Frequent monoallelic or skewed expression for developmental genes in CNS-derived cells and evidence for balancing selection

Sergio Branciamore[a,1], Zuzana Valo[a], Min Li[b], Jinhui Wang[c], Arthur D. Riggs[a,1], and Judith Singer-Sam[a,1]

[a]Diabetes and Metabolism Research Institute, City of Hope Comprehensive Cancer Center, Duarte, CA 91010; [b]Division of Biostatistics, City of Hope Comprehensive Cancer Center, Duarte, CA 91010; and [c]Integrative Genomics Core, City of Hope Comprehensive Cancer Center, Duarte, CA 91010

Contributed by Arthur D. Riggs, August 31, 2018 (sent for review May 21, 2018; reviewed by Andrew Chess and Soojin V. Yi)

Cellular mosaicism due to monoallelic autosomal expression (MAE), with cell selection during development, is becoming increasingly recognized as prevalent in mammals, leading to interest in understanding its extent and mechanism(s). We report here use of clonal cell lines derived from the CNS of adult female F_1 hybrid (C57BL/6 X JF1) mice to characterize MAE as neural stem cells (*nscs*) differentiate to astrocyte-like cells (*asls*). We found that different subsets of genes show MAE in the two populations of cells; in each case, there is strong enrichment for genes specific to the respective developmental state. Genes that exhibit MAE are 22% of *nsc*-specific genes and 26% of *asl*-specific genes. Moreover, the promoters of genes with MAE have reduced CpG dinucleotides but increased CpG differences between the two parental mouse strains. Extending the study of variability to wild populations of mice, we found evidence for balancing selection as a contributing force in evolution of those genes showing developmental specificity (i.e., expressed in either *nsc* or *asl*), not just for genes showing MAE. Furthermore, we found that genes showing skewed allelic expression (SKE) were similarly enriched among cell type-specific genes and also showed a heightened probability of balancing selection. Thus, developmental stage-specific genes and genes with MAE or SKE seem to make up overlapping classes subject to selection for increased diversity. The implications of these results for development and evolution are discussed in the context of a model with stochastic epigenetic modifications taking place only during a relatively brief developmental window.

evolution | Gstm5 | overdominance | differentiation | schizophrenia

Somatic cellular mosaicism is becoming increasingly recognized as a significant aspect of mammalian development, phenotypic variation, and disease (1–5). Epigenetic as well as genetic mosaicism is prevalent in the CNS, where it has implications for both normal function and neuropsychiatric disorders, especially since there can be cell selection during development (1). X chromosome inactivation (XCI) is a well-known example of an epigenetic process resulting in monoallelic expression, with resulting cellular mosaicism due to random selection and inactivation early in differentiation of most genes on one X chromosome (2). In any given cell, choice of either the maternal or paternal X chromosome is random, but after the decision is made, it is mitotically inherited with great stability. More recently, evidence has been accumulating that many autosomal genes also show monoallelic expression and that genes with monoallelic autosomal expression (MAE) are enriched for those that affect membrane proteins and development (3, 5). There are about 100 imprinted autosomal genes with expression that is determined by parental source (that is, their expression depends on whether they are paternally or maternally derived) (6). Even more genes show random MAE, where choice of allele is apparently stochastically determined (2, 4, 7, 8). Since mitotically persistent random MAE and skewed expression are likely to impact development as well as genetic inheritance and evolution, it is important to understand their extent and mechanism(s). Despite this, random MAE's mechanism and evolutionary impetus remain largely unexplored. There have been a number of recent studies on MAE in humans and mice (reviewed in refs. 2 and 3), but significant puzzles remain. How is MAE affected by developmental changes? How does MAE affect development? How are alleles showing MAE regulated differently, although in the same nucleus? Are there specific sequences or patterns correlated with MAE? Does MAE have a special role in evolution? Even less is known about skewed allelic expression (SKE): selective mechanisms acting on MAE genes also are likely to be operating on SKE genes.

For several mouse genes, it is now known that stochastic epigenetic variability, sometimes influenced by the environment, can result in more than one phenotype from the same genotype (9–11). The best-studied example is mice carrying the agouti viable yellow gene: they are sometimes normal with agouti coat color, and they are sometimes yellow and obese. It is known that the insertion of a retroposon potentially provides an alternate promoter and that DNA methylation of this promoter correlates with phenotype (12). Methylation in the very early embryo is incomplete and stochastic, not always silencing the promoter. At or soon after implantation, the methylation state of the

Significance

While most mammalian genes are expressed from both chromosomal copies, many autosomal genes randomly express only one allele in a given cell, resulting in somatic cellular mosaicism. To better understand the mechanisms, developmental aspects, and evolution of autosomal monoallelic expression (MAE), we used nucleotide polymorphism differences between hybrid mice to analyze MAE of clonal neural stem cell lines as they differentiated to astrocytes. We found that genes showing MAE are highly enriched among developmental stage-specific genes. Genes showing strong skewed expression are similarly enriched. We also found evidence suggestive of balancing selection not just for genes with MAE but also, for developmental stage-specific genes.

Author contributions: S.B., A.D.R., and J.S.-S. designed research; S.B., Z.V., J.W., and J.S.-S. performed research; S.B., M.L., A.D.R., and J.S.-S. analyzed data; and S.B., A.D.R., and J.S.-S. wrote the paper.

Reviewers: A.C., Mount Sinai School of Medicine; and S.V.Y., Georgia Institute of Technology.

The authors declare no conflict of interest.

Data deposition: The genomic JF1 SNPs reported in this paper have been deposited in the European Nucleotide Archive, https://www.ebi.ac.uk/ena (accession no. ERZ772629).

[1]To whom correspondence may be addressed. Email: ariggs@coh.org, SBranciamore@coh.org, or jsam@coh.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1808652115/-/DCSupplemental.

Published online October 15, 2018.

retroelement is locked. The key point here is that the retroelement introduces epigenetic variability dependent on stochastic DNA methylation during a developmental window. Theoretical modeling, based on somatically heritable stochastic epigenetic modifications taking place only during specific developmental windows of opportunity (10, 11), led to the surprising prediction (10) that genes associated with stochastic epigenetic modification can show overdominance (i.e., higher relative fitness of heterozygotes relative to either homozygote). Overdominance is associated with balancing selection (i.e., the selective process by which genetic polymorphisms are maintained in the gene pool). Loci associated with balancing selection have been recently reported for humans (13, 14) and *Drosophila* (15). Classical examples include sickle cell anemia, HLA (16), and olfactory receptor loci (17). It recently has been reported that some human genes with MAE, identified by a chromatin modification signature, show excessive variability and selection for maintenance of heterozygosity (balancing selection) (18). Do mouse genes exhibiting MAE similarly show increased variability and balancing selection? We report here evidence in support of this possibility for genes with MAE and for cell type-specific genes. We also report on changes in genes exhibiting MAE during development. Most of the genes showing MAE or skewed expression in astrocyte-like cells (*asl*s) are in membrane-associated or cell-to-cell signaling categories. We have found increased sequence diversity in genes with MAE or SKE and in cell type-specific/developmental genes. This diversity may not only affect development but also, may provide some evolutionary advantage, perhaps related to the establishment of individual cell identities.

Results

Experimental Overview. We performed transcriptome-wide profiling of four previously described F_1(B6 X JF1) clonal cell lines (2A1, 2A5, 3A1, and 4A5) as they underwent in vitro differentiation along the astrocytic pathway (4). Cells were sampled at day 0 (before the start of differentiation) and at 1–4 d thereafter. A heat map describing changes in expression for each cell line is shown in Fig. 1*A*. Remarkably, most changes are visible by day 1 of differentiation, becoming more pronounced by days 3 and 4. Principal component analysis verified that about 85% of developmental changes occur by the first day after induction of differentiation (Fig. 1*B*). The astrocytic cells show considerable overlap in gene expression with astrocytes from adult mice (19) (Dataset S1).

Establishment of a JF1 genomic SNP library allowed us to analyze transcriptome-wide allele-specific expression. As previously described, a cutoff of fragments per killobase per million mapped reads (fpkm) ≥ 3 and a pooled SNP depth of coverage $\geq$10 allowed us to analyze ~10,000 genes in the four cell lines (8).

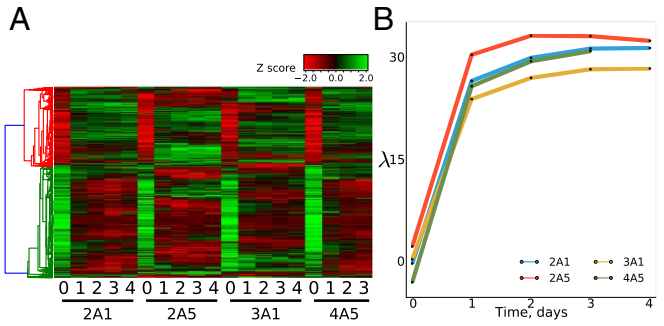

Fig. 1. (*A*) Heat map showing expression changes during in vitro differentiation of *nsc*s to astrocytes. Results are shown for each cell line and day of differentiation. (*B*) Principal component analysis. In this constructed space, the λ_1 component mainly associated with differentiation is shown to rise to near maximal levels by day 1 after induction of differentiation. Dataset S1 has a complete list of genes used in our study and their expression levels.

Definitions for Computational Analysis. We define neural stem cells (*nsc*s) and the *asl*s derived from them. *nsc* refers to the cells before induction of differentiation (day 0), whereas *asl* refers to days 1–4 postinduction. The probability of each gene being expressed from the C57BL/6 (B6) allele was computed as the simple ratio between the SNP counts assigned to the B6 haplotype and the total number of SNP counts observed at the locus (details are in ref. 8). To measure the bias in expression for one allele with respect to the other, we define the monoallelic skew variable (σ) for each gene as

$$\sigma_i = |P_i^{B6} - 0.5|, \quad [1]$$

where σ is bound between $[0, 0.5]$ and is close to 0 for genes with biallelic expression, whereas it approaches 0.5 for genes with MAE.

It is important to note that for a gene to be considered within the MAE set, we use strict criteria; it must have a large and statistically significant value of $\sigma \geq 0.35$ and be both monoallelically expressed at *fpkm* ≥ 3 and biallelically expressed at *fpkm* ≥ 3 in at least one cell line at one stage. These criteria minimize contributions from statistical noise due to low expression and from apparent MAE due to erroneous SNP calls; they also provide assurance that both alleles have the potential to be expressed at similar levels. X-linked genes provide a control (8) but are excluded from analysis of MAE. We then define MAE^{nsc} and MAE^{asl} as sets that contain genes displaying monoallelic expression in nsc or asl, respectively. Analogously, we define BAE^{nsc} and BAE^{asl} for the biallelically expressed genes.

***nsc* and *asl*—Specific Genes Preferentially Show MAE.** Overall, we found 712 genes (7.3%) showing MAE and 9,078 genes showing biallelic autosomal expression (BAE) (Fig. 2). Considering the two developmental stages independently, we observed the following: nsc: MAE^{nsc} is equal to 389 (4.6%), and BAE^{nsc} is equal to 8,160 (95.4%); asl: MAE^{asl} is equal to 587 (6.4%), and BAE^{asl} is equal to 8,629 (93.6%) (Fig. 2 *A*, *b* and *c* and *B*, *b* and *c*). It has been reported for both mice and humans (2) that MAE is enriched in cell type-specific genes. To investigate this in our system, we divided the MAE and BAE sets into three subsets: those that belong exclusively to nsc or asl and those that belong to both developmental stages (Fig. 2 *A*, *d* and *e* and *B*, *d* and *e*; *Materials and Methods* has a formal definition of the set, and Dataset S2 has gene lists). We then investigated if the MAE set is enriched in cell type-specific genes [i.e., genes expressed in only one developmental stage ($MAE^{nsc'}$ *vs.* $BAE^{nsc'}$ and $MAE^{asl'}$ *vs.* $BAE^{asl'}$) compared with genes that are in common ($MAE^{\cap}$ *vs.* $BAE^{\cap}$)]. The results in Fig. 2 *A*, *f*–*h* clearly show that the fraction of genes with MAE changes in different subsets: from 3.3% in the case of genes that are in common in the two developmental stages to 21.8% and 26.0% for genes present only in nsc and asl, respectively. These highly statistically significant observations (P value $<< 10^{-5}$) clearly show that the probability of MAE is strongly increased and can be a high percentage in genes that are specific to a particular cell type.

Differences in Allele-Specific Expression Between *nsc* and *asl*. Gene ontology (GO) analysis (20) was performed to explore the differences in the types of genes expressed among nsc and asl. Genes showing MAE specifically associated with nsc or asl or in common in both cell line are listed in Dataset S2. Genes enriched in nsc include mostly those involved in cell division

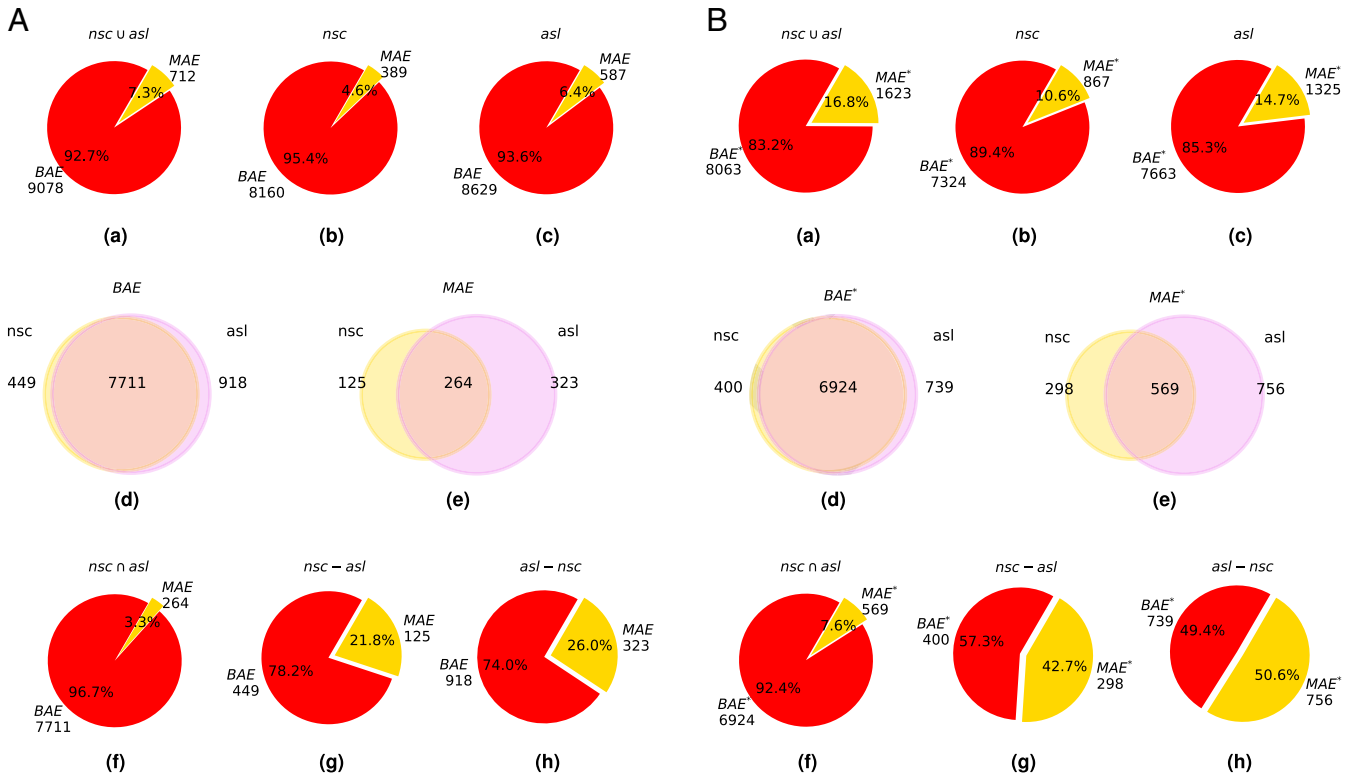

Fig. 2. Distribution of autosomal genes showing monoallelic or biallelic expression (MAE or BAE) in undifferentiated *nsc* and *asl*. In *A*, *MAE* is defined as $\sigma \geq 0.35$; in *B*, *MAE** is defined as $\sigma \geq 0.25$. (*a*–*c*) Counts and percentages of *MAE* (*MAE**) and *BAE* (*BAE**) in the union set of *nsc*, *asl* (*nsc* ∪ *asl*), *nsc*, and *asl*, respectively. (*d* and *e*) The distribution of *BAE* (*BAE**) and *MAE* (*MAE**), respectively, in *nsc* and *asl*. (*f*) Counts and percentages of *MAE* (*MAE**) and *BAE* (*BAE**) in the intersection set of *nsc*, *asl* (*nsc* ∩ *asl*; i.e., genes that are expressed in both *nsc* and *asl*). (*g* and *h*) Counts and percentages of *MAE* (*MAE**) and *BAE* (*BAE**) for genes that are expressed only in *nsc* or *asl*, respectively.

and DNA replication, while genes in *asl* cells are enriched for those involved in differentiated activity of neural cells, such as ion channels, transporters, and cell surface components (*SI Appendix*). Ontology analysis confirmed the enrichment of genes involved in cell adhesion and cell–cell signaling as well as extracellular matrix organization and calcium ion binding.

Our results confirmed previous findings of monoallelic expression of members of both the annexin and glutathione transferase (*Gst*) gene families *nsc* and *asl* (8). For the latter gene family, we were able to confirm monoallelic expression of *Gstm5* in individual hippocampal cells of postnatal mice (Fig. 3). In addition, a number of known imprinted genes show the expected pattern of reciprocal expression in *nsc* lines 2A1, 3A1, and 4A5 (paternal JF1 allele) vs. 2A5 (paternal B6 allele): *Commd1*, *Impact*, *Igf2*, *Mest*, *Ndn*, *Peg3*, *Peg10*, *Peg12*, *Plagl1*, *Sgce*, *Snrpn*, and *Zrsr1*. Examining the remaining genes with MAE, we noted a small category that increases somewhat in expression as they transition from BAE in *nsc* to MAE in astrocytes, including *Aldoc* and pituitary tumor-transforming 1 (*Pttg1*). *Aldoc* codes for a brain-specific glycolytic enzyme (21). *Pttg1* is involved in mitosis and has tumorigenic properties (22).

Another interesting category includes genes that show a relatively low level of monoallelic expression in *nsc* (or no expression at all) but are induced to high levels of MAE in astrocytes (Dataset S2). Some of these genes contribute to inflammation, such as oncostatin M (*Osmr*), a required gene for glioblastoma development (23), and chitinase-3-like protein 1 (*Chi3l1*; also termed *Chil1*) (24). *Chi3l1* is of particular interest, as it has been implicated as part of the known link between the immune system and schizophrenia (reviewed in ref. 25). Others are extracellular or membrane components, a group enriched among astrocytic genes as described above: these include *Cpxm1* (a Zinc carboxypeptidase) (26), integrin alpha chain 7 (*Itga7*) (27), and transmembrane 4 superfamily member 1 (*Tm4sf1*), a member of the tetraspanin family (www.ncbi.nlm.nih.gov/gene/17112). MAE of *Aldoc*, *Chi3l1*, *Cpxm1*, *Itga7*, *Osmr*, and *Tm4sf1* was independently confirmed by qRT-PCR (Dataset S3).

Dynamic Changes. We next examined the dynamics of changes in allele-specific expression for each gene during differentiation. The transition of each gene i is considered from a given *nsc* cell line to the same cell line j after differentiation to *asl* at day 3. Comparing day 0 with day 3, it can be seen (Fig. 4) that approximately one-half (0.46) of genes that display MAE in *nsc* continue to do so after differentiation to *asl*. The frequency of transition from BAE to MAE is 0.001 and to loss of expression is 0.17. The cases of BAE in *asl* cells are unchanged in most cases (0.83) from the *nsc* state, while a small percentage transitions from lack of expression to MAE (0.01). Although the percentage is small, it should be emphasized that 44% of MAE in *asl* is from previously silent genes (*SI Appendix*), and a high percentage (26%) of *asl*-specific genes shows MAE when $\sigma \geq 0.35$ (at least 85% expression from one allele) (Fig. 2 *A*, *g* and *h* and *B*, *g* and *h*). If σ minimum value is reduced to 0.25 (at least 75% from one allele), then even more *asl*-specific genes show MAE (50.6%) (Fig. 2 *B*, *g* and *h*). It is important to note that the differences between cell lines have persisted through numerous mitoses, consistent with other studies and with the hypothesis that stochastically variable epigenetic states, either preexisting in the *nsc* or established during differentiation, can then become fixed and mitotically persistent.

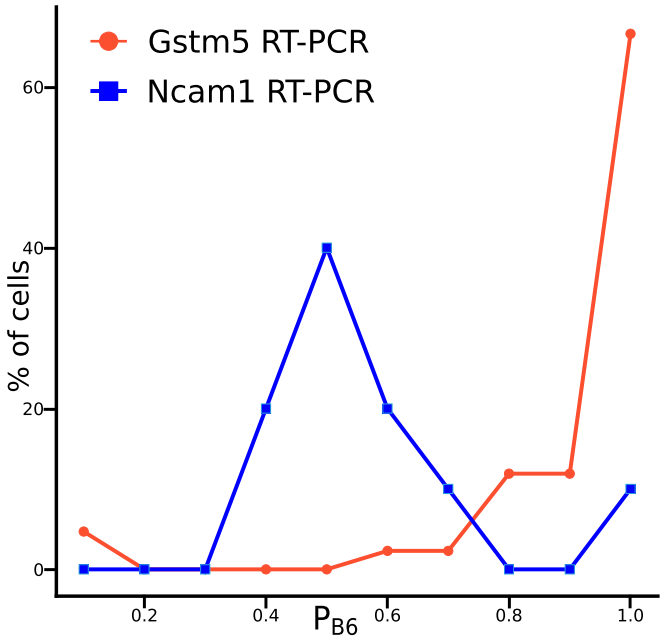

Fig. 3. Allele-specific expression of *Gstm5* and *Ncam1* in individual hippocampal cells of postnatal day 7 B6 X JF1 hybrid mice. The *x* axis shows the fractional expression of the B6 allele. *Gstm5* assay, *n* = 42 cells; *Ncam1* assay, *n* = 25 cells.

Genes with MAE and Genes Expressed in Either *nsc* or *asl* Show Decreased CpG Frequency and Increased CpG Variability. To explore possible sequence differences between the genes showing monoallelic vs. biallelic expression, we defined various regions (Fig. 5). For example, we defined the region from 1 kb upstream to 0.5 kb downstream of the transcription starting site as 5′ proximal. Given the well-known role of CpG dinucleotides in the epigenetic control of gene expression by DNA methylation, we also checked the distribution of $CpG_{o/e}$, where *o*/*e* equals observed/expected. Similarly, we considered all other possible dinucleotides ($NpM_{o/e}$).

We first found that there is a large, statistically significant difference between MAE and BAE in the distribution of dinucleotides in the vicinity of the 5′-proximal region (Fig. 6*A*). We then found a similar difference for genes that are cell type specific (either *asl* or *nsc*) and those that are expressed by both cell types (Fig. 6*B*). In particular, $CpG_{o/e}$ shows the largest observed difference, with cell type-specific genes characterized by a lower $CpG_{o/e}$ than those expressed in common. Other dinucleotides show less pronounced but significant differences (Fig. 6)

We also checked the differences in dinucleotide variability (θ) defined as the number of dinucleotide changes per dinucleotide count between the two parental mouse strains, B6 and JF1. We then compared the distribution of θ in MAE vs. BAE genes and cell type-specific genes vs. those expressed in common. Overall variability was found to be significantly greater for genes that display MAE (Fig. 7), predominately at CpG dinucleotides, but again, the difference was found to be due mainly to differences between cell type-specific genes and those expressed in common (Fig. 7*B*). We conclude that cell type-specific genes (and genes with MAE) show lower $CpG_{o/e}$ at the promoter and more variability at CpGs between B6 and JF1. Taken together, our observations suggest that particular evolutionary mechanisms may be involved for genes displaying cell type-specific expression, with genes showing MAE or skewed expression as an overlapping group. This possibility is explored in the following section.

MAE and Cell Type-Specific Genes Show Increased Variability and Balancing Selection. We first investigated the role of balancing selection in the evolution of the genes that we identified as showing MAE. The possibility of such a role has been previously suggested theoretically (10, 11), and it received some experimental corroboration recently in human (18). Using data from 28 individuals from four natural populations of mice of *Mus musculus domesticus* (28), we performed a genome-wide scan for regions with an excess of SNPs with intermediate frequencies, which is a signature of balancing selection (i.e., selection for heterozygosity) (29). To investigate the possibility of balancing selection, we used Tajima's *D*, which results from an equation in widespread use (29) to detect evidence of balancing selection in natural populations. Briefly, under the infinite site model assumption, Tajima showed that the quantity D should vanish under conditions of neutral evolution. A positive value of the D statistic results from an increase in the intermediate frequencies of polymorphic nucleotides and suggests that a sequence or a genomic region is subject to balancing selection or another type of diversifying selection. A similar pattern can also be observed if the sampled population was formed from a recent admixture of two different populations (29). Negative values are associated with an excess of rare alleles, purifying selection, and/or expanding populations.

For each gene analyzed, we delineated six gene neighborhoods encompassing the 5′-distal to the 3′-distal region as defined in Fig. 5. We then computed the D statistic for 1-kb windows in each neighborhood and recorded the largest value of D observed in each window. As for the previous analysis, we divided the active genes into two sets, MAE and BAE. We then considered as putative regions under balancing selection Tajima's *D* statistic peaks with values of *D* larger than the 95th percentile. We used the Fisher exact test to assess whether the number of putative peaks under balancing selection in a given neighborhood differed significantly between the *MAE* and *BAE* sets. We observed that the *MAE* set is associated with peaks with a larger *D* statistic than the *BAE* set in introns, exons, and 3′-proximal regions (Fig. 8*A*). Because genes showing MAE are enriched among cell type-specific genes (Fig. 2 *A*, *g* and *h* and *B*, *g* and *h*), we next compared the Tajima's *D* among genes showing monoallelic or biallelic expression in nsc or asl only vs. genes expressed in both cell types. The results are reported in Fig. 8 *B*–*D*. When we compared cell type-specific MAE and BAE

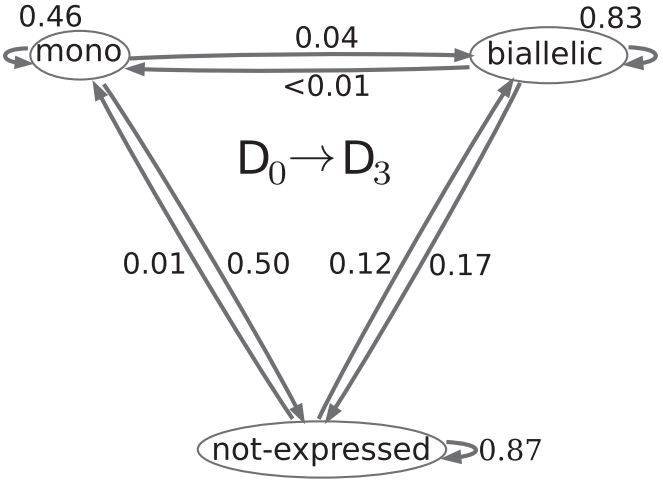

Fig. 4. Probability of S_{D_0} change to S_{D_3} ($S_{D_0} \rightarrow S_{D_3}$), where *S* is one of three possible states of expression: monoallelic, biallelic, or not expressed. D_0 refers to *nsc* before induction of differentiation, and D_3 refers to *asl* after 3 d of differentiation.

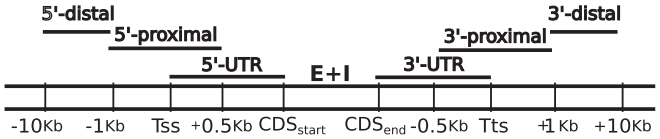

Fig. 5. Gene neighborhood scheme. E + I is the coding portion of exons plus introns, and CDS$_{start}$ and CDS$_{end}$ represent the Coding DNA Sequences starting and end points, respectively.

genes, we observed that the differences were marginal, with the only remaining significant difference being for the 3′-proximal region and introns (P value <0.05) (Fig. 8*B*). When we compared genes showing biallelic expression, we again observed a higher Tajima's D for cell type-specific genes, especially at introns (Fig. 8*D*). We conclude that the increased variability is a property of astrocyte- or neuronal pathway-specific genes, not just genes showing MAE.

Properties of Genes Showing SKE. In earlier sections, we described σ as a monoallelic skew variable and considered the *MAE* set as including only cases where σ was $\geq$0.35. However, as shown in Fig. 9, the expression bias represented by σ is a smooth continuous function. We, therefore, were interested to see if some of the properties observed for MAE vs. BAE genes are a more general property of genes with skewed expression. We asked whether the enrichment for monoallelic expression observed for cell type-specific genes is only the tip of the iceberg (distributionally speaking) of the more general property of expression bias. To explore this possibility, we computed the average value of σ for each gene and divided the genes into two groups: those expressed in only one cell type (*nsc* or *asl*) and those expressed in both cell types. The results shown in Fig. 9 clearly show a distributional difference between the two groups, with biased genes (high $\overline{\sigma}$ value) more likely present among cell type-specific genes (Wilcoxon rank sum test P value $\ll 10^{-5}$). We further investigated whether the distributional difference in $\overline{\sigma}$ observed between the two groups of genes was still present if genes showing the most stringently described MAE were removed from the analysis. To do this, we removed the 712 genes defined as showing MAE (Fig. 2) from both groups of genes and tested whether the cell type-specific genes remained more biased in gene expression compared with the genes expressed in common. The results are reported in *SI Appendix* and show that cell type-specific genes are skewed toward higher σ, even if the *MAE* gene set is removed (Wilcoxon rank sum test P value $\ll 10^{-5}$). Following a similar strategy, we performed GO analysis (20) to see if the ontology categories typically enriched for monoallelic expression also persist for genes showing moderately skewed expression. The results of the analysis of GO-ranked genes (reported in *SI Appendix*) show that skewed genes (high $\overline{\sigma}$) are enriched for categories, such as "extracellular region," "integral component of the membrane," "cell adhesion," and "sensory perception," just as are genes showing MAE. This category remains enriched even if the most skewed genes are removed from the GO ranked list (*SI Appendix*). This suggests that skewed expression may provide some selective advantage for genes involved in development and communication between cells, at least for the nervous system.

Discussion

Nag et al. (5) reported cell type-specific preferences for MAE in mouse lymphoblastoid cell lines using chromatin "signatures" to identify genes with MAE and suggested a link between MAE and cell identity. Here, we report direct measurements of monoallelic expression in mouse *nsc*s and in vitro-generated astrocytes. The use of clonal cells derived from a hybrid of two distantly related mouse subspecies (J1 and B6) has allowed us to distinguish alleles of ∼10,000 genes and follow their ratio of expression during in vitro differentiation of *nsc* to *asl* cells (Fig. 2). Our results confirm and extend previous studies in mouse and human (5, 7, 8), which have established that (*i*) a subset of autosomal genes can be categorized as displaying MAE; (*ii*) the expression state of a given gene, MAE, BAE, or SKE, is persistent in cell culture; and (*iii*) the MAE category is enriched for cell type-specific genes, many of which are membrane-associated genes involved in cell–cell signaling and development. We find that 26% of astrocyte-specific genes show MAE using a strict criterion ($\geq$0.85% expression of a given allele; $\sigma \geq 0.35$). Use of a more relaxed criterion ($\geq$0.75%; $\sigma \geq 0.25$) reveals SKE for an even higher percentage of *asl*-specific genes (56%) (Fig. 2*B*).

Several reports have suggested that MAE is important for normal development as well as disease (3, 30–34). Chess (2), who first found that odorant genes show MAE, has reviewed random MAE and XCI and suggested that MAE and resulting cellular mosaicism have important implications, not only for development but also for evolution (2). We (10, 11) and others (18) have made similar suggestions.

It is widely known that many genotypes display incomplete penetrance (i.e., variable phenotypes). A dramatic example is the observation of incomplete concordance among identical twins for major human genetic disorders, such as schizophrenia and bipolar disorder (35). A very recent and relevant example is bistable epigenetic obesity in mice and humans (9). It also is well known that XCI takes place in the early embryo at about

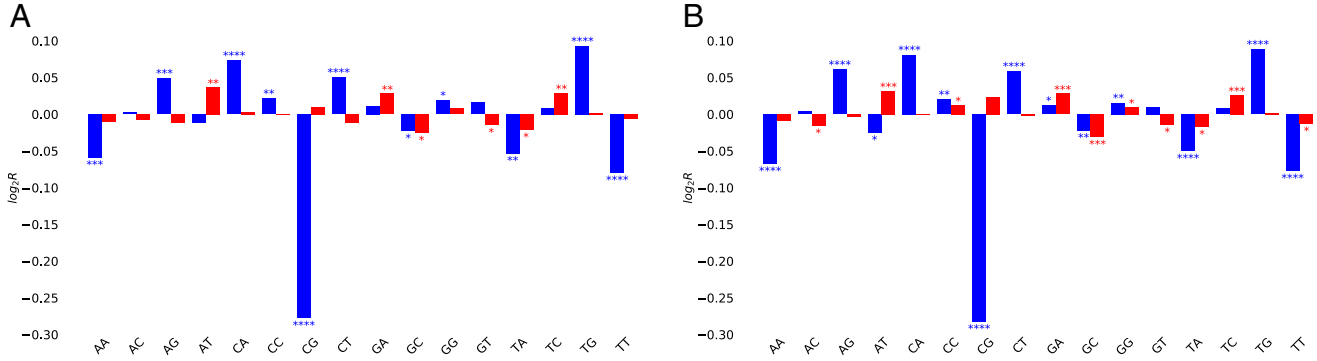

Fig. 6. Comparison (R) of the mean observed to expected dinucleotide distribution for (*A*) genes with MAE vs. BAE and (*B*) genes expressed specifically in *nsc* or *asl* vs. genes expressed in both cell types. Results are shown for the 5′-proximal (blue) and 3′-proximal (red) regions (Fig. 5 shows gene neighborhoods). *$P < 10^{-2}$; **$P < 10^{-5}$; ***$P < 10^{-10}$; ****$P < 10^{-15}$ (Mann–Whitney U rank test).

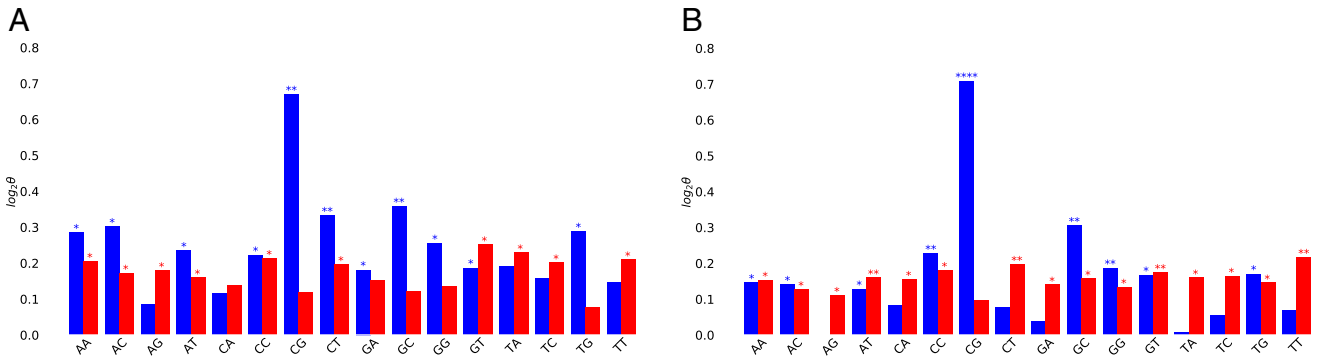

Fig. 7. Comparison (θ) of the mean dinucleotide variability distribution for (*A*) genes with MAE vs. BAE and (*B*) genes expressed specifically in *nsc* or *asl* vs. genes expressed in both cell types. Results are shown for the 5′-proximal (blue) and 3′-proximal (red) regions (Fig. 5 shows gene neighborhoods). *P value $< 10^{-2}$; **$P < 10^{-5}$; ****$P < 10^{-15}$ (Mann–Whitney U rank test).

the time of implantation, and after it is established, it is very stable; that is, monoallelic expression resulting from stochastically random XCI during a developmental window is somatically heritable (2). Somatic heritability is due to epigenetic marks, which include DNA methylation, laid down during a developmental window. Reasoning from XCI and the prevalence of variable phenotypes, we proposed a stochastic epigenetic modification mechanism and did quantitative mathematical modeling (10, 11). The key assumptions of the stochastic epigenetic modification model are (*i*) stochastic and incomplete epigenetic marking during a relatively short developmental window in the early embryo, (*ii*) high-fidelity somatic heritability after the window of variability is closed, and (*iii*) the DNA sequence of *cis*-located elements, which can be due to retroposons or other mutations, affects the probability (ρ) of epigenetic marking of the *cis*-located gene that is subject to selection. Another assumption is that dosage compensation is absent or incomplete, and this has been confirmed for most genes with MAE (8, 36). A significant finding reported here is that skewed expression rather than strict MAE is quite prevalent. With respect to the stochastic epigenetic modification model, it is not critical whether or not at the single-cell level there is complete silencing. The fundamental assumption is that the ratio of expression between the maternal and paternal alleles follows a stochastic rule during the window of opportunity in early development, becoming fixed when the window closes. Quantitative analytical and computational modeling of stochastic epigenetic modification (11) showed that the fixation of genes in a population could be greatly accelerated by stochastic epigenetic modification. Another unanticipated result of our modeling was that overdominance is predicted for reasonable values of ρ (e.g., 0.75). This means that the heterozygote has a selective advantage and will be maintained in the population, increasing genetic diversity.

Whether or not balancing selection is widespread in mammals has been controversial due to lack of convincing evidence for more than just a handful of genes. However, by analyzing SNPs in two human populations, Andrés et al. (13) obtained strong evidence of excess variation and balancing selection for at least 60 human genes. In addition, the laboratory of Gimelbrant and coworkers (7, 18) analyzed nucleotide diversity in genes having a chromatin signature suggestive of MAE. They concluded that human MAE genes contribute disproportionately to genetic diversity and are strongly enriched among those genes identified by Andrés et al. (13) as showing balancing selection. We confirm for mouse the findings of Savova et al. (18) for human that genes with MAE are more variable than genes with BAE. Moreover, we find that *nsc* or *asl* cell type-specific genes are also more variable than genes expressed in both cell types, overlapping with the MAE set of genes. It remains to be determined by future work whether genes specific to the neuronal lineage are more variable than genes in other cell lineages.

Eckersley-Maslin et al. (36) found that overall DNA methylation was no different between genes showing MAE vs. BAE in embryonic stem cells or neuroprogenitor cells, but they did not analyze CpGs or other dinucleotide frequencies. A potential link of MAE to DNA methylation is suggested by a study using the MAUD assay: it showed that a set of genes with a distinct DNA methylation pattern in mouse brain was enriched for genes with MAE in *nsc*s and also for genes showing lineage-specific differentiation within the developing CNS (34).

Savova et al. (18) have reported that, when MAE is compared with BAE, there is an increase in the density of CpGs in the gene body and in particular, in the coding portion of the genes. When

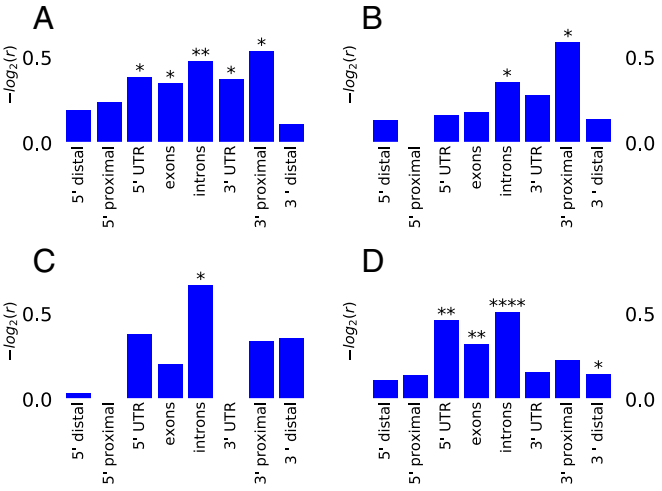

Fig. 8. Balancing selection characterizes both MAE genes and cell type-specific genes (expressed in either *nsc* or *asl* but not in both). The Fisher test was performed, comparing the top 95th percentile Tajima's D (calculated from natural mouse populations) observed in different gene neighborhoods as defined in Fig. 5 and *Material and Methods*. r is the odds ratio of the number of top 95th percentile Tajima's D observed in different groups. *$P < 0.05$; **$P < 10^{-3}$; ****$P < 10^{-10}$ (Fisher test). (*A*) MAE vs. BAE. (*B*) MAE vs. BAE for cell type-specific genes. (*C*) MAE: cell type-specific genes vs. both cell types. (*D*) All expression (MAE or BAE): cell type-specific vs. genes expressed in both cell types.

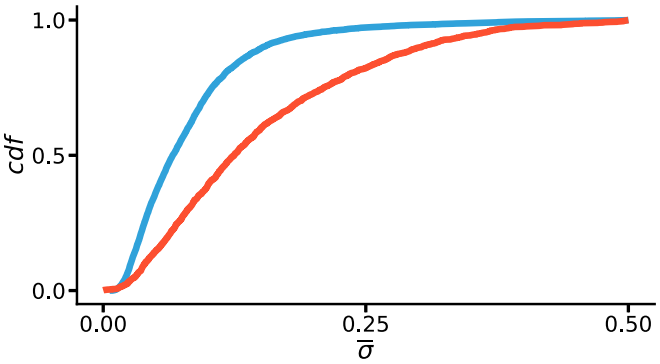

Fig. 9. Developmental genes are associated with SKE expression. The cumulative density distribution is shown for the variable mean of σ (Eq. **1**) for genes expressed in both cell types (blue) or only one cell type (red).

we analyzed the 5′-proximal region, we observed that, in contrast to the gene body, $CpG_{o/e}$ is lower for mouse genes that show MAE vs. those that show BAE, and there is increased dinucleotide variability, especially at CpGs (Fig. 6). This difference persists if housekeeping genes are removed from the analysis (*SI Appendix*). This finding led us to extend our variability analysis to different mouse strains by using Tajima's *D* statistic, which yields a positive value for gene regions under balancing selection (29). We found that the gene body (coding region plus introns) and 3′-proximal regions (but not promoters and 5′-proximal regions) of the MAE genes show a significant positive Tajima's *D* statistic compared with BAE genes (Fig. 8*A*). Housekeeping genes are known to have G + C (GC)- and CpG-rich promoters, with expression levels correlated with GC richness (37). Highly expressed housekeeping genes are also known to be under purifying selection (38), which would yield a negative Tajima's *D* statistic. We, therefore, searched for sites with low Tajima's *D* statistic in BAE or housekeeping genes and found no enrichment compared with MAE genes. This is in agreement with Savova et al. (18), who found that purifying selection similarly affects MAE and BAE genes. Therefore, the relative increase in Tajima's *D* statistic in MAE genes is not likely due to a decrease in *D* score of BAE genes, and it is suggestive of balancing selection. We next found that the difference seen between MAE and BAE genes was mainly due to differences between cell type-specific genes and those genes expressed in both cell types (Fig. 8*D*). We conclude that there is an important, previously unappreciated difference in the variability and Tajima's *D* statistic in the coding regions and introns of cell type-specific genes compared with those expressed by both cell types (including housekeeping genes). Cell type-specific genes are enriched for developmentally important, cell–cell, and cell–environment sensing genes. It is these genes, not just MAE genes, that our data suggest have a positive Tajima's *D* statistic and are likely to be under balancing selection, resulting in increased genetic diversity.

We also addressed the question of whether SKE is subject to similar selection pressure as MAE. Our analysis, shown in Fig. 9, suggests that it is, since cell type-specific genes are skewed toward higher allele-specific expression bias, even when genes meeting our strict definition of MAE are removed (*SI Appendix*). Genes showing SKE were also found to be similarly enriched for the same GO categories as cell type-specific genes (*SI Appendix*). We suggest that skewed expression, resulting from stochastic epigenetic modification during a short developmental window, is evolutionarily important for giving each cell a unique cellular identity, with MAE genes being extreme examples of SKE genes.

Materials and Methods

Single-cell RT-PCR. All mouse procedures were approved by the Institutional Animal Care and Use Committee at City of Hope (American Association for Laboratory Animal Care approval 000720). After euthanasia of postnatal (7.5 d) B6 × JF1 mice and dissection of hippocampi, the Worthington Papain Dissociation System was used to isolate neurons. Cells were diluted to one cell per 2 μL and distributed to 48-well plates. Wells containing single cells were identified microscopically. After addition of Tri-zol and ethanol (50 μL each), RNA was transferred to a Zymo spin collection column (Direct-zol RNA Microprep Kit; Zymo). Each RNA sample was eluted with 12 μL of water and added to a Bioneer Accupower Dual-Hotstart RT-PCR pellet. After addition of primers, the mixture was transferred to 0.65-mL thin-walled Eppendorf tubes, and RT-PCR was performed for 35 cycles. The percentage B6 expression was determined as previously described (8). *Ncam1* and *Gstm5* cDNA primers are listed in *SI Appendix*.

High-Throughput Sequencing. Clonal *nsc* lines were cultured and differentiated, and RNA was prepared for high-throughput (Illumina) sequencing as previously described (4, 8). A JF1 genomic SNP library was also created by use of Illumina sequencing. Briefly, short reads from JF1 DNA-seq were mapped to the mouse genome (GRCm38/mm10) using BWA (v. 0.7.5a) with default settings. Pileup of the sequences was created by samtools (v.0.1.19). The Genome Analysis Toolkit (v.1.4-17) was used with a minimum phred-scaled confidence threshold of calling a variant at 30 and other parameters at default settings. The JF1 DNA SNPs were further filtered with depth of coverage ≥5 and quality ≥30 with consensus probability ≥0.9. Although we have obtained JF1 DNA SNPS for the whole genome, monoallelic expression was assessed using JF1 SNPs located in the exons of Refseq genes. Short reads from RNA-seq were mapped to the mouse genome using TopHat (v.2.014). We allowed the reads to have a maximum of one hit in the genome, and reads with more than seven mismatches were discarded. The inner distance between mate pairs was set at 200 bp, and other parameters were set at default. Expression of Refseq genes was measured as fragments per kilobase of transcript per million mapped reads and was calculated using Cufflinks (v.2.2.1) with default settings.

During the course of this study, another JF1 genomic SNP library was independently published (39).

Mathematical Definitions.

Monoallelic expression. The probability for each gene *i* to be expressed from the B6 allele is

$$P_i^{B6} = \frac{C_i^{B6}}{C_i^{tot}}, \quad [2]$$

where $C_i^{tot} = C_i^{B6} + C_i^{JF1}$ is the total number of SNP counts observed at locus *i* and C_i^{B6} and C_i^{JF1} are the numbers of SNP counts assigned to the *B6* or *JF1* haplotype, respectively (ref. 8 has details).

Moreover, for each gene *i*, the exact binomial test was used to check the null hypothesis $H_0 : P_i^{B6} = P_i^{JF1} = 0.5$, and the probability p_i of rejecting the hypothesis that expression *i* was biallelic was computed. Finally, $FPKM_i$, obtained from RNA-seq experiments, was used as a measure of the level of expression of each gene. Defining *S* as the set of all of the samples, a gene *i* shows *MAE* for a sample $k \in S$ if $\sigma_{ik} \geq 0.35$, $p_{ik} < 0.05$, $C_{ik}^{tot} \geq 10$, $FPKM_{ik} \geq 3$. We define the set *MAE* as the set of all genes with *MAE* in at least one sample and defined as *BAE* genes that are expressed (i.e., with $FPKM \geq 3$) but that do not display MAE in any sample. Finally, we define MAE^{nsc} and MAE^{asl} as the sets of genes that show *mae* in *nsc* or *asl*, respectively, but not both. Similarly, BAE^{nsc} and BAE^{asl} are genes that are expressed (i.e., with $FPKM \geq 3$) in *asl* or *nsc*, respectively, but that do not display MAE in any sample. For cSNP-seq analysis,

$$BAE = AE - MAE, \quad [3]$$

where *AE* is the set of all autosomal expressed genes in our experiments; $MAE^{cell-type}$ and $BAE^{cell-type}$ are the sets of genes showing autosomal monoallelic and biallelic expressions, respectively; and cell type can be *nsc* or *asl*. Formally,

$$MAE^{nsc'} = MAE^{nsc} - MAE^{asl}, \quad [4]$$

$$MAE^{asl'} = MAE^{asl} - MAE^{nsc}, \quad [5]$$

and

$$MAE^{\cap} = MAE^{asl} \cap MAE^{nsc}. \quad [6]$$

Analogously, we can define $BAE^{nsc'}$, $BAE^{asl'}$, and $BAE^{\cap}$ for *BAE* (Fig. 2). We define NEG_{ik} as not expressed (with $i \in ALLG$ and $k \in S$) if its

expression cannot be classified as monoallelic (MAE_{ik}) or biallelic (BAE); formally,

$$NEG_{ik} = \neg BAE_{ik} \wedge \neg MAE_{ik}, \quad [7]$$

where $ALLG$ is the set of all genes i in our experiments with $FPKM_{ik} \geq 3$ in at least one sample k. The results are shown in Fig. 4, where each data point (i.e., gene, cell line, and developmental state) is considered independently.

Observed/expected dinucleotide frequency.

$$NpM_{o/e} = \frac{f(NpM)}{f(N) \times f(M)}, \quad N, M \in \{A, T, C, G\}, \quad [8]$$

with $f(NpM)$, $f(N)$, $f(M)$ equal to the frequency, in the region of interest, of the dinucleotide NpM, respectively (Fig. 6)

Tajima's D statistic.

$$D = \frac{\Pi - S/a}{\sqrt{V(\Pi - S/a)}}, \quad [9]$$

where Π is the average number of pairwise mismatches observed in a set of sequences, S is the number of segregating sites in a sample of n sequences, $a = \sum_{i=1}^{n-1} 1/i$, and $V(\cdot)$ is the operator for the variance.

ACKNOWLEDGMENTS. A.D.R. holds the Samuel Rahbar Chair for Diabetes and Drug Discovery.

1. McConnell MJ, et al. (2017) Intersection of diverse neuronal genomes and neuropsychiatric disease: The brain somatic mosaicism network. *Science* 356:eaal1641.
2. Chess A (2016) Monoallelic gene expression in mammals. *Annu Rev Genet* 50:317–327.
3. Gendrel A, Marion-Poll L, Katoh K, Heard E (2016) Random monoallelic expression of genes on autosomes: Parallels with x-chromosome inactivation. *Semin Cell Dev Biol* 56:100–110.
4. Wang J, Valo Z, Smith D, Singer-Sam J (2007) Monoallelic expression of multiple genes in the cns. *PLoS One* 2:e1293.
5. Nag A, Vigneau S, Savova V, Zwemer L, Gimelbrant A (2015) Chromatin signature identifies monoallelic gene expression across mammalian cell types. *G3 (Bethesda)* 5:1713–1720.
6. Morison IM, Ramsay JP, Spencer HG (2005) A census of mammalian imprinting. *Trends Genet* 21:457–465.
7. Gimelbrant A, Hutchinson J, Thompson B, Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* 318:1136–1140.
8. Li S, et al. (2012) Transcriptome-wide survey of mouse cns-derived cells reveals monoallelic expression within novel gene families. *PLoS One* 7:e31751.
9. Dalgaard K, et al. (2016) Trim28 haploinsufficiency triggers bi-stable epigenetic obesity. *Cell* 164:353–364.
10. Branciamore S, Rodin AS, Gogoshin G, Riggs AD (2015) Epigenetics and evolution: Transposons and the stochastic epigenetic modification model. *AIMS Genet* 2: 148–162.
11. Branciamore S, Rodin AS, Riggs AD, Rodin SN (2014) Enhanced evolution by stochastically variable modification of epigenetic marks in the early embryo. *Proc Natl Acad Sci USA* 111:6353–6358.
12. Dolinoy DC, Weinhouse C, Jones TR, Rozek LS, Jirtle RL (2010) Variable histone modifications at the avy metastable epiallele. *Epigenetics* 5:637–644.
13. Andrés AM, et al. (2009) Targets of balancing selection in the human genome. *Mol Biol Evol* 26:2755–2764.
14. Leffler EM, et al. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339:1578–1782.
15. Croze M, et al. (2017) A genome-wide scan for genes under balancing selection in drosophila melanogaster. *BMC Evol Biol* 17:15.
16. Hedrick PW (2011) Population genetics of malaria resistance in humans. *Heredity* 107:283–304.
17. Alonso S, López S, Izagirre N, de la Rúa C (2008) Overdominance in the human genome and olfactory receptor activity. *Mol Biol Evol* 25:997–1001.
18. Savova V, et al. (2016) Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nat Genet* 48:231–237.
19. Cahoy J, et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function. *J Neurosci* 28:264–278.
20. Beissbarth T, Speed TP (2004) Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 20:1464–1465.
21. Paolella G, Buono P, Mancini F, Izzo P, Salvatore F (1986) Structure and expression of mouse aldolase genes. brain-specific aldolase c amino acid sequence is closely related to aldolase a. *Eur J Biochem* 156:229–235.
22. Vlotides G, Eigler T, Melmed S (2007) Pituitary tumor-transforming gene: Physiology and implications for tumorigenesis. *Endocr Rev* 28:165–186.
23. Jahani-Asl A, et al. (2016) Control of glioblastoma tumorigenesis by feed-forward cytokine signaling. *Nat Neurosci* 19:798–806.
24. Wiley C, et al. (2015) Role for mammalian chitinase 3-like protein 1 in traumatic brain injury. *Neuropathology* 35:95–106.
25. Horvath S, Mirnics K (2014) Immune system disturbances in schizophrenia. *Biol Psychiatry* 75:316–323.
26. Lei Y, Xin X, Morgan D, Pintar J, Fricker L (1999) Identification of mouse cpx-1, a novel member of the metallocarboxypeptidase gene family with highest similarity to cpx-2. *DNA Cell Biol* 18:175–185.
27. Burkin D, Kaufman S (1999) The alpha7beta1 integrin in muscle development and disease. *Cell Tissue Res* 296:183–190.
28. Harr B, et al. (2016) Genomic resources for wild populations of the house mouse, mus musculus and its close relative mus spretus. *Sci Data* 3:160075.
29. Hartl DL, Clark AG (1989) *Principles of Population Genetics*, eds Hartl DL, Clark AG (Sinauer Associates, Sunderland, MA), 2nd Ed.
30. Jeffries AR, et al. (2013) Random or stochastic monoallelic expressed genes are enriched for neurodevelopmental disorder candidate genes. *PLoS One* 8: e85093.
31. Do C, et al. (2016) Mechanisms and disease associations of haplotype-dependent allele-specific dna methylation. *Am J Hum Genet* 98:934–955.
32. Eckersley-Maslin M, Spector D (2014) Random monoallelic expression: Regulating gene expression one allele at a time. *Trends Genet* 30:237–44.
33. Zwemer LM, et al. (2012) Autosomal monoallelic expression in the mouse. *Genome Biol* 13:R10.
34. Wang J, et al. (2010) Dual DNA methylation patterns in the cns reveal developmentally poised chromatin and monoallelic expression of critical genes. *PLoS One* 5:e13843.
35. Torrey EF, Bowler AE, Taylor EH, Gottesman II (1994) *Schizophrenia and Manic-Depressive Disorder: The Biological Roots of Mental Illness as Revealed by the Landmark Study of Identical Twins* (Basic Books, New York), Vol 41, pp 424–425.
36. Eckersley-Maslin M, et al. (2014) Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev Cel* 28:351–365.
37. Zhu J, He F, Hu S, Yu J (2008) On the nature of human housekeeping genes. *Trends Genet* 24:481–484.
38. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21:236–239.
39. Takada T, et al. (2013) The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. *Genome Res* 23:1329–1338.