Title

Risk alleles of genes with monoallelic expression are enriched in gain-of-function variants and depleted in loss-of-function variants for neurodevelopmental disorders

Journal

Molecular Psychiatry, 22(12)

ISSN

1359-4184

Authors

Savova, V
Vinogradova, S
Pruss, D
et al.

Publication Date

2017-12-01

DOI

10.1038/mp.2017.13

Peer reviewed

HHS Public Access

Author manuscript

Mol Psychiatry. Author manuscript; available in PMC 2017 November 30.

Published in final edited form as:

Mol Psychiatry. 2017 December ; 22(12): 1785–1794. doi:10.1038/mp.2017.13.

Risk alleles of genes with monoallelic expression are enriched in gain-of-function variants and depleted in loss-of-function variants for neurodevelopmental disorders

Virginia Savova, PhD[1], **Svetlana Vinogradova, PhD**[1], **Danielle Pruss, BS**[1], **Alexander A. Gimelbrant, PhD**[1], and **Lauren A. Weiss, PhD**[2]

[1]Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 450 Brookline Ave., Boston, MA 02215

[2]Department of Psychiatry and Institute for Human Genetics, University of California San Francisco, 401 Parnassus Ave, San Francisco, CA 94117

Abstract

Over 3,000 human genes can be expressed from a single allele in one cell, and from the other allele – or both – in neighboring cells. Little is known about the consequences of this epigenetic phenomenon, monoallelic expression (MAE). We hypothesized that *MAE increases expression variability*, with potential impact on human disease. Here, we use a chromatin signature to infer MAE for genes in lymphoblastoid cell lines and human fetal brain tissue. We confirm that across clones, MAE status correlates with expression level, and that in human tissue datasets, MAE genes show increased expression variability. We then compare mono- and biallelic genes at three distinct scales. In the human population, we observe that genes with polymorphisms influencing expression variance are more likely to be MAE ($P < 1.1 \times 10^{-6}$). At the trans-species level, we find gene expression differences and directional selection between humans and chimpanzees more common among MAE genes ($P < 0.05$). Extending to human disease, we show that MAE genes are underrepresented in neurodevelopmental CNVs ($P < 2.2\times10^{-10}$) suggesting that pathogenic variants acting via expression level are less likely to involve MAE genes. Using neuropsychiatric SNP and SNV data, we see that genes with pathogenic expression-altering or loss-of-function variants are less likely MAE ($P < 7.5\times10^{-11}$) and genes with only missense or gain-of-function variants are more likely MAE ($P < 1.4\times10^{-6}$). Together, our results suggest that MAE genes tolerate a greater range of expression level than BAE genes and this information may be useful in prediction of pathogenicity.

Correspondence to: Alexander A. Gimelbrant; Lauren A. Weiss.

Correspondence: Dr AA Gimelbrant, Dana-Farber, Cancer Institute, Department of Genetics, Harvard Medical School, Boston, MA 02215, USA or Dr LA Weiss, Department of Psychiatry and Institute for Human Genetics, University of California San Francisco, Langley Porter Psychiatric Institute, Nina Ireland Lab, Box F-0984, 401 Parnassus Avenue, Room A101, San Francisco, CA 94143-0984, USA. gimelbrant@mail.dfci.harvard.edu or Lauren.Weiss@ucsf.edu.

Supplementary information is available at Molecular Psychiatry's website.

We have no financial conflicts of interest to disclose.

INTRODUCTION

In recent years, we and others have discovered that a large fraction of human genes (10 - 30%) are subject to monoallelic expression in a tissue-dependent manner. The known biological properties of this epigenetic mechanism have been reviewed recently[1], and the most relevant aspects are summarized in brief: 1) MAE is a cell-type specific property for each gene; the majority of MAE genes show variable patterns across cell-types with constitutive biallelic expression (BAE) in some cell types and MAE in others. 2) Allelic choice for each gene undergoing MAE is independent. Multiple genes are simultaneously subject to MAE in a given cell and its clonal progeny, and undergo independent choice at each locus. 3) MAE is highly mitotically stable over multiple cell divisions. 4) MAE status of genes strongly correlates with a specific gene-body chromatin signature; a polymorphism-independent approach based on this (Monoallelic Gene Inference from Chromatin; MaGIC) can be used in complex samples to predict MAE status in tissues of interest[2, 3].

MAE is known to generate otherwise similar cells with different functional properties based on allelic composition. For example, the MAE status of *Tlr4* in a heterozygous NULL results in two distinct cellular populations with respect to LPS response[4]. However, sources of MAE-related cell-to-cell variation beyond functional allelic differences have remained relatively unexplored. It is rarely noted that while other types of monoallelic expression (e.g. olfactory allelic exclusion, imprinting, X-inactivation) involve obligatory choice of only one allele in any single cell or clone, a typical MAE gene can be expressed biallelically in some clonal lineages and monoallelically in others within the same tissue- or cell-type. In light of overall evidence of random clonal make-up of a given tissue[5], this observation strongly suggests that different individuals will have different proportions of monoallelic- and biallelic-expressing cells for a specific MAE gene in a relevant tissue. The consequences of this epigenetic source of variation for cell-to-cell variability and individual variation are currently unknown.

Specifically, it has been shown that MAE genes in a monoallelic state have lower expression levels than in a biallelic state in the scenarios tested to date[6]. However, assessing the consequences of silencing one allele for overall gene expression levels in individual cells and in tissues across the population has been technically challenging due to methodological issues and small numbers of confirmed MAE genes. The relationship between monoallelic state and expression range in individual clones has not been tested in human cells beyond individual examples[7], and other evidence remains equivocal, as one group has argued for possible cases of dosage compensation in mouse NPCs[8].

Further, it is not known what effect expression variability across cells has on tissue-level expression variability. One possibility is that MAE is a mechanism for exerting tight control over expression level in a developing tissue by lowering expression level in a fraction of cells[6]. It has been previously proposed that a stochastic model with limiting amounts of transcription or chromatin factors could, by chance, generate a population with zero, one, or two active alleles consistent with MAE. This model carries the strong implication of tight control in overall expression level of MAE genes and would predict that expression

variability across individuals would be unlikely. However, it is not clear how this model applies to mitotically stable MAE.

We hypothesized that MAE is likely to increase *expression level variability* not only among cells in the same tissue but also among tissues in different individuals as a result of clonal composition. Based on previous observations[5], we assume that clonal composition is subject to random developmental events in otherwise identical genotypes. Consequently, proportions of monoallelic and biallelic active cells for any given MAE gene are likely to differ across individuals. If monoallelic expression is lower than biallelic expression, individuals with increased proportions of monoallelic clonal lineages will exhibit lower expression level in the tissue overall than individuals with increased proportions of biallelic lineages.

Even small gene expression level differences are known to have a major impact on human variation and in evolution. For example, common DNA polymorphisms associated with expression level (eQTLs) have been shown to have major functional enrichment in genome-wide association studies for a variety of complex heritable traits, and copy number variants which have a large impact on the expression of the genes within them are often pathogenic for neurodevelopmental disorders[9-14]. Similarly, recent data suggests that epigenetic and regulatory changes contribute disproportionately to evolutionary difference, particularly in the brain[15-17]. Therefore, if MAE impacts expression variance, it may have consequences for the function and evolution of MAE genes, and leave a specific signature on human variation and disease.

To assess the validity of our hypotheses, we first assessed expression differences in MAE genes with different allelic state (monoallelic or biallelic) in human clonal cell lines. We then used a previous catalog of MAE genes[18] identified in lymphoblastoid cell lines (LCLs) using the MaGIC method, and generated a novel set of such genes in human fetal brain by the same method, in order to investigate the extent and consequences of MAE expression level variation in existing datasets. We examined human expression data to show that the variance of MAE genes is indeed higher than for BAE genes, and used genetic mapping of human expression variability data to show that MAE genes are more likely to have significant expression variability quantitative trait loci (evQTLs) than BAE genes.

To assess the long-term consequences of the MAE-induced expression variation, we analyzed data describing species expression level differences between humans and non-human primates. Genes with MAE status have higher inter-species expression variance and are more likely to have acquired differences in expression level during evolution compared with genes classified as constitutive BAE. We also performed an analysis of pathogenic neurodevelopmental copy number variant (CNV) data in humans to show that MAE genes are represented in polymorphic CNVs in the population but specifically depleted in pathogenic CNVs, suggesting that they are genes for which a range of expression level is tolerated in humans. We support this disease-related hypothesis by analyzing common and rare variants associated with schizophrenia and autism, where we find that risk factors predicted to act on expression level underrepresent MAE genes and those predicted to act via potential gain-of-function are enriched for such genes. Thus, tissue-specific epigenetic characteristics of genes may be complementary to current information about mutational

impact and evolutionary conservation in future research and clinical prediction of genetic variant pathogenicity.

MATERIALS AND METHODS

Categorization of genes as MAE or BAE based on histone marks

Analysis was performed on a set of MAE and BAE genes compiled from a recently published newly-available genome-wide dataset[2]. Briefly, the dataset indicates the monoallelic or biallelic status of genes in six human cell types represented by seven cell lines, on the basis of an experimentally tested specific chromatin signature (co-occurrence of H3K27me3 silencing mark and H3K36me3 active mark on the gene body), and expression data. We selected the GM12878 (lymphoblast) cell line for our analysis of MAE in blood. To predict that a gene is MAE, we required classified MAE status with expression rank of 50% or more after quantile rank normalization of expression levels. In GM12878 this corresponds to RPKM ~1, or 5-10 transcripts per cell. The threshold was used to exclude any genes which are monoallelic only in non-physiologically relevant expression levels. To consider a gene BAE, we required expression rank of 50% and inferred BAE status. Thus, if a gene was inferred monoallelic only at low expression (rank < 50%), it was not included in the MAE set, nor was it considered to be positively BAE, and was therefore excluded from consideration. In addition, short genes (<2,500) were excluded from the analysis because the gene body chromatin signature signal was found to be less reliable for such genes[3]. To generate predictions for fetal brain, we used the ChIP-Seq dataset from Epigenome Roadmap[19] which was analyzed according to the method previously described[2]. Briefly, H3K27me3 and H3K36me3 signal on the gene body was integrated and normalized to input signal. The resulting data was quantile-normalized and processed with a prediction model trained on an analogous dataset derived from human LCLs, for which the MAE and BAE gene status was experimentally verified[7].

Experimental determination of expression level differences of MAE genes in monoallelic compared to biallelic clones

We analyzed a previously published dataset (GSE52090) of RNA-seq on two lymphoblastic clonal cell lines. Details on the derivation of clonal cell lines and the RNA-seq library preparation can be found in the original publication[2]. Briefly, the clonal cell lines were derived from the GM12878 cell line obtained from the Coriell depository through single-cell sorting, and polyA-selected libraries were prepared from Trizol-extracted RNA and sequenced on an Illumina HiSeq 2000. The RSEM package[20] was used to compute expected read counts on hg19 with UCSC-derived gene models. The DESeq R-package (10.1186/gb-2010-11-10-r106) was used to estimate size factors for each library and for variance stabilization to allow for cross-clonal comparison of expression levels for genes with at most two-fold difference in expression (the expected dosage difference between monoallelic and biallelic state of an MAE gene).

To compare expression levels of the MAE genes, we selected all genes which were determined to be monoallelically expressed in one clone, and biallelically expressed in the other as group 1 (different state), and all genes either monoallelically or biallelically

expressed in both clones as group 2 (consistent state). Further subgroups were examined according to specific state in each clone (monoallelic in clone DF1 and biallelic in DF2; monoallelic in clone DF2 and biallelic in DF1; consistent -- monoallelic in both; consistent -- biallelic in both). A gene was determined to be monoallelically expressed if the proportion of reads with one variant genotype exceeded 2/3 (>67%) and the *P*-value of the binomial test for bias was less than 0.05. Benjamini-Hochberg FDR correction was used to correct for multiple comparisons. Biallelic expression was determined by equivalence testing, as described elsewhere[2]. Briefly, an interval of equivalence is defined for bias of +/−17%, and genes with bias estimate which confidently falls within this interval are considered biallelically expressed.

Human expression variance dataset

Gene expression data from the GTEx consortium [21] was downloaded from the GTEx portal (http://www.gtexportal.org/, version V6p) and processed using YARN [22] (bioconductor.org/packages/yarn). In total, we analyzed 393 whole blood samples, 118 EBL-transformed cell lines samples, and 108 frontal cortex samples. The full list of samples used is available in Table S1. We applied smooth quantile normalization [23] that normalizes samples in a tissue aware manner and returns a log-transformed normalized matrix. We filtered out genes with mean coverage across samples lower than 10 reads per gene and calculated the standard deviation as a gene-wise estimation of dispersion. To compare dispersions between genes with different levels of expression, we split genes into 5 interquartile bins according to the level of expression. We sampled 300 genes from each bin (if either of the MAE or BAE genes in the bin contained less than 300 genes, we sampled this amount of genes across this bin) and then applied Wilcoxon test to test for higher dispersion levels of MAE genes. The results of these calculation are available in Table S2.

evQTL Dataset

We utilized published data on human expression variance quantitative trait loci (evQTLs) in order to assess genetic regulation of expression variance[24]. In brief, the largest dataset used in this study was gene expression omnibus (GEO) dataset GSE6536, which contained measurement of LCL mRNA expression data from HapMap individuals from Illumina human whole-genome expression array (WG-6 version 1)[25]. These data included 16,992 genes (19,440 probes). Genotype data were obtained from the HapMap project for 210 unrelated CEU and YRI parent samples overlapping with the expression data. Genotype data were also extracted from the 1,000 genomes project (1KGP) for individuals overlapping with expression data. Regression models were used to identify significant SNP effects on variance $P < 1 \times 10^{-8}$ (equivalent to Bonferroni adjusted $P < 0.01$). For these putative significant loci, $P_{permutation} < 0.001$ and F-K test $P < 0.01$ for homogeneity of variance were also required for the final set of 166 evQTLs from GSE6536 considered in our study. We also consulted the lists of eQTLs reported in the original manuscript to be sure they did not bias our analysis[25].

Human and non-human primate expression dataset

In order to examine inter-species variability and evidence for natural selection, we utilized data from a previous study[26]. In brief, five each human, chimpanzee, and rhesus macaque

LCLs were used. Peptide expression was measured by stable isotope labeling by amino acids in cell culture (SILAC) using high-resolution quantitative mass spectrometry in order to compare protein expression levels. RNA-seq data was also collected from the same samples for comparisons of mRNA expression. Both mRNA and protein expression were measured in at least three individuals from all three species for 3,390 genes. For this set of genes, supplementary data was available including the results of a test of protein level species differences, mRNA level species differences, protein and mRNA variance, RPKM for mRNA measurements, and the best fitting selection model for both protein and mRNA: 1 = expression level pattern consistent with directional selection along human lineage, 2 = expression level pattern consistent with directional selection along chimpanzee lineage, 3 = undetermined pattern, 4 = patterns with no significant difference between mean expression levels; 5 = evidence for relaxation of constraint along human lineage, 6 = evidence of relaxation of constraint along chimpanzee lineage.

NDD CNV datasets

In order to test the hypothesis that genes in pathogenic copy number variants would be depleted for MAE genes, we used the datasets in Jeffries *et al*[27] (Supplementary File_S1; refs. [28-35]) and added several more recent datasets containing lists of pathogenic CNVs implicated in neurodevelopmental disorders[8, 36-38]. For each dataset, we extracted unique genes from the CNV regions reported and aligned these NDD CNV gene lists with the genes for which MAE/BAE status could be inferred from LCLs and from fetal brain (above).

NDD GWAS/SNV datasets

In order to test the hypothesis that variants likely to act via gene expression would show underrepresentation compared to variants likely to act via gain-of-function mechanisms, we examined data from several recent neuropsychiatric disease studies. First, we utilized the supplementary data provided by a schizophrenia genome-wide association study including annotation of associated SNPs (threshold $P < 5 \times 10^{-8}$)[39]. In brief, the relevant annotations were 1) genes in credible regions of an associated SNP (protein-coding genes based on GENCODE with 20 kb flanking region on each side or where there is no intersecting gene, the nearest gene within 500 kb), 2) blood eQTLs in Wright *et al*[40] with a transcript within 1Mb and $P_{eQTL}<1\times10^{-4}$; 3) brain eQTLs (meta-analysis of GSE8919, GSE15745, GSE30272, GSE15222[41-44]) with a transcript within 1Mb and $P_{eQTL}<1\times10^{-4}$, and 4) missense variants within the credible regions.

Next, we used supplementary data from two exome sequencing studies of SNVs for ASDs. Unfortunately, although missense mutations can act via gain-of-function or loss of function, there are no validated methodologies to predict without onerous experimentation. Thus we used several approaches to enrich our data under the assumption that genes acting via loss-of-function would show some enrichment for frameshift, nonsense, or splice site mutations (likely to result in no expression of an allele) and genes acting exclusively via gain-of-function would show only missense mutations but no enrichment for frameshift, nonsense, or splice site mutation in affected individuals. The first study provided an integrated analysis of *de novo*, inherited and case-control loss-of-function variant counts, as well as *de novo* missense variants predicted to be damaging with q < 0.3 by TADA (transmission and *de*

novo association), based on examination of over 3,800 cases and 9,900 controls [45]. We defined likely loss-of-function genes as those showing a *de novo* loss-of-function mutation (LoF: frameshift, nonsense, donor/acceptor splice site mutations) in at least one case and no LoF mutations in controls. We defined potential gain-of-function genes as those showing no LoF *de novo* mutations in cases with association signal thus coming primarily from missense (probably damaging, as defined by PolyPhen-2[45]) variants. (Note that the 'probably damaging' category in PolyPhen-2 does not refer specifically to loss of function, but only to the likelihood of altered function.) The second study only included *de novo* SNVs identified in 2,500 simplex families comparing cases and sibling controls[47]. We defined likely loss-of-function genes as those with *de novo* likely gene disrupting mutations (LGD: frameshift, nonsense, and splice site) in at least one case and none in control siblings and potential gain-of-function genes as those with at least two *de novo* missense mutations in cases and none in siblings and no *de novo* LGD mutations in cases. We tested the full gene list examined compared to the genome-wide LCL and fetal brain MAE/BAE genes and found no difference in representation, showing that exome sequencing adequately covered the genome. In order to identify genes likely to act in a recessive manner on autism, we utilized a previous dataset of genes with either rare homozygous or compound heterozygous nonsense or splice site (LoF) mutations in autism cases, but not in controls, provided in a supplementary table[48].

Haploinsufficiency, constraint, and OMIM datasets

In order to support our hypotheses about haploinsufficiency and gain-of-function, we utilized several published or public datasets. A previous publication included a supplementary table providing their predictions of haploinsufficient (HI) and haplosufficient (HS) genes[49]. We used predictions of at least 0.9 to indicate likely HI status and not more than 0.1 to indicate likely HS status. A recent study derived constraint scores in order to interpret pathogenicity of *de novo* mutations[50]. We downloaded a table of Z-scores for constraint including synonymous, missense, and loss-of-function mutations (http://atgu.mgh.harvard.edu/webtools/gene-lookup/lookupGene; downloaded 1/05/15). In order to predict gain-of-function, we calculated the difference between constraint Z-scores for missense and loss-of-function mutations and considered the top 10% of genes, under the assumption that genes more constrained for missense than loss-of-function mutations are most likely to act via gain-of-function. In order to identify human disease genes primarily acting via gain-of-function, we searched the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim). We used search terms 'lossoffunction' and 'gainoffunction' and retained unique genes annotated with only gain-of-function.

Statistical analysis

ANOVA test and Tukey's 'Honest Significant Difference' method was used to assess the differences in expression level between groups of different/same state genes in the analysis of expression change between two clonal cell lines. To confirm differences in MAE and BAE gene variance in the GTEx data, we used Wilcoxon test. To assess differential representation of MAE and BAE genes in evQTL targets and disease datasets, we used the chi-square test (or a one-sided Fisher exact test when expected counts were <5 for any cell). Meta-analysis of the 7 NDD CNV studies was performed using the Mantel-Haenszel method

as implemented in the R package (Meta, v 4.0). The two-tailed test of population proportion (z-test) was used to assess the difference in high-variance versus low-variance genes in evolutionary comparisons. A chi-square test was used to assess differences in MAE/BAE distribution across the three evolutionary model categories (directional, expression-constrained, and relaxed).

RESULTS

Inference of MAE genes in fetal brain

Our classification analysis of existing chromatin data from fetal brain dataset predicted 4690 genes as MAE and 7408 as BAE with rank of 50% or higher in expression. The relatively high number of genes with this expression rank (>12,000) is likely due to fact that the transcriptome is generally very broad in this developing tissue (76% of assayed genes)[51]. This, and the fact that moderate to low expression genes are more likely to be MAE, may contribute to the different estimates of MAE proportion in LCLs and fetal brain (0.13 vs. 0.39). However, it is also possible that the discrepancy is partially due to the fact that expression in this dataset was measured by microarray, which is more sensitive than RNA-seq for lower-expressed genes. The predictions are very significantly enriched with monoallelic genes from an existing human neuronal dataset of experimentally measured monoallelic expression[27]. While the study had an idiosyncratically low level of confirmed MAE as a whole (<2%), out of the 145 genes positively identified as MAE, 107, or 74% were predicted MAE by our classifier. The expected baseline rate in our dataset was 39% ($P < 2.2 \times 10^{-16}$). Further examination revealed that confirmed MAE genes from the Jeffries study were uniformly distributed within the predicted region based on the chromatin data, suggesting that the overall very low calling rate, rather than biological differences is likely to be responsible for the relatively small overlap of predicted and confirmed genes (Figure S1).

For a deeper validation of our predictions, we used experimental data from two independent approaches: we compared our predictions to published RNA-FISH in mouse fetal brain, and to genome-wide allele-specific RNA-Seq in multiple mouse clonal neuronal progenitor cells (NPCs). It shows that our chromatin inference is as accurate as experimentally-determined MAE status (Figure 1).

Both of these datasets show a good agreement between the genes' MAE status as inferred from the chromatin signature in human fetal brain and the MAE status of the orthologous mouse genes experimentally determined in mouse fetal brain and NPC clones. Considering the overall significant conservation of MAE status between human and mouse, this analysis provides stronger evidence than any plausible experiment using human samples.

Relationship between monoallelic transcription and mRNA level

Level of mRNA is the main output of the cell's transcription regulatory machinery. We have previously observed that the specific allelic state in an MAE gene between clonal cell lineages can have significant consequences for the transcript levels: for example, clones of human lymphoblastoid cells expressing a single allele of APP gene showed lower transcript level than clones with both alleles expressed equally, as measured by real-time PCR[7].

However, these observations were limited to a small number of genes. RNA-Seq analysis of mRNA from similar clones[2] provided means to assess whether this quantitative relationship holds genome-wide.

We analyzed a total of 42 genes with different status (monoallelic and biallelic) in two clones (DF1 and DF2) derived from the same human LCL (GM12878), and 617 genes with consistent monoallelic or biallelic status in both clonal lines. We found that genes with consistent biallelic or monoallelic status showed less expression level difference than genes with different status ($P < 0.01$; Figure 2a). Further, for genes with different status, expression in the monoallelic state was lower overall ($P < 1.1 \times 10^{-16}$; Figure 2b, Table S3). The effect was more pronounced in one direction, which was likely to be due to difference in coverage in the two sequencing libraries (higher coverage in the DF1 library is likely to lead to fewer false detections of allelic bias), and to lower average difference in bias between the mono- and biallelic state in the direction exhibiting less pronounced effect (0.21 versus 0.25). In contrast, having the same status (both monoallelic versus both biallelic) was not associated with any difference in expression level ($P < 0.95$).

Genes with MAE status show higher inter-individual expression variability in humans

Previous studies of clone-specific phenomena, such as X-inactivation, have shown that developmental processes lead to variability in clonal composition [5]. Since MAE genes are more variable in their expression level across clones, we hypothesized that they would also make a greater contribution to expression variability across individuals than BAE genes. To test this prediction, we took advantage of a large-scale publicly available gene expression RNA-seq dataset measuring expression in multiple primary human tissues – GTEx [48]. We predicted gene status as MAE and BAE using specific chromatin signature, as previously described[2]. As the source of chromatin data, we used the closest available cell type: to analyze GTEx expression data from 393 whole blood samples, genes were classified as MAE or BAE using chromatin data obtained in GM12878 lymphoblasts by the ENCODE project[52]. For the brain, 108 GTEx samples of human cortex were used, and genes were classified as MAE or BAE using ChIP-Seq data obtained from human fetal brain samples in the Epigenome Roadmap[19]. Standard deviation was consistently and significantly higher in all tested tissues for MAE genes compared to BAE genes, for all bins (Figure 3a and Table S2, Table S3). We thus conclude that expression variation between individuals is significantly higher for MAE compared with BAE genes.

Targets of evQTLs are enriched for MAE genes

Next, we hypothesized that if MAE-associated expression variation is functionally important within the human population, genes undergoing MAE would also be subject to germline polymorphism associated with expression variance. We thus analyzed data on expression variance quantitative trait loci (evQTLs) in human LCLs to determine whether genes with major evQTLs are also more likely to be MAE in LCLs[24]. We performed analyses using data based on chromatin signature in LCLs, extensively validated using allele-specific RNA-Seq in clonal LCLs[2]. However, we also wanted to assess any brain-specific effects using the classification described above (also see Methods and Table S1). Of the 166 genes with reported *cis* evQTLs, 49 were found in our dataset with MAE/BAE status in LCLs and 104

were found with MAE/BAE status in fetal brain. Compared with the expected proportion of 0.13 in the LCL dataset, we found 18 (0.36) evQTL targets to be MAE compared with 31 BAE ($P < 1.1 \times 10^{-6}$). Compared with the expected proportion of 0.39 in the fetal brain dataset, we found 73 (0.7) evQTL targets to be fetal brain MAE compared with 31 BAE ($P < 7.8 \times 10^{-11}$; Figure 3b). In order to rule out the possibility that eQTLs targeting MAE genes would appear to be evQTLs, we assessed eQTLs in this dataset[25]. A small number of evQTL target genes also showed a significant eQTL (2 MAE and 6 BAE), and our results remain unchanged excluding these eQTL targets.

The analysis above may be subject to confounding factors, specifically expression level. For example, highly expressed genes (like housekeeping genes) show low variation between individuals. The same genes tend also to be BAE. This will generate an association between variation in gene expression, evQTLs and MAE. To address this possibility, we performed a regression analysis with expression level in GM12878 as a nuisance variable. MAE/BAE status was shown to be the significant predictor of the presence of evQTL in the gene ($P < 1.6 \times 10^{-5}$) but the effect of expression level wasn't significant ($P < 0.6$).

Genes with MAE status are overrepresented in inter-species variance and directional expression level selection and underrepresented in expression-constrained genes

Standing variation under little constraint in a population can provide a substrate for directional selection; thus we hypothesized that if MAE genes have greater expression variance within a population, they might also show increased directional selection during evolution. To assess whether MAE gene expression variability might play a role in inter-species differences, we examined data from a recent publication describing human-chimpanzee differences in gene expression at the RNA and protein level in LCLs[26]. Supplementary data provided with this publication included 3,390 LCL expressed genes, of which 3,230 (72 MAE, 3,158 BAE) intersected with genes meeting our criteria of length > 2,500 and expression level above 50th percentile in LCLs. To compare the inter-species expression level change for MAE genes to BAE genes, we examined the proportion of MAE genes in high and low variance genes, as defined by the original publication[26]. Genes with MAE status were marginally overrepresented in high variance genes (RNA $P < 0.04$ two-sided; protein $P < 0.082$, one-sided). The weak significance values were largely due to the relatively small proportion of genes inferred to be MAE in LCLs in the overall dataset (2% of the protein and 1.3% of the RNA dataset). We found that these results were more robust for fetal brain predictions, where more MAE genes were assayed (14% of such genes, RNA $P < 2 \times 10^{-14}$; Protein $P < 7 \times 10^{-9}$), despite LCL expression data.

Next, we examined the distribution of models to which BAE and MAE fit best, combining human and chimpanzee lineage selection into three categories: 1) directional selection along human or chimpanzee lineage, 2) no mean expression difference, indicative of constraint, 3) relaxation of constraint along human or chimpanzee lineage[26]. We found an overall difference in the RNA model assignments of LCL and fetal brain MAE and BAE genes ($P < 0.05$). Specifically, MAE genes are overrepresented in directional selection for human and chimpanzee lineages and underrepresented in genes with no mean LCL expression difference (Figure 4). The same pattern held true for fetal brain prediction and LCL protein

expression (Table S4). Although intuitively, we might expect within-species variance to differ across these selection categories, we found that for BAE genes, average variance increased approximately 10% from relaxed (0.00050) to directional (0.00055) to equal-expression (0.00059). For MAE genes, average variance increased approximately 2-fold between these same categories (0.0013 to 0.0024 to 0.0048). Within each category, variance for MAE genes was higher than for BAE genes (each $P_{one\text{-}sided} < 0.05$). In the Khan *et al*[26] paper from which these data were derived, a conclusion reached was that variation was buffered for protein compared to RNA expression. In contrast, we found differences between BAE and MAE genes to be consistent between RNA and protein expression (above).

Genes with MAE status are underrepresented in pathogenic CNVs, but not in polymorphic CNVs

Given that MAE genes were found to vary in expression level more than BAE genes, and given that those differences seem to have functional significance in humans, we asked how MAE genes might intersect with genomic variation influencing gene expression. A previous study assessed distribution of 212 genes that were identified as MAE by allele-specific expression analysis in patient-derived clonal neuronal progenitor cells[53]. It reported overrepresentation of those MAE genes in known pathogenic CNVs from multiple datasets. We used our much larger gene set to assess enrichment of MAE genes in the same neuropsychiatric datasets[8, 28-35, 38], as well as several additional neuropsychiatric CNV datasets[36, 37, 54, 55]. The overall rate of fetal brain MAE is 0.39. In the set of control CNVs, the proportion is 0.37 ($P < 0.21$). Unlike the aforementioned study[27], we found that each of the NDD datasets showed *decreased* proportions of MAE genes (0.26-0.37 with DerSimonian-Laird estimator of heterogeneity tau^2= 0 across NDD datasets; $P < 2.2 \times 10^{-10}$ Mantel-Haenszel meta-analysis compared with fetal-brain expressed, and $P < 0.02$ compared with control CNVs from dbVAR, Figure 5a). We further validated the result with a higher expression level cutoff ($P < 0.003$, Table S5).

Genes with MAE status are depleted in predicted haploinsufficient gene sets

To further test our model that MAE variants are unlikely to act via haploinsufficiency, we utilized a recent study predicting haploinsufficient genes in the human genome[49]. We compared genes with probability of haploinsufficiency (HI) greater or equal to 0.9 to genes with HI less than or equal to 0.1 [strongly predicted to be haplosufficient (HS)] out of all genes scored. As our model predicts, genes likely to be HI show reduced proportion of fetal brain genes with MAE signature (0.28; $P < 3.6 \times 10^{-6}$), and further, genes likely to be HS show increased proportion of fetal brain genes with MAE signature (0.4; $P < 0.052$) (Figure 5b, 5c).

Disease risk nucleotides acting on gene expression (eQTLs) are dramatically depleted of MAE genes

In order to examine an independent set of neuropsychiatric risk genes not derived from CNVs, we analyzed data from schizophrenia genome-wide association meta-analysis[39]. Within the full list of 347 unique genes potentially implicated across 108 schizophrenia loci, we found only a modest and non-significant decrease in brain MAE genes (0.35) in the complete dataset. However, this study also provided a filtered subset, based on an overlap

between risk loci and eQTLs in both blood and brain, as these are more likely to be functionally significant variants (via expression level). In genes having a blood eQTL, inferred brain MAE genes were dramatically depleted (0.20, $P < 1.7 \times 10^{-4}$). In reported brain eQTLs, inferred brain MAE genes were found at only a small fraction (0.09, $P < 0.0056$) (Figure 5b).

However, based on the functional importance implied by directional selection in humans, we hypothesized that representation of MAE genes would vary based on the predicted role of the variant allele. Although both CNVs and eQTLs are likely to act via expression level changes, protein coding mutations can act via additional mechanisms, such as gain-of-function. There were a small number of common protein coding missense variants in the same set of potentially implicated genes in schizophrenia association regions (N=11), and brain MAE genes appear to be strongly overrepresented in these protein-altering candidates (0.75, $P < 0.042$, one-sided; Figure 5c). We therefore wanted to more specifically examine our prediction that genetic variation in MAE genes acting via loss-of-function and via gain-of-function would show opposite patterns.

Autism genes enriched for loss-of-function single nucleotide variants compared to autism genes enriched only for missense variants suggest a novel hypothesis

To examine rare single nucleotide variants (SNVs), we analyzed data from two recent autism spectrum disorder (ASD) exome sequencing studies[45, 47]. The first provides an analysis of both *de novo* and inherited rare exome variants in >3,800 case and >9,900 control subjects, identifying ASD-associated genes with FDR q-values < 0.3. ASD associated genes were suggestively depleted overall in fetal brain MAE genes (0.29, $P < 0.07$). From these probable ASD risk genes, we examined genes with *de novo* loss-of-function (LoF) mutations (equivalent to deletion CNV) in cases but not controls, and we found brain MAE genes similarly (but non-significantly) underrepresented (0.2, $P < 0.025$; Figure 5b). However, ASD-associated genes driven by missense variants (with no observed proband *de novo* LoF mutations) showed (non-significant) overrepresentation of genes with MAE chromatin signature in brain (0.43, $P < 0.48$; Figure 5c). The second study examined only *de novo* variants in >2,500 ASD simplex cases and sibling controls. When we restricted to genes with at least 1 *de novo* LoF mutation in cases and no LoF *de novo* mutations in controls, we similarly see an underrepresentation of MAE genes (0.33, $P < 0.069$; Figure 5b). However, when we examine genes with at least 2 *de novo* proband missense mutations with no *de novo* LoF proband mutations and no sibling missense mutations observed, we see overrepresentation of MAE genes (0.46, $P < 0.24$; Figure 5c). Although missense variants can act via either gain-of-function (GoF) or LoF mechanisms, we enriched the missense categories examined here for GoF by excluding genes showing any evidence of likely LoF (e.g. frameshift, nonsense) *de novo* variants in affected probands. Although none of these observations are statistically convincing on their own, together these data suggest a hypothesis that MAE genes may be more likely to act via GoF (i.e. protein function is important but reducing amount 0.5X does not lead to disease), while BAE genes may be more likely to act via LoF (inactivating one of two copies leads to disease).

Genes more constrained for missense than for LoF mutation are enriched for inferred MAE status

In order to test our hypothesis utilizing an independently-defined set of genes likely to act via GoF, we utilized a study providing constraint scores for genes across the genome[50]. Constraint was estimated separately for missense and for LoF mutations (http://atgu.mgh.harvard.edu/webtools/gene-lookup/lookupGene), and we infer that genes showing higher levels of constraint for missense mutations than for LoF mutations are most likely to act via GoF. We thus compared MAE ratio for the top 10% of genes showing higher constraint for missense mutations than for loss-of-function mutations compared with all genes scored. We find a higher proportion of brain MAE genes (0.45; $P < 3.7 \times 10^{-4}$) among this missense-constrained gene set. To further support the hypothesis that MAE genes are likely to be pathogenic via GoF, a search of OMIM using the term 'gain-of-function' and excluding the term 'loss-of-function' shows an increased proportion of fetal brain predicted MAE genes in the group annotated only with 'gain-of-function' (0.55, $P < 7.8 \times 10^{-5}$).

Genes with predicted MAE status are enriched among recessive autism genes

Another class of genes involved in disease, but likely to be haplosufficient, are those acting via a recessive model. Although recessive mutations are likely to act via LoF, the recessive model requires both alleles to be affected, with no fully functional allele remaining, so are unlikely to be highly dosage sensitive. We tested our prediction that genes likely to act recessively on autism would also be more likely to be in the MAE class utilizing recently published data[48]. Autosomal genes with either homozygous or compound heterozygous loss-of-function SNVs (nonsense or splice site) have increased likelihood to be classified as MAE in fetal brain (0.69, $P < 9.5 \times 10^{-4}$).

Meta-analyses

In order to confirm the overall pattern of eQTL/LoF/HI underrepresentation and GoF/HS/recessive overrepresentation of fetal brain MAE genes relevant to neurodevelopmental disorders, we performed a meta-analysis across the schizophrenia GWAS, autism exome, constraint and HI/HS prediction, and OMIM categories. Overall, fetal brain MAE genes are underrepresented in predicted pathogenic genes likely to affect expression level or be haploinsufficient ($P < 7.5 \times 10^{-11}$) (Figure 5b). In contrast, fetal brain MAE genes are overrepresented in predicted pathogenic genes likely to incur gain of function or be haplosufficient ($P < 1.4 \times 10^{-6}$) (Figure 5c).

DISCUSSION

In light of recent data indicating that somatic MAE affects a large proportion of expressed genes in multiple tissues, we asked how potential differences in expression level between monoallelic and biallelic cells affect variation in the population, and what the implications for evolution and disease would be. Our focus was on two tissue types: blood as the best studied MAE system in human, and brain as a system with particular sensitivity to small regulatory effects. We used a published and validated chromatin signature method to infer monoallelic expression in human fetal brain and found that the results of the method are in agreement with experimentally measured monoallelic expression in NPCs. The application

of this method to the neural lineage has already been validated in mouse [3]. Notably, for analyses where we were able to utilize both LCL and fetal brain data, results were consistent, suggesting that large-scale patterns for this class of genes are evident across tissues, much as has been found for genetic regulation of gene expression[21]. However, because of technical differences, direct comparisons cannot be made between MAE predictions in LCLs and fetal brain.

We first experimentally determined that the monoallelic state shows lower expression than the biallelic state for the same gene in human cells by performing a meticulous technical analysis of genome-wide RNA-seq data from two LCL clonal cell lines and comparing expression of the same gene across tissues with MAE compared with BAE status. Similar results have been reported previously for mouse NPCs[6]. One methodological difficulty in extending the analyses performed on mouse NPCs to human data is that in previous studies the biallelic state was defined merely as the absence of confirmed monoallelic state, establishing low coverage as a potentially significant source of noise. This is particularly concerning in human data where only few diagnostic loci are available to distinguish between the two alleles. We therefore improved on the existing methodology by utilizing separate statistical tests to positively predict both biallelic and monoallelic expression for our comparisons.

We then showed that MAE genes have higher variance than BAE genes in humans, by taking advantage of a large dataset encompassing transcriptome-wide measurement of multiple tissues derived from dozens of individuals. Given that MAE genes differ in expression level across clones, this variability is likely to be due, at least in part, to clonal composition differences during development. Interestingly, these observations suggest that expression variation is tolerated for MAE genes, from individual cells to the whole organism. However, the question remained whether this variation has functional consequences.

We hypothesized that if high levels of expression variance are tolerated, or even adaptive, variance might be maintained in the population through genetically-encoded regulation, such as evQTLs. Thus, we examined human targets of regulation in expression variance, and found that indeed genes with MAE status are also more frequently the target of genetic regulation of expression variance in humans. Further experimentation would be necessary to determine whether these genetic *cis* evQTLs act on expression variance via MAE or through an independent mechanism.

We further conjectured that variance in expression level of MAE genes could provide substrate for evolutionary selection if these genes are functionally significant. If so, MAE genes would be more likely to show cross-species differences in expression level and evidence for directional selection compared with BAE genes. We analyzed data on human-chimpanzee gene expression at the RNA and protein level. We found support for our hypothesis that MAE genes would show greater inter-species variance in expression than BAE genes. In addition, MAE genes more often fit a model of directional selection in either species and less often showed similar expression level in humans and chimpanzees than BAE genes. Together, these data suggest that MAE increases variation in gene expression level, which provides a substrate for evolutionary change. Genes undergoing MAE might be

particularly likely to differentiate humans from non-human primates, as evidenced from an underrepresentation of genes with similar expression levels. These data also support our model that MAE expression variance is not only tolerated, but also adaptive, and is consistent with our recent finding that at the nucleotide level, MAE genes show increased diversity and evidence for balancing selection compared with BAE genes[56].

Finally, we wanted to assess whether MAE status is predictive of pathogenic effects associated with expression level changes. Copy number variants, or deletions/duplications >1kb, have been shown to affect expression level of genes contained within them, and be major risk factors for neurodevelopmental and neuropsychiatric disease, such as autism and schizophrenia. Thus our model would predict that MAE genes can tolerate expression changes and would be non-pathogenic if contained in a CNV. Indeed, analysis of the large dataset of MAE genes we predicted in fetal brain provides strong evidence that these genes are underrepresented in pathogenic CNVs compared to polymorphic CNVs or all expressed genes. A previous study reached an opposite conclusion[27]. Notably, that study assessed a small fraction (<2%) of the number of MAE genes analyzed here. More importantly, the overall proportion of MAE genes in the CNV datasets was calculated without taking into account that only a fraction of the genes in those datasets were classified as MAE or BAE. When the same data were re-analyzed correcting for the number of genes positively identified as MAE or BAE, there was no significant over- or under-representation of MAE genes in that study among neurodevelopmental CNVs (Table S6).

Our CNV results are supported by analysis of eQTLs overlapping schizophrenia risk loci and loss-of-function exome variants in autism, within which genes with MAE chromatin signature are also dramatically depleted. Both CNVs and loss-of-function variants are likely to act via haploinsufficiency. In contrast, genes with schizophrenia missense polymorphisms and autism missense or recessive variants are overrepresented in MAE genes. Missense variants may act through gain-of-function mechanisms (functional properties not amount of protein lead to disease), hence although MAE genes can be functionally important (and pathogenic), this seems to be mediated either by specific protein sequence changes which may incur gain-of-function or by complete knock-out.

Thus, our observations might be generally useful in interpreting genetic variation in the context of human disease. A recent study of stochastic monoallelic expression suggests that it could contribute to variance around predicted genotype-phenotype relationships such as penetrance, expressivity, and discordant monozygotic twins[53]. We further propose that specific genotype-phenotype relationships might be more accurately predicted in the population with knowledge of their (predicted) MAE status. SNVs in genome or exome sequencing results without convincing data available in research or clinical databases can be difficult to interpret. Current strategies include prediction of how disruptive an amino acid substitution is likely to be (e.g. PolyPhen), evolutionary conservation or overall constraint as an indicator of functional importance, and tissue-specific expression. We propose that known tissue-specific MAE status could be used as an additional prediction tool, for example gain-of-function or recessive changes in genes of the MAE class might be more likely to be pathogenic, and heterozygous loss-of-function changes in MAE genes, less likely to be pathogenic.

In conclusion, we have used a variety of data to support the hypothesis that MAE is an important epigenetic regulatory mechanism influencing expression level variance both within the population and across species. This carries implications for a critical role for MAE in evolution, maintenance of variation in the population, and with future follow-up could prove valuable to prediction of pathogenicity in neurodevelopmental disorders.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge helpful discussion from Drs. Artem Artemov, Rahul Deo, Aditi Deshpande, Ophir Klein, Andrey Mironov, Ludmila Pawlikowska, Jonathan Pritchard, Erika Yeh, and Noah Zaitlen. We acknowledge funding from R21MH105745 (Weiss/Gimelbrant).

References

1. Savova V, Vigneau S, Gimelbrant AA. Autosomal monoallelic expression: genetics of epigenetic diversity? Current opinion in genetics & development. 2013; 23(6):642–648. [PubMed: 24075575]
2. Nag A, Savova V, Fung HL, Miron A, Yuan GC, Zhang K, et al. Chromatin signature of widespread monoallelic expression. Elife. 2013; 2:e01256. [PubMed: 24381246]
3. Nag A, Vigneau S, Savova V, Zwemer LM, Gimelbrant AA. Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types. G3. 2015; 5(8):1713–1720. [PubMed: 26092837]
4. Pereira JP, Girard R, Chaby R, Cumano A, Vieira P. Monoallelic expression of the murine gene encoding Toll-like receptor 4. Nat Immunol. 2003; 4(5):464–470. [PubMed: 12665857]
5. Wu H, Luo J, Yu H, Rattner A, Mo A, Wang Y, et al. Cellular resolution maps of × chromosome inactivation: implications for neural development, function, and disease. Neuron. 2014; 81(1):103–119. [PubMed: 24411735]
6. Gendrel AV, Attia M, Chen CJ, Diabangouaya P, Servant N, Barillot E, et al. Developmental dynamics and disease potential of random monoallelic gene expression. Developmental cell. 2014; 28(4):366–380. [PubMed: 24576422]
7. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science (New York, NY. 2007; 318(5853):1136–1140.
8. Eckersley-Maslin MA, Thybert D, Bergmann JH, Marioni JC, Flicek P, Spector DL. Random monoallelic gene expression increases upon embryonic stem cell differentiation. Developmental cell. 2014; 28(4):351–365. [PubMed: 24576421]
9. Kindt AS, Navarro P, Semple CA, Haley CS. The genomic signature of trait-associated variants. BMC genomics. 2013; 14:108. [PubMed: 23418889]
10. Stranger BE, Raj T. Genetics of human gene expression. Current opinion in genetics & development. 2013; 23(6):627–634. [PubMed: 24238872]
11. Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, et al. Cis and trans effects of human genomic variants on gene expression. PLoS genetics. 2014; 10(7):e1004461. [PubMed: 25010687]
12. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Human molecular genetics. 2009; 18(R1):R1–8. [PubMed: 19297395]
13. Nord AS, Roeb W, Dickel DE, Walsh T, Kusenda M, O'Connor KL, et al. Reduced transcript expression of genes affected by inherited and de novo CNVs in autism. European journal of human genetics : EJHG. 2011; 19(6):727–731. [PubMed: 21448237]
14. Gamazon ER, Nicolae DL, Cox NJ. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS genetics. 2011; 7(2):e1001292. [PubMed: 21304891]

15. Prendergast JG, Chambers EV, Semple CA. Sequence-level mechanisms of human epigenome evolution. Genome biology and evolution. 2014; 6(7):1758–1771. [PubMed: 24966180]
16. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, et al. Dynamics of DNA methylation in recent human and great ape evolution. PLoS genetics. 2013; 9(9):e1003763. [PubMed: 24039605]
17. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. Nature. 2014; 508(7495):199–206. [PubMed: 24695229]
18. Savova V, Patsenker J, Vigneau S, Gimelbrant AA. dbMAE: the database of autosomal monoallelic expression. Nucleic acids research. 2016; 44(D1):D753–756. [PubMed: 26503248]
19. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28(10):1045–1048. [PubMed: 20944595]
20. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011; 12:323. [PubMed: 21816040]
21. Consortium GTEx. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science (New York, NY. 2015; 348(6235):648–660.
22. Paulson, J., Chen, C-Y., Lopes-Ramos, CM., Kuijjer, ML., Platig, J., Sonawane, AR., et al. Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. 2016. bioRxiv
23. Hicks, SC., Okrah, K., Paulson, JN., Quackenbush, J., Irizarry, RA., Corrada Bravo, H. Smooth Quantile Normalization. 2016. bioRxiv
24. Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. Genetics. 2013; 193(1):95–108. [PubMed: 23150607]
25. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nature genetics. 2007; 39(10):1217–1224. [PubMed: 17873874]
26. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science (New York, NY. 2013; 342(6162):1100–1104.
27. Jeffries AR, Collier DA, Vassos E, Curran S, Ogilvie CM, Price J. Random or stochastic monoallelic expressed genes are enriched for neurodevelopmental disorder candidate genes. PloS one. 2013; 8(12):e85093. [PubMed: 24386451]
28. Mowry BJ, Gratten J. The emerging spectrum of allelic variation in schizophrenia: current evidence and strategies for the identification and functional characterization of common and rare variants. Molecular psychiatry. 2013; 18(1):38–52. [PubMed: 22547114]
29. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008; 455(7210):232–236. [PubMed: 18668039]
30. Murdoch JD, State MW. Recent developments in the genetics of autism spectrum disorders. Current opinion in genetics & development. 2013; 23(3):310–315. [PubMed: 23537858]
31. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nature genetics. 2011; 43(9):838–846. [PubMed: 21841781]
32. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464(7289):704–712. [PubMed: 19812545]
33. de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, Tsang P, et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Human molecular genetics. 2007; 16(23):2783–2794. [PubMed: 17666407]
34. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nature genetics. 2010; 42(5):400–405. [PubMed: 20364138]
35. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, et al. The fine-scale and complex architecture of human copy-number variation. American journal of human genetics. 2008; 82(3):685–695. [PubMed: 18304495]

36. Rosenfeld JA, Coe BP, Eichler EE, Cuckle H, Shaffer LG. Estimates of penetrance for recurrent pathogenic copy-number variations. Genetics in medicine : official journal of the American College of Medical Genetics. 2013; 15(6):478–481. [PubMed: 23258348]
37. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014; 505(7483):361–366. [PubMed: 24352232]
38. Leitersdorf E, Chakravarti A, Hobbs HH. Polymorphic DNA haplotypes at the LDL receptor locus. American journal of human genetics. 1989; 44(3):409–421. [PubMed: 2563635]
39. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421–427. [PubMed: 25056061]
40. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nature genetics. 2014; 46(5):430–437. [PubMed: 24728292]
41. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, et al. A survey of genetic human cortical gene expression. Nature genetics. 2007; 39(12):1494–1499. [PubMed: 17982457]
42. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics. 2010; 6(5):e1000952. [PubMed: 20485568]
43. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature. 2011; 478(7370):519–523. [PubMed: 22031444]
44. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, et al. Genetic control of human brain transcript expression in Alzheimer disease. American journal of human genetics. 2009; 84(4):445–458. [PubMed: 19361613]
45. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014; 515(7526):209–215. [PubMed: 25363760]
46. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4):248–249. [PubMed: 20354512]
47. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014; 515(7526):216–221. [PubMed: 25363768]
48. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron. 2013; 77(2):235–242. [PubMed: 23352160]
49. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS genetics. 2010; 6(10):e1001154. [PubMed: 20976243]
50. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nature genetics. 2014; 46(9):944–950. [PubMed: 25086666]
51. Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, et al. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron. 2009; 62(4):494–509. [PubMed: 19477152]
52. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. [PubMed: 22955616]
53. Jeffries AR, Perfect LW, Ledderose J, Schalkwyk LC, Bray NJ, Mill J, et al. Stochastic choice of allelic expression in human neural stem cells. Stem cells. 2012; 30(9):1938–1947. [PubMed: 22714879]
54. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nature genetics. 2014; 46(10):1063–1071. [PubMed: 25217958]
55. Olson H, Shen Y, Avallone J, Sheidley BR, Pinsky R, Bergin AM, et al. Copy number variation plays an important role in clinical epilepsy. Annals of neurology. 2014; 75(6):943–958. [PubMed: 24811917]

56. Savova V, Chun S, Sohail M, McCole RB, Witwicki R, Gai L, et al. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. Nature genetics. 2016; 48(3):231–237. [PubMed: 26808112]

57. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic acids research. 2015; 43(Database issue):D250–256. [PubMed: 25428351]

58. Zwemer LM, Zak A, Thompson BR, Kirby A, Daly MJ, Chess A, et al. Autosomal monoallelic expression in the mouse. Genome biology. 2012; 13(2):R10. [PubMed: 22348269]

59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. [PubMed: 25516281]

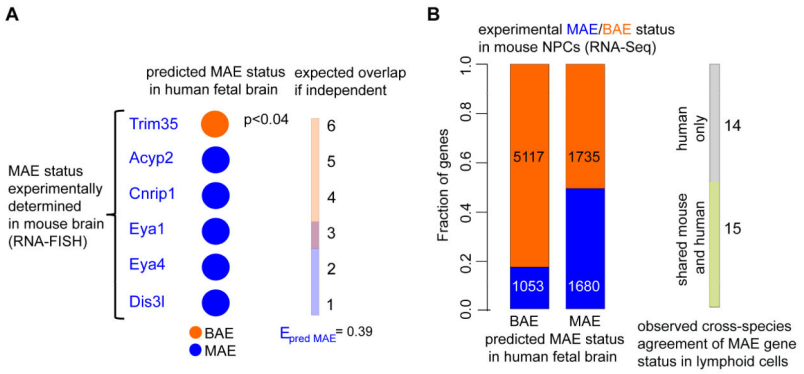

Figure 1. Experimental confirmation of fetal brain MAE status in mouse neural cell lines and tissue

(A) Six mouse orthologues confirmed MAE by RNA-FISH in mouse brain [6]. Each circle represents a gene, colored by status (MAE=blue, BAE=orange). Expected fraction of MAE under the null hypothesis based on our prediction dataset (0.39) is indicated in the vertical bar graph. Orthologous genes were defined using OrthoDB database [57].

(B) MAE gene status in experimental RNA-seq data on mouse neuronal progenitor clonal cell lines [6]. MAE status was determined as previously described [2]. Fraction of measured MAE genes (blue) and BAE genes (orange) among predicted BAE (left) and MAE (right) orthologues. Expected fraction of MAE under the alternative hypothesis of human-mouse differences but not predicted-experimental or NPC-brain differences contributing based on empirical data comparing mouse and human experiments from the same tissue [58] is shown in the vertical bar graph.

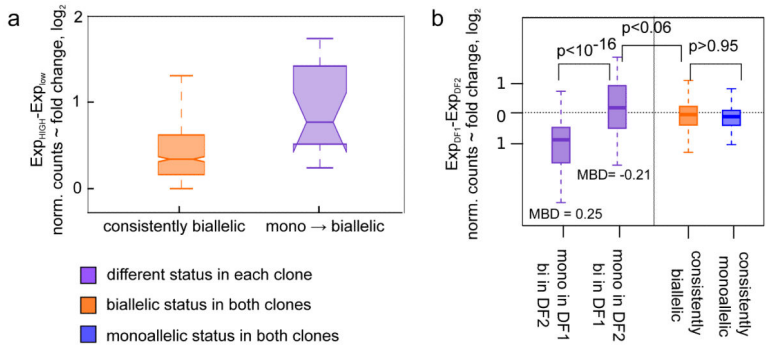

Figure 2. Expression differences in MAE genes between two clones

Expression differences are shown according to monoallelic state in each clone as measured in DESeq-"normalized counts" and approximately equivalent to $\log_2$ fold-change [59].

(A) Absolute difference in expression levels for "consistently biallelic" genes – genes with biallelic status in both clones, and for "mono → biallelic" genes – genes with monoallelic status in one clone and biallelic status in the other clone.

(B) Difference in expression levels between clone DF1 and clone DF2 for four groups of genes defined by allelic expression: "consistently biallelic", as in A; "consistently monoallelic" – genes with monoallelic status in both clones; "mono in DF1, bi in DF2" – genes with monoallelic status in DF1 and biallelic status in DF2; "mono in DF2, bi in DF1" – genes with monoallelic status in DF2 and biallelic status in DF1. Mean bias difference (MBD) between DF1 and DF2 is shown for each of the groups below.

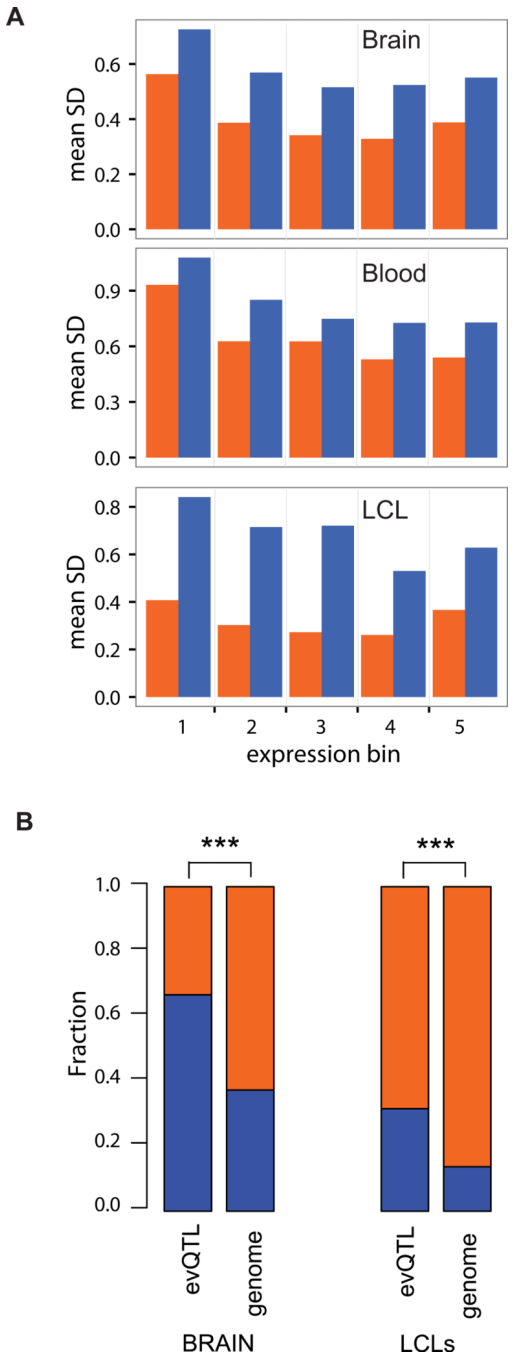

Figure 3. Genes with MAE status show elevated expression level variation and are overrepresented among genes targeted by evQTL variants across tissues

(A) LCL: lymphoblastic cell lines; Blood: whole peripheral blood samples; Brain: brain cortex samples [48]. MAE: genes predicted to be MAE in the tissue by chromatin signature. BAE: genes predicted to be BAE in the tissue by chromatin signature. Shown are standard deviations for MAE and BAE classified genes. See also Table S2.

(B) Left: proportion of MAE and BAE genes as predicted by chromatin signature in fetal brain genome-wide (genome), and among genes targeted by evQTLs (dataset from Hulse and Cai)[24]; right: same for predicted MAE and BAE genes in LCLs.

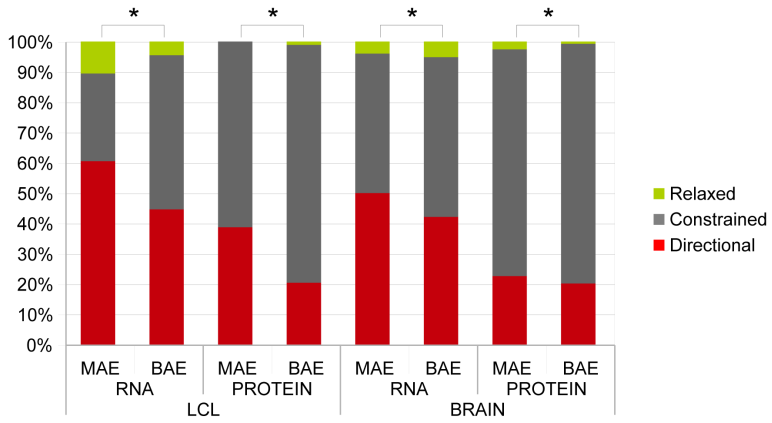

Figure 4. Genes with inferred MAE status are more often subject to directional selection on expression level between human and chimpanzee

Proportion of genes identified as subject to conservational constraint (*grey*), directional (*red*) and relaxed (*green*) selection in RNA and protein expression levels in blood [26]. LCL: genes predicted MAE by chromatin signature in LCLs. BRAIN: genes predicted MAE by chromatin signature in fetal brain.

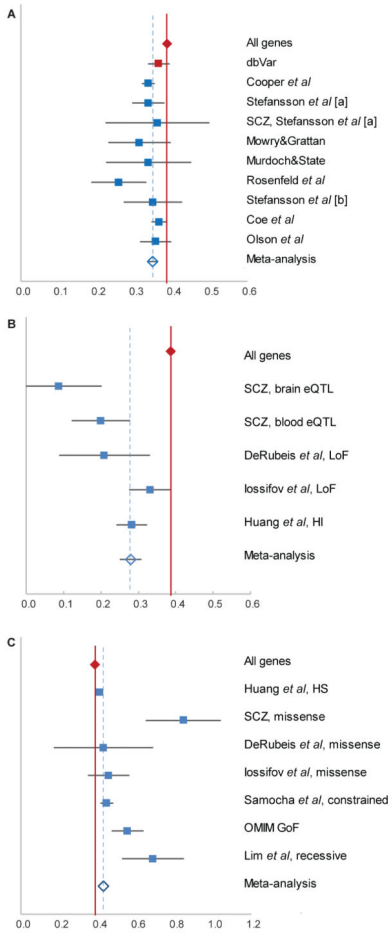

Figure 5. Mechanism of gene pathogenicity is associated with MAE status

(A) Proportion of MAE genes in CNV datasets representing controls (dbVar, red square) and pathogenic neurodevelopmental CNVs (individual datasets: blue squares, meta-analysis: blue diamond) compared with the genomewide proportion (red diamond). Proportion is calculated as percent of unique genes contained in a CNV with MAE status based on all genes with either status assigned. SCZ, schizophrenia: CNVs associated with schizophrenia. 95% confidence intervals for each estimate are shown with horizontal bars. The red vertical line represents the proportion of all genes, and the dotted blue line represents the estimated proportion in neurodevelopmental CNVs estimated by meta-analysis. Studies analyzed: Cooper et al [31], Stefansson et al [a] [37], Stefansson et al [b] [29], Mowry and Grattan [28], Murdoch and State [30], Rosenfeld et al [36], Coe et al [54], Olson et al [55].

(B) Proportion of MAE genes in datasets representing haploinsufficient, expression-altering, and loss-of-function pathogenic genes (individual datasets: blue squares, meta-analysis: blue diamond) compared with the genomewide proportion (red diamond). Proportion is calculated as percent of unique genes in a dataset with MAE status based on all genes with either status predicted. HI, haploinsufficient: genes with haploinsufficiency scores at or above 0.9. SCZ, schizophrenia PGC meta-analysis: annotated genes in credible regions of SNP association [36]. eQTL, expression quantitative trait locus: defined as in the study [36]. LoF, loss-of-function: genes with *de novo* frameshift, nonsense, splice mutations in at least

one autism case and no controls. 95% confidence intervals for each estimate are shown with horizontal bars. The red vertical line represents the proportion of all genes, and the dotted blue line represents the estimated proportion in pathogenic genes estimated by meta-analysis. Studies analyzed: brain and blood eQTLs [39]; DeRubeis et al [45], Iossifov et al [47], Huang et al [49].

(C) Proportion of MAE genes in datasets representing recessive, missense, and gain-of-function pathogenic genes and haplosufficient genes (individual datasets: blue squares, meta-analysis: blue diamond) compared with the genomewide proportion (red diamond). Proportion is calculated as percent of unique genes in a dataset with MAE status based on all genes with either MAE or BAE status predicted. HS, haplosufficient: genes with haploinsufficiency scores at or below 0.1. Missense: genes with no *de novo* LoF mutations in autism cases and either $q<0.3$ or at least two *de novo* missense mutations in cases and no missense mutations in controls. Constrained: genes in the top 10% showing higher constraint for missense variants than loss-of-function variants. 95% confidence intervals for each estimate are shown with horizontal bars. The red vertical line represents the proportion of all genes, and the dotted blue line represents the estimated proportion in pathogenic genes estimated by meta-analysis. Studies analyzed: same as in panel B, and also Samocha et al [50], Lim et al [48].