Supplementary Methods, Tables and Figures

1.1 Autism Genetic Resource Exchange (AGRE)

The Autism Genetic Resource Exchange (AGRE; http://www.agre.org) has a collection of DNA samples and clinical information from families with autism spectrum disorders (ASDs) [1]. We have genotyped DNA samples from 943 families (4,444 individuals) from the entire AGRE collection (as of August 2007). These AGRE families include 917 multiplex families, 24 simplex families and 2 families without ASD diagnosis (not used in analysis).

The AGRE annotation database classifies three diagnostic categories based on the Autism Diagnostic Interview-Revised (ADI-R) [2]: autism, broad spectrum (patterns of impairment along the spectrum of pervasive developmental disorders, including PDD-NOS and Asperger's syndrome) or Not Quite Autism (individuals who are no more than one point away from meeting autism criteria on any or all of the social, communication, and/or behavior domains and meet criteria for "age of onset"; or, individuals who meet criteria on all domains, but do not meet criteria for the "age of onset"). In our analysis, AGRE patients with "Autism" (n=1,684), "Broad Spectrum" (n= 171) or "Not Quite Autism" (n=79) phenotype annotation were treated as a single ASD group. Among them, 11 subjects had autism diagnoses assigned by ADOS (Autism Diagnostic Observation Schedule) [3] without ADI-R (Autism Diagnostic Interview-Revised).

The age of onset and age of assessment for ASD subjects with different diagnostic categories were given in detail below. The Ravens estimated non-verbal IQ scores are available for a subset of AGRE individuals: the median score is 100 in multiplex families (708 ASD subjects) and 98 in simplex families (49 ASD subjects). 387 ASD subjects in multiplex families and 28 ASD subjects in simplex families cannot be tested on the Ravens (annotated as "Ravens-untestable" in AGRE annotation database) due to either low functioning or behavior.

Multiplex	**Number of individuals**		**Median**	**Mean**	**SD**	**Range**
Autism	1358					
		Age of Onset	1.25	1.25	0.68	<1 - 5 years
		Age of Assessment	7.12	8.11	4.68	2 - 46 years
NQA	68					
		Age of Onset	1.5	1.82	1.15	<1 - 6 years
		Age of Assessment	5.44	6.84	4.35	2 - 24 years
BroadSpectrum	136					
		Age of Onset	1.5	1.73	1.02	<1 - 5 years
		Age of Assessment	6.19	8.18	6.25	2 - 44 years
						2 - 44 years
			Median	Mean	SD	Range
Ravens estimated non-verbal IQ	708		100	100	18	38 - 143
Simplex	**Number of individuals**		**Median**	**Mean**	**SD**	**Range**
Autism	105					
		Age of Onset	1.5	1.36	0.72	<1 - 3.5 years
		Age of Assessment	9.98	9.57	4.52	3 - 30 years
NQA	3					
		Age of Onset	2	1.6	0.57	1 - 2 years
		Age of Assessment	7.49	9.94	5.78	5-16 years
BroadSpectrum	13					
		Age of Onset	1.5	1.92	1.25	<1 - 5 years
		Age of Assessment	6.88	10.4	9.38	3 - 31 years

		Median	Mean	SD	Range
Ravens estimated non-verbal IQ	49	98	96	22	38 - 134

The self-identified race/ethnicity information for these AGRE individuals is listed below. However, in our association analysis, we used multi-dimensional scaling on genotype data and applied stringent criteria to identify all subjects with European ancestry, and we excluded subjects of other ancestry from the association test (see detailed QC procedure below).

AGRE self-identified ancestry	**Number of subjects**
American Indian/Alaskan Native	10
Asian	103
Black or African American	99
More Than One Race	262
Native Hawaiian or other Pacific Islander	28
Unknown	448
White	3,494

1.2 ASD and control subjects in (Autism Case-Control) ACC cohort

The ASD subjects within the ACC cohort were provided by researchers from multiple collaborative projects across the US, as well as the Children's Hospital of Philadelphia (CHOP), where all samples were genotyped. All ASD subjects utilized for the case-control analysis were diagnosed with the ADOS (Autism Diagnostic Observation Schedule), ADI (Autism Diagnostic Interview) or ADI-R (Autism Diagnostic Interview-Revised) diagnostic tools. The "Best Diagnosis" provided by collaborators are used to select ASD subjects for genotyping, which is a composite measure based on both ADI and ADOS. After excluding subjects who have not been genotyped, subjects without genotype data in the database (due to chip failure), subjects without phenotype annotation, and subjects with missing diagnosis data (when "Best diagnosis" is set as "MISSING"), we were left with 1,453 samples that met the study criteria of either positive ADI/ADI-R, ADOS or both.

The average age of the study subjects was 10.3 ± 6.6 years, and the average age for ADI diagnosis was 8.4 ± 4.7 years, the average age for ADOS diagnosis was 9.9 ± 7.2 years, and the average age of IQ test is 10.9 ± 6.7 years. Only 1,204 subjects of European ancestry were used in the study (see QC section below). The majority (83.1%) of subjects

were males. Almost all (94.5%) DNA samples were extracted from whole blood, while others were from cell lines.

The IQ distribution, when known, is given below.

	NVIQ				VIQ			
Level	**Number**	**Median**	**Mean**	**SD**	**Number**	**Median**	**Mean**	**SD**
Autism/AUT	572	89	85	27.7	562	72	75	29.8
ASD/PDD-NOS/Asperger	29	100	98	18.8	36	106	105	24.8

The control group used in the discovery phase included 7,077 children of self-reported Caucasian ancestry (average age was 8.8 ± 5.4 SD years; 52.08% males, 47.65% females and 0.27% unknown). All control subjects had no history of ASDs, and had not demonstrated symptoms to be referred to diagnostic testing. The CHOP controls were recruited by CHOP nursing and medical assistant staff under the direction of CHOP clinicians within the CHOP Health Care Network, including four primary care clinics and several group practices and outpatient practices that included well child visits. All DNA samples were extracted from whole blood. Although these control subjects were all self-identified Caucasians, we combined these subjects with cases and used multi-dimensional scaling to infer a homogeneous group of subjects of European ancestry during our quality control procedure (see QC section below).

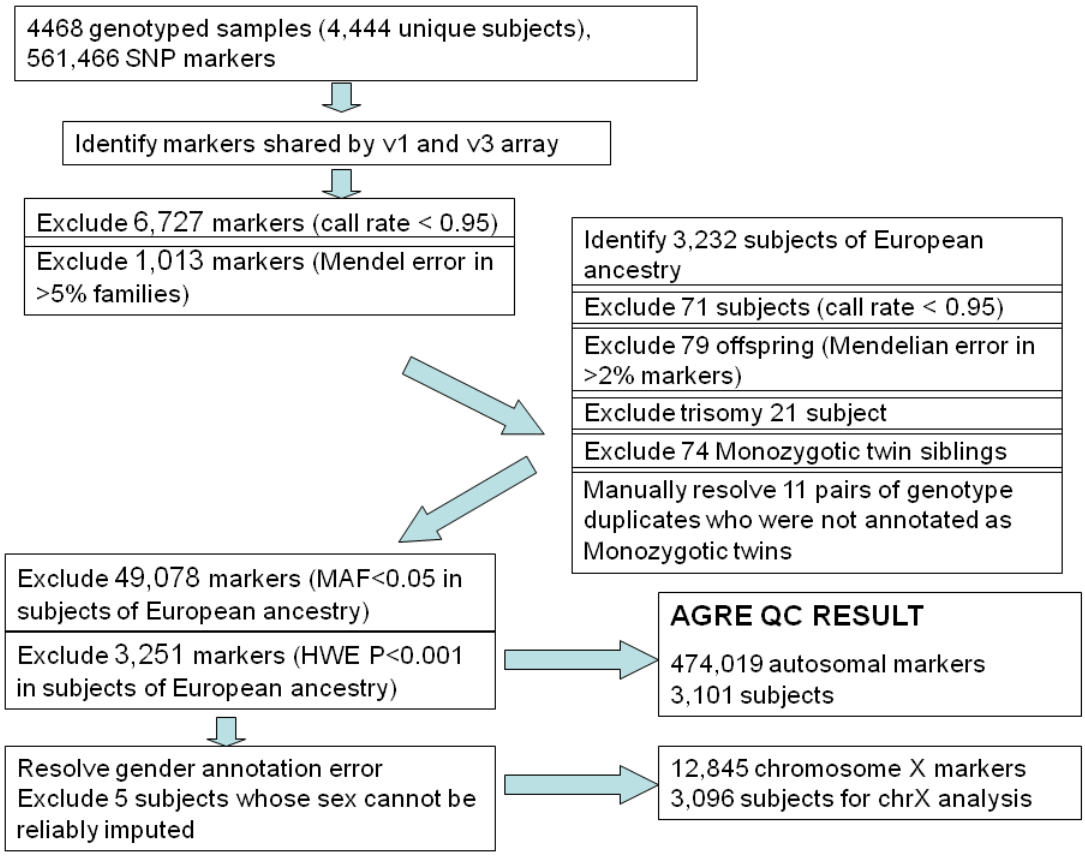
1.3 Genotyping platform for discovery cohorts

Individuals in the AGRE cohort and the ACC cohort were genotyped utilizing the Illumina HumanHap550 SNP genotyping array, which contains more than 550,000 tag SNPs, selected on the basis of HapMap Phase I and Phase II data to capture the haplotype diversity across the human genome. Among the several cohorts used in our study, the samples from AGRE were genotyped using DNA extracted from Epstein-Barr Virus (EBV)-transformed lymphoblastoid cell lines, while almost all subjects in the other cohorts (both ASD cases and control subjects) were genotyped using DNA extracted from whole blood.

The genotyping experiments for AGRE families and the ACC subjects were performed at the Center for Applied Genomics, Children's Hospital of Philadelphia. Most of the AGRE samples (n=4,163) were genotyped on the Illumina HumanHap550 version 3 arrays, but a small subset of AGRE samples (n=291) were genotyped by the version 1 arrays. The only difference between version 1 and version 3 arrays is the replacement of ~10K SNP markers in the new version of arrays by Illumina.

1.4 Quality control (QC) overview for AGRE data set

An overview of the quality control (QC) procedure for the AGRE data set (autosomal markers) is given in the figure below. More detailed QC procedure is described in the following two sections. Since the PDT software cannot be used on sex chromosomes, we have applied X-APL on chromosome X markers in a separate analysis, and the QC procedure is described in section 1.6.9.



1.5 Quality control for the selection of subjects in association analysis in the AGRE cohort

Stringent quality control (QC) measure was applied on the genotyped AGRE subjects for subsequent association analysis. The various aspects of QC were described in detail below:

1.5.1 Low genotype call rate

The call rate is calculated based on the number of "No Call" genotypes with default genotyping calling algorithm as implemented in the Illumina BeadStudio software. The call rate per individual was assessed by the PLINK software [4]. A total 24 samples have been genotyped twice due to the low call rate in the first batch of genotyping. Altogether, 47 unique individuals in AGRE data set were excluded from analysis due to low call rate.

1.5.2 Mendelian error

Due to the availability of family data, we were able to check the familial relationships between the AGRE samples with known pedigree information. Samples with excessive Mendelian errors could indicate potential paternity problems, sample mislabeling, or sample handling problem during the genotyping experiments, and should be excluded from downstream association analysis.

This analysis was performed with respect to offspring, that is, whenever a Mendelian error is present, the offspring gets a count of Mendelian error, while the parents do not get such a count. When one offspring in a large nuclear family has Mendelian problems (for example, due to sample mislabeling for this individual), this procedure ensures that only this offspring is excluded, while other offspring and the parents are still kept in the analysis. The Mendelian error rate per individual was assessed by the PLINK software [4]. A total of 79 samples (as offspring) are identified who had >2% markers with Mendelian inconsistency with respect to parental genotype data, and were excluded from our association test.

1.5.3 Monozygotic twins

In the AGRE collection, 70 families contain MonoZygotic (MZ) twins, including those with triplets and quartets siblings. We have removed 74 individuals from the analysis, such that only one MZ twin sibling in each family is kept in the analysis.

1.5.4 Genotype duplicates who were not annotated as monozygotic twins

We next checked genotype duplicates, that is, two subjects with almost identical genotypes, who were not annotated as monozygotic twins in the AGRE annotation, some of whom were even present in two different families. As expected, when two duplicates were present in two different families, they can be readily detected by Mendelian

inconsistency and usually we can infer which sample is being mislabeled into the wrong family. The complete list of duplicated individuals who were not annotated as MZ twins is given below, and these issues were manually examined and resolved.

individual 1	individual 2	Notes
AU026402	AU013801	AU013801 is singleton and not used in any analysis
AU001201	AU000803	Mendelian error for AU000803, excluded from analysis
AU043603	AU033402	Mendelian error for AU043603, excluded from analysis
AU1242302	AU1214302	Mendelian error for AU1214302, excluded from analysis
AU1364302	AU1378304	Mendelian error for AU1364302 and excluded from analysis; AU137804 excluded from analysis
AU1644304	AU1655201	AU1655201 is parent and this family has no children passing QC
AU1070301	AU1008201	AU1008201 is parent and this family has no children passing QC
AU1953302	AU1953303	Both individuals were excluded from analysis
AU1791301	AU1791302	family AU1791 excluded from analysis
AU1833302	AU1833303	This pair of MZ twin is NOT annotated in the AGRE phenotype database; AU1833303 is manually excluded from analysis
AU037803	AU035502	Mendelian error for AU037803, excluded from analysis; AU035502 is singleton and not used in analysis

1.5.5 Chromosome 21 trisomy

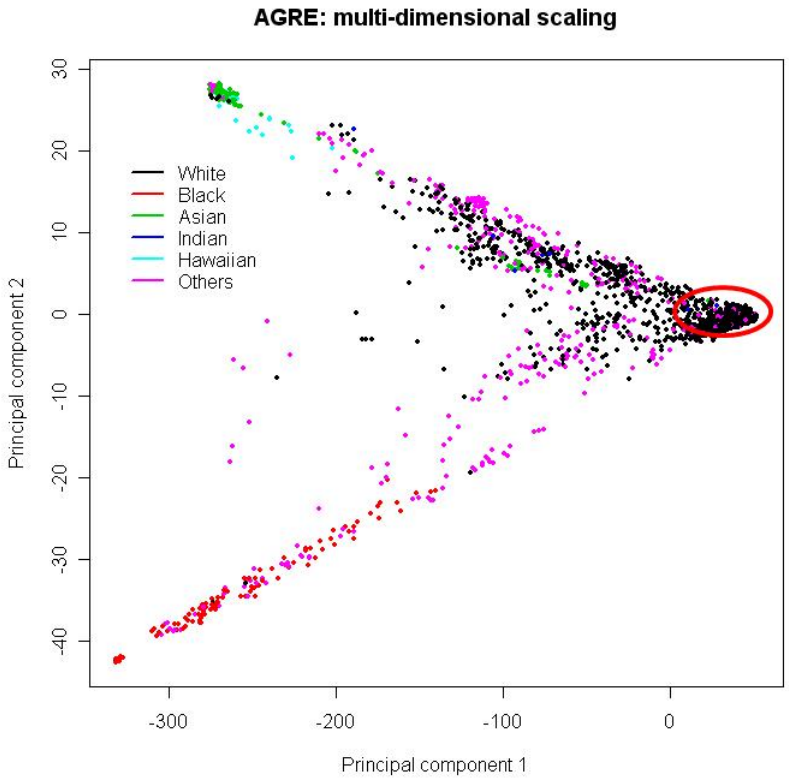
Using the PennCNV algorithm [5], we have identified three subjects with chromosome 21 trisomy, including AU075307, AU1227303 and AU015804. The individual AU015804

was annotated as "non-idiopathic autism" in the AGRE phenotype database, and was excluded from our association analysis.

1.5.6 Inferring individuals of European ancestry

Although family-based study design protects against population stratification, it may lead to allelic heterogeneity and mask truly associated signals. We have examined individuals of inferred European ancestry for association signals in all our discovery cohorts and replication cohorts.

We used Multi-Dimensional Scaling (MDS), as implemented in the PLINK software [4], for inferring population structure in the AGRE data set (See figure below). Comparing self-identified ancestry with the MDS-inferred ancestry confirmed the reliability of MDS to identify genetically inferred individuals of European ancestry. These individuals are clustered towards the right side of the triangle, as defined by that Principal component 1 is more than -10, and that Principal component 2 is between -2 and 2 (the "head" of the triangle, see red circle below).



A total of 3232 individuals were inferred as having European ancestry using the above procedure.

1.5.7 Genome-wide homozygosity

The abnormal homozygosity levels in genome-wide genotype data may indicate low-quality samples, genotyping failures, sample contamination/mixing up, or cell-line artifacts. We used PLINK to calculate runs of homozygosity levels in autosome markers passing QC, using all default parameters. No systematic problems were detected in these 3,101 subjects, with mean=23.4Mb, median=22.9Mb, minimum=4.3Mb and

maximum=70.5Mb. Furthermore, for each individual, we also examined the fraction of autosome markers passing QC that have homozygous genotypes. The fraction range from 63.7% to 67.7%, with mean=65.0%, indicating the lack of extreme outliers.

1.5.8 Final counts of subjects passing QC

Applying the QC measures mentioned in all the previous sections, we were left with 3,101 individuals for association analysis on autosomes.

1.6 Quality control for selection of SNPs in association analysis in the AGRE cohort

1.6.1 Overlap of the HumanHap550 v1 and v3 arrays

Since a small portion of the individuals in the AGRE cohort are genotyped by the HumanHap550 v1 array (n=291) while others are genotyped by the v3 array, our analysis only concerns on the markers shared by the v1 and v3 array: The HumanHap550 v1 array contains 555,352 markers, while the v3 array contains 561,466 markers, including 545,080 markers that are shared by the two arrays.

1.6.2 Mitochondria and sex chromosome markers

To analyze the AGRE data set by the PDT software, we have excluded markers from X, Y and Mitochondria chromosomes to restrict our association analysis to autosomal markers and markers in pseudo-autosomal regions (chrXY markers). This left us with 531,689 autosomal markers and 15 chrXY markers from the above step. The analysis on chromosome X was performed by the X-APL software in parallel, and the QC procedures were described in section 1.6.9 below. The analysis on chromosome Y markers cannot be performed on the family-based cohorts, but we have attempted the analysis on the ACC case-control cohort (described in section 1.6.10). The analysis on chrXY markers were performed as if they were autosomal markers (described in section 1.6.11).

1.6.3 NoCall rate per marker

Markers with call rate less than 95% were excluded from analysis. The call rates were calculated by the PLINK software [4]. A total of 6,727 markers were excluded from association analysis in this step.

1.6.4 Mendelian error

Markers with excessive Mendelian error (in >5% families) were excluded from analysis, since they may indicate genotyping failure, SNP clustering failure or the presence of SNPs within common copy number variation regions. Based on per-individual Mendelian error rate calculated by the PLINK software [4], a total of 492 markers does not meet this threshold and should be excluded.

1.6.5 Minor Allele Frequency (parents of European ancestry)

Markers with Minor Allele Frequency (MAF) less than 5% were excluded from our analysis. This procedure is restricted on AGRE founders passing QC and used in our association analysis, and the MAF are calculated by the PLINK software on the founders (parents) of the AGRE collection. A total of 49,078 markers were excluded from association analysis in this step.

1.6.6 Hardy-Weinberg Equilibrium (parents of European ancestry)

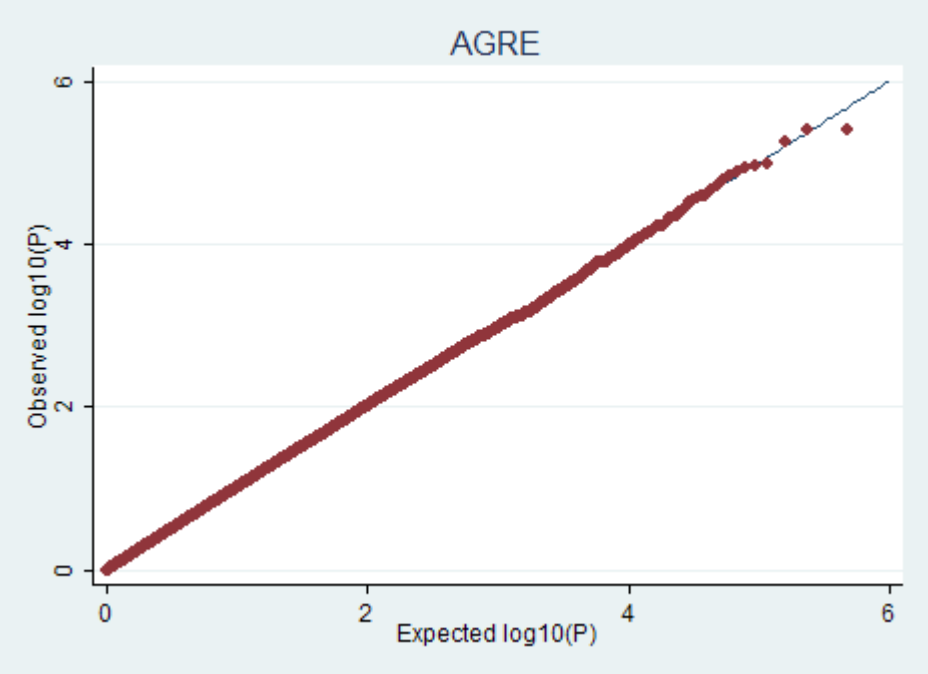
Markers with Hardy-Weinberg Equilibrium P-value less than 0.001 excluded from analysis, since these markers may have genotyping failure, or are located in common CNV regions. This procedure is restricted on AGRE founders passing QC and used in our association analysis, and the MAF are calculated by the PLINK software on the founders (parents) of the AGRE collection. A total of 3,251 markers were excluded from association analysis in this step.

1.6.7 Final counts of autosomal SNPs passing QC

After the above QC procedure for selection of SNPs, a total of 474,019 autosomal SNPs were used in subsequent association analysis by PDT.

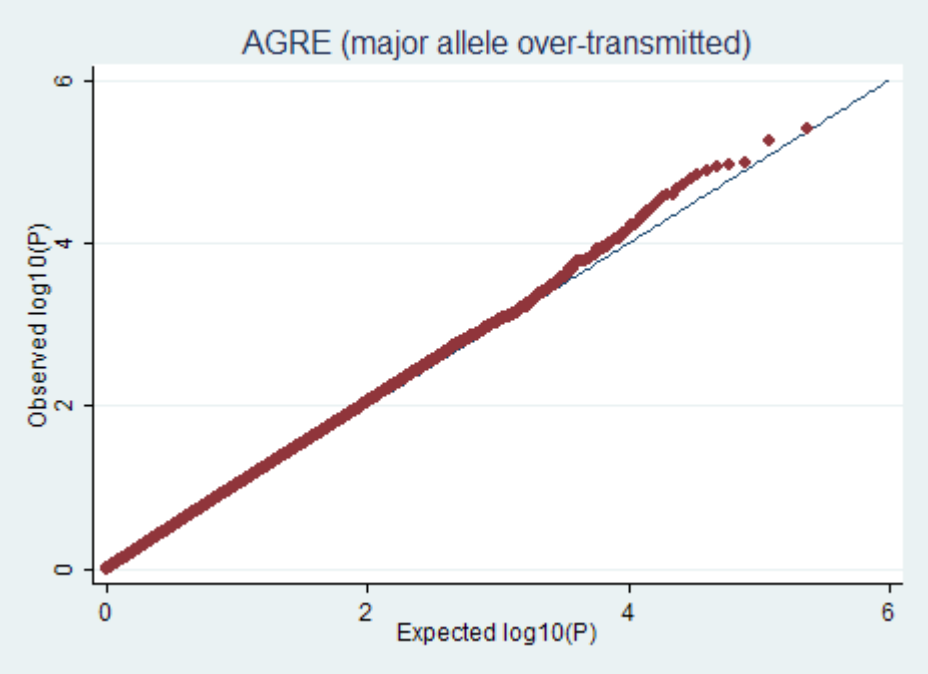
1.6.8 Examination of SNP association results

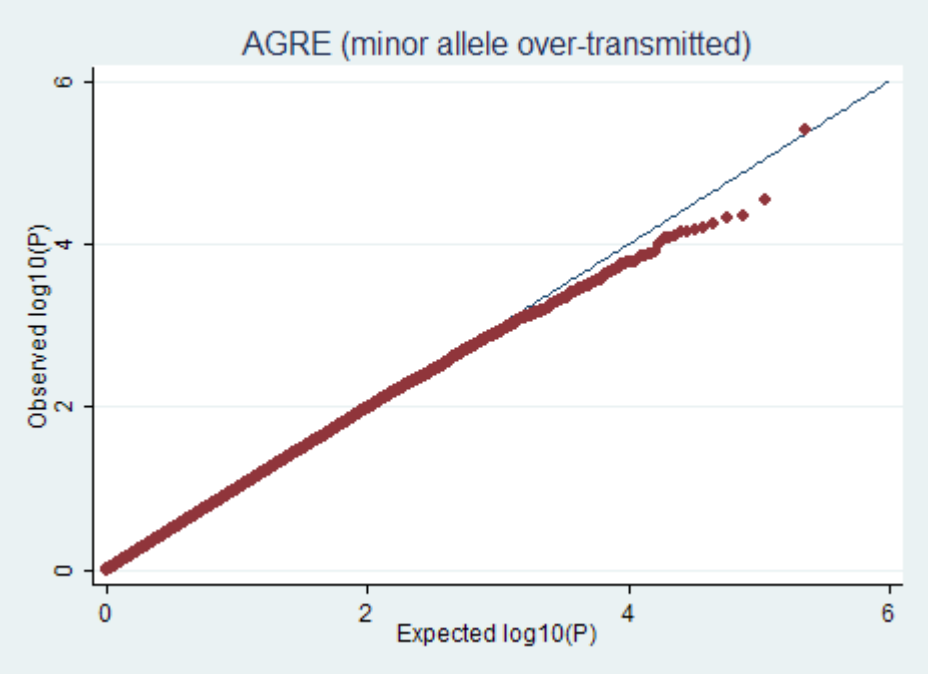
To detect potential problems with the SNP association tests, we examined the QQ-plot of P-values and did not find evidence of systematic inflation of P-values. The PDT test is based on Z-score test statistic rather than chi2 test statistic, which is used in conventional genomic inflation factor calculations [6]. The median P-value is 0.493, which can be transformed to a 1-df chi2 value of 0.47, so the genomic inflation factor [6] is calculated as 1.031.



Furthermore, to detect whether there is systematic differences in over-transmission of major versus minor alleles, we also plotted the QQ-plot for SNPs with over-transmitted

major alleles and minor alleles, respectively. The transformed genomic control inflation factors are calculated as 1.096 and 1.018, respectively. No systematic difference can be observed in the comparison of the two figures, while slight differences in inflation might be due to the bias of genotype calling algorithms: genotypes with major alleles (including homozygotes of major alleles and heterozygotes) are more likely to be called erroneously, so major alleles are more likely to have higher counts of over-transmission [7].





1.6.9 Handling SNPs in chromosome X

The PDT method used in autosomal analysis cannot be applied on chrX SNPs, so we handled chrX SNPs separately in a different analysis by X-APL [8]. We used PLINK to impute the sex for all subjects and eliminated the five subjects whose sex cannot be reliably imputed (the number of subjects drop from 3,101 to 3,096). Furthermore, we have identified father-mother switches based on PLINK imputation and lack of Mendelian inconsistency, and therefore manually switched the father and mother. The QC procedures on SNPs were identical to those applied on autosomal markers, and the final analysis includes 12,845 markers.

1.6.10 Handling SNPs in chromosome Y

Although we did not analyze markers on the Y chromosome (chrY) in the AGRE cohort, we have performed case-control association analysis on chrY markers in the ACC cohort using PLINK [4], treating these markers as if they were chromosome X markers. A total of 10 chrY markers were identified in the HumanHap550 array and were analyzed in our study: rs1865680, rs2032590, rs2032597, rs2032612, rs2032617, rs2032621, rs2032624, rs2032652, rs2058276, rs3848982. (The SNP rs2032635 in HumanHap550 v1 array was removed from HumanHap550 v3 array by design). The association analysis was performed on 989 subjects with ASDs and 3,391 control subjects, all of whom were male subjects.

1.6.11 Handling SNPs in pseudo-autosomal regions of sex chromosomes

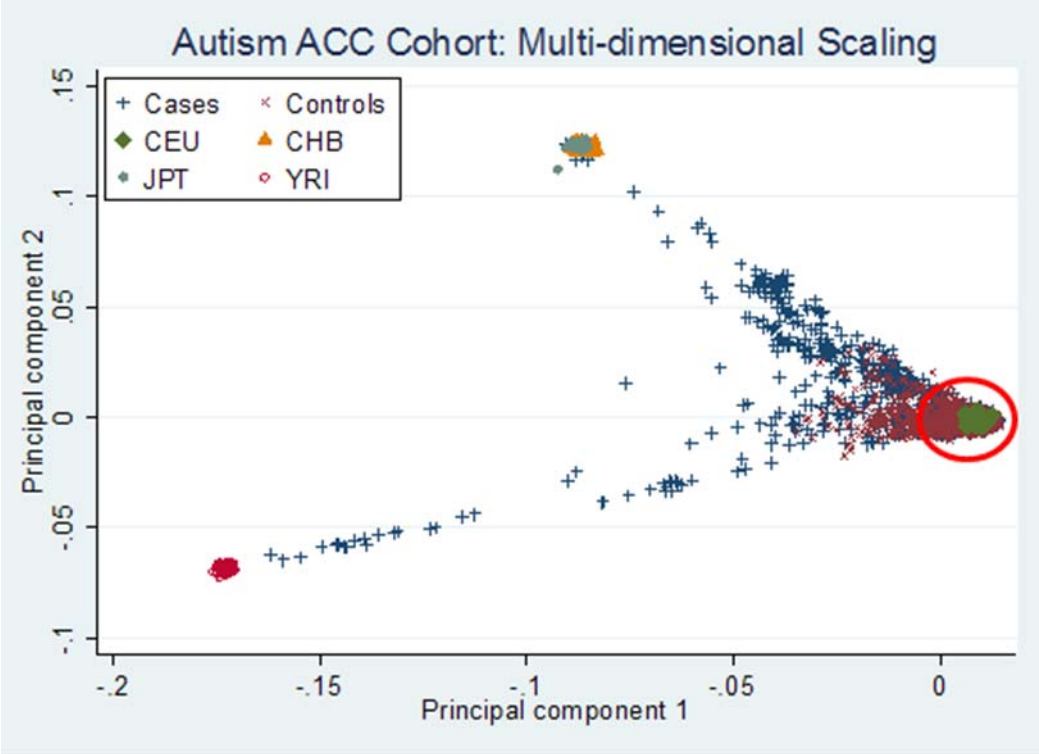
We have also analyzed SNP association in pseudo-autosomal regions of sex chromosomes (chrXY markers), by treating these markers as if they were autosomal markers. A total of 15 chrXY markers were annotated in the HumanHap550 array and were analyzed in our study: rs17148876, rs17148878, rs1764581, rs17653586, rs17719702, rs17792825, rs17842869, rs17842890, rs17842893, rs2738388, rs4933045, rs5949188, rs5983854, rs5989732, rs6567787.

1.7 Quality control for the ACC cohort

The quality control procedure for the ACC cohort is largely similar to those performed on the AGRE cohort. Here we describe several different aspects of QC that were applied on the ACC cohort.

1.7.1 Population stratification

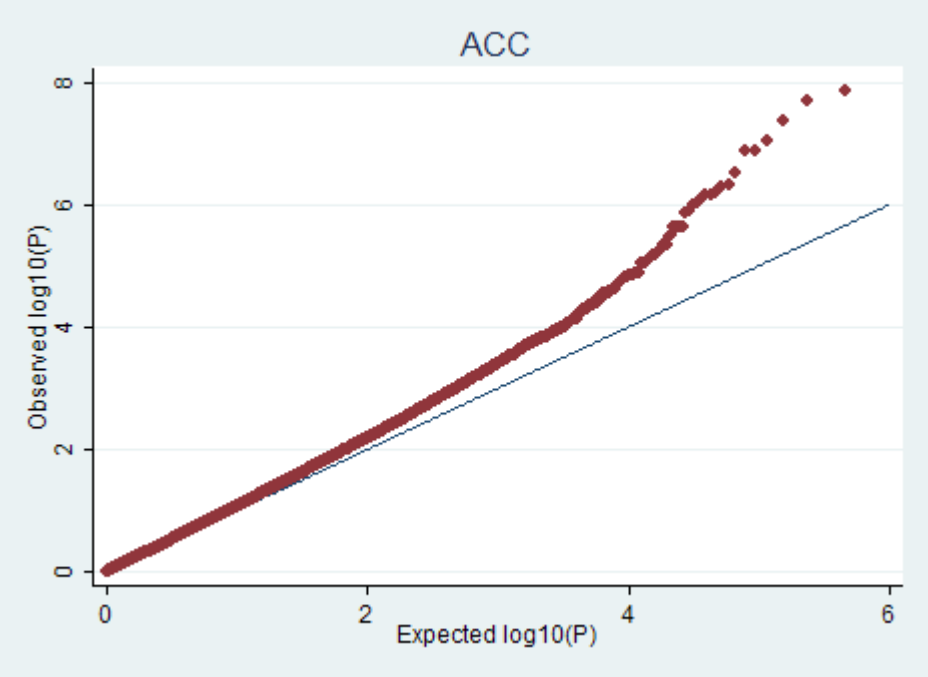
We applied the PLINK software for generation of genome-wide IBS estimates between all subjects (including both cases and controls), and then generated multi-dimensional scaling (MDS) plots for visual examination of population outliers. To help interpret the population genetic analysis, we have included 270 spike-in HapMap individuals as positive controls (labeled as CEU, CHB, JPT, YRI below) into the MDS analysis. The individuals of European ancestry are selected by the Principal component 1 of more than -0.01 and Principal component 2 of less than 0.03 (see red circle below).



The quality of the data for ACC cohort was screened by a series of routine analyses. Individual SNPs were excluded from further analysis if they deviated from Hardy-Weinberg equilibrium with a P-value of less than 0.001, an individual SNP genotype yield of less than 95%, or a minor allele frequency of less than 5%. In addition, subjects were also removed if their genotype yield is less than 95% (excluding 26 subjects). These procedures were identical as those applied in the AGRE data set.

Furthermore, we have plotted the QQ-plot for the ACC cohort to check whether there is systematic inflation of P-values indicating potential stratification (see below). The genomic inflation factor (lambda value) is calculated as 1.12, which is similar to other

studies using population-based control subjects (that is, control subjects were not collected/genotyped in the same batch as the case subjects). To further address the concerns on population stratification, we have also applied EigenStrat software [9] to re-perform all association tests on the case and control subjects passing the QC threshold above. The P-values for the SNPs reported in Table 2 are all within 10-fold differences, further implicating the effectiveness of MDS approach in removing population outliers. Nevertheless, in the combined analysis, we followed the well-established standards proposed by de Bakker et al [10], by scaling down the Z-scores for the ACC cohort by 1.058 (the square root of lambda value).



1.7.2 Detection and elimination of cryptic relatedness and duplicated genotyping

We have calculated genome-wide IBS estimates for all pairwise comparisons among all case subjects and control subjects. To detect cryptic relatedness and potential duplicated genotyping within our data sets, we have applied a two-step procedure to calculate pairwise IBD estimates between all individuals. First, we examined MDS and only keep in our data sets those individuals of inferred European ancestry, with call rates greater than 95%; second, we re-calculate genome-wide IBS estimates and re-calculate the IBD estimates among individuals of inferred European ancestry. This two-step procedure ensures that allele frequency differences between populations do not lead to biases in IBD estimations. We applied a stringent threshold for detecting cryptic relatedness: any pairs of subjects with IBD>0.15 were processed such that only one of the subjects remained in the final association test.

1.7.3 Final counts of subjects passing QC

The QC procedure resulted in the use of 1,204 cases, 6,491 controls and 480,530 SNPs in the subsequent association analysis.

1.8 CAP cohort

The Collaborative Autism Project (CAP) cohort was used as a replication case series, which is genotyped by the HumanHap1M array with ~1 million markers, including all

markers in the 550K array used in the discovery phase. The data set included 1,537 samples from 487 families with children affected with autism; after QC, association analyses were performed on 1,390 subjects in 447 families of European ancestry. The detailed sample ascertainment scheme is described in detail below.

The autism patients and their affected and unaffected family members were ascertained as part of through four clinical groups at the Miami institute for Human Genomics (MIHG, Miami, Florida), University of South Carolina (Columbia, South Carolina), W.S. Hall Psychiatric Institute (Columbia, South Carolina) and Vanderbilt Center for Human Genetics Research (Vanderbilt University, Nashville, Tennessee). Participating families were enrolled through a multi-site study of autism genetics and recruited via support groups, advertisements, and clinical and educational settings. All participants and families were ascertained using a standard protocol. These protocols were approved by appropriate Institutional Review Boards. Written informed consent was obtained from parents as well as from minors who were able to give informed consent; in individuals unable to give assent due to age or developmental problems, assent was obtained whenever possible.

Core inclusion criteria were as follows: (1) Chronological age between 3 and 21 years of age; (2) Presumptive clinical diagnosis of autism; (3) Expert clinical determination of

autism diagnosis using DSM-IV criteria supported by the Autism Diagnostic Interview (ADI-R [2]), in the majority of cases and all available clinical information. The ADI-R is a semi-structured diagnostic interview which provides diagnostic algorithms for classification of autism. All ADI-R interviews were conducted by formally trained interviewers who have achieved reliability according to established methods. Thirty-eight individuals were missing an ADI-R. For those cases we implemented a best estimate procedure using all available information from the research record and data from other assessment procedures. This information was reviewed by a clinical panel led by a clinical psychologist with over 20 years of autism-specific experience (MLC) and included two other psychologists and a pediatric medical geneticist—all of whom were experienced in autism. Following review of case material the panel discussed the case until a consensus diagnosis was obtained. Only those cases in which a consensus diagnosis of autism was reached were included. Minimal developmental level of 18 months as determined by the Vineland Adaptive Behavior Scale (VABS) or the VABS-II or IQ equivalent > 35. These minimal developmental levels assure that ADI-R results are valid and reduce the likelihood of including individuals with severe mental retardation only. We excluded participants with severe sensory problems (e.g., visual impairment or hearing loss), significant motor impairments (e.g., failure to sit by 12 months or walk by 24 months), or identified metabolic, genetic, or progressive neurological disorders.

1.9 CART cohort

The CART (Center for Autism Research and Treatment) cohort was used as a replication case series, which is genotyped on the Illumina HumanCNV370 arrays. This array is a SNP genotyping array with >300K SNP markers, but is also supplemented by non-polymorphic markers for the purpose of detecting copy number variations (CNVs). The majority of the markers from the HumanHap550 array are also present in this array. For the remaining markers that were not in the array but used in the discovery cohorts, we subsequently used genotype imputation (described in section 1.13 below) to infer the genotypes on untyped SNP markers.

The case subjects in the replication data were comprised of participants recruited through the UCLA Autism Evaluation Clinic, parent groups and organizations for families of children with autism, and a database maintained by the MIND (Medical Investigation of Neurodevelopmental Disorders) Institute at University of California Davis. Autism Spectrum Disorder diagnoses were confirmed through an assessment at the Autism Evaluation Clinic that included the Autism Diagnostic Interview-Revised (ADI-R [2]), the Autism Diagnostic Observation Schedule (ADOS [11]), and DSM-IV criteria. At UC Davis, diagnoses were verified through a review of medical records, or when records were incomplete or contradictory, through direct assessments utilizing the ADOS and ADI-R. DNA was prepared from blood or swabs acquired either in the homes of families or at the MIND Institute. A total of 110 families were recruited and genotyped, but some families

have multiple affected subjects. For association analysis, only one ASD subjects per family were chosen in a random fashion. Multi-dimensional scaling analysis excluded two subjects due to evidence of having African American ancestry.

The control subjects in the replication cohort were retrieved from the Illumina iControlDB database without evidence of documented diseases. In total we retrieved 733 subjects genotyped by the same platform, and were of self-declared Caucasian ancestry. To ensure that the ancestry background of cases and control subjects are comparable to each other, we used the IBD/IBS clustering procedure in PLINK and selected 540 control subjects for the association analysis, based on a 1:5 genetic matching scheme.

1.10 Association tests used in the study

1.10.1 Pedigree Disequilibrium Test (PDT)

The association analysis for the AGRE cohort is performed by the PDT software version 6, which implements the Pedigree Disequilibrium Test [12,13]. Custom scripts were used to convert the standard genotype data into formats that can be read by the PDT software, to zero out Mendelian errors (since PDT was unable to handle Mendelian errors correctly), and to pad parental genotype data as missing data for parents whose genotype information were not available. All default parameters were used in the association

analysis. The PDT needs either: (1) both parents genotypes and one or more affected offspring, or (2) a discordant (one affected, one unaffected) sibpair. Other families were not used in the analysis. The test statistic is given as Z-score, and the P-value is calculated based on the Z-score.

1.10.2 X-association in the presence of linkage (X-APL)

The PDT method cannot be used for analysis of chromosome X markers in the AGRE cohort. The X-linked sibling transmission/disequilibrium test (XS-TDT) and the reconstruction-combined TDT for X-chromosome markers (XRC-TDT) cannot be applied on multiplex families, while the XPDT (extension of PDT for chrX) have low power in the presence of missing parental genotype information. Therefore, we applied the X-APL approach [8], which uses singleton or multiplex families and properly infers missing parental genotypes in linkage regions by considering identity-by-descent parameters for affected siblings.

The X-APL [8] software version 1.1 was used in the association analysis for markers in chromosome X. Custom scripts were used to convert the standard genotype data into formats that can be read by the software, to zero out Mendelian errors, and to pad parental genotype data as missing data for parents whose genotype information were not available. All default parameters were used in the association analysis. The test statistic is given as Z-score, and the P-value is calculated based on the Z-score.

1.10.3 Family-Based Association Test (FBAT)

To cross-check the association results calculated by the PDT software, we have also applied a different algorithm as implemented in the FBAT (Family-based association test) software [14] on the AGRE data set. Similar to PDT, the FBAT software can use both nuclear family information and discordant sibpair information in the association test. We have adopted all default parameters in the FBAT software (FBAT automatically zero out Mendelian errors detected in families), with additive model, bi-allelic test. The association results for the most significantly associated markers in Table 2 are listed below. The results are overall similar, but slightly more significant than the PDT results. The Z-score indicates the direction of association: positive value indicates over-transmission of major allele.

Marker	minor/major allele	MAF	Z	P
rs4307059	C/T	0.368	4.495	6.96E-06
rs7704909	C/T	0.374	4.221	0.000024
rs12518194	G/A	0.37	4.363	0.000013
rs4327572	T/C	0.371	4.283	0.000018
rs1896731	C/T	0.351	-2.826	4.71E-03
rs10038113	C/T	0.413	-2.894	0.003808

1.10.4 FBAT assuming linkage

We also tested a different FBAT model, by taking into account of potential linkage, when testing for association. These results are largely concordant with those generated by default parameters. The Z-score indicates the direction of association: positive value indicates over-transmission of major allele.

Marker	minor/major allele	MAF	Z	P
rs4307059	C/T	0.368	4.468	7.89E-06
rs7704909	C/T	0.374	4.166	0.000031
rs12518194	G/A	0.37	4.350	0.000014
rs4327572	T/C	0.371	4.292	0.000018
rs1896731	C/T	0.351	-2.766	0.005668
rs10038113	C/T	0.413	-2.818	0.004829

1.10.5 Association tests for case-control studies

The PLINK software was used for case-control comparisons. In Table 2 of the manuscript we presented the results for allelic association test, here we provide the P-values for all the genetic models that were tested.

SNP	Allelic P-value	Cochran-Armitage P-value	Genotypic P-value	Dominant P-value	Recessive P-value
rs4307059	0.000222	0.000246	0.000496	0.000118	0.06256
rs7704909	0.000623	0.000569	0.002445	0.001413	0.02119
rs12518194	0.001014	0.000973	0.003687	0.001592	0.03847
rs4327572	0.002025	0.002015	0.007765	0.003729	0.04304
rs1896731	0.001647	0.001633	0.001872	0.000394	0.2567
rs10038113	0.002426	0.002416	0.009487	0.01389	0.01203

1.10.6 Additional QC vanity check on the most significant SNPs

We examined minor allele frequency (MAF), Hardy-Weinberg equilibrium, Non-call rate, Mendelian error rate on the six markers in four cohorts. The MAF and HWE are calculated for founders only. None of the QC measures flag potential genotyping error.

AGRE cohort:

SNP	Founder MAF	Founder HWE P-value	Missing rate	Mendelian error rate (fraction of trios)
rs4307059	0.3845	0.4758	0.01161	0
rs7704909	0.3904	0.9068	0	0.001019
rs12518194	0.3868	0.9531	0.000323	0
rs4327572	0.3882	0.7688	0.00129	0
rs1896731	0.3437	0.6211	0	0
rs10038113	0.4036	0.224	0	0

ACC cohort:

SNP	Case MAF	Control MAF	HWE P-value (control)	Missing rate
rs10038113	0.4273	0.3943	0.8557	0.00026
rs12518194	0.3552	0.3908	0.4338	0.001559
rs1896731	0.3771	0.3438	0.6204	0.00039
rs4307059	0.3485	0.3892	0.5612	0.025991
rs4327572	0.3588	0.3922	0.8146	0.00117
rs7704909	0.3588	0.3958	0.2033	0.00078

CAP cohort:

SNP	Founder MAF	Founder HWE P-value	Missing rate	Mendelian error rate (fraction of trios)
rs4307059	0.3554	0.826	0.005036	0
rs7704909	0.364	0.6124	0	0
rs12518194	0.3556	0.7698	0	0
rs4327572	0.3573	0.5592	0.000719	0
rs1896731	0.3607	0.4673	0	0
rs10038113	0.4101	0.6283	0.000719	0

CART cohort:

SNP	Case MAF	Control MAF	HWE P-value (control)	Missing rate
rs1896731	0.3286	0.3284	1	0.006173
rs12518194	0.3148	0.4007	0.4729	0.001543
rs4307059	0.3148	0.4019	0.4721	0.007716
rs4327572	0.3148	0.4028	0.4741	0

1.10.7 Meta-analysis of association results

The conventional weighted Z-score approach was used for calculating the combined P-values in meta-analysis. The weight for each cohort is calculated as sqrt(N_i/N_{total}) and then summed together, while the sign of the Z-score indicates the direction of association. For the family-based cohorts, the Z-scores can be directly calculated by the PDT

software. For case-control cohort, the Z-score is calculated as the square root of chi2 values, and the sign is assigned based on the odds ratio.

Based on the recently proposed standard for GWAS meta-analysis by de Bakker et al [10], the following specific adjustments were made to the meta-analysis procedure: (1) Since the study design in the ACC cohort contains asymmetric number of cases and controls, the effective sample size may be over-estimated by pooling the total sample size together; therefore, we used the method proposed by de Bakker et al to infer sample size when a symmetric design is used, based on the non-centrality parameter calculated by the Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/). This approach effectively down-weights the contribution of ACC Z-scores to the combined analysis. (2) To address the potential stratification based on the genomic control lambda value [6], we have scaled down the Z-scores by the square root of lambda value, before calculating the sum of weighted Z-scores.

1.11 In situ hybridization on human fetal brain

We evaluated the mRNA distribution for *CDH10* and *CDH9* in the developing human brain by *in situ* hybridization, as previously described [15,16]. Multiple sagittally sectioned human fetal brains, each between 19 and 20 weeks gestation, were obtained from the Developmental Brain and Tissue Bank at the University of Maryland.

The riboprobe for *CDH9* corresponds to bps 39-2972 of NM_016279.3, and the riboprobe for *CDH10* corresponds to bps 1701-3205 of NM_006727. These riboprobes were hybridized against multiple sections from separate brains. Sense controls were tested on adjacent sections, which gave no signal.

1.12 CNV validation

We attempted validation of CNVs by two experimental techniques (TaqMan quantitative PCR and multiplex ligation-dependent probe amplification assay, or MLPA) at two different sites. The detailed experimental protocol was described below:

For MLPA, we used the Universal Probe Library (UPL) system from Roche (Roche, Indianapolis, IN) to perform genomic quantitative PCR. Primers to be used with the UPL probes were designed using gene sequences from the UCSC Genome Browser and the ProbeFinder application. Ten microliters reactions were assembled with 25 ng DNA, 400 nM of each primer, 100 nM UPL probe (Roche, NJ), and 1xPlatinum Quantitative PCR SuperMix-Uracil-D-Glycosylase (UDG), with Rox (Invitrogen, CA). All reactions were performed in triplicate with an ABI PrismTM 7900HT sequence detection system (Applied Biosystems, CA). In addition to the genes to be assayed for copy number, we

included four reference genes (COBL, GUSB, PPIA, SNCA) on each plate and used qBase for subsequent analysis. The primer sequences were given below:

Primer name	Sequence	Universal Probe Library probe number
1 Fw	ttccatgcagtgatgtgacc	61
1 Rv	ggcgtggaacaccaagtatt	
2-1_Fw	tctctctttccccactgtgc	55
2-1_Rv	ctggagtctgatgcccaag	
2-2_Fw	ccctcccttaacccccctac	81
2-2_Rv	catggacacagggagaggat	
3-3_Fw	ccccttttctccagagttca	35
3-3_Rv	aaatcctcacttcggggaac	
3-5_Fw	tggctcagcacagcaaaa	33
3-5_Rv	tcccttcaaaggggggaag	
COBL_Fw	atggacgtcaccgtggtc	19
COBL_Rv	ggctctcacctcccattg	
GUSB_Fw	cccagtgatgggagtgct	77
GUSB_Rv	cacagccctcagccaaag	
PPIA_Fw	ttgcttgagcctagagtgagc	59
PPIA_Rv	gcctctgcctacctttgaga	
SNCA_Fw	ggctttgtcaaaaaggacca	76
SNCA_Rv	ccagcttataaatgtaacacaaaacg	

QPCR (TaqMan) probes were custom-designed using Primer Express 3.0 (Applied Biosystems, Foster City, CA) with default parameters. Amplification primers were synthesized by IDT (Coralville, IA) and probes were made by Applied Biosystems with FAM as the reporter and a non-fluorescent quencher (NFQ) and minor-groove binder (MGB). Reactions were set up in triplicate using 10 ng of genomic DNA in a 10 ul reaction which contained 200 nM concentration of probe, 900 nM of each amplification primer and 1 X of Real-time PCR Master Mix (Applied Biosystems). The samples were amplified on an Applied Biosystems 7900HT Sequence Detection System using standard cycling conditions and data collected and analyzed with SDS 2.3 software. Standard curves were constructed using serial 2-fold dilutions of genomic DNA from an individual without CNVs and used to estimate amounts of DNA in the experimental samples from their cycle threshold (Ct) values. Ratios of amounts were calculated from the assay designed in the area of predicted CNVs and a nearby region in 5p14 without CNVs. This normalized amount was then compared to the value in a control calibrator sample to produce a fold change ratio (normal =1) and multiplied by 2 to generate a copy number (normal = 2).

Primer info (Sup Figure 3, panel b):

Forward Primer: AAGCTGAGCAATTTGTTCTTCA (Tm=59°C)
Reverse primer: GCATCCTCTCCCTCAGAATTA (Tm=58°C)
Probe: TGCACTGTCATCCATACACACTTGGTC (Tm=68°C)

Primers info (Sup Figure 3, panel d):

Forward primer: TCATCCTCCTTAGCCACCTC (Tm=59°C)

Reverse primer: GGGAAGCAGAATAACCAAGC (Tm=59°C)

Probe: AGGTCATCCTCTTCAATGCACTCATCA (Tm=68°C)

1.13 Genotype imputation

We used the Markov Chain Haplotyping (MACH) software (http://www.sph.umich.edu/csg/abecasis/MaCH/index.html) for genotype imputation on markers that are not present in the genotyping platforms. The software version 1.0.16 was used in the study, and the default two-step procedure was adopted for imputation. The software requires several input files for SNPs and phased haplotypes, so we used the HapMap phased haplotypes (release 22) on CEU subjects, as downloaded from the HapMap database (http://www.hapmap.org).

The software generates several files, including mlinfo, mlprob, mlqc and mlgeno files, from the imputation. We analyzed the mlinfo file, and used the recommended Rsq threshold of 0.3 to flag unreliable markers used in imputation analysis, and removed these markers from association tests. (A total of 61.8K and 86.1K imputed SNPs were removed from the AGRE and the ACC cohort, respectively.) We also analyzed the mlqc

file, which provides a per-genotype posterior probability for each imputed calls, and we used the 0.9 threshold to flag unreliable calls (by recoding them as NoCall genotype).

For family-based studies, the imputed genotype calls were directly used in association analysis: unreliable calls were further zeroed out as NoCall genotypes during the Mendelian check step in association tests. For case-control association tests, to take into account of imputation uncertainty, we have used the SNPTEST software (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html), which can explicitly use the imputation uncertainty information. We converted the mlprob files from MACH imputation results into the SNPTEST input format, where each SNP call has three posterior probabilities associated with it (corresponding to AA, AB and BB genotypes), and these quantitative information were used in association tests (through `-proper` arguments in SNPTEST).

In addition, we have also explored the differences in using quantitative values of genotype imputation probability for association tests in SNPTEST, versus directly using the most likely genotype calls (>0.9 posterior probability) in PLINK. We have found that these results are largely identical: The Pearson's correlation coefficient for logP values is 0.96 between the two tests, and the Spearman's rank correlation coefficient is 0.94. Nevertheless, in the study, we report the results generated by SNPTEST additive model, as it properly handles genotype imputation uncertainty.

The identical suite of QC procedures for genotyped markers was applied on the imputed genotype data, after removing markers that cannot be reliably imputed. In total, 1.4 million imputed autosomal markers (out of 2.3 million markers with phase information in HapMap CEU population) passed the QC for both the AGRE and ACC cohort. Although no additional markers pass genome-wide significance, we provide all loci with $P<1x10^{-5}$ in Supplementary Table 4.

1.14 Pathway-based association analysis

Two types of analysis were performed in our study to determine whether a gene set/pathway is collectively associated with the trait of interest (ASDs). These procedures are described in detail below.

The first approach requires P-values for all SNPs (but not the raw genotype data), so it can be applied to the combined P-values from two discovery cohorts. In the approach, we used the RefGene annotation from the UCSC genome browser database, and assigned each SNP to its overlapping gene or its closest gene (if the SNP is located in an intergenic region). We then summarize the significance of each gene using a Simes method [17-19] that computes a single P-value from multiple SNPs. For L SNPs ranked by their P-value, $p_{(1)}$,

…, $p_{(L)}$, the Simes P-value is calculated as $\min\{p_{(i)}*L/i\}$, where $1 \le i \le L$. The Simes method provides an overall P-value for the entire collection of L hypotheses. Finally, we can then test whether the distribution of P-values differs between a group of genes (for example, cadherin genes) and all other genes using a non-parametric rank sum test.

The second approach was a previously described pathway-association approach [20]. This method requires the use of raw SNP genotyping data and the shuffling of phenotypes, so it can only be applied on the ACC case-control data set. Briefly, we associate each SNP to the overlapping gene/genes, or its closest gene in the genome if it does not overlap with a gene. For each gene, we assigned the highest statistic value (chi-square value) among all SNPs mapped to the gene as the statistic value of the gene. For all the N genes that are represented by SNPs in the GWA study, we sorted their statistic values from the largest to the smallest, denoted by $r_{(1)}$, …, $r_{(N)}$. For any given gene set S composed of N_H genes, we then calculated a weighted Kolmogorov-Smirnov-like running sum statistic[21] that reflects the overrepresentation of genes within the set S at the top of the entire ranked list of genes in the genome,

$$ES(S) = \max_{1 \le j \le N} \left\{ \sum_{G_{j^*} \in S, j^* \le j} \frac{\left| r_{(j^*)} \right|^p}{N_R} - \sum_{G_{j^*} \notin S, j^* \le j} \frac{1}{N - N_H} \right\}$$

where $N_R = \sum_{G_{j^*} \in S,} \left| r_{(j^*)} \right|^p$ and p (default value=1) is a parameter that gives higher weight to genes with extreme statistic values. To adjust for differences in gene size (hence the

different number of SNPs located within/nearby each gene), as well as the linkage disequilibrium between SNPs within the same gene, we conducted a two-step correction procedure. In the first step, we permuted the disease labels of all samples ensuring the same number of individuals in each phenotype group for case-control studies. During each permutation (denoted by π), we repeated the calculation of enrichment score as described above as $ES(S, \pi)$. In the second-step, we calculated normalized enrichment score (*NES*) as a Z-score, defined as ($ES(S)$-mean($ES(S, \pi)$)) / SD($ES(S, \pi)$), so that different gene sets are directly comparable to each other. The P-value can be calculated from the π permutations, based on the fraction of permutations that generates a more extreme enrichment scores. In our study, 2,500 permutation cycles were used on the raw genotype data from the ACC cohort.

The pathway association analysis were performed on a few artificial gene sets: (1) we combined the 25 cadherin genes (*CDH1* , *CDH2*, *CDH3*, … , *CDH26*, except *CDH14* which is not annotated in the human genome) into a group; (2) we also combined 25 cadherins and 8 neurexin family members (*NRXN1* , *NRXN2*, *NRXN3*, *CNTNAP1*, *CNTNAP2*, *CNTNAP3*, *CNTNAP4*, *CNTNAP5*) into a single cell-adhesion gene set.

Furthermore, we also explored combining 68 protocadherins into a group to test for association. However, no enrichment signal was detected for this group of genes, many of which are tightly clustered at the same genomic region. These genes include *PCDH1,*

PCDH10, PCDH11X, PCDH11Y, PCDH12, PCDH15, PCDH17, PCDH18, PCDH19, PCDH20, PCDH21, PCDH24, PCDH7, PCDH8, PCDH9, PCDHA1, PCDHA10, PCDHA11, PCDHA12, PCDHA13, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHA8, PCDHA9, PCDHAC1, PCDHAC2, PCDHB1, PCDHB10, PCDHB11, PCDHB12, PCDHB13, PCDHB14, PCDHB15, PCDHB16, PCDHB2, PCDHB3, PCDHB4, PCDHB5, PCDHB6, PCDHB7, PCDHB8, PCDHB9, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA12, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, PCDHGC3, PCDHGC4, PCDHGC5.

1.15 URLs for the software/webserver used in the study

Below we give the URL and citation for the software used in the study:

BeadStudio: http://www.illumina.com/pages.ilmn?ID=35

Database for Genomic Variants [22]: http://projects.tcag.ca/variation/

Eigenstrat [9]: http://genepath.med.harvard.edu/~reich/Software.htm

FBAT (Family-based association test) [23]: http://biosun1.harvard.edu/~fbat/fbat.htm

GNF SymAtlas [24,25]: http://symatlas.gnf.org

HaploView [26]: http://www.broad.mit.edu/mpg/haploview/

HapMap [27]: http://www.hapmap.org

PDT (Pedigree Disequilibrium Test) [12]:

http://www.mihg.org/weblog/core_resources/2007/11/statistical-programming-core.html

PLINK [4]: http://pngu.mgh.harvard.edu/~purcell/plink/

Pathway-association software [20]: http://www.openbioinformatics.org/gengen/

PennCNV [5]: http://www.openbioinformatics.org/penncnv/

PhastCons [28]: http://compgen.bscb.cornell.edu/phast/phastCons-HOWTO.html

Regional association Plot [29]: http://www.broad.mit.edu/node/555

SNPExpress [30]: http://people.genome.duke.edu/~dg48/SNPExpress/

SNPTEST [31]: http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html

UCSC Genome Browser [32]: http://genome.ucsc.edu/

WGAViewer [33]: http://people.genome.duke.edu/~dg48/WGAViewer/

X-APL (X-Association in the Presence of Linkage) [8]:

http://www.mihg.org/weblog/core_resources/2007/11/statistical-programming-core.html

Supplementary Table 1. Association analysis on 10 markers in the Y chromosome in the ACC cohort. The association analysis was performed on 989 male subjects with ASDs and 3,391 male control subjects. The allele frequency and odds ratio were calculated with respect to A1 (allele 1).

SNP	Position	Missing rate	A1	A2	A1 Freq (cases)	A1 Freq (controls)	CHISQ	P	OR
rs2058276	2728456	0.0041	A	G	0.4823	0.4889	0.13	0.7159	0.97
rs1865680	6928118	0.0039	G	A	0.414	0.4237	0.30	0.5857	0.96
rs2032597	13357186	0.0037	C	A	0.2053	0.1532	15.05	0.0001045	1.43
rs2032590	13529007	0.003	G	T	0.001011	0.000592	0.20	0.6582	1.71
rs2032624	13535818	0.0256	A	C	0.3962	0.4167	1.28	0.2572	0.92
rs3848982	20176596	0.0041	A	G	0.06079	0.08148	4.61	0.03176	0.73
rs2032612	20325879	0.008	T	C	0	0	NA	NA	NA
rs2032621	20332126	0.0039	C	T	0	0	NA	NA	NA
rs2032617	20355649	0.0062	T	G	0	0.000892	0.88	0.3478	0.00
rs2032652	20376701	0.0037	C	T	0.06275	0.08235	4.08	0.04329	0.75

Supplementary Table 2. Association analysis on 15 markers in pseudoautosomal regions in the discovery cohorts. These markers were analyzed in the same procedure as autosomal markers. The allele frequency, OR (Odds Ratio) and the Z-score were calculated with respect to A1 (allele 1).

SNP	A1	A2	HWE P-value (ACC control)	Missing rate (ACC)	A1 Freq (ACC cases)	A1 Freq (ACC controls)	P (ACC)	OR (ACC)	Missing rate (AGRE)	HWE P-value (AGRE parents)	A1 Freq (AGRE parents)	P (AGRE)	Z (AGRE)	P (combined)
rs4933045	A	G	0.50	0.03	0.32	0.32	0.94	1.00	0.03	0.56	0.31	0.20	1.28	0.37
rs2738388	T	G	0.83	0.00	0.21	0.22	0.16	0.93	0.00	0.36	0.23	0.85	-0.19	0.26
rs17792825	A	G	0.07	0.00	0.15	0.17	0.09	0.90	0.00	0.75	0.16	0.56	-0.59	0.11
rs17719702	C	T	0.44	0.00	0.34	0.34	0.91	1.01	0.00	1.00	0.35	0.85	0.18	0.84
rs17148878	T	C	0.66	0.05	0.15	0.16	0.22	0.93	0.02	0.74	0.16	0.50	-0.68	0.19
rs17148876	T	C	0.86	0.00	0.12	0.12	0.32	0.93	0.00	0.58	0.12	0.43	0.79	0.85
rs5989732	T	G	0.04	0.01	0.14	0.13	0.27	1.08	0.01	0.25	0.15	1.00	0.00	0.44
rs5949188	C	A	0.31	0.01	0.26	0.28	0.04	0.90	0.01	0.23	0.25	0.07	1.82	0.80
rs17842869	T	C	0.55	0.00	0.16	0.16	0.76	1.02	0.00	0.69	0.16	0.43	0.80	0.46
rs17842890	G	A	0.27	0.00	0.03	0.04	0.27	0.87	0.00	0.28	0.04	0.27	1.10	0.95
rs17842893	A	G	0.27	0.00	0.03	0.04	0.30	0.88	0.00	0.29	0.03	0.25	1.16	0.98
rs17653586	T	G	0.59	0.00	0.16	0.15	0.07	1.12	0.00	0.83	0.13	0.94	0.08	0.17
rs1764581	T	C	0.18	0.01	0.44	0.44	0.97	1.00	0.01	0.21	0.45	0.48	-0.71	0.66
rs6567787	T	C	0.65	0.00	0.22	0.21	0.22	1.07	0.00	0.33	0.21	0.98	-0.02	0.39
rs5983854	C	A	0.46	0.00	0.45	0.43	0.09	1.08	0.00	0.39	0.42	0.13	-1.53	0.84

Supplementary Table 3. Imputation-based analysis on four cohorts identifies additional SNPs with $P<1x10^{-4}$ on the 5p14.1 region. The MACH Rsq measure estimates the squared correlation between imputed and true genotypes, and a threshold of 0.3 is recommended to flag poorly imputed SNPs. The ACC and CART cohorts were analyzed by SNPTEST, which takes into account of imputation uncertainty. A1 and A2 refer to allele 1 and allele 2, respectively, and Freq, Z-scores and Odds Ratios were calculated with respect to A1 allele.

(a) Imputed SNPs

SNP	MACH Rsq	A1	A2	AGRE			ACC			CAP		CART			P(all)
				Freq	Z	P	Z	P	OR	Z	P	Z	P	OR	
rs12521388	0.9625	A	G	0.35	-4.15	3.3E-05	-3.35	8E-04	0.85	-2.68	0.007	-2.31	0.021	0.69	1.52E-09
rs10942147	0.9625	A	G	0.35	-4.15	3.3E-05	-3.36	7E-04	0.85	-2.65	0.008	-2.31	0.021	0.69	1.59E-09
rs4701511	0.9732	C	A	0.36	-4.15	3.3E-05	-3.43	6E-04	0.84	-2.61	0.009	-2.06	0.039	0.75	1.69E-09
rs4475231	0.9977	T	C	0.35	-4.19	2.8E-05	-3.29	1E-03	0.86	-2.60	0.009	-2.43	0.015	0.68	1.81E-09
rs17482975	0.9832	T	C	0.35	-4.04	5.4E-05	-3.38	7E-04	0.85	-2.72	0.007	-2.29	0.022	0.69	1.90E-09
rs7705715	0.9919	T	C	0.36	-4.15	3.3E-05	-3.42	6E-04	0.85	-2.57	0.01	-2.04	0.041	0.74	2.04E-09
rs13187934	0.9802	T	C	0.35	-3.99	6.5E-05	-3.37	7E-04	0.86	-2.72	0.007	-2.29	0.022	0.69	2.29E-09
rs7380139	0.9575	A	G	0.35	-4.10	4.2E-05	-3.31	9E-04	0.85	-2.63	0.008	-2.33	0.019	0.69	2.38E-09
rs4701259	0.9826	A	G	0.36	-4.10	4.1E-05	-3.33	8E-04	0.86	-2.61	0.009	-2.20	0.028	0.71	2.63E-09
rs12519594	0.9778	A	G	0.35	-4.10	4.2E-05	-3.24	0.001	0.86	-2.63	0.008	-2.38	0.017	0.70	3.01E-09
rs13176113	0.9828	A	G	0.36	-4.15	3.3E-05	-3.40	6E-04	0.85	-2.43	0.015	-2.05	0.04	0.74	3.08E-09
rs6452304	0.9568	T	C	0.35	-4.00	6.4E-05	-3.32	9E-04	0.85	-2.63	0.008	-2.33	0.019	0.69	3.28E-09
rs6452305	0.9568	A	C	0.35	-4.00	6.4E-05	-3.32	9E-04	0.85	-2.63	0.008	-2.33	0.019	0.69	3.33E-09
rs13166776	0.9955	C	T	0.35	-4.10	4.1E-05	-3.21	0.001	0.86	-2.54	0.011	-2.46	0.014	0.69	3.84E-09
rs6873221	0.9755	A	G	0.45	2.84	0.00445	3.67	2E-04	1.18	4.50	7E-06	-0.68	0.497	0.91	1.05E-08
rs9293194	0.9794	A	C	0.45	2.81	0.00492	3.62	3E-04	1.18	4.55	5E-06	-0.67	0.5	0.91	1.21E-08
rs11739167	0.9783	T	C	0.45	2.81	0.00492	3.62	3E-04	1.18	4.35	1E-05	-0.67	0.5	0.91	1.95E-08
rs1346536	0.9838	G	A	0.45	2.60	0.00936	3.63	3E-04	1.18	4.55	5E-06	-0.67	0.5	0.91	2.42E-08
rs6894102	0.9753	C	T	0.36	3.15	0.00163	3.25	0.001	1.16	3.95	8E-05	0.11	0.909	1.00	2.87E-08
rs10058083	0.9486	A	G	0.36	3.15	0.00163	3.29	1E-03	1.17	3.95	8E-05	-0.11	0.915	0.99	3.15E-08
rs6891206	0.9916	T	C	0.40	-3.44	0.00058	-3.12	0.002	0.87	-2.38	0.017	-3.09	0.002	0.62	4.09E-08
rs922551	0.9476	G	A	0.36	-3.10	0.00196	-3.57	3E-04	0.84	-1.84	0.065	-2.37	0.017	0.67	1.67E-07
rs409649	0.9583	G	A	0.34	-2.83	0.00469	-3.61	3E-04	0.84	-2.83	0.005	-0.68	0.495	0.91	1.80E-07
rs11740209	0.9915	C	T	0.38	-2.98	0.00293	-2.76	0.006	0.88	-2.76	0.006	-1.34	0.179	0.82	1.25E-06
rs12697669	0.9625	A	C	0.39	-3.11	0.00185	-2.36	0.018	0.89	-2.33	0.02	-2.70	0.007	0.65	2.27E-06
rs12173236	0.9713	T	C	0.20	1.87	0.06162	3.28	0.001	1.20	3.00	0.003	-0.33	0.74	0.95	1.37E-05
rs7720426	0.9916	A	G	0.20	1.87	0.06162	3.23	0.001	1.20	3.05	0.002	-0.32	0.75	0.95	1.40E-05
rs12187724	0.9889	C	A	0.36	1.60	0.10941	2.58	0.01	1.13	4.07	5E-05	0.48	0.631	1.08	1.55E-05

rs12187661	0.9939	T	C	0.44	1.31	0.1896	2.59	0.009	1.12	3.66	2E-04	2.38	0.017	1.42	1.59E-05
rs10063934	0.9799	A	G	0.36	1.48	0.13801	2.67	0.007	1.13	4.07	5E-05	0.42	0.675	1.03	1.66E-05
rs12659830	0.9868	T	G	0.36	1.34	0.18029	2.70	0.007	1.13	4.16	3E-05	0.46	0.643	1.08	1.84E-05
rs10214380	0.9958	T	C	0.44	-2.72	0.00658	-1.99	0.046	0.91	-2.08	0.037	-3.41	6E-04	0.59	1.86E-05
rs12516367	0.9621	C	T	0.20	1.84	0.06632	3.25	0.001	1.19	2.63	0.009	-0.43	0.671	0.94	3.45E-05
rs1330642	0.9461	T	C	0.21	1.83	0.06704	3.40	6E-04	1.21	2.45	0.014	-0.55	0.58	0.92	3.51E-05
rs6898772	0.951	C	T	0.20	1.84	0.06632	3.24	0.001	1.20	2.63	0.009	-0.43	0.665	0.94	3.52E-05
rs367519	0.9659	C	T	0.40	-2.22	0.02652	-2.23	0.025	0.90	-2.66	0.008	-1.23	0.22	0.84	6.01E-05
rs2619942	0.9764	A	G	0.40	-2.22	0.02652	-2.20	0.027	0.91	-2.66	0.008	-1.22	0.223	0.84	6.52E-05
rs2619941	0.982	A	G	0.40	-2.22	0.02652	-2.19	0.028	0.91	-2.66	0.008	-1.22	0.223	0.84	6.75E-05
rs374014	0.9876	G	A	0.40	-2.22	0.02652	-2.18	0.029	0.90	-2.66	0.008	-1.22	0.222	0.84	6.97E-05
rs10491401	0.991	T	C	0.40	-2.22	0.02652	-2.17	0.029	0.90	-2.66	0.008	-1.21	0.226	0.84	7.14E-05
rs443439	0.981	A	G	0.40	-2.17	0.02999	-2.20	0.027	0.89	-2.66	0.008	-1.23	0.219	0.84	7.36E-05
rs2619940	0.981	C	T	0.40	-2.20	0.02801	-2.18	0.029	0.91	-2.66	0.008	-1.22	0.221	0.85	7.38E-05
rs437316	0.9826	A	G	0.40	-2.17	0.02999	-2.19	0.028	0.90	-2.66	0.008	-1.23	0.22	0.84	7.50E-05
rs438137	0.9832	A	C	0.39	-2.28	0.02239	-2.06	0.039	0.91	-2.63	0.009	-1.02	0.307	0.87	9.76E-05
rs12521681	0.9629	A	G	0.39	-2.67	0.00769	-1.93	0.053	0.92	-2.19	0.028	-1.26	0.206	0.85	9.81E-05

(b) Genotyped SNPs

SNP	A1	A2	AGRE			ACC			CAP		CART			P(all)
			Freq	Z	P	Z	P	OR	Z	P	Z	P	OR	
rs4307059	T	C	0.62	4.40	1E-05	3.69	2E-04	1.1909	2.52	0.012	2.40	0.0166	1.46	2.07E-10
rs7704909	T	C	0.61	4.31	2E-05	3.42	6E-04	1.1708	2.61	0.009	2.05	0.0397	1.36	9.94E-10
rs12518194	A	G	0.61	4.36	1E-05	3.29	0.001	1.1641	2.60	0.009	2.37	0.0179	1.46	1.07E-09
rs4327572	T	C	0.39	-4.24	2E-05	-3.09	0.002	0.8673	-2.68	0.007	-2.42	0.0155	0.68	2.71E-09
rs1896731	T	C	0.66	-3.14	0.002	-3.15	0.002	0.8658	-3.95	8E-05	-0.01	0.9958	1.00	4.80E-08
rs10038113	T	C	0.60	-3.19	0.001	-3.03	0.002	0.8726	-4.19	3E-05	0.75	0.4519	1.14	7.36E-08
rs12521157	T	C	0.39	-3.26	0.001	-3.43	6E-04	0.8524	-1.95	0.051	-2.39	0.0168	0.66	1.25E-07
rs6894838	T	C	0.41	-3.48	5E-04	-2.24	0.025	0.9031	-2.41	0.016	-2.77	0.0057	0.65	9.06E-07
rs10065041	T	C	0.41	-3.00	0.003	-2.81	0.005	0.8798	-2.76	0.006	-1.31	0.1895	0.82	1.01E-06
rs4701260	A	G	0.41	1.63	0.103	2.52	0.012	1.12	3.76	2E-04	2.36	0.0184	1.42	6.79E-06
rs7447989	A	G	0.65	-1.46	0.144	-2.79	0.005	0.8803	-4.15	3E-05	-0.46	0.6458	0.93	1.03E-05
rs1366465	T	C	0.63	-1.90	0.057	-2.87	0.004	0.8749	-2.40	0.016	-2.92	0.0034	0.61	1.03E-05
rs10072518	T	C	0.53	-1.54	0.123	-3.31	9E-04	0.8636	-2.85	0.004	-1.14	0.2558	0.84	1.29E-05
rs12514304	T	G	0.19	1.78	0.076	3.01	0.003	1.185	2.93	0.003	-0.32	0.7459	0.94	4.14E-05
rs423116	T	C	0.41	-2.39	0.017	-2.14	0.032	0.9076	-2.66	0.008	1.16	0.2473	1.19	5.30E-05

Supplementary Table 4. Association results for genotyped SNPs within/nearby cadherins and protocadherins (other than *CDH9*/*CDH10*) among the top 1000 most significant SNPs in the combined analysis of the discovery cohorts. A1 and A2 refer to allele 1 and allele 2, respectively, and the allele frequencies below are calculated based on allele 1 in AGRE parents or in ACC control subjects.

SNP	Chr	Position	Closest gene	SNP-gene distance	A1	A2	A1_Freq (AGRE parents)	P (AGRE)	A1_Freq (ACC control)	P (ACC)	Odds Ratio (ACC)	P (combined)
rs3775330	4	30337382	*PCDH7*	0	A	G	0.8639	0.084181	0.8731	0.001745	0.82	0.000765
rs2879041	4	33041141	*PCDH7*	2283622	T	G	0.91786	0.818546	0.90831	1.18E-05	1.48	0.001058
rs17547161	4	133152254	*PCDH10*	1137666	A	G	0.90241	0.056483	0.897	0.001865	0.81	0.000521
rs3857321	5	21926009	*CDH12*	0	A	G	0.8037	0.000854	0.7827	0.07409	1.10	0.000537
rs6452027	5	21937473	*CDH12*	0	T	C	0.8033	0.00075	0.7798	0.07461	1.10	0.000496
rs13162273	5	21953276	*CDH12*	0	A	C	0.7956	0.002102	0.7808	5.06E-02	1.11	0.000645
rs2026410	10	56015517	*PCDH15*	0	T	C	0.1772	0.252904	0.1699	0.000258	1.23	0.000793
rs11647166	16	60923063	*CDH8*	295526	A	G	0.05486	0.235271	0.06127	0.000449	0.69	0.001033
rs318203	16	62807919	*CDH11*	730265	A	G	0.8875	0.275189	0.9021	0.000135	0.77	0.000587
rs11862535	16	82218967	*CDH13*	0	A	G	0.4485	0.013527	0.4397	0.01603	1.11	0.000845
rs11564334	18	23735780	*CDH2*	49153	A	G	0.6663	0.012173	0.6715	0.01791	0.90	0.000858
rs8098920	18	23755999	*CDH2*	28934	A	G	0.4773	0.010397	0.4623	0.02113	1.11	0.000883
rs11083238	18	23777488	*CDH2*	7445	T	C	0.4965	0.015573	0.5134	0.009464	0.89	0.000587
rs11564410	18	23888092	*CDH2*	0	A	G	0.2794	0.027227	0.2572	0.009571	1.14	9.79E-04
rs9965582	18	23951510	*CDH2*	0	A	G	0.2533	0.001612	0.2795	0.08672	0.92	0.000999
rs7505845	18	62637464	*CDH19*	215268	A	G	0.2194	0.063983	0.2254	0.000639	0.83	0.000262
rs6131030	20	44241393	*CDH22*	0	A	G	0.4199	0.001213	0.4106	0.000815	0.86	6.46E-06
rs1321001	20	44250143	*CDH22*	0	T	G	0.8437	0.011272	0.8483	0.01366	0.86	0.000623

Supplementary Table 5. Top association results (P<0.01) for genotyped SNPs within or surrounding prominent ASD loci previously implicated in linkage studies, cytogenetic studies and candidate gene association studies. This list of potential ASD loci was compiled from a recent review paper [34], including 8 "promising" genes and 18 "probable" genes.

(a) Significant SNPs within or surrounding ASD candidate loci on autosomes are summarized below:

SNP	Chr	Closest gene	SNP-gene distance	A1	A2	A1 Freq (AGRE)	P (AGRE)	A1 Freq (ACC)	P (ACC)	Odds Ratio (ACC)	P (combined)
rs10495983	2	*NRXN1*	67954	T	C	0.1195	0.00256896	0.1166	0.02217	0.85	0.000307
rs11889255	2	*NRXN1*	57977	T	G	0.1214	0.005084097	0.1177	0.01489	0.84	0.000351
rs10495985	2	*NRXN1*	54727	T	C	0.8826	0.002420425	0.8865	0.01763	1.19	0.000231
rs11891766	2	*NRXN1*	21019	A	G	0.1198	0.00887239	0.1164	0.0137	0.83	0.000511
rs7604754	2	*NRXN1*	0	T	C	0.1068	0.071860611	0.1067	0.004754	0.80	0.001402
rs17494917	2	*NRXN1*	0	A	G	0.8646	0.048899829	0.8596	0.02106	1.17	0.003311
rs2078232	2	*NRXN1*	0	A	C	0.1127	0.15840894	0.1151	3.38E-05	0.72	0.000104
rs970896	2	*NRXN1*	0	A	C	0.2396	0.295821428	0.2527	0.004915	0.86	0.007117
rs10490237	2	*NRXN1*	0	T	G	0.8647	0.324748158	0.8617	0.000414	1.28	0.001556
rs4467312	2	*NRXN1*	0	T	C	0.7206	0.652554154	0.7424	0.000738	0.85	0.006898
rs10183349	2	*NRXN1*	0	T	C	0.68	0.350503564	0.655	0.000185	1.20	0.001043
rs858937	2	*NRXN1*	0	T	C	0.8834	0.008496761	0.8762	0.1482	1.11	0.005816
rs12616608	2	*NRXN1*	562234	A	G	0.832	0.007517517	0.8447	0.2179	0.93	0.008474
rs2953300	2	*NRXN1*	592699	T	C	0.186	0.007090926	0.1645	0.007533	1.17	0.00024
rs952893	2	*NRXN1*	616553	A	G	0.8207	0.001659334	0.8348	0.4157	0.95	0.008184
rs6758434	2	*NRXN1*	641162	A	C	0.7981	0.002289891	0.8089	0.04787	0.90	0.000645
rs7569104	2	*NRXN1*	646423	T	C	0.1899	0.002935112	0.1811	0.02541	1.13	0.000392
rs4146703	2	*NRXN1*	650079	A	G	0.7934	0.005581081	0.8029	0.07404	0.91	0.001986
rs6714367	2	*NRXN1*	654680	T	C	0.8305	0.001250982	0.8446	0.07318	0.90	0.000686
rs1028145	2	*NRXN1*	668368	T	G	0.8584	8.24332E-05	0.8632	0.8058	0.98	0.005797
rs4971757	2	*NRXN1*	675074	A	G	0.8664	0.000594437	0.8677	0.4011	0.95	0.004341
rs4353689	2	*NRXN1*	675652	A	C	0.1269	0.001190066	0.1253	0.2783	1.07	0.003733
rs2354387	2	*NRXN1*	681044	T	C	0.8232	0.003188074	0.822	0.1675	0.92	0.003508
rs1516194	2	*NRXN1*	684059	T	G	0.1665	0.010675669	0.1681	0.1203	1.09	0.005382
rs11125373	2	*NRXN1*	686489	A	G	0.2044	0.00480634	0.2079	0.1145	1.09	0.002923

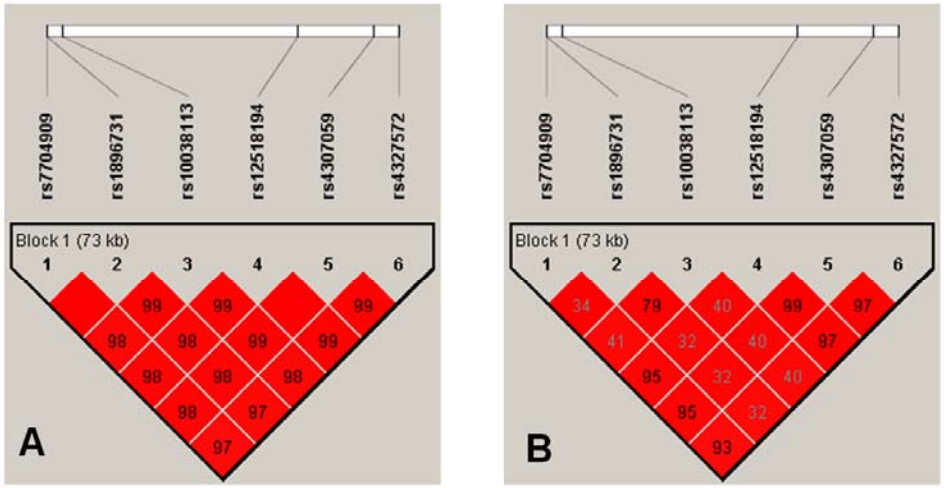
rs10202118	2	*NRXN1*	690650	T	C	0.7537	0.054147244	0.763	0.02912	0.89	0.004811
rs6712068	2	*NRXN1*	824642	A	G	0.7824	0.169636846	0.7752	0.006414	1.16	0.004445
rs4971785	2	*NRXN1*	1006007	T	C	0.4665	0.015101492	0.4403	0.05213	1.09	0.002889
rs75775	3	*OXTR*	9432	T	G	0.1317	0.167462647	0.1247	0.004034	1.20	0.003116
rs4839797	6	*GRIK2*	0	T	C	0.09385	0.026019454	0.1	0.04654	0.86	0.003978
rs2782908	6	*GRIK2*	0	A	G	0.653	0.217719853	0.6495	0.006077	1.14	0.005692
rs9390897	6	*GRIK2*	758634	A	G	0.6799	0.758075774	0.6897	7.32E-05	0.83	0.008647
rs1367645	6	*GRIK2*	774787	A	G	0.94168	0.137809694	0.93105	0.01327	1.27	0.006152
rs2205681	6	*GRIK2*	1041317	A	G	0.91407	0.005932629	0.91248	0.1318	1.13	0.003976
rs522447	6	*GRIK2*	1094701	A	G	0.90591	0.005951226	0.90131	0.03858	1.18	0.001039
rs513091	6	*GRIK2*	1111819	A	C	0.1209	0.000834405	0.1227	0.03655	0.86	0.00023
rs9404359	6	*GRIK2*	1270169	T	C	0.07665	0.008650422	0.07922	0.1336	0.88	0.005231
rs2399931	6	*GRIK2*	1276273	A	G	0.92518	0.004135609	0.92239	0.1228	1.14	0.002864
rs1155126	6	*GRIK2*	1283682	T	G	0.8461	0.011702895	0.8359	0.08919	1.11	0.004154
rs10264684	7	*CNTNAP2*	0	T	C	0.1193	0.070182264	0.1145	0.04229	1.15	0.008385
rs17170932	7	*CNTNAP2*	0	T	C	0.7957	0.063198268	0.7759	0.05066	1.11	0.008989
rs11971331	7	*EN2*	63597	A	G	0.7906	0.032859445	0.8115	0.07177	0.90	0.007274
rs2785079	10	*PTEN*	53854	A	G	0.1792	0.037645042	0.1944	0.01999	0.87	0.002503
rs1855970	10	*PTEN*	94189	T	G	0.8555	0.003603458	0.851	0.02425	1.16	0.000437
rs2108636	12	*CACNA1C*	2823	T	G	0.2367	0.003731608	0.2493	0.1541	0.93	0.00351
rs7972947	12	*CACNA1C*	0	A	C	0.2038	0.019070685	0.2189	0.1204	0.92	0.008154
rs4765898	12	*CACNA1C*	0	A	G	0.6734	0.039253592	0.6479	0.002878	1.15	0.000505
rs2238034	12	*CACNA1C*	0	T	C	0.7613	0.056032717	0.7385	0.01058	1.14	0.002094
rs2370419	12	*CACNA1C*	0	A	G	0.07443	0.133075953	0.06006	0.002356	1.30	0.00161
rs4076021	15	*GABRB3*	229543	T	C	0.90824	0.297545791	0.8989	0.007551	1.24	0.009643
rs751994	15	*GABRB3*	0	T	C	0.2965	0.023734212	0.2907	0.1227	1.08	0.009754
rs1863455	15	*GABRB3*	0	T	C	0.8874	0.009559274	0.8865	0.1976	0.92	0.008815
rs11652097	17	*ITGB3*	14491	T	C	0.3902	0.127124429	0.3989	0.007181	0.88	0.003525
rs2056131	17	*ITGB3*	0	T	C	0.3152	0.229719758	0.3022	0.008369	1.13	0.007625
rs10514919	17	*ITGB3*	0	T	G	0.2525	0.024986625	0.2477	0.08547	0.91	0.006991
rs999323	17	*ITGB3*	0	A	G	0.6944	0.085578978	0.6883	0.03919	1.11	0.009361

(b) Significant SNPs within or surrounding ASD candidate loci on the X chromosome are summarized below:

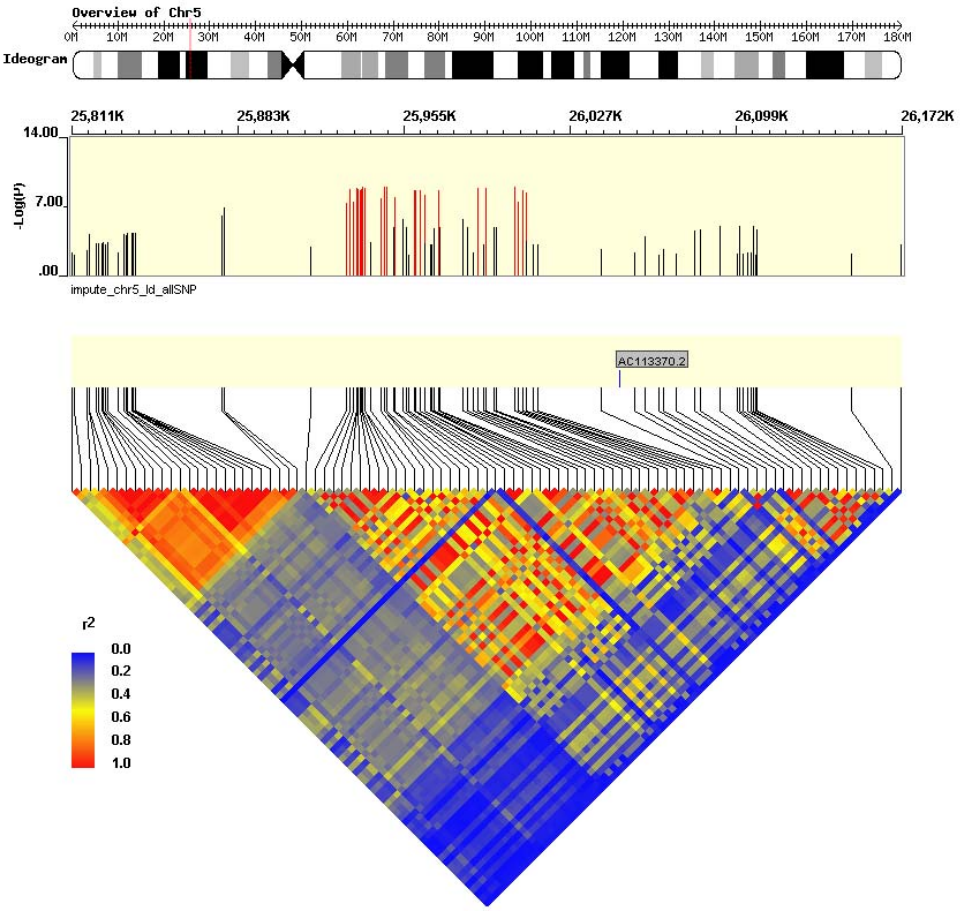
SNP	Closest gene	SNP-gene distance	A1	A2	A1_freq (AGRE)	P (AGRE)	A1_freq (ACC)	P (ACC)	Odds Ratio (ACC)	P (combined)
rs11798405	*NLGN4X*	877282	A	G	0.907172	0.006729	0.90051	1.10E-05	1.66	8.96E-07

rs878252	*NLGN4X*	221323	T	C	0.480405	0.030916	0.4949	0.004555	0.85	0.000584
rs11094994	*NLGN4X*	0	T	C	0.262923	0.021081	0.2533	0.002878	0.81	0.000274
rs4826722	*NLGN4X*	100865	A	G	0.234249	0.012679	0.2364	0.1762	0.91	0.009322
rs4826723	*NLGN4X*	115449	T	C	0.235548	0.01695	0.2372	0.1188	0.90	0.007383
rs5951989	*FMR1*	411643	T	C	0.780632	0.15267	0.7903	0.004058	1.24	0.002813

Supplementary Figure 1. The linkage disequilibrium between the six SNPs in Table 2 of the manuscript. Both D' measure (A) and r^2 measure (B) are shown. The figure is generated by Haploview [26].

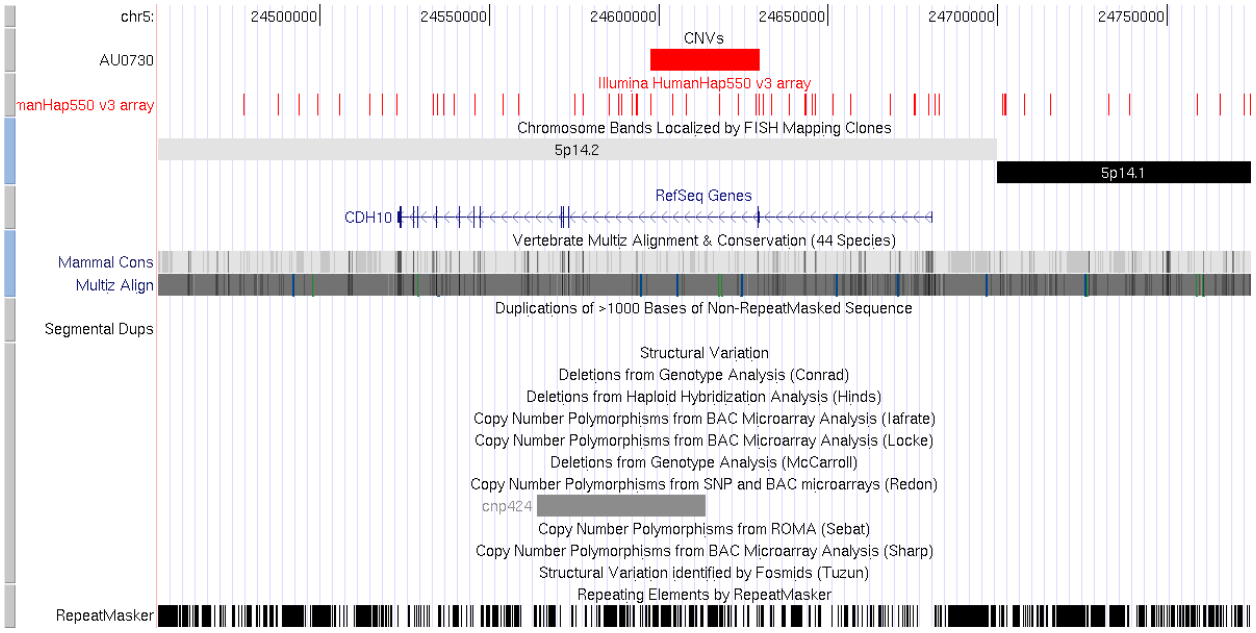


Supplementary Figure 2. The linkage disequilibrium plot for all the genotyped/imputed SNPs on 5p14.1 region, with their combined P-values (as –log10 values) on four cohorts (markers with $P<1x10^{-7}$ are highlighted in red color). All the most significant SNPs in this region fall within the same LD block. The figure is generated by WGAViewer [33].

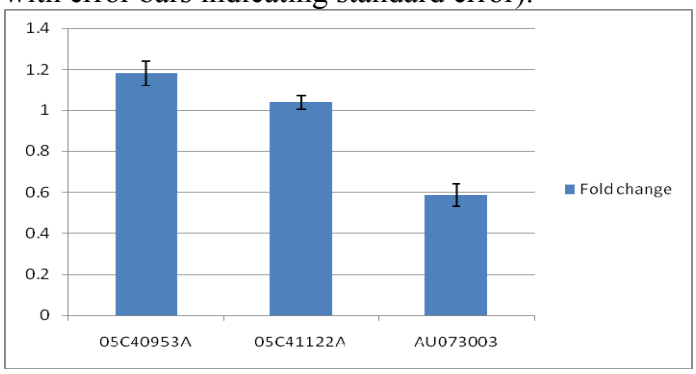


Supplementary Figure 3. CNVs between *CDH10* and *CDH9* in our study, as visualized in the UCSC Genome Browser. We attempted experimental validation of intergenic CNVs by quantitative PCR (QPCR) and by multiplex ligation-dependent probe amplification (MLPA) assay. Each panel below illustrates one CNV loci overlapping or between *CDH10* and *CDH9*, and the red bar in each panel represents the location and coordinate of the CNVs. The marker coverage in the SNP genotyping array is illustrated by red vertical lines in the track.

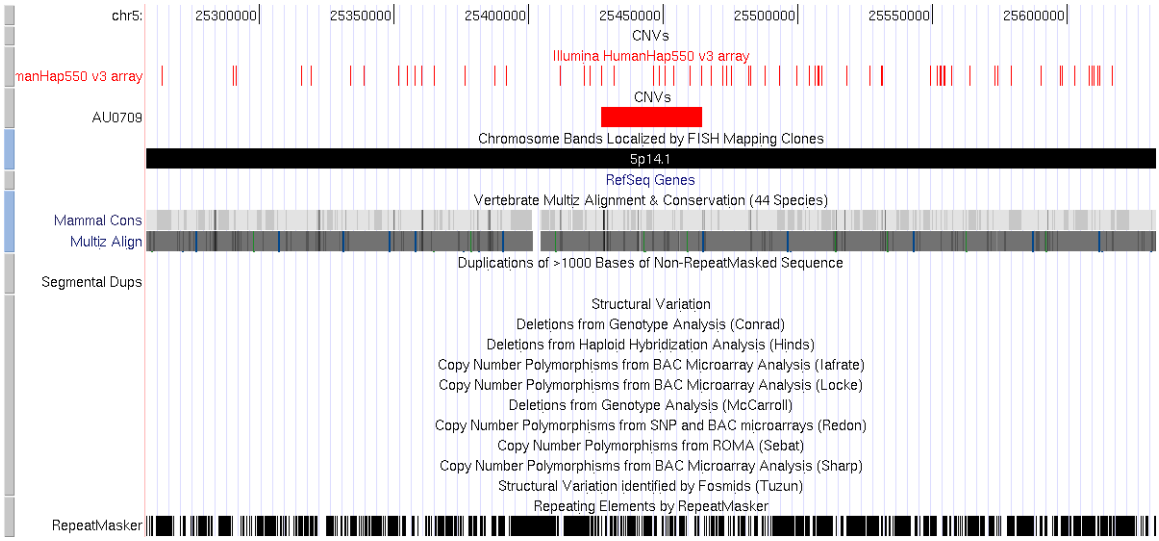
Panel (a):



CNV Range: chr5:24597380-24629650. The CNV has been previously described in HapMap subjects (Redon et al). We attempted experimental validation using the MLPA assay (a representative series is shown below with two control subjects and one proband, with error bars indicating standard error).

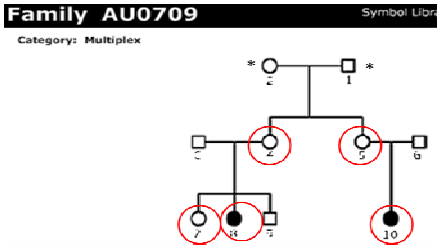


Panel (b):

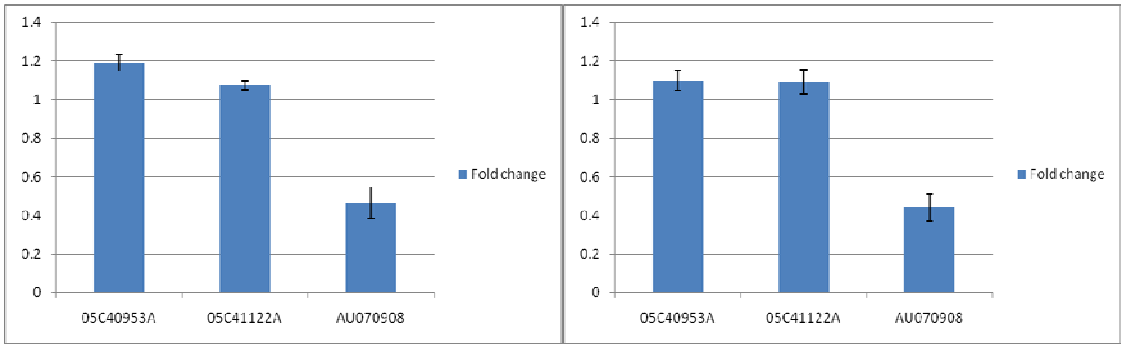


Range: chr5:25426837-25464318. We validated the CNV by both QPCR and MLPA.

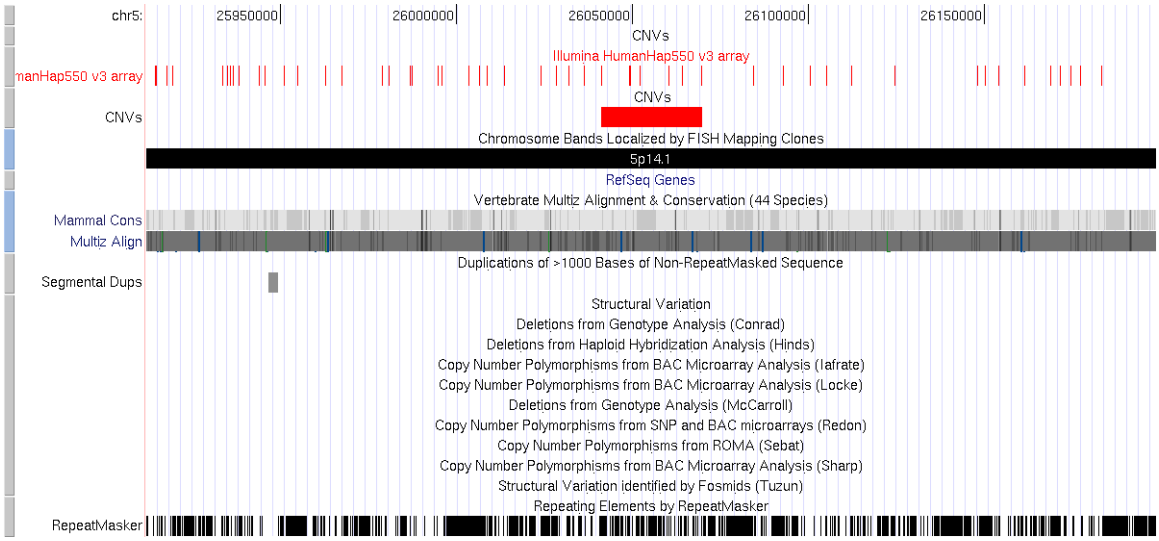
QPCR results: The CNV is detected in only one family (Red circle: subject carrying the deletions; Star: DNA not available).



MLPA results (two probes were used and both validate the CNV)

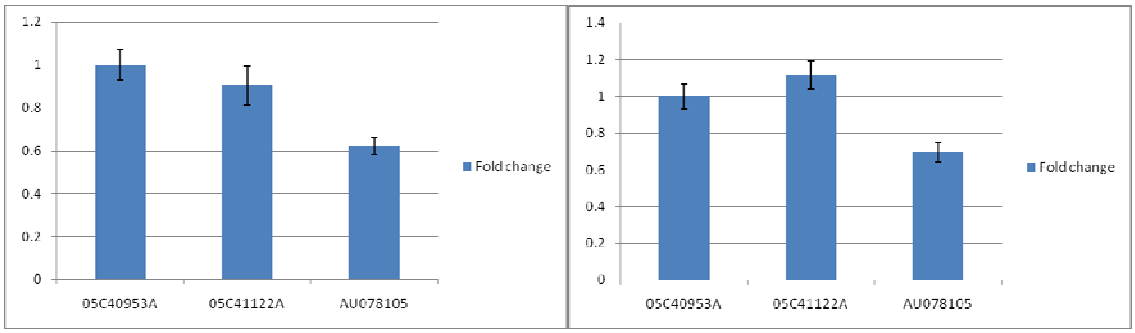


Panel (c):

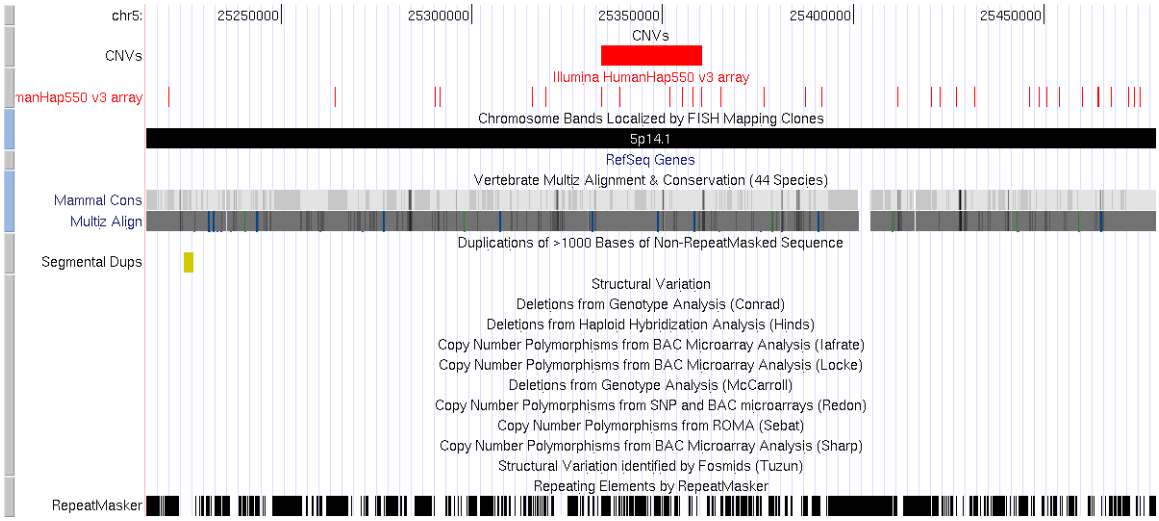


Region: chr5:26041000-26069708

This CNV is validate by MLPA (two probes were used and both validate the CNV).



Panel (d):



Region: chr5:25333737-25360219

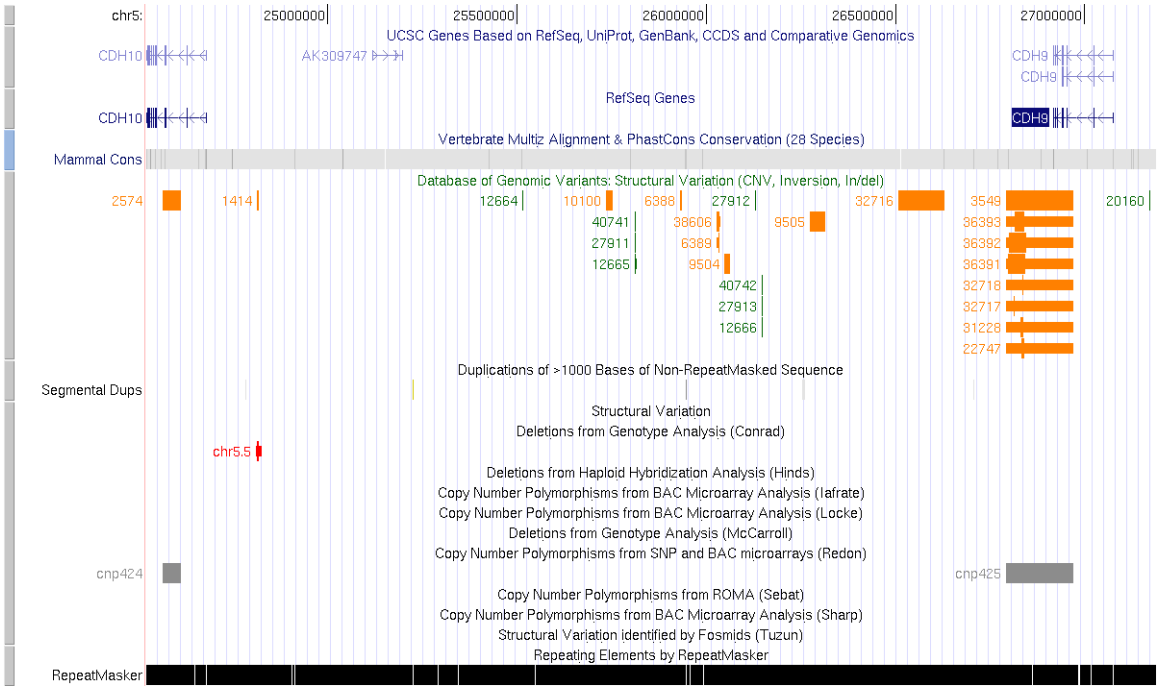
We attempted experimental validation by TaqMan QPCR on multiple subjects including controls, but all CN estimates have very large variations from 1.8 to over 30 (the CN for four subjects in this family were between 2.3 to 2.7), indicating a potential assay error. Nevertheless, since it is detected in related subjects, we list the CNV information here, but caution that it failed validation.

Panel (e):

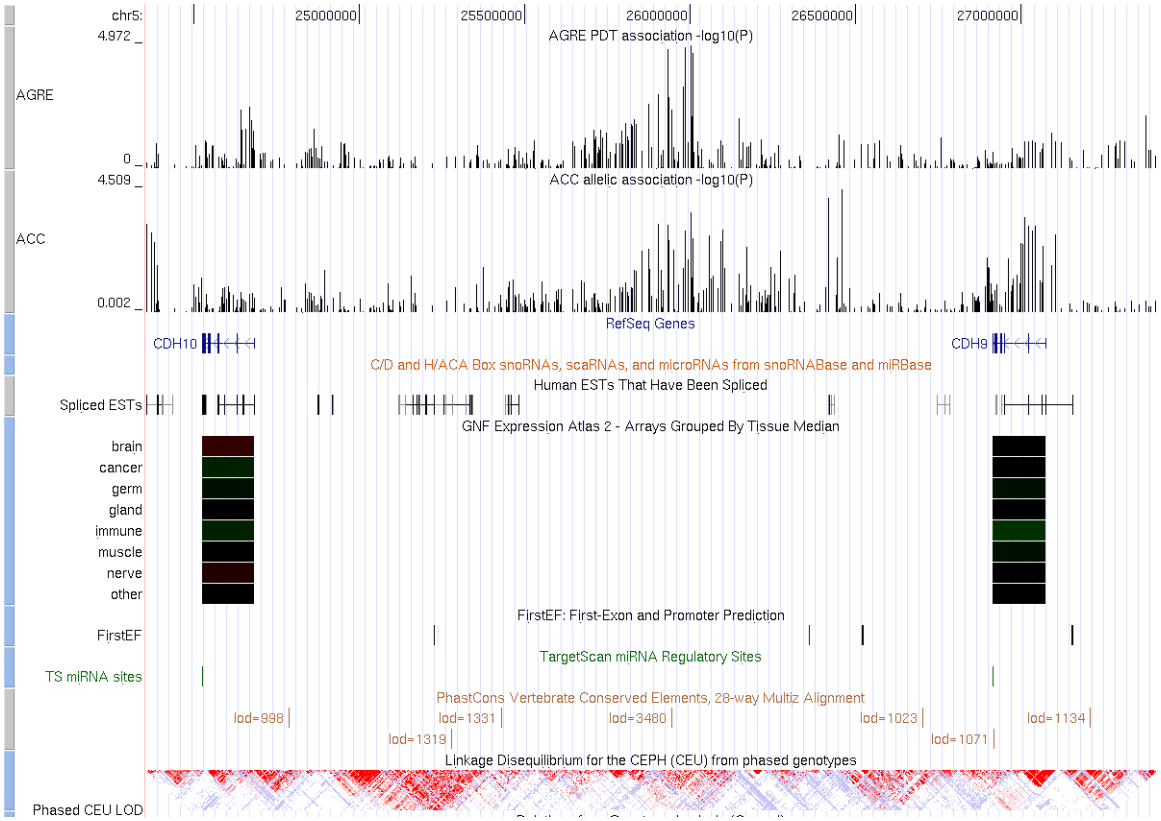
Region: chr5:26955104-26963260

The CNV has been previously described in HapMap subjects (Redon et al).

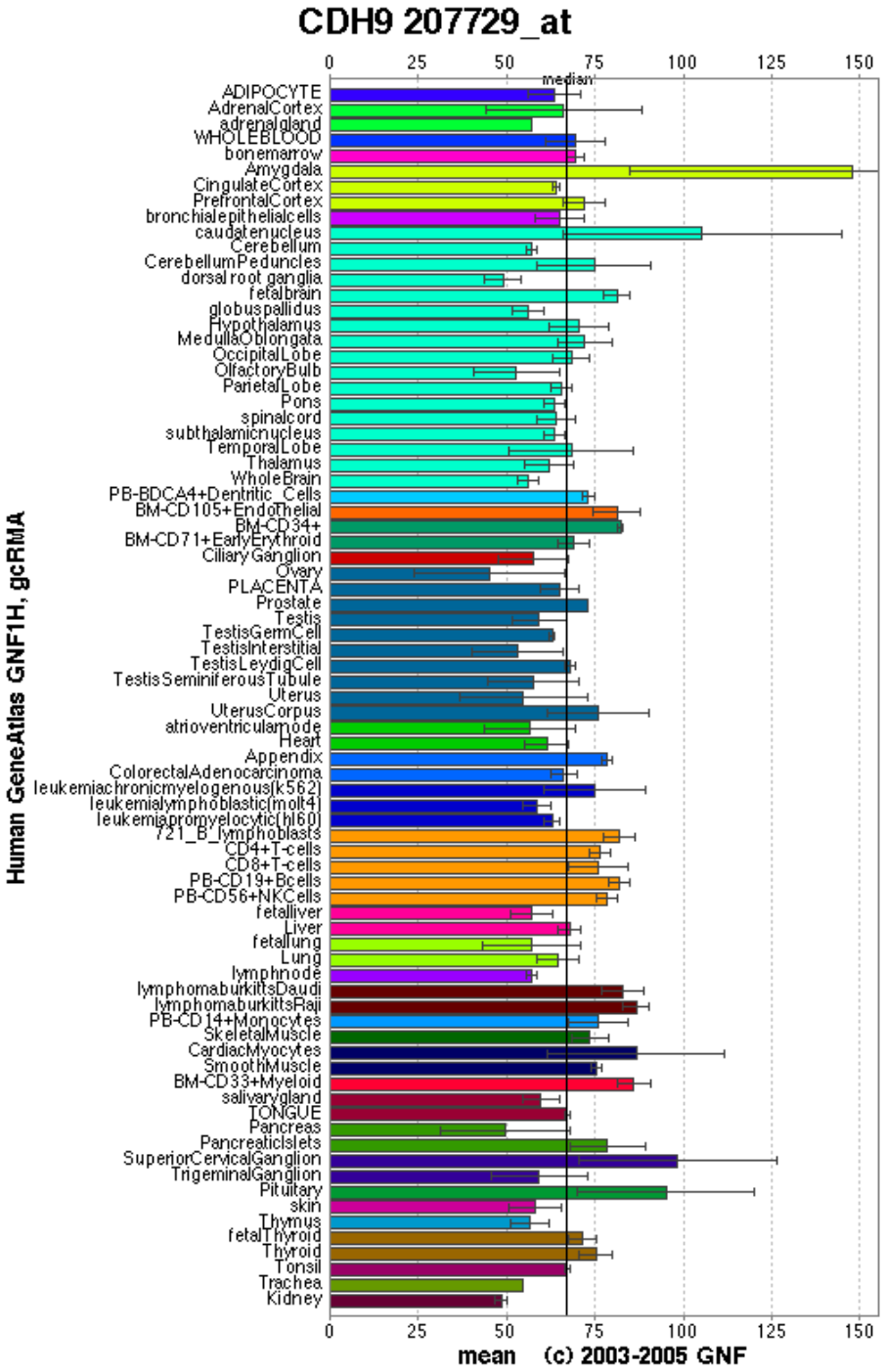
Supplementary Figure 4. Previously reported CNVs between *CDH10* and *CDH9*, as annotated in the UCSC Genome Browser annotation databases. Two tracks were displayed in the browser, including the Database of Genomic Variants (http://projects.tcag.ca/variation/) track, as well as "Structural Variation" track compiled from nine previous publications. Both tracks indicate that no common CNVs were identified between *CDH10* and *CDH9*, although a CNV that disrupts *CDH9* 3' region has been detected in multiple subjects. Therefore, unless a very small CNV exists that evades detection by current technical platforms, the top SNP association result is unlikely to be due to the linkage disequilibrium with a CNV.



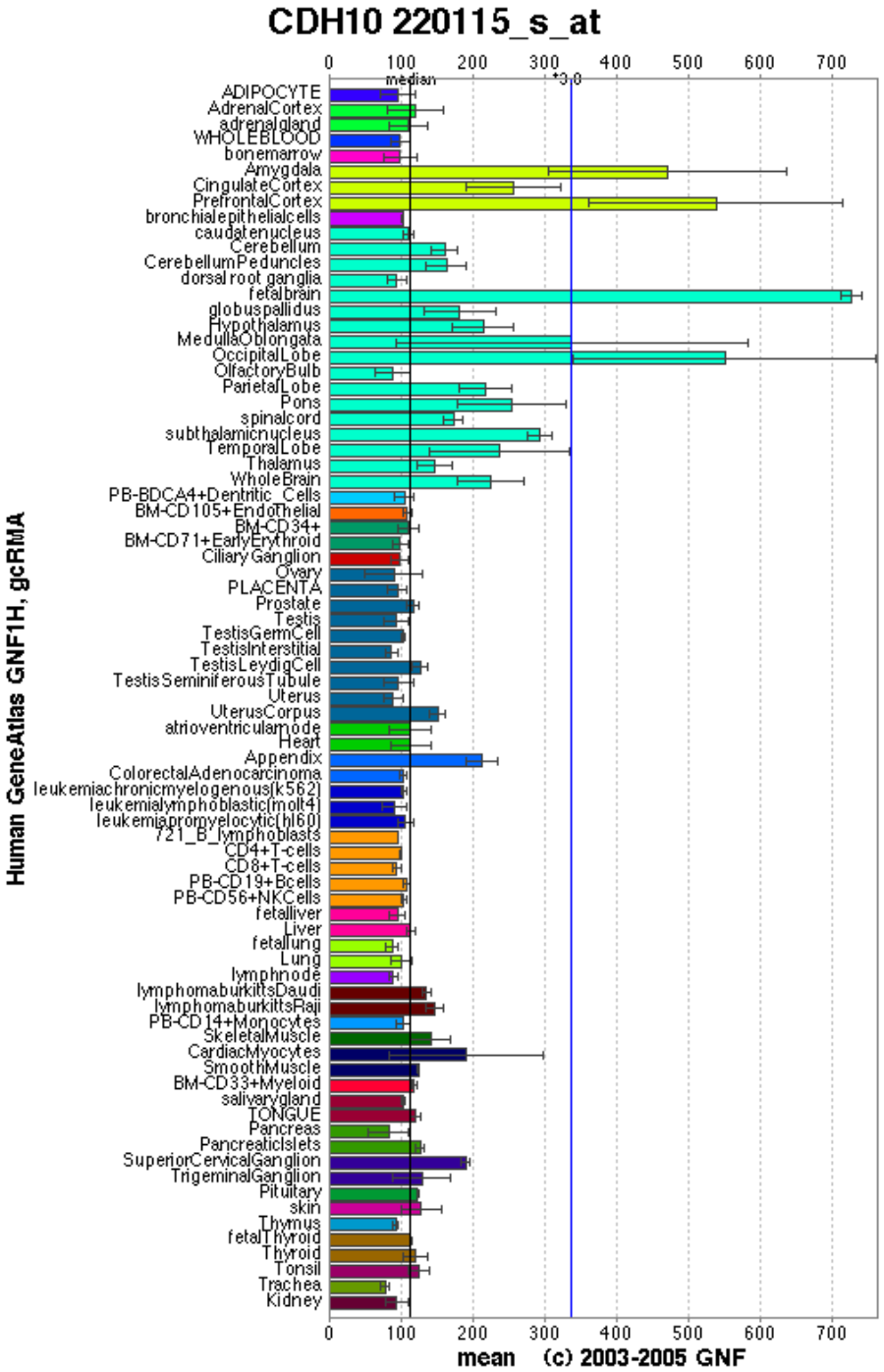
Supplementary Figure 5. Genome browser shot of the genomic region between *CDH10* and *CDH9*. The SNP association results for the AGRE cohort and ACC cohort are displayed as vertical lines that represent $-\log_{10}(P)$ values. There are no known microRNAs or small nuclear RNA in this region, as shown in the Genome Browser track. There are no spliced human Expressed Sequence Tags (ESTs) that overlap with the LD block, as shown in the "Human EST" track [35]. The expression values (color changes from red to black to green with decreasing expression) for different groups of human tissues are displayed in the "GNF Expression Atlas 2" track [25]. The predicted transcription start sites are displayed in the "FirstEF" track [36], and no such sites overlap with the LD block. The predicted microRNA targets were displayed in the "TargetScan" track [37], and none of them overlap with the LD block. The conserved genomic elements are displayed in the PhastCons track [28] with LOD scores.



Supplementary Figure 6. The tissue-specific gene expression levels for *CDH9* (probe identifier: 207729_at), based on the GNF SymAtlas database on 79 human tissues. The black line represents median value.



Supplementary Figure 7. The tissue-specific gene expression levels for *CDH10* (probe identifier: 220115_s_at), based on the GNF SymAtlas database on 79 human tissues. The black line and blue line represent median value and its 3 fold value, respectively.



Supplementary References

1 Geschwind, D.H. *et al.*, The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet* 69 (2), 463-466 (2001).

2 Lord, C., Rutter, M., & Le Couteur, A., Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 24 (5), 659-685 (1994).

3 Lord, C. *et al.*, The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* 30 (3), 205-223 (2000).

4 Purcell, S. *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81 (3), 559-575 (2007).

5 Wang, K. *et al.*, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17, 1665-1674 (2007).

6 Devlin, B. & Roeder, K., Genomic control for association studies. *Biometrics* 55 (4), 997-1004 (1999).

7 Mitchell, A.A., Cutler, D.J., & Chakravarti, A., Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* 72 (3), 598-610 (2003).

8 Chung, R.H., Morris, R.W., Zhang, L., Li, Y.J., & Martin, E.R., X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. *Am J Hum Genet* 80 (1), 59-68 (2007).

9 Price, A.L. *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38 (8), 904-909 (2006).

10 de Bakker, P.I. *et al.*, Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17 (R2), R122-128 (2008).

11 Lord, C. *et al.*, Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* 19 (2), 185-212 (1989).

12 Martin, E.R., Monks, S.A., Warren, L.L., & Kaplan, N.L., A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 67 (1), 146-154 (2000).

13 Martin, E.R., Bass, M.P., & Kaplan, N.L., Correcting for a potential bias in the pedigree disequilibrium test. *Am J Hum Genet* 68 (4), 1065-1067 (2001).

14 Horvath, S., Xu, X., & Laird, N.M., The family based association test method: strategies for studying general genotype--phenotype associations. *Eur J Hum Genet* 9 (4), 301-306 (2001).

15 Abrahams, B.S. *et al.*, Genome-wide analyses of human perisylvian cerebral cortical patterning. *Proc Natl Acad Sci U S A* 104 (45), 17849-17854 (2007).

16 Easterday, M.C. *et al.*, Neural progenitor genes. Germinal zone expression and analysis of genetic overlap in stem cell populations. *Dev Biol* 264 (2), 309-322 (2003).

17 Simes, R.J., An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751-754 (1986).

18 Sarkar, S. & Chang, C.K., Simes' method for multiple hypothesis testing with positively dependent test statistics. *J Am Stat Assoc* 92, 1601-1608 (1997).

19 Chen, B.E., Sakoda, L.C., Hsing, A.W., & Rosenberg, P.S., Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genet Epidemiol* 30 (6), 495-507 (2006).

20 Wang, K., Li, M., & Bucan, M., Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet* 81 (6) (2007).

21 Hollander, M. & Wolfe, D.A., *Nonparametric Statistical Methods*. (Wiley, New York, 1999).

22 Iafrate, A.J. *et al.*, Detection of large-scale variation in the human genome. *Nat Genet* 36 (9), 949-951 (2004).

23 Laird, N.M., Horvath, S., & Xu, X., Implementing a unified approach to family-based tests of association. *Genet Epidemiol* 19 Suppl 1, S36-42 (2000).

24 Su, A.I. *et al.*, Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 99 (7), 4465-4470 (2002).

25 Su, A.I. *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101 (16), 6062-6067 (2004).

26 Barrett, J.C., Fry, B., Maller, J., & Daly, M.J., Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263-265 (2005).

27 Frazer, K.A. *et al.*, A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449 (7164), 851-861 (2007).

28 Siepel, A. *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15 (8), 1034-1050 (2005).

29 Saxena, R. *et al.*, Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316 (5829), 1331-1336 (2007).

30 Heinzen, E.L. *et al.*, Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits. *PLoS Biol* 6 (12), e1000001 (2008).

31 Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P., A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39 (7), 906-913 (2007).

32 Kuhn, R.M. *et al.*, The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 37 (Database issue), D755-761 (2009).

33 Ge, D. *et al.*, WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res* 18 (4), 640-643 (2008).
34 Abrahams, B.S. & Geschwind, D.H., Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9 (5), 341-355 (2008).
35 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., & Wheeler, D.L., GenBank. *Nucleic Acids Res* 36 (Database issue), D25-30 (2008).
36 Davuluri, R.V., Grosse, I., & Zhang, M.Q., Computational identification of promoters and first exons in the human genome. *Nat Genet* 29 (4), 412-417 (2001).
37 Lewis, B.P., Burge, C.B., & Bartel, D.P., Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120 (1), 15-20 (2005).