5% of the sample for the high-scoring group, and in the lowest (least impaired) 29.9% of the sample for the low scoring group. For the nonsocial study, the equivalent cutoffs for the high and low groups were 13.7 and 25.4%. The choice of cut-off was guided by quantitative genetic research in TEDS showing that heritability of autistic-like traits is high regardless of cutoff (Ronald et al. 2006a) and by statistical genetic simulations that show that such cutoffs balance the power obtained in DNA pooling studies from using extreme cutoffs and from using large samples (Sham et al. 2002). The cutoffs are less extreme for the low groups because of the lack of variation at the low end of the distribution (see Fig. 1).

If both twins fell in the extreme group, the more extreme-scoring child was selected to be included in the high group. In opposite-sex pairs, only the male twin was



**Fig. 1** Histograms showing the distributions of the social (*top figure*) and nonsocial (*bottom figure*) autistic-like trait scales in the male-only unrelated TEDS sample used in Stage 1. *Dotted lines* indicate the cutoffs employed for selecting the high-scoring and low-scoring groups in each study

included. Similarly, in the low group the lowest-scoring twin was selected. Thus children in the high group were unrelated to each other, and the same was true for the low group; moreover, children in the high and low groups were unrelated to each other. The final  $N$  in the high and low social groups were 373 and 372, respectively. The final  $N$  in the high and low nonsocial groups were 434 and 436, respectively. Thirty percent of the children in the high social group were also in the high nonsocial group, and 11% of the children in the low social group were also in the low nonsocial group. If we had excluded the children who appeared in both high groups, our two GWAS would have been studies of social-but-not-nonsocial autistic-like traits and nonsocial-but-not-social rather than separate GWAS of social and nonsocial autistic-like traits, which in fact correlate modestly in the population ( $r_{ph} = .15$ ) (Ronald et al. 2006a; Ronald et al. 2005). Moreover, this slight overlap of subjects in the two studies is conservative in the sense that it works against our hypothesis that we will find different SNPs associated with social and nonsocial autistic-like traits.

#### Stage 2: unselected community sample

An unselected sample was constructed from TEDS children not included in Stage 1, therefore including individuals who were independent of the sample used in Stage 1. Because the sample was unselected, it could be used for both the social and nonsocial studies. The same exclusion criteria were employed as for the initial sample. Initially a male-only sample was used; 1,411 had social autistic-like trait scores and 1,379 nonsocial autistic-like trait scores. This sample was subsequently extended to provide more power in Stage 2 by also genotyping females. The sample  $N$  with males and females with the social and nonsocial scores and DNA was 3,341 and 3,308, respectively. The overlap in individuals between the two studies was 3,155 indicating that a total of 3,494 individuals were genotyped. The community sample replication provides a test of the QTL hypothesis that those SNPs exhibiting allele frequency differences between the low and high parts of the distribution will be associated with variation across the entire quantitative trait distribution.

#### Measures

Teachers rated the twins' autistic-like traits on DSM-IV-based social and non-social scales (Ronald et al. 2005) when the twins were age 7. The questionnaire was designed to include items that were relevant for assessing the types of social and nonsocial behaviors notable in autism but that also would be seen in the general population. The majority of items were derived from the DSM IV autism criteria (see

Ronald et al. 2005). Each item was rated as *Not true* (0), *Somewhat true* (1) or *Certainly true* (2). The social scale had 10 items and therefore a range of 0–20; the non-social scale had 6 items and therefore a range of 0–12. Figure 1 shows the distributions of the social and nonsocial autistic-like trait scales in the male-only unrelated TEDS sample used for selecting high and low groups in Stage 1.

The DSM-IV-based social and non-social scales showed moderate internal consistency with Cronbach's alphas of .72 and .51 for the social and nonsocial scales, respectively. Another index of reliability is the MZ twin correlation because reliability creates an upper limit for MZ twin correlations. MZ twin correlations for both the social and nonsocial scales were moderate to high for teacher ratings (.60–.77).

#### Design and procedures

##### Stage 1 DNA pool construction

A total of 20 biologically independent DNA pools were constructed to represent high and low scorers for both social and nonsocial autistic-like traits (five pools per group, with approximately 74 subjects in each social pool and about 86 subjects in each nonsocial pool). Genomic DNA for each individual was extracted from buccal swabs (Freeman et al. 2003), suspended in EDTA TE buffer (0.01 M Tris-HCl, 0.001 M EDTA, pH 8.0) and quantified in triplicate using PicoGreen<sup>®</sup> dsDNA quantification reagent (Cambridge Bioscience, UK). Upon obtaining reliable triplicate readings ( $\pm 0.5$  ng/ $\mu$ l), equimolar quantities of DNA for each individual were added to their pool. Differing DNA concentrations for the individuals were compensated by adding different volumes for the individuals; the minimum volume added was 1  $\mu$ l to avoid compromise due to pipetting errors. Therefore, the amount of DNA an individual contributed to each group of pools was established as the amount contained within 1  $\mu$ l of the most concentrated individual from that group.

##### SNP microarray allelotyping of pooled DNA

Each of the 20 DNA pools was allelotyped using the GeneChip<sup>®</sup> Mapping 500 K Array set in accordance with the standard protocol for individual DNA samples (see the GeneChip<sup>®</sup> Mapping 500 K Assay Manual for full protocol). Each microarray was scanned using the GeneChip<sup>®</sup> Scanner 3000 with High-Resolution Scanning Upgrade, which was controlled using GeneChip<sup>®</sup> Operating software (GCOS) v1.4. Cell intensity (.cel) files were analyzed using GTYPE. Each of the 20 DNA pools was assayed on a separate microarray set; for quality control checks, a reference DNA individual provided by the manufacturer

(sample number 100103) was also assayed on a separate microarray set.

#### *Generation of SNP-MaP allele frequency estimates*

Relative allele signal (RAS) scores, calculated using the 10 K MPAM Mapping algorithm, have been shown to be reliable and valid indices of allele frequency in pooled DNA (Brohede et al. 2005; Butcher et al. 2004; Craig et al. 2005; Kirov et al. 2006; Liu et al. 2005; Meaburn et al. 2005, 2006; Simpson et al. 2005). Details of how probesets on Affymetrix Mapping GeneChip® microarrays are used to calculate allele frequency estimates are described in Appendix 1. Allele frequency estimates for the 500 K microarray set were calculated manually from the raw probe intensity data exported as a .txt file.

#### *SNP selection*

To screen SNPs in Stage 1 SNP-MaP analysis, we derived a rank-based composite score using five equally weighted criteria. The rationale and derivation of this composite score is presented in Appendix 1 (see also Butcher et al. 2008). Briefly, the five criteria were: (1) greater average allele frequency difference between low and high autistic-like trait groups, (2) smaller average variance of the low and high groups (i.e., variance across the 5 DNA pooled allele frequency estimates for each group), (3) smaller average variance within each microarray i.e., variance across the multiple probe sets that form the microarray's allele frequency estimate (to account for probe-specific errors), (4) greater number of successful replicate pools, and (5) greater minor allele frequency, as indexed by the average of the low and high autistic-like trait groups. The fourth criterion was included because the more data we had from the replicate pools, the more accurate the allele frequency estimates were likely to be. The fifth criterion was included because we had more power to detect common alleles. We used this composite to choose the top SNPs with the highest composite scores in each of the two studies.

#### *Stage 2 individual genotyping of unselected replication sample*

The Stage 2 sample of 3,494 individuals from TEDS were genotyped using Applied Biosystems' SNPlex™ genotyping system and analyzed using *GeneMapper v4.0* software (Applied Biosystems). SNPlex is a capillary electrophoresis-based multiplex genotyping system capable of genotyping up to 48 SNPs per sample per well (Tobler et al. 2005). A SNPlex multiplex was successfully designed for the top 47 SNPs as indexed by the composite score described

above: 23 SNPs for social traits; 24 for nonsocial traits. In addition to the TEDS individuals, DNA from 88 Centre de d'Etude du Polymorphisme Humain (CEPH) individuals who have been genotyped as part of the HapMap Project (The International HapMap Consortium 2003, 2005) were obtained from the Coriell Institute to assess genotyping quality and error rate. Reference genotypes of CEPH individuals for the selected SNPs were downloaded from HapMart, the data mining tool for downloading HapMap data (<http://hapmart.hapmap.org/BioMart/martview>).

SNPs passing quality control (see below) were tested for additive genetic effects (coding genotypes as 0, 1 or 2) using a Pearson correlation ( $r$ ) between additive genotypic values and quantitative trait scores.

#### *Genotyping quality control*

The following sequential criteria were applied for the genotyping quality control: SNPs were omitted from analysis if either poor genotype clusters prevented GeneMapper software from making calls or a SNP showed more than one genotype mismatch between CEPH genotypes deposited in HapMap and those derived using in-house genotyping methods. Individuals were omitted if their SNP call rate was <65% (1 SD below the average). Finally, for each SNP, individual genotypes were omitted if their peak heights were <25% of the average peak height for that genotypic group as measured across the entire sample; we apply this procedure because poor quality samples often exhibit high background noise that SNPlex can mistake as heterozygotes. This leads to an apparent excess of heterozygotes that inflates the number of false positives in Hardy–Weinberg equilibrium tests.

## **Results**

### *Stage 1 results*

SNP-MaP allele frequencies for the 20 DNA pools were calculated and analyzed separately for the social and nonsocial studies. In order to increase the reliability of SNP-MaP allele frequency estimates, we required allele frequency estimates from a minimum of 3 (out of 5) pools for both high and low groups. We also excluded SNPs with minor allele frequencies lower than .05 (according to CEPH allele frequencies from HapMap) as power to detect association in this range is greatly reduced. After these exclusion criteria, the autosomal genome-wide screen consisted of 433,813 SNPs for the social study and 435,457 SNPs for the nonsocial study.

The average allele frequency for the low and high groups was calculated for each SNP. The correlation

between allele frequency estimates for the low and high groups was .986 and .988 for the social and nonsocial SNP-MaP studies respectively, indicating that the rank order of allele frequencies was highly reliable overall—a test analogous to genome control. Accordingly, allele frequency differences between the low and high groups were small: For social autistic-like traits, 76% of SNPs exhibited between-group differences smaller than .05, with a mean between-group absolute difference of .035 (range: .00–.40); for nonsocial autistic-like traits, 78% of SNPs exhibited between-group differences smaller than .05, with a mean between-group absolute difference of .030 (range: .00–.30).

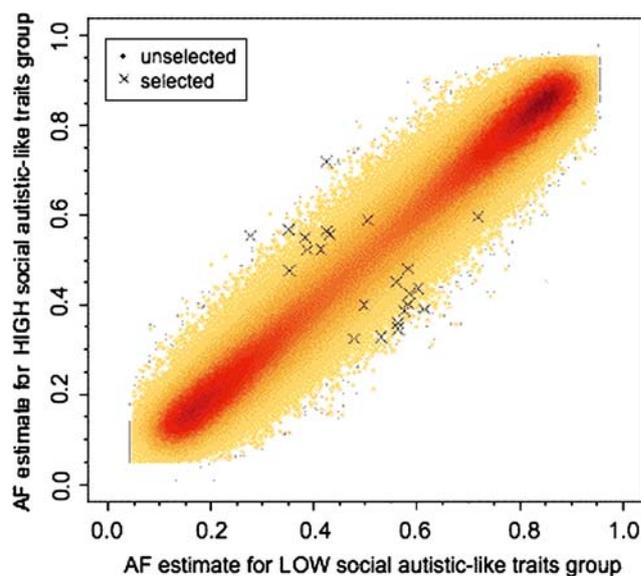
As explained in “[Design and procedures](#)”, Stage 1 was used to screen SNPs on the basis of a ranked composite score which took into account the between-group allele frequency difference, variance between- and within- biological replicate microarrays, number of successfully assayed arrays, and minor allele frequency. Due to financial restrictions, we were limited to genotyping individuals in Stage 2 using a single SNPlex multiplex probe set of 47 SNPs: 23 SNPs in the social study and 24 SNPs in the nonsocial study. These SNPs represent the highest composite scores in Stage 1 for each of the two studies. None of the SNPs selected from the social SNP-MaP study overlapped with those selected from the nonsocial SNP-MaP study. The mean absolute difference between low and high SNP-MaP allele frequency estimates for the 23 social SNPs was .17 (ranging from .08 to .29); for the 24 nonsocial SNPs the mean absolute difference was .16 (ranging from .08 to .26). Figures 2 and 3 place the selected SNPs in the context of the full dataset for the social and nonsocial studies, respectively, by plotting the average allele frequency of the low scoring groups against that of the high (impaired) scoring groups.

### Stage 2 results

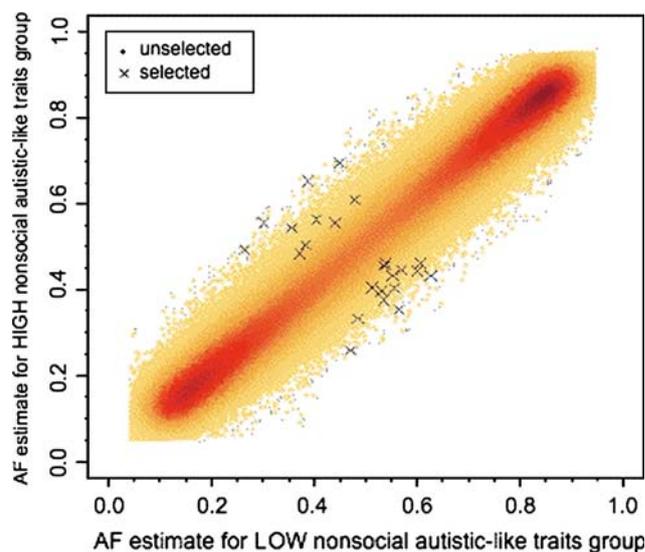
The 47 SNPs nominated in Stage 1 in the two studies were individually genotyped in the unselected sample in order to test the QTL hypothesis directly by assessing the degree to which the SNPs nominated in Stage 1 are associated with quantitative autistic-like traits throughout the distribution. With 23 tests for the social study and 24 tests for the nonsocial study and an alpha of 0.05, one significant result would be expected in each study on the basis of chance alone.

### Individual genotyping quality control

In our SNPlex analysis, three out of 47 SNPs (rs6701037 [selected for social], and rs1546377 and rs7894025 [selected for nonsocial]) exhibited poor call rates across plates due to poor genotype clustering and were omitted



**Fig. 2** Scatterplot showing allele frequencies (AF) of all SNPs for high versus low social groups from Stage 1. The 23 top-ranked SNPs (*crosses*) are shown against the background of 433,813 unselected autosomal SNPs comparing AF for the low (*x*-axis) and high i.e. impaired (*y*-axis) social groups. The figure also displays the density of SNPs as the density map changes from *light* (sparse clusters) though to *dark* (dense clusters). AF differences are small with the majority of differences occurring for SNPs with minor allele frequencies of .10–.25, which reflects the representation of SNPs with these allele frequencies on the Affymetrix microarray. The more extreme deviants (than those selected) failed to meet the 5 selection criteria, as outlined in the “[Design and procedures](#)”



**Fig. 3** Scatterplot showing allele frequencies (AF) of all SNPs for high versus low nonsocial groups from Stage 1. The 24 top-ranked SNPs (*crosses*) are shown against the background of 435,457 unselected SNPs comparing AF for the low (*x*-axis) and the high (*y*-axis) nonsocial groups. The figure also displays the density of SNPs—see Fig. 1 for further details

**Table 1** Number of SNPs and individuals remaining after SNP, sample and genotype quality control procedures

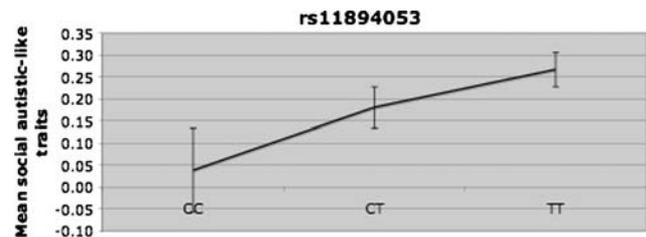
	Social study	Nonsocial study
Number of SNPs genotyped	23	24
Number of SNPs after		
Excluding SNPs with low call rate	22	22
Excluding SNPs with inaccurate calls	22	20
<i>Male-only sample</i>		
Number of males with trait score (Genotypes; % complete)	1,411 (29,430; 94.8%)	1,379 (25,983; 94.2%)
Number of males after	1,356	1,329
Excluding individuals with low call rate	(28,642; 92.3%)	(25,478; 92.4%)
Excluding low quality genotypes	(27,162; 87.5%)	(24,684; 89.5%)
<i>Male and female sample</i>		
Number of males and females with trait score (Genotypes; % complete)	3,341 (69,518; 94.6%)	3,308 (62,421; 94.3%)
Number of males and females after	3,218	3,190
Excluding individuals with low call rate	(67,792; 92.2%)	(61,223; 92.5%)
Excluding low quality genotypes <sup>a</sup>	(64,900; 88.3%)	(59,071; 89.3%)

<sup>a</sup> The number of individuals at this point is the same as the previous filter; only genotypes are filtered out

from further analyses. We omitted two other SNPs (rs6903663 and rs9654873 [both selected for nonsocial]) from analyses because of inaccurate calls, that is, they showed poor concordance between in-house derived genotyping of 88 CEPH individuals and genotypes deposited in HapMap International HapMap (Frazer et al. 2007). Using this comparison between in-house genotyping and genotypes deposited in HapMap, the remaining 42 SNPs (22 SNPs for social, 20 SNPs for nonsocial) displayed genotyping error rates of <1%. These errors were caused by homozygotes being erroneously called as heterozygotes.

55 (3.9%) individuals and 50 (3.6%) individuals showing low call rates (<65%) across SNPs were omitted from the social and nonsocial analyses respectively. We also excluded an additional 4.8 and 2.9% of genotypes from the social and non-social studies respectively, whose peak heights were <25% of the average peak height for that SNP across the study. After excluding the 5 aforementioned SNPs, samples with poor call rates and genotypes with low peak heights, we observed 27,162 (87.5% completeness) and 24,684 (89.5% completeness) genotypes to perform association analysis for the social and nonsocial autistic-like traits, respectively, in the male-only sample. Table 1 outlines the number of SNPs and individuals remaining after SNP, sample and genotype quality control procedures.

Our conservative criteria improved observed genotypic distributions under Hardy–Weinberg equilibrium, tightened genotype clusters in SNPlex, and left the distributions of social and nonsocial autistic-like trait scores unchanged.



**Fig. 4** Genotype-by-phenotype plot for rs11894053 (correlated in the male-only sample) illustrating the effect of genotype (*x*-axis) on standardized social autistic-like trait scores (*y*-axis)

#### Results of Stage 2 in an independent male-only sample

One SNP (rs11894053) in the social study correlated .06 ( $P = .02$ ) in the male-only sample, and no SNPs in the nonsocial study showed nominally significant correlations ( $P < .05$ ) in the male-only sample. Figure 4 plots the results for rs11894053 in terms of standardised mean quantitative trait social score (age- and sex-regressed) for the three SNP genotypes in the male-only sample. The SNP shows an additive pattern. The homozygotes differ by .23 SD. Stage 1 and 2 results for all 42 SNPs are shown in Table 2.

Squaring the correlation of  $r = 0.06$  to estimate effect size indicates that this association accounted for only 0.36% of the variance in teacher-rated social autistic-like traits.

#### Results of Stage 2 including both males and females

Next, we explored which of the SNPs nominated in Stage 1 replicated in the whole sample including females as well as

**Table 2** Summary of Stage 1 (screening) and Stage 2 (testing) results for 22 SNPs in the social autistic-like trait study and 20 SNPs in the nonsocial autistic-like trait study for males only

Study	AffyID	dbSNP RS ID	Chr	Genomic relation	Stage 1 (screening with SNP-MaP)				Stage 2 (testing with individual genotyping)				Correct direction?		
					Strand	Alleles	Low	High	L–H diff	Strand	QTL <i>r</i>	QTL <i>p</i>		N	H–W
Social study	SNP_A-1992893	rs10503871	8p12	Intron (GTF2E2)	–	A/G	0.35	0.57	–0.22	+	0.03	0.15	1241	0.71	Incorrect
	SNP_A-1884519	rs11894053	2p21	Intergenic	–	C/T	0.48	0.32	0.16	+	0.06	0.02	1254	0.87	Correct
	SNP_A-2209521	rs1251584	1p31.1	Intergenic	+	C/T	0.59	0.42	0.17	–	0.02	0.19	1253	0.91	Correct
	SNP_A-2190550	rs12703774	7q34	Intron (SLC37A3)	–	A/G	0.53	0.33	0.20	+	–0.01	0.39	1250	0.86	Correct
	SNP_A-2255602	rs17622673	6q16.3	Intergenic	+	C/T	0.62	0.39	0.23	+	0.02	0.28	1256	0.13	Correct
	SNP_A-1787434	rs17733015	2p23.2	Intron (ITSN2)	–	A/G	0.35	0.48	–0.13	+	0.00	0.49	1219	0.52	Incorrect
	SNP_A-2230316	rs1821383	15q14	Intergenic	+	A/G	0.58	0.39	0.19	+	–0.05	0.03	1252	0.93	Incorrect
	SNP_A-1789039	rs248232	5q35.5	Intergenic	+	C/T	0.58	0.40	0.18	+	0.01	0.42	1181	<0.01	Correct
	SNP_A-2219703	rs343373	1p34.2	Intron (FOXJ3)	–	A/C	0.58	0.48	0.10	–	–0.01	0.33	1260	0.98	Correct
	SNP_A-2194849	rs4671615	2p14	Intergenic	+	G/T	0.43	0.56	–0.13	+	–0.04	0.10	1251	0.93	Correct
	SNP_A-1896445	rs4736702	8q24.22	Intergenic	–	C/G	0.38	0.55	–0.17	+	–0.03	0.17	1272	0.89	Correct
	SNP_A-1861345	rs6562153	13q21.31	Intergenic	–	A/G	0.56	0.45	0.11	+	0.02	0.27	1206	0.48	Correct
	SNP_A-1811839	rs7211602	17q12	Intergenic	+	A/C	0.50	0.59	–0.09	+	–0.01	0.35	1154	0.14	Correct
	SNP_A-2005876	rs7247108	19p13.2	Intergenic	–	C/T	0.29	0.55	–0.26	+	0.01	0.33	1239	0.10	Correct
	SNP_A-2062011	rs7309433	12q13.2	Intergenic	+	A/G	0.41	0.52	–0.11	+	–0.01	0.31	1089	0.13	Correct
	SNP_A-2032325	rs7638018	3q22.1	Intron (TF)	–	C/T	0.39	0.52	–0.13	+	0.02	0.21	1236	0.01	Incorrect
	SNP_A-2077176	rs7758499	6q23.3	Intergenic	+	A/G	0.42	0.72	–0.30	+	–0.03	0.11	1279	0.74	Correct
	SNP_A-2273108	rs7773834	6q22.31	Intergenic	+	C/G	0.72	0.60	0.12	+	0.00	0.45	1261	0.99	Correct
	SNP_A-2002514	rs7973458	12p13.31	Intergenic	+	C/T	0.42	0.56	–0.14	+	0.01	0.37	1278	0.42	Correct
	SNP_A-2310304	rs854633	17q12	Intergenic	+	C/T	0.60	0.44	0.16	–	0.01	0.35	1261	0.04	Correct
	SNP_A-2173622	rs9347088	6q27	Intergenic	–	A/G	0.50	0.40	0.10	+	–0.03	0.16	1243	0.15	Incorrect
	SNP_A-1812640	rs9923476	16q24.1	Intergenic	+	C/G	0.56	0.35	0.21	+	0.01	0.37	1227	0.63	Correct
SNP_A-2080138	rs10836588	11p12	Intergenic	+	C/T	0.36	0.54	–0.18	+	0.03	0.19	1259	0.18	Incorrect	
SNP_A-2243503	rs11218615	11q22.1	Intergenic	–	A/G	0.39	0.50	–0.12	+	–0.01	0.38	1268	0.98	Correct	
SNP_A-2046249	rs12433492	14q11.2	Intron	+	A/C	0.27	0.49	–0.22	+	0.00	0.44	1257	0.27	Correct	
SNP_A-2204779	rs12578517	12p12.3	Intergenic	+	C/T	0.61	0.46	0.15	+	0.03	0.18	1263	0.27	Correct	
SNP_A-4255314	rs12624227	2p25.1	Intergenic	–	A/G	0.57	0.45	0.13	+	0.00	0.44	1031	1.00	Correct	
SNP_A-1976029	rs1565924	4p15.33	Intergenic	–	C/T	0.57	0.35	0.22	+	0.04	0.11	1141	0.03	Correct	
SNP_A-2001794	rs1945427	11p15.1	Intron (NELL1)	–	A/G	0.63	0.43	0.20	–	0.03	0.15	1209	0.01	Correct	
SNP_A-2018270	rs2284550	21q22.11	Intron (IFNAR2)	+	A/G	0.56	0.43	0.13	+	0.01	0.32	1228	1.00	Correct	
SNP_A-2128100	rs236394	6p21.2	Intron (CPNE5)	–	A/T	0.38	0.48	–0.11	–	–0.01	0.35	1266	0.01	Correct	
SNP_A-2255970	rs2400954	12q21.31	Intron (PPFIA2)	+	C/T	0.54	0.46	0.08	+	0.02	0.22	1251	0.57	Correct	
SNP_A-2043020	rs2818309	9q33.1	Intergenic	+	C/T	0.29	0.56	–0.26	–	–0.03	0.18	1270	0.34	Correct	
SNP_A-2015285	rs2822859	21q11.2	Intron	–	C/T	0.54	0.40	0.14	+	0.00	0.47	1246	0.28	Correct	
Nonsocial study	SNP_A-2005876	rs7247108	19p13.2	Intergenic	–	C/T	0.29	0.55	–0.26	+	0.01	0.33	1239	0.10	Correct
	SNP_A-2062011	rs7309433	12q13.2	Intergenic	+	A/G	0.41	0.52	–0.11	+	–0.01	0.31	1089	0.13	Correct
	SNP_A-2032325	rs7638018	3q22.1	Intron (TF)	–	C/T	0.39	0.52	–0.13	+	0.02	0.21	1236	0.01	Incorrect
	SNP_A-2077176	rs7758499	6q23.3	Intergenic	+	A/G	0.42	0.72	–0.30	+	–0.03	0.11	1279	0.74	Correct
	SNP_A-2273108	rs7773834	6q22.31	Intergenic	+	C/G	0.72	0.60	0.12	+	0.00	0.45	1261	0.99	Correct
	SNP_A-2002514	rs7973458	12p13.31	Intergenic	+	C/T	0.42	0.56	–0.14	+	0.01	0.37	1278	0.42	Correct
	SNP_A-2310304	rs854633	17q12	Intergenic	+	C/T	0.60	0.44	0.16	–	0.01	0.35	1261	0.04	Correct
	SNP_A-2173622	rs9347088	6q27	Intergenic	–	A/G	0.50	0.40	0.10	+	–0.03	0.16	1243	0.15	Incorrect
	SNP_A-1812640	rs9923476	16q24.1	Intergenic	+	C/G	0.56	0.35	0.21	+	0.01	0.37	1227	0.63	Correct
	SNP_A-2080138	rs10836588	11p12	Intergenic	+	C/T	0.36	0.54	–0.18	+	0.03	0.19	1259	0.18	Incorrect

Table 2 continued

AffyID	dbSNP RS ID	Chr	Genomic relation	Stage 1 (screening with SNP-MaP)			Stage 2 (testing with individual genotyping)							
				Strand	Alleles	Low	High	L–H diff	Strand	QTL <i>r</i>	QTL <i>p</i>	N	H–W	Correct direction?
SNP_A-1788579	rs4849166	2q13	Intron (PSD4)	+	A/T	0.48	0.26	0.22	+	0.03	0.11	1240	0.15	Correct
SNP_A-22277950	rs6901498	6q13	Intergenic	+	C/G	0.60	0.40	0.20	+	-0.04	0.08	1247	0.55	Incorrect
SNP_A-1984789	rs6919059	6p25.3	Intron (GMD5)	+	A/G	0.52	0.41	0.11	+	0.00	0.48	1252	0.21	
SNP_A-1867474	rs7410608	22q13.33	Intron (SAPS2)	-	C/G	0.39	0.65	-0.26	+	-0.01	0.38	1261	0.09	
SNP_A-2146324	rs7723053	5p15.33	Intergenic	+	C/T	0.45	0.56	-0.11	+	0.01	0.36	1251	<0.01	
SNP_A-1947292	rs7835148	8q24.23	Intergenic	+	A/G	0.46	0.70	-0.24	+	0.01	0.41	1254	0.99	
SNP_A-2237902	rs7928155	11p12	Intergenic	-	C/T	0.48	0.61	-0.13	+	-0.03	0.19	1243	0.74	Correct
SNP_A-1818651	rs979037	6p23	Intergenic	+	A/C	0.49	0.33	0.16	+	-0.02	0.28	1247	0.96	Incorrect

Note: In Stage 1, the allele whose frequency is shown in the 'Low' and 'High' columns is the allele that is alphabetically first in the 'Alleles' column (e.g., A rather than G in the first row). Note that the alleles measured in Stage 1 (using Affymetrix microarrays) may not be the same as the alleles individually genotyped in Stage 2 (using the SNPlex<sup>TM</sup> assay) because the Affymetrix and SNPlex platforms may be based on different DNA strands (e.g., for the SNP in the first row, alleles on different strands were assessed in the Affymetrix microarray and in the SNPlex), which was of course taken into account in determining the direction of the association in Stage 1 and Stage 2. The allele frequency difference ('L–H') shown in the 9th column was calculated by subtracting the allele frequency of the high (H) social group from the low (L) social group (and similarly for the nonsocial groups). In Stage 2 these 42 SNPs were individually genotyped in an unselected sample in order to test the QTL (quantitative trait loci) hypothesis. 'QTL *r*' (column 12) is the correlation between additive genotypic values (0, 1, 2) and the autistic trait score. Genotypic values in Stage 2 were scored as follows: AA = 0, AB = 1, BB = 2. QTL *p* (column 13) is the *P* value of the QTL *r* for the unselected sample of individuals; *p* is one-tailed because only the direction of association expected on the basis of Stage 1 results was accepted as significant. The male-only N is shown in column 14. The last column ('correct direction') indicates whether the direction of allele frequency differences is the same in Stage 2 as it was in Stage 1. We did not compare directions of effect between Stage 1 and Stage 2 for SNPs yielding correlations in the range -0.01 and 0.01 in Stage 2. Two SNPs deviated significantly (*P* < 0.01) from Hardy–Weinberg (H–W) expectation, which may be due in part to our large sample size. As described in the main text, SNP rs11894053 is the only SNP yielding a nominally significant association (*P* < 0.05) in the expected direction in this male-only sample in Stage 2. SNPMap = Single Nucleotide Polymorphism Microarrays and Pooling (see also "Materials and methods" section); *Chr* chromosome

males. Although this meant mixing males and females together, the advantage of using the whole sample for the Stage 2 replication was that it provided greater power to detect effects. Power to detect a QTL explaining 0.5% of the variance with the male-only sample was 62–72% ( $N = 1031$ – $1278$ ), whereas power to detect a QTL explaining 0.5% of the variance with the male and female sample was 94–97% ( $N = 2519$ – $3047$ ). In the social study, the SNP (rs11894053) that correlated in the male-only sample also correlated .04 ( $P = .02$ ) in the whole sample, and another SNP (rs17622673) correlated .03 ( $P = .05$ ) in the whole sample. For the nonsocial study, one SNP (rs12578517) correlated .03 ( $P = .03$ ) in the whole sample.

#### *Additional analysis using database of families with diagnosed ASD probands*

To test our hypothesis that SNPs associated with autistic-like traits in the general population will also be associated with diagnosed ASD, each of the 3 SNPs that were nominally significant in Stage 2 (rs11894053, rs17622673, rs12578517) were tested individually in a transmission disequilibrium test (TDT) analysis using the AGRE ASD database (see Appendix 2). The TDT analysis tested for over-transmission of the risk allele from heterozygous parents to affected offspring. However all 3 Chi-square tests were nonsignificant ( $P = .51$ – $.91$ ).

## Discussion

In this first genome-wide association study of autistic-like traits in the general population, we found one SNP associated with social autistic-like traits in the male-only Stage 2 sample that was nominally significant ( $P < 0.05$ ). When using the whole sample of males and females, the same SNP and another one were both nominally significant with social autistic-like traits, and one SNP was significant  $P < .05$  with nonsocial autistic-like traits. With 42 SNPs nominated in the SNP-MaP stage using pooled DNA for low versus high social and nonsocial autistic-like trait groups we would expect two SNPs to remain significant; therefore we only found one more SNP to be nominally significant with  $P < .05$  than that expected by chance alone. Importantly, no SNP associations emerged in Stage 2 that accounted for more than 0.4% of the variance in either study. In sum, despite studying two highly heritable traits and employing a three-stage design, we did not find any associations of the effect size we had power to detect.

The nominally significant SNP (rs11894053) in the male-only sample is in an intergenic region at 2p21 that maps to a hypothetical protein BC007901. According to a recent review (Abrahams and Geschwind 2008), one gene

on 2p, *NRXN1* (neurexin 1), has been implicated in clinical ASD but is located at 2p16, which is not in linkage disequilibrium with rs11894053. As described above, when using the whole sample, two additional SNPs became nominally significant. rs17622673 was associated with social autistic-like traits in the male and female sample; it is located in an intergenic region on 6q16.3 downstream from *GRIK2* (glutamate receptor, ionotropic, kainate 2). rs12578517 was associated with nonsocial autistic-like traits in the whole sample and is located on 12p12.3, upstream from *PTPRO* (protein tyrosine phosphatase, receptor type, O). However, none of these 3 SNPs were associated with diagnosed ASD in the AGRE sample.

We had hypothesised that SNPs selected from the pooling stage (which compared high versus low scoring groups) would also show a significant association across the dimension of autistic-like traits in the individual differences replication stage. In an additional analysis, we hypothesised that SNPs associated with autistic-like traits in Stages 1 and 2 would also be associated with diagnosed ASD. The small number of significant SNPs (at  $P < 0.05$ ) found in Stage 2, as well as the lack of significant association of these SNPs in the AGRE ASD sample, deserves discussion. Although the selection criteria for the SNPs for Stage 2 were carefully explored and have been applied successfully to identify SNPs in previous GWAS SNP-MaP studies (e.g., Butcher et al. 2008), it is possible that some aspect of the selection criteria meant that the most optimal SNPs for Stage 2 were not identified here. For example, we purposely biased our selection criteria in Stage 2 towards common alleles because we had greater power to detect them, which meant discriminating against less common alleles. DNA pooling has considerable advantages and has been shown to work effectively with the Affymetrix 500 K array (e.g., Docherty et al. 2007). Results might have been improved if more arrays per pool had been used. Ultimately, individual genotyping would have provided greater information in Stage 1. Financial considerations meant that the number of SNPs that could be followed up in the individual genotyping stage was constrained to approximately 20 SNPs per study. Therefore, financial constraints that limited the sensitivity and breadth of the genome-wide scan, as well as the sample size, were possible reasons for the lack of significant findings in Stages 1 and 2.

Indeed, power relies on the design employed and the sample sizes. Because Stage 1 was a screening stage that employed a rank-based composite to select SNPs, the power available in this stage is not explicit. However, it is thought that pooling retains 60–70% of the power of individual genotyping (Barratt et al. 2002). Power in Stage 2 of the design with both males and females included, as described in the “Results” section, was >90%. In Stage 3, with 777 families and with a genotypic relative risk of 1.5, we could

expect to have 77% power to detect an additive SNP with a MAF of 0.05, using a nominal significance level ( $p < .05$ ).

Regarding the additional TDT analyses of the AGRE database, heterogeneity within ASD samples (for example, due to different types of diagnoses and comorbid features) is often cited as a problem in ASD samples and it is possible that this could explain the null findings in the AGRE sample. It is noteworthy that using three different types of analysis on the SNPs at each stage of the study (allelic low–high group association, a QTL additive model, TDT analysis) made the test of significance at 3 consecutive stages extremely stringent. Finally, which SNPs were tested in the AGRE sample depended on the results of Stages 1 and 2. Further research is needed, but the present data did not support the hypothesis that the same genes influence autistic-like traits assessed dimensionally in the general population and diagnosed ASD in clinical samples.

At the phenotypic level, reliable and valid measurement of autistic-like traits is vital. Quantitative measures of autistic-like traits are still in development—there is no gold standard as there is with ASD diagnoses. The measures employed in this study were relatively short questionnaires and were only moderately reliable (Ronald et al. 2005). An additional issue to consider is the choice of rater when assessing autistic-like traits as it has been shown that parent, teacher and self reports show only modest agreement and may in part pick up on somewhat different genetic influences when assessing autistic-like traits in children (Ronald et al. 2008). Indeed, in a previous study, different linkage signals were found for parent and teacher ratings of a quantitative assessment of autistic behaviors in ASD families (Duvall et al. 2007).

A male-only sample was initially selected for Stages 1 and 2 in order to avoid conflating the genders in the analysis, and because autistic-like traits and ASD are both more common in males and there is some evidence for sex-specific genetic effects (e.g., Stone et al. 2004). Therefore Stage 1 will have missed SNPs associated specifically with autistic-like traits in females. Females were added to the Stage 2 sample in order to increase power in the analysis. These data do not distinguish between the possibilities that some SNPs showed a sex-specific effect or that they were easier to identify simply because the sample of males and females offered greater power than the male-only sample.

A more general consideration concerning the design of the study is that relevant SNPs not captured by the Affymetrix 500 K array, other polymorphisms (e.g., copy number variation, microsatellites), as well as rare alleles, as mentioned above, may have passed through our screen unnoticed. This limitation is likely to be avoided in future studies because of the advent of newer microarrays offering more comprehensive coverage of both SNPs and copy number variations and the availability of larger samples. Both social and nonsocial teacher-rated autistic traits show high heritability (63–74%)

(Ronald et al. 2005) which led us to expect to be able to find SNPs associated with these traits. Three recent GWAS studies have identified common SNP variants associated with autism (Ma et al. 2009; Wang et al. 2009; Weiss et al. 2009). In positional candidate gene studies, several common variants have been found to be associated with autism, for example, a common allele in the promoter region of the MET gene (Campbell et al. 2006) and a common polymorphism in the contactin-associated protein-like 2 gene (CNTNAP2; Alarcon et al. 2008). These previous findings support the hypothesis that common variants play a role in the etiology of autism. In conclusion, this first GWAS study of social and non-social autistic-like traits in the general population joins one of the other published GWAS studies of autism to date, a family-based study, in reporting largely negative findings (Arking et al. 2008).

**Acknowledgments** TEDS is funded by MRC grant G0500079. AR was funded by Autism Speaks during this project. We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families. The Autism Genetic Resource Exchange is a program of Autism Speaks and is supported, in part, by grant 1U24MH081810 from the National Institute of Mental Health to Clara M. Lajonchere (PI). AGRE Affymetrix 5.0 data was generated at the Broad Institute and provided to AGRE by Dr. Mark Daly and the Autism Consortium. Maja Bucan, Ph.D., AGRE Scientific Steering Committee Chair, Department of Psychiatry, University of Pennsylvania, PA; Daniel H. Geschwind, M.D., Ph.D., AGRE Chief Scientific Advisor, Department of Neurology, UCLA School of Medicine, CA; W. Ted Brown, M.D., Ph.D., NYS Institute for Basic Research, NY; Rita M. Cantor, Ph.D., UCLA School of Medicine, CA; John N. Constantino, M.D., Washington University School of Medicine; T. Conrad Gilliam, Ph.D., University of Chicago, IL; Martha Herbert, M.D., Ph.D. Harvard University, MA; David Ledbetter, Ph.D., Emory University, GA; Stanley F. Nelson, M.D., UCLA School of Medicine, CA; Carole Samango-Sprouse, Ed.D., George Washington University, DC; Gerard Schellenberg, Ph.D., University of Washington, WA; Jonathan Shestack, Autism Speaks, Los Angeles, CA; Matthew State, M.D., Ph.D., Yale University School of Medicine, CT; Rudolph Tanzi, Ph.D., Harvard Medical School, MA.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## Appendix 1

The following sections give an overview of how we calculated allele frequency estimates using the measurement structure of Affymetrix DNA genotyping microarrays with pooled DNA of groups selected for low and high social (or nonsocial) autistic-like traits. We also detail how these measurements (and their derivatives) were implemented to form a five-criterion rank-based composite score for each SNP used to select SNPs from Stage 1 (pooled DNA using Affymetrix 500 K microarrays) for individual genotyping on an independent and representative sample of social (or

nonsocial) autistic-like traits. The same approach is described in Butcher et al. (2008).

#### Calculation of allele frequency estimates using pooled DNA

As with other applications of pooled DNA, allele frequency estimates ( $p$ ) are calculated as the proportion of fluorescent intensity corresponding to allele  $A$  to the fluorescent intensities corresponding to the sum of alleles  $A$  and  $B$ :

$$p = \frac{A}{A + B}. \quad (1)$$

Affymetrix microarrays, however, measure alleles  $A$  and  $B$  numerous times using multiple unique probes (oligos) scattered across the microarray's surface. The fluorescent intensity values for these probes are contained in cell files (.cel) that are produced separately for each microarray. The lowest level at which an allele frequency estimate can be measured using Eq. 1 is at the "quartet" level. A quartet contains four 25 bp probes with variations at a consistent location within each probe. The four variations are either (1) a perfect match to allele  $A$  of the SNP ( $PM^A$ ), (2) a perfect match to allele  $B$  ( $PM^B$ ), (3) a mismatch to allele  $A$  ( $MM^A$ ) or (4) a mismatch to allele  $B$  ( $MM^B$ ). Numerous quartets represent each SNP on the microarray with variation achieved through "off-sets" and/or by designing quartets on sense or anti-sense strands. Off-sets refer to the shifting of the SNP interrogation site to a different position within the quartets' probes. Sequence design occurs either exclusively on the sense or anti-sense strand, or on both strands depending on the SNP. Therefore, microarray measurement is defined as taking place at the  $i$ th quartet of the  $j$ th replicate in group  $k$ , for SNP  $s$ .

We denote the four probes within a quartet as:

Perfect match allele  $A = PM_{ijks}^A$

Perfect match allele  $B = PM_{ijks}^B$

Mismatch allele  $A = MM_{ijks}^A$

Mismatch allele  $B = MM_{ijks}^B$ .

We then transform  $PM_{ijks}$  probe intensities by subtracting an estimate of non-specific hybridization (the average intensity of the two mismatch probes) to derive best estimates for allele  $A$  ( $A_{ikjs}$ ) and allele  $B$  ( $B_{ikjs}$ ):

$$A_{ikjs} = \text{MAX} \left( PM_{ijks}^A - \left\{ \frac{MM_{ikjs}^A + MM_{ikjs}^B}{2} \right\}, 0 \right) \quad (2)$$

$$B_{ikjs} = \text{MAX} \left( PM_{ijks}^B - \left\{ \frac{MM_{ikjs}^A + MM_{ikjs}^B}{2} \right\}, 0 \right). \quad (3)$$

Transformed values are then substituted into Eq. 1 to provide an allele frequency estimate for the  $i$ th quartet:

$$p_{ijks} = \frac{A_{ijks}}{A_{ijks} + B_{ijks}}. \quad (4)$$

The allele frequency estimate for the  $j$ th replicate is the simple arithmetic mean of  $I$  quartets with 'interpretable data', i.e., the denominator of Eq. 4  $\neq 0$ :

$$p_{jks} = \begin{cases} \frac{1}{I} \sum_{i=1}^I p_{ijks}, & \text{if } I \geq 0.7I_{\max} \\ NULL, & \text{if } I < 0.7I_{\max} \end{cases} \quad (5)$$

where  $I_{\max} \in \{6, 10\}$  and is dependent on SNP,  $s$ . As can be seen from Eq. 5 we only accepted allele frequency estimates from replicates with either 5/6 or 7/10 quartets with interpretable data.

Variance across quartets is thus:

$$p_{jks}^{\text{var}} = \frac{1}{I-1} \sum_{i=1}^I (p_{ijks} - p_{jks})^2. \quad (6)$$

The allele frequency estimate for the  $k$ th group is simply the arithmetic mean of  $J$  replicates, thus:

$$p_{ks} = \frac{1}{J} \sum_{j=1}^J p_{jks}, \quad (7)$$

with variance of allele frequency estimates across replicates:

$$p_{ks}^{\text{var}} = \frac{1}{J-1} \sum_{j=1}^J (p_{jks} - p_{ks})^2, \quad (8)$$

where  $J$  is the number of replicates within the  $k$ th group.

The following sections detail how the data acquired from the above equations is used to create a rank-based composite score (based on five criteria) for each SNP used to select SNPs from Stage 1 (pooled DNA using Affymetrix 500 K microarrays) for individual genotyping on an independent and representative sample assessed on social and nonsocial autistic-like traits.

#### Criterion (1)

##### *Allele frequency difference between low and high groups*

Using Eq. 7 we calculated allele frequency estimates separately for the two groups, low social autistic-like traits ( $k^{\text{low}}$ ) and high social autistic-like traits ( $k^{\text{high}}$ ). The same was done for the high and low nonsocial groups. The allele frequency difference between groups at SNP  $s$  is the absolute difference between group estimates, thus:

$$p_s^{\text{diff}} = |p_{k^{\text{low}},s} - p_{k^{\text{high}},s}|. \quad (9)$$

For the composite, allele frequency differences were standardized separately by array type (NspI or StyI) and

weighted positively (to prioritize larger allele frequency differences) which we denote:

$$+Zp_s^{\text{diff}}.$$

Criterion (2)

#### Variance of allele frequency estimates across replicates

Equation 8 was applied to each group, averaged, standardized separately by array type, summed then weighted negatively (to prioritize low variance scores), which we denote:

$$-\frac{1}{K} \sum_{k=1}^K Zp_{ks}^{\text{var}}.$$

Criterion (3)

#### Variance of allele frequency estimates across quartets

Equation 6 was applied to each replicate, averaged across all replicates, standardized separately by array type, summed then weighted negatively (to prioritize low variance scores), which we denote:

$$-\frac{1}{J} \sum_{j=1}^J Zp_{jks}^{\text{var}}.$$

Criterion (4)

#### Number of replicates

We denote the number of replicates at SNP  $s$ , for low and high groups at SNP  $s$  as  $J_{k^{\text{low}}_s}$  and  $J_{k^{\text{high}}_s}$ , respectively. We took the arithmetic mean of these values, standardized the result separately by array type, summed them then weighted them positively (to prioritize SNPs with more replicates) to give:

$$+\frac{1}{K}(ZJ_{k^{\text{low}}_s} + ZJ_{k^{\text{high}}_s}).$$

Criterion (5)

#### Minor allele frequency

Minor allele frequency was calculated as:

$$p_s^{\text{min}} = \begin{cases} \frac{1}{2}(p_{k^{\text{low}}_s} + p_{k^{\text{high}}_s}) & \text{if } \frac{1}{2}(p_{k^{\text{low}}_s} + p_{k^{\text{high}}_s}) \leq 0.5 \\ 1 - \frac{1}{2}(p_{k^{\text{low}}_s} + p_{k^{\text{high}}_s}) & \text{if } \frac{1}{2}(p_{k^{\text{low}}_s} + p_{k^{\text{high}}_s}) > 0.5 \end{cases}.$$

For the composite, minor allele frequencies were standardized separately by array type and positively weighted (to prioritize common allele frequencies) which we denote:

$$+Zp_s^{\text{min}}.$$

Composite

The composite measure ( $C$ ) was the simple summation of the standardized information from criterion 1–5:

$$C = +Zp_s^{\text{diff}} - \frac{1}{K} \sum_{k=1}^K Zp_{ks}^{\text{var}} - \frac{1}{J} \sum_{j=1}^J Zp_{jks}^{\text{var}} + \frac{1}{K}(ZJ_{k^{\text{low}}_s} + ZJ_{k^{\text{high}}_s}) + Zp_s^{\text{min}}.$$

In future experiments, different composite scores may be created by assigning weights to the different criteria, or add or remove different criteria. At the time of writing, the criteria used were believed to be the most informative for detecting QTLs of small effect size.

## Appendix 2

### Additional TDT analysis using database of families with diagnosed ASD probands

High-density single nucleotide polymorphism data from 500 K Affymetrix arrays on 777 families were analysed. These were contributed by the Autism Consortium to the Autism Genetic Resource Exchange (AGRE). From the sample of 2883 individuals, 4 had missing phenotypic information and 5 were removed for low genotyping (maximum SNP missingness rate >0.1). The final sample was 63% male and contained 2,874 individuals.

Distortion in the transmission of SNP alleles was tested for by the transmission disequilibrium test (TDT; Spielman et al. 1993) implemented in PLINK (Purcell et al. 2007). Thresholds were set to exclude minor allele frequencies <.01, maximum minor allele frequencies of 1.0, maximum SNP missingness rate of 0.1 and maximum individual missingness rate of 0.1. Over-transmission of the risk allele from heterozygous parents to affected offspring was tested for. However all 3 Chi-square tests were nonsignificant ( $P = .51-.91$ ).

## References

- Abrahams BS, Geschwind DH (2008) Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9(5):341–355
- Alarcon M, Cantor RM, Liu J, Gilliam TC, Geschwind DH (2002) Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *Am J Hum Genet* 70(1):60–71
- Alarcon M, Abrahams BS, Stone JL, Duvall JA, Perederer JV, Bomar JM, Sebat J, Wigler M, Martin CL, Ledbetter DH, Nelson SF, Cantor RM, Geschwind DH (2008) Linkage, association, and

- gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet* 82(1):150–159
- Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M, Rea A, Guy M, Lin S, Cook EH et al (2008) A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am J Hum Genet* 82(1):160–164
- Bailey A, Palferman S, Heavey L, Le Couteur A (1998) Autism: the phenotype in relatives. *J Autism Dev Disord* 28(5):369–392
- Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001) The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord* 31(1):5–17
- Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Am J Hum Genet* 66:393–405
- Bishop DV, Maybery M, Maley A, Wong D, Hill W, Hallmayer J (2004) Using self-report to identify the broad phenotype in parents of children with autistic spectrum disorders: a study using the Autism-Spectrum Quotient. *J Child Psychol Psychiatry* 45(8):1431–1436
- Brohede J, Dunne R, McKay JD, Hannan GN (2005) PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays. *Nucleic Acids Res* 33(17):e142
- Brune CW, Kim SJ, Salt J, Leventhal BL, Lord C, Cook EH Jr (2006) 5-HTTLPR genotype-specific phenotype in children and adolescents with autism. *Am J Psychiatry* 163(12):2148–2156
- Butcher LM, Meaburn E, Liu L, Fernandes C, Hill L, Al Chalabi A, Plomin R, Schalkwyk L, Craig IW (2004) Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav Genet* 34(5):549–555
- Butcher LM, Davis OS, Craig IW, Plomin R (2008) Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500 K single nucleotide polymorphism microarrays. *Genes Brain Behav* 7(4):435–446
- Buxbaum JD, Silverman J, Keddache M, Smith CJ, Hollander E, Ramoz N, Reichert JG (2004) Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. *Mol Psychiatry* 9(2):144–150
- Campbell DB, Sutcliffe JS, Ebert PJ, Militeri R, Bravaccio C, Trillo S, Elia M, Schneider C, Melmed R, Sacco R, Persico AM, Levitt P (2006) A genetic variant that disrupts MET transcription is associated with autism. *Proc Natl Acad Sci USA* 103(45):16834–16839
- Constantino JN, Todd RD (2000) Genetic structure of reciprocal social behavior. *Am J Psychiatry* 157(12):2043–2045
- Constantino JN, Todd RD (2003) Autistic traits in the general population: a twin study. *Arch Gen Psychiatry* 60(5):524–530
- Constantino JN, Davis SA, Todd RD, Schindler MK, Gross MM, Brophy SL, Metzger LM, Shoushtari CS, Splinter R, Reich W (2003) Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *J Autism Dev Disord* 33(4):427–433
- Craig DW, Huentelman MJ, Hu-Lince D, Zismann VL, Kruer MC, Lee AM, Puffenberger EG, Pearson JM, Stephan DA (2005) Identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics* 6:138
- Darvasi A, Soller M (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138(4):1365–1373
- Department for Education and Skills (2002) Special educational needs in England: January, 2002 Bulletin
- Docherty SJ, Butcher LM, Schalkwyk LC, Plomin R (2007) Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics* 8:214
- Duvall JA, Lu A, Cantor RM, Todd RD, Constantino JN, Geschwind DH (2007) A quantitative trait locus analysis of social responsiveness in multiplex autism families. *Am J Psychiatry* 164(4):656–662
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861
- Freeman B, Smith N, Curtis C, Huckett L, Mill J, Craig IW (2003) DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav Genet* 33(1):67–72
- Happé F, Ronald A, Plomin R (2006) Time to give up on a single explanation for autism. *Nat Neurosci* 9(10):1218–1220
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6(2):95–108
- Hoekstra RA, Bartels M, Verweij CJ, Boomsma DI (2007) Heritability of autistic traits in the general population. *Arch Pediatr Adolesc Med* 161(4):372–377
- Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC (2006) Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* 7:27
- Knight J, Sham P (2006) Design and analysis of association studies using pooled DNA from large twin samples. *Behav Genet* 36(5):665–677
- Liu QR, Drgon T, Walther D, Johnson C, Poleskaya O, Hess J, Uhl GR (2005) Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci USA* 102(33):11864–11869
- Liu XQ, Paterson AD, Szatmari P (2009) Genome-wide Linkage Analyses of Quantitative and Categorical Autism Subphenotypes. *Biol Psychiatry* 64:561–570
- Ma D, Salyakina D, Jaworski JM, Konidari I, Whitehead PL, Andersen AN, Hoffman JD, Slifer SH, Hedges DJ, Cukier HN, et al (2009) A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet* 73:263–273
- McCauley JL, Olson LM, Dowd M, Amin T, Steele A, Blakely RD, Folstein SE, Haines JL, Sutcliffe JS (2004) Linkage and association analysis at the serotonin transporter (SLC6A4) locus in a rigid-compulsive subset of autism. *Am J Med Genet* 127B(1):104–112
- Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al Chalabi A, Plomin R, Craig I, Schalkwyk LC (2005) Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* 6(1):52
- Meaburn E, Butcher LM, Schalkwyk LC, Plomin R (2006) Genotyping pooled DNA using 100 K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res* 34(4):e27
- Mulder EJ, Anderson GM, Kema IP, Brugman AM, Ketelaars CE, de Bildt A, van Lang ND, den Boer JA, Minderaa RB (2005) Serotonin transporter intron 2 polymorphism associated with rigid-compulsive behaviors in Dutch individuals with pervasive developmental disorder. *Am J Med Genet B Neuropsychiatr Genet* 133B(1):93–96
- Norton N, Williams NM, O'Donovan MC, Owen MJ (2004) DNA pooling as a tool for large-scale association studies in complex traits. *Ann Med* 36(2):146–152

- Nurmi EL, Dowd M, Tadevosyan-Leyfer O, Haines JL, Folstein SE, Sutcliffe JS (2003) Exploratory subsetting of autism families based on savant improves evidence of genetic linkage to 15q11–q13. *J Am Acad Child Adolesc Psychiatry* 42(7):856–863
- Office for National Statistics (2002) Living in Britain: results from the 2000/01 General Household Survey
- Oliver BR, Plomin R (2007) Twins' early development study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Res Hum Genet* 10(1):96–105
- Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL et al (2007) Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am J Hum Genet* 80(1):126–139
- Pickles A, Starr E, Kazak S, Bolton P, Papanikolaou K, Bailey A, Goodman R, Rutter M (2000) Variable expression of the autism broader phenotype: findings from extended pedigrees. *J Child Psychol Psychiatry* 41(4):491–502
- Posserud MB, Lundervold AJ, Gillberg C (2006) Autistic features in a total population of 7–9-year-old children assessed by the ASSQ (Autism Spectrum Screening Questionnaire). *J Child Psychol Psychiatry* 47(2):167–175
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature* 405(6788):847–856
- Ronald A, Happé F, Plomin R (2005) The genetic relationship between individual differences in social and nonsocial behaviours characteristic of autism. *Dev Sci* 8:444–458
- Ronald A, Happé F, Bolton P, Butcher LM, Price TS, Wheelwright S, Baron-Cohen S, Plomin R (2006a) Genetic heterogeneity between the three components of the autism spectrum: a twin study. *J Am Acad Child Adolesc Psychiatry* 45:691–699
- Ronald A, Happé F, Price TS, Baron-Cohen S, Plomin R (2006b) Phenotypic and genetic overlap between autistic traits at the extremes of the general population. *J Am Acad Child Adolesc Psychiatry* 45:1206–1214
- Ronald A, Happé F, Plomin R (2008) A twin study investigating the genetic and environmental aetiologies of parent, teacher and child ratings of autistic-like traits and their overlap. *Eur Child Adolesc Psychiatry* 17:473–483
- Scourfield J, Martin N, Lewis G, McGuffin P (1999) Heritability of social cognitive skills in children and adolescents. *Br J Psychiatry* 175:559–564
- Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66(5):1616–1630
- Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. *Nat Rev Genet* 3(11):862–871
- Shao Y, Cuccaro ML, Hauser ER, Raiford KL, Menold MM, Wolpert CM, Ravan SA, Elston L, Decena K, Donnelly SL et al (2003) Fine mapping of autistic disorder to chromosome 15q11–q13 by use of phenotypic subtypes. *Am J Hum Genet* 72(3):539–548
- Simpson CL, Knight J, Butcher LM, Hansen VK, Meaburn E, Schalkwyk LC, Craig IW, Powell JF, Sham PC, Al-Chalabi A (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res* 33(3):e25
- Skuse DH, Mandy WP, Scourfield J (2005) Measuring autistic traits: heritability, reliability and validity of the social and communication disorders checklist. *Br J Psychiatry* 187:568–572
- Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516
- Stone JL, Merriman B, Cantor RM, Yonan AL, Gilliam TC, Geschwind DH, Nelson SF (2004) Evidence for sex-specific risk alleles in autism spectrum disorder. *Am J Hum Genet* 75:1117–1123
- Sutcliffe JS, Delahanty RJ, Prasad HC, McCauley JL, Han Q, Jiang L, Li C, Folstein SE, Blakely RD (2005) Allelic heterogeneity at the serotonin transporter locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors. *Am J Hum Genet* 77(2):265–279
- Sykes NH, Lamb JA (2007) Autism: the quest for the genes. *Expert Rev Mol Med* 9(24):1–15
- Szatmari P, MacLean JE, Jones MB, Bryson SE, Zwaigenbaum L, Bartolucci G, Mahoney WJ, Tuff L (2000) The familial aggregation of the lesser variant in biological and nonbiological relatives of PDD probands: a family history study. *J Child Psychol Psychiatry* 41(5):579–586
- The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426(6968):789–796
- The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437(7063):1299–1320
- Tobler AR, Short S, Andersen MR, Paner TM, Briggs JC, Lambert SM, Wu PP, Wang Y, Spoonde AY, Koehler RT et al (2005) The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech* 16(4):398–406
- Tordjman S, Gutknecht L, Carlier M, Spitz E, Antoine C, Slama F, Carsalade V, Cohen DJ, Ferrari P, Roubertoux PL et al (2001) Role of the serotonin transporter gene in the behavioral expression of autism. *Mol Psychiatry* 6(4):434–439
- Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM, et al (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459(7246):528–533
- Weiss LA, Arking DE, Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly MJ, Chakravarti A (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 461(7265):802–808
- Yang MS, Gill M (2007) A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence. *Int J Dev Neurosci* 25(2):69–85