A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder

Jillian P. Casey[1], Tiago Magalhaes[2,3,4], Judith M. Conroy[1], Regina Regan[1], Naisha Shah[1], Richard Anney[5], Denis C. Shields[1], Brett S. Abrahams[6], Joana Almeida[7], Elena Bacchelli[8], Anthony J. Bailey[9], Gillian Baird[10], Agatino Battaglia[11], Tom Berney[12], Nadia Bolshakova[5], Patrick F. Bolton[13], Thomas Bourgeron[14], Sean Brennan[5], Phil Cali[15], Catarina Correia[2,3,4], Christina Corsello[16], Marc Coutanche[63], Geraldine Dawson[17,18], Maretha de Jonge[19], Richard Delorme[20], Eftichia Duketis[21], Frederico Duque[7], Annette Estes[22], Penny Farrar[23], Bridget A. Fernandez[24], Susan E. Folstein[25], Suzanne Foley[63], Eric Fombonne[26], Christine M. Freitag[21], John Gilbert[27], Christopher Gillberg[28], Joseph T. Glessner[29], Jonathan Green[30], Stephen J. Guter[15], Hakon Hakonarson[29, 31], Richard Holt[23], Gillian Hughes[5], Vanessa Hus[16], Roberta Igliozzi[11], Cecilia Kim[29], Sabine M. Klauck[32], Alexander Kolevzon[33], Janine A. Lamb[34], Marion Leboyer[35], Ann Le Couteur[12], Bennett L. Leventhal[36,37], Catherine Lord[16], Sabata C. Lund[38], Elena Maestrini[8], Carine Mantoulan[39], Christian R. Marshall[41], Helen McConachie[12], Christopher J. McDougle[42], Jane McGrath[5], William M. McMahon[43], Alison Merikangas[5], Judith Miller[43], Fiorella Minopoli[8], Ghazala K. Mirza[23], Jeff Munson[44], Stanley F. Nelson[45], Gudrun Nygren[28], Guiomar Oliveira[7], Alistair T. Pagnamenta[23], Katerina Papanikolaou[46], Jeremy R. Parr[12], Barbara Parrini[11], Andrew Pickles[47], Dalila Pinto[41], Joseph Piven[48], David J. Posey[42], Annemarie Poustka[32]‡, Fritz Poustka[21], Jiannis Ragoussis[23], Bernadette Roge[39], Michael L. Rutter[49], Ana F. Sequeira[2,3,4], Latha Soorya[33], Inês Sousa[23], Nuala Sykes[23], Vera Stoppioni[50], Raffaella Tancredi[11], Maïté Tauber[39], Ann P. Thompson[40], Susanne Thomson[38], John Tsiantis[46], Herman Van Engeland[19], John B. Vincent[51], Fred Volkmar[52], Jacob A.S. Vorstman[19], Simon Wallace[63], Kai Wang[29], Thomas H. Wassink[53], Kathy White[63], Kirsty Wing[23], Kerstin Wittemeyer[54], Brian L. Yaspan[38], Lonnie Zwaigenbaum[55], Catalina Betancur[56*], Joseph D. Buxbaum[33,57*], Rita M. Cantor[45*], Edwin H. Cook[15*], Hilary Coon[43*], Michael L. Cuccaro[27*], Daniel H. Geschwind[6*], Jonathan L. Haines[38*], Joachim Hallmayer[58*], Anthony P. Monaco[23*], John I. Nurnberger Jr[42*], Margaret A. Pericak-Vance[27*], Gerard D. Schellenberg[59*], Stephen W. Scherer[41, 60*], James S. Sutcliffe[38*], Peter Szatmari[40*], Veronica J. Vieland[61*], Ellen M. Wijsman[62*], Andrew Green[1], Michael Gill[5*], Louise Gallagher[5*], Astrid Vicente[2,3,4*], & Sean Ennis[1]*†

[1]School of Medicine and Medical Science University College, Dublin 4, Ireland. [2]Instituto Nacional de Saude Dr Ricardo Jorge, Av Padre Cruz 1649-016, Lisbon, Portugal. [3]BioFIG—Center for Biodiversity, Functional and Integrative Genomics, Campus da FCUL, C2.2.12, Campo Grande, 1749-016 Lisboa, Portugal. [4]Instituto Gulbenkian de Cîencia, Rua Quinta Grande, 2780-156 Oeiras, Portugal. [5]Autism Genetics Group, Department of Psychiatry, School of Medicine, Trinity College, Dublin 8, Ireland. [6]Program in Neurogenetics, Department of Neurology and Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine at UCLA. [7]Hospital Pediátrico de Coimbra, 3000 – 076 Coimbra, Portugal. [8]Department of Biology, University of Bologna, 40126 Bologna, Italy. [9]Department of Psychiatry, University of British Columbia, V6T 2A1, Canada. [10]Newcomen Centre, Guy's

Hospital, London SE1 9RT, UK. [11]Stella Maris Institute for Child and Adolescent Neuropsychiatry, 56128 Calambrone (Pisa), Italy. [12]Institute of Neuroscience, and Institute of Health and Society, Newcastle University, Newcastle Upon Tyne, NE1 7RU, UK. [13]Department of Child and Adolescent Psychiatry, Institute of Psychiatry, London SE5 8AF, UK. [14]Human Genetics and Cognitive Functions, Institut Pasteur; University Paris Diderot-Paris 7, CNRS URA 2182, Fondation FondaMental, 75015 Paris, France. [15]Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60612,, USA. [16]Autism and Communicative Disorders Centre, University of Michigan, Ann Arbor, Michigan 48109-2054, USA. [17]Autism Speaks, New York 10016, USA. [18]Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-3366, USA. [19]Department of Child and Adolescent Psychiatry, University Medical Center, Utrecht 3508 GA, The Netherlands. [20]INSERM U 955, Fondation FondaMental, APHP, Hôpital Robert Debré, Child and Adolescent Psychiatry, 75019 Paris, France. [21]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, J.W. Goethe University Frankfurt, 60528 Frankfurt, Germany. [22]Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington 98195, USA. [23]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK. [24]Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St John's Newfoundland A1B 3V6, Canada. [25]Department of Psychiatry, University of Miami School of Medicine, Miami, FL 33136, USA. [26]Division of Psychiatry, McGill University, Montreal, Quebec H3A 1A1, Canada. [27]The John P. Hussman Institute for Human Genomics, University of Miami School of Medicine, Miami, Florida 33136, USA. [28]Gillberg Neuropsychiatry Centre, Sahlgrenska Academy, University of Gothenburg, S41345 Gothenburg, Sweden. [29]The Center for Applied Genomics, Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [30]Academic Department of Child Psychiatry, Booth Hall of Children's Hospital, Blackley, Manchester M9 7AA, UK. [31]Department of Pediatrics, Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. [32]Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [33]The Seaver Autism Center for Research and Treatment and Department of Psychiatry, Mount Sinai School of Medicine, New York 10029, USA. [34]Centre for Integrated Genomic Medical Research, University of Manchester, Manchester M13 9PT, UK. [35]INSERM U995, Department of Psychiatry, Groupe Hospitalier Henri Mondor-Albert Chenevier, AP-HP; University Paris 12, Fondation FondaMental, Créteil 94000, France. [36]Nathan Kline Institute for Psychiatric Research (NKI), 140 Old Orangeburg Road, Orangeburg, New York 10962, USA. [37]Department of Child and Adolescent Psychiatry, New York University and NYU Child Study Center, 550 First Avenue, New York, New York 10016, USA. [38]Department of Molecular Physiology and Biophysics, Vanderbilt Kennedy Center, and Centers for Human Genetics Research and Molecular Neuroscience, Vanderbilt University, Nashville, Tennessee 37232, USA. [39]Octogone/CERPP (Centre d'Eudes et de Recherches en Psychopathologie), University de Toulouse Le Mirail, Toulouse Cedex 31058, France. [40]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario L8N 3Z5, Canada. [41]The Centre for Applied

Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Ontario M5G 1L7, Canada. [42]Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. [43]Psychiatry Department, University of Utah Medical School, Salt Lake City, Utah 84108, USA. [44]Department of Psychiatry and Behavioural Sciences, University of Washington, Seattle, Washington 98195, USA. [45]Department of Human Genetics, University of California—Los Angeles School of Medicine, Los Angeles, California 90095, USA. [46]University Department of Child Psychiatry, Athens University, Medical School, Agia Sophia Children's Hospital, 115 27 Athens, Greece. [47]Department of Medicine, School of Epidemiology and Health Science, University of Manchester, Manchester M13 9PT, UK. [48]Carolina Institute for Developmental Disabilities, CB3366 , University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3366. [49]Social, Genetic and Developmental Psychiatry Centre, Institute Of Psychiatry, London SE5 8AF, UK. [50]Neuropsichiatria Infantile, Ospedale Santa Croce, 61032 Fano, Italy. [51]Centre for Addiction and Mental Health, Clarke Institute and Department of Psychiatry, University of Toronto, Toronto, Ontario M5G 1X8, Canada. [52]Child Study Centre, Yale University, New Haven, Connecticut 06520, USA. [53]Department of Psychiatry, Carver College of Medicine, Iowa City, Iowa 52242, USA. [54]Autism Centre for Education and Research, School of Education, University of Birmingham, B15 2TT. [55]Department of Pediatrics, University of Alberta, Edmonton, Alberta T6G 2J3, Canada. [56]INSERM U952 and CNRS UMR 7224 and UPMC Univ Paris 06, Paris 75005, France. [57]Departments of Genetics and Genomic Sciences and Neuroscience, Mount Sinai School of Medicine, New York 10029, USA. [58]Department of Psychiatry, Division of Child and Adolescent Psychiatry and Child Development, Stanford University School of Medicine, Stanford, California 94304, USA. [59]Pathology and Laboratory Medicine, University of Pennsylvania, Pennsylvania 19104, USA. [60]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A1, Canada. [61]Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital and The Ohio State University, Columbus, Ohio 43205, USA. [62]Departments of Biostatistics and Medicine, University of Washington, Seattle, Washington 98195, USA. [63]Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford OX3 7JX, UK.

† To whom correspondence should be addressed. Dr. Sean Ennis, Health Sciences Centre, University College Dublin, Ireland. Email: Sean.Ennis@ucd.ie

Abstract

Autism spectrum disorder (ASD) is a highly heritable disorder of complex and heterogeneous etiology. It is primarily characterized by altered cognitive ability including impaired language and communication skills and fundamental deficits in social reciprocity. Despite some notable successes in neuropsychiatric genetics, overall, the high heritability of ASD (~90%) remains poorly explained by common genetic risk variants. Instead, early studies of rare variation, in particular copy number variation, have suggested that it may account for a significant proportion of the genetic basis of ASD. We present a large scale analysis to identify candidate genes which may contain low-frequency recessive variation contributing to ASD while taking into account the potential contribution of population differences to the genetic heterogeneity of ASD. Our strategy, homozygous haplotype (HH) mapping, aims to detect homozygous segments of identical haplotype structure that are shared at a higher frequency amongst ASD patients compared to parental controls. The analysis was performed on 1,402 Autism Genome Project (AGP) trios genotyped for 1 million single nucleotide polymorphisms (SNPs). We identified 25 known and 1,218 novel candidate genes in the discovery analysis including *ABHD14A*, *CADM2*, *CHRFAM7A*, *GRIK2*, *GRM3*, *EPHA3*, *FGF10*, *PDZK1*, *KCND2*, *IMMP2L* and *FOXP2*. Furthermore, 10 of the previously reported ASD genes and 300 of the novel candidates identified in the discovery analysis were replicated in an independent sample of 1,182 trios. Our results demonstrate that regions of HH are significantly enriched for previously reported ASD candidate genes and the observed association is independent of gene size (Odds Ratio 2.10). Our findings highlight the applicability of HH mapping in complex disorders such as ASD and offer an alternative approach to the analysis of genome-wide association data.

Introduction

Extended runs of homozygosity (ROH) have recently been highlighted as a genomic feature that may be useful to map recessive disease genes in outbred populations (Hildebrandt et al. 2009; Lencz et al. 2007; Yang et al. 2010). Furthermore, even in complex disorders, we expect to find an unusually high number of affected individuals sharing a haplotype in the region surrounding a disease mutation (Durand et al. 2007; Lesch et al. 2008; Wong et al. 2002). Therefore, a rare pathogenic variant and surrounding haplotype is often enriched in frequency in a group of affected individuals compared with the haplotype frequency in a cohort of unaffected controls (The International HapMap Consortium 2003). We propose that homozygous haplotypes (HH) that are shared by multiple affected individuals may be important for the discovery of recessive disease genes in complex disorders. We have extended the traditional homozygosity mapping method by analysing the haplotype within shared ROH regions to identify homozygous segments of identical haplotype that are present uniquely or at a higher frequency in ASD probands compared to parental controls (Fig. 1). Such regions are termed risk homozygous haplotypes (rHHs). We postulate that rHHs may contain low frequency recessive variants that contribute to ASD risk in a subset of ASD patients.

Allelic and locus heterogeneity are major challenges in the identification of ASD risk loci (Lamb et al. 2000). In cases where distinct populations share the same risk allele, differences in allele frequency and LD structure between populations may result in the risk allele segregating on different haplotype backgrounds. Correction for population substructure in large scale studies minimises the rate of false-positives and dilution of a population-specific signal. Furthermore, Nothnagel and colleagues recently reported that the distribution of SNP-defined ROHs is highly structured across European populations and highlighted the importance of accounting for ancestry when undertaking ROH-based analyses (Nothnagel et

al. 2010). Our analysis accounts for population effects by separating the samples into groups of common ancestry and applying the HH mapping strategy to each population group independently. It also involves the use of parental controls to address variation in low frequency alleles across populations (Cardon and Palmer 2003). The genetic ancestry of the sample set was examined by principal component analysis (PCA), Hopach clustering (van der Laan 2002) and genetic distance F_{st} calculations (Supplementary Material 1, Supplementary Fig. 1 and 2, Supplementary Tables 2 and 3). Ten distinct population clusters were identified ranging in size from 27 to 289 probands (Fig. 2). Population clusters with a minimum of 50 probands were selected for the discovery analysis (n = 5). Each cluster was analysed independently by HH mapping and the genes identified in each population cluster were then compared. In this manner (taking a gene-centric approach) it is possible to identify genes that may confer risk across different populations regardless of whether the causal haplotype is population-specific or not.

Subjects and Methods

Cohort description

The samples used in the HH analysis were collected as part of an international consortium, the Autism Genome Project (AGP). Informed consent was obtained from all participants. The AGP sample set is a trio based collection, comprising an affected proband and two parents, grouped into the three distinct diagnostic classes of autism; strict, broad and spectrum. Affected individuals were diagnosed using the Autism Diagnostic Interview-Revised (ADI-R) and/or the Autism Diagnostic Observation Schedule (ADOS). A detailed description of the AGP sample set is provided in Supplementary Material 1. Raw genotype data for the ASD trios is deposited at NCBI dbGAP (accession phs000267.v1.p1). The HH analysis was performed on trios in the autism spectrum diagnostic category (n = 2,584 trios). The ASD

spectrum trios were further subdivided into stage 1 and stage 2 collections. In the current study, the 1,402 stage 1 trios were used for the initial discovery analysis and the 1,182 stage 2 trios were used for the independent replication study.

Genotyping and quality control

Stage 1 samples were genotyped using the Illumina 1M-single array while the stage 2 samples were genotyped on a combination of 1M and 1M-duo chips. The 1M SNP array contains 1,072,820 SNP markers at an average inter-marker distance of 2.7 kb while the 1M-duo chip contains almost 1,199,187 SNPs with a mean marker spacing of 1.5 kb. The 1,003,736 SNPs that were genotyped on both the 1M and 1M-duo platforms were considered for the analysis. Possible gender miscalls were assessed through analysis of chromosome X genotypes in PLINK (version 1.04) (http://pngu.mgh.harvard.edu/~purcell/plink/) (Purcell et al. 2007). Duplicates were identified by calculating identity by state (IBS) values using PLINK and one sample from each duplicate pair was removed. After quality control and filtering for autosomal markers, 887,716 SNPs and 7,719 individuals were retained for analysis (Supplementary Material 1).

Ancestry analysis

The population structure of the sample set was assessed through principal component analysis (PCA), Hopach hierarchical clustering and F_{st} calculations. PCA was performed with EIGENSTRAT from EIGENSOFT (version 3.0) (Price et al. 2006) using 70,175 independent autosomal SNPs with a minor allele frequency > 5% and a call rate of 100%. Tracy-Widom statistics indicated that the first 8 principal components (PCs) were significantly contributing to the population structure of the sample set. Individuals were assigned to clusters based on similarity in eigenvalues for the first 8 PCs using the Hopach

hybrid hierarchical clustering algorithm available in the R package (van der Laan 2002). Hierfstat was used to calculate pair-wise F_{st} metrics for the clusters using 5,000 SNPs randomly chosen from the panel of 70,175 SNPs used for PCA (Goudet 2005). Clusters displaying high similarity (F_{st} value $< 1 \times 10^{-4}$) were merged, resulting in 10 population clusters for the HH analysis (Supplementary Material 1 and Supplementary Table 3).

Homozygous haplotype analysis

Long series of consecutive homozygous SNPS, referred to as runs of homozygosity (ROHs), were identified in each sample using the 'homozyg' function in PLINK. A threshold of 100 consecutive homozygous SNPs spanning at least 1Mb at a minimum density of 1 SNP/50kb was implemented. Using the homozyg-group function in PLINK, samples (min 3) with overlapping ROHs were pooled and subdivided into haplotype groups (min 95% allelic identity). Each haplotype within the overlapping ROH region is referred to as a homozygous haplotype (HH). For each overlapping ROH, the frequency of each HH within cases and controls was evaluated using the Fisher's exact test (R script). HH that were significantly more common in ASD probands compared to parental controls (p value < 0.05) were considered risk homozygous haplotypes (rHH) (Supplementary Tables 1a-m). All subsequent analysis was limited to these regions. The rHH regions were annotated using the Illumina 1M annotation file (Human Genome build 36.1 RefSeq). In cases where rHH were located in intergenic regions, the nearest centromeric and telomeric gene was noted.

Inspection of LD structure and copy number variants

Patterns of linkage disequilibrium (LD) within the rHH were visualised and analysed in Haploview using HapMap CEU as a reference (Barrett et al. 2005). LD was measured as r^2 values and calculated between each pair of SNPs. The *Tagger* algorithm (Haploview) was

used to determine the number of tagging SNPs within each rHH at an r^2 threshold of 0.8. Genes located in rHH comprising < 10 tagging SNPs were noted. For each stage 1 population group, the samples contributing to significant rHH were inspected for CNV content as detected by Pinto and colleagues (Pinto et al. 2010). If present, rHH identified as CNVs were removed prior to application of the Fisher's exact test.

Replication study

A replication study was undertaken using the AGP follow up stage 2 data set (freezes 5-8) which comprises 1,182 ASD trios genotyped on a combination of the Illumina 1M and 1M-duo platforms. The stage 2 data was cleaned with the stage 1 samples to ensure that the same markers would be used in both analyses. The stage 1 and stage 2 probands were clustered to identify ancestry-matched replication groups. The 4 population clusters (C3-C6) with ≥50 probands in both stage 1 and 2 analyses were considered for the replication study. The same HH mapping method was applied to the four stage 2 population clusters. The rHH genes identified in the discovery (stage 1) and replication (stage 2) analyses were then compared to identify overlap.

Statistical analyses

Comparison of ROH burden. For each population group, the number and length of autosomal ROH in ASD probands and parental controls was compared using a paired t-test. ASD probands were compared to their mothers and fathers separately.

Identification of HHs that are more prevalent in ASD. A Fisher's exact test was used to identify HH that were more prevalent in ASD probands compared to parental controls at a

5% Fisher's significance level. As we are assuming a recessive model a one-sided test was used.

Comparison of results across the 5 stage 1 population groups. Each of the 5 population clusters were analysed independently and the rHH genes subsequently compared to identify overlaps. A gene was considered a candidate in multiple population clusters regardless of whether the position of the rHH in each population differed (to allow for population-specific effects). In this manner, it is possible to identify genes that may confer susceptibility to ASD across multiple populations even though the underlying risk allele may be population-specific. Genes located in/near rHH over-represented in ASD probands in at least two population groups were noted.

Enrichment for previously reported ASD genes. A χ^2-test with Yates correction factor was used to determine if the rHH genes displayed enrichment for previously reported ASD candidate genes. To account for a potential bias towards large genes, a logistic multiple regression was performed in STATA including both ASD gene status and gene size as covariates.

Results

Identification of risk homozygous haplotypes (rHH)

Runs of homozygosity (ROH) of at least 1Mb and 100 SNPs were identified in samples from each of the 5 population clusters using PLINK (Methods and Supplementary Material 1). A paired t-test demonstrated that ASD probands did not have a higher genome-wide burden of ROH compared to parental controls (Supplementary Fig. 3). The findings from the current ASD study reflect those of a related neuropsychiatric condition, bipolar disorder, where it

was recently reported that individuals with bipolar disorder do not have an excess of runs of homozygosity (Vine et al. 2009). A 1-sided Fisher's exact test was applied to identify HH that were more prevalent in ASD probands compared to parental controls at a 5% significance threshold. Such HH are considered regions of interest and are referred to as risk homozygous haplotypes (rHH). Since low frequency variants will rarely be present at a high enough frequency to survive multiple testing they will not be detected with the correction methods currently available. Therefore unless otherwise stated, *p* values have not been corrected for multiple testing. Genes located within or overlapping the rHH were noted. In cases where the rHH was intergenic, the neighbouring centromeric and telomeric genes were used. We found that, on average, 76% of the rHH were genic (Supplementary Fig. 4).

Haplotype sharing in homozygous segments

To determine if excess HH sharing is also likely to occur in a non-disease cohort we compared the number of HHs that are more common in ASD probands compared to parental controls and the number of HHs that are more common in parental controls compared to ASD cases. We observed that ASD probands shared a significantly higher number of homozygous segments of identical haplotype compared to the parental control group (paired t-test *p* value = 0.008) (Fig. 3). This finding suggests that, although ASD probands may not have a higher burden of homozygous segments compared to parental controls, the probands display a much higher degree of haplotype sharing within overlapping homozygous regions. Excess haplotype sharing often indicates the presence of a disease locus (Bahlo et al. 2006), an observation that forms the basis of the current study. Two distinct types of rHH were identified in the affected cohort; 1) homozygous segments with a haplotype that is present in ASD probands but absent in parental controls and 2) homozygous segments with a haplotype that is present at a higher frequency in ASD cases compared to parental controls. A summary

of the rHH results are shown in Table 1. The average rHH size across the 5 population clusters was 541.27 kb, 24-fold larger than the average haplotype block (Gabriel et al. 2002). Visual inspection of LD structure showed that the rHH extended across multiple LD blocks and 90% of rHH contained at least 10 tagging SNPs, indicating that it is unlikely that the observed HH sharing is a consequence of long-range LD.

Genes located in rHH across multiple population clusters

We identified 192 genes (1 gene in 4 populations, 15 in 3 populations and 176 in 2 populations) that are in or near rHH regions significantly more prevalent among the ASD probands in two or more population clusters (Supplementary Table 4). In 4 of the 5 population clusters, an rHH was found in an evolutionary conserved intergenic region on 3p12.1 in the vicinity of *CADM2*, a novel ASD candidate gene (Fig. 4). The rHHs adjacent to *CADM2* were identified in 23/1019 ASD cases and 11/2031 parental controls (χ p value = 1.9×10^{-5}, OR = 4.26 (2.1, 8.6)). Recent studies have highlighted the presence of long-range regulatory mutations in several diseases and suggest that *cis*-regulatory domains of developmental genes may extend over megabases of DNA flanking its coding sequences (Benko et al. 2009; Kleinjan and van Heyningen 2005). *CADM2* is a member of the synaptic cell adhesion molecule (SynCAM) immunoglobulin superfamily which has potential roles in early postnatal development of the central nervous system (specifically synapse formation) and is predominantly expressed in neurons of the developing and adult brain (Thomas et al. 2008). The encoded protein localises to both excitatory and inhibitory neurons which is intriguing given that there is compelling evidence for dysfunction in neuronal communication and excitatory/inhibitory imbalance in autism (Persico and Bourgeron 2006; Sudhof 2008). Of interest, *CADM2* contains a neurexin domain. Rare variation contributing to autism has previously been identified in neurexin and neurexin-binding genes (Arking et al. 2008;

Betancur et al. 2009; Kim et al. 2008). In our study, the location of the rHH in relation to *CADM2* differed between each of the 4 population clusters. This may indicate population-specific risk loci arising from different founder events involving the gene or a control element proximal to the gene. Given the recent implication of synaptic genes in ASD susceptibility (Durand et al. 2008; Zoghbi 2003) and the strong functional candidacy of *CADM2*, its involvement in ASD warrants further investigation.

Other novel ASD candidate genes located in rHH across multiple population clusters include *GRIK2;* an ionotropic glutamate receptor associated with autosomal recessive mental retardation and autism (Jamain et al. 2002; Motazacker et al. 2007), *GRM3*; a metabotropic glutamate receptor that impacts on aspects of cognition dependent on hippocampal and prefrontal cortical function (Egan et al. 2004), *EPHA3;* an ephrin receptor involved in axon guidance and synaptic plasticity, *FGF10;* a fibroblast growth factor involved in regulating the onset of neurogenesis and cortical brain size (Sahara and O'Leary 2009), *ABHD14A*; a cerebellar abhydrolase protein with a role in granule neuron development (Hoshino et al. 2003), *NTS*; a brain and gastrointestinal peptide involved in passive avoidance behaviour and the modulation of dopaminergic neurotransmission and serotonin levels (Shugalev et al. 2008), *SCAMP5*; a brain-enriched membrane trafficking protein located 10kb centromeric to the breakpoint of a 15q de novo balanced translocation in a patient with autism (Castermans et al. 2010), *CHRFAM7A*; a nicotinic acetylcholine receptor in the postsynaptic membrane that has previously been associated with schizophrenia, bipolar disorder, dementia, Alzheimer's disease and attention-deficit hyperactivity disorder and *KCND2;* a brain-specific voltage-gated ion channel component that regulates neurotransmitter release and neuronal excitability at the glutamatergic synapse where *SHANK3* and *NLGN* products are formed.

ASD-specific rHH across multiple populations

Of particular interest in this study are the rHHs that are ASD-specific; shared by multiple ASD patients but not present in any of the parental controls. We identified 8 ASD-specific rHH, implicating 12 genes (*POLR3C*, *CD160*, *ZNF364*, *PDZK1*, *NUDT17*, *GBE1*, *HTR1E*, *C10orf95*, *CUECDC2*, *CHORDC1*, *MGAT4C* and *C12orf50*) and a cluster of defensins (8p23.1), which are significant in at least two population clusters. The ASD-specific rHH at 1q21.1 is shared by three population clusters and the maximal overlapping region contains a single gene; *PDZK1*. PDZK1 encodes a scaffold protein that connects plasma membrane proteins and regulatory components. The PDZK1 protein is part of a complex that includes SYNGAP1, KLHL17 and NMDA receptors and this complex is important for maintaining the integrity of actin cytoskeleton structures in neurons (Chen and Li 2005). In addition, the ASD-specific rHH at 1q21.1 overlaps with a rare deletion associated with schizophrenia in two independent studies (Tam et al. 2009).

Enrichment for genes previously implicated in ASD

The genes located in rHH regions were compared to autosomal genes that have previously been implicated in ASD in association studies, expression analyses and chromosomal anomaly studies, as reviewed by Yang and Gill 2007 (Yang and Gill 2007). We extended the literature search to 2009 using the same search criteria provided by Yang and Gill (Supplementary Table 5). Of the 1,243 genes identified in the study, 25 are published ASD candidate genes (corrected $p = 8.0 \times 10^{-5}$, OR = 2.36 (1.55, 3.61)) (Table 2). However, considerable caution is required in interpreting such findings, since both rHH genes and previously reported ASD candidate genes tend to be larger ($p < 10^{-3}$). Accordingly, we performed a logistic multiple regression including both ASD gene status and gene size as covariates. While the Odds Ratio decreased slightly, it remained highly significant, indicating that there is an association between rHH status and ASD status that is independent of gene

size (correct p = 0.001, OR = 2.10, (1.36, 3.25). In addition, 9 ASD candidates (*BTN2A1*, *FOXP2*, *GBE1*, *GRIK2*, *IMMP2L*, *LRFN5*, *LRRN3*, *ROBO1*, *SLC4A10*) are amongst the 192 genes located in rHH in two or more population clusters (corrected $p = 1.3 \times 10^{-5}$, OR = 4.47 (2.29, 8.76)).

Replication study

To investigate our findings further we performed a replication study on an additional 1,182 independent AGP trios. The replication cohort (stage 2) was clustered with the discovery sample set (stage 1) to identify ancestry-matched replication groups (Supplementary Material 1). The four population clusters (C3-C6) with at least 50 probands in both the discovery and replication sample sets were considered for the replication study (Supplementary Table 6). The same analytical strategy was applied to the replication sample set and identified 1,190 genes in rHH regions. The rHH genes identified in the discovery (number of genes = 1,086) and replication (number of genes = 1,190) analyses of population clusters C3-C6 were subsequently compared. We found that 28.5% (310/1086 genes) of the rHH genes identified in the discovery analysis occurred in a rHH in at least one of the replication population clusters (Supplementary Table 7). In particular, the replication study provided further evidence for the possible involvement of the novel candidate genes *ABHD14A*, *CADM2*, *EPHA3*, *FGF10*, *GRIK2*, *GRM3*, and *KCND2*. *SCAMP5* and *CHRFAM7A* (found in rHH in multiple population clusters in the discovery analysis) did not replicate in their corresponding replication clusters while *NTS* was significant in one of two ancestry-matched replication groups. Furthermore, 10 previously identified ASD candidate genes (*ALAS1*, *CSMD3*, *FOXP2*, *GABRG1*, *GBE1*, *GRM8*, *FBXO33*, *IMMP2L*, *SLC4A10* and *ACO2*) were located in rHH regions in both the discovery and replication HH mapping analyses (corrected p = 0.001, OR = 3.06 (1.51, 6.01)). The ASD-specific rHH at 1p21.1 was also significant in two of four

population clusters in the replication study providing further evidence of an ASD risk gene at this locus.

rHH located in significant ASD linkage peaks

The rHH regions identified in the discovery (stage 1) and replication (stage 2) analyses were compared to the genomic regions that have previously displayed significant linkage (LOD score > 3.3) with ASD. We found that 31 rHH identified in the discovery and replication analyses are located under significant ASD linkage peaks (Supplementary Table 8). Of interest, three of the ASD linkage peaks harbour rHH from multiple population groups.

Firstly, three population groups in the replication analysis (stage 2: C4, C5, C6) have an overlapping rHH within the 7q22.1 ASD linkage peak (International Molecular Genetic Study of Autism Consortium (IMGSAC) 2001). The rHHs at 7q22.1 range in size from 545 kb to 808 kb. The maximal shared region is 545.3 kb and includes 19 genes. Three of the genes at this locus (*CYP3A4*, *CYP3A5* and *CYP3A7*) are members of the cytochrome P450 superfamily that plays important roles in hormone synthesis and breakdown, cholesterol synthesis, vitamin D metabolism and the metabolism of drugs and toxic compounds. Data from the KEGG databases suggests that members of the cytochrome P450 family such as *CYP3A4* and *CYP3A5* are involved in 5-HT (serotonin) catabolism (Jia et al. 2004). Moreover, in the brain, mRNA expression appears to be specific to neurons in the cerebral cortex and basal ganglia, regions of the brain that are consistently implicated in ASD (Dutheil et al. 2008). In addition, genetic polymorphisms of cytochrome P450 enzymes have been linked to ASD and schizophrenia (Currenti 2010).

Secondly, a rHH shared by two populations in the discovery analysis is located within the significant ASD linkage peak at 15q13.1-q14 (Lauritsen et al. 2006; Liu et al. 2008; Philippe et al. 1999). This particular locus has been implicated in several neuropsychiatric

disorders. The overlapping rHH region identified in the current study is 579 kb and contains 5 genes, only two of which are coding (*CHRFAM7A* and *ARHGAP11B*). Of most relevance to ASD is *CHRFAM7A* which encodes a nicotinic acetylcholine receptor in the postsynaptic membrane. *CHRFAM7A* has previously been associated with schizophrenia, bipolar disorder, dementia, Alzheimer's and attention-deficit hyperactivity disorder (Feher et al. 2009; Flomen et al. 2006; Manchia et al. 2010; Martin et al. 2007; Sinkus et al. 2009). This is the first implication of *CHRFAM7A* in ASD.

Thirdly, we noted a rHH at 20q11.21-a13.12 that was significantly more common in ASD patients compared to parental controls in two population groups in the replication study. The shared rHH region includes 11 genes. The strongest functional candidate at this locus is *MYH7B* which encodes a myosin heavy chain protein that maintains excitatory synaptic function. Reduction of MyH7B in rats causes a
profound alteration in dendritic spine structure and excitatory synaptic strength (Rubio et al. 2011). The resulting abnormal spine phenotype was strikingly similar to the morphological changes induced by over-expression of SynGAP1, a gene linked to mental retardation and ASD (Hamdan et al. 2009; Pinto et al. 2010). The advantage of the rHH analysis is that the ASD-specific rHH regions are much smaller (~0.53 Mb) than ASD linkage regions which range from 2-242Mb and may facilitate identification of the causative gene within these loci.

Discussion

Despite some notable successes in neuropsychiatric genetics, overall the high heritability of ASD (~90%) remains poorly explained by common genetic risk variants. Instead, early studies of rare variation, in particular copy number variation, have suggested that rare variants may account for a significant proportion of the genetic basis of ASD. The study of excess haplotype sharing within homozygous regions offers a complimentary approach to the analysis of GWAS data in the study of complex disease and particularly focuses on the

contribution of low-frequency variation to disease risk. We report the first genome-wide rHH mapping study to identify novel recessive candidate genes and loci involved in ASD susceptibility. The strategy was applied to one of the largest ASD trio collections, subdivided into 10 distinct population clusters. We identified 307 rHH regions containing 1,243 genes for further investigation. In cases where ancestry-matched replication groups were available, almost one third of the discovery phase rHH genes (310/1,086) were confirmed. Importantly, we identified novel ASD genes that merit further study including *ABHD14A*, *CADM2*, *CHRFAM7A*, *EPHA3*, *FGF10*, *GRIK2*, *GRM3* and *KCND2*. The current study also provides further support for 25 previously reported ASD genes. Interestingly 192 of the rHH genes were significant in multiple population clusters. The variation in haplotype of the rHH across the different populations may suggest the existence of common risk genes but population-specific risk alleles. The findings of the current study serve as a starting point to screen for causal variants and elucidate the underlying pathogenesis of ASD. In particular, sequence analysis will be more economically feasible since the search for causal variants is limited to 1,243 genes and specifically identifies the patients carrying the rHH of interest, allowing a more targeted follow-up approach.

Sample size as a limiting factor

Sample size is an important feature of the HH mapping approach and is a limiting factor in the current study. A number of the population clusters in the ASD study are of modest sample size ($n < 100$) and 5 population clusters could not be included because of insufficient sample numbers. Larger collections of genetically homogenous samples would increase the power to detect low frequency events. Similarly, due to the small sample sizes, we have not corrected for multiple testing. The HH sharing strategy aims to identify genes that may contain low-frequency disease variants. Given the small sample sizes, such events are highly unlikely to survive correction for multiple testing. Although this may be considered a weakness it is

important to note that correcting for multiple testing will not change the relative order of the rHH results [Zaykin and Zhivotovsky, 2005]. The distributions of ranks will remain the same with HHs that show the greatest difference in frequency between ASD cases and parental controls remaining as the most important findings. We acknowledge that by not correcting for multiple testing, it is possible that there are some false positive results within the findings. To address this we have focused on the genes that occur in rHHs in multiple population clusters and replicate in the stage 2 sample set, as they are less likely to be false positives.

Genomic structure underlying rHH regions

Analysis of Log R ratios and B allele frequencies confirmed that the rHH regions are not attributed to copy number deletions. To ascertain whether some unknown genomic architecture contributed to the observed rHH findings and also to reduce the risk of false positives, we undertook rHH mapping in two additional disease data sets from the Wellcome-Trust Case-Control Consortium; bipolar disorder (BD) (cases=1,875 and controls = 2,954) and coronary artery disease (CAD) (cases= 1,963 and controls = 2,978). We hypothesised that genome architecture could possibly be contributing to the results of the ASD rHH analysis if 1) the rHH in the BD and CAD studies showed a significant overlap with the rHH identified in the current ASD study and 2) the BD and CAD rHH genes showed a significant enrichment for previously identified ASD genes. There was a 2% and 2.6% overlap between BD-ASD and CAD-ASD rHH regions respectively, neither of which are greater than expected by chance (J.P.C., unpublished data). Furthermore the BD and CAD rHH genes did not display an enrichment for known ASD candidate genes (BD corrected p = 0.815, CAD corrected p = 1.000) suggesting that the rHH identified in the current study are more likely to be related to the ASD phenotype than to the genomic architecture in the rHH regions.

Homozygous haplotype sharing in complex disorders

There are very few analytical strategies designed to identify rare recessive disease genes for complex traits. The homozygous haplotype sharing strategy addresses this issue and proposes a novel concept for searching for genes that may contain low-frequency disease variants. The most important feature of the HH mapping approach presented in this thesis is the concept; analysis of the haplotype within shared homozygous segments provides an additional level of information that has been overlooked in ROH-based analyses and excess sharing of a homozygous haplotype amongst patients may support the presence of a rare recessive disease mutation in the region. In the future, the strategy itself will undoubtedly benefit from further modifications and improvements, particularly in the areas of modelling, simulation and statistics for rare genetic events.

We applied the HH mapping approach to one of the largest international ASD cohorts (4,206 samples) and identified novel and known ASD candidate genes which were replicated in an independent sample set. We also found that homozygous haplotypes over-represented in ASD patients were significantly enriched for previously identified ASD candidate genes, further validating our approach. Although HH mapping does not identify causative alleles, the regions reported in the current study provide narrow genomic intervals containing highly plausible candidate genes for further investigation. The findings reported in this study suggest that the analysis of homozygous haplotype sharing may be an important tool in uncovering some of the missing heritability in a variety of complex disorders.

Acknowledgements

The authors acknowledge the families participating in the study and the main funders of the Autism Genome Project Consortium (AGP): Autism Speaks (USA), the Health Research Board (HRB; Ireland), The Medical Research Council (MRC; UK), Genome Canada/Ontario Genomics Institute, and the Hilibrand Foundation (USA). Additional support for individual groups was provided by the National Children's Research Centre (NCRC) Our Lady's Children's Hospital Crumlin Ireland, US National Institutes of Health (NIH grants HD055751, HD055782, HD055784, HD35465, MH52708, MH55284, MH57881,

MH061009, MH06359, MH066673, MH080647, MH081754, MH66766, NS026630, NS042165, NS049261), the Canadian Institute for Advanced Research (CIFAR), the Canadian Institutes for Health Research (CIHR), Assistance Publique–Hôpitaux de Paris (France), Autistica, Canada Foundation for Innovation/Ontario Innovation Trust, Deutsche Forschungsgemeinschaft (grant Po 255/17-4) (Germany), EC Sixth FP AUTISM MOLGEN, Fundação Calouste Gulbenkian (Portugal), Fondation de France, Fondation FondaMental (France), Fondation Orange (France), Fondation pour la Recherche Médicale (France), Fundação para a Ciência e Tecnologia (Portugal), the Hospital for Sick Children Foundation and University of Toronto (Canada), INSERM (France), Institut Pasteur (France), the Italian Ministry of Health (convention 181 of 19.10.2001), the John P Hussman Foundation (USA), McLaughlin Centre (Canada), Ontario Ministry of Research and Innovation (Canada), the Seaver Foundation (USA), the Swedish Science Council, The Centre for Applied Genomics (Canada), the Utah Autism Foundation (USA) and the Wellcome Trust core award 075491/Z/04 (UK). We acknowledge support from the Autism Genetic Resource Exchange (AGRE) and Autism Speaks. We gratefully acknowledge the resources provided by the AGRE consortium* and the participating AGRE families. AGRE 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). We wish to acknowledge Wellcome Trust Case-Control Consortium for providing data sets that were used as part of this study. J.P.C is supported by an EMBARK postgraduate award from the Irish Research Council for Science, Engineering and Technology (IRCSET).

*The AGRE Consortium:
Dan Geschwind, M.D., Ph.D., UCLA, Los Angeles, CA;
Maja Bucan, Ph.D., University of Pennsylvania, Philadelphia, PA;
W.Ted Brown, M.D., Ph.D., F.A.C.M.G., N.Y.S. Institute for Basic Research in Developmental Disabilities, Staten Island, NY;
Rita M. Cantor, Ph.D., UCLA School of Medicine, Los Angeles, CA;
John N. Constantino, M.D., Washington University School of Medicine, St. Louis, MO;
T.Conrad Gilliam, Ph.D., University of Chicago, Chicago, IL;
Joachim Hallmayer, Ph.D., Stanford University, Stanford, CA;
Martha Herbert, M.D., Ph.D., Harvard Medical School, Boston, MA
Clara Lajonchere, Ph.D, Autism Speaks, Los Angeles, CA;
David H. Ledbetter, Ph.D., Emory University, Atlanta, GA;
Christa Lese-Martin, Ph.D., Emory University, Atlanta, GA;
Janet Miller, J.D., Ph.D., Autism Speaks, Los Angeles, CA;
Stanley F. Nelson, M.D., UCLA School of Medicine, Los Angeles, CA;
Gerard D. Schellenberg, Ph.D., University of Washington, Seattle, WA;
Carol A. Samango-Sprouse, Ed.D., George Washington University, Washington, D.C.;
Jonathan Shestack, Autism Speaks, NY, NY;
Sarah Spence, M.D., Ph.D., UCLA, Los Angeles, CA;
Matthew State, M.D., Ph.D., Yale University , New Haven, CT.
Rudolph E. Tanzi, Ph.D., Massachusetts General Hospital, Boston, MA.

Ethical standards and conflict of interest

The experiments comply with the current laws of the country in which they were performed. The authors declare that they have no conflict of interest.

References

Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M, Rea A, Guy M, Lin S, Cook EH, Chakravarti A (2008) A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. Am J Hum Genet 82: 160-4

Bahlo M, Stankovich J, Speed TP, Rubio JP, Burfoot RK, Foote SJ (2006) Detecting genome wide haplotype sharing using SNP or microsatellite haplotype data. Hum Genet 119: 38-50

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-5

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez MP, Vekemans M, Roest Crollius H, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, Fitzpatrick DR, Lyonnet S (2009) Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet 41: 359-64

Betancur C, Sakurai T, Buxbaum JD (2009) The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders. Trends Neurosci 32: 402-12

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598-604

Castermans D, Volders K, Crepel A, Backx L, De Vos R, Freson K, Meulemans S, Vermeesch JR, Schrander-Stumpel CT, De Rijk P, Del-Favero J, Van Geet C, Van De Ven WJ, Steyaert JG, Devriendt K, Creemers JW (2010) SCAMP5, NBEA and AMISYN: three candidate genes for autism involved in secretion of large dense-core vesicles. Hum Mol Genet 19: 1368-78

Chen Y, Li M (2005) Interactions between CAP70 and actinfilin are important for integrity of actin cytoskeleton structures in neurons. Neuropharmacology 49: 1026-41

Currenti SA (2010) Understanding and determining the etiology of autism. Cell Mol Neurobiol 30: 161-71

Durand CM, Betancur C, Boeckers TM, Bockmann J, Chaste P, Fauchereau F, Nygren G, Rastam M, Gillberg IC, Anckarsater H, Sponheim E, Goubran-Botros H, Delorme R, Chabane N, Mouren-Simeoni MC, de Mas P, Bieth E, Roge B, Heron D, Burglen L, Gillberg C, Leboyer M, Bourgeron T (2007) Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. Nat Genet 39: 25-7

Durand CM, Chaste P, Fauchereau F, Betancur C, Leboyer M, Bourgeron T (2008) Alterations in synapsis formation and function in autism disorders. Med Sci (Paris) 24: 25-8

Dutheil F, Beaune P, Loriot MA (2008) Xenobiotic metabolizing enzymes in the central nervous system: Contribution of cytochrome P450 enzymes in normal and pathological human brain. Biochimie 90: 426-36

Egan MF, Straub RE, Goldberg TE, Yakub I, Callicott JH, Hariri AR, Mattay VS, Bertolino A, Hyde TM, Shannon-Weickert C, Akil M, Crook J, Vakkalanka RK, Balkissoon R, Gibbs RA, Kleinman JE, Weinberger DR (2004) Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. Proc Natl Acad Sci U S A 101: 12604-9

Feher A, Juhasz A, Rimanoczy A, Csibri E, Kalman J, Janka Z (2009) Association between a genetic variant of the alpha-7 nicotinic acetylcholine receptor subunit and four types of dementia. Dement Geriatr Cogn Disord 28: 56-62

Flomen RH, Collier DA, Osborne S, Munro J, Breen G, St Clair D, Makoff AJ (2006) Association study of CHRFAM7A copy number and 2 bp deletion polymorphisms with schizophrenia and bipolar affective disorder. Am J Med Genet B Neuropsychiatr Genet 141B: 571-5

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-9

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5: 184-186

Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, Pellerin S, Dobrzeniecka S, Cote M, Perreau-Linck E, Carmant L, D'Anjou G, Fombonne E, Addington AM, Rapoport JL, Delisi LE, Krebs MO, Mouaffak F, Joober R, Mottron L, Drapeau P, Marineau C, Lafreniere RG, Lacaille JC, Rouleau GA, Michaud JL (2009) Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N Engl J Med 360: 599-605

Hildebrandt F, Heeringa SF, Ruschendorf F, Attanasio M, Nurnberg G, Becker C, Seelow D, Huebner N, Chernin G, Vlangos CN, Zhou W, O'Toole JF, Hoskins BE, Wolf MT, Hinkes BG, Chaib H, Ashraf S, Schoeb DS, Ovunc B, Allen SJ, Vega-Warner V, Wise E, Harville HM, Lyons RH, Washburn J, Macdonald J, Nurnberg P, Otto EA (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet 5: e1000353

Hoshino J, Aruga J, Ishiguro A, Mikoshiba K (2003) Dorz1, a novel gene expressed in differentiating cerebellar granule neurons, is down-regulated in Zic1-deficient mouse. Brain Res Mol Brain Res 120: 57-64

International Molecular Genetic Study of Autism Consortium (IMGSAC) (2001) A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. Am J Hum Genet 69: 570-81

Jamain S, Betancur C, Quach H, Philippe A, Fellous M, Giros B, Gillberg C, Leboyer M, Bourgeron T (2002) Linkage and association of the glutamate receptor 6 gene with autism. Mol Psychiatry 7: 302-10

Jia Y, Yu X, Zhang B, Yuan Y, Xu Q, Shen Y (2004) No association between polymorphisms in three genes of cytochrome p450 family and paranoid schizophrenia in northern Chinese Han population. Eur Psychiatry 19: 374-6

Kim HG, Kishikawa S, Higgins AW, Seong IS, Donovan DJ, Shen Y, Lally E, Weiss LA, Najm J, Kutsche K, Descartes M, Holt L, Braddock S, Troxell R, Kaplan L, Volkmar F, Klin A, Tsatsanis K, Harris DJ, Noens I, Pauls DL, Daly MJ, MacDonald ME, Morton CC, Quade BJ, Gusella JF (2008) Disruption of neurexin 1 associated with autism spectrum disorder. Am J Hum Genet 82: 199-207

Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76: 8-32
Lamb JA, Moore J, Bailey A, Monaco AP (2000) Autism: recent molecular genetic advances. Hum Mol Genet 9: 861-8
Lauritsen MB, Als TD, Dahl HA, Flint TJ, Wang AG, Vang M, Kruse TA, Ewald H, Mors O (2006) A genome-wide search for alleles and haplotypes associated with autism and related pervasive developmental disorders on the Faroe Islands. Mol Psychiatry 11: 37-46
Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A 104: 19942-7
Lesch KP, Timmesfeld N, Renner TJ, Halperin R, Roser C, Nguyen TT, Craig DW, Romanos J, Heine M, Meyer J, Freitag C, Warnke A, Romanos M, Schafer H, Walitza S, Reif A, Stephan DA, Jacob C (2008) Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. J Neural Transm 115: 1573-85
Liu XQ, Paterson AD, Szatmari P (2008) Genome-wide linkage analyses of quantitative and categorical autism subphenotypes. Biol Psychiatry 64: 561-70
Manchia M, Viggiano E, Tiwari AK, Renou J, Jain U, De Luca V, Kennedy JL (2010) Smoking in adult attention-deficit/hyperactivity disorder: interaction between 15q13 nicotinic genes and Temperament Character Inventory scores. World J Biol Psychiatry 11: 506-10
Martin LF, Leonard S, Hall MH, Tregellas JR, Freedman R, Olincy A (2007) Sensory gating and alpha-7 nicotinic receptor gene allelic variants in schizoaffective disorder, bipolar type. Am J Med Genet B Neuropsychiatr Genet 144B: 611-4
Meldrum BS (2000) Glutamate as a neurotransmitter in the brain: review of physiology and pathology. J Nutr 130: 1007S-15S
Motazacker MM, Rost BR, Hucho T, Garshasbi M, Kahrizi K, Ullmann R, Abedini SS, Nieh SE, Amini SH, Goswami C, Tzschach A, Jensen LR, Schmitz D, Ropers HH, Najmabadi H, Kuss AW (2007) A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. Am J Hum Genet 81: 792-8
Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19: 2927-2935
Persico AM, Bourgeron T (2006) Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. Trends Neurosci 29: 349-58
Philippe A, Martinez M, Guilloud-Bataille M, Gillberg C, Rastam M, Sponheim E, Coleman M, Zappella M, Aschauer H, Van Maldergem L, Penet C, Feingold J, Brice A, Leboyer M (1999) Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. Hum Mol Genet 8: 805-12
Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL,

Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368-72

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-9

Purcell AE, Jeon OH, Zimmerman AW, Blue ME, Pevsner J (2001) Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. Neurology 57: 1618-28

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-75

Rubio MD, Johnson R, Miller CA, Huganir RL, Rumbaugh G (2011) Regulation of synapse structure and function by distinct myosin II motors. J Neurosci 31: 1448-60

Sahara S, O'Leary DD (2009) Fgf10 regulates transition period of cortical stem cell differentiation to radial glia controlling generation of neurons and basal progenitors. Neuron 63: 48-62

Serajee FJ, Zhong H, Nabi R, Huq AH (2003) The metabotropic glutamate receptor 8 gene at 7q31: partial duplication and possible association with autism. J Med Genet 40: e42

Shuang M, Liu J, Jia MX, Yang JZ, Wu SP, Gong XH, Ling YS, Ruan Y, Yang XL, Zhang D (2004) Family-based association study between autism and glutamate receptor 6 gene in Chinese Han trios. Am J Med Genet B Neuropsychiatr Genet 131B: 48-50

Shugalev NP, Stavrovskaya AV, Ol'shanskii AS, Hartmann G, Lenard L (2008) Serotoninergic mechanisms of the effects of neurotensin on passive avoidance behavior in rats. Neurosci Behav Physiol 38: 517-21

Sinkus ML, Lee MJ, Gault J, Logel J, Short M, Freedman R, Christian SL, Lyon J, Leonard S (2009) A 2-base pair deletion polymorphism in the partial duplication of the alpha7 nicotinic acetylcholine gene (CHRFAM7A) on chromosome 15q14 is associated with schizophrenia. Brain Res 1291: 1-11

Sudhof TC (2008) Neuroligins and neurexins link synaptic function to cognitive disease. Nature 455: 903-11

Tam GW, Redon R, Carter NP, Grant SG (2009) The role of DNA copy number variation in schizophrenia. Biol Psychiatry 66: 1005-12

The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-96

Thomas LA, Akins MR, Biederer T (2008) Expression and adhesion profiles of SynCAM molecules indicate distinct neuronal functions. J Comp Neurol 510: 47-67

van der Laan MJaPKS (2002) A new algorithm for hybrid hierarchical clustering with visualisation and the bootstrap. J Stat Planning and Inference 117: 275-303

Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Robinson M, Lawrence J, Anjorin A, Sklar P, Gurling HM, Curtis D (2009) No evidence for excess runs of homozygosity in bipolar disorder. Psychiatr Genet 19: 165-70

Wong W, Newell EW, Jugloff DG, Jones OT, Schlichter LC (2002) Cell surface targeting and clustering interactions between heterologously expressed PSD-95 and the Shal voltage-gated potassium channel, Kv4.2. J Biol Chem 277: 20423-30

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: 69-85

Yang TL, Guo Y, Zhang LS, Tian Q, Yan H, Papasian CJ, Recker RR, Deng HW (2010) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. J Clin Endocrinol Metab 95: 3777-82

Zoghbi HY (2003) Postnatal neurodevelopmental disorders: meeting at the synapse? Science 302: 826-30

Figures

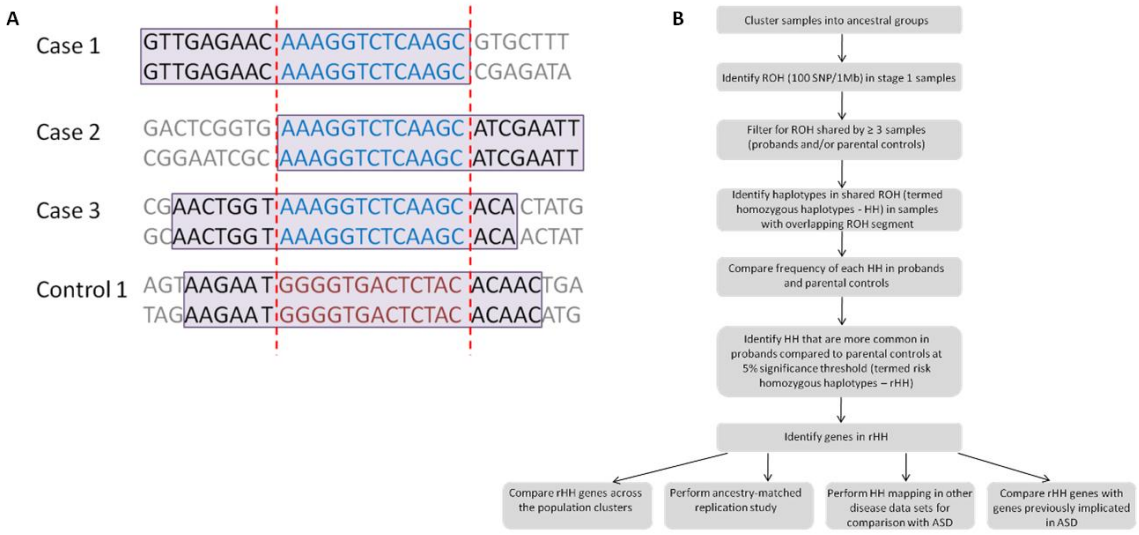


Fig 1 The principles and analytical approach of homozygous haplotype mapping. A: The schematic outlines the principle of homozygous haplotype (HH) mapping. SNP genotype data is collected on each case and control. Homozygous and heterozygous SNPs are shown in black and grey respectively. Firstly, runs of homozygosity (ROH) are identified in the samples (outlined in purple boxes). The overlapping ROH region shared by a minimum of 3 individuals (shown between red dashed lines) is considered for the HH analysis. The haplotypes within the overlapping ROH region are identified and a fisher's test applied to determine if a particular HH is significantly more common in ASD cases compared to parental controls. Only the haplotypes of those individuals who have an ROH in the region in question are considered. In the above example all 4 individuals (3 ASD cases and 1 parental control) have an overlapping ROH. However the haplotype in the overlapping ROH may differ. The 3 ASD cases have haplotype A (blue) while the parental control has haplotype B (red). Haplotype A is shared at a higher frequency in ASD cases compared to parental controls (apply fisher's test) and is termed a risk homozygous haplotype (rHH). This is an example of a rHH that is specific to ASD probands; B: Flowchart of homozygous haplotype analysis of ASD cohort. The discovery analysis was performed on 1182 AGP trios from the AGP stage 1 collection. The replication study involved an additional 1042 AGP trios from

the stage 2 collection. The stage 1 and 2 samples were clustered together to 1) separate stage 1 and 2 individuals into population clusters of similar ancestry and 2) classify stage 2 individuals into the joint ancestry-matched population clusters for the stage 2 replication study. The same rHH mapping strategy was applied to the discovery (stage 1) and replication (stage 2) data sets independently. The genes located in homozygous haplotypes significantly more common in ASD cases compared to parental controls were identified in each analysis. The rHH candidate genes were then compared for the ancestry-matched groups that had at least 50 probands in both the discovery and replication sample sets. To assess the contribution of genomic architecture to the rHH findings in ASD, the same strategy was applied to two additional disease data sets; bipolar disorder (BD) and coronary artery disease (CAD). The location of the rHH in ASD, BD and CAD were compared.

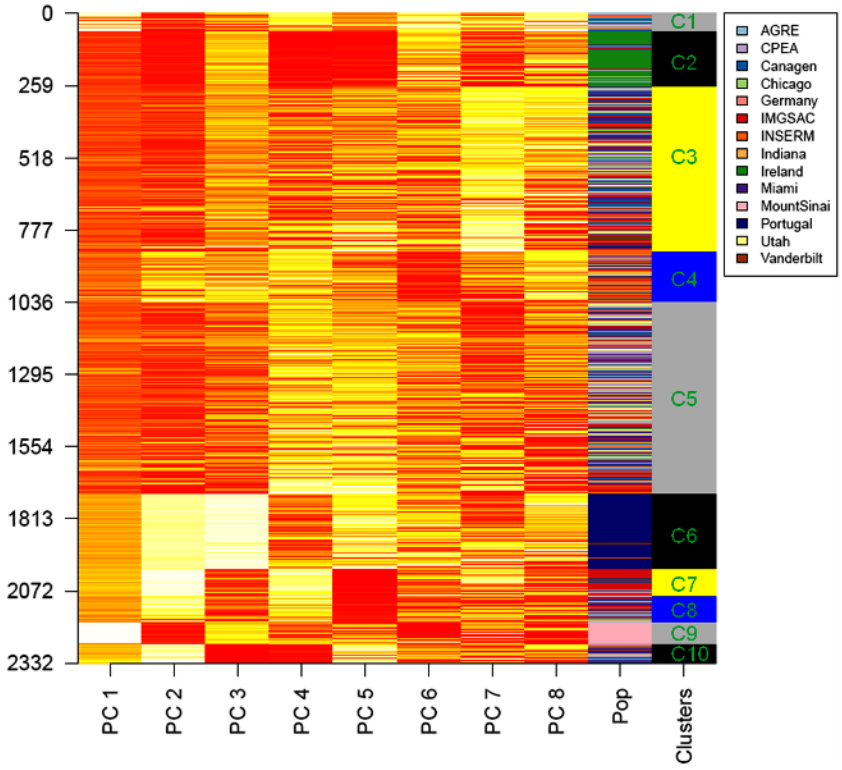


Fig 2 Genetic ancestry of AGP sample set. Principal component analysis (PCA) of 2584 ASD proband samples (discovery stage 1=1,402 samples, replication stage 2=1,182 samples) was performed in EIGENSOFT. Tracy-Widom statistics indicated that the first 8 principal components (PCs) were significantly contributing to the genetic variation of the sample set (Supplementary Table 1). The Hopach hierarchical clustering algorithm was applied to eigenvalues (y-axis) from the first 8 PCs (x-axis) (van der Laan 2002). In the 'Pop' column each sample is colored according to the AGP site at which it was collected (see legend). Hopach clustering, non-parametric bootstrapping and genetic distance calculations (Supplementary Tables 2 and 3) identified 10 ancestral population clusters labelled C1 to C10. The 5 population clusters with a minimum of 50 probands (C2-C6) were used in the discovery analysis (n = 1,019 trios).

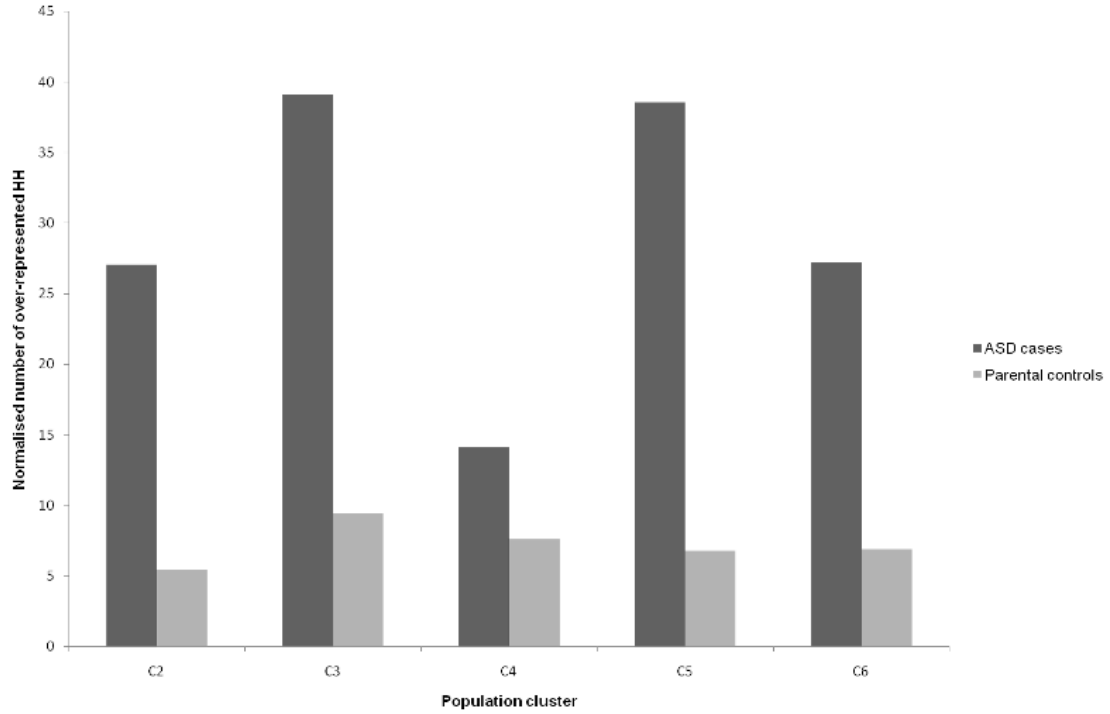


Fig 3 Comparison of HH sharing in ASD cases and parental controls. The normalised number of rHH (HH that are more common in one group compared to the other) in 5 population clusters with a minimum of 50 probands. The dark grey bars represent the number of HH that are more common (Fisher's exact test right p value < 0.05) in ASD probands compared to parental controls. Such regions are referred to as rHH throughout the paper. The light grey bars denote the number of HH that are more common (Fisher's exact test left p value < 0.05) in parental controls compared to ASD probands. To account for differences in sample size, counts have been normalised to a group of 100 samples (Supplementary Material 1). The number of rHH identified in ASD probands is significantly greater than the number of rHH identified in parental controls across the 5 population clusters (paired t-test p value = 0.008).

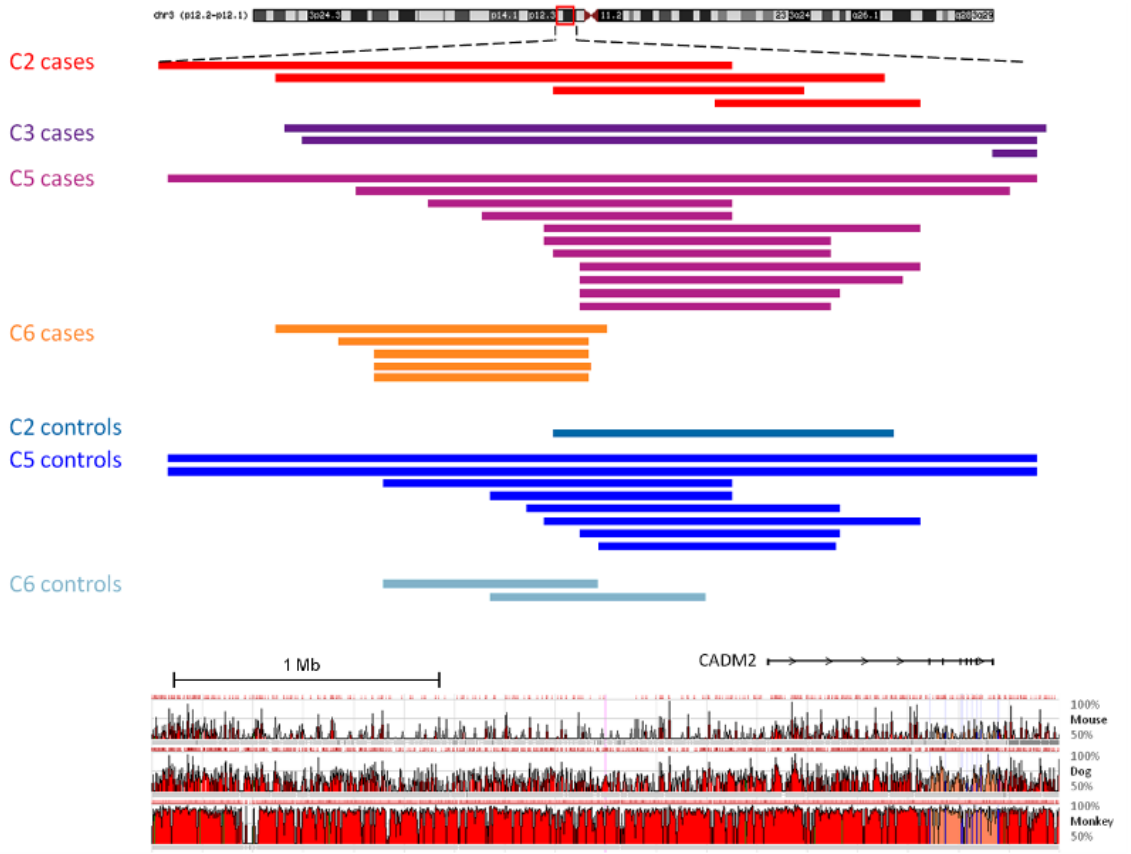


Fig 4 rHH identified in 4 population clusters in the vicinity of *CADM2*. An rHH located in a non-coding evolutionary-conserved region on 3p12.1 was identified in 4 of the 5 population clusters. The colored bars represent the run of homozygosity in each patient / parental control carrying the rHH. For each population cluster, the rHH is the shared ROH segment. The ROH profile is presented with a conservation plot (ECR browser; conservation throughout mouse, dog and rhesus monkey of fragments >350 bp at 75% identity indicated in red). rHHs adjacent to CADM2 were identified in 23/1019 ASD cases and 11/2031 parental controls (χ corrected *p* value = 1.9×10^{-5} , OR = 4.26 (2.1,8.6)).

Table 1 Summary of rHH results for population clusters

Cluster	**No. of ASD**	**No. of parental**	**Significant rHH**			**Total no. of rHH**
			Total	ASD-specific	Enriched	

	probands	controls				genes
C2	148	294	40	23	17	243
C3	289	584	99	44	55	341
C4	85	170	11	5	6	79
C5	280	560	100	48	52	417
C6	217	434	57	18	39	372
Total	**1019**	**2024**	**307**	**138**	**169**	**1452***

A summary of results for each of the ancestry-matched population clusters. The number of homozygous haplotypes over-represented (5% significance level, risk homozygous haplotypes (rHH)) in the ASD cohort is further subdivided into ASD-specific (only present in probands) and enriched (more common in probands than controls). The number of genes implicated by the rHH in each population cluster is shown in the final column. A total of 307 rHH were identified across the 5 population clusters. These regions contained or were adjacent to 1452 genes (* when genes that are found in more than one population cluster are considered only once, the final number of genes is 1243).

Table 2 Previously identified ASD candidate genes located in rHH

Pop.	No. genes identified in rHH regions	No. rHH genes previously implicated in ASD	ASD candidate genes located in rHH regions
Discovery Stage 1			
C2	243	7	**BTN2A1**, CSMD3, FNTA, **FOXP2**, GABRA2, GABRG1, **GBE1**
C3	341	12	ALAS1, FBXO33, **GBE1**, **IMMP2L**, **LRFN5**, **LRRN3**, NRXN1, PRCP, PTGS2, REEP3, **ROBO1**, **SLC4A10**
C4	79	-	-
C5	417	5	ACO2, **FOXP2**, **GRIK2**, HERC2, TSPAN12
C6	372	11	**BTN2A1**, DOCK4, **GBE1**, **GRIK2**, GRM8, **IMMP2L**, **LRFN5**, **LRRN3**, **ROBO1**, **SLC4A10**, WDR75
Replication Stage 2			
C3	529	8	ALAS1, FOXP2, GBE1, GRIK2, **GRM8**, IMMP2L, SLC4A10, STOM
C4	70	1	**CSMD3**
C5	731	9	ACO2, **CSMD3**, FBXO33, GABRG1, **GRM8**, NAGLU, PCDH10, SEPHS2, SLC6A4
C6	46	-	-

A number of genes that have previously been implicated in ASD were found to be located in rHH regions in both the discovery and replication HH mapping studies. In the discovery analysis, 25 previously reported ASD candidate genes occurred in rHH regions. Nine ASD candidate genes were located in rHH in more than one population group and are shown in bold. Another 16 ASD candidates were located in rHH in a single population group and may represent population-specific susceptibility genes. We also found that 16 previously implicated ASD genes were located in rHH regions in the replication study. Two ASD candidate genes were located in rHH in more than population one group and 14 ASD genes

were population-specific. Ten ASD candidate genes (*ALAS1*, *CSMD3*, *FOXP2*, *GABRG1*, *GBE1*, *GRM8*, *FBXO33*, *IMMP2L*, *SLC4A10* and *ACO2*) occurred in rHH regions in both the discovery and replication HH mapping analyses.

Supplementary Material 1

Supplementary Material and Methods

Patient Cohort

The samples used in the HH analysis were collected as part of an international consortium, the Autism Genome Project (AGP). Informed consent was obtained from all participants. The AGP sample set is a trio based collection, comprising an affected proband and two parents, grouped into the three distinct diagnostic classes of autism; strict, broad and spectrum. Affected individuals were diagnosed using the Autism Diagnostic Interview-Revised (ADI-R) and/or the Autism Diagnostic Observation Schedule (ADOS). To qualify for the strict class,

affected individuals met criteria for autism on both the ADI-R and the ADOS diagnostic instruments. The broad class included individuals who met ADI-R criteria for autism and ADOS criteria for ASD, but not autism, or vice versa. ADI-R-based diagnostic classification of subjects as ASD followed criteria published by Risi *et al*. 2006. Specifically, individuals who almost met ADI criteria for autism were classified as ASD if (1) they met criteria on social and either communication or repetitive behavior domains; or (2) met criteria on social and within 2 points of criteria for communication, or met criteria on communication and within 2 points of social criteria, or within 1 point on both social and communication domains (Risi et al. 2006). Finally, the spectrum class included all individuals who were classified as ASD on both the ADI-R and ADOS or who were not evaluated on one of the instruments but were diagnosed with autism on the other instrument. The HH analysis was performed on trios in the autism spectrum diagnostic category (n = 2,584 trios). The ASD spectrum trios were further subdivided into stage 1 and stage 2 collections. In the current study, the 1,402 stage 1 trios were used for the initial discovery analysis and the 1,182 stage 2 trios were used for the independent replication study.

SNP Genotyping

Samples for the discovery (stage 1) and replication (stage 2) analyses are part of the Autism Genome Project (AGP) sample collection. Stage 1 samples (AGP freezes 1-3) were genotyped using the Illumina 1M-single array while the stage 2 samples (AGP freezes 4-8) were genotyped on a combination of 1M and 1M-duo chips. The 1M platform contains a total of 1,072,820 SNPs with a mean marker spacing of 2.7 kb. Stage 2 trios were genotyped on a combination of the Illumina 1M and 1M duo arrays. The 1M duo chip contains almost 1,199,187 SNPs with a mean marker spacing of 1.5 kb. Samples were processed according to the manufacturer's recommended protocol. Bead Chips were scanned on the Illumina BeadArray Reader using the default settings. Analysis and intra-chip normalization were performed using Illumina's BeadStudio software v.3.3.7 using a GenCall cut-off of 0.1. Built-in sample independent and sample-dependent controls were inspected to assess the quality of the experiment. Genotype calling was performed according to the manufacturer's protocols and involved the use of technical controls (Peiffer et al. 2006). Given that the AGP samples were genotyped on a combination of arrays, only the 1,003,768 markers common to both platforms were considered for the HH study.

Quality Control

The AGP data set contains a number of multiplex families but the HH analysis involves only 1 affected proband per family. Therefore, the genotype data of families with more or less than 3 members was examined to determine which family members would be included in the study. We identified 37 AGP families with more than 3 members. Of these, 35 families consisted of affected sib-pairs and two unaffected parents. In each of these families the sib with the lower call rate was excluded. A single family had three children, two of which were excluded. Two trios were highly related (based on identity by state analysis) which resulted in the random exclusion of one family. One trio consisted of an affected child but parental genotype data was unavailable due to low quality. This trio was removed.

The AGP implemented quality control measures on the ASD trio data prior to data release. We enforced additional quality control parameters, outlined below, for the HH analysis. All quality control steps were

performed in PLINK. After quality control and filtering for autosomal markers, 887,716 SNPs and 7719 individuals were retained for analysis. The total genotyping rate for remaining individuals in the cleaned data set was 99.27%.

Quality Control of SNP genotype data

	Plink threshold	**# SNPs removed**	**# SNPs after QC**	**# Samples removed**	**# Samples after QC**
Missingness per SNP[1]	0.05	0	1,003,768	-	7764
Missingness per individual	0.05	-	1,003,768	45	7719
Hardy-Weinberg equilibrium	0.001	85,286	918,482	-	7719
Mendel error per SNP[1]	0.05	0	918,482	-	7719
Mendel error per family[2]	0.1	-	918,482	0	7719
Non-autosomal SNPs	-	30,736	887,716	-	7719
Data after quality control			887,716		7719

[1] Note that missingness per SNP and Mendel errors per SNP were zeroed out during the AGP quality control process

[2] 13 families with > 10,000 mendel errors were removed by the AGP

Clustering method

The AGP samples were separated into population clusters using principal component analysis, Hopach (van der Laan 2002) clustering and Fst calculations. Firstly, EIGENSOFT (Price et al. 2006) was used to obtain principal components for the 2,584 study samples (stage 1 = 1,402 samples, stage 2 = 1,182 samples). The principal component analysis (PCA) was applied to the ASD proband genotypes. To assess the population structure through PCA, we selected 70,175 autosomal SNPs with a minor allele frequency > 5% and a SNP call rate of 100%. To avoid LD-effects the SNPs were thinned using the 'indep-pairwise' option in PLINK. A window of 1,500 SNPs was selected for LD-pruning on the basis that it corresponds to the largest known high-LD region when mapped with the Hap550 Panel. The step size for LD-pruning was 150 SNPs (10% of the window size). All SNPs within the 1,500 window were required to have an $r^2 < 0.2$. SNPs located within 24 known regions of long-range LD were also removed (Price et al. 2008). The outlier removal algorithm available in EIGENSOFT (removes individuals more than 6 standard deviations from the mean) was applied to the first 5 iterations (default) and resulted in the exclusion of 150 individuals. Self-reported ancestry information was available for 43 of the excluded samples; 72% are of Asian or African origin. Inspection of SNP loadings on all axes deemed significant by the Tracy-Widom method of Patterson and colleagues revealed that no axes were dominated by single high-LD regions of the genome (Patterson et al. 2006). Tracy-Widom statistics were calculated in EIGENSOFT in order to select the number of principal components (PCs) that would be used for subsequent hierarchical clustering. The transition phase for our data occurred between PC8 and PC9 (2.0×10^{-16} to 2.5×10^{-6}). Accordingly, we have used the first eight PCs for Hopach hierarchical clustering.

Although PCA is useful to visually inspect population substructure within a sample set, it does not infer discrete population clusters or assign samples to subpopulations. The Hopach clustering algorithm (available in R) was used to define groups of similar ancestry (using the PCA results) by identifying individuals with similar eigenvalues across the first 8 PCs. The following methodology was used:

1. Run Hopach with all individuals (ASD probands) using the euclidean metric as the distance measurement and apply 20,000 bootstraps
2. Identify the most representative element of each cluster termed the medoid
3. Iteratively run Hopach with the fixed medoids from each step, eliminating in each round the medoids of clusters with the least number of members; members of eliminated clusters will fall into the next closest cluster
4. Stop when all clusters have more than 75 members
5. Run Hopach with only the medoids of the selected clusters (clusters with at least 75 members)
6. Assign every element to the medoid with the highest bootstrap value
7. Order the individuals within each cluster based on the distance to the cluster medoid
8. Calculate F_{ST} for all pair-wise cluster combinations
9. Choose one single medoid from clusters with low F_{ST}
10. Run Hopach on fixed medoids from selected clusters

Steps 1-7 yielded 19 clusters. The homogeneity of each cluster was calculated using the F_{ST} metric. F_{ST} values were calculated using 5,000 SNPs (randomly chosen from the 70,175 SNPs used in the PCA analysis) in the R package hierfstat(Goudet 2005). F_{ST} values were calculated for every pair-wise combination of clusters. Clusters with F_{ST} values < $1e^{-04}$ were collapsed and a single medoid selected from the merged group. After F_{st} analysis 10 medoids were used to perform Hopach hierarchical clustering, assigning samples to their closest medoid.

Within each of the 10 clusters the samples are ordered based on their distance to the medoid. QQ plots of 'distance from the medoid' were examined and individuals were excluded based on heuristically defined thresholds (data not shown). For each cluster we removed individuals that were highly distant from the medoid, thereby creating more homogenous clusters. During this process, 105 individuals were removed (4.3% of all individuals) and 2,333 individuals (1,151 stage 1 and 1,182 stage 2) were retained for analysis. Self-reported ancestry information was available for a subset of the samples and was used to examine the accuracy of the clustering process. Given the known genetic homogeneity of the Portuguese (Pato et al. 1997), Irish (Hill et al. 2000) and Costa Rican (Mathews et al. 2004) populations, one would expect the samples from each population to fall into the same cluster. We observed that 99.2% of individuals collected in Portugal were assigned to cluster 6; 87.5% of individuals collected in Ireland were assigned to cluster 2 and 81.1% of individuals collected in Costa Rica were assigned to cluster 9.

The PCA and Hopach analysis showed that the samples collected at Mount Sinai formed a distinct population cluster (Supplementary Fig. 1). Further investigation into the ancestry of the samples revealed that the individuals were of Costa Rican origin. All of the Mount Sinai samples are part of the AGP stage 2 collection

and therefore were not analysed in the stage 1 discovery study. However this cluster group is of particular interest as the Costa Rican population is considered an isolate with a high level of genetic homogeneity (Mathews et al. 2004). The rHH mapping was applied to the 78 Costa Rican trios and identified 20 regions with a homozygous haplotype that was significantly more common in ASD probands compared to parental controls. The candidate loci contain 113 genes including 2 previously identified ASD genes (*GRIK2* and *NAGLU*).

Identification of runs of homozygosity (ROH)

Long series of consecutive homozygous SNPS, referred to as runs of homozygosity (ROHs), were identified in each subject using the 'Runs of Homozygosity' program in PLINK (version 1.04) (http://pngu.mgh.havard.edu/~pucell/plink/contact.shtml). ROH detection was performed on the quality controlled data and only considered autosomal SNPs. A threshold of 100 consecutive homozygous SNPs spanning at least 1Mb was implemented to define an ROH. This requirement is similar to the criteria used by Nathnagel et al. 2010, Nalls et al. 2009, Jakkula et al. 2008 and Gibson et al. 2006 (Gibson et al. 2006; Jakkula et al. 2008; Nalls et al. 2009; Nothnagel et al. 2010). In addition a minimum density of 1 SNP per 50 kb was added, allowing for centromeric and SNP-poor regions to be algorithmically excluded from the analysis. The autosomal genome of each individual was scanned for ROH using a sliding window of 50 SNPs, allowing at most five missing genotypes and one heterozygote call per ROH.

Shared ROH and haplotype analysis

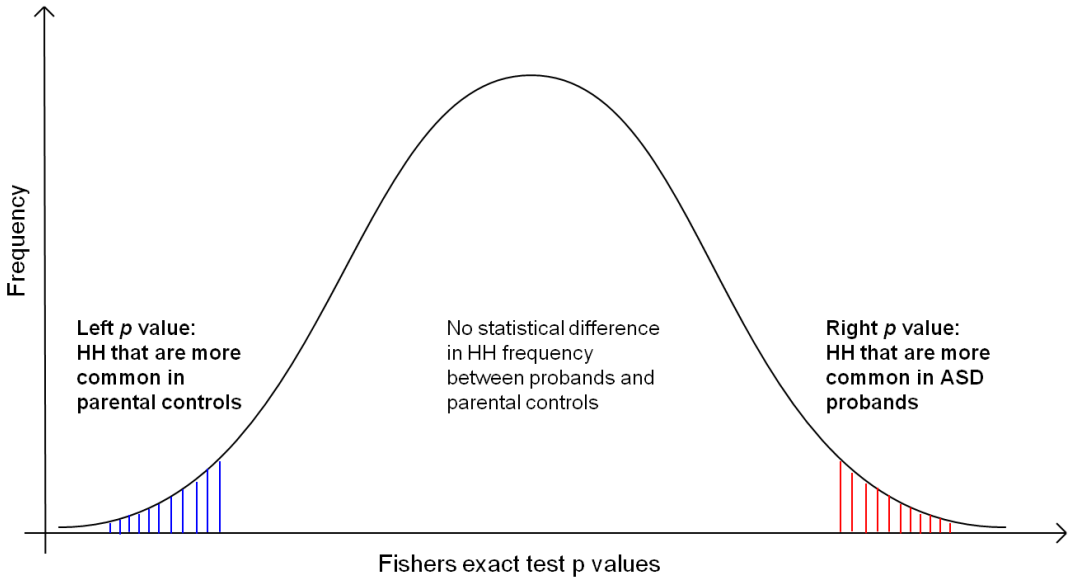
Pools of overlapping ROH segments shared by a minimum of 3 individuals were defined using the 'homozyg-group' function in PLINK. During this process, the haplotype of the overlapping ROH segments was compared. The haplotypes within the shared ROH region, termed homozygous haplotypes (HH), were declared a match if at least 95% of the homozygous sites were allelically identical. A 5% leniency threshold was allowed to ensure that long stretches of allelically matching SNPs were not broken by occasional genotyping errors. HH frequency between ASD cases and parental controls was evaluated at each overlapping ROH using the Fisher's exact test. HH with a right p value < 0.05 were considered regions of interest and termed risk homozygous haplotypes (rHH).

Comparison of rHH sharing in ASD probands and parental controls

To determine if the phenomenon of rHH sharing was equally likely to occur in a non-disease cohort, we compared the number of rHH regions identified in probands and parental controls as outlined below. The figure illustrates the regions under comparison.

1. Within each overlapping ROH segment identify the homozygous haplotypes (HH) of the individuals sharing the ROH segment
2. Identify HH's that are more common (Fisher's exact test right p value < 0.05) in ASD probands compared to parental controls. These regions are termed risk homozygous haplotypes (rHH's)
3. Count the number of rHH regions identified in ASD probands (rHH-probands)
4. Identify HH's that are significantly more common (Fisher's exact test left p value < 0.05) in parental controls compared to ASD probands. These are termed rHH-parental

5. Count the number of rHH-parental regions
6. To account for differences in sample size (the number of cases and parental controls is not equal within each population cluster and sample sizes also differ across population clusters) the raw counts were 'normalised' for a sample size of 100 individuals. The raw counts are also provided below.
7. Use a paired t-test to test determine if the number of rHH-probands and rHH-parental differ significantly across all population clusters.



Comparison of rHH in ASD parents and parental controls. HH with a right *p* value < 0.05 are more common in ASD cases compared to parental controls and are termed rHH-probands. HH with a left *p* value < 0.05 are more common in parental controls compared to ASD cases and are termed rHH-parental. Under a normal distribution one would expect the number of rHH-probands to equal the number of rHH-parental.

The number of rHH in ASD probands and parental controls

Population cluster	No. of ASD cases in the population cluster	No. of parental controls in the population cluster	No. rHH-probands		No. rHH-parental	
			RC	NC	RC	NC
C2	148	294	40	27.03	16	5.44
C3	289	584	113	39.10	55	9.42
C4	85	170	12	14.12	13	7.64
C5	280	560	108	38.57	38	6.78
C6	217	434	59	27.19	30	6.91

The size of each population cluster and the number of rHH in probands and parental controls is provided. RC and NC refer to raw counts and normalized counts (to a sample size of 100) respectively.

Inspection of LD structure

Patterns of linkage disequilibrium (LD) within the 307 rHH were visualised and analysed in Haploview using HapMap CEU as a reference (Barrett et al. 2005). LD was measured as r^2 values and calculated between each pair of SNPs. The *Tagger* algorithm (Haploview) was used to determine the number of tagging SNPs within each rHH at an r^2 threshold of 0.8. Genes located in rHH comprising <10 tagging SNPs were noted.

Copy number variation

Similar to previous studies, we used robust criteria for defining runs of homozygosity as a means of reducing false-positives due to copy number variation (Gibson et al. 2006; Jakkula et al. 2008; Nalls et al. 2009; Nothnagel et al. 2010). Furthermore, copy number variants (CNVs) were identified in the AGP stage 1 data set using iPattern and QuantiSNP. The algorithms infer putative CNVs from log R ratio and B allele frequency. For each population cluster in the discovery analysis (stage 1), the samples contributing to significant rHH were inspected for CNV content as detected by Pinto and colleagues (Pinto et al. 2010). In circumstances where, for a particular individual, a region called as homozygous was deemed to be a CNV by Pinto and colleagues, the individual in question was excluded prior to applying the Fisher's exact test. None of the rHHs were called as CNVs in the individuals of interest (ie the specific individuals carrying the rHH). It is possible that some of the rHHs may be hemizygous deletions that were not called as CNVs using the Illumina platform and Quanti-SNP/iPattern algorithms. However, a hemizygous deletion exposing a recessive rare variant would have a similar biological and phenotypic affect as being homozygous for the rare variant. Although CNVs were not examined in the four population clusters used for the replication analysis, the contribution of CNVs had no effect in the discovery phase. Whether the individuals in question have a hemizygous deletion or are homozygous for a recessive variant, the region/patients still warrant investigation.

Replication study

A replication study was undertaken using the AGP follow up stage 2 data set (freezes 5-8) which comprises 1,182 ASD trios genotyped on a combination of the Illumina 1M and 1M-duo platforms. The stage 2 data was cleaned with the stage 1 samples to ensure that the same markers would be used in both analyses. The stage 1 and stage 2 probands were clustered to identify ancestry-matched replication groups. The 4 population clusters (C3-C6) with ≥50 probands in both stage 1 and 2 analyses were considered for the replication study. The same HH mapping method was applied to the four stage 2 population clusters. The rHH genes identified in the discovery (stage 1) and replication (stage 2) analyses were then compared to identify overlap.

Wellcome Trust Case Control Cohorts

To assess the potential contribution of genomic architecture to the HH analyses, rHH mapping was performed in two additional disease datasets available from the Wellcome Trust Case Control Consortium; bipolar disorder (n = 1,998) and coronary artery disease (n = 1,988). The 1958 British birth cohort (n = 1,504) and National Blood Service control cohort (n = 1,500) were used as controls. Further information for these data sets is provided by Burton and colleagues (The Wellcome Trust Case Control Consortium 2007). The data can be accessed through (http://www.ebi.ac.uk/ega/page.php?page=studies&name=WTCCC). The cases and controls were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip), which comprises 500,568 SNPs. After

implementing the same quality control criteria used for the autism HH mapping study, 478,224 SNPs were retained for analysis. ROH were identified using the same parameters applied to the AGP data.

References for Supplementary Material 1

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-5

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15: 789-95

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5: 184-186

Hill EW, Jobling MA, Bradley DG (2000) Y-chromosome variation and Irish origins. Nature 404: 351-2

Jakkula E, Rehnstrom K, Varilo T, Pietilainen OP, Paunio T, Pedersen NL, deFaire U, Jarvelin MR, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L (2008) The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83: 787-94

Mathews CA, Reus VI, Bejarano J, Escamilla MA, Fournier E, Herrera LD, Lowe TL, McInnes LA, Molina J, Ophoff RA, Raventos H, Sandkuijl LA, Service SK, Spesny M, Leon PE, Freimer NB (2004) Genetic studies of neuropsychiatric disorders in Costa Rica: a model for the use of isolated populations. Psychiatr Genet 14: 13-23

Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB (2009) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 5: e1000415

Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19: 2927-2935

Pato CN, Azevedo MH, Pato MT, Kennedy JL, Coelho I, Dourado A, Macedo A, Valente J, Ferreira CP, Madeira J, Gago da Camara J, Moniz M, Correia C (1997) Selection of homogeneous populations for genetic study: the Portugal genetics of psychosis project. Am J Med Genet 74: 286-8

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-48

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368-72

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-9

Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D (2008) Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 83: 132-5; author reply 135-9

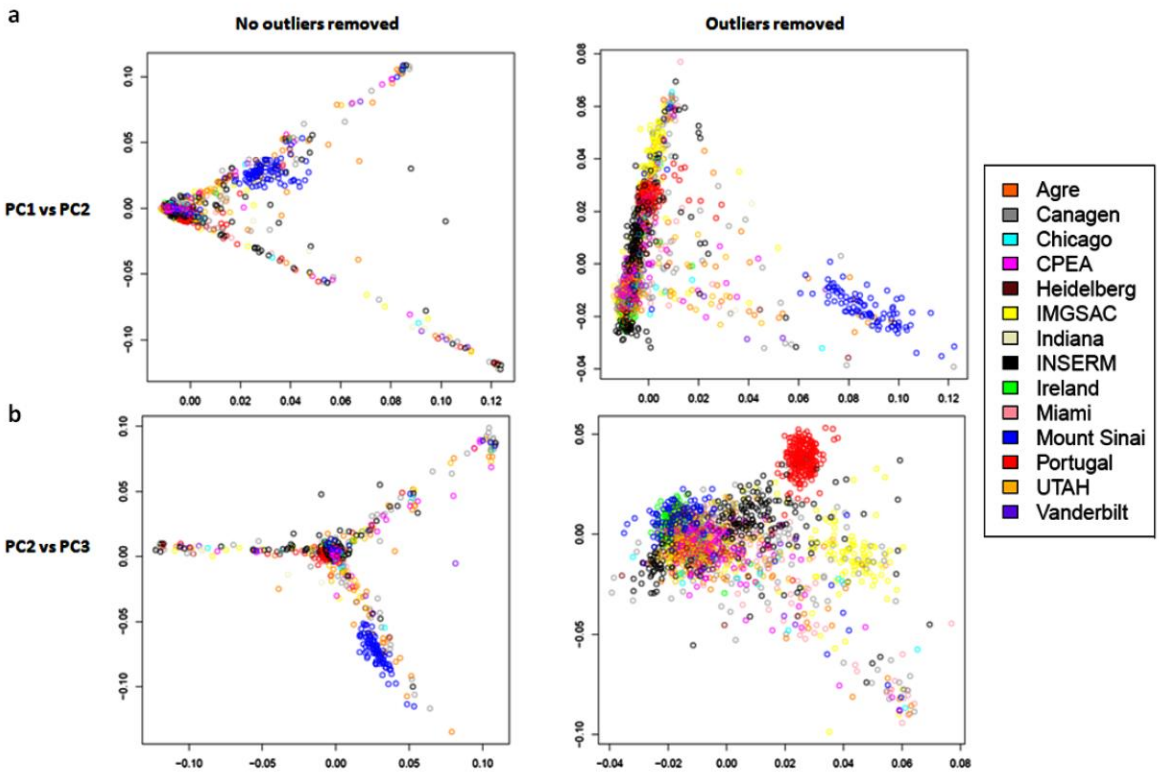
Risi S, Lord C, Gotham K, Corsello C, Chrysler C, Szatmari P, Cook EH, Jr., Leventhal BL, Pickles A (2006) Combining information from multiple sources in the diagnosis of autism spectrum disorders. J Am Acad Child Adolesc Psychiatry 45: 1094-103

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-78

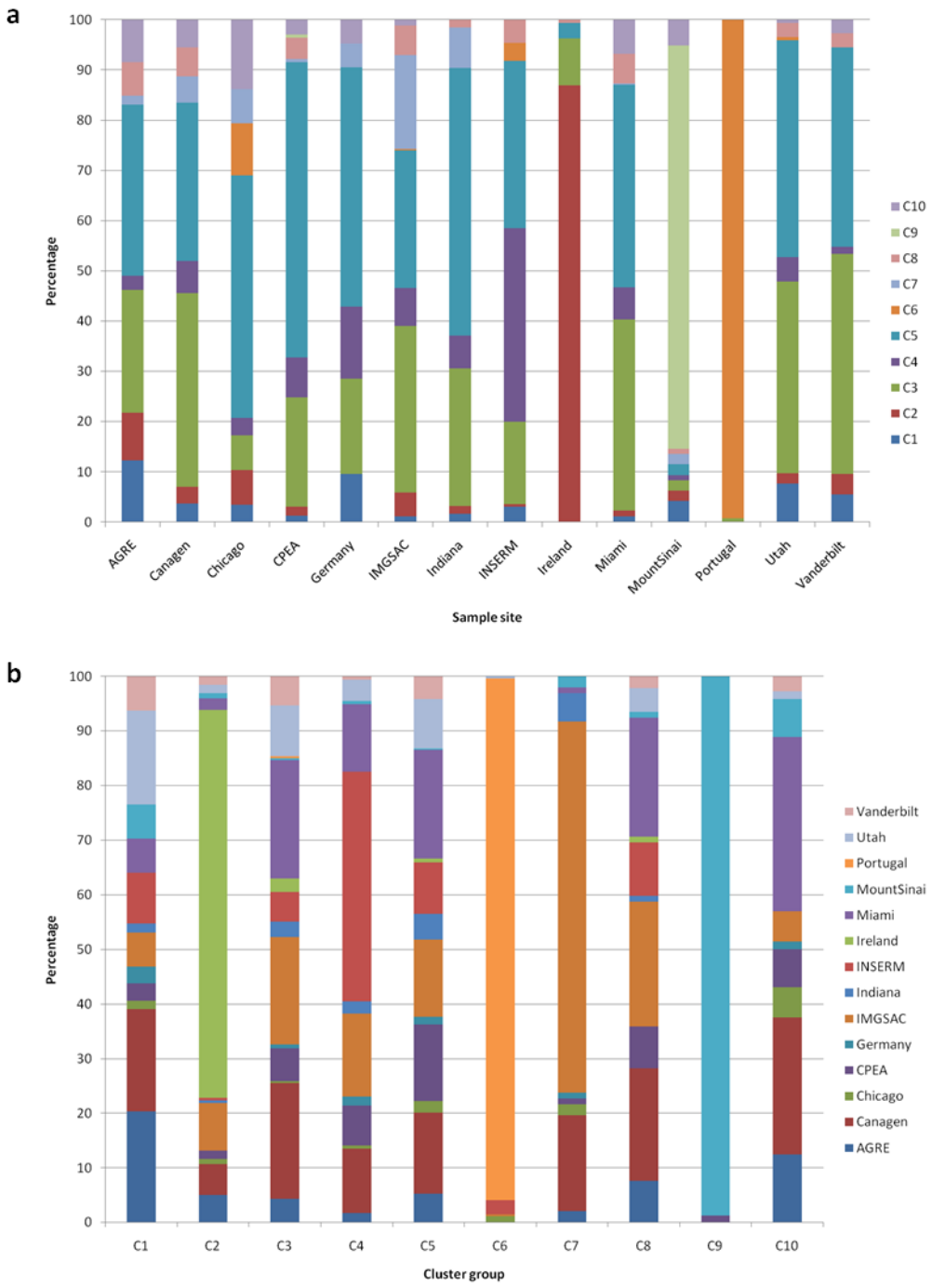
van der Laan MJaPKS (2002) A new algorithm for hybrid hierarchical clustering with visualisation and the bootstrap. J Stat Planning and Inference 117: 275-303

Supplementary Material 2

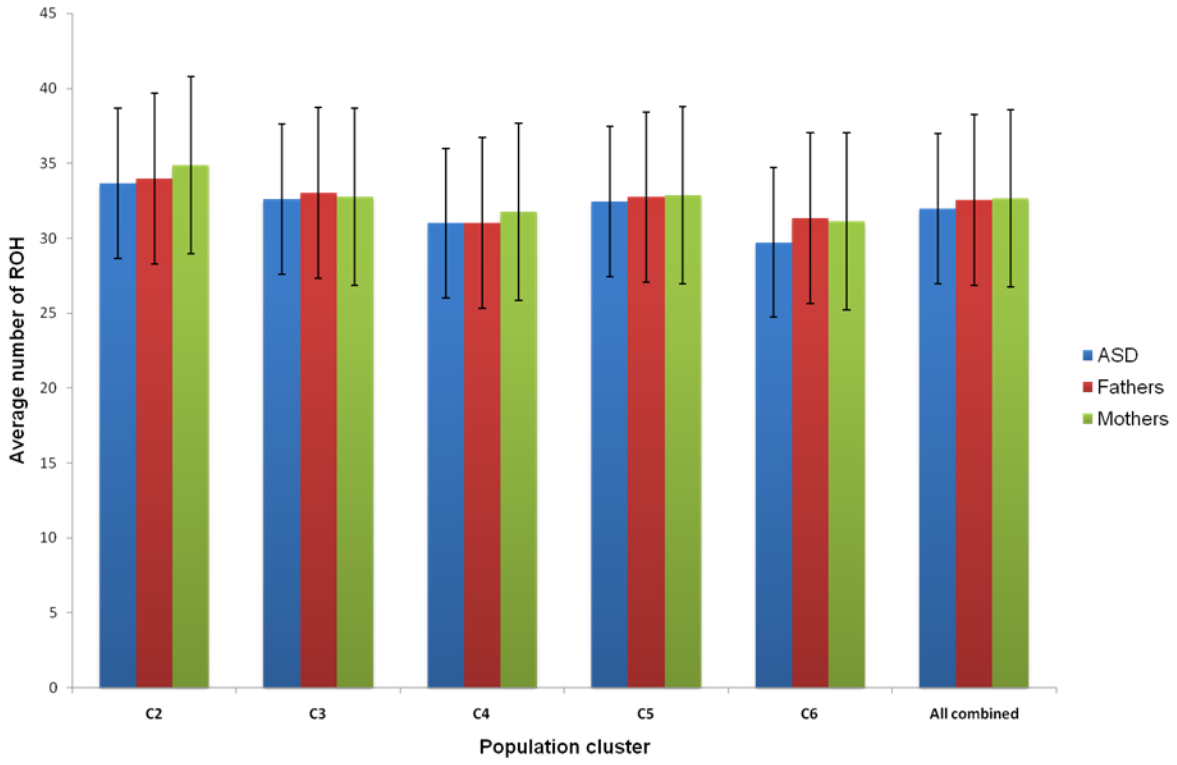
Supplementary Figures



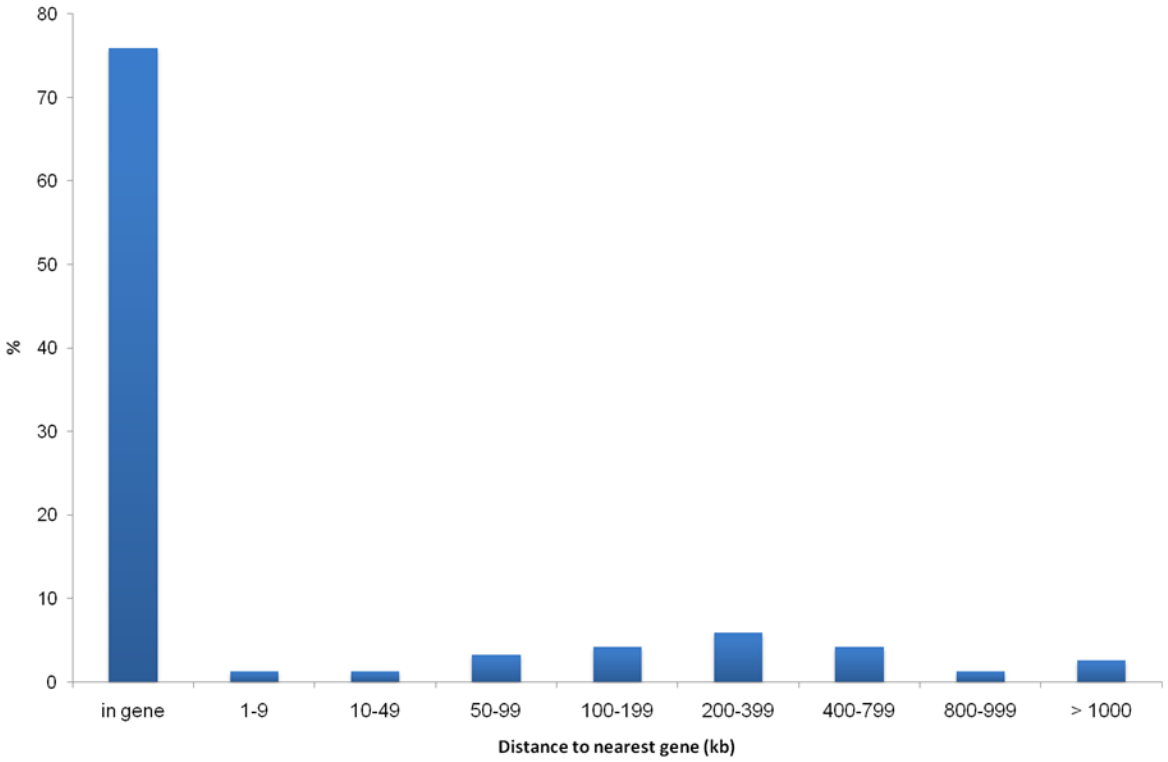
Supplementary Fig. 1 Principal component analysis of AGP stage 1 and stage 2 ASD probands. Principal component analysis of the 2,548 ASD probands (stage 1 = 1,402 and stage 2 = 1,182 samples) was performed in EIGENSOFT using 70,175 SNPs. The analysis was performed without and with the default outlier removal algorithm available in EIGENSOFT. Row **a** represents PC1 plotted against PC2, without and with outlier removal. When outliers are allowed to remain in the sample set, the resulting genetic map forms a triangular structure with samples of self-reported African and Asian origin at the extremities. Removal of outliers generates a more linear map with the Mount Sinai samples of Costa Rican origin (blue) forming a distinct group. The samples collected at Mount Sinai are of individuals from Cost Rica. Row **b** represents PC2 plotted against PC3, without and with outlier removal. When outliers are removed, the second and third principal components show clear separation of the predominantly Portuguese samples (red) from the remainder of the sample set.



Supplementary Fig. 2 Classification of each AGP site by cluster membership and site-composition of each population cluster. A: The x-axis refers to the 14 AGP sample collection sites. Each colour denotes one of the 10 population clusters to which samples can be assigned. The percentage of samples assigned to each population cluster is illustrated on the y-axis. The analysis showed that samples collected at the Portuguese, Irish and Mount Sinai sites had the greatest genetic homogeneity. Samples from the remaining AGP sites have more varied ancestral backgrounds. B: The x-axis represents each of the 10 population clusters. Each colour defines the site from which the sample was collected.



Supplementary Fig. 3 Comparison of genome-wide ROH burden in ASD probands and parental controls. Each of the 5 population clusters (C2-C6) are denoted on the x-axis. In addition, all of the AGP trios were combined into a single group for statistical analysis. The y-axis represents the number of autosomal ROH (≥ 1Mb) identified per individual. With the exception of cluster 6 (C6), there was no statistical difference in ROH number between probands (red), fathers (blue) and mothers (green) in each of the individual population groups after correcting for multiple testing (paired t-test). In the combined analysis, parental controls had a significantly higher number of homozygous regions (father p = 0.012, mothers p = 0.006) compared to their off-spring (fathers = 32.56 ROH, mothers = 32.65 ROH, ASD probands = 31.98 ROH). However it should be noted that the higher ROH number observed in parental controls in the combined analysis stems primarily from C6, which was the only group to show significant differences in the individual population group analysis. The two-tail p values were calculated using a paired t-test. Error bars, ±1 standard deviation (s.d.).



Supplementary Fig. 4 Percentage of rHH located in genic and intergenic regions. The genomic positions of the 307 rHH regions identified in the discovery analysis were visually inspected in the UCSC genome browser (build hg18). In circumstances where a rHH was located in an intergenic region, the distance to the nearest RefSeq gene was calculated. The x-axis denotes the distance (in kb) of the rHH to the nearest gene. The y-axis represents the percentage of rHH regions that are within a specified distance to the nearest gene. We found that 75.9% of rHH are located in or directly overlap genes.

Supplementary Material 3

Table 1a Stage 1 Cluster 2

Chr	Start	End	Length (kb)	Distance to gene	ASD probands	Parental controls	RPV	Genes
6	49,098,263	49,613,996	515.73	0	5	0	0.004022	MUT C6orf139 C6orf141
12	109,807,801	109,841,394	33.59	0	6	1	0.006641	CCDC63 MYL2
2	56,476,501	56,903,511	427.01	9.69	4	0	0.012233	EFEMP1
2	186,573,851	187,698,829	1124.98	0	4	0	0.012233	ZSWIM2 FLJ44048 ITGAV LEREPO4 CALCRL KIAA1946
5	45,293,214	45,605,533	312.32	0	4	0	0.012233	HCN1
7	113,495,023	113,560,041	65.02	0	5	1	0.01755	FOXP2
2	81,832,890	81,865,855	32.97	1103.39	7	3	0.019011	CTNNA2
6	27,406,258	27,991,536	585.28	0	7	3	0.019011	ZNF184 HIST1H2BL ZNF204 LOC441136 FKSG83 HIST1H4J HIST1H4L HIST1H2AL HIST1H1B HIST1H2BN HIST1H2BO HIST1H4K OR2B2 HIST1H2AM HIST1H3J HIST1H2AK HIST1H2AI HIST1H2BM HIST1H3H HIST1H2AJ
3	161,899,602	162,422,511	522.91	0	6	2	0.019193	PPM1L NMD3 B3GALT3 ARL14
12	109,955,099	109,966,246	11.15	0	19	20	0.028634	CUTL2
6	57,437,845	58,412,859	975.01	0	9	6	0.029441	PRIM2A GUSBL2
8	42,440,767	43,371,581	930.81	0	9	6	0.029441	CHRNB3 POTE8 RNF170 C8orf40 CHRNA6 FNTA HOOK3 THAP1 SLC20A2 FLJ23356
8	112,673,262	112,788,744	115.48	515.59	8	5	0.033465	CSMD3
1	189,400,412	189,696,127	295.72	1066.99	3	0	0.037036	RGS18 FAM5C
2	188,367,601	188,777,903	410.30	0	3	0	0.037036	TFPI GULP1
3	82,530,472	82,849,478	319.01	636.83	3	0	0.037036	GBE1
3	111,574,317	111,976,626	402.31	296.93	3	0	0.037036	PVRL3
4	45,651,745	46,035,520	383.78	0	3	0	0.037036	GABRG1 GABRA2
5	44,763,463	45,605,533	842.07	0	3	0	0.037036	HCN1 LOC441070 MRPS30
5	87,250,540	87,351,008	100.47	175.77	3	0	0.037036	MGC33214
5	109,630,568	110,510,389	879.82	0	3	0	0.037036	TSLP LOC91137 WDR36 MAN2A1
7	86,394,765	86,583,884	189.12	0	3	0	0.037036	KIAA1324L DMTF1
7	90,498,620	90,512,128	13.51	0	3	0	0.037036	PFTK1
7	112,499,405	112,887,943	388.54	0	3	0	0.037036	GPR85
8	6,786,975	8,059,517	1272.54	0	3	0	0.037036	DEFA1 PJCG6 DEFB109 DEFB103A DEFB105B DEFB4 DEFA5 LOC401447 LOC349196 DEFA3 DEFB105A DEFA4 DEFB106A SPAG11 DEFB104A

8	85,256,860	85,270,500	13.64	0	3	0	0.037036	RALYL
10	92,917,247	94,207,391	1290.14	0	3	0	0.037036	HECTD2 TNKS2 C10orf13 CPEB3 39146 IDE PCGF5 PPP1R3C MARCH5 BTAF1
10	104,183,824	104,203,492	19.67	0	3	0	0.037036	C10orf95 CUEDC2
11	28,973,078	29,091,959	118.88	661.45	3	0	0.037036	METT5D1
11	65,826,722	66,922,336	1095.61	0	3	0	0.037036	PELI3 FBXL11 B3GNT6 SYT12 PC DPP3 ADRBK1 BRMS1 MRPL11 RBM14 SLC29A2 NPAS4 FLJ22531 MGC33486 ACTN3 SPTBN2 RHOD ANKRD13D RAD9A FLJ10786 BBS1 ZDHHC24 CCS MGC15912 RBM4 RBM4B POLD4 CLCF1 RCE1 RIN1 SSH3 LRFN4 CTSF CD248
12	86,933,325	86,953,161	19.84	0	3	0	0.037036	C12orf50
15	27,762,156	28,736,917	974.76	0	3	0	0.037036	TJP1 CHRFAM7A FAM7A2 KIAA1018
12	109,842,617	110,282,119	439.50	0	6	3	0.041714	MYL2 CUTL2 FAM109A
5	87,114,798	87,541,020	426.22	0	5	2	0.044824	MGC33214 CCNH
6	25,792,384	27,479,172	1686.79	0	5	2	0.044824	HIST1H3D BTN3A2 ZNF322A HIST1H2BK GUSBL1 HIST1H4G HIST1H2BD SCGN ZNF204 HIST1H2AH HIST1H1C HIST1H4D HFE HMGN4 SLC17A2 HIST1H1A HIST1H3B HIST1H1D HIST1H4H BTN3A3 SLC17A1 BTN1A1 HIST1H2BF BTN3A1 LOC441136 BTN2A3 PRSS16 TRIM38 SLC17A4 HIST1H2BC HIST1H2AC FKSG83 SLC17A3 HIST1H2BG HIST1H2BJ HIST1H2BB HIST1H2AE HIST1H2BH HIST1H3C HIST1H2AG HIST1H4E HIST1H3G HIST1H1E BTN2A2 ABT1 BTN2A1 HIST1H2BE HIST1H1T HIST1H2AB HIST1H4C HIST1H2BA HIST1H2AA HIST1H3F HIST1H3E HIST1H2AD HIST1H3A HIST1H4A HIST1H4F HIST1H4B HIST1H4I HIST1H2BI
7	85,409,655	85,971,962	562.31	139.20	5	2	0.044824	GRM3
3	47,894,539	49,109,725	1215.19	0	4	1	0.045077	MAP4 NME6 PRKAR2A ARIH2 COL7A1 QRICH1 CAMP UQCRC1 ZNF589 SCOTIN PLXNB1 PFKFB4 IMPDH2 SLC25A20 IHPK2 FBXW12 WDR6 NCKIPSD CELSR3 C3orf60 PH-4 TREX1 SLC26A6 CCDC51 CDC25A UCN2 DALRD3 CCDC72 TMEM89 QARS
3	84,712,995	84,803,672	90.68	287.15	4	1	0.045077	CADM2
6	28,023,066	28,407,274	384.21	0	14	14	0.046537	ZNF307 OR2B6 ZNF165 PGBD1 ZNF192 ZNF193 ZNF187 C6orf194 ZNF435 ZNF323
1	50,375,082	50,449,091	74.01	0	9	7	0.048188	ELAVL4

List of rHHs for discovery stage 1 population cluster 2. The number of ASD probands and parental controls is 148 and 294 respectively.

Table 1b Stage 1 Cluster 3

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
19	21,902,059	22,229,909	327.85	0	9	2	0.001213	ZNF208 ZNF257 ZNF676
3	81,361,789	81,497,220	135.43	124.32	7	1	0.002373	GBE1
19	22,345,947	22,439,127	93.18	0	7	1	0.002373	ZNF676
6	63,763,803	64,019,651	255.85	24.16	8	2	0.00308	GLULD1
1	192,278,312	193,031,949	753.64	1429.59	5	0	0.003884	CDC73 KCNT2
12	86,455,427	86,685,143	229.72	212.80	5	0	0.003884	C12orf50
19	21,743,257	22,229,909	486.65	0	6	1	0.006407	ZNF100 ZNF208 ZNF257 ZNF43 ZNF676
8	33,792,698	34,651,643	858.95	224.31	11	6	0.007043	DUSP26 UNC5D
1	71,811,511	72,227,919	416.41	0	7	2	0.007652	NEGR1
5	43,307,304	43,430,535	123.23	0	7	2	0.007652	HMGCS1 CCL28 MGC42105
1	141,500,438	144,800,611	3300.17	0	4	0	0.011843	NBPF20 HFE2 TXNIP RBM8A POLR3C NOTCH2NL NBPF11 PDZK1 CD160 ZNF364 LOC401131 LIX1L PIAS3 PDE4DIP ANKRD35 PEX11B NUDT17 ITGA10 SEC22L1 NUDT4P1 POLR3GL GNRHR2
4	150,853,641	150,879,725	26.08	339.81	4	0	0.011843	DCAMKL2
5	42,908,813	43,908,839	1000.03	0	4	0	0.011843	SEPP1 ZNF131 PAIP1 NNT FLJ32363 HMGCS1 FLJ21657 FLJ10246 MGC42105 LOC153684 CCL28 LOC389289
8	63,683,668	63,746,387	62.72	0	4	0	0.011843	FAM77D
11	82,200,997	82,514,027	313.03	0	4	0	0.011843	RAB30 PRCP FLJ25416 LOC143543 PCF11
12	48,856,172	49,869,773	1013.60	0	4	0	0.011843	LARP4 TFCP2 LIMA1 DIP2B SLC11A2 ATF1 TMPRSS12 METTL7A POU6F1 C12orf22 LETMD1
8	51,548,704	52,431,945	883.24	0	15	12	0.012232	SNTG1 PXDNL
6	63,048,965	63,218,749	169.78	0	11	7	0.012833	KHDRBS2
5	43,777,059	44,417,455	640.40	0	10	6	0.014358	NNT FGF10
11	48,012,023	48,093,566	81.54	0	98	155	0.015111	PTPRJ
2	136,752,259	136,915,796	163.54	160.06	18	17	0.017029	CXCR4
1	192,320,984	192,829,261	508.28	830.42	8	4	0.017291	CDC73
6	127,274,638	127,399,975	125.34	0	7	3	0.018321	RSPO3
1	144,437,974	144,970,329	532.36	0	6	2	0.018509	NBPF11 PDZK1 PRKAB2
2	161,599,628	162,501,529	901.90	0	6	2	0.018509	PSMD14 TBR1 SLC4A10 TANK
2	175,420,188	176,187,475	767.29	0	6	2	0.018509	ATP5G3 CHN1 KIAA1715 ATF2
6	57,332,078	58,887,738	1555.66	0	6	2	0.018509	PRIM2A GUSBL2

8	33,792,698	34,219,082	426.38	215.72	11	8	0.021624	DUSP26
5	43,419,259	44,277,240	857.98	0	9	6	0.02835	PAIP1 NNT FLJ32363 FLJ21657 FGF10 CCL28
12	108,836,744	109,733,160	896.42	0	22	25	0.031307	ARPC3 TA-PP2C FLJ21127 FLJ40142 ATPBD1C IFT81 MGC15619 GIT2 PPP1CC RAD9B TCHP ANKRD13 VPS29 ATP2A2 ANAPC7 CCDC63 C12orf24
5	44,417,455	44,427,899	10.44	0	8	5	0.032215	FGF10
22	29,863,843	30,495,508	631.67	0	8	5	0.032215	DEPDC5 MGC50372 SFI1 MGC17330 PISD EIF4ENIF1 LIMK2 DRG1 ZNF278 RNF185 PLA2G3
1	48,983,157	50,363,319	1380.16	0	3	0	0.036027	AGBL4 ELAVL4 C1orf165
1	183,860,062	185,089,587	1229.53	0	3	0	0.036027	PTGS2 C1orf27 HMCN1 PLA2G4A TPR PDC OCLM PRG4
1	194,578,136	195,601,646	1023.51	0	3	0	0.036027	CFHR5 ASPM ZBTB41 CFH CFHR1 KCNT2 CRB1 CFHR2 F13B CFHR3
2	50,128,278	50,416,464	288.19	0	3	0	0.036027	NRXN1
2	73,354,961	73,411,911	56.95	0	3	0	0.036027	EGR4
3	58,197,472	58,244,752	47.28	0	3	0	0.036027	ABHD6
3	79,096,568	79,799,230	702.66	0	3	0	0.036027	ROBO1
3	82,652,949	83,121,967	469.02	759.31	3	0	0.036027	GBE1
3	86,250,563	87,187,589	937.03	0	3	0	0.036027	VGLL3 CADM2
3	88,792,595	88,876,114	83.52	363.25	3	0	0.036027	EPHA3
4	47,179,868	47,186,827	6.96	0	3	0	0.036027	ATP10D
4	47,278,803	47,447,311	168.51	0	3	0	0.036027	CORIN ATP10D
4	53,115,929	53,453,028	337.10	0	3	0	0.036027	FLJ12684 USP46 RASL11B SCFD2
4	74,336,605	74,995,082	658.48	0	3	0	0.036027	RASSF6 AFP IL8 ALB CXCL1 ANKRD17 PF4V1 CXCL6 AFM
4	87,403,190	88,286,528	883.34	0	3	0	0.036027	PTPN13 AFF1 MAPK10 SLC10A6 MGC26744
4	113,135,436	113,462,714	327.28	0	3	0	0.036027	FLJ39370 C4orf16 TIFA
4	150,097,597	150,879,725	782.13	339.81	3	0	0.036027	NR3C2 DCAMKL2
6	64,406,400	64,690,688	284.29	0	3	0	0.036027	PHF3
6	87,117,509	87,555,414	437.91	148.33	3	0	0.036027	HTR1E
6	87,158,118	88,498,382	1340.26	0	3	0	0.036027	HTR1E CGA C6orf163 C6orf162 GJB7 ORC3L SLC35A1 C6orf166 RARSL C6orf165
7	117,507,112	118,050,616	543.50	0	3	0	0.036027	ANKRD7 LSM8
8	63,683,668	64,263,375	579.71	0	3	0	0.036027	FAM77D YTHDF3 TTPA GGH
8	89,257,570	90,417,220	1159.65	0	3	0	0.036027	MMP16 RIPK2
8	90,280,129	90,435,943	155.81	403.17	3	0	0.036027	RIPK2

10	111,325,760	112,024,436	698.68	0	3	0	0.036027	MXI1 XPNPEP1 ADD3
11	67,166,650	67,168,406	1.76	0	3	0	0.036027	ACY3
11	82,200,997	83,233,715	1032.72	0	3	0	0.036027	PCF11 DLG2 RAB30 PRCP MDS025 FLJ25416 LOC143543 ANKRD42 FLJ37266
11	88,709,382	88,732,304	22.92	0	3	0	0.036027	NOX4
11	89,713,776	89,761,523	47.75	117.60	3	0	0.036027	CHORDC1
12	37,905,772	38,468,873	563.10	0	3	0	0.036027	C12orf40 KIF21A ABCD2 SLC2A13
12	86,266,753	86,662,843	396.09	235.10	3	0	0.036027	C12orf50 MGAT4C
13	67,853,476	67,988,075	134.60	1151.01	3	0	0.036027	PCDH9 KLHL1
13	95,178,898	96,219,346	1040.45	0	3	0	0.036027	UGCGL2 HSP90AB6P DNAJC3
14	39,410,936	39,678,075	267.14	439.57	3	0	0.036027	FBXO33
14	40,366,328	40,601,146	234.82	545.95	3	0	0.036027	LRFN5
15	39,680,412	39,957,870	277.46	0	3	0	0.036027	PLA2G4B MAPKBP1 TYRO3 SPTBN5
10	38,144,493	39,137,918	993.43	0	33	44	0.039453	LOC399744 ZNF25 ZNF37A ZNF248 LOC158160 ZNF33A
6	127,393,971	127,651,788	257.82	0	6	3	0.040223	RNF146 RSPO3
14	38,446,110	39,136,999	690.89	0	6	3	0.040223	SEC23A CTAGE5 TRAPPC6B SIP1 FBXO33 PNN MIA2
16	31,271,994	33,900,037	2628.04	0	6	3	0.040223	ERAF ZNF720 TP53TG3 FLJ43855 FLJ46121 ZNF267 LOC441762 ITGAX ARMC5 SLC5A2 C16orf58 COX6A2 TGFB1I1 LOC440366 MGC3020 ITGAD HERC2P4
18	50,218,771	50,619,139	400.37	0	6	3	0.040223	C18orf26 RAB27B C18orf54
2	137,348,531	138,151,484	802.95	0	5	2	0.04331	HNMT CXCR4
3	51,916,683	52,778,629	861.95	0	5	2	0.04331	TMEM113 DNAH1 STAB1 NISCH DUSP7 ALAS1 PTK9L PB1 NEK4 SEMA3G GNL3 TNNC1 GLT8D1 PHF7 ABHD14A GLYCTK WDR51A PPM1M PARP3 GPR62 PCBP4 BAP1 RPL29 TLR9 ABHD14B IQCF1 ACY1 RNU3IP2 SPCS1 NT5DC2
8	50,243,165	50,863,839	620.67	91.97	5	2	0.04331	C8orf22 SNTG1
10	65,113,760	65,144,918	31.16	58.87	5	2	0.04331	REEP3
12	84,317,328	84,993,929	676.60	0	5	2	0.04331	MGAT4C PAMCI CART1 NTS
13	82,134,429	82,320,867	186.44	1028.48	5	2	0.04331	SLITRK1
4	65,142,376	65,686,929	544.55	0	4	1	0.04368	SRD5A2L2 EPHA5
4	170,894,603	171,301,088	406.49	0	4	1	0.04368	FLJ20534 AADAT MFAP3L
5	59,605,280	59,771,622	166.34	0	4	1	0.04368	PDE4D
5	60,920,937	60,943,558	22.62	25.84	4	1	0.04368	FLJ37543
5	87,978,624	87,991,694	13.07	0	4	1	0.04368	MEF2C

6	33,076,576	33,614,045	537.47	0	4	1	0.04368	KIFC1 HLA-DPB1 COL11A2 VPS52 HLA-DPA1 HLA-DOA SLC39A7 RXRB DAXX TAPBP ZNF297 RING1 SYNGAP1 HSD17B8 HLA-DPB2 ZBTB9 BAK1 WDR46 RGL2 CUTA RPS18 PHF1 B3GALT4 PFDN6
6	86,901,777	87,122,490	220.71	0	4	1	0.04368	SYNCRIP HTR1E
6	126,487,654	127,399,975	912.32	0	4	1	0.04368	RSPO3 C6orf173 C6orf75
7	110,268,060	110,620,072	352.01	0	4	1	0.04368	IMMP2L LRRN3
8	49,752,138	49,995,542	243.40	0	4	1	0.04368	EFCAB1 SNAI2
8	76,867,960	77,431,275	563.32	226.34	4	1	0.04368	HNF4G ZFHX4
8	78,178,550	78,265,972	87.42	102.72	4	1	0.04368	PXMP3
12	85,870,225	86,230,340	360.12	113.41	4	1	0.04368	MGAT4C
15	80,196,915	81,291,105	1094.19	0	4	1	0.04368	DKFZp666G057 FLJ22795 CPEB1 AP3B2 FSD2 EFTUD1 RPS17 LOC440295
17	47,711,626	48,342,881	631.26	119.47	4	1	0.04368	CA10
12	37,239,340	37,409,891	170.55	0	57	87	0.044666	CPNE8
3	17,320,301	18,154,991	834.69	0	9	7	0.04634	TBC1D5 SATB1
4	8,979,912	9,719,703	739.79	0	9	7	0.04634	SLC2A9 DUB4 DRD5 WDR1
6	64,156,969	64,298,475	141.51	67.13	9	7	0.04634	GLULD1 PTP4A1
7	101,901,824	102,206,948	305.12	0	19	22	0.049348	POLR2J POLR2J3 RASA4 POLR2J2 MGC119295 MGC35361

List of rHHs for discovery stage 1 population cluster 3. The number of ASD probands and parental controls is 289 and 584 respectively.

Table 1c Stage 1 Cluster 4

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
15	41,232,597	41,328,431	95.83	0	7	3	0.017672	EPB42 TGM5 CCNDBP1 TMEM62
1	144,435,002	144,437,974	2.97	1.11	3	0	0.036164	PDZK1
4	68,897,629	69,950,973	1053.34	0	3	0	0.036164	YTHDC1 TMPRSS11E UGT2B17 UGT2B15 UGT2A3 UGT2B10 UGT2B7
14	59,580,228	60,662,379	1082.15	0	3	0	0.036164	C14orf39 SIX1 MNAT1 TRMT5 SLC38A6 PPM1A SIX6 SIX4 DHRS7 C14orf135 FLJ46156
14	70,413,084	70,877,171	464.09	0	3	0	0.036164	PCNX RNU56B
15	40,705,374	41,328,431	623.06	0	3	0	0.036164	TTBK2 UBR1 EPB42 CDAN1 TGM5 CCNDBP1 TMEM62 CEP27
17	40,905,309	42,021,315	1116.01	0	10	8	0.037651	CRHR1 LRRC37A LOC641522 ARL17P1 LOC474170 KIAA1267 C17orf69 MAPT STH PLEKHM1 IMP5
2	95,754,597	96,143,069	388.47	0	5	2	0.043	LOC400986 LOC150763 ADRA2B TRIM43
1	149,434,910	149,777,892	342.98	0	4	1	0.043649	PIP5K1A CGN PIK4CB POGZ PSMD4 ZNF687 SELENBP1 RFX5 PSMB4
7	64,726,242	65,907,667	1181.43	0	9	7	0.044513	VKORC1L1 ASL TPST1 LOC285908 KCTD7 RABGEF1 LOC441242 GUSB RCP9
1	146,292,078	148,168,936	1876.86	0	15	16	0.047417	NBPF1 FLJ39739 NBPF15 FCGR1A HIST2H2AA MTMR11 HIST2H2BF BOLA1 SF3B4 LOC388692 NBPF14 SV2A HIST2H4 HIST2H2AB HIST2H2BE LOC440686 PPIAL4

List of rHHs for discovery stage 1 population cluster 4. The number of ASD probands and parental controls is 85 and 170 respectively.

Table 1d Stage 1 Cluster 5

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
15	62,070,001	62,758,988	688.99	0	6	0	0.001323	TRIP4 KIAA0101 SNX1 CSNK1G1 ZNF609 PPIB DAPK2 SNX22 FAM96A
1	197,287,610	197,374,504	86.89	0	7	1	0.00248	PTPRC
2	195,680,123	195,954,766	274.64	275.33	8	2	0.003232	SLC39A10
4	159,418,343	160,096,823	678.48	0	5	0	0.004018	FLJ11155 RXFP1 LGR7 PPID FLJ25371 LOC201725 ETFDH
13	55,315,271	56,053,311	738.04	559.74	5	0	0.004018	FLJ40296
6	62,736,112	63,231,998	495.89	0	16	11	0.004426	KHDRBS2
1	196,371,095	197,238,929	867.83	0	6	1	0.006654	NEK7 PTPRC ATP6V1G3
6	64,248,405	64,410,979	162.57	0	12	8	0.012058	PTP4A1 PHF3
15	75,114,626	75,631,959	517.33	0	12	8	0.012058	HMG20A C15orf5 TSPAN3 LRRN6A
3	44,048,839	44,110,360	61.52	244.59	4	0	0.01217	C3orf23
4	143,712,663	144,492,808	780.15	0	4	0	0.01217	USP38 FLJ44477 INPP4B GAB1
7	119,484,469	120,644,047	1159.58	0	4	0	0.01217	FLJ21986 KCND2 ING3 TSPAN12
13	55,044,198	56,053,311	1009.11	0	4	0	0.01217	FLJ40296 LOC387930
15	74,610,702	75,631,959	1021.26	0	11	7	0.013546	ZNF291 HMG20A C15orf5 PSTPIP1 TSPAN3 RCN2 LRRN6A
15	80,364,060	81,361,118	997.06	0	5	1	0.01749	DKFZp666G057 FLJ22795 CPEB1 AP3B2 FSD2 HOMER2 RPS17 LOC440295
19	20,972,627	21,567,334	594.71	0	5	1	0.01749	ZNF429 ZNF430 ZNF431 LOC115648 ZNF714 ZNF708 ZNF493
22	39,580,958	40,299,777	718.82	0	8	4	0.018051	TOB2 LOC63929 EP300 PHF5A RANGAP1 TEF RBX1 CSDC2 ACO2 ZC3H7B POLR3H DNAJB7 L3MBTL2 ST13
15	25,695,483	26,723,446	1027.96	0	7	3	0.019054	OCA2 HERC2 GOLGA8G LOC440248
1	196,901,248	197,238,929	337.68	0	6	2	0.019175	PTPRC
4	158,117,004	158,490,178	373.17	0	6	2	0.019175	GLRB GRIA2 PDGFC
5	12,371,620	12,831,499	459.88	0	12	9	0.019866	CTNND2
3	84,130,486	84,922,139	791.65	936.18	11	8	0.022769	CADM2
11	46,319,861	47,140,693	820.83	0	15	14	0.028831	FLJ20294 C11orf49 CKAP5 MDK KIAA0652 ZNF408 DGKZ ARHGAP1 LRP4 FLJ32675 F2 CHRM4
1	769,185	1,102,984	333.80	0	9	6	0.029633	C1orf159 SAMD11 FLJ22639 TTLL10 G1P2 PLEKHN1 NOC2L AGRN HES4 KLHL17
12	48,313,612	48,445,133	131.52	0	8	5	0.033548	TEGT FMNL3 PRPF40B
7	64,407,931	65,140,426	732.50	0	20	22	0.034617	ZNF92 VKORC1L1 LOC441242 GUSB ASL
11	46,659,496	47,140,693	481.20	0	18	19	0.035053	C11orf49 CKAP5 ZNF408 LRP4 F2 ARHGAP1
1	23,099,753	23,758,067	658.31	0	3	0	0.036772	HNRPR AOF2 ID3 LUZP1 HTR1D E2F2 DDEFL1 TCEA3 ZNF436 LOC148898 EPHB2

1	75,222,557	76,259,472	1036.92	0	3	0	0.036772	MSH4 LHX8 SLC44A5 ASB17 ACADM RABGGTB ST6GALNAC3
1	77,732,262	78,757,176	1024.91	0	3	0	0.036772	FAM73A PTGFR FUBP1 GIPC2 ZZZ3 DNAJB4 USP33 AK5 NEXN C1orf118
1	95,973,006	96,716,582	743.58	487.64	3	0	0.036772	RWDD3 PTBP2
1	105,124,662	105,883,977	759.32	1516.81	3	0	0.036772	AMY1C PRMT6
1	144,299,541	144,435,121	135.58	0	3	0	0.036772	POLR3C CD160 ZNF364 PDZK1 NUDT17
1	220,581,960	221,039,808	457.85	0	3	0	0.036772	KIAA1822L TAF1A C1orf80 C1orf58 FLJ43505
2	163,006,664	163,529,784	523.12	0	3	0	0.036772	KCNH7
2	209,626,244	209,753,019	126.78	399.63	3	0	0.036772	MAP2
3	159,069,541	159,689,477	619.94	0	3	0	0.036772	RSRC1 SHOX2
3	166,729,254	166,737,972	8.72	235.41	3	0	0.036772	
3	171,239,540	171,505,135	265.60	0	3	0	0.036772	PRKCI PHC3 GPR160
4	43,253,273	43,600,576	347.30	270.10	3	0	0.036772	KCTD8
4	73,785,834	73,790,935	5.10	132.45	3	0	0.036772	
4	158,949,317	159,314,236	364.92	0	3	0	0.036772	C4orf18
4	159,680,172	160,100,680	420.51	0	3	0	0.036772	RXFP1 LGR7 PPID FLJ25371 LOC201725 ETFDH
5	63,335,845	63,929,058	593.21	0	3	0	0.036772	RNF180 R7BP HTR1A
5	137,975,870	138,840,606	864.74	0	3	0	0.036772	SIL1 MATR3 DNAJC18 CTNNA1 HSPA9B SLC23A1 LRRTM2 PACAP LOC340061 PAIP2
6	67,964,673	68,917,030	952.36	0	3	0	0.036772	BAI3 LOC442229
6	86,835,803	87,545,776	709.97	157.97	3	0	0.036772	HTR1E SYNCRIP
6	87,158,118	88,193,273	1035.16	0	3	0	0.036772	HTR1E CGA C6orf163 C6orf162 GJB7 C6orf165
6	87,181,057	87,539,106	358.05	164.64	3	0	0.036772	HTR1E
6	103,693,415	103,970,247	276.83	1068.76	3	0	0.036772	GRIK2 HACE1
6	112,638,353	112,645,014	6.66	0	3	0	0.036772	LAMA4
7	124,021,266	124,500,240	478.97	0	3	0	0.036772	GPR37 POT1 LOC401398 LOC154872
9	94,608,051	94,950,329	342.28	0	3	0	0.036772	ZNF484 FGD3 NINJ1 ANKRD19 SUSD3 C9orf89
10	57,260,571	58,024,931	764.36	0	3	0	0.036772	ZWINT
10	104,112,832	104,381,024	268.19	0	3	0	0.036772	SUFU CUEDC2 NFKB2 C10orf77 PSD ACTR1A GBF1 C10orf95 FBXL15
10	116,359,696	117,429,024	1069.33	0	3	0	0.036772	ABLIM1 TRUB1 ATRNL1 KIAA1600
11	38,567,581	39,222,671	655.09	869.66	3	0	0.036772	LRRC4C

11	63,034,906	63,041,086	6.18	0	3	0	0.036772	LGALS12
11	89,679,112	90,782,519	1103.41	83.26	3	0	0.036772	CHORDC1
12	85,816,468	86,066,705	250.24	39.66	3	0	0.036772	MGAT4C
13	82,941,754	83,029,226	87.47	320.12	3	0	0.036772	SLITRK1
14	62,845,982	63,971,869	1125.89	0	3	0	0.036772	SGPP1 SYNE2 C14orf150 PPP2R5E GPHB5 ESR2 MTHFD1
15	49,269,214	49,791,852	522.64	0	3	0	0.036772	SCG3 CYP19A1 GLDN DMXL2
15	74,240,308	75,645,624	1405.32	0	3	0	0.036772	ZNF291 HMG20A C15orf5 PSTPIP1 ETFA TSPAN3 ISL2 TYRO3P RCN2 LRRN6A C15orf27
16	67,264,434	67,640,318	375.88	0	3	0	0.036772	CDH1 FLJ12331 CDH3
16	68,698,616	69,252,788	554.17	0	3	0	0.036772	MGC34761 SF3B3 PDPR DDX19-DDX19L AARS ST3GAL2 DDX19B MGC34647 EXOSC6 FUK COG4 DDX19A
17	58,307,001	58,335,701	28.70	69.42	3	0	0.036772	RNF190
19	20,738,115	21,567,334	829.22	0	3	0	0.036772	ZNF85 ZNF429 ZNF430 ZNF626 ZNF431 LOC115648 ZNF714 ZNF708 ZNF493
2	203,330,045	203,885,861	555.82	0	7	4	0.037635	ALS2CR13 ALS2CR15 WDR12 ALS2CR8 ALS2CR16 NBEAL1 CYP20A1 ALS2CR14
7	56,241,393	57,182,269	940.88	0	7	4	0.037635	CHCHD2 LOC401357 LOC441233
15	42,039,369	42,169,690	130.32	0	7	4	0.037635	FRMD5
2	63,117,639	63,607,779	490.14	0	13	12	0.039176	LOC51057 OTX1 EHBP1
8	91,653,445	92,104,497	451.05	0	13	12	0.039176	EFCBP1 TMEM64 TMEM55A
3	51,902,638	52,172,567	269.93	0	6	3	0.041568	DUSP7 IQCF1 ABHD14A WDR51A PARP3 GPR62 PCBP4 RPL29 ABHD14B ACY1 RNU3IP2
6	145,554,235	146,237,874	683.64	0	6	3	0.041568	EPM2A UTRN FBXO30 SHPRH FLJ44955
7	86,318,949	86,513,455	194.51	0	6	3	0.041568	KIAA1324L GRM3
20	32,227,844	32,991,499	763.66	0	6	3	0.041568	ASIP TP53INP2 ACSS2 NCOA6 ITCH MAP1LC3A DYNLRB1 GGTL3 CDC91L1 AHCY HMG4L GSS FLJ38773 EIF2S2
6	64,088,441	64,188,110	99.67	0.6	5	2	0.044581	GLULD1
11	47,311,870	47,393,778	81.91	0	5	2	0.044581	MYBPC3 SPI1 SLC39A13
11	55,346,872	56,320,038	973.17	0	5	2	0.044581	TRIM51 LRRC55 OR9G4 OR5T3 OR5W2 OR5AS1 OR8I2 OR8H1 OR5R1 OR5M1 OR5AP2 OR5M11 OR5I1 OR8K3 OR5D16 OR5T2 OR8K1 OR8J1 OR5M10 OR5AR1 OR9G1 OR8K5 OR5J2 OR5T1 OR8U8 OR8J3 OR5L2 OR10AG1 OR5F1 OR8H2 OR5M9 OR5D18 OR5M8 OR8H3 OR5M3
12	78,368,911	78,968,742	599.83	0	5	2	0.044581	PAWR PPP1R12A SYT1
12	84,539,578	85,552,180	1012.60	0	5	2	0.044581	MGAT4C PAMCI NTS
14	59,329,346	60,617,792	1288.45	0	5	2	0.044581	C14orf39 SIX1 MNAT1 TRMT5 PPM1A SIX6 SLC38A6 SIX4 RTN1 DHRS7 FLJ46156 C14orf135
3	86,260,271	86,361,641	101.37	59.63	4	1	0.044777	CADM2
4	73,414,568	73,892,737	478.17	0	4	1	0.044777	ADAMTS3

6	75,687,853	75,696,912	9.06	40.77	4	1	0.044777	COL12A1
6	121,833,792	121,844,011	10.22	21.22	4	1	0.044777	GJA1
7	62,401,114	63,433,606	1032.49	0	4	1	0.044777	ZNF679 LOC401354
7	86,516,112	86,930,778	414.67	0	4	1	0.044777	CROT ABCB4 C7orf23 TP53AP1 DMTF1
7	113,495,023	114,272,235	777.21	0	4	1	0.044777	FOXP2 PPP1R3A MDFIC
9	10,962,720	11,942,643	979.92	361.22	4	1	0.044777	PTPRD TYRP1
9	34,421,079	34,990,289	569.21	0	4	1	0.044777	DNAI1 DNAJB5 ARID3C UNQ470 LOC389715 CCL21 GALT IL11RA CNTFR C9orf25 DCTN3 OPRS1 CCL19 C9orf23 CCL27
12	84,292,391	84,647,202	354.81	75.26	4	1	0.044777	PAMCI CART1
15	63,015,499	63,108,476	92.98	0	4	1	0.044777	SPG21 MTFMT ANKDD1A
15	73,884,645	74,153,931	269.29	0	4	1	0.044777	UBE2Q2 NRG4 C15orf27 FBXO22
17	24,403,986	24,417,436	13.45	0	4	1	0.044777	PIPOX MYO18A
18	24,706,102	25,390,077	683.98	694.66	4	1	0.044777	CDH2
15	72,831,248	73,236,727	405.48	0	18	20	0.046775	CYP1A2 PPCDC SCAMP2 LMAN1L C15orf17 COX5A C15orf39 MPI RPP25 CSK SCAMP5 ULK3 CPLX3
8	91,274,840	91,889,608	614.77	0	16	17	0.047569	CALB1 TMEM64 EFCBP1
6	64,248,405	64,662,833	414.43	0	9	7	0.048318	PHF3 PTP4A1

List of rHHs for discovery stage 1 population cluster 5. The number of ASD probands and parental controls is 280 and 560 respectively.

Table 1e Stage 1 Cluster 6

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
3	50,612,473	51,987,040	1374.57	0	9	4	0.008235	DOCK3 GRM2 IQCF2 VPRBP CISH MAPKAPK3 RAD54L2 IQCF1 ARMET PARP3 GPR62 TEX264 PCBP4 ABHD14B ABHD14A RNU3IP2 RBM15B
1	145,543,907	145,830,625	286.72	0	4	0	0.012119	GJA8 ACP6 GJA5 BCL9
3	87,489,936	87,708,778	218.84	81.51	4	0	0.012119	POU1F1
8	6,960,131	8,152,381	1192.25	0	4	0	0.012119	PJCG6 DEFB109 DEFB103A DEFB105B DEFB4 LOC401447 DEFA5 LOC349196 DEFB105A DEFB106A SPAG11 DEFB104A
16	66,472,895	66,981,007	508.11	0	14	11	0.014661	UNQ2446 DPEP2 NFATC3 SLC7A6 RBM35B PSMB10 DUS2L PRMT7 SMPD3 LCAT SLC12A4 SLC7A6OS LYPLA3 CTRL RCD-8 DPEP3 PSKH1 DDX28
7	118,009,548	118,879,916	870.37	339.53	10	6	0.014858	ANKRD7 KCND2
5	44,048,229	44,427,899	379.67	0	9	5	0.016422	FGF10
4	64,838,474	65,257,505	419.03	0	5	1	0.017391	SRD5A2L2
7	64,053,719	66,102,199	2048.48	0	5	1	0.017391	FLJ25037 ZNF92 VKORC1L1 ASL TPST1 LOC285908 KCTD7 RABGEF1 LOC441242 FLJ10099 GUSB SBDS H-plk ZNF117 RCP9 ERV3 RSAFD1
4	32,457,740	33,101,619	643.88	0	6	2	0.019034	PCDH7
3	89,180,521	89,604,354	423.83	0	15	13	0.019269	EPHA3
15	70,120,662	70,754,145	633.48	0	64	95	0.021826	MYO9A PARP6 GOLGA HEXA PKM2 ARIH1 GRAMD2 BRUNOL6 HIGD2BP SENP8 GOLGA6
3	48,202,466	48,359,392	156.93	0	11	8	0.022393	NME6 CAMP ZNF589 FBXW12 CDC25A
15	70,754,136	70,947,067	192.93	0	41	55	0.02437	ADPGK BBS4 HIGD2BP
5	43,856,692	44,427,899	571.21	0	8	5	0.033226	NNT FGF10
1	72,305,094	72,616,534	311.44	0	3	0	0.036696	NEGR1
2	62,637,346	63,744,765	1107.42	0	3	0	0.036696	EHBP1 LOC51057 TMEM17 OTX1 MDH1 LOC388955
2	130,854,382	131,262,790	408.41	0	3	0	0.036696	CFC1 PTPN18 GPR148 FLJ38377
2	189,519,319	190,210,280	690.96	0	3	0	0.036696	COL5A2 ASNSD1 WDR75 COL3A1 SLC40A1
5	104,186,930	104,270,540	83.61	192.54	3	0	0.036696	RAB9P1
6	26,449,041	27,411,906	962.87	0	3	0	0.036696	ZNF322A HIST1H2BK GUSBL1 HIST1H2AH HMGN4 BTN3A2 BTN3A3 BTN1A1 BTN3A1 BTN2A3 PRSS16 FKSG83 HIST1H2BJ HIST1H2AG BTN2A2 ABT1 BTN2A1 HIST1H4I
6	56,430,348	57,199,487	769.14	0	3	0	0.036696	DST C6orf65 KIAA1586 ZNF451 RAB23 BAG2
7	63,337,247	64,533,726	1196.48	0	3	0	0.036696	ZNF680 ZNF588 FLJ25037 ZNF92 ZNF679 ZNF273 ZNF138 LOC168474 H-plk ZNF117 ERV3
7	120,773,596	120,910,659	137.06	0	3	0	0.036696	FAM3C
10	80,888,507	81,974,337	1085.83	0	3	0	0.036696	SFTPD SFTPA2 SFTPA1 C10orf56 MBL1P1 ANXA11 C10orf57 PLAC9
11	57,116,992	58,215,812	1098.82	0	3	0	0.036696	OR9Q1 SERPING1 TXNDC14 GLYAT OR9I1 OR10W1 CNTF CTNND1 OR6Q1 OR5B21 OR1S1 OR5B12 LPXN HEAB OR9Q2 OR1S2 OR10Q1 LOC399898 OR5B17 ZFP91 YPEL4 MED19 OR5B3 OR5B2 C11orf31 ZDHHC5

12	55,608,851	55,788,449	179.60	0	3	0	0.036696	RDH16 SDR-O STAT6 TAC3 ADMR MYO1A KIAA0286 ZBTB39 NAB2
15	72,814,933	73,405,467	590.53	0	3	0	0.036696	CYP1A2 PPCDC COMMD4 SCAMP2 LMAN1L C15orf17 COX5A C15orf39 MPI RPP25 CSK LOC554175 SCAMP5 ULK3 CPLX3
16	30,463,540	31,271,994	808.45	0	3	0	0.036696	ZNF668 ITGAM VKORC1 CTF1 MYST1 RNF40 STX4A FUS SRCAP BCL7C FBXL19 SETD1A STX1B2 PRSS36 ITGAX FBS1 PRSS8 LOC283932 LOC493829 MGC13024 ZNF688 ZNF646 MGC3121 PYCARD PHKG2 ZNF689 HSD3B7 BCKDK MGC13138 FLJ32130 PYDC1 LOC90835
17	24,254,805	24,915,990	661.19	0	3	0	0.036696	MYO18A PHF12 CRYBA1 NUFIP2 TIAF1 SEZ6 LOC116236 TAOK1 PIPOX
11	55,235,017	56,424,204	1189.19	0	7	4	0.037346	TRIM51 LRRC55 OR9G4 OR5T3 OR5D13 OR5W2 OR5AS1 OR8I2 OR8H1 OR5R1 OR5M1 OR5AP2 OR5M11 OR5I1 OR8K3 OR5D16 OR5T2 OR8K1 OR8J1 OR5M10 OR5AR1 OR9G1 OR8K5 OR5J2 OR5T1 OR8U8 OR5AK2 OR4C6 OR5D14 OR5D18 OR8J3 OR5L2 OR10AG1 OR5F1 OR8H2 OR5M9 OR5L1 OR5M8 OR8H3 OR5M3
15	28,157,206	29,159,268	1002.06	0	7	4	0.037346	CHRFAM7A FAM7A2 KIAA1018 TJP1 MTMR10 TRPM1
19	42,119,719	42,355,878	236.16	0	7	4	0.037346	ZNF420 ZNF568 ZNF585A
7	64,407,931	64,533,726	125.80	0	13	12	0.038496	ZNF92
1	153,467,814	154,218,253	750.44	0	24	29	0.04007	CLK2 ASH1L MSTO1 RLN3R2 GBA SCAMP3 FDPS PKLR RIT1 DAP3 C1orf104 GON4 HCN3 ARHGEF2 C1orf2 RUSC1 KIAA0907 SYT11 YY1AP1
3	48,202,466	48,349,454	146.99	0	6	3	0.04132	CAMP ZNF589 NME6 CDC25A
8	34,159,707	34,602,339	442.63	582.73	6	3	0.04132	UNC5D DUSP26
19	23,276,616	24,152,378	875.76	0	6	3	0.04132	LOC388524 ZNF91 ZNF675 ZNF254 ZNF539 ZNF681
3	83,609,285	84,152,607	543.32	938.22	5	2	0.044381	CADM2 GBE1
3	161,899,602	163,005,957	1106.36	0	5	2	0.044381	PPM1L NMD3 C3orf57 B3GALT3 ARL14
4	32,457,740	33,292,809	835.07	1700.22	5	2	0.044381	PCDH7 CENTD1
6	101,589,100	101,953,275	364.18	0.31	5	2	0.044381	ASCC3 GRIK2
7	110,270,913	111,083,465	812.55	0	5	2	0.044381	IMMP2L LRRN3 DOCK4
7	126,670,433	127,542,247	871.81	0	5	2	0.044381	GCC1 SND1 GRM8 LOC168850 PAX4 LRRC4 ARF5 NAG8 FSCN3
11	88,097,077	89,081,987	984.91	0	5	2	0.044381	NOX4 GRM5 PSMAL TYR
19	42,063,572	42,669,449	605.88	0	5	2	0.044381	ZNF420 ZNF383 ZNF569 ZNF568 ZNF345 HKR1 ZNF585B ZNF585A ZNF570
2	148,171,771	148,672,202	500.43	0	4	1	0.044635	ORC4L ACVR2A
2	161,604,026	162,336,552	732.53	0	4	1	0.044635	PSMD14 TBR1 SLC4A10 TANK
3	79,129,159	79,494,882	365.72	0	4	1	0.044635	ROBO1
3	131,726,810	131,981,992	255.18	0	4	1	0.044635	PIK3R4 FLJ35880
5	139,219,077	139,329,099	110.02	0	4	1	0.044635	NRG2
7	119,113,463	119,244,423	130.96	456.54	4	1	0.044635	KCND2

8	67,642,365	68,530,031	887.67	0	4	1	0.044635	C8orf45 LOC286187 CSPP1 CPA6 VCPIP1 ARFGEF1 SGK3 C8orf44 COPS5
8	89,831,029	89,933,336	102.31	422.19	4	1	0.044635	MMP16
10	104,188,229	104,203,492	15.26	0	4	1	0.044635	C10orf95
14	41,025,847	41,091,740	65.89	55.35	4	1	0.044635	LRFN5
6	62,760,137	63,231,998	471.86	0	12	11	0.045056	KHDRBS2

List of rHHs for discovery stage 1 population cluster 6. The number of ASD probands and parental controls is 217 and 434 respectively.

Table 1f Stage 2 Cluster 3

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
17	55,582,582	56,166,117	583.54	0	8	2	0.003403	C17orf64 PPM1D USP32 BCAS3 APPBP2 CA4
3	52,444,825	52,453,468	8.64	0	16	11	0.004822	SEMA3G
17	24,675,154	24,908,731	233.58	0	6	1	0.006922	TAOK1 NUFIP2 LOC116236
17	54,871,776	55,146,813	275.04	0	6	1	0.006922	DHX40 CLTC PTRH2 TMEM49 YPEL2
17	54,922,333	56,166,117	1243.78	0	7	2	0.008337	DHX40 C17orf64 PPM1D ABC1 USP32 LOC51136 TMEM49 RPS6KB1 CLTC PTRH2 CA4 BCAS3 TUBD1 APPBP2
3	96,327,642	96,396,002	68.36	999.32	10	5	0.008465	NSUN3
1	28,451,582	29,102,336	650.75	0	9	4	0.008821	PHACTR4 RAB42 YTHDF2 GMEB1 OPRD1 SESN2 MED18 RCC1 TAF12 EPB41 TRSPAP1
10	37,886,527	38,295,995	409.47	0	9	4	0.008821	ZNF25 ZNF248
14	45,406,142	45,820,357	414.22	613.99	8	3	0.008855	C14orf106 RPL10L
3	97,650,079	98,674,274	1024.20	0	15	11	0.008927	ARL6
15	54,359,468	55,084,899	725.43	0	17	14	0.010674	TEX9 MNS1 SUHW4 TCF12
1	28,083,990	28,113,579	29.59	0	4	0	0.012509	C1orf38 RPA2
1	75,399,916	76,250,268	850.35	0	4	0	0.012509	MSH4 SLC44A5 ASB17 ACADM RABGGTB ST6GALNAC3 LHX8
1	106,114,035	106,236,632	122.60	1164.3	4	0	0.012509	NTNG1 PRMT6
1	188,843,959	189,445,799	601.84	0	4	0	0.012509	FAM5C
1	193,058,963	193,433,937	374.97	1174.90	4	0	0.012509	KCNT2
2	117,554,690	117,676,894	122.20	611.83	4	0	0.012509	DDX18
3	96,453,317	97,442,615	989.30	1166.40	4	0	0.012509	NSUN3 ARL6
5	100,354,764	101,574,538	1219.77	23.05	4	0	0.012509	SLCO4C1 ST8SIA4
7	125,255,643	125,957,858	702.22	0	4	0	0.012509	GRM8
10	116,666,250	117,410,105	743.86	0	4	0	0.012509	TRUB1 ATRNL1
11	65,818,599	65,820,702	2.10	0	4	0	0.012509	MGC33486
12	43,849,059	43,961,112	112.05	0	4	0	0.012509	TMEM16F PLEKHA9
13	63,524,248	64,076,454	552.21	309.55	4	0	0.012509	OR7E156P
14	59,264,894	59,662,154	397.26	0	4	0	0.012509	RTN1 FLJ46156 C14orf135
14	66,967,424	67,018,827	51.40	0	4	0	0.012509	FLJ33387
12	108,800,318	109,802,238	1001.92	0	9	5	0.017497	ARPC3 TA-PP2C FLJ21127 FLJ40142 ATPBD1C IFT81 MGC15619 GIT2 PPP1CC RAD9B TCHP ANKRD13 VPS29 ATP2A2 CCDC63 ANAPC7 GLTP C12orf24

2	184,907,717	184,985,412	77.70	0	5	1	0.018059	C2orf10
6	75,045,174	75,681,347	636.17	169.42	5	1	0.018059	COL12A1 CD109
7	114,083,340	114,134,390	51.05	0	5	1	0.018059	FOXP2
12	86,455,427	86,755,189	299.76	146.12	5	1	0.018059	C12orf50
14	34,700,070	34,902,417	202.35	0	5	1	0.018059	NFKBIA KIAA0391 PSMA6
15	42,132,228	42,775,902	643.67	0	5	1	0.018059	FRMD5 CASC4 EIF3S1 KIAA1840 CTDSPL2 B2M
1	49,425,456	50,595,152	1169.70	0	13	10	0.018448	AGBL4 ELAVL4 FAF1
3	88,948,580	89,192,371	243.79	46.99	7	3	0.019848	EPHA3
5	137,444,450	137,918,000	473.55	0	7	3	0.019848	GFRA3 WNT8A JMJD1B BRD8 CDC25C ETF1 EGR1 CDC23 REEP2 KIF20A FAM53C NME5
9	122,399,025	123,168,845	769.82	0	6	2	0.019887	CEP1 C5 PSMD5 TRAF1 PHF19 STOM CDK5RAP2 RAB14 GSN FBXW2
12	33,403,499	34,717,841	1314.34	0	6	2	0.019887	ALG10 SYT10
12	109,770,460	109,796,312	25.85	0	6	2	0.019887	CCDC63
15	42,350,178	42,775,902	425.72	0	6	2	0.019887	CASC4 EIF3S1 KIAA1840 CTDSPL2 B2M
3	48,149,214	48,481,308	332.09	0	17	16	0.022942	NME6 CAMP ZNF589 PLXNB1 FBXW12 TREX1 CCDC51 CDC25A CCDC72
17	38,094,923	38,315,794	220.87	0	11	8	0.024064	RAMP2 EZH1 CNTNAP1 BECN1 PSME3 G6PC FLJ31222 CNTD AOC2 LOC90586 AOC3 WNK4 VPS25 CCDC56
5	12,232,005	12,682,482	450.48	274.89	14	12	0.024498	CTNND2
12	109,824,626	109,842,617	17.99	0	26	30	0.026818	MYL2
3	51,388,422	52,237,490	849.07	0	10	7	0.027384	GRM2 IQCF2 DUSP7 ALAS1 VPRBP RAD54L2 IQCF1 ARMET ABHD14A DOCK3 WDR51A PARP3 GPR62 TEX264 PCBP4 RPL29 TLR9 ABHD14B ACY1 RNU3IP2 RBM15B
4	151,437,641	152,368,667	931.03	0	10	7	0.027384	LRBA MAB21L2 SH3D19 RPS3A RNU73
10	37,272,335	38,295,995	1023.66	0	10	7	0.027384	ANKRD30A ZNF25 ZNF248
3	52,154,126	52,230,784	76.66	0	18	18	0.027688	ALAS1 TLR9 WDR51A
3	51,326,096	51,808,661	482.57	0	82	127	0.029595	GRM2 DOCK3 VPRBP RAD54L2 ARMET TEX264 IQCF2 RBM15B
7	83,835,135	84,832,369	997.23	0	9	6	0.031045	SEMA3D SEMA3A
12	110,118,664	110,140,615	21.95	0	66	99	0.03335	CUTL2
3	50,682,083	51,368,414	686.33	0	89	141	0.034344	DOCK3
2	196,786,698	196,941,944	155.25	0	8	5	0.034993	HECW2
17	55,107,560	56,166,117	1058.56	0	8	5	0.034993	C17orf64 PPM1D ABC1 USP32 LOC51136 TMEM49 RPS6KB1 PTRH2 CA4 BCAS3 CLTC TUBD1 APPBP2
20	33,262,612	33,350,919	88.31	0	8	5	0.034993	MMP24 C20orf128 C20orf127 ITGB4BP C20orf44
1	72,070,354	72,154,184	83.83	0	3	0	0.037527	NEGR1

1	103,248,767	104,152,300	903.53	0	3	0	0.037527	RNPC3 AMY2B AMY2A AMY1A AMY1B COL11A1 AMY1C
1	104,787,242	104,790,148	2.91	746.81	3	0	0.037527	AMY1C
1	174,672,616	175,107,669	435.05	0	3	0	0.037527	PAPPA2 ASTN
2	62,875,554	63,747,093	871.54	0	3	0	0.037527	LOC51057 EHBP1 OTX1 MDH1 LOC388955
2	96,440,369	96,949,663	509.29	0	3	0	0.037527	ARID5A FLJ10081 LOC51252 LMAN2L ANKRD39 SEMA4C CNNM3 CNNM4 BRRN1 ANKRD23
2	162,047,297	162,557,972	510.68	0	3	0	0.037527	TBR1 SLC4A10 DPP4
3	83,742,555	84,916,203	1173.65	942.12	3	0	0.037527	CADM2 GBE1
3	88,545,295	88,561,356	16.06	256.54	3	0	0.037527	C3orf38
3	122,754,748	122,768,183	13.44	7.20	3	0	0.037527	POLQ ARGFX
4	32,237,710	32,250,599	12.89	1480.19	3	0	0.037527	PCDH7
4	72,813,981	73,790,935	976.95	0	3	0	0.037527	ADAMTS3 NPFFR2 GC
4	144,415,986	144,566,689	150.70	0	3	0	0.037527	GAB1 USP38
4	152,674,789	153,073,117	398.33	0	3	0	0.037527	PET112L
5	44,431,772	44,682,210	250.44	7.23	3	0	0.037527	FGF10 MRPS30
5	87,114,798	88,222,710	1107.91	0	3	0	0.037527	MGC33214 MEF2C CCNH
5	104,401,279	104,967,637	566.36	0	3	0	0.037527	NUDT12 RAB9P1
5	136,962,383	137,980,941	1018.56	0	3	0	0.037527	KLHL3 C5orf5 GFRA3 PKD2L2 WNT8A NPY6R JMJD1B BRD8 CDC25C ETF1 HSPA9B HNRPA0 MYOT EGR1 CDC23 REEP2 KIF20A FAM53C NME5
6	47,446,050	47,813,001	366.95	0	3	0	0.037527	CD2AP GPR115 GPR111 TNFRSF21
6	50,056,018	50,119,755	63.74	0	3	0	0.037527	CRISP1
6	79,372,896	79,913,349	540.45	0	3	0	0.037527	IRAK1BP1 PHIP HMGN3
6	140,202,811	141,369,259	1166.45	466.04	3	0	0.037527	CITED2 NMBR
7	109,435,065	109,603,374	168.31	486.97	3	0	0.037527	IMMP2L
10	23,412,062	23,440,399	28.34	0	3	0	0.037527	MSRB2
10	41,677,980	42,727,244	1049.26	0	3	0	0.037527	BMS1L MGC16291 ZNF11B DUXAP3
10	73,556,312	74,383,141	826.83	0	3	0	0.037527	OIT3 C10orf42 CBARA1 DNAJB12 DDIT4 C10orf104 ASCC1 PLA2G12B
12	32,883,406	34,717,841	1834.44	0	3	0	0.037527	PKP2 SYT10 ALG10
12	82,713,396	83,756,207	1042.81	42.55	3	0	0.037527	SLC6A15 TMTC2
12	86,886,923	87,959,308	1072.39	0	3	0	0.037527	C12orf29 KITLG DUSP6 TMTC3 CEP290 C12orf50
15	39,008,359	39,634,468	626.11	0	3	0	0.037527	CHP INOC1 OIP5 LTK EXDL1 NUSAP1 CHAC1 RTF1 NDUFAF1 RPAP1 ITPKA DLL4 TYRO3

15	62,135,911	63,011,807	875.90	0	3	0	0.037527	TRIP4 RBPMS2 KIAA0101 SNX1 PLEKHQ1 CSNK1G1 OAZ2 ZNF609 PPIB ANKDD1A SNX22 FAM96A C15orf20
16	28,240,912	29,582,827	1341.92	0	3	0	0.037527	LOC440350 EIF3S8 LOC440348 ATXN2L LAT BOLA2 SULT1A4 SULT1A3 SULT1A1 ATP2A1 LOC112869 CLN3 P8 SPN NFATC2IP LOC400509 SBK1 RABEP2 SPIN1 SULT1A2 SH2B TUFM CD19 APOB48R IL27 LAT1-3TM
17	15,272,242	15,766,813	494.57	0	3	0	0.037527	TRIM16 FAM18B2 MEIS3P1 MGC51025 NBLA10383 ADORA2B CDRT1 ZNF286
17	19,598,673	20,640,829	1042.16	0	3	0	0.037527	SPECC1 MGC87631 AKAP10 ULK2 ALDH3A1
17	21,679,693	22,775,097	1095.40	0	3	0	0.037527	FAM27L WSB1 KSR1
2	131,246,068	131,877,774	631.71	0	7	4	0.039087	PLEKHB2 POTE2 LOC130074 ARHGEF4 FLJ38377 LOC401010
11	47,180,838	47,309,887	129.05	0	7	4	0.039087	MADD NR1H3 MYBPC3 ACP2 DDB2
15	42,742,701	43,146,450	403.75	0	7	4	0.039087	C15orf43 SORD KIAA1840 RNF36 B2M
3	48,516,020	48,548,315	32.30	0	13	12	0.041496	PFKFB4 SCOTIN
3	50,215,342	50,312,036	96.69	0	6	3	0.042986	IFRD2 GNAI2 SEMA3B C3orf45 HYAL3 SLC38A3 NAT6
6	26,247,375	26,379,692	132.32	0	6	3	0.042986	HIST1H3D HIST1H4G HIST1H2BD HIST1H4D HIST1H1D HIST1H2BF HIST1H2BG HIST1H2AE HIST1H2BH HIST1H4E HIST1H3G HIST1H1E HIST1H2BE HIST1H3F HIST1H2AC HIST1H3E HIST1H2AD HIST1H4F
10	48,481,839	49,146,922	665.08	0	6	3	0.042986	PTPN20B FRMPD2
10	69,638,195	70,065,576	427.38	0	6	3	0.042986	MAWBP RUFY2 SLC25A16 FLJ14437 HNRPH3 CXXC6 ATOH7
17	54,834,525	54,842,320	7.80	0.65	6	3	0.042986	YPEL2
2	186,264,709	187,087,060	822.35	0	15	15	0.04325	FLJ44048 FSIP2 LEREPO4
2	95,754,597	96,206,266	451.67	0	10	8	0.043659	LOC400986 LOC150763 ADRA2B ASTL DUSP2 TRIM43 STARD7
8	50,614,095	51,050,357	436.26	0	66	101	0.044008	C8orf22 SNTG1
8	50,863,839	51,100,596	236.76	0	67	103	0.04519	SNTG1
1	45,283,888	46,517,353	1233.47	0	4	1	0.045903	RAD54L C1orf190 TESK2 GPBP1L1 IPP POMGNT1 MAST2 NASP PRDX1 MMACHC TMEM69 PIK3R3 GLOXD1 MUTYH TSPAN1 UROD CCDC17 TOE1 AKR1A1 LRRC41
1	46,507,379	46,668,228	160.85	0	4	1	0.045903	NSUN4 RAD54L LRRC41 UQCRH FAAH
1	144,300,134	144,417,650	117.52	0	4	1	0.045903	POLR3C CD160 ZNF364 NUDT17
2	27,246,534	28,468,298	1221.76	0	5	2	0.045903	FLJ13646 BRE SLC5A6 C2orf16 IFT172 RBKS SLC4A1AP SLC30A3 PPM1G SNX17 DNAJC5G GTF3C2 TRIM54 SUPT7L CAD MRPL33 NRBP1 FNDC4 KRTCAP3 MPV17 EIF2B4 FTHL3 C2orf28 ZNF512 GCKR XAB1 FOSL2 ZNF513 UCN
2	95,401,820	96,507,994	1106.17	0	5	2	0.045903	FAHD2A TRIM43 LOC400986 LOC150763 ADRA2B BRRN1 ASTL DUSP2 WDR39 STARD7 CSEN ASCC3L1 TMEM127 KIAA1754L ARID5A
2	161,613,247	162,618,649	1005.40	0	4	1	0.045903	PSMD14 TBR1 SLC4A10 TANK DPP4
2	188,256,322	188,469,784	213.46	128.86	5	2	0.045903	TFPI
2	193,043,904	193,754,540	710.64	276.01	5	2	0.045903	TMEFF2

3	85,964,415	86,069,410	105.00	0	4	1	0.045903	CADM2
3	86,365,276	87,069,510	704.23	0	4	1	0.045903	VGLL3 CADM2
3	130,420,327	131,114,469	694.14	0	4	1	0.045903	C3orf47 MBD4 PLXND1 H1FOO TMCC1 COPG C3orf37 IFT122 C3orf25 RHO H1FX
3	139,501,031	140,467,990	966.96	0	5	2	0.045903	PIK3CB FLJ46210 LOC389151 FOXL2 BPESC1 FAM62C MRPS22 CEP70 MRAS FAIM TXNDC6
5	136,403,949	136,454,266	50.32	0	4	1	0.045903	SPOCK
5	139,228,155	139,390,871	162.72	0	5	2	0.045903	NRG2
6	63,908,236	64,662,893	754.66	0	4	1	0.045903	GLULD1 PHF3 PTP4A1
6	65,486,749	65,508,791	22.04	0	4	1	0.045903	EGFL11
6	81,978,615	82,529,275	550.66	0	5	2	0.045903	FAM46A
6	100,986,372	101,953,275	966.90	0	4	1	0.045903	ASCC3 GRIK2 SIM1
7	86,233,862	86,554,038	320.18	0	4	1	0.045903	KIAA1324L GRM3 DMTF1
7	113,691,297	114,134,390	443.09	0	4	1	0.045903	FOXP2
8	7,157,150	8,146,665	989.52	0	5	2	0.045903	DEFB109 DEFB103A DEFB105B DEFB4 LOC401447 DEFB105A DEFB106A SPAG11 DEFB104A
8	34,246,490	34,640,824	394.33	669.51	5	2	0.045903	UNC5D DUSP26
8	116,967,812	117,306,506	338.69	217.41	4	1	0.045903	TRPS1 EIF3S3
12	84,317,328	84,985,902	668.57	0	4	1	0.045903	MGAT4C PAMCI CART1 NTS
17	24,290,020	24,908,731	618.71	0	4	1	0.045903	MYO18A CRYBA1 NUFIP2 TIAF1 SEZ6 TAOK1 PHF12 PIPOX LOC116236

List of rHHs for replication stage 2 population cluster 3. The number of ASD probands and parental controls is 306 and 606 respectively.

Table 1g Stage 2 Cluster 4

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
8	111,983,792	112,098,047	114.26	927.19	5	0	0.003824	KCNV1
12	85,024,867	85,824,120	799.25	0	5	0	0.003824	MGAT4C
20	32,597,214	33,277,616	680.40	0	13	10	0.014686	TP53INP2 ACSS2 NCOA6 C20orf31 MAP1LC3A TRPC4AP PROCR GGTL3 GSS CDC91L1 MYH7B HMG4L C20orf127 MMP24
3	163,680,829	163,739,058	58.23	1108.26	5	1	0.016805	C3orf57
19	22,522,613	23,028,643	506.03	0	6	2	0.018203	ZNF91 ZNF676
13	54,564,640	55,195,142	630.50	0	9	6	0.027241	LOC387930
7	98,715,203	99,260,475	545.27	0	12	10	0.027287	CYP3A5 CYP3A7 CYP3A4 ZNF655 PTCD1 G10 ZNF498 ARPC1B ZFP95 CPSF4 ZNF394 LOC285989 DKFZp727G131 ARPC1A PDAP1 MYH16 ATP5J2 CYP3A43
2	152,387,708	152,918,690	530.98	0	7	4	0.035632	STAM2 CACNB4 FMNL2 ARL5
2	187,211,091	187,384,199	173.11	0	3	0	0.03624	KIAA1946 ITGAV ZSWIM2
10	48,221,579	49,074,485	852.91	0	3	0	0.03624	PTPN20B FRMPD2
12	121,217,128	121,532,326	315.20	0	3	0	0.03624	RSN ZCCHC8 IL31 LRRC43 VPS33A B3GNT4 DIABLO
15	63,023,928	63,135,537	111.61	0	3	0	0.03624	RASL12 SPG21 MTFMT ANKDD1A OSTbeta
20	33,183,694	33,277,616	93.92	0	18	20	0.038777	PROCR C20orf31 C20orf127 MMP24
1	50,671,566	51,362,243	690.68	0	6	3	0.039844	FAF1 CDKN2C RNF11
2	194,041,941	195,037,928	995.99	1191.85	4	1	0.043788	TMEFF2 SLC39A10
8	112,667,537	113,416,432	748.90	0	4	1	0.043788	CSMD3
4	71,771,262	72,211,493	440.23	0	11	10	0.048828	RUFY3 GRSF1 DCK MOBKL1A SAS10

List of rHHs for replication stage 2 population cluster 4. The number of ASD probands and parental controls is 93 and 186 respectively.

Table 1h Stage 2 Cluster 5

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
2	88,475,282	89,868,877	1393.60	0	6	0	0.001338	RPIA LOC651928 FLJ25369 EIF2AK3 GeneID:28881 GeneID:28874 GeneID:3514 GeneID:28913 GeneID:28883 GeneID:28912 GeneID:28299 GeneID:28919 GeneID:28885 GeneID:28950 GeneID:28930 GeneID:28906 GeneID:28875 GeneID:28935 GeneID:28869 GeneID:28940 GeneID:28903 GeneID:28876 GeneID:28902 GeneID:28915 GeneID:28931 GeneID:28896 GeneID:28933 GeneID:28937 GeneID:28923 GeneID:28870 GeneID:28892 GeneID:28900
7	98,719,135	99,526,758	807.62	0	10	3	0.001561	CYP3A5 CYP3A7 CYP3A4 ZNF655 AZGP1 PTCD1 G10 ZNF498 ARPC1B ZFP95 CPSF4 CYP3A43 ZNF394 LOC285989 DKFZp727G131 ARPC1A ZKSCAN1 COPS6 PDAP1 ZNF38 MYH16 ZNF3 TRIM4 ATP5J2
3	49,581,979	50,133,778	551.80	0	8	2	0.003286	RBM5 MST1R BSN IHPK1 CAMKV APEH RNF123 RBM6 UBE1L TRAIP MON1A C3orf54 LOC389118 GMPPB MST1 AMIGO3
2	187,270,600	187,582,053	311.45	0	5	0	0.004049	ZSWIM2 KIAA1946
6	26,866,458	27,724,244	857.79	0	5	0	0.004049	HIST1H2BK ZNF184 GUSBL1 ZNF204 HIST1H2AH LOC441136 PRSS16 FKSG83 HIST1H2BJ HIST1H2AG HIST1H4I HIST1H2BL
7	119,435,323	119,559,697	124.37	141.26	5	0	0.004049	KCND2
16	31,453,119	34,101,480	2648.36	0	5	0	0.004049	TP53TG3 ERAF ZNF720 FLJ43855 FLJ46121 ZNF267 LOC441762 MGC34800 LOC440366 MGC3020 HERC2P4
16	69,623,579	69,648,589	25.01	0	5	0	0.004049	HYDIN
16	31,343,178	31,537,650	194.47	0	8	3	0.008558	ERAF ARMC5 SLC5A2 C16orf58 COX6A2 TGFB1I1 ITGAD MGC3020
8	91,886,818	92,104,497	217.68	0	24	24	0.010967	EFCBP1 TMEM55A
2	137,728,653	138,329,393	600.74	0	4	0	0.012225	HNMT
2	187,270,600	188,286,647	1016.05	0	4	0	0.012225	ZSWIM2 TFPI CALCRL KIAA1946
3	88,819,045	90,559,061	1740.02	0	4	0	0.012225	EPHA3
6	27,599,278	28,141,484	542.21	0	4	0	0.012225	ZNF184 HIST1H2BL OR2B2 OR2B6 HIST1H4J HIST1H4L HIST1H2AL HIST1H1B HIST1H2BN HIST1H2BO HIST1H4K ZNF165 HIST1H2AM HIST1H3J HIST1H2AK HIST1H2AI HIST1H2BM HIST1H3H HIST1H2AJ
8	35,449,698	36,091,383	641.69	0	4	0	0.012225	UNC5D
14	39,330,910	39,413,454	82.54	359.54	4	0	0.012225	FBXO33
16	69,623,579	70,512,901	889.32	0	4	0	0.012225	CALB2 AP1G1 HYDIN PHLPPL ZNF19 CHST4 TAT KIAA0174 MARVELD3 LOC55565 ZNF23 FLJ11171
22	39,690,162	40,199,571	509.41	0	4	0	0.012225	TOB2 EP300 PHF5A RANGAP1 TEF RBX1 ZC3H7B L3MBTL2 ACO2
5	21,899,711	22,127,725	228.01	0	9	5	0.016897	CDH12
2	21,903,121	22,470,306	567.19	982.67	5	1	0.017599	APOB
2	177,863,367	178,794,325	930.96	0	5	1	0.017599	AGPS PDE11A DRB1 DKFZp451M2119 OSBPL6 FLJ30990 FLJ13946
3	161,253,089	162,285,936	1032.85	0	5	1	0.017599	IL12A TRIM59 KPNA4 PPM1L IFT80 SMC4L1 ARL14 B3GALT3
4	132,420,036	133,099,653	679.62	1190.27	5	1	0.017599	PCDH10
22	39,391,662	39,591,261	199.60	0	5	1	0.017599	GPR24 SLC25A17 ST13 DNAJB7
1	145,614,797	145,640,129	25.33	0	8	4	0.018275	ACP6

2	88,378,005	88,457,593	79.59	0	8	4	0.018275	FLJ10916 FLJ25369
2	96,178,964	96,348,789	169.83	0	8	4	0.018275	WDR39 STARD7 ASCC3L1 TMEM127 DUSP2 KIAA1754L
8	85,996,580	87,104,406	1107.83	0	6	2	0.019329	CA13 REXO1L1 LRRCC1 E2F5 LOC138046 CA2 PSKH2 CA3 CA1
12	83,763,772	84,609,030	845.26	0	6	2	0.019329	LRRIQ1 SLC6A15 PAMCI CART1
17	42,663,046	42,758,056	95.01	0	6	2	0.019329	ITGB3 MYL4 C17orf57
7	62,754,013	62,951,888	197.88	374.39	14	12	0.023519	LOC401354 ZNF679
8	92,065,293	92,104,497	39.20	0	24	27	0.026284	TMEM55A
15	43,503,089	43,527,684	24.60	0	10	7	0.026424	C15orf48 SPATA5L1
3	50,264,857	50,584,628	319.77	0	18	18	0.0265	TUSC2 C3orf18 TMEM115 IFRD2 CACNA2D2 HYAL1 ZMYND10 TUSC4 SEMA3B RASSF1 C3orf45 GNAI2 HYAL3 HYAL2 HEMK1 CYB561D2 NAT6
3	96,569,961	96,763,711	193.75	1241.64	18	18	0.0265	NSUN3
3	47,573,667	48,359,392	785.73	0	15	14	0.029561	MAP4 NME6 SMARCC1 CAMP ZNF589 DHX30 CSPG5 FBXW12 CDC25A
12	109,837,939	111,176,330	1338.39	0	9	6	0.030012	MYL2 ATXN2 BRAP C12orf51 CUTL2 ALDH2 C12orf30 FAM109A LNK ACAD10 C12orf47 C12orf8 TRAFD1 MAPKAPK5 TMEM116
6	27,871,573	27,882,803	11.23	0.43	12	10	0.03147	HIST1H2BL
3	48,388,541	48,599,044	210.50	0	8	5	0.033899	COL7A1 SCOTIN PLXNB1 PFKFB4 TREX1 CCDC51 UCN2 FBXW12 CCDC72
10	22,384,234	22,788,915	404.68	0	8	5	0.033899	COMMD3 SPAG6 DNAJC1 PCGF4
12	109,815,399	109,842,617	27.22	0	8	5	0.033899	CCDC63 MYL2
8	100,176,109	101,062,189	886.08	0	14	13	0.034333	VPS13B RGS22 COX6C
3	47,894,539	47,965,492	70.95	0	18	19	0.036101	MAP4
1	51,113,315	51,798,972	685.66	0	3	0	0.036856	FAF1 RNF11 EPS15 CDKN2C
1	72,124,825	72,843,006	718.18	0	3	0	0.036856	NEGR1
1	80,053,594	80,418,816	365.22	1151.24	3	0	0.036856	IFI44
1	174,702,190	175,083,033	380.84	0	3	0	0.036856	PAPPA2
2	83,328,343	83,454,796	126.45	916.52	3	0	0.036856	LOC388965
2	116,108,246	116,797,239	688.99	0	3	0	0.036856	DPP10
2	136,776,376	137,286,278	509.90	184.18	3	0	0.036856	CXCR4
2	185,024,702	185,917,555	892.85	0	3	0	0.036856	C2orf10 FSIP2
2	200,418,951	200,738,211	319.26	0	3	0	0.036856	FLJ22555 FLJ38973 DNAPTP6 FLJ37953
2	201,044,874	202,076,828	1031.95	0	3	0	0.036856	BZW1 MGC39518 CFLAR CASP10 TRAK2 AOX2 KCTD18 ALS2CR2 SGOL2 NIF3L1 ORC2L AOX1 DNAPTP6 ALS2CR12 NDUFB3 CASP8 PPIL3 ALS2CR11 CLK1
3	42,160,566	42,200,955	40.39	0	3	0	0.036856	TRAK1

3	55,781,036	56,755,043	974.01	0	3	0	0.036856	CAST1 ARHGEF3 C3orf63 CCDC66
4	45,336,416	45,401,994	65.58	330.55	3	0	0.036856	GABRG1
4	104,049,242	104,710,262	661.02	0	3	0	0.036856	LOC133308 BDH2 CENPE LOC150159 TACR3
4	128,298,777	129,531,273	1232.50	0	3	0	0.036856	SLC25A31 LARP2 PGRMC2 PDZD6 MGC33302 PLK4 FLJ21106 HSPA4L
5	29,950,078	30,154,264	204.19	0	3	0	0.036856	CDH6
5	41,828,046	42,653,249	825.20	0	3	0	0.036856	GHR FBXO4 LOC285636 OXCT1
5	42,908,813	43,430,535	521.72	0	3	0	0.036856	SEPP1 ZNF131 HMGCS1 FLJ10246 MGC42105 LOC153684 CCL28 LOC389289
6	50,728,920	51,055,980	327.06	0	3	0	0.036856	TFAP2B TFAP2D
6	97,656,784	98,289,833	633.05	0	3	0	0.036856	C6orf167 KIAA1900
7	117,515,414	118,009,548	494.13	0	3	0	0.036856	ANKRD7 LSM8
8	113,722,276	114,824,894	1102.62	0	3	0	0.036856	CSMD3
10	89,485,344	89,564,941	79.60	0	3	0	0.036856	ATAD1 PAPSS2
10	103,495,326	104,528,601	1033.28	0	3	0	0.036856	C10orf76 GBF1 SUFU CUEDC2 NFKB2 NOLC1 PPRC1 C10orf77 PSD ACTR1A ELOVL3 TRIM8 SFXN2 FGF8 KCNIP2 HPS6 ARL3 MGEA5 LDB1 C10orf95 C10orf26 PITX3 NPM3 FBXL15
11	66,577,304	67,022,738	445.43	0	3	0	0.036856	FBXL11 RPS6KB2 AIP ADRBK1 PPP1CA RHOD ANKRD13D RAD9A TBC1D10C POLD4 CLCF1 CORO1B CABP4 FLJ21749 SSH3 PTPRCAP GPR152 PITPNM1
11	104,383,013	105,008,868	625.86	0	3	0	0.036856	GRIA4 ICEBERG COP1 INCA CASP5 CASP1
11	107,128,834	107,347,799	218.97	0	3	0	0.036856	SLC35F2 RAB39
12	56,266,216	56,413,599	147.38	0	3	0	0.036856	OS9 GALGT GEFT DTX3 KIF5A CENTG1 PIP5K2C SLC26A10
12	86,332,557	87,755,618	1423.06	0	3	0	0.036856	C12orf50 C12orf29 KITLG TMTC3 CEP290
13	55,454,251	56,463,195	1008.94	149.86	3	0	0.036856	FLJ40296
13	56,016,237	57,077,004	1060.77	26.78	3	0	0.036856	FLJ40296 PCDH17
13	82,265,803	82,987,208	721.41	362.14	3	0	0.036856	SPRY2 SLITRK1
15	42,752,594	42,888,343	135.75	0	3	0	0.036856	RNF36 B2M KIAA1840
15	43,215,739	43,233,448	17.71	0	3	0	0.036856	DUOX1
16	30,102,515	30,389,995	287.48	0	3	0	0.036856	CD2BP2 SULT1A3 LOC51333 SEPHS2 SEPT1 XTP3TPA ZNF553 LOC595101 TBC1D10B MYLPF CORO1A ITGAL LOC440354
16	30,463,540	31,189,252	725.71	0	3	0	0.036856	ZNF668 VKORC1 CTF1 MYST1 RNF40 STX4A FUS SRCAP BCL7C FBXL19 SETD1A STX1B2 PRSS36 FBS1 PRSS8 ITGAM LOC283932 LOC493829 MGC13024 ZNF688 ZNF646 MGC3121 PYCARD PHKG2 ZNF689 HSD3B7 BCKDK MGC13138 FLJ32130 PYDC1 LOC90835
16	68,099,877	69,171,509	1071.63	0	3	0	0.036856	MGC34761 NFAT5 NQO1 WWP2 PDPR DDX19-DDX19L SF3B3 AARS ST3GAL2 LOC348174 DDX19B EXOSC6 FUK COG4 DDX19A NOB1P CYB5-M
17	17,658,534	18,810,083	1151.55	0	3	0	0.036856	DRG2 FLJ35934 LOC654346 LOC220594 FLJ36492 TOM1L2 PRPSAP2 FBXW10 LRRC48 ALKBH5 FLII FAM18B C17orf39 ATPAF2 SREBF1 SLC5A10 MYO15A SMCR8 SHMT1 TOP3A LLGL1 LOC339240 SMCR7 FAM106A

17	24,936,408	25,562,658	626.25	0	3	0	0.036856	CORO6 SSH2 GIT1 SLC6A4 FLJ46247 CCDC55 ANKRD13B
17	53,856,419	54,731,591	875.17	0	3	0	0.036856	MTMR4 PRR11 PPM1E TEX14 SEPT4 FAM33A TRIM37 GDPD1 C17orf71 RNF43 C17orf47 RAD51C
17	59,176,841	59,462,855	286.01	0	3	0	0.036856	CD79B DDX42 SCN4A ICAM2 PSMC5 CSH1 CCDC47 CSH2 GH2 SMARCD2 FTSJ3 CSHL1 GH1 ERN1
18	49,437,677	50,668,904	1231.23	0	3	0	0.036856	C18orf26 DCC STARD6 RAB27B MBD2 C18orf54 POLI
19	23,233,776	23,502,953	269.18	0	3	0	0.036856	ZNF91 ZNF675
2	95,865,668	96,163,207	297.54	0	7	4	0.03795	LOC400986 LOC150763 ADRA2B ASTL
2	96,178,964	96,215,344	36.38	0	7	4	0.03795	STARD7 DUSP2
5	43,763,299	44,427,899	664.60	0	7	4	0.03795	NNT FGF10
7	101,901,824	102,533,314	631.49	0	7	4	0.03795	POLR2J POLR2J3 RASA4 POLR2J2 MGC119295 MGC35361 FBXL13 LRRC17 SVH NAPE-PLD
11	87,914,378	88,846,098	931.72	0	7	4	0.03795	GRM5 TYR NOX4
12	87,217,094	87,933,820	716.73	0	7	4	0.03795	KITLG DUSP6 TMTC3
13	54,819,860	55,342,530	522.67	1270.52	7	4	0.03795	LOC387930 FLJ40296
1	51,787,699	52,131,576	343.88	0	15	15	0.041522	NRD1 OSBPL9 EPS15
2	195,105,186	195,239,425	134.24	990.35	6	3	0.041839	SLC39A10
3	111,862,930	112,636,191	773.26	0	6	3	0.041839	PVRL3 CD96
5	21,754,685	22,196,981	442.30	0	6	3	0.041839	CDH12 PMCHL1
5	43,419,259	44,427,899	1008.64	0	6	3	0.041839	PAIP1 NNT FLJ32363 FLJ21657 FGF10 CCL28
5	87,118,032	88,222,225	1104.19	0	6	3	0.041839	MGC33214 MEF2C CCNH
11	47,193,935	47,329,450	135.52	0	6	3	0.041839	MADD MYBPC3 NR1H3 ACP2 DDB2
16	69,199,938	69,251,501	51.56	0	6	3	0.041839	MGC34647
16	70,454,745	70,783,287	328.54	0	6	3	0.041839	KIAA0174 DHODH PKD1L3 TXNL4B PMFBP1 DHX38 LOC55565 HPR HP
19	22,603,507	22,867,619	264.11	0	6	3	0.041839	ZNF676 ZNF91
3	97,926,792	98,847,728	920.94	0	10	8	0.042171	EPHA6
10	104,441,749	104,528,601	86.85	0	10	8	0.042171	SFXN2 ARL3 C10orf26
17	37,820,177	38,800,671	980.49	0	10	8	0.042171	ATP6V0A1 NAGLU RAMP2 EZH1 AARSD1 RUNDC1 NBR1 TMEM106A VAT1 TUBG1 PTRF LOC162427 CNTNAP1 BECN1 BRCA1 IFI35 TUBG2 RPL27 PSME3 NBR2 PLEKHH3 ARL4D G6PC COASY FLJ31222 RND2 HSD17B1 CCR10 CNTD AOC2 LOC90586 AOC3 WNK4 TBPIP MLX VPS25 CCDC56
8	51,753,354	52,121,962	368.61	0	24	29	0.04328	SNTG1
1	188,025,395	188,170,120	144.73	163.30	5	2	0.044799	FAM5C
2	56,560,262	57,572,231	1011.97	555.83	5	2	0.044799	EFEMP1 VRK2
3	139,253,842	140,113,284	859.44	0	5	2	0.044799	DBR1 PIK3CB ARMC8 FAM62C A4GNT DZIP1L CEP70 MRAS FOXL2 FAIM TXNDC6

7	85,404,705	85,971,962	567.26	139.20	5	2	0.044799	GRM3
7	125,261,168	125,379,474	118.31	486.41	5	2	0.044799	GRM8
11	58,216,649	58,433,650	217.00	0	5	2	0.044799	GLYATL2 GLYAT GLYATL1
12	42,267,835	42,280,252	12.42	35.84	5	2	0.044799	ADAMTS20
15	46,897,611	47,030,097	132.49	0	5	2	0.044799	SHC4 CRI1
15	49,535,902	50,134,220	598.32	0	5	2	0.044799	SCG3 MAPK6 LEO1 TMOD3 DMXL2 TMOD2 LYSMD2
20	32,078,884	33,184,387	1105.50	0	5	2	0.044799	ASIP TP53INP2 ACSS2 EIF2S2 NCOA6 C20orf31 RALY ITCH MAP1LC3A TRPC4AP DYNLRB1 GGTL3 GSS CDC91L1 MYH7B AHCY HMG4L FLJ38773
1	120,147,172	121,186,536	1039.36	0	4	1	0.044933	NOTCH2 LOC440607 ADAM30 REG4
2	95,559,457	96,728,940	1169.48	0	4	1	0.044933	ARID5A FLJ10081 TRIM43 LOC400986 LOC150763 ADRA2B BRRN1 ASTL DUSP2 WDR39 STARD7 ASCC3L1 TMEM127 KIAA1754L LMAN2L
2	131,390,755	131,876,316	485.56	0	4	1	0.044933	PLEKHB2 POTE2 LOC130074 ARHGEF4
2	186,580,593	187,582,053	1001.46	0	4	1	0.044933	ZSWIM2 ITGAV LEREPO4 FLJ44048 KIAA1946
3	121,788,190	122,066,323	278.13	0	4	1	0.044933	NDUFB4 HGD RABL3 GTF2E1
4	9,040,539	10,033,421	992.88	0	4	1	0.044933	SLC2A9 WDR1 KIAA1729 DUB4 DRD5
4	143,118,002	143,706,064	588.06	0	4	1	0.044933	INPP4B
4	167,936,695	168,637,578	700.88	0	4	1	0.044933	SPOCK3
5	49,497,985	50,531,671	1033.69	0	4	1	0.044933	EMB PARP8 ISL1
6	81,639,989	82,562,424	922.44	0	4	1	0.044933	FAM46A BCKDHB
6	120,680,743	121,198,990	518.25	243.34	4	1	0.044933	C6orf170
8	110,830,952	111,049,047	218.10	0	4	1	0.044933	FLJ20366 KCNV1
12	121,778,693	121,962,514	183.82	0	4	1	0.044933	DENR VPS37B CCDC62 HIP1R GPR81
16	68,099,877	68,663,700	563.82	0	4	1	0.044933	NFAT5 NQO1 WWP2 LOC348174 NOB1P PDPR CYB5-M
16	69,459,011	69,500,701	41.69	0	4	1	0.044933	VAC14
17	21,679,693	22,898,583	1218.89	0	4	1	0.044933	FAM27L WSB1 KSR1
17	42,306,776	43,231,021	924.25	0	4	1	0.044933	GOSR2 CDC27 C17orf57 NPEPPS RPRML ITGB3 TBX21 MYL4 KPNB1 OSBPL7 WNT9B TBKBP1
17	58,273,040	59,350,157	1077.12	0	4	1	0.044933	CYB561 CCDC44 LYK5 RNF190 LIMD2 DDX42 ACE MAP3K3 PSMC5 WDR68 CSH1 CCDC47 KCNH6 CSH2 GH2 SMARCD2 FTSJ3 CSHL1 GH1
2	94,739,002	95,785,152	1046.15	0	16	17	0.048658	MAL MRPS5 TEKT4 FAHD2A TRIM43 CSEN ZNF514 ZNF2 PROM2 LOC400986

List of rHHs for replication stage 2 population cluster 5. The number of ASD probands and parental controls is 410 and 820 respectively.

Table 1i Stage 2 Cluster 6

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
3	85,259,739	85,918,597	658.86	0	4	0	0.011439	CADM2
14	105,041,834	106,040,558	998.72	0	3	0	0.035663	TMEM121 LOC652848 GeneID:28445 GeneID:28429 KIAA0125 GeneID:28505 GeneID:3502 GeneID:28467 GeneID:28455 C14orf80 GeneID:3501 GeneID:3500 GeneID:28394 GeneID:3493 GeneID:28426 GeneID:3503 GeneID:28452 GeneID:28472 GeneID:28449 GeneID:28451 GeneID:28447 GeneID:3494 GeneID:28466 GeneID:28473 GeneID:3497 GeneID:28432 GeneID:28468 GeneID:28444 GeneID:28395 GeneID:28465 GeneID:28450 GeneID:28493 GeneID:28496 GeneID:28439 GeneID:28475 GeneID:3495 GeneID:28457 GeneID:28400 GeneID:28474 GeneID:3507
1	51,584,518	52,515,082	930.56	0	3	0	0.035663	EPS15 NRD1 RAB3B OSBPL9 BTF3L4 ZFYVE9 TXNDC12 KTI12
16	31,052,720	31,068,872	16.15	0	3	0	0.035663	PRSS8 PRSS36
13	57,143,152	57,611,859	468.71	0	3	0	0.035663	PCDH17
4	148,562,367	148,683,852	121.49	0	3	0	0.035663	EDNRA
8	85,947,968	86,879,277	931.31	0	5	2	0.04169	CA13 REXO1L1 LRRCC1 E2F5 LOC138046 CA2 CA3 CA1
2	88,337,248	88,428,892	91.64	79.47	5	2	0.04169	FLJ10916
7	98,629,828	99,297,192	667.36	0	4	1	0.042715	CYP3A5 CYP3A7 CYP3A4 ZNF655 PTCD1 G10 ZNF498 ARPC1B ZFP95 CPSF4 CYP3A43 ZNF394 LOC285989 DKFZp727G131 ARPC1A MYH16 PDAP1 ATP5J2 SMURF1
14	40,742,945	40,785,318	42.37	361.78	4	1	0.042715	LRFN5

List of rHHs for replication stage 2 population cluster 6. The number of ASD probands and parental controls is 54 and 108 respectively.

Table 1j Stage 2 Cluster 7

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
17	37,591,382	38,306,940	715.56	0	8	1	0.000732	STAT5B STAT3 ATP6V0A1 NAGLU RAMP2 EZH1 TUBG1 PTRF LOC162427 CNTNAP1 BECN1 TUBG2 PSME3 LGP1 PLEKHH3 STAT5A COASY FLJ31222 HSD17B1 CCR10 CNTD AOC2 LOC90586 AOC3 WNK4 TBPIP MLX VPS25 HCRT G6PC CCDC56
11	48,114,445	49,119,524	1005.08	0	4	0	0.011567	PTPRJ FOLH1 OR4A47 OR4B1 OR4X2 OR4X1 OR4C3 OR4S1
6	145,776,621	146,666,727	890.11	0	7	3	0.016987	SHPRH EPM2A GRM1 FBXO30 FLJ44955
5	42,102,114	42,573,041	470.93	0	6	2	0.017515	GHR FBXO4
11	47,326,019	47,327,183	1.16	0	6	2	0.017515	MYBPC3
7	62,754,013	62,793,854	39.84	0	7	4	0.034202	LOC401354
16	45,668,620	46,564,588	895.97	0	10	8	0.035653	PHKB CDA08 NETO2 ABCC12
3	46,732,266	46,942,044	209.78	0	3	0	0.035859	TESSP5 TESSP2 MYL3 PTHR1 CCDC12 TSP50
7	56,800,546	57,182,269	381.72	0	3	0	0.035859	CHCHD2 LOC441233 LOC401357
11	47,321,775	47,327,183	5.41	0	3	0	0.035859	MYBPC3
3	50,352,627	51,370,611	1017.98	0	14	14	0.037333	DOCK3 C3orf18 TMEM115 CISH CACNA2D2 MAPKAPK3 ZMYND10 TUSC4 HEMK1 CYB561D2
1	50,726,686	50,900,112	173.43	0	5	2	0.042204	FAF1
8	33,667,649	34,118,720	451.07	90.67	5	2	0.042204	DUSP26
11	47,337,169	48,093,566	756.40	0	5	2	0.042204	AGBL2 FNBP4 C1QTNF4 CUGBP1 PTPRJ RAPSN NDUFS3 SPI1 PSMC3 MTCH2 NUP160 SLC39A13 KBTBD4

List of rHHs for replication stage 2 population cluster 7. The number of ASD probands and parental controls is 63 and 126 respectively. Population cluster 7 was not considered in the replication study due to the small sample size (< 50 probands) of population cluster 7 in the discovery analysis.

Table 1k Stage 2 Cluster 9

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
3	161,697,163	162,052,889	355.73	0	4	0	0.011716	KPNA4 PPM1L ARL14
5	100,001,929	100,173,287	171.36	0	8	4	0.016247	UNQ1912 ST8SIA4
5	87,118,032	87,867,883	749.85	0	5	1	0.016609	MGC33214 CCNH MEF2C
6	126,531,019	127,087,814	556.80	0	7	3	0.017496	C6orf173 C6orf75
13	88,064,459	88,888,381	823.92	934.59	3	0	0.036086	SLITRK5
5	44,429,326	45,501,683	1072.36	0	3	0	0.036086	HCN1 FGF10 LOC441070 MRPS30
20	31,929,715	33,238,861	1309.15	0	3	0	0.036086	RALY ASIP TP53INP2 ACSS2 EIF2S2 NCOA6 C20orf31 ITCH MAP1LC3A TRPC4AP DYNLRB1 GGTL3 GSS CDC91L1 MYH7B AHCY HMG4L CHMP4B PROCR FLJ38773
6	102,967,356	103,611,156	643.80	342.71	3	0	0.036086	GRIK2
3	161,697,163	162,422,511	725.35	0	3	0	0.036086	KPNA4 PPM1L NMD3 B3GALT3 ARL14
14	44,823,687	44,981,331	157.64	31.54	3	0	0.036086	C14orf106
13	56,610,488	57,143,152	532.66	0	3	0	0.036086	FLJ40296 PCDH17
12	38,572,369	39,885,270	1312.90	0	3	0	0.036086	LRRK2 CNTN1 SLC2A13 C12orf40
8	35,009,528	35,499,150	489.62	0	3	0	0.036086	UNC5D
12	42,645,806	43,289,841	644.04	0	3	0	0.036086	TMEM117 NELL2
10	50,645,864	51,696,046	1050.18	0	3	0	0.036086	PARG MSMB TIMM23 LOC387680 ASAH2 NCOA4 OGDHL
1	101,173,046	101,504,049	331.00	0	3	0	0.036086	EDG1 SLC30A7 DPH5
2	197,840,817	198,757,353	916.54	0	5	2	0.042795	SF3B1 PREI3 BOLL C2orf11 PLCL1 HSPD1 COQ10B ANKRD44 HSPE1 MARS2
6	57,407,051	58,085,974	678.92	0	4	1	0.043503	PRIM2A GUSBL2
17	37,858,625	38,866,376	1007.75	0	4	1	0.043503	ATP6V0A1 NAGLU RAMP2 EZH1 AARSD1 RUNDC1 NBR1 TMEM106A VAT1 TUBG1 LOC162427 CNTNAP1 BECN1 BRCA1 IFI35 ARL4D TUBG2 RPL27 PSME3 NBR2 PLEKHH3 G6PC COASY FLJ31222 RND2 HSD17B1 CCR10 CNTD AOC2 LOC90586 AOC3 WNK4 TBPIP MLX VPS25 CCDC56
8	35,415,415	35,499,150	83.74	0	4	1	0.043503	UNC5D
6	118,677,207	119,499,535	822.33	0	4	1	0.043503	SLC35F1 C6orf204 MCMDC1 ASF1A C6orf60 PLN MAN1A1

List of rHHs for replication stage 2 population cluster 9. The number of ASD probands and parental controls is 78 and 156 respectively. The individuals in population cluster 9 are of Costa Rican origin. There was no corresponding population cluster in the discovery analysis.

Table 1l Stage 1 Clusters with <50 probands

Stage 1 Cluster 1 : List of rHHs for stage 1 population cluster 1 (# probands = 27, # parental controls = 54)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
1	49,153,491	50,362,385	1208.89	0	5	1	0.01632	AGBL4 ELAVL4
7	98,700,955	99,222,904	521.95	0	4	1	0.043081	CYP3A5 CYP3A7 CYP3A4 ZNF655 PTCD1 G10 ZNF498 ARPC1B ZFP95 CPSF4 ZNF394 LOC285989 DKFZp727G131 ARPC1A PDAP1 MYH16 ATP5J2

Stage 1 Cluster 7 : List of rHHs for stage 1 population cluster 7 (# probands = 34, # parental controls = 68)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
14	104,868,710	106,019,705	1151.00	0	4	0	0.010913	TMEM121 LOC652848 GeneID:28445 MTA1 GeneID:28429 PACS2 CRIP1 KIAA0125 GeneID:28505 GeneID:3502 GeneID:28467 GeneID:28455 C14orf80 GeneID:3501 GeneID:3500 GeneID:28394 GeneID:3493 GeneID:28426 GeneID:3503 GeneID:28452 GeneID:28472 GeneID:28449 GeneID:28451 GeneID:28447 GeneID:3494 GeneID:28473 GeneID:3497 GeneID:28432 GeneID:28468 GeneID:28444 GeneID:28395 GeneID:28450 GeneID:28493 GeneID:28496 GeneID:28439 GeneID:28475 GeneID:3495 CRIP2 GeneID:28457 GeneID:28466 GeneID:28400 GeneID:28474 GeneID:3507
3	51,956,061	51,998,060	42.00	0	8	5	0.025579	ABHD14A PARP3 GPR62 PCBP4 ABHD14B ACY1

Stage 1 Cluster 8 : List of rHHs for stage 1 population cluster 8 (# probands = 43, # parental controls = 86)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
5	130,413,182	131,397,140	983.96	0	7	3	0.017822	KIAA1961 RAPGEF6 CDC42SE2 HINT1 LOC90624 ACSL6
16	31,537,650	31,559,652	22.00	0	15	15	0.034721	ZNF720 MGC3020
15	80,364,060	81,769,606	1405.55	0	3	0	0.036231	HDGFRP3 DKFZp666G057 FLJ22795 BNC1 BTBD1 CPEB1 TM6SF1 AP3B2 FSD2 HOMER2 FAM103A1 C15orf40 RPS17 LOC440295
5	138,011,435	138,307,955	296.52	0	3	0	0.036231	SIL1 CTNNA1 HSPA9B LRRTM2
22	26,950,996	27,786,733	835.74	0	3	0	0.036231	XBP1 CHEK2 PITPNB HS747E2A FLJ33814 HSC20
15	80,364,060	81,769,606	1405.55	0	3	0	0.036231	HDGFRP3 DKFZp666G057 FLJ22795 BNC1 BTBD1 CPEB1 TM6SF1 AP3B2 FSD2 HOMER2 FAM103A1 C15orf40 RPS17 LOC440295
14	39,607,151	40,425,223	818.07	635.78	3	0	0.036231	FBXO33 LRFN5
22	26,950,996	27,583,364	632.37	0	6	3	0.039816	CHEK2 PITPNB XBP1 FLJ33814 HSC20
3	80,270,653	80,370,296	99.64	1541.57	4	1	0.043772	ROBO1
6	34,659,064	35,503,982	844.92	0	4	1	0.043772	C6orf107 TCP11 PPARD SCUBE3 DEF6 ZNF76 TAF11 ANKS1A SNRPC C6orf106
22	27,870,297	27,942,508	72.21	0	4	1	0.043772	EMID1 KREMEN1

Stage 1 Cluster 10 : List of rHHs for stage 1 population cluster 10 (# probands = 28, # parental controls = 58)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
3	52,183,938	52,199,565	15.63	7.59	8	4	0.010168	ALAS1
9	30,047,012	30,925,448	878.44	1149.15	3	0	0.032011	ACO1 LRRN6C
3	48,638,050	48,696,044	57.99	0	7	5	0.045651	NCKIPSD CELSR3 SLC26A6

List of rHHs for the four stage 1 population clusters (C1, C7, C8, C10) that had <50 probands

.

Table 1m Stage 2 Clusters with <50 probands

Stage 2 Cluster 1 : List of rHHs for stage 2 population cluster 1 (# probands = 36, # parental controls = 72)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
13	54,865,754	54,903,571	37.82	0	5	1	0.015203	LOC387930
8	35,785,310	35,816,890	31.58	13.59	3	0	0.034973	UNC5D
8	35,521,243	36,550,989	1029.75	0	3	0	0.034973	FKSG2 UNC5D
15	42,161,179	42,361,941	200.76	0	3	0	0.034973	FRMD5 CASC4
19	23,389,292	23,541,400	152.11	19.18	4	1	0.04143	ZNF91 ZNF675
13	54,865,754	55,710,950	845.20	902.10	4	1	0.04143	FLJ40296 LOC387930

Stage 2 Cluster 2 : List of rHHs for stage 2 population cluster 1 (# probands = 49, # parental controls = 100)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
7	71,579,375	72,522,124	942.75	0	3	0	0.034101	LOC441251 POM121 LOC541473 NSUN5C LOC442582 LOC441257 LOC389517 TRIM50C NSUN5 CALN1 FKBP6 BAZ1B LOC442578 SBDSP FZD9 TRIM50A GTF2IRD2P
7	62,175,915	62,793,854	617.94	532.43	3	0	0.034101	ZNF679 LOC401354
19	11,332,074	11,535,294	203.22	0	3	0	0.034101	RGL3 CNN1 MGC20983 ELOF1 ELAVL3 PRKCSH SITPEC LPPR2 EPOR ZNF653 FLJ35119
1	188,689,667	189,140,954	451.29	0	4	1	0.040383	FAM5C

Stage 2 Cluster 8 : List of rHHs for stage 2 population cluster 1 (# probands = 49, # parental controls = 98)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
19	22,521,422	23,151,535	630.11	0	4	45	0.011347	ZNF91 ZNF676
7	73,638,615	74,882,758	1244.14	0	19	38	0.019784	GTF2IRD1 GTF2I GTF2IRD2B NCF1 WBSCR16 LOC441257 PMS2L2 GTF2IRD2 LOC442582 PMS2L5 DKFZP434A0131 LOC541473 WBSCR20B
2	193,931,779	194,842,340	910.56	1163.89	3	46	0.035522	TMEFF2 SLC39A10
10	104,199,067	104,203,492	4.43	0	3	46	0.035522	C10orf95
11	64,924,857	65,129,156	204.30	0	3	46	0.035522	FKSG44 MAP3K11 KCNK7 MALAT1 LTBP3 FAM89B SCYL1 SSSCA1 TncRNA
5	44,622,918	44,663,855	40.94	198.37	3	46	0.035522	FGF10 MRPS30

7	62,754,013	62,893,151	139.14	0	9	43	0.060389	LOC401354
12	37,919,825	38,305,515	385.69	0	5	45	0.042453	KIF21A C12orf40 ABCD2

Stage 2 Cluster 10 : List of rHHs for stage 2 population cluster 1 (# probands = 44, # parental controls = 86)

Chr	Start	End	Length (kb)	Distance to gene (kb)	ASD probands	Parental controls	RPV	Genes
22	26,647,401	27,583,364	935.96	0	4	0	0.011951	PITPNB CHEK2 XBP1 FLJ33814 HSC20
10	22,583,581	22,677,387	93.81	0	8	5	0.030435	COMMD3 PCGF4 SPAG6
3	163,721,879	164,724,724	1002.85	0	3	0	0.037019	C3orf57 SI
4	69,547,255	70,072,336	525.08	0	3	0	0.037019	UGT2B15 UGT2A3 UGT2B10 UGT2B7 UGT2B11
2	126,059,795	126,212,264	152.47	670.47	4	1	0.044579	CNTNAP5

List of rHHs for the four stage2 population clusters (C1, C2, C8, C10) that had <50 probands

Supplementary Table 2 EIGENSOFT principal component statistics

Principal component	Percentage of explanation	Tracy-Widom statistic
1	9.14	0
2	5.36	0
3	2.24	0
4	1.72	0
5	1.62	3.51×10^{-319}
6	1.57	1.30×10^{-184}
7	1.49	1.26×10^{-23}
8	1.48	1.97×10^{-16}
9	1.47	2.48×10^{-6}
10	1.47	7.89×10^{-5}
11	1.46	0.00054

EIGENSOFT uses Tracy-Widom statistics to determine the significance of each principal component (PC) identified by principal component analysis (PCA) (Patterson et al. 2006). The transition phase occurs between PC8 and PC9 indicating that the first 8 PCs are significantly contributing to the population structure of the AGP sample set. Accordingly, eigen values from the first 8 PCs were used for Hopach hierarchical clustering to assign individuals to specific population groups.

Supplementary Table 3 The highest and lowest ranking F_{ST} values

Hopach Cluster A	Hopach Cluster B	F_{ST} value
D	E	5.06E-07
E	G	5.66E-06
C	I	2.84E-05
N	O	3.55E-05
I	L	4.73E-05
G	K	6.15E-05
G	L	6.57E-05
M	N	7.04E-05
E	F	7.44E-05
I	K	8.09E-05
F	G	1.01E-04
E	I	1.04E-04
E	K	1.04E-04
J	K	1.05E-04
J	L	1.09E-04
E	L	1.12E-04
K	L	1.31E-04
I	J	1.39E-04
E	J	1.55E-04
G	I	1.94E-04
D	L	1.96E-04
F	I	1.97E-04
M	O	1.98E-04
G	J	2.08E-04
F	J	2.24E-04
D	I	2.28E-04
B	J	2.29E-04

D	K	2.50E-04
D	G	2.67E-04
F	K	2.74E-04
F	L	3.09E-04
B	F	3.30E-04
D	F	3.49E-04
D	J	3.74E-04
H	I	3.97E-04
H	Q	4.04E-04
B	G	4.13E-04
…		
R	S	1.81E-02
B	R	1.84E-02

Hopach Cluster refers to the 19 provisional clusters yielded by the Hopach algorithm. Clusters 1-19 are labelled A-S respectively. To assess the similarity of the clusters we calculated pair-wise genetic distance (F_{ST}) values of every cluster against all other clusters. The R package hierfstat was used to calculate the F_{ST} metrics using 5,000 SNPs randomly chosen from the panel of 70,175 SNPs used for PCA (Goudet 2005). Cluster R shows the highest degree of dissimilarity with all other clusters. We identified three groups of clusters with low F_{ST} values ($F_{st} < 1E^{-04}$) and of high similarity; clusters [D,E,F,G], [C,I,J,K, L] and [M,N,O]. The three groups of highly similar clusters were collapsed and a medoid selected from each merged group. The 10 medoids from clusters A, B, [D,E,F,G], H, [C,I,J,K,L], [M,N,O], P, Q, R and S were used to re-cluster all samples by Hopach hierarchical clustering. During this process each sample was assigned to one of the 10 clusters, based on similarity to the medoid.

Supplementayry Table 4 192 genes located in rHH in two or more population groups

Gene	C2	C3	C4	C5	C6	Total
CADM2	x	x^R		x	x^R	4
ABHD14A		x^R		x	x	3
ABHD14B		x^R		x	x	3
ASL			x	x	x	3
C10orf95	●			x^R	●	3
GBE1	x	x^R			x	3
GPR62		x^R		x	x	3
GUSB			x	x	x	3
IQCF1		x		x	x	3
KHDRBS2		x		x	x	3
LOC441242			x	x	x	3
PARP3		x^R		x	x	3
PCBP4		x^R		x	x	3
PDZK1		x	●	x		3
RNU3IP2		x^R		x	x	3
VKORC1L1			x	x	x	3
ABT1	x				x	2
ACY1		x^R		x		2
ANKRD7		x			x	2
AP3B2		x		x		2
ARL14	x				x	2
B3GALT3	x				x	2
BTN1A1	x				x	2
BTN2A1	x				x	2
BTN2A2	x				x	2
BTN2A3	x				x	2
BTN3A1	x				x	2
BTN3A2	x				x	2
BTN3A3	x				x	2
C12orf50	●	x^R				2
C14orf135			x	x		2
C14orf39			x	x		2
C15orf17				x	x	2
C15orf39				x	x	2
C6orf162		x		x		2
C6orf163		x		x		2
C6orf165		x		x		2
CAMP	x				●	2
CART1		x^R		x^R		2
CCDC63	●	x^R				2
CD160		x^R		x		2
CDC25A	x				●	2
CGA		x		x		2
CHORDC1		x		x		2
CHRFAM7A	x				x	2
COX5A				x	x	2
CPEB1		x		x		2
CPLX3				x	x	2
CSK				x	x	2

CUEDC2	●			x[R]		2
CYP1A2				x	x	2
DEFA5	x				x	2
DEFB103A	x				x	2
DEFB104A	x				x	2
DEFB105A	x				x	2
DEFB105B	x				x	2
DEFB106A	x				x	2
DEFB109	x				x	2
DEFB4	x				x	2
DHRS7			x	x		2
DKFZp666G057		x		x		2
DMTF1	x			x		2
DUSP26		x[R]			x	2
DUSP7		x[R]		x		2
EHBP1				x	x	2
ELAVL4	x	x[R]				2
EPHA3		x[R]			x	2
FAM7A2	x				x	2
FBXW12	x				●	2
FGF10		●[R]			x	2
FKSG83	x				x	2
FLJ22795		x		x		2
FLJ46156			x	x		2
FOXP2	●			x		2
FSD2		x		x		2
GJB7		x		x		2
GLULD1		x[R]		x		2
GRIK2				x	x	2
GRM3	x			x[R]		2
GUSBL1	x				x	2
GUSBL2	x	x				2
HIST1H2AG	x				x	2
HIST1H2AH	x				x	2
HIST1H2BJ	x				x	2
HIST1H2BK	x				x	2
HIST1H4I	x				x	2
HMGN4	x				x	2
HTR1E		x		x		2
IMMP2L		x[R]			x	2
ITGAX		x			x	2
KCND2				x[R]	x	2
KCTD7			x		x	2
KIAA1018	x				x	2
KIAA1324L	x			x		2
LMAN1L				x	x	2
LOC285908			x		x	2
LOC349196	x				x	2
LOC401447	x				x	2
LOC440295		x		x		2
LOC51057				x	x	2
LRFN5		x			●[R]	2

LRRC55				x	x	2
LRRN3		x			x	2
MGAT4C		x^R		x		2
MMP16		x			x	2
MNAT1			x	x		2
MPI				x	x	2
MYO18A				●	x	2
NEGR1		x^R			x	2
NMD3	x				x	2
NME6	x				●	2
NNT		x			x	2
NOX4		●			x	2
NTS		x^R		x		2
NUDT17		x^R		x		2
OR10AG1				x	x	2
OR5AP2				x	x	2
OR5AR1				x	x	2
OR5AS1				x	x	2
OR5D16				x	x	2
OR5D18				x	x	2
OR5F1				x	x	2
OR5I1				x	x	2
OR5J2				x	x	2
OR5L2				x	x	2
OR5M1				x	x	2
OR5M10				x	x	2
OR5M11				x	x	2
OR5M3				x	x	2
OR5M8				x	x	2
OR5M9				x	x	2
OR5R1				x	x	2
OR5T1				x	x	2
OR5T2				x	x	2
OR5T3				x	x	2
OR5W2				x	x	2
OR8H1				x	x	2
OR8H2				x	x	2
OR8H3				x	x	2
OR8I2				x	x	2
OR8J1				x	x	2
OR8J3				x	x	2
OR8K1				x	x	2
OR8K3				x	x	2
OR8K5				x	x	2
OR8U8				x	x	2
OR9G1				x	x	2
OR9G4				x	x	2
OTX1				x	x	2
PAMCI		x^R		x^R		2
PHF3		x^R		x		2
PIPOX				●	x	2
PJCG6	x				x	2

POLR3C		x[R]		x		2
PPCDC				x	x	2
PPM1A			x	x		2
PPM1L	x				x	2
PRIM2A	x	x				2
PRSS16	x				x	2
PSMD14		x[R]			x	2
PTP4A1		x[R]		x		2
RABGEF1			x		x	2
RCP9			x		x	2
ROBO1		x			x	2
RPL29		x[R]		x		2
RPP25				x	x	2
RPS17		x		x		2
SCAMP2				x	x	2
SCAMP5				x	x	2
SIX1			x	x		2
SIX4			x	x		2
SIX6			x	x		2
SLC38A6			x	x		2
SLC4A10		x[R]			x	2
SLITRK1		x		x[R]		2
SPAG11	x				x	2
SRD5A2L2		x			x	2
SYNCRIP		x		x		2
TANK		x[R]			x	2
TBR1		x[R]			x	2
TJP1	x				x	2
TPST1			x		x	2
TRIM51				x	x	2
TRMT5			x	x		2
ULK3				x	x	2
UNC5D		x[R]			x	2
WDR51A		x[R]		x		2
ZNF322A	x				x	2
ZNF364		x[R]		x		2
ZNF589	x				●	2
ZNF679				x[R]	x	2
ZNF92				x	●	2

We found that 192 genes are located in rHH regions in 2 or more population clusters. An 'x' indicates that the gene was located in a homozygous haplotype that was significantly more common in ASD probands compared to parental controls in the population cluster in question (5% significance threshold). Any rHH that contains <10 tagging SNPs (Haploview CEU data) are marked with a ●. Genes that occurred in a rHH in both the discovery analysis and the ancestry-matched stage 2 replication study are denoted with [R]. Note that cluster 2 did not have a sufficient sample size for the replication study.

Supplementary Table 5 Candidate genes previously implicated in ASD

Gene	Band	Reference
Association		
PTGS2	1q25.2-q25.3	Yoo et al. 2008
MARK1	1q41	Maussion et al. 2008
DISC1	1q42.1	Kilpinen et al. 2008
CENTG2	2p24.3-p24.1	Wassink et al. 2005
NPAS2	2q11.2	Nicholas et al. 2007
SLC25A12	2q24	Ramoz et al. 2004; Ramoz et al. 2008
FAM130A2	2q24.3	Maestrini et al. 2009
NOSTRIN	2q24.3	Maestrini et al. 2009
STK39	2q24.3	Ramoz et al. 2008
ZNF533	2q31.2-q31.3	Maestrini et al. 2009
ITGA4	2q31.3	Correia et al. 2009; Conroy et al. 2009
DLX1	2q32	Liu et al. 2009
DLX2	2q32	Liu et al. 2009
NRP2	2q33.3	Wu et al. 2007
OXTR	3p25	Wu et al. 2005
DRD3	3q13.3	de Krom et al. 2009
HTR3C	3q27.1	Rehnstrom et al. 2009
EGF	4q25	Toyoda et al. 2007
TDO2	4q31-q32	Nabi et al. 2004
CDH10	5p14.1	Wang et al. 2009
PRLR	5p14-p13	Yrigollen et al. 2008
PITX1	5q31	Philippi et al. 2007
F13A1	6p25.3-p24.3	Hu et al. 2006
JARID2	6p24-p23	Weiss et al. 2009
PRL	6p22.2-p21.3	Yrigollen et al. 2008
GLO1	6p21.3-p21.1	Junaid et al. 2004
HLA-A	6p21.3	Torres et al. 2006
HLA–DRB1	6p21	Torres et al. 2006; Warren et al. 1996
GRIK2	6q16.3-q21	Jamain et al. 2002; Shuang et al. 2004
LAMB1	7q22	Hutcheson et al. 2004
CUX1	7q22.1	Maestrini et al. 2009
LHFPL3	7q22.1	Maestrini et al. 2009
FOXP2	7q31	Gong et al. 2004
IMMP2L	7q31	Maestrini et al. 2009
MET	7q31	Campbell et al. 2006
WNT2	7q31	Wassink et al. 2001
DOCK4	7q31.1	Maestrini et al. 2009
LRRN3	7q31.1	Maestrini et al. 2009
NRCAM	7q31.1-q31.2	Bonora et al. 2005; Marui et al 2009
TSPAN12	7q31.31	Maestrini et al. 2009
FEZF1	7q31.32	Maestrini et al. 2009
GRM8	7q31.3-q32.1	Serajee et al. 2003

SLC13A1	7q31-q32	Maestrini et al. 2009
UBE2H	7q32	Vourc'h et al. 2003
SMO	7q32.3	Maestrini et al 2009
PLEXNA4	7q32.3	Maestrini et al. 2009
CNTNAP2	7q35-q36	Arking et al. 2008; Alarcon et al. 2008
EN2	7q36	Cheh et al. 2006; Benayed et al. 2005; Gharani et al. 2004
DBH	9q34	Robinson et al. 2001
CTNNA3	10q22.2	Weiss et al. 2009
GAS2	11p14.3-p15.2	Weiss et al. 2009
BDNF	11p13	Nishimura et al. 2007; Philippe et al. 2002
HTR3A	11q23.1	Anderson et al. 2009
AVPR1A	12q14-q15	Wassink et al. 2004;Yirmiya et al. 2006
DAO	12q24	Chung et al. 2007
FBXO33	14q21.1	Wang et al. 2009
LRFN5	14q21.2	Wang et al. 2009
GABRA5	15q11.2-q12	Menold et al. 2001
GABRB3	15q11.2-q12	Buxbaum et al. 2002; McCauley et al. 2004
UBE3A	15q11-q13	Nurmi et al. 2001
ABAT	16p13.2	Barnby et al. 2005
PRKCB1	16p11.2	Philippi et al. 2005;Lintas et al. 2009
PER1	17p13.1-17p12	Nicholas et al. 2007
SLC6A4	17q11.1-q12	Cook et al. 1997; Yirmiya et al. 2001; Kim et al. 2002; Sutcliffe et al. 2005
OMG	17q11.2	Vourc'h et al. 2003
NOS2A	17q11.2-q12	Kim et al. 2009
HOXB1	17q21.3	Ingram et al. 2000
PLAUR	19q13	Campbell et al.2008
MIF	22q11.23	Grigorenko et al. 2008
ADA	20q12-q13.11	Bottini et al. 2001
Copy number variation and rearrangements		
CLIC4	1p36.11	Castermans et al. 2003
DPYD	1p22	Marshall et al. 2008
NRXN1	2p16.3	Szatmari et al. 2007
SLC4A10	2q23-q24	Sebat et al. 2007
SCN7A	2q24.3	Morrow et al. 2008
WDR75	2q32.2	Conroy et al. 2008
SUMF1	3p26.2	Glessner et al. 2009
CNTN4	3p26-p25	Roohi et al. 2008
CNTN3	3p12.3	Morrow et al. 2008
C3orf58	3q24	Morrow et al. 2008
SLC9A9	3q24	Morrow et al. 2008
NLGN1	3q26.31	Glessner et al. 2009
GABRG1	4p12	Vincent et al. 2006
PCDH10	4q28.3	Morrow et al. 2008
RNF8	6p21.2	Morrow et al. 2008
PARK2	6q25.2-q27	Glessner et al. 2009
UPK3B	7q11.2	Edelmann et al. 2007

ELN	7q11.23	Herguner and Mukaddes 2006
GTF2I	7q11.23	Depienne et al. 2007
GTF2IRD1	7q11.23	Edelmann et al. 2006
CFTR	7q31.2	Warburton et al. 2000
CADPS2	7q31.3	Sadakata et al. 2007
CREB3L2	7q34	Rossi et al. 2008
GIMAP2	7q36.1	Rossi et al. 2008
CSMD3	8q23.3	Floris et al. 2008
GATA3	10p15	Verri et al. 2004
JMJD1C	10q21.2	Castermans et al. 2007
REEP3	10q21.3	Castermans et al. 2007
GRID1	10q22	Glessner et al. 2009
DCAMKL1	13q13	Reddy 2005
NBEA	13q13	Castermans et al. 2003
SNRPN	15q11.2	Sabry & Farag 1998; Hou & Wang 1998; Wang et al. 2008
HERC2	15q13	Filipek et al. 2003
PTPN9	15q24.2	Smith et al. 2000
A2BP1	16p13.3	Sebat et al. 2007
CSNK1D	17q25	Vazna et al. 2007
DPH1	17q25	Vazna et al. 2007
HIRA	22q11.2	Michaelis et al. 1998; Carratala et al. 1998
SHANK3	22q13.3	Moessner et al. 2008; Durand et al. 2007
ACR	22q13.33	Anderlid et al. 2002
Expression		
ARID1A	1p35.3	Baron et al. 2006
L1TD1	1p31.3	Hu et al. 2006
CHRNB2	1q21.3	Martin-Ruiz et al. 2004
TLK1	2q31.1	Baron et al. 2006
LANCL1	2q33-q35	Baron et al. 2006
SERPINE2	2q33-q35	Purcell et al. 2001
GPR55	2q37	Hu et al. 2006
CHL1	3p26.1	Hu et al. 2006
ALAS1	3p21.1	Purcell et al. 2001
FAM107A	3p21.1	Purcell et al. 2001
CCR1	3p21	Purcell et al. 2001
MITF	3p14.2-p14.1	Purcell et al. 2001
ROBO1	3p12	Anitha et al. 2008
GBE1	3p12.3	Baron et al. 2006
ROBO2	3p12.3	Anitha et al. 2008
PLA1A	3q13.13-q13.2	Hu et al. 2006
CHST2	3q24	Hu et al. 2006
TIPARP	3q25.31	Baron et al. 2006
SCHIP1	3q25.33	Hu et al. 2006
CD38	4p15	Hu et al. 2006
GABRA2	4p12	Fatemi et al. 2009
PDLIM5	4q22	Purcell et al. 2001

SPARCL1	4q22.1	Purcell et al. 2001
SLC1A3	5p13	Purcell et al. 2001
IL6ST	5q11	Hu et al. 2006
SKP1A	5q31	Baron et al. 2006
CD74	5q32	Baron et al. 2006
GRIA1	5q33	Purcell et al. 2001
GABRA1	5q34-q35	Fatemi et al. 2009
HLA-G	6p21.3	Purcell et al. 2001
HSPA1A	6p21.3	Purcell et al. 2001
BTN2A1	6p22.1	Baron et al. 2006
CD83	6p23	Hu et al. 2006
PCDHA4	6q14–q15	Purcell et al. 2001
TRAF3IP2	6q21	Baron et al. 2006
PTPRK	6q22.2-23.1	Hu et al. 2006
STX1A	7q11.23	Nakamura et al. 2008
CYP3A5P2	7q21.3-q22.1	Purcell et al. 2001
GNB2	7q21.3-q22.1	Baron et al. 2006
ACHE	7q22	Purcell et al. 2001
RELN	7q22	Fatemi et al. 2005
ATXN7L1	7q22.1	Hu et al. 2006
ZNHIT1	7q22.1	Baron et al. 2006
FNTA	8p22-q11	Hu et al. 2006
TNFRSF10B	8p22-p21	Baron et al. 2006
CLU	8p21-p12	Purcell et al. 2001
PAG1	8q21.13	Hu et al. 2006
CYP7B1	8q21.3	Hu et al. 2006
SIAHBP1	8q24.2–qter	Baron et al. 2006
SIT1	9p13-p12	Baron et al. 2006
STOM	9q34.1	Baron et al. 2006
ITGB1	10p11.2	Baron et al. 2006
EGR2	10q21.1	Hu et al. 2006
KIAA0913	10q22.2	Purcell et al. 2001
CD44	11p13	Baron et al. 2006
PRCP	11q14	Baron et al. 2009
CEP57	11q21	Baron et al. 2006
ELMOD1	11q22.3	Hu et al. 2006
ROBO3	11q24.2	Anitha et al. 2008
ROBO4	11q24.2	Anitha et al. 2008
ITGB7	12q13.13	Baron et al. 2006; Hu et al. 2006
PLA2G1B	12q23-q24.1	Purcell et al. 2001
ALOX5AP	13q12	Hu et al. 2006
FLT1	13q12	Hu et al. 2006
ZFYVE26	14q24.1	Purcell et al. 2001
IGHG3	14q32.33	Baron et al. 2006
HDC	15q21-q22	Purcell et al. 2001
PDE8A	15q25.3	Baron et al. 2006

IL32	16p13.3	Hu et al. 2006
GRIN2A	16p13.2	Barnby et al. 2005
GSPT1	16p13.1	Baron et al. 2006
SEPHS2	16p11.2	Purcell et al. 2001
COTL1	16q24.1	Hu et al. 2006
P2RX5	17p13.3	Hu et al. 2006
FLOT2	17q11-q12	Baron et al. 2006
STARD3	17q11-q12	Baron et al. 2006
GFAP	17q21	Purcell et al. 2001
NAGLU	17q21	Hu et al. 2006
CCL3L1	17q21.1	Hu et al. 2006
SSTR2	17q24	Purcell et al. 2001
GNA15	19p13.3	Baron et al. 2006
TPM4	19p13.1	Baron et al. 2006
ARHGEF1	19q13.13	Baron et al. 2006
APOE	19q13.2	Purcell et al. 2001
KCNN4	19q13.2	Baron et al. 2006
CHRNA4	20q13.2-q13.3	Martin-Ruiz et al. 2004
SAMSN1	21q11	Hu et al. 2006
CCT8	21q22.11	Baron et al. 2006
MORC3	21q22.13	Baron et al. 2006
IGLL3	22q11.2	Baron et al. 2006
NIPSNAP1	22q12.2	Baron et al. 2006
ACO2	22q13.2	Hu et al. 2006

Yang and colleagues published a list of candidate genes that were previously implicated in autism spectrum disorder (Yang and Gill 2007). The candidate gene list was extended to 2009 using the same criteria implemented by Yang and colleagues. The candidate genes are subdivided into three categories (association, copy number variation and expression) according to the type of analysis by which they were identified.

Supplementary Table 6 Cluster groups for stage 1 and stage 2 analyses

	Stage 1		Stage 2	
Cluster group	**Number of probands**	**Number of parental controls**	**Number of probands**	**Number of parental controls**
C1	27	54	36	72
C2	148	294	49	100
C3*	289	584	306	606
C4*	85	170	93	186
C5*	280	560	410	820
C6*	217	434	54	108
C7	34	68	63	126
C8	43	86	49	98
C9	0	0	78	156
C10	28	58	44	86

The number of individuals in each population cluster, divided into probands and the parents that formed the control group, are listed. The population clusters range in size from 0 to 289 probands. For replication studies, the Autism Genome Project (AGP) stage 2 samples were clustered with the AGP stage 1 samples to form ancestry-matched replication groups.The stage 2 data sets range in size from 36 to 410 probands. Sufficient replication groups (minimum 50 probands in stage 1 and stage 2) were available for 4 population clusters, as indicated with a *. Parental sample size is not always twice the case sample size due to the removal of samples during the quality control process.

Supplementary Table 7 Replication study for cluster 3 to cluster 6 using independent AGP trios

	Stage 1					Stage 2			
Gene	**C3**	**C4**	**C5**	**C6**		**C3**	**C4**	**C5**	**C6**
CADM2	x		x	x		x			x
ABHD14A	x		x	x		x			
ABHD14B	x		x	x		x			
CART1	x		x			x		x	
EPHA3	x			x		x		x	
FGF10	x			x		x		x	
GPR62	x		x	x		x			
IQCF1	x		x	x		x			
MGAT4C	x		x			x	x		
NEGR1	x			x		x		x	
PAMCI	x		x			x		x	
PARP3	x		x	x		x			
PCBP4	x		x	x		x			
RNU3IP2	x		x	x		x			
UNC5D	x			x		x		x	
ACSS2			x				x	x	
ACY1	x		x			x			
ADRA2B		x				x		x	
ANKRD7	x			x				x	
C10orf95			x	x				x	
C12orf50	x					x		x	
C14orf135		x	x			x			
CAMP				x		x		x	
CCDC63	x					x		x	
CD160	x		x			x			
CDC25A				x		x		x	
CDC91L1			x				x	x	
DUSP26	x			x		x			
DUSP7	x		x			x			
EHBP1			x	x		x			
FBXW12				x		x		x	
GBE1*	x			x		x			
GGTL3			x				x	x	
GLULD1	x		x			x			
GRIK2*			x	x		x			
GRM3			x			x		x	
GRM8*				x		x		x	
GSS			x				x	x	
HMG4L			x				x	x	
IMMP2L*	x			x		x			
ITGAX	x			x				x	
KCND2*			x	x				x	
LOC389289	x			x				x	
LOC400986		x				x		x	
LOC90835				x		x		x	
LRFN5*	x			x					x
MAP1LC3A			x				x	x	
MEF2C	x					x		x	

MYBPC3			x			x		x	
MYO18A			x	x		x			
NCOA6			x				x	x	
NME6				x		x		x	
NNT	x			x				x	
NOX4	x			x				x	
NTS	x		x			x			
NUDT17	x		x			x			
OTX1			x	x		x			
PHF3	x		x			x			
PIPOX			x	x		x			
POLR3C	x		x			x			
PRSS36				x				x	x
PRSS8				x				x	x
PSMD14	x			x		x			
PTP4A1	x		x			x			
RPL29	x		x			x			
SLC39A10			x				x	x	
SLC4A10*	x			x		x			
SLITRK1	x		x					x	
SNTG1	x					x		x	
TANK	x			x		x			
TP53INP2			x				x	x	
TRIM43		x				x		x	
ZNF676	x						x	x	
ZNF679			x	x				x	
ZNF91				x			x	x	
AARS			x					x	
ACADM			x			x			
ACO2*			x					x	
ACP6				x				x	
ACTR1A			x					x	
ADAMTS3			x			x			
AGBL4	x					x			
AHCY			x					x	
ALAS1	x					x			
AMY1C			x			x			
ANAPC7	x					x			
ANKDD1A			x			x			
ANKRD13	x					x			
ARL14				x				x	
ARMC5	x							x	
ARMET				x		x			
ARPC3	x					x			
ASB17			x			x			
ASCC3				x		x			
ASIP			x					x	
ATP2A2	x					x			
ATPBD1C	x					x			
ATRNL1			x			x			
B3GALT3				x				x	
BCKDK				x				x	

BCL7C				x				x	
C10orf77			x					x	
C12orf24	x					x			
C15orf20			x			x			
C16orf58	x							x	
C18orf26	x							x	
C18orf54	x							x	
C3orf57				x			x		
C8orf22	x					x			
CCL28	x							x	
COG4			x					x	
COL12A1			x			x			
COX6A2	x							x	
CRYBA1				x		x			
CSNK1G1			x			x			
CTF1				x				x	
CTNND2			x			x			
CUEDC2			x					x	
CXCR4	x							x	
DDX19-DDX19L			x					x	
DDX19A			x					x	
DDX19B			x					x	
DEFB103A				x		x			
DEFB104A				x		x			
DEFB105A				x		x			
DEFB105B				x		x			
DEFB106A				x		x			
DEFB109				x		x			
DEFB4				x		x			
DMTF1			x			x			
DMXL2			x					x	
DNAJB7			x					x	
DOCK3				x		x			
DRD5	x							x	
DUB4	x							x	
DYNLRB1			x					x	
EFCBP1			x					x	
EIF2S2			x					x	
ELAVL4	x					x			
EP300			x					x	
ERAF	x							x	
EXOSC6			x					x	
FAM96A			x			x			
FBS1				x				x	
FBXL15			x					x	
FBXL19				x				x	
FBXO33*	x							x	
FKSG83				x				x	
FLJ10246	x							x	
FLJ21127	x					x			
FLJ21657	x							x	

FLJ32130				x				x	
FLJ32363	x							x	
FLJ38377				x		x			
FLJ38773			x					x	
FLJ40142	x					x			
FLJ40296			x					x	
FLJ43855	x							x	
FLJ46121	x							x	
FLJ46156		x	x						
FOXP2*			x			x			
FRMD5			x			x			
FUK			x					x	
FUS				x				x	
GAB1			x			x			
GBF1			x					x	
GIT2	x					x			
GLYAT				x				x	
GRM2				x		x			
GRM5				x				x	
GUSBL1				x				x	
HERC2P4	x							x	
HIST1H2AG				x				x	
HIST1H2AH				x				x	
HIST1H2BJ				x				x	
HIST1H2BK				x				x	
HIST1H4I				x				x	
HMGCS1	x							x	
HNMT	x							x	
HSD3B7				x				x	
HSPA9B			x			x			
IFT81	x					x			
INPP4B			x					x	
IQCF2				x		x			
ITCH			x					x	
ITGAD	x							x	
ITGAM				x				x	
KCNT2	x					x			
KIAA0101			x			x			
KIAA1324L			x			x			
L3MBTL2			x					x	
LHX8			x			x			
LOC116236				x		x			
LOC150763		x				x			
LOC153684	x							x	
LOC283932				x				x	
LOC387930			x				x		
LOC388955				x		x			
LOC401354			x					x	
LOC401447				x		x			
LOC440366	x							x	
LOC441762	x							x	
LOC493829				x				x	

LOC51057			x			x			
LSM8	x							x	
MDH1				x		x			
MGC119295	x							x	
MGC13024				x				x	
MGC13138				x				x	
MGC15619	x					x			
MGC3020	x							x	
MGC3121				x				x	
MGC34647			x					x	
MGC34761			x					x	
MGC35361	x							x	
MGC42105	x							x	
MSH4			x			x			
MTFMT			x				x		
MYST1				x				x	
NFKB2			x					x	
NRG2				x		x			
NUFIP2				x		x			
OAZ2			x			x			
PAIP1	x							x	
PCDH7				x		x			
PDPR			x					x	
PHF12				x		x			
PHF5A			x					x	
PHKG2				x				x	
POLR2J	x							x	
POLR2J2	x							x	
POLR2J3	x							x	
PPIB			x			x			
PPM1L				x				x	
PPP1CC	x					x			
PRMT6			x			x			
PRSS16				x				x	
PSD			x					x	
PYCARD				x				x	
PYDC1				x				x	
RAB27B	x							x	
RAB9P1				x		x			
RABGGTB			x			x			
RAD54L2				x		x			
RAD9B	x					x			
RANGAP1			x					x	
RASA4	x							x	
RBM15B				x		x			
RBPMS2			x			x			
RBX1			x					x	
RNF190			x					x	
RNF40				x				x	
RTN1			x			x			
SCG3			x					x	
SEMA3G	x					x			

SEPP1	x							x	
SETD1A				x				x	
SEZ6				x		x			
SF3B3			x					x	
SLC2A9	x							x	
SLC44A5			x			x			
SLC5A2	x							x	
SNX1			x			x			
SNX22			x			x			
SPAG11				x		x			
SPG21			x				x		
SRCAP				x				x	
ST13			x					x	
ST3GAL2			x					x	
ST6GALNAC3			x			x			
STX1B2				x				x	
STX4A				x				x	
SUFU			x					x	
TA-PP2C	x					x			
TAOK1				x		x			
TBR1				x		x			
TCHP	x					x			
TEF			x					x	
TEX264				x		x			
TGFB1I1	x							x	
TIAF1				x		x			
TLR9	x					x			
TMEM55A			x					x	
TOB2			x					x	
TP53TG3	x							x	
TRIP4			x			x			
TRUB1			x			x			
TYR				x				x	
TYRO3	x					x			
USP38			x			x			
VGLL3	x					x			
VKORC1				x				x	
VPRBP				x		x			
VPS29	x					x			
WDR1	x							x	
WDR51A			x			x			
ZC3H7B			x					x	
ZNF131	x							x	
ZNF248	x					x			
ZNF25	x					x			
ZNF267	x							x	
ZNF364			x			x			
ZNF589				x				x	
ZNF609			x			x			
ZNF646				x				x	
ZNF668				x				x	
ZNF675				x				x	

ZNF688				x				x	
ZNF689				x				x	
ZNF720	x							x	

Four (C3-C6) anestry-matched stage 1 and stage 2 sample sets had sufficient sample sizes (≥50 probands) for the replication study. A total of 1,086 and 1,190 genes were identified in C3-C6 discovery (stage 1) and replication (stage 2) analyses respectively. The 310 genes that are significant in both stage 1 and stage 2 analyses are listed. Previously reported ASD candidate genes are denoted with a *.

Supplementary Table 8 Risk homozygous haplotypes in genomic regions showing significant linkage with ASD

	Discovery Stage 1			Replication Stage 2		
Locus	**Pop**	**Genomic position**	**Genes**	**Pop**	**Genomic position**	**Genes**
1p21.1	C5	chr1:105124663-105883977	MIR548H3	C3	chr1:103248768-104152300	COL11A1, RNPC3, AMY2B, AMY2A, AMY1A, AMY1B, AMY1C
2q31.1	C3	chr2:175420189-176187475	CHN1, ATF2, ATP5G3			
3q25-27	C5	chr3:159069542-159689477	SHOX2, RSRC1	C5	chr3:161253090-162285936	LOC401097, IFT80, SMC4, KPNA4, ARL14, PPM1L, B3GALNT1
	C2	chr3:161899603-162422511	PPM1L, B3GALNT1	C4	chr3:163680830-163739058	Intergenic
	C6	chr3:161899603-163005957	PPM1L, B3GALNT1, NMD3, C3orf57, OTOL1			
	C5	chr3:171239541-171505135	CPR160, PHC3, PRKC1			
7q22.1				C6	chr7:98629829-99297192	KPNA7, MYH16, ARPC1A, ARPC1B, BUD31, PTCD1, PDAP1, ATP5J2-PTCD1, CPSF4, ZNF789, ZNF394, ATP5J2, ZKSCAN5, FAM200A, ZNF655, ZNF498, CYP3A5, CYP3A7, CYP3A4, CYP3A43
				C5	chr7:98719136-99526758	MYH16, ARPC1A, ARPC1B, BUD31, PTCD1, PDAP1, ATP5J2-PTCD1, CPSF4, ZNF789, ZNF394, ATP5J2, ZKSCAN5, FAM200A, ZNF655, ZNF498, CYP3A5, CYP3A7, CYP3A4, CYP3A43, OR2AE1, TRIM4, GJC3, AZGP1, AZGP1P1, ZKSCAN1, ZSCAN21, ZNF3, COPS6
				C4	chr7:98715204-99260475	MYH16, ARPC1A, ARPC1B, BUD31, PTCD1, PDAP1, ATP5J2-PTCD1, CPSF4, ZNF789, ZNF394, ATP5J2, ZKSCAN5, FAM200A, ZNF655, ZNF498, CYP3A5, CYP3A7, CYP3A4, CYP3A43
15q13.1-q14	C5	chr15:25695484-26723446	OCA2, HERC2, GOLGA8G, GOLGA8F			
	C2	chr15:27762157-28736917	TJP1, FAM7A1, FAM7A2, FAM7A3, LOC653075, CHRFAM7A, ARHGAP11B			
	C6	chr15:28157207-29159268	FAM7A1, FAM7A2, FAM7A3, LOC653075, CHRFAM7A, ARHGAP11B, FAN1M			

			MTMR10, TRPM1			
15q14-q21.1	C3	chr15:39680413-39957870	MGA, MAPKBP1, JMJD7, PLA2G4B, SPTBN5	C3	chr15:39008360-39634468	DLL4, CHAC1, INO80, EXD1, CHP, LOC729082, OIP5, NUSAP1, NDUFAF1, RTF1, ITPKA, LTK, RPAP1
	C4	chr15:41232598-41328431	TMEM62, CCNDBP1, EPB42, TGM5	C3	chr15:42350179-42775902	CASC4, CTDSPL2, LOC645212, EIF3J, SPG11, PATL2
	C5	chr15:42039370-42169690	FRMD5	C5	chr15:42752595-42888343	PATL2, B2M, TRIM69
				C5	chr15:43215740-43233448	DUOX1
				C5	chr15:43503090-43527684	C15orf48
				C5	chr15:46897612-47030097	SHC4, EID1
15q21.1-q22.2	C5	chr15:49269215-49791852	CYP19A1, GLDN, DMXL2, SCG3	C5	chr15:49535903-50134220	DMXL2, SCG3, LYSMD2, TMOD2, TMOD3, LEO1, MAPK6
				C3	chr15:54359469-55084899	TEX2, MNS1, ZNF280D, LOC145783, TCF12
20q11.21-q13.12	C5	chr20:32227845-32991499	ASIP, AHCY, ITCH, DYNLRB1, MAP1LC3A, PIGU, NCOA6, TP53INP2, HMGB3P1, GGT7, ACSS2, GSS	C5	chr20:32078885-33184387	RALY, EIF2S2, ASIP, AHCY, ITCH, DYNLRB1, MAP1LC3A, PIGU, NCOA6, TP53INP2, HMGB3P1, GGT7, ACSS2, GSS, MYH7B, TRPC4AP, EDEM2
				C4	chr20:32600000-33277616	MAP1LC3A, PIGU, NCOA6, TP53INP2, HMGB3P1, GGT7, ACSS2, GSS, MYH7B, TRPC4AP, EDEM2, PROCR
				C3	chr20:33262613-33350919	MMP24, EIF6, FAM83C

Thirty-one risk homozygous haplotypes identified in the stage 1 (C2-C6) and stage 2 (C3-C6) analyses are located under significant (LOD > 3.3) ASD linkage peaks.

References (Supplementary Material 3)

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5: 184-186

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190

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: 69-85