Patterns of inherited deletions in autistic spectrum disorders

Ning Lei[a,b,1], Chang Chan[b,1], Tengting Lim[c], Yingleong Chan[c], Shoichi Metsugi[b], Veronica Demtchouk[a], Iulia Kotenko[a], Joseph Chan[d], Raul Rabadan[d], Gunaretnam Rajagopal [e], Arnold J Levine[b,e,2,] and Daniel A. Notterman[a,2]

[a]Department of Molecular Biology, Princeton University, Princeton, NJ, 08544; [b]Simons Center for Systems Biology, Institute for Advanced Studies, Princeton, NJ, 08540; [c]Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, 02115; [d]Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, New York, 10032; and [e]Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ, 08903

[1]contributed equally
[2]contributed equally

Classification: Biological Sciences (Genetics)

- Designed research: AL, DN, GR
- Performed research: CC, NL, TL, YC, SM, VD, IK
- Contributed new reagents or analytic tools: CC, JC, RR, NL, TL, YC
- Analyzed data: CC, NL, TL, YC, SM
- Wrote the paper: NL, CC, DN, AL

Corresponding author:

Daniel Notterman
Department of Molecular Biology
229 Lewis Thomas Laboratory
Princeton University
Princeton, NJ
Dan1@Princeton.edu
609-258-7185 (voice)
609-258-4575 (fax)

Abstract

The core features of the autistic spectrum disorders are pervasive and severe impairment of communication, social interactions, cognitive ability, and behavioral repertoire. Several studies have pointed to spontaneous or *de novo* copy number variants (CNVs, mostly deletions) associated with autism and implicating genes expressed in the central nervous system (CNS). To provide additional information concerning the role of inherited structural variants in autism, we carried out a family-based association study using the Autism Genetic Research Exchange (AGRE) database of multiplex autism families (www.AGRE.org). Working with pedigree, phenotype, and SNP data residing in the AGRE database of approximately 4500 individuals in 900 multiplex families and with several comparable databases of control subjects, we examined the frequency and location of CNV deletions in the AGRE families. We identified four genes inherited in these families and that with incomplete penetrance are associated with an ASD. These CNV deletions were confirmed by PCR and sequencing in DNA samples from some of these subjects. Of the four genes that we identified, two of them were previously shown to be associated with an ASD: *Neurexin-1-alpha* (*NRXN1*, Ch. 2p16.3) and *Contactin-associated protein-like 2* (*CNTNAP2*, Ch. 7q35-q36). Deletions in two novel susceptibility genes were uncovered: *Neural Cell Adhesion Molecule 2* (*NCAM2*, Ch. 21q21.1) and *Protein Tyrosine Phosphatase, Receptor Type D* (*PTPRD*, Ch. 9p23-p24.3). Penetrance was incomplete, and ranged from 50 to 85%; as predicted, the endpoints of the deletion mutations were unique to most families but within a family were stable.

Introduction

While it is straightforward to enumerate the core features of the autistic spectrum disorders (pervasive and severe impairment of communication, social interactions, cognitive ability, and behavioral repertoire), these disorders display great clinical and etiologic heterogeneity, and may in fact represent different etiologies and pathological mechanisms linked by some common phenotypic elements. A comparison of monozygotic versus dizygotic twin studies and sibling recurrence rates have long established that autism is largely genetically determined (1). However, linkage and association studies suggest that many different genetic loci can be involved in the phenotype of autistic spectrum disorders (ASD) and inheritance patterns are often complex and resist simple genetic interpretations. Even though on average, autism has a higher recurrence risk among siblings of an affected child than in the general population ($\lambda_s > 100$), in many families (simplex), autism arises in a single proband, usually male. Several studies have pointed to spontaneous or *de novo* copy number variants (CNVs, mostly deletions) associated with these autistic offspring. These CNVs implicate genes involved in a number of biological processes affecting developmental and functional aspects of the central nervous system (CNS). Indeed, *de novo* deletions (and less commonly duplications) have been identified in many cases of simplex autism (2-4). It has been estimated that at least 10 to 50 % of sporadic autism is due to this sort of structural variant (3, 5).

It was initially observed that *de novo* CNVs were less common in familial autism (3); however, CNVs at several loci have now been implicated in parents and siblings of multiplex families with more than one affected child, suggesting that they have inherited a common, variably penetrant, autism-susceptibility region (6, 7). These observations can be reconciled by a model (5) in which most cases of autism are caused by a *de novo* (germline) mutation in the

proband and selection usually removes this mutation within a single generation. On occasion, the mutation has incomplete penetrance and therefore occurs in an individual who does not manifest a complete autistic phenotype (but who may display isolated or attenuated deviations in social, communicative or cognitive function that do not significantly limit reproductive fitness). If the mutation is located on an autosome, then it will be passed on to offspring with a probability of 50%. If a child receiving this mutation does not have a suppressor gene that confers poor penetrance or an environment that attenuates the phenotype then he will display autistic features.

To provide additional information concerning the role of inherited structural variants in autism, we carried out a family-based association study using the Autism Genetic Research Exchange (AGRE) database of multiplex autism families (www.AGRE.org). Working with pedigree, phenotype, and SNP data residing in the AGRE database of approximately 4500 individuals in 900 multiplex families and with several comparable databases of control subjects, we examined the frequency and location of CNV deletions in the AGRE families. We identified four genes that can be inherited in these families and with incomplete penetrance are associated with autism or an ASD, and possibly other disorders. These CNV deletions were confirmed by PCR and sequencing in DNA samples from some of these subjects. Of the four genes that we identified, two of them were previously shown to be associated with an ASD: *Neurexin-1-alpha* (*NRXN1*, Ch. 2p16.3) and *Contactin-associated protein-like 2* (*CNTNAP2*, Ch. 7q35-q36). Deletions in two novel susceptibility genes were uncovered: *Neural Cell Adhesion Molecule 2* (*NCAM2*, Ch. 21q21.1) and *Protein Tyrosine Phosphatase, Receptor Type D* (*PTPRD*, Ch. 9p23-p24.3). Unaffected carrier parents were identified in approximately 25 families. Similar, but non-identical deletions in these genes were also found in control populations but at a lower

frequency than in the dataset of the AGRE families. Unexpectedly, the CNVs were evenly divided between mothers and fathers. Penetrance was incomplete, and ranged from 50 to 85%; as predicted, the endpoints of the deletion mutations were unique to most families (except for 2 families with deletions in *PTPRD*) but within a family were stable. Most striking was that even in families where there was a clear pattern of segregation of an inherited deletion with the autistic phenotype, a number of affected children did not harbor the familial variant. This was true for *NRXN1* and *CNTNAP2,* genes that have been previously judged to represent autism susceptibility genes, as well as for the new genes identified here, *NCAM2* and *PTPRD*. Further analysis with several of these families revealed deletions at other loci, also segregating within the families, which may participate in additional autism risk. These observations are most consistent with a two-gene or multi-gene hypothesis to explain many cases of autism.

Results

Identification of Deletions in the AGRE Multiplex Collection

We performed two related analyses of pedigree and genotype data from the AGRE collection of 943 families, most of which had more than one affected child. An ASD phenotype was assigned after testing with the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS). Most subjects were genotyped with the Illumina HumanHap550 BeadChip; genotypes from several individuals were excluded after quality tests (e.g., individuals with no or bad genotypes, see Supporting Information Materials and Methods), and the number of subjects analyzed at this stage was 4562. Using cnvPartition v1.2.1 (Illumina, San Diego), with a window size of 1 megabase, we identified CNV deletions encompassing 10 or more SNPs and performed a family-based genome-wide association study using autistic children as case and non-autistic children as control. We looked for association with deletions having the highest penetrance, using the odds ratio. We sought both intronic and exonic deletions. Twenty-five deletions were identified (Table 1), including deletions that disrupt genes coding *NRXN1* and *CNTNAP2*, both of which have recently been found by several independent groups to be associated with autism (8-16). The mean size of these CNVs was approximately 53 kb (*NCAM2*) 84 kb (*PPTRD*), 159 kb (*NRXN1*), and 180 kb (*CNTNAP2*). To account for the prevalence of these 25 deletions in a normal population, we tested for their presence in a control group of approximately 6317 individuals without developmental or neuropsychiatric illness that we assembled from genotype data derived from several studies (see Methods). This comparison was performed using PennCNV (17) to detect deletions in the AGRE and control groups because we found this algorithm to be somewhat more congruent with the results of our PCR validation (see Materials & Methods) and also more applicable to non-

Illumina genotype data. Barnard's Exact Test (18) was employed to evaluate significance ($p < 0.1$) and these results were then confirmed with Fisher's Exact Test. This comparison resulted in a list of four deletion regions that are enriched in families with autism, but found significantly less often in the control group (Table 2). For *NRXN1*, there are eight families with inherited intronic or exonic deletions in *NRXN1* (deleted exons encoding *NRXN1*α) with at least one of the children with autism in each family inheriting the deletion (*NRXN1* deletions shown in Supplemental Figure 1B and pedigrees shown in Supplemental Figure 2B). For *CNTNAP2*, three families have an intronic deletion and one family has an exonic deletion (*CNTNAP2* deletions shown in Supplemental Figure 1A and pedigrees shown in Supplemental Figure 2A). An additional exonic deletion spanning only 6 SNPs in *CNTNAP2* was identified in a family (i.e., AU0880), in which one autistic son obtained the deletion from his mother (Supplemental Figure 1A and Figure 2A). These findings in genes that are known to play a role in autism provide confidence that the approach used in this family-based genome wide association study is valid and fruitful. Further supporting the assertion that these inherited deletions are family-specific (i.e., do not represent single ancestral events that are now distributed across the population) is that the deletion in each of the families has unique breakpoints as mapped by the SNP analysis and confirmed by PCR analysis (Supplemental Figure 1).

Our analysis also yielded two genes segregating in autistic families in which deletions have not been previously detected in autism. Inherited deletions in PTPRD, a protein tyrosine phosphatase receptor expressed in the CNS and regulating neuron axon guidance was found in members of 10 families. An additional three probands with deletions in *PTPRD* were identified, of which two are sporadic (e.g., AU050603 and AU070703) and the third is ambiguous due to the absence of the father's genotype data (e.g., AU074204). *PTPRD* was recently associated

with attention-deficit/hyperactivity disorder (ADHD) (19), but has not been previously associated with an ASD. *NCAM2* was identified as a candidate gene though linkage analysis in autism characterized by developmental regression, but specific abnormalities in this gene have not previously been identified (20). Pedigrees of the families affected by deletions in *NCAM2* or *PTPRD* are shown in Figure 2. Surprisingly, in several families in which an inherited deletion segregated, males displayed an autistic phenotype, but did not appear to inherit the family deletion. The magnitude of gender bias in these families was otherwise similar to that recorded in the literature: of 58 children classified with ASD, 47 were male (4.3:1). However, of 14 individuals with an ASD and no familial deletion all were male and every female contained the family deletion. This finding was significant ($p=0.021$, Barnard's Exact Test). While there could be many reasons for this sexual dimorphism, one simple interpretation is that some mutations predisposing to ASD that are not detected by this Illumina chip analysis (small deletions and point mutations, etc.) can impact males but are not penetrant in females (thus the 4:1 ratio). This is consistent with females' displaying relative resistance to autism, with the phenotype expressed only in the context of a large deletion or other major mutation.

Validation of Deletion Endpoints by Sequencing

To accurately determine the breakpoints for the predicted deletions, we developed sets of nested PCR primers adjacent to the 5' and 3' ends of each predicted deletion. In the presence of a deletion, a PCR product will be produced and it can be sequenced while in the absence of a deletion the PCR primers will be too far apart and produce no product. Following amplification, PCR products were sequenced to confirm the deletion breakpoints. Table 2 provides the sequence-validated breakpoints for selected families for each implicated gene.

In every case so far examined (37 individuals in 13 families), sequencing confirmed the presence of the predicted deletion. The validated breakpoint endpoints were often predicated by PennCNV but in 13 of 37 cases, the true breakpoint was shifted by several kb from that identified by PennCNV (data not shown). Consistent with their arising independently, the precise endpoints for these deletions differ between families, but they remain the same within individual families, confirming parent-to-child-transmission without any additional mutations in the locus. The inherited deletions in *PTPRD* occurred across several different exons and introns, but all were mapped to the 5' untranslated region (UTR) of the gene (Figure 1B), implying that the functional effect of these deletions would involve regulation of expression (such as a ribosome binding site, etc). Deletions in *NCAM2* were always intronic (Figure 1A). The effect of these intronic deletions on gene splicing and regulation still requires study. However, in at least one case studied employing lymphocytes in cell culture from the family, an intronic deletion in *CNTNAP2* was associated with errors in splicing that resulted in deletion of several exons (AU1412301, data not shown). Since neither *NCAM2* nor *PTPRD* is more than minimally expressed in the cultured lymphocytes from which DNA samples were derived (data not shown), testing the functional significance of these intronic and 5' UTR deletions will require additional studies and reagents.

Pedigree Patterns of Inherited Deletions

Although deletions in each of the four loci that we identified segregate with the ASD phenotype, and are enriched in these families, the association of the deletions with the ASD phenotypes is complex: penetrance was incomplete (from 50 to 85 %), and in many families, some children with an ASD had the mutation, while others (only males) did not (Figure 2 and Supplemental Figure 2). For example, inherited deletions were uncovered in this study in

families that implicate *CNTNAP2*, which has been previously associated with ASD through common variants (8-10), a complex rearrangement disrupting the *CNTNAP2* gene (11), and even a homozygous deletion (12). We detected a heterozygous, familial, intronic as well as exonic deletion in four families (*CNTNAP2* Pedigrees are shown in Supplemental Figure 2A). In two families, the unaffected carrier was the mother, and in the other two, the father. For this gene, penetrance in offspring was complete (all children with the deletion were affected); however, 3 male children, each from a different family, met ADOS criteria for autism, but did not contain a detectable deletion at this locus. The absence of the familial deletion was confirmed for AU019303 and AU038303 by PCR (Supplemental Figure 1A). In the additional family AU0880 in which the mother and an affected child had a deletion in *CNTNAP2* (predicted to span 6 SNPs and contain exon 2), a pair of dizygotic twin's were classified with ASD but neither child had the familial deletion under study. In this case, however, both of these children had Fragile X syndrome, known to be associated with syndromic autism. Thus the family was carrying one deletion in *CNTNAP2* and had offspring with Fragile X syndrome (presumably spontaneous) which explains the ASD phenotypes. Of interest, in two of the three children affected with autism but without a *CNTNAP2* deletion, we detected an inherited deletion in another gene previously associated with ASD, such as *catenin alpha 3* (*CTNNA3*, Ch. 10q21.3) in family AU0193 (21), or *ataxin 2-binding protein 1* (*A2BP1*, Ch. 16p13.3) in family AU0383 (22), suggesting that mutations in a second or additional genes may segregate within these families and contribute either individually or in combinations to the autistic phenotype (Supplemental Figure 2A).

With respect to *NRXN1,* which has also been connected to autism, we identified 8 families in which the deletion was inherited (Supplemental Figure 1B and 2B). In 5 of the 8

families, it was transmitted by an apparently unaffected father, and in 3 families by the mother. While penetrance was high (13 classified with ASD of 15 children with their familial deletion, 87 %), 3 children were classified as autistic, but did not manifest the *NRXN1* deletion.

We observed 3 families with inherited deletions in *NCAM2*, also termed olfactory cell adhesion molecule (*OCAM,* Figure 1A and 2A). Of 8 children with the deletion, 5 (63 %) were autistic; conversely, 3 children displayed an ASD, but not the familial deletion in *NCAM2* (Figure 1A). Of these 3 children, one (AU073005) had an apparently *de novo* deletion in *LPS-responsive vesicle trafficking, beach and anchor containing* (*LRBA,* Ch. 4q31.3), whose close related subfamily member, *neurobeachin* (*NBEA*, Ch.13q13), has been previously associated with autism (23).

The pedigrees associated with the novel autism susceptibility gene, *PTPRD*, continue this pattern. Ten families with a deletion implicating this gene were identified (Figure 1B and 2B). Of the 10 families, in 6 cases the deletion was transmitted by the father and in 4 cases by the mother. Of 23 children who inherited the deletion, 19 (83 %) were classified with an ASD; conversely, 5 children (all males, each from a different family) were classified with an ASD but did not inherit the familial deletion that was detected in this study.

Of interest, in addition to the familial *PTPRD* deletion in family AU0700, CNV deletion mutations in *NCAM2* were observed in two of the autistic sisters (e.g., AU070003 and AU070004, Figure 2B). Similarly, one of the children (AU073003) on the autistic spectrum who had an inherited deletion of *NCAM2* also had a *de novo* deletion in the promoter of *PTPRD*, suggesting that in this case, we may have identified two deletions, each of which contribute to the autistic phenotype. Thus there are a growing number of examples, just examining these large deletions, which suggest that at least two or more genes associated with the ASD phenotypes can

be found in families that demonstrate inherited ASD. In addition, we did not attempt identify other type of mutations, such as point mutations, duplications, or inversions, so our analysis is likely a substantial underestimate of the extent of associated mutations segregating in these families. The extensive variation of the ASD phenotype could then result from one, two or more mutations contributing to the phenotype. As we learn more about the genetics of these families these become testable ideas.

Discussion

The results of this family association study confirm the prior findings that *CNTNAP2* and *NRXN1* are autism susceptibility genes and continue the emerging theme that genes coding proteins involved in the synaptic adhesion complex are often implicated in autism and related disorders (24). We also provide definitive evidence that deletions in the range of 54-84 kb, affecting two new susceptibility genes, *NCAM2* and *PTPRD* segregate within the multiplex families and are associated with a penetrance of approximately 0.75. It is anticipated that the deletions we found associated with ASD have incomplete penetrance. The lower bound for the penetrance of a deletion is calculated by comparing the number of autistic people having the deletion in a family with the total number of people in the family with deletions. The average lower bound for these deletions is about 50%. The reason this is a lower bound is because most of these deletions are inherited, and by including parents, who are almost always not autistic, we are biased towards having more people with the deletion who are not autistic. The upper bound can be calculated by looking at the number of autistic children having the deletion in a family to the total number of children in the family having the deletion. The average upper bound for these deletions is 80%. The reason this is an upper bound is due to selection bias (AGRE has

selected for families with multiple autistic children) and stoppage (parents may be more likely to stop having additional children after having an autistic child).

NCAM2 was identified as a positional candidate for autism because of proximity to an interval on chromosome 21q that was linked to susceptibility to the regressive phenotype of autism (20). *PTPRD*, has not been previously linked to autism, and while its role in the CNS is not clear, there is evidence that it too, functions, in part, as an adhesion molecule (25, 26).

NCAM2 was originally described as a membrane-associated protein mainly expressed in olfactory sensory neurons of rodents; hence in these species it may be referred to as OCAM-olfactory cell adhesion molecule (27). In humans, *NCAM2* was identified in an exon-trapping experiment designed to identify genes potentially involved in the Down Syndrome phenotype, since the gene is located on chromosome 21. Unlike the case in rodents, where expression is mainly in the olfactory sensory regions, in humans it is expressed widely in fetal and adult brain (amygdala, caudate nucleus, corpus callosum, hippocampus, substantia nigra, subthalamic nucleus, and thalamus) (28). NCAM2 consists of several domains. Five extracellular Ig homology regions are followed by two fibronectin type 3 homology regions. Alternative splicing of the transcript from *NCAM2* results in either a glycosylphosphatidylinositol (GPI)-anchored isoform of the protein or a transmembrane isoform that, in addition to the extracellular domain, contains a 20-25 amino acid-long transmembrane region followed by a 106-119 amino acid-long cytoplasmic tail, potentially used for signaling. Its amino acid sequence is approximately 45% homologous with neural cell adhesion molecule 1 (NCAM1, Ch. 11q23.1) also involved in neuronal adhesion functions in the CNS (reviewed in (27)).

Studies suggest that NCAM2 modulates neurite outgrowth, axonal guidance, synapse formation in the olfactory system and formation of dendritic bundles in mouse (29). Its role in

human CNS development remains to be elucidated, although participation in the development synaptic adhesion complex seems likely based on its homology to NCAM and its expression pattern. A potential role for *NCAM2* in the genetic susceptibility to autism was first implied by a linkage study performed on a subset of AGRE multiplex families (357 families), in which identified a locus that is 600Kb away from *NCAM2* under the dominant mode of inheritance for autism (20). In our studies with more than 900 multiplex families, we describe three families with distinct heritable deletions, each affecting the first intron of *NCAM2*. Together with the prior genome-wide association study this provides strong evidence that mutations in *NCAM2* contribute to the pathogenesis of autism.

The second novel autism susceptibility gene that we report, *PTPRD*, is a Type IIα receptor-like phosphatase, expressed in diverse tissues derived form the neural crest, including the hippocampus, the olivary nucleus and the mitral and glomerular layer of the olfactory bulb, where expression is highest in mid-gestation (26). It has an extracellular domain that contains cell adhesion motifs, in an intracellular region with two phosphatase domains (30). Together with *Protein Tyrosine Phosphatase, Receptor Type S* (*PTPRS,* Ch. 19p13.3), it plays an essential role in axon targeting during axonogenesis. Mice deficient in *Ptprd* display poor eating and learning impairment in the Morris water maze and radial arm maze tasks. Consistent with an effect on synaptic plasticity, the magnitudes of long term potentiation at CA1 and CA3 synapses was greater in the *Ptprd*-deficient animals. Thus, PTPRD has a function in axonogenesis and synaptic plasticity that is consistent with a pathogenetic role in neurodevelopment disorders, including autism (30-35). However, until this report, *PTPRD* has not been previously associated with autism, although prior genome-wide association studies have risk alleles of *PTPRD* in restless legs syndrome and ADHD (19, 33). Interestingly, the analysis of ADHD also identified

linkage with *CNTNAP2*, which others and we have found to be an autism susceptibility gene. This leads to speculation that these shared causative variants convey a generalized risk for abnormal neurodevelopment that is then channeled by co-existing rare or common genetic variants and the environment into a particular phenotype (19). *PTPRD* has also been considered as a tumor suppressor, implicated in a variety of solid tumors, including, melanoma, colon, and malignant glioma (36), extending an interesting link between autism and many tumor suppressor genes that regulate growth and respond to nutrients (e.g., *PTEN*, *TSC-1*, *TSC-2* and others).

Approximately 24 % of the children affected with an ASD in this study did not appear to display the familial CNV. This was true for the two genes (*NRXN1* and *CNTNAP2*) that have previously been identified as autism susceptibility genes and for the two novel risk genes that we report (*NCAM2* and *PTPRD*). In several of these apparently anomalous cases, a deletion in an additional autism susceptibility gene was found to be segregating in the family. Several additional CNV deletions and duplications were found to segregate in these families and some of these implicated genes are established susceptibility genes (e.g., *A2BP1*, *CTNNA3*, and *MDGA2*) while others are attractive candidates by virtue of being expressed in the synapse, or associated with other neurodevelopmental or psychiatric disorders (e.g., *ARHGEH10*, *CLN8*, *CTNNB1*, *FHIT*, *KRT3*, *LRBA*, *MACROD2*, *NEURL*, *NRIP1*, *PTPRK*, *SH3PXD2A* and *RASSF4*). Furthermore, this follow-up analysis was limited to deletions that mapped to plausible autism susceptibility gene candidates. Other types of mutation were not evaluated, and may also be involved in some of the anomalous cases as well.

It is interesting to speculate on the link between CNVs and autism and other disorders such as schizophrenia (37). Genomic contexts associated with fragile sites are associated with the accumulation of deletions and duplications. However, only 6 of the validated breakpoint

endpoints (approximately 25%) in our samples are located within repeat regions, such as low copy repeats LCR(38) short interspersed repetitive elements (SINE, such as alu sequences) (39), long interspersed repetitive elements (LINE), simple (TA)n, or long terminal repeats (LTR) (data not shown). Duplications and deletions may also occur stochastically. Regardless of the mechanism or mechanisms by which they are produced, the probability that a specific gene or class of genes will be affected is partly a function of the size of that gene (or the sum of the sizes of group of genes that compose a class). Other considerations being equal, larger genes will present a great target area for mutation and will be affected more often than smaller genes. To learn if gene size could play a role in the accumulation of mutations in genes affecting neurological function, genes and their lengths were obtained (for Homo sapiens) and genes participating in neural function, development, and disease were annotated either by using the Gene Ontology database (www.geneontology.org) or by Ingenuity Pathway Analysis (Ingenuity Systems, Redwood City). To test whether neurological genes were enriched for large genes, we examined the proportion of human neurological genes larger than any given gene length (Supplemental Figure 3). The resulting plot confirms that most genes affecting neurological functions are large in length. The rank sum *p*-value for the difference between the lengths of neurological and non-neurological genes was calculated to be 1.37E-17. Moreover, an enumeration of the 100 largest genes in the human genome indicates that many have functions in the CNS Supplemental Figure 3. Indeed, *PTPRD*, which we report to be an autism susceptibility gene is the second largest gene in humans, and *CNTNAP2* and *NRXN1*, which we confirm as susceptibility genes, are the largest and the 42nd largest, respectively. This result encourages us to consider that other large genes that encode for proteins with function in the CNS are plausible candidates for susceptible genes in disorders such as autism and schizophrenia.

Materials and Methods

See *Supporting information* for full methods.

AGRE Multiplex Collection. SNP genotype data using the Illumina HumanHap550 was provided for 4562 individuals from approximately 900 families. Phenotype information and EBV-transformed lympboblast cells were provided for most individuals classified as having an ASD and for many of their first-degree relatives. Details about this resource are provided at http://www.AGRE.org.

Assembled Control Group. The control group consisting of 6317 individuals was comprised of three different genotype collections: 2026 normal individuals probed by the Illumina HumanHap550 BeadChip(40), 1259 European-American individuals that were used as a control group for a study of schizophrenia (https://dbgap.ncbi.nlm.nih.gov, Genome-Wide Association Study of Schizophrenia); and 3032 individuals from a case control study of diabetes (https://dbgap.ncbi.nlm.nih.gov, GENEVA Diabetes Study). The latter two studies were sampled with the Affymetrix Genome-Wide SNP Array 6.0.

Phenotype information and EBV-transformed lympboblast cells of unaffected indivudals were provided by Coriell Institute for Medical Research at http://www.coriell.org.

PCR and Sequencing. PCR was performed as suggested by the manufacturer (Phusion® PCR Master Mix F-531L; http://www.NEB.com) and DNA products specific to deletion carriers from AGRE collection were sequenced with gene-specific primers (GENEWIZ, Inc.; http://www.genewiz.com).

Acknowledgement
This project was supported by the Simons Foundation and the Nancy Laurie Marks Foundation. We gratefully acknowledge the resources provided by the AGRE Consortium and the participating AGRE families. The 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). CC acknowledges support from the Charles L. Brown Membership at the Institute for Advanced Study.

References

1. Folstein S & Rutter M (1977) Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry* 18:297-321.
2. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT*, et al.* (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 358:667-675.
3. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C*, et al.* (2007) Strong association of de novo copy number mutations with autism. *Science* 316:445-449.
4. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE*, et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459:569-573.
5. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH*, et al.* (2007) A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci U S A* 104:12831-12836.
6. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y*, et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science* 321:218-223.
7. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI*, et al.* (2009) Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* 5:e1000536.
8. Alarcon M, Abrahams BS, Stone JL, Duvall JA, Perederiy JV*, et al.* (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet* 82:150-159.
9. Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K*, et al.* (2008) A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am J Hum Genet* 82:160-164.
10. Bakkaloglu B, O'Roak BJ, Louvi A, Gupta AR, Abelson JF*, et al.* (2008) Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am J Hum Genet* 82:165-173.
11. Poot M, Beyer V, Schwaab I, Damatova N, Van't Slot R*, et al.* (2009) Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay and autism spectrum disorder. *Neurogenetics*.
12. Jackman C, Horn ND, Molleston JP, & Sokol DK (2009) Gene associated with seizures, autism, and hepatomegaly in an Amish girl. *Pediatr Neurol* 40:310-313.
13. Kim HG, Kishikawa S, Higgins AW, Seong IS, Donovan DJ*, et al.* (2008) Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* 82:199-207.
14. Sudhof TC (2008) Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455:903-911.
15. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J*, et al.* (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 39:319-328.

16. Yan J, Noltner K, Feng J, Li W, Schroer R*, et al.* (2008) Neurexin 1alpha structural variants associated with autism. *Neurosci Lett* 438:368-370.
17. Wang K, Li M, Hadley D, Liu R, Glessner J*, et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665-1674.
18. Mehta C & Senchaudhuri P (2003) Conditional versus unconditional exact test for comparing two binomials.
19. Elia J, Gai X, Xie HM, Perin JC, Geiger E*, et al.* (2009) Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry*.
20. Molloy CA, Keddache M, & Martin LJ (2005) Evidence for linkage on 21q and 7q in a subset of autism characterized by developmental regression. *Mol Psychiatry* 10:741-746.
21. Wang K, Zhang H, Ma D, Bucan M, Glessner JT*, et al.* (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459:528-533.
22. Martin CL, Duvall JA, Ilkin Y, Simon JS, Arreaza MG*, et al.* (2007) Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. *Am J Med Genet B Neuropsychiatr Genet* 144B:869-876.
23. Castermans D, Wilquet V, Parthoens E, Huysmans C, Steyaert J*, et al.* (2003) The neurobeachin gene is disrupted by a translocation in a patient with idiopathic autism. *J Med Genet* 40:352-356.
24. Biederer T & Stagi M (2008) Signaling by synaptogenic molecules. *Curr Opin Neurobiol* 18:261-269.
25. Schaapveld R, Wieringa B, & Hendriks W (1997) Receptor-like protein tyrosine phosphatases: alike and yet so different. *Mol Biol Rep* 24:247-262.
26. Schaapveld RQ, Schepens JT, Bachner D, Attema J, Wieringa B*, et al.* (1998) Developmental expression of the cell adhesion molecule-like protein tyrosine phosphatases LAR, RPTPdelta and RPTPsigma in the mouse. *Mech Dev* 77:59-62.
27. Kulahin N & Walmod PS (2008) The Neural Cell Adhesion Molecule NCAM2/OCAM/RNCAM, a Close Relative to NCAM. *Neurochem Res*.
28. Paoloni-Giacobino A, Chen H, & Antonarakis SE (1997) Cloning of a novel human neural cell adhesion molecule gene (NCAM2) that maps to chromosome region 21q21 and is potentially involved in Down syndrome. *Genomics* 43:43-51.
29. Walz A, Mombaerts P, Greer CA, & Treloar HB (2006) Disrupted compartmental organization of axons and dendrites within olfactory glomeruli of mice deficient in the olfactory cell adhesion molecule, OCAM. *Mol Cell Neurosci* 32:1-14.
30. Uetani N, Chagnon MJ, Kennedy TE, Iwakura Y, & Tremblay ML (2006) Mammalian motoneuron axon targeting requires receptor protein tyrosine phosphatases sigma and delta. *J Neurosci* 26:5872-5880.
31. Sun QL, Wang J, Bookman RJ, & Bixby JL (2000) Growth cone steering by receptor tyrosine phosphatase delta defines a distinct class of guidance cue. *Mol Cell Neurosci* 16:686-695.
32. Uetani N, Kato K, Ogura H, Mizuno K, Kawano K*, et al.* (2000) Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. *EMBO J* 19:2775-2785.

33. Schormair B, Kemlink D, Roeske D, Eckstein G, Xiong L*, et al.* (2008) PTPRD (protein tyrosine phosphatase receptor type delta) is associated with restless legs syndrome. *Nat Genet* 40:946-948.
34. Elchebly M, Wagner J, Kennedy TE, Lanctot C, Michaliszyn E*, et al.* (1999) Neuroendocrine dysplasia in mice lacking protein tyrosine phosphatase sigma. *Nat Genet* 21:330-333.
35. Wallace MJ, Batt J, Fladd CA, Henderson JT, Skarnes W*, et al.* (1999) Neuronal defects and posterior pituitary hypoplasia in mice lacking the receptor tyrosine phosphatase PTPsigma. *Nat Genet* 21:334-338.
36. Veeriah S, Brennan C, Meng S, Singh B, Fagin JA*, et al.* (2009) The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proc Natl Acad Sci U S A* 106:9435-9440.
37. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S*, et al.* (2009) Common variants conferring risk of schizophrenia. *Nature* 460:744-747.
38. Edelmann L, Pandita RK, & Morrow BE (1999) Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet* 64:1076-1086.
39. de Smith AJ, Walters RG, Coin LJ, Steinfeld I, Yakhini Z*, et al.* (2008) Small deletion variants have stable breakpoints commonly associated with alu elements. *PLoS One* 3:e3104.
40. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H*, et al.* (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* 19:1682-1690.

Figure Legends

Figure 1. Detection and validation of inherited deletions in A) *NCAM2* and B) *PTPRD* in AGRE families with ASD. The span and location of deletions predicted by PennCNV is indicated by the colored bars. Position is indicated as base pairs from the telomeric end of the p arm of the chromosome, annotated to *NCAM2* or *PTPRD* by RefSeq assembly Hg18 (March 2006). Individual identification codes (AGRE ID_gender_affected status; M: male, F: female, A: autistic and N: normal) are indicated in the column to the left of the figure. Color code: children with ASD (Autistic, red); parent and unaffected siblings (Carrier, green), validated family deletion (Validated, blue). For selected family, PCR was performed for the carrier parent and siblings (family pedigrees shown in Figure 2) as described in Materials and Methods and shown in the bottom gel, where absence of a PCR product indicates the deletion was not present. Unrelated, unaffected individuals denoted N1 and N2, were included as negative controls. * indicates sporadic deletions detected in children with ASD but not in their parents or siblings. ** indicates a deletion with undetermined inheritance due to absent paternal genotype.

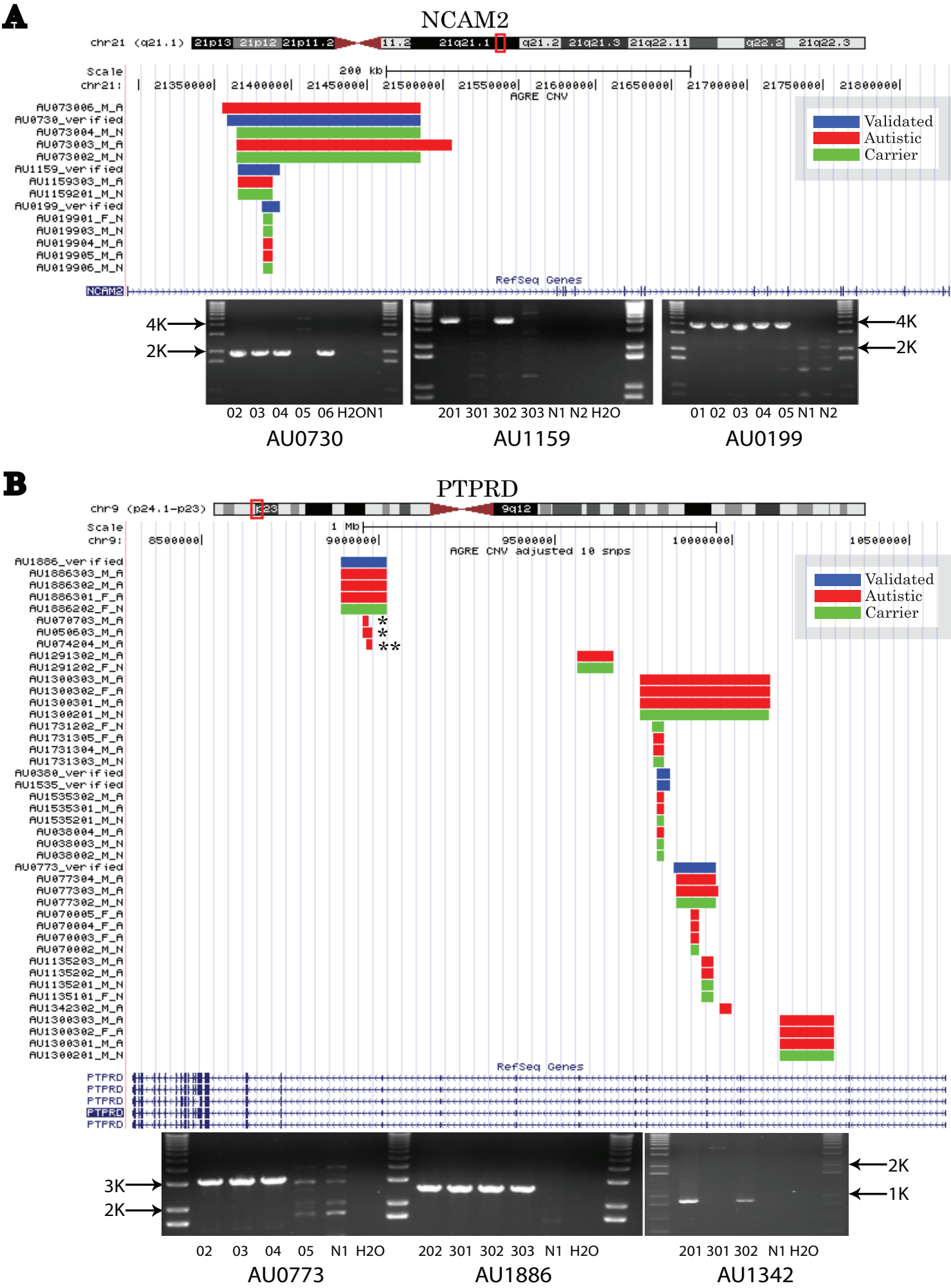
Figure 2. Pedigrees of families with inherited deletions in *NCAM2* A) and *PTPRD* B) segregating with ASD. Individuals predicted by PennCNV to harbor a familial deletion are underscored. A red insert indicates individuals for whom PCR validation was performed. Various other inherited or spontaneous CNV deletions segregating with ASD in these families were predicted by PennCNV and the implicated genes are indicated under individuals. Black:

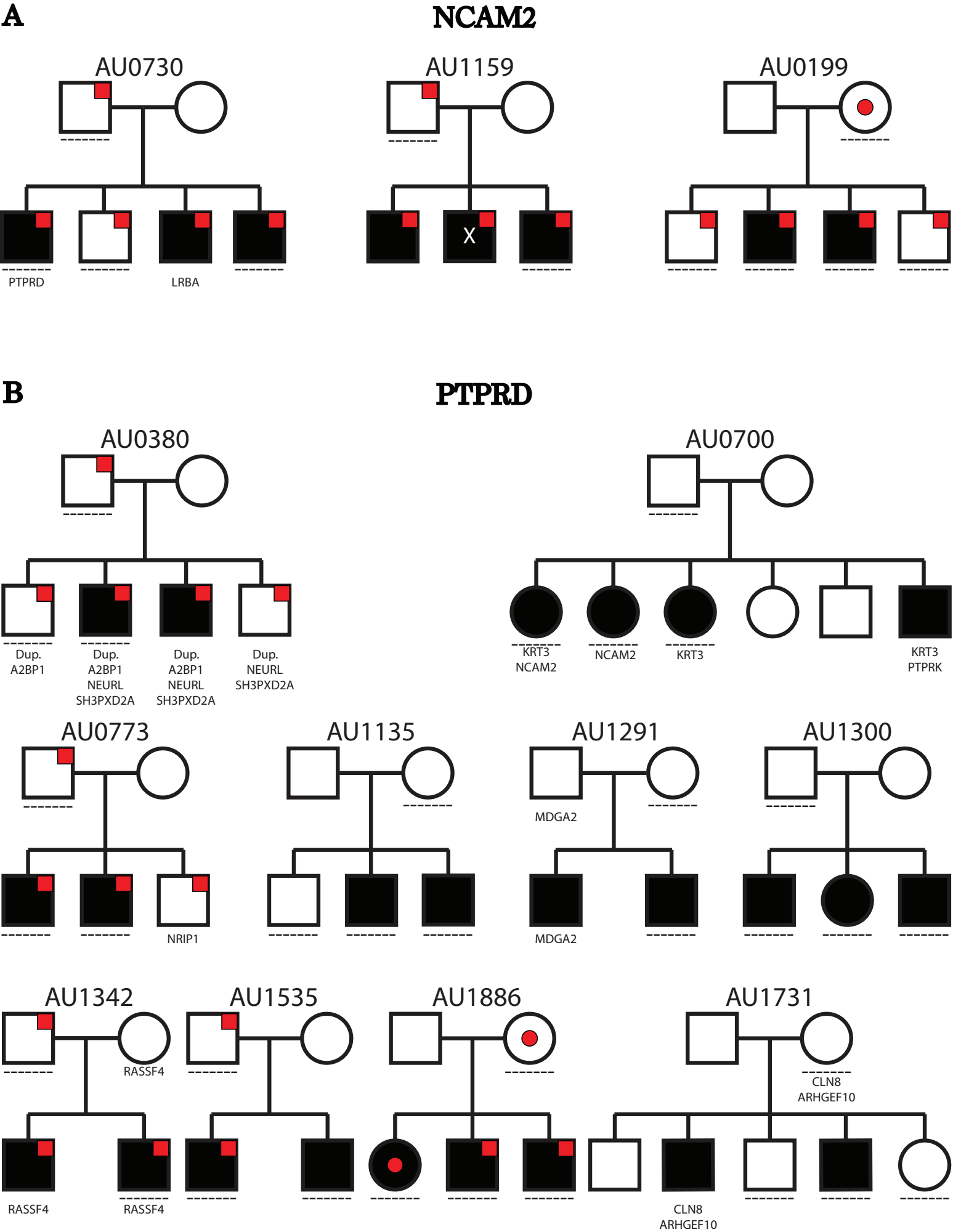
ASD, White: unaffected parents and siblings. Square: male, Circle: female. X: no genotype available.

Table Legends

Table 1. Twenty-five deletion regions segregating with ASD. Deletions were identified with cnvPartition v1.2.1 software, employing a 1 Mb window. Deletion window start and end positions indicated the 1 Mb chromosomal region in which the deletions were found, assigning position from the telomeric end of the p arm of the chromosome. Implicated genes were annotated by RefSeq assembly Hg18 (March 2006). The location of some genes has been slightly revised in subsequent editions. NA--Deletions in apparent inter-genic region.

Table 2. Genes with deletions more prevalent in individuals with ASD than in control individuals ($p < 0.1$). Validated breakpoints were unique to the families and similar within families, confirming independent occurrence and parent-to-child transmission. The breakpoint position was referenced in base-pairs to the telomeric end of the p arm. N: AGRE cases with an inherited or sporadic (*de novo*) deletion in the gene. ^ Includes 1 sporadic deletion in a child with an ASD. ^^ Includes at least 2 sporadic cases. * indicates identical breakpoints validated in the families.


A
NCAM2
chr21 (q21.1)
Scale
chr21:
AGRE CNV
Validated
Autistic
Carrier
RefSeq Genes
NCAM2
4K
2K
02 03 04 05 06 H2ON1
201 301 302 303 N1 N2 H2O
01 02 03 04 05 N1 N2
AU0730
AU1159
AU0199
B
PTPRD
chr9 (p24.1-p23)
AGRE CNV adjusted 10 snps
RefSeq Genes
PTPRD
3K
2K
1K
02 03 04 05 N1 H2O
202 301 302 303 N1 H2O
201 301 302 N1 H2O
AU0773
AU1886
AU1342


A
NCAM2
AU0730
PTPRD
LRBA
AU1159
X
AU0199
B
PTPRD
AU0380
Dup.
A2BP1
Dup.
A2BP1
NEURL
SH3PXD2A
Dup.
A2BP1
NEURL
SH3PXD2A
Dup.
NEURL
SH3PXD2A
AU0700
KRT3
NCAM2
NCAM2
KRT3
KRT3
PTPRK
AU0773
NRIP1
AU1135
AU1291
MDGA2
MDGA2
AU1300
AU1342
RASSF4
RASSF4
RASSF4
AU1535
AU1886
AU1731
CLN8
ARHGEF10
CLN8
ARHGEF10

Chr.	Deletion window start position	Deletion window end position	Number of families in which deletion segregates with ASD	Number of children with ASD having deletion	Number of unaffected children having deletion	Implicated genes
2p22.3	34500000	36100000	10	13	0	NA
2q16.3	50100000	51100000	8	12	2	Neurexin-1-alpha (NRXN1)
2q13	109300000	111300000	11	17	2	Nephrocystin-1 isoform 3 (NPHP1)
2q32.3	193100000	194800000	6	10	1	NA
3p26.1	6200000	7200000	6	8	1	Glutamate receptor, metabotropic 7 (GRM7)
4q32.3	168100000	170200000	8	12	0	Annexin A10 (ANXA10)
5q23.1	119500000	121400000	12	12	1	Ferritin mitochondrial (FTMT)
6q14.3	86100000	87100000	7	8	1	Sorting nexin 14 (SNX14); Synaptotagmin binding, cytoplasmic RNA interacting protein (SYNCRIP)
6q16.1	94600000	96200000	11	17	2	NA
7p21.1	16600000	18500000	7	11	1	Sorting nexin 13 (SNX13)
7q31.1	111000000	112100000	8	11	0	Dedicator of cytokinesis 4 (DOCK4)
7q35-36.1	145200000	146200000	4	6	0	Contactin-associated protein-like 2 (CNTNAP2)
8p23.2-23.1	5800000	7100000	5	7	0	Microcephalin 1 (MCPH1)
9p23-24.1	9700000	11600000	14	24	3	Protein tyrosine phosphatase, receptor type, D (PTPRD)
9p23	12300000	13300000	7	8	1	NA
11p12	37300000	39200000	5	10	0	NA
11p11.12	49600000	51300000	16	17	3	NA
11q22.1	98100000	99100000	6	11	1	Contactin 5 (CNTN5)
13q21.33	68700000	70700000	18	22	1	Kelch-like protein 1 (KLHL1)
15q21.3	51200000	52800000	3	6	0	WD repeat domain 72 (WDR72)
15q24.2-24.3	73700000	75000000	5	6	0	Neuregulin 4 (NRG4); Ubiquitin-conjugating enzyme E2Q family member 2 (UBE2Q2); S-phase cyclin A-associated protein in the ER (SCAPER)
15q25.2-25.3	82300000	84200000	5	7	0	Phosphodiesterase 8A (PD8A); Solute carrier family 28 (sodium-coupled nucleoside transporter), member 1 (SLC28A1)
16p13.3-13.2	6100000	7900000	8	13	2	Ataxin 2-binding protein 1 (A2BP1)
18q22.1	62800000	64800000	10	14	1	Dermatan-sulfate epimerase-like protein (DSEL)
21p21.1	20400000	21700000	6	8	1	Neural cell adhesion molecule 2 (NCAM2)

Table 1

Gene	Chr.	N	Combined control (Barnard's *p*-value)	Family validated by sequencing	Sequence-validated breakpoints	Deletion size (position)
NRXN1	2p16.3	9^	18	AU1043	Chr2p: 50,806,568-50,732,693	73,874bp (within intron2)
			(0.0649)	AU1210	Chr2p: 51,236,318-50,782,877	453,442bp (including exon1&2)
CNTNAP2	7q35-q36.1	4	4	AU0193	Chr7q: 146,120,426-146,437,026	316,599bp (including exon3,4&5)
			(0.0317)	AU0383	Chr7q:145,846,735-146,047,595	200,861bp (within intron1)
				AU1412	Chr7q: 145,625,946-145,723,501	97,554bp (within intron1)
PTPRD	9p23-p24.1	13	30	AU0380*	Chr9p: 9,823,339-9,788,114	25,224bp (within intron4)
		^^	(0.0552)	AU1535*		
				AU0773	Chr9p: 9,954,852-9,835,049	119,802bp (including exon4)
				AU1342	Chr9p: 9,999,966-9,965,232	34,735bp (within intron3)
				AU1886	Chr9p: 9,022,233-8,892,625	129,607bp (including exon10)
NCAM2	21q21.1	3	0	AU0199	Chr21q: 21,380,639-21,392,752	12,112bp (within intron1)
			(0.00221)	AU0730	Chr21q: 21,358,402-21,485,493	127,090bp (within intron1)
				AU1159	Chr21q: 21,364,949- 21,392,916	27,966bp (within intron1)

Table 2.