Profilng allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage

Changhoon Lee1,7, Eun Yong Kang2, Michael J. Gandal1,3, Eleazar Eskin2,4,5 and Daniel H. Geschwind1,3,4,6*

One fundamental but understudied mechanism of gene regulation in disease is allele-specific expression (ASE), the preferential expression of one allele. We leveraged RNA-sequencing data from human brain to assess ASE in autism spectrum disorder (ASD). When ASE is observed in ASD, the allele with lower population frequency (minor allele) is preferentially more highly expressed than the major allele, opposite to the canonical pattern. Importantly, genes showing ASE in ASD are enriched in those downregulated in ASD postmortem brains and in genes harboring de novo mutations in ASD. Two regions, 14q32 and 15q11, containing all known orphan C/D box small nucleolar RNAs (snoRNAs), are particularly enriched in shifts to higher minor allele expression. We demonstrate that this allele shifting enhances snoRNA-targeted splicing changes in ASD-related target genes in idiopathic ASD and 15q11-q13 duplication syndrome. Together, these results implicate allelic imbalance and dysregulation of orphan C/D box snoRNAs in ASD pathogenesis.

ASE is a form of genetic regulation in which the expression of mRNA at a specific locus is biased toward the specific allele1–10. An extreme case of ASE is genomic imprinting, resulting in monoallelic expression (MAE), whereby a particular allele is completely silenced11,12. But most ASE represents more subtle shifts in the allelic ratio rather than complete silencing of one allele13,14 as evidenced by the observations of preferentially reduced expression of mutant disease alleles and that a substantial proportion of cis-acting expression quantitative trait loci (cis-eQTL) are mediated by ASE12,14. Indeed, cis-eQTL and ASE are complementary mechanisms involved in the regulation of gene expression15. Further, as it relies primarily on within-sample expression comparisons, rather than comparisons between samples, analysis of ASE is less influenced by genetic, environmental and technical confounders15,16.

A number of diseases can be attributed to abnormalities in ASE17–19. For example, Prader–Willi syndrome and Angelman syndrome are rare neurodevelopmental disorders resulting from deletions within the highly imprinted region, 15q11–q13. Conversely, recurrent maternally derived duplications of this region lead to dup15q, a rare but penetrant syndromic form of ASD comprising approximately 1% of cases20. Although duplications generally increase gene expression17, varied patterns are observed across genes in the region impacted by dup15q due to complex local regulatory mechanisms18,19.

In addition to known imprinted regions, ASE extends to over 5% of autosomal genes and is especially enriched in transmembrane receptors and cell surface molecules17, Random monoallelic expression (RMAE), which reflects MAE not associated with parent-of-origin imprinting, may be important for cell-type-specific gene dosage effects18,21. Moreover, emerging evidence suggests that RMAE is involved in neurodevelopmental disorders, as recently observed for developmental dyspraxia caused by RMAE in FOXP2 (ref. 22). In this case, ASE may contribute to disease susceptibility by exacerbating the deleterious effects of a mutation via haploinsufficiency, or mosaic somatic expression.

Except for one notable study, which showed that ASE in several specific genes may play a role in a subset of ASD cases4, the role of genome-wide ASE has not been explored in a sufficiently powered cohort. The overall paucity of ASE studies in brain disorders, such as ASD, is likely due to several challenges including limited access to human brain tissue, small sample sizes, incomplete transcriptome-wide genotyping data and reference bias in mapping transcription4,15,19,23,24,25. Reference bias is attributable to reduced mapping of alternative-allele-containing RNA-sequencing (RNA-seq) reads (due to a greater number of mismatches) and can generate
up to 40% false positive signals if not properly addressed\textsuperscript{11,13,23}. However, when these issues are properly addressed, ASE can be a highly robust approach to investigate the relationship between epi-genetic variation and gene expression underlying diseases\textsuperscript{15}. Here we applied an optimized ASE pipeline in a relatively large sample of postmortem brains from individuals with idiopathic ASD and controls. We found that most loci showing ASE are shared between ASD and control brains, which is in contrast to large ASE differences observed between the cerebral cortex and cerebellum\textsuperscript{15}. A small proportion of loci manifest reproducible ASE changes between ASD and controls, which may represent a form of genetic regulation related to disease pathogenesis.

**Results**

**ASE identification in brain.** We first developed an optimized ASE identification pipeline (Fig. 1; Methods) to overcome potential challenges, such as reference bias in mapping transcripts, lack of transcriptome-wide genotyping data and allelic expression bias introduced by ethnicity. We collected 263 postmortem brain tissue samples from frontal cortex (Brodmann area (BA)9), temporal cortex (BA41-42-22 (BA41)) and cerebellar vermis from 96 individuals (40 controls and 56 ASD cases including 8 cases with dup15q). Gene expression profiling was performed using RNA-seq\textsuperscript{1}, and genotyping was conducted on arrays (Methods).

To overcome reference bias, we generated a reference genome masked for single nucleotide polymorphisms (SNPs) that could contribute to mapping bias. We compiled ~40 million possible variants based on the union of SNPs called from RNA-seq data (Supplementary Fig. 1a) and SNPs already identified from microarray probe sets and the 1000 Genomes Project\textsuperscript{26,27} (Methods). We prepared a custom human reference genome after masking these SNPs (Methods) and removed 11.13% of SNPs that showed biased mapping from a simulated dataset\textsuperscript{18} (Methods). To collect transcriptome-wide genotyping data from our brain samples, we performed imputation from genotyping array and RNA-seq-based genotyping (Methods; Supplementary Fig. 1b,c), and ancestry was identified for all samples (Methods; Supplementary Fig. 1d).

For investigation of group-level ASE patterns across cases and brain regions, 1,163,249 heterozygous SNPs in 21,929 genes were used following quality control (Methods), and their significance was quantified using a linear mixed model to account for sample variation and covariates (Supplementary Table 1; Methods). The values of excluded covariates showed balanced distributions between control and ASD (Supplementary Fig. 1e; Methods).

We filtered out all SNPs showing ancestry-associated variation to avoid the potential effects of population stratification (Methods). ASE was also assessed within each individual sample for quality-controlled SNPs using Fisher’s exact test (Methods). Both individual and group-level ASE analyses underwent multiple testing correction for independent SNPs after linkage disequilibrium (LD)-based SNP pruning (Methods). Among the alleles showing ASE, we also filtered out reference-biased SNPs related to proximal indels (Methods). As expected, SNPs within the same haplotype blocks show similar allelic expression patterns (Supplementary Fig. 1f,g).

**Common ASE patterns in brain tissues.** For external validation, we first compared genes with evidence of ASE with those in dbMAE, an integrated database compiled from ASE analyses across human and mouse tissues\textsuperscript{22}. Although this database contains data from mostly non-neural tissues, genes manifesting ASE in our study showed significant overlap with those in dbMAE (Supplementary Fig. 2a). Furthermore, although assessed separately, there was strong concordance in common ASE genes between brain regions in our data (Fig. 2a and Supplementary Table 2). For example, the overlap between genes showing ASE in the two cortical regions, BA9 and BA41, is very high (odds ratio (OR) = 221.00 and 7.22 \times 10^{-16} \times (Methods).)

Furthermore, although assessed separately, there was strong concordance in common ASE genes between brain regions in our data (Fig. 2a and Supplementary Table 2). For example, the overlap between genes showing ASE in the two cortical regions, BA9 and BA41, is very high (odds ratio (OR) = 221.00 and 7.22 \times 10^{-16} \times (Methods).)

The distribution of expression of imbalanced alleles was significantly lower than that of balanced alleles (Kolmogorov–Smirnov test, \textit{P} < 2 \times 10^{-16}, the means of log\textsubscript{2} reads per kilobase of SNP area
ASE patterns shared among cases and controls. a, Venn diagram comparing the overlap of ASE genes from three brain tissues: BA9, BA41 and vermis. b, Enrichment analyses of genes exhibiting ASE with ASD-relevant (SFARI gene list; Methods) and cell-type-specific gene lists (Methods)\(^3\). Across all three brain regions, ASE genes showed strong overlap with known imprinted genes (Methods) as well as targets of FMRP (ref. \(^2\)), HuR (ref. \(^2\)) and RBFOX1 (ref. \(^2\)). Plots show ORs and FDR-corrected \(P\) values for enrichment, if significant. c, Enrichment analysis of genes exhibiting ASE with candidate psychiatric risk genes. Likely gene-disrupting de novo variants were considered from studies of schizophrenia (SCZ), intellectual disability (ID) and ASD. The datasets for SCZ, ID and ASD1 were gene lists containing de novo likely gene disrupting mutations from Iossifov et al.\(^3\), and the data for ASD2 represent risk genes integrating de novo copy number variations (FDR \(\leq 0.01\)) from Sanders et al.\(^3\) (Methods). The risk variants from GWAS\(^3\) were considered for ADHD, BPD, MDD, SCZ and ASD (Methods). In this case, ASD1 and ASD2 represent the GWAS datasets from the Cross-Disorder Group of the Psychiatric Genomics\(^3\) and Grove et al.\(^3\), respectively. If significant, the plot shows ORs and FDR-corrected \(P\) values for de novo variant datasets and FDR-corrected \(P\) values for GWAS. The sample numbers of BA9, BA41 and vermis are 67, 64 and 64, respectively. Ver, vermis.

ASE patterns shared across idiopathic ASD cases and controls. Similar to controls, idiopathic ASD cases showed tissue-specific patterns of ASE across the three brain regions (Fig. 3a and Supplementary Fig. 3a,b). Since the two cortical regions showed strong overlap in ASE and cortical manifestations of transcriptional dysregulation in ASD are more prominent than those in cerebellum, we focused on the cerebral cortex for most of the subsequent case–control comparisons. We found that SNPs showing ASE are broadly and evenly distributed throughout the genome in both control and idiopathic ASD samples (Fig. 3b). The majority of genes exhibiting ASE (70.41\%) are shared across case and control groups, providing further confidence in the reproducibility of these results (Fig. 3c and Supplementary Table 3). Nevertheless, fewer genes showing ASE are observed in idiopathic ASD compared with controls (Fig. 3c), and this was recapitulated when studying the rate of ASE at the level of each individual sample (Fig. 3d).

Comparing biological pathway enrichment in cases and controls revealed overlap and differences in ASE gene function. Genes exhibiting ASE in both idiopathic ASD and controls are enriched in phosphoprotein, RNA splicing and neuron projection pathways (Fig. 3c and Supplementary Table 4). Genes exhibiting control-specific ASE are significantly enriched for pathways related to growth cone and synapse (Supplementary Fig. 3c and Supplementary Table 4).

Although there are fewer idiopathic ASD-specific ASE genes than control-specific ASE genes, those showing idiopathic ASD-specific ASE are significantly enriched for pathways involved in transport (Supplementary Fig. 3d and Supplementary Table 4). Genes exhibiting ASE in both idiopathic ASD and controls show significant enrichment in genes harboring de novo ASD risk variants (Supplementary Fig. 3e; \(OR = 1.68, P = 5 \times 10^{-4}\) for ASD1; \(OR = 3.83, P = 0.001\) for ASD2). However, control-specific and idiopathic ASD-specific ASE genes do not show any enrichment either with rare de novo ASD risk variants or with the common variants from GWAS data (Supplementary Fig. 3e).
Quantitative allelic imbalance in idiopathic ASD. ASE is not always completely monoallelic and may involve subtle shifts in allelic balance, whereby one allele is quantitatively favored over the other\(^4\). To investigate this full spectrum of ASE in control and idiopathic ASD, we compared the distributions of the minor allele expression fraction for SNPs in which these calls are high quality (Methods; Fig. 4a). As expected, the majority of heterozygous alleles show no evidence of ASE (herein called ‘balanced’) with the fraction near 50%. A much smaller number of alleles show complete MAE and accordingly their frequencies are either 0% or 100%. Finally, we define ‘imbalanced’ alleles as those that lie between the balanced and MAE distribution ranges. Controls had a higher density of balanced alleles than ASD cases, while idiopathic ASD cases showed more imbalanced alleles (Fig. 4a and Supplementary Fig. 4a). However, controls showed significantly more major allele MAE for genes harboring low-frequency minor alleles (minor allele frequency (MAF) < 0.05) compared with idiopathic ASD (Supplementary Fig. 4b). These patterns were recapitulated in the pseudoautosomal region (PAR) of the chromosome X (Supplementary Fig. 4c).

We next investigated whether the differential ASE pattern was related to the MAF genome wide (Fig. 4b and Supplementary Fig. 4b). The disease-specific patterns (balanced alleles: control > idiopathic ASD; imbalanced alleles: control < idiopathic ASD) were largely stable across the range of MAFs. However, we observed that complete MAE was limited to rare alleles. Furthermore, preferential major allele expression becomes stronger as the MAF decreases (Fig. 4c). Major alleles exhibiting MAE are enriched in genes that are more tolerant of loss of function\(^{11}\) (LoF; Methods) and missense mutations compared with those showing balanced expression (chi-squared test \(P = 0.0134\); Fig. 4d). This suggests that a bias toward major allele expression may provide a buffer against the consequences of potentially deleterious (rarer) mutations. Remarkably, this preferential pattern of major allele expression shows a significant interaction with idiopathic ASD (Fig. 4e). On average, minor alleles are preferentially more highly expressed in idiopathic ASD than in controls (minor/major allele ratio: 0.79 in ASD and 0.75 in control; chi-squared test \(P = 7.33 \times 10^{-11}\)). Together, these results suggest that rare and potentially pathogenic alleles are more likely to be unmasked in idiopathic ASD brain by the patterns of ASE.

MAE occurs primarily from the major allele. Overall, we observed a strong bias toward expression of the major allele for sites in genes harboring rare minor alleles and MAE alleles. However, unlike global ASE patterns (Fig. 3c), MAE patterns differed between cases and controls (Fig. 5a). Although MAE SNPs were observed across the genome (Supplementary Fig. 5), there was an approximately two-fold enrichment on chromosome 15 harboring a number of known imprinted loci (Supplementary Table 5). Reasoning that this could provide a convergent mechanism underlying the co-occurrence of dup15q syndrome and idiopathic ASD, we next investigated MAE in eight cases of dup15q. We observed a substantial enrichment of minor allele MAE in ASD and dup15q compared with controls (Fig. 5b; 2.5- and 2.7-fold, respectively). This genome-wide enrichment pattern persisted when SNPs were analyzed at individual genes (Supplementary Fig. 6a,b).

Overall, minor allele-predominant MAE was significantly enriched on chromosomes 14 and 15 (Fig. 5c), with ASD and dup15q showing stronger patterns compared with controls. Loci on chromosome 15 exhibited a substantially higher minor allele MAE fraction in both ASD and dup15q than controls (3.9- and 1.6-fold, respectively; chi-squared test \(P < 2.2 \times 10^{-16}\) and \(P < 0.0184\), respectively; Fig. 5c and Supplementary Table 5). Chromosomes 3, 6 and 7 also showed higher fractions of minor allele MAE in ASD and dup15q samples than controls, of which only chromosome 3 was significant in both (4.2- and 9.8-fold, respectively; chi-squared test \(P = 0.0022\) and \(P = 8.89 \times 10^{-16}\), respectively). However, we focused on chromosomes 14 and 15 for further analysis since chromosome 3 had an overall lower degree of minor allele MAE.

Pathway analysis of minor allele MAE genes showed strong enrichment for known imprinted genes, as expected (Fig. 5d). Interestingly, these MAE genes were enriched for distinct pathways from typical ASE genes that favor major allele expression (Fig. 2b), suggesting that orthogonal biological processes are involved. For example, genes showing minor allele MAE showed no overlap with astrocyte markers\(^{27}\), PSD\(^{29}\) or ASD-specific downregulated genes\(^{28}\) (Supplementary Fig. 6c). However, they did show enrichment for genes harboring known rare mutations increasing risk for ASD (Simons Foundation Autism Research Initiative (SFARI) genes\(^{36}\); Methods) whereas other classes of genes showing ASE did not. Similar to genes that manifest ASE (Fig. 2c), major allele MAE genes show significant enrichment with genes associated with the rare de novo ASD risk variants (Supplementary Fig. 6d). However, MAE genes observed in control and ASD samples showed some enrichment for common variants from GWAS data of BPD and SCZ (Supplementary Fig. 6d).

Allele shift regions enriched in minor allele MAE in ASD and dup15q and snoRNAs. We have so far demonstrated that ASD and dup15q are associated with an increased rate of genes showing minor allele MAE, which are enriched on chromosomes 14 and 15 and overlap known ASD risk genes, especially those harboring deleterious mutations that act via haploinsufficiency. Interestingly, the genomic regions showing MAE allele shifts from major to minor allele in both ASD and dup15q relative to controls are located on just a few chromosomes (Table 1) and overlap regions known to harbor ASD-associated copy-number variants (CNVs)\(^{37}\). Among common MAE alleles, the frequencies of the allele shift in ASD and dup15q are 15.52% and 2.62%, respectively. To understand the effect of this apparent MAE allele shift in ASD, we compiled a set of high-confidence genomic regions which we called ‘allele shift rich regions’. We defined these regions as having such allele shifts in both ASD and dup15q relative to controls. This analysis identifies only two genomic regions, 14q32 (chr14:101302,638-101,544,745) and 15q11 (chr15:25,223,730-25,582,395). Both are known imprinted loci that have many small splice junctions (Fig. 6) and show continuous high expression (Supplementary Fig. 7) as single transcript units\(^{38,40}\). Both regions also contain tandem repeats of multiple orphan C/D box snoRNA genes\(^{39,40}\), which are located within repeated introns of MEG8 (ref. \(^{41}\)), SNURP-SNFPN and SNHG14 transcripts. Indeed, all known orphan C/D box snoRNA genes are located within these two

---

**Fig. 3 | ASE patterns distinguish control and idiopathic ASD brains.** a. The comparison of ASE genes from BAV3, BA41 and vermis within control and idiopathic ASD groups. b. Manhattan plots for control and idiopathic ASD cortex show a broad distribution of genomic-loci exhibiting ASE. Sample numbers of control and idiopathic ASD brains are 69 and 62, respectively. c. Venn diagram showing comparison of ASE genes across groups in cortex. d. ASE rates for individual samples from idiopathic ASD and control groups (Methods). Idiopathic ASD samples (n = 32) show lower overall rates of ASE compared with controls (n = 31). For the violin plots, ASE rates were calculated per chromosome for each cortical sample (Methods). The Wilcoxon rank sum test was two-sided. e. GO analyses are shown for common ASE genes between control (n = 69) and idiopathic ASD groups (n = 62). In the upper interactive graph, the bubble color indicates the P value of the GO term, and bubble size indicates its frequency\(^{41}\). The P values and other results of the GO analysis are in Supplementary Table 4. Highly similar GO terms are linked by edges, and the line width indicates the degree of similarity. In the lower figure, the white, yellow and orange boxes represent \(P > 10^{-5}\), \(10^{-5} < P \leq 10^{-4}\) and \(10^{-4} < P \leq 10^{-3}\), respectively. Ctl, control.
regions, including SNORD113 and SNORD114 at 14q32 (Fig. 6a), and SNORD64, SNORD107, SNORD108, SNORD109, SNORD115 and SNORD116 at 15q11 (Fig. 6b). These snoRNA genes are highly expressed in the brain\cite{9,40} and control alternative splicing of specific target genes without complementarity to ribosomal RNA within their sequences\cite{9-11}. 

![Diagram](image_url)
To begin to understand the impact of the allele shifts in these regions, we next assessed whether the relevant genes were differentially expressed in ASD or dup15q cases versus controls. Analysis of RNA-seq data showed regional downregulation of 15q11 in ASD ($P = 0.0016$; Fig. 7a). SNORD116-24 downregulation was observed in ASD and dup15q from small noncoding RNA-seq data (sncRNA-seq) data42 ($P = 0.0347$ and 0.0019, respectively; Fig. 7b).

As snoRNA genes regulate the splicing of downstream targets, allele shifts would change potential binding and may alter the splicing patterns of snoRNA targets. To test this hypothesis, we used a known splicing target list41 of the snoRNAs to investigate whether these genes show splicing changes based on RNA-seq data from postmortem ASD brain1. The accuracy of these splicing target predictions has been confirmed in previously published studies43,44, providing

---

**Fig. 4 | Quantitative allelic imbalance in idiopathic ASD and control cortex.** a, The distribution of minor allele expression fraction for idiopathic ASD (red) and control (blue) groups for autosomal SNPs. The majority of loci show balanced expression. The density plot near 0% and 100% is zoomed in to show patterns of MAE. b, SNP fractions of balanced, imbalanced or MAE expression per MAF. c, SNP fractions showing preferential major or minor allele expression per MAF. In b and c, open circles are control, and closed circles are idiopathic ASD. d, The comparison of SNP numbers, which possibly can cause LoF mutations, amino acid changes or synonymous mutations in control cortex. Major-allele MAE can act to buffer deleterious LoF and missense mutations. Y axes show SNP numbers. e, SNP counts showing preferential expression of the major and minor alleles in control and idiopathic ASD. Their ratio is expressed as Minor/Major.
experimentally validated snoRNA targets including some studies in neuronal cells. From receiver operating characteristic analysis, the splicing target prediction showed a 90–100% true positive rate within a 0.02–0.30% false positive rate range. Indeed, we find that the snoRNA target genes show more splicing changes in ASD and dup15q compared with controls (Fisher’s exact tests: OR = 1.40 and...
The snoRNA targeting sites were located preferentially proximal to alternatively spliced junctions of their target genes in ASD (Supplementary Fig. 8a), and their splicing changes show significant correlations with the expression changes of their specific snoRNAs (Supplementary Fig. 8b,c). When one snoRNA has two different targets, their splicing changes also show strong correlation (Supplementary Fig. 8d).

Furthermore, among snoRNA targets, genes showing altered splicing changes in ASD and dup15q have two different datasets of de novo ASD risk variants (in ASD, OR = 2.30 (P = 0.04) for the first and OR = 7.42 (P = 0.005) for the second; in dup15q, OR = 2.14 (P = 0.05) for the first and OR = 5.97 (P = 0.02) for the second; Fig. 7e).

Discussion
This study provides a large-scale, genome-wide investigation of ASE patterns in human brain samples from people with ASD and ASDdup. The results are consistent with the bioinformatic predictions above and potentially implicate disruption of snoRNA-mediated splicing of synaptic ASD risk genes in the pathophysiology of ASD. Moreover, the shared patterns of MAE allele shifting in ASD and dup15q provide a potential convergent biological mechanism linking idiopathic ASD and dup15q syndrome. Similar to ASE genes (Fig. 2c), the snoRNA targets that change splicing in ASD and dup15q showed significant enrichment with the other target genes without splicing changes (Fig. 7d; Methods). These results are consistent with the bioinformatic predictions above and potentially implicate disruption of snoRNA-mediated splicing of synaptic ASD risk genes in the pathophysiology of ASD. Moreover, the shared patterns of MAE allele shifting in ASD and dup15q provide a potential convergent biological mechanism linking idiopathic ASD and dup15q syndrome. Similar to ASE genes (Fig. 2c), the snoRNA targets that change splicing in ASD and dup15q showed significant enrichment with two different datasets of de novo ASD risk variants (in ASD, OR = 2.30 (P = 0.04) for the first and OR = 7.42 (P = 0.005) for the second; in dup15q, OR = 2.14 (P = 0.05) for the first and OR = 5.97 (P = 0.02) for the second; Fig. 7e).

Discussion
This study provides a large-scale, genome-wide investigation of ASE patterns in human brain samples from people with ASD and ASDdup.
We identify distinct patterns of ASE in ASD brain samples compared with those from controls, including overall fewer sites showing ASE in ASD. However, when ASE does occur, ASD samples show a preferential minor allele predominance, rather than the usual pattern of major allele predominance, particularly in instances of pure MAE. Loci exhibiting ASE were also enriched for genes that harbor de novo ASD risk variation, consistent with the expectation that such genes could be highly dosage sensitive. It follows from this that ASE could increase ASD risk from such de novo mutations in highly dosage-sensitive genes. Moreover, since snoRNA target genes showing splicing changes are enriched in genes showing ASE and de novo ASD-associated risk variation, snoRNA-mediated splicing changes could enhance ASD risk in dosage-sensitive genes that harbor de novo risk variants. In ASD, loci showing minor allele MAE were enriched for known ASD-risk genes and genes associated with the PSD. Furthermore, we identified two orphan C/D box snoRNA rich regions at 14q32 and 15q11 strongly enriched for ASD-related MAE changes. These loci were also enriched for minor allele MAE in dup15q, demonstrating convergence between known genetic risk factors and DNA methylation changes in brain. The allele shift toward minor allele MAE in ASD highlights the potential importance of this understudied mode of gene regulation in ASD pathogenesis, and the snoRNA-mediated splicing changes point to potential biological disease mechanisms in brain.

Our results were enabled by an optimized pipeline that we developed to perform a large-scale genome-wide ASE investigation in ASD, overcoming multiple challenges. To maximize transcriptome-wide SNP collection, we combined results from multiple methods including SNP array, imputation and RNA-seq-based genotyping. This is especially important for identifying SNPs with MAE, which often are missed by RNA-seq-based genotyping. We generated a masked reference genome to filter out all potential reference-biased SNPs during mapping to increase the accuracy of ASE measurement and to reduce false positive results. We further filtered SNPs showing association with ethnicity to identify generalizable ASE patterns across populations. We used the largest number of human postmortem brain samples from people with ASD, including multiple brain regions per individual to bolster reproducibility of results. Finally, we employed a linear mixed-model approach that accounts for sample-level variation to quantify ASE patterns across ASD and control groups. This optimized work-flow can guide future large-scale genome-wide ASE studies in disease.

Preferential major allele expression is the most common pattern observed in most cases of MAE, presumably because it can provide a buffer against unexpected, potentially deleterious rare SNP risks that might increase risk for disease. Consistent with this interpretation, rare SNPs show more preferential major allele expression than common SNPs. Common SNPs, in general, show more balanced expression patterns, as expected for non-delaetorius alleles that have been through the filter of natural selection. Since rare alleles are more likely to be deleterious than common alleles, the decrease in preferential major allele expression in ASD and dup15q, which could typically protect against the effect of deleterious rare alleles, may contribute to the risk of developing ASD. This is supported by the overlap of genes showing this pattern in ASD postmortem brain with genes harboring known ASD risk variants.

We found a notable convergence between idiopathic ASD and dup15q, both of which show enrichment of allele shift from major allele MAE to minor allele MAE at specific loci on chromosomes 14 and 15. 15q11–q13 duplication causes hypermethylation at 15q11–q13 (ref. 19) as well as a genome-wide increase of minor allele MAE. Both idiopathic ASD and dup15q have increased minor allele MAE compared with control and the same trends of regional and snoRNA expression changes at the two regions showing a high frequency of shifts from the major to minor alleles. Idiopathic ASD shows extensive allele shifting at the 15q11 locus, which suggests that strong allele-specific methylation may be regulating gene expression in a similar fashion as the 15q11–q13 duplication itself. Because idiopathic ASD and dup15q share similar expression patterns within both of the allele shift rich regions, the splicing changes targeted by snoRNA may be related to disease pathogenesis.

An allele shift rich region was identified at 14q32 which also contains 2 dense clusters of over 50 microRNAs (miRNAs) within the delta-like 1 homolog-type 3 iodothyronine deiodinase (DLK1-DIO3) domain locus. These clusters are maternally expressed (Fig. 6) and share an upstream imprinting control region with the other genes in the allele shift rich region. Several of these miRNAs have been shown to be downregulated in cancer and SCZ, and their targets are enriched in axon guidance pathways. This could represent a point of convergence between ASD and SCZ, which are known to have significant phenotypic and genetic overlap.

These analyses highlight preferential minor allele expression and orphan C/D box snoRNA-mediated splicing changes as two forms of genetic regulation altered in ASD. Our ASE identification approach incorporates both allele frequencies and quantitative measures of allelic imbalance. We identified strong enrichment of minor allele-containing transcripts in ASD, which were further enriched for known ASD risk genes, including PSD components, RBFOX1 targets, and FMR1 targets. However, we recognize that the studies of PSD components, and of FMRP and RBFOX1 targets, rely on postnatal data. Prenatal studies could identify other pathways, including chromatin modifiers. However, we note that even using these postnatal data, we did identify overlap with de novo mutation-containing ASD risk genes, which include chromatin modifiers and transcriptional regulators. Furthermore, we identify evidence supporting orphan C/D box snoRNA-mediated expression changes that show overlapping expression changes with the 15q11–q13 duplication itself. Because idiopathic ASD shows extensive allele shifting at the 15q11 locus, which suggests that strong allele-specific methylation may be regulating gene expression in a similar fashion as the 15q11–q13 duplication itself. Because idiopathic ASD and dup15q share similar expression patterns within both of the allele shift rich regions, the splicing changes targeted by snoRNA may be related to disease pathogenesis.

Fig. 7 | Characterization of allele-shift-rich regions in ASD and dup15q. a, Regional expression changes of the shared allele-shift-rich regions across ASD and dup15q. The x axes are log_{10}(reads per kilobase of gene model per million mapped reads) (log_{10}(RPKM)), and two-tailed unpaired t-test P = 0.6397, P = 0.0016, P = 0.1538 and P = 0.0152 for 14q32 in ASD, 15q11 in ASD, 14q32 in dup15q and 15q11 in dup15q, respectively. Significantly downregulated regions are marked with an asterisk. 14q32 in ASD also has a lower mean than control, similar to the others. The minimum, first quartile, median, third quartile and maximum expression values of the boxplots are shown for 14q32 (control: 1.671, 2.048, 2.270, 2.442 and 2.806; ASD: 0.7351, 2.0643, 2.3355, 2.5048 and 2.9375; dup15q: 0.9286, 1.8464, 2.1859, 2.2984 and 3.0108, respectively) and 15q11 loci (control: 2.039, 2.890, 3.130, 3.302 and 3.524; ASD: 0.3255, 2.5385, 2.8784, 3.2264 and 3.7170; dup15q: 1.114, 2.537, 3.105, 3.208 and 3.430, respectively). b, snoRNA gene expression changes in ASD and dup15q. Yellow and blue backgrounds indicate 14q32 and 15q11, respectively. Among snoRNA genes, 51 genes (RPKM ≥ 1) are selected as expressed. Based on a linear mixed-model based differential expression gene analysis, significantly downregulated SNORD116-24 genes (ASD: P = 0.0347; dup15q: P = 0.0019) are marked with asterisks (*P ≤ 0.05; **P ≤ 0.001). c, The numbers of snoRNA target genes and genes with splicing changes in ASD and dup15q. Two-sided Fisher’s exact test P values are shown at the bottom of tables. d, Gene set enrichment analysis for snoRNA target genes. Among snoRNA target genes, we compared splice changes and the other genes in ASD and dup15q brain samples. The labels ‘mono major’ and ‘mono minor’ are major and minor allele MAE genes, respectively. The labels ‘expression up’ and ‘expression down’ represent significantly up- and downregulated genes in idiopathic ASD. e, Enrichment of snoRNA target genes with candidate psychiatric disorder risk genes, as defined in Fig. 2c. Plots show ORs and P values if significant. For the GWAS datasets defined in Fig. 2c, plots show FDR-corrected P values if significant (Methods). For a, b, d and e, RNA-seq sample numbers of control, ASD and dup15q are 69, 62 and 15, respectively, Dup, participants with dup15q.
splicing changes in ASD brain, providing evidence of an association between snoRNA and ASD.

Among all forms of small noncoding RNAs, miRNAs have received the most focus as a target for ASD and other psychiatric disease studies38,42. Although the orphan C/D box snoRNA is a small family of snoRNAs existing only at chromosomes 14q32 and 15q11, this family controls the splicing of a unique set of about 400 target genes that are distributed across the genome. Although we could not find enrichment for common ASD risk in these targets, it may be due to simple lack of power of current GWAS studies in...
ASD. The most recent ASD GWAS contained over 45,000 participants and yet identified an order of magnitude lower significant loci than a similar size study of SCZ, consistent with substantially less power in ASD. Once larger, more conclusive GWAS results are available, future studies can identify whether this group of genes is enriched in a common variation that could contribute to polygenic risk for ASD. The allele shift rich regions in both idiopathic ASD and dup15q are directly overlapping with these orphan C/D box snoRNA repeat loci. Remarkably, we identified splicing changes in the downstream targets of these snoRNA genes, providing a potential mechanistic link between genetic, epigenetic and transcriptomic changes in ASD. These results highlight genetic and epigenetic regulation in ASD, identifying biological mechanisms warranting further investigation.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41593-019-0461-9.

Received: 12 December 2017; Accepted: 9 July 2019; Published online: 26 August 2019

References
1. Elsabbagh, M. et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res. 5, 160–179 (2012).
2. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 14, 1109–1120 (2015).
3. Voinoagu, I. et al. Transciptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474, 380–384 (2011).
4. Parikshak, N. N. et al. Genome-wide changes in lncRNA, splicing, and dup15q are directly overlapping with these orphan C/D box snoRNA repeat loci. Remarkably, we identified splicing changes in the downstream targets of these snoRNA genes, providing a potential mechanistic link between genetic, epigenetic and transcriptomic changes in ASD. These results highlight genetic and epigenetic regulation in ASD, identifying biological mechanisms warranting further investigation.

5. Nardone, S. et al. DNA methylation analysis of the autistic brain reveals molecular pathology. Nature 474, 380–384 (2011).
6. Parikshak, N. H. et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).
7. Darnell, J. C. et al. FMRP stalls ribosomal translation on mRNAs linked to synaptic function and autism. Cell 146, 247–261 (2011).
8. Mukherjee, N. et al. Integrative regulatory mapping indicates that the RNA-binding protein Hu couples pre-mRNA processing and mRNA stability. Mol. Cell 43, 327–339 (2011).
9. Weyn-Vanhentenryck, S. M. et al. HTIS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep. 6, 1139–1152 (2014).
10. Zhang, Y. et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J. Neurosci. 34, 1129–1149 (2014).
11. Pisani, F. et al. The connection between autism spectrum disorders and schizophrenia: a major role for immune activation in the pathogenesis of ASD. Mol. Psychiatry 24, 1057–1064 (2016).
12. Castel, S. E., Levy-Moonshine, A., Mohammad, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 16, 195 (2015).
13. de la Chapelle, A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. Oncogene 28, 3345–3348 (2009).
14. Scolos, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & Lassalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. Mol. Autism 2, 19 (2011).
Acknowledgements
We thank S. Mangul for helpful advice about his published ASE data-filtering method. We also thank N. N. Parikshak, Y. E. Wu and T. G. Belgard for providing published RNA-seq and snRNA-seq data and helpful advice. We also thank K. Y. Choe for proofreading. This work was supported by Simons Foundation grant no. 401457 to D.H.G., Simons Foundation Bridge to Independence Award to M.J.G., and National Institutes of Health grants no. U01MH116489, no. R01MH110927 and no. R01MH109912 to D.H.G.

Author contributions
D.H.G. and C.L. designed the study and D.H.G. supervised the project. C.L. analyzed the data. E.Y.K. helped with generating the masked reference file under guidance from E.E., who helped oversee ASE analysis. C.L., D.H.G. and M.J.G. wrote the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0461-9.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.H.G.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

Samples, RNA-seq and genotyping. As in our previous study, human postmortem brain samples were acquired from the Autism Tissue Program at the Harvard Brain Bank, the University of Maryland Brain and Tissue Bank (a Brain and Tissue Repository of the National Institutes of Health NeuroBioBank), the UK Brain Bank for Autism and Related Developmental Research and the Medical Research Council (MRC) London Neurodegenerative Diseases Brain Bank. A total of 263 brain samples were collected from BA9, BA41 and vermis of 40 controls and 56 ASD cases including 8 cases with dup15q syndrome. Though most sample information was already published, we provide it again including additional information (Supplementary Table 1). After RNA-seq library preparation with rRNA depletion (Illumina TruSeq v2 with RiboZero Gold (Illumina)), RNA-seq was performed using Illumina HiSeq2000 to generate 50-base pair (bp) paired-end reads as published. Sample genotyping was performed using Illumina Omni2.5 SNP array.

Preparing a masked reference and RNA-seq data mapping. To avoid reference bias, we first prepared a masked reference genome. We compiled the set of SNPs (39,683,000 SNPs) from multiple sources including those present on SNP arrays (Omni2.5 SNP: 2,372,783 SNPs; Affymetrix: 905,721 SNPs), those present in 1000 Genomes (38,270,182 SNPs) and those identified in the brain samples from RNA-seq-based SNP calling (3,163,431 SNPs).

RNA-seq-based SNP calling was performed using GATK and Samtools. To increase their SNP calling accuracy, we optimized bam files following the recommended GATK pipeline (Supplementary Fig. 1a). After updating mapping quality scores, we removed headers and read groups. After reordering chromosomes and marking duplicate reads, we recalibrated quality scores. Possible false positive SNP calling data were filtered out based on the common outputs of GATK and Samtools. We removed indels and retained only single base variants (for example, SNPs). When multiple SNPs were present at the same genomic coordinate, we randomly kept only one for downstream analysis. Using these collected SNPs, we masked the 19p reference sequence with randomly selected third alleles to minimize systematic mapping bias. We mapped all RNA-seq data to this masked reference genome using TopHat2 (ref. 55; Ensembl v73 annotations and the following parameters: tophat -g 10 -p 8 × 99 --no-unc-juncs -G).

Filtering out reference-biased SNPs. For further prevention of potential reference bias mapping by TopHat2, we generated a simulated RNA-seq dataset consisting of all possible 50-bp reads overlapping each SNP for both reference and alternative alleles (100 reads total per SNP). We mapped these simulated reads to the masked reference genome as described above. We identified any SNPs that showed reference bias (despite mapping to the masked reference genome) using chi-squared test (P ≤ 0.05 and read depth < 10). These reference-biased SNPs were removed from downstream analysis.

Genotyping. RNA-seq data were imputed to 1000 Genomes using Mach/Minimac. Imputed SNPs were filtered out by their output quality (filtered out based on R² and score values: R² < 0.5 and score > 3) and Hardy–Weinberg equilibrium (HWE; P < 1.0 × 10⁻⁶) to yield 10,992,184 high-quality imputed SNPs.

To generate additional genotyping coverage, we performed RNA-seq based genotyping using the following parameters:

- Homozygous SNP: a₁ ≥ 4, a₁ > 10 × a₂, and a₃ < 0.5 × a₁
- Heterozygous SNP: a₂ ≥ 2, a₁ ≤ 10 × a₂, and a₃ < 0.5 × a₂

We called the SNPs based on the first-, second- and third-most-abundant allele counts (a₁, a₂ and a₃) in mapping transcripts, respectively. We filtered out the genotyped SNPs if they did not pass the HWE filter (P < 1.0 × 10⁻⁶). Compared with the previously published RNA-seq-based genotyping methods by Quinn et al., these genotyping parameters showed much higher accuracy with sensitivity exceeding 95% (Supplementary Fig. 1b) and specificity in the range of 90–95% (Supplementary Fig. 1c; see Quinn et al. for accuracy tests).

We integrated the three types of genotyping data, RNA-seq array, imputation and RNA-seq-based genotyping. If we observed discordance among these methods, we weighted the data in the following order, favoring SNP array > imputation > RNA-seq-based genotyping.

Ancestry identification. Using the integrated genotyping data, we identified ancestry by a multidimensional scaling plot with HapMap3 populations. We categorized our samples into European, Mexican, Asian, African and ambiguous groups (Supplementary Fig. 1d).

Group-based ASE identification. We first studied ASE patterns within each tissue group (BA9, BA41, vermis and cortex (BA9 and BA41)), considering samples from brains of people with idiopathic ASD and control samples. In this part of our study, we excluded dup15q samples from the other ASD cases since they have known structural chromosomal alterations that cause ASD. We refer to the remaining ASD cases as ‘idiopathic ASD’, for which the cases are unknown. Control brain samples were acquired from individuals without a diagnosis of ASD or any known structural chromosomal alteration.

First, we collected SNPs if ‘good’ SNPs were present in 80% of samples. SNPs were defined as ‘good’ SNPs if they had the third and fourth allele counts less than 5% of the major allele. We also collected SNPs considered present in at least 20% of samples. ‘Present’ is defined as the percentage of individuals that have more than ten allele counts. ‘Good’ and ‘present’ filters account for expression levels and RNA-seq read mapping error, respectively.

For each heterozygous SNP, we counted the number of reads mapping to each allele. To normalize for differences in library size, results were converted to RPSM values defined as:

\[ \text{RPSM} = \frac{\text{allele mapped read no.} + 1}{\text{total uniquely mapped read no.}} \times \frac{\text{SNP area length}^{\text{bp}}}{1,000,000} \]

The SNP area length is calculated as (read length × 2) – 1, and mapped reads can reach an SNP within the SNP area.

For each group, ASE was quantified at SNP level using a linear mixed model (log(RPSM) = allele + age + sex + sequencing batch + RNA integrity number + brain bank + ancestry, rand = [biological replicate/individual]). Fixed effects included allele, age, sex, sequencing batch, RIN, brain bank and ancestry. Biological replicate and individual were included as random effects. Biological replicate represents the RNA-seq data replicates. To ensure that groups were balanced across replicates, we removed these covariates from the mixed model.

The AT and GC dropout values are the percentages of misaligned reads at low GC (GC < 50%) or high GC (GC > 90%) conditions, respectively. In BA9 samples, the distributions of the covariate values were balanced between control and ASD (Supplementary Fig. 1e). When we identified ASE from an additional linear mixed model including the covariates, the covariates did not show significant P values without affecting the ASE results.

We investigated both autosomal and chromosome X. For no PARs on chromosome X, we considered only female data. PARs were defined as: PAR1(chrX:60,001-2,699,520, chrY:10,001-2,649,520), PAR2(chrX:154,931,044-155,260,560, chrY:59,034,050-39,363,566) and PAR3(chrX:88,400,000-92,000,000, chrY:3,440,000-5,750,000).

Based on the P values of the allele ancestry covariate from the linear mixed model regression, we filtered out SNPs showing ethnicity-biased allele expression. After LD-based SNP pruning using PLINK (R² cutoff = 0.2), we counted the number of LD-independent SNPs and identified ASE SNPs passing Bonferroni correction.

To filter out potential indel-inducing reference bias from the identified ASEs, we considered 1,450,137 imputed indels identified during the above genotyping step. Among them, we selected only accurately 1,040,317 indel outputs from Mach/Minimac (deletions (≤ 5bp) and insertions (1bp)). We further filtered them out by the imputation quality (R² < 0.5 and score > 3) and HWE (P < 1.0 × 10⁻⁶), missing genotype ratio (< 0.05) and MAF (< 0.01) to yield high-quality imputed indels. To test indel-inducing reference bias, we selected 457,817 indels that showed heterozygous types in at least one sample.

Based on the masked reference genome, we generated a simulated RNA-seq dataset (50mer fastq file), which overlaps each indel area. It contains both indel-containing and non-containing sequences. We aligned the fastq file to the masked reference genome using TopHat2 and identified possible indel-inducing reference biases (chi-squared test P ≤ 0.05) from their proximal reads (read depth ≥ 10). Once SNPs showed the reference biases from at least one sample with heterozygote genotypes from both indels and SNPs, we filtered them out from the ASE results. To genotype SNPs, we integrated SNP array, imputation and RNA-seq-based genotyping as described above.

To compare ASE across groups, Manhattan plots were generated with the P values of the allele covariate from the linear mixed model regression using the qman R package (https://cran.r-project.org/web/packages/qman/; Fig. 3b). If ASE SNPs were located in gene bodies, we called the genes as ASE genes. The ASE genes were compared using area-proportional Venn diagrams (Fig. 3c).

Individual-based ASE identification. We next studied ASE within individual samples, comparing differences in allele counts for heterozygous SNPs. We used only ‘good’ SNPs in at least 80% of samples (as defined above) without considering the present filter. P values for each SNP were calculated using Fisher’s exact test. After LD-based SNP pruning, ASE SNPs were identified by Bonferroni correction (as described above). We filtered out indel-inducing reference-biased SNPs as described above.

For each cortical sample, ASE rates were calculated per chromosome as the number of ASE SNPs/total number of tested SNPs. Differences in the ASE rate were tested using Wilcoxon rank sum test (Fig. 3d).
Evaluation of identified ASE. To test ASE identification accuracy, we identified haplotype blocks using PLINK with the SNPs showing no ethnicity-biased allelic expression. We calculated deviations of log(fold change) values (the beta values of allele covariance from the group-based ASE identifications) per haplotype block in cortex and compared their distributions with normal distributions (Supplementary Fig. 1f,g). For further validation, we compared ASE genes identified in our study with those present in dbMAE (https://mae.hms.harvard.edu/) using Fisher's exact test (Supplementary Fig. 2a). Of note, dbMAE mostly includes nonneuronal tissues, and we considered its human and mouse data. Nevertheless, we observed strong overlap between them, providing validation of our ASE genes. For additional external validation, we generated data as following GTEX data analysis. Similar to the method for GTEX data, we considered only heterozygous SNPs and followed the same P value cutoff of GTEX data. We considered the ASE SNPs once their P values were below 0.005 at the previous individual-based ASE identification. Similar to the GTEX data calculation, ASE data was also tissue (Supplementary Fig. 2c). To compare the GTEX data with our data, we compared the mean of ASE rates from the tissues.

ASE and expression comparison. We calculated the ‘less expressed allele expression fraction per heterozygous SNP’ from one sample, UMB303 (tissue: BA41). The fraction is a2/(a1 + a2). a1 is the more expressed allele read count, and a2 is the less expressed allele read count. If the SNP had more than ten mapped reads, we calculated log(RPSM). Based on the less expressed allele expression fraction, we grouped SNPs into ‘imbalanced’ and ‘other’ groups. The fractions of the imbalance group output or less than 30%. The fractions of the other group were greater than 30%. We compared their log(RPSM) values using Kolmogorov–Smirnov tests (Supplementary Fig. 2d).

Gene set enrichment. Gene set enrichment analyses were performed using logistic regression accounting for gene length as a covariate. Heatmaps were prepared showing HR and false discovery rate (FDR) for SNP P values for different tissues, if significant. For the known ASD risk gene list (SFARI gene), we selected genes from categories 1, 2 and 3 corresponding to gene scoring criteria (https://gene.sfari.org/autdb/GS_Home.do; Supplementary Table 6). The gene list was used in a previous publication.

For these gene set enrichment studies, we also used known imprinted genes (www.geneimprint.com); PSD4, FMRP28, HuR29 and RBFOX1 target genes; cell marker genes (neuron, astrocyte, oligodendrocyte, microglia, and endothelial); and expression up- and down-regulated genes in ASD cortex (Supplementary Table 6). The HuR target gene list was collected from photoactivatable ribonucleoside cross-linking and immunoprecipitation (PAR-CLIP) data instead of RNA-binding proteins immunoprecipitation complementary DNA array (RIP-chip) data since PAR-CLIP data identified RNA-binding protein binding sites more precisely than RIP-chip (Supplementary Table 6).

For the de novo variant data, we used the risk genes containing rare de novo likely gene-disrupting mutations in SCZ, ID and ASD from the dataset of Lossio et al. (Supplementary Table 6). Since this list is not based on the most stringent statistical thresholding (353 genes for ASD), we used additional risk gene lists from other work to cross-reference. This includes Sanders et al., who identified risk genes after using the transmission and de novo association model for the risk variants from psychiatric diseases, we calculated the log(fold change) from the linear mixed model regression outputs. We considered 51 snoRNAs that have RPMs greater than or equal to 1 (Fig. 7b).

Regional expression and splice junction analysis. To study regional expression changes in allele shift rich regions, we calculated their log(RPMS)-values from uniquely mapped reads of RNA-seq data and compared them between groups (Fig. 7a).

snoRNA target splicing change analysis. For orphan C/D box snoRNAs at the allele shift rich regions, we relied on previously identified splicing targets. We identified splicing-changing genes that show β > 2 s.d. alternative splicing change (beta values from the previously published linear mixed model regression data) in ASD and dup15q versus controls. Using Fisher’s exact test, we assessed whether the snoRNA target genes show more alternative splicing changes in ASD and dup15q versus other non-snoRNA targets (Fig. 7c). Among the snoRNA target genes, we performed a gene set enrichment analysis for genes showing significant splicing changes (Fig. 7d).

Summary of statistical methods. Sample sizes. No statistical methods were used to predict sample sizes, but our sample sizes are larger than those reported in previous publications that studied ASE.

Normality of data distribution. For ASE identification and quantitative allele imbalance study, normality was not formally tested, but data distribution was assumed to be normal.

Randomization. To avoid reference bias, we randomized allele selection for a masked reference preparation.

Blinding. Data analysis was not performed blind to the metadata information of the brain samples.

Accession codes. Raw RNA-seq and sncRNA-seq data from brain samples were deposited to the PsychENCODE Knowledge Portal (https://doi.org/10.7303/syn4587609).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The detailed description of brain samples is provided in Supplementary Table 1. For each tissue, group-based ASE identification results are available in Supplementary Tables 2 and 3. We also include gene lists that we used for gene set enrichment and GO analysis, which include brain-expressed genes that we used for our study (Supplementary Table 6). Raw next-generation sequencing data from human postmortem brain samples are available from published RNA-seq and sncRNA-seq studies. They have been deposited to the PsychENCODE Knowledge Portal (https://doi.org/10.7303/syn4587609).
Code availability
The R code for the ASE identification using a linear mixed model is provided in the Supplementary Software.

References
50. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
51. Li, H. et al. The Sequence Alignment/Map format and SAM tools. Bioinformatics 25, 2078–2079 (2009).
52. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2015).
53. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. Bioinformatics 31, 782–784 (2015).
54. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).
55. Quinn, É. M. et al. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One 8, e58815 (2013).
56. Veerappa, A. M., Padakannaya, P. & Ramachandra, N. B. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct. Integr. Genom. 13, 285–293 (2013).
57. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
58. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn—a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9, 488 (2008).
59. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. 11, e1004219 (2015).
60. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GoRilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009).
61. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput. Biol. 3, e39 (2007).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
63. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
64. Mortazavi, A., Williams, B. A., McCue, K., Schaefler, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628 (2008).
65. Lee, C., Mayfield, R. D. & Harris, R. A. Altered gamma-aminobutyric acid type B receptor subunit 1 splicing in alcoholics. Biol. Psychiatry 75, 765–773 (2014).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
  - Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
- We did not collect data since the data were already published (Parikshak et al., Nature 2016; Wu et al., Nat. Neurosci., 2016).

Data analysis
- R and python were used to analyze the data. For the data process, we also used GATK (version 2.7-2), Samtools (version 0.1.19), TopHat2 (version 2.0.9), Bowtie2 (version 2.1.0), Mach (version 1.0), Minimac (version beta-2013.7.17), Picard (version 1.100), PLINK (version 0.08), MAGMA (version 1.07b), FASTX-Toolkit (version 0.0.13.2), and htsq-ccount (version 0.6.1p1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that we analyzed were already published at the previous studies (Parikshak et al., Nature 2016; Wu et al., Nat. Neurosci., 2016). We could access their raw data directly from the groups.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No statistical method was used to predetermine the sample size, but the sample size used in our study is comparable with the previous studies (e.g. Parikshak et al., Nature 2016; Wu et al., Nat. Neurosci., 2016) |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | To ensure if groups were balanced with respect to age and brain bank, we removed samples based on age (<=7) and brain banks (The UK Brain Bank for Autism and Related Developmental Research and the MRC London Neurodegenerative Diseases Brain Banks). |
| Replication | The genes with allele specific expression which were reported in this manuscript showed significant overlap with the previous study (Savova et al. Nucleic Acids Res., 2016). The findings of allele specific expression were confirmed using multiple analytical approaches on the same dataset showing the robustness of our findings. |
| Randomization | To avoid reference bias, we prepared a masked reference based on allele randomization. |
| Blinding | Data analysis was not performed blind to the metadata information of the samples. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems | Methods
--- | ---
| n/a | n/a |
| [ ] Antibodies | [x] ChIP-seq |
| [ ] Eukaryotic cell lines | [x] Flow cytometry |
| [ ] Palaeontology | [x] MRI-based neuroimaging |
| [ ] Animals and other organisms | |
| [ ] Human research participants | |
| [ ] Clinical data | |

Human research participants

Policy information about human research participants

Population characteristics | As described in the previous study (Parikshak et al., Nature 2016), human postmortem brain samples were acquired from the Autism Tissue Program at the Harvard Brain Bank, the University of Maryland Brain and Tissue Bank, the UK Brain Bank for Autism and Related Developmental Research, and the MRC London Neurodegenerative Diseases Brain Bank. A total of 263 brain samples were collected from BA9, BA41-42-22, and vermis of 40 controls and 56 autism spectrum disorder cases including 8 cases with dup15q syndrome.

Recruitment | No recruitment was processed since we used publically available datasets.

Ethics oversight | No ethics oversight was required because we analyzed open-source datasets.

Note that full information on the approval of the study protocol must also be provided in the manuscript.