Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome

Stefan F. Pinter,*,†,‡,1 David Colognori,*,†,‡,1 Brian J. Beliveau,‡ Ruslan I. Sadreyev,*,†,‡ Bernhard Payer,*,†,‡,3 Eda Yildirim,*,†,‡,2 Chao-ting Wu,‡ and Jeannie T. Lee*,†,‡,§,4

*Howard Hughes Medical Institute, Chevy Chase, Maryland 20815-6789, †Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, ‡Department of Genetics and §Department of Pathology, Harvard Medical School, Boston, MA 02115

ORCID IDs: 0000-0003-4750-1403 (S.F.P.); 0000-0003-1314-3118 (B.J.B.); 0000-0002-4694-2082 (B.P.)

ABSTRACT In mammals, several classes of monoallelic genes have been identified, including those subject to X-chromosome inactivation (XCI), genomic imprinting, and random monoallelic expression (RMAE). However, the extent to which these epigenetic phenomena are influenced by underlying genetic variation is unknown. Here we perform a systematic classification of allelic imbalance in mouse hybrids derived from reciprocal crosses of divergent strains. We observe that deviation from balanced biallelic expression is common, occurring in ~20% of the mouse transcriptome in a given tissue. Allelic imbalance attributed to genotypic variation is by far the most prevalent class and typically is tissue-specific. However, some genotype-based imbalance is maintained across tissues and is associated with greater genetic variation, especially in 5′ and 3′ termini of transcripts. We further identify novel random monoallelic and imprinted genes and find that genotype can modify penetrance of parental origin even in the setting of large imprinted regions. Examination of nascent transcripts in single cells from inbred parental strains reveals that genes showing genotype-based imbalance in hybrids can also exhibit monoallelic expression in isogenic backgrounds. This surprising observation may suggest a competition between alleles and/or reflect the combined impact of *cis*- and *trans*-acting variation on expression of a given gene. Our findings provide novel insights into gene regulation and may be relevant to human genetic variation and disease.

KEYWORDS gene expression; genotype; random monoallelic expression; imprinting; allele-specific RNA-seq

BIALLELIC gene expression affords diploid organisms a safeguard against single-hit mutations that may prove detrimental in development or cause disease. Exceptions to this rule have been limited to genes involved in parental resource allocation (imprinting) (Bartolomei and Ferguson-Smith 2011; Ferguson-Smith 2011) and developmental processes that necessitate expression of a single allele to achieve either cell heterogeneity [random monoallelic expression (RMAE), *e.g.*, olfactory receptor genes)] (Gimelbrant *et al.* 2007; Chess 2012; Zwemer *et al.* 2012) or sex chromosome dosage compensation [X-chromosome inactivation (XCI)] (Wutz and Gribnau 2007; Payer and Lee 2008; Starmer and Magnuson 2009; Disteche 2012; Lee and Bartolomei 2013). Although the notion that genotypic variation can modify phenotypic expression has long been appreciated, the extent to which divergence between parental genomes influences gene expression in offspring has not been rigorously addressed, partly because allelic imbalance due to genetic variation and imprinting cannot be distinguished in human studies (Cheung *et al.* 2010; Heap *et al.* 2010; Montgomery *et al.* 2010; Pickrell *et al.* 2010; Battle *et al.* 2014). Inbred mouse lines represent a powerful tool to address this question, given the possibility of reciprocal cross-design and *a priori* knowledge of genotype. Previous studies in mice have focused largely on RMAE (Zwemer *et al.* 2012; Eckersley-Maslin *et al.* 2014; Gendrel *et al.* 2014), X-linked (Yang *et al.* 2010; Yildirim *et al.* 2011; Pinter *et al.* 2012), and imprinted genes (Babak *et al.* 2008; Deveale *et al.* 2012; Prickett and Oakey 2012), specifically excluding allelic imbalance attributed to genotypic variation

doi: 10.1534/genetics.115.176263
Manuscript received December 8, 2014; accepted for publication March 27, 2015; published Early Online April 9, 2015.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176263/-/DC1
Sequencing data have been submitted to the GEO database at NCBI as series GSE58524.
[1]These authors contributed equally to this work.
[2]Present address: Department of Cell Biology, Duke University School of Medicine, Durham, NC 27710
[3]Present address: Centre for Genomic Regulation (CRG), Dr. Aiguader, 88 08003 Barcelona, Spain
[4]Corresponding author: E-mail: lee@molbio.mgh.harvard.edu

[henceforth, genotypically imbalanced expression (GTIE)] with some exceptions (Goncalves *et al.* 2012; Lagarrigue *et al.* 2013). Interestingly, recent single-cell RNA sequencing (RNA-Seq) studies have suggested stochastic and uncoordinated expression of single alleles (Deng *et al.* 2014; Marinov *et al.* 2014; Borel *et al.* 2015), as suggested previously (Raj *et al.* 2006; Raj and van Oudenaarden 2008; Dar *et al.* 2012), though high levels of technical noise precluded conclusive quantitative analysis of this phenomenon (Deng *et al.* 2014).

Here we analyze transcriptomes of hybrid mice mated in reciprocal crosses to measure and classify allelic imbalance due to genotype, imprinting, or RMAE, including attenuated random imbalance. We find that genotype-based imbalance is (1) more prevalent than imprinting or RMAE, (2) generally tissue-specific but typically consistent in allelic preference when present in two tissues, and (3) associated with elevated genetic variation in proximal noncoding sequences. Surprisingly, GTIE genes often show monoallelic expression in inbred parental lines, potentially revealing a predisposition for imbalanced expression in outbred hybrids. Because hybrid genomes share one nucleoplasm and hence their *trans*-acting factors, differences in allelic representation linked to genotype can be attributed to variants acting in *cis*. Our data demonstrate that a large number of genes contain *cis*-acting variants that skew their expression in favor of one allele, an observation that may prove relevant to the study of human genetic variation and disease.

Materials and Methods

Mouse crosses and RNA-Seq

Parental and F_1 hybrid tail-tip fibroblasts (TTFs) of *Mus musculus castaneus* and *M.m. musculus* origins were prepared from 2- to 5-day-old pups or obtained from Payer *et al.* (2013). TTFs and mouse embryonic fibroblasts (MEFs) from similar reciprocal crosses were immortalized by SV-40 T-antigen (Brown *et al.* 1986) and either subcloned by limiting dilution or used as a heterogeneous population. The sex and strain of each parent, as well as the sex of the offspring from which cells were isolated, were varied to obtain each of the possible combinations listed in Figure 1E. Total RNA from TTFs was harvested using the mirVana RNA Isolation Kit (Ambion, Austin, TX), followed by polyA selection of messenger RNA (mRNA) and directional RNA-Seq library generation on the Apollo 324 System (IntegenX, Pleasanton, CA) at the Biopolymers Facility (Harvard Medical School, Boston, MA). Parental and F_1 hybrid RNA libraries (Supporting Information, Table S1) were sequenced to a depth of typically ~60 million paired-end 50-nt reads per sample. Read pairs were aligned to both parental genomes, allowing for a maximum of three mismatched or gapped bases per read (Tophat v2.0.8), followed by removal of PCR duplicates and multiply mapping reads. Read pairs considered allele-specific aligned (1) to only one parental genome, (2) with both ends to one but only a single end to the other parental genome, or (3) with fewer mismatches/insertions/deletions to one parental genome than to the other. Allele-specific pairs and pairs aligning equally well to both genomes were used for transcript assembly and quantitation (Cufflinks v2.1.1) using Ensembl Release 66.

Classification of allelic imbalance

Exonic variants for each annotated transcript were queried for allele-specific counts and aggregated across the transcript as in Deveale *et al.* (2012). To systematically identify genes preferentially expressed from one allele over the other, as well as genes exhibiting balanced expression within a given sample, a number of qualifying criteria were applied to each gene (File S1). Transcripts lacking sufficient expression and assembly support and showing only spurious allelic coverage were excluded. For each assessable gene, the skew (R) and the cumulative probability (P) under a binomial distribution were calculated either under the assumption of equal probability for mapping to either allele or, where sufficient parental data were available, by empirically estimating technical bias for each gene and accounting for this bias by adjusting the null probability. A number of genes (486 in liver samples, 1008 in TTFs, mostly pseudo- or predicted genes) skewing significantly in the direction of the nonrepresented allele in the parental controls were excluded from further analysis.

To estimate the likelihood of detecting significantly skewed genes ($P < 0.05$, cumulative binomial probability) reproducibly across a number of samples, each allele-specific read was first randomly and repeatedly redistributed between the two alleles for each data set to build a null model based on actual sample sizes (number of allelic reads for each gene in each experiment). Genes skewing significantly more often across experiments than expected under the null model then were assessed in 2×2 contingency tables (forward/reverse crosses *vs.* genotypes) to determine whether allelic skew associates with parental or genetic origin. Genes associating significantly (two-tailed $P < 0.05$, Barnard's exact test) and unanimously with parental or genetic origin were classified as "imprinted" or showing "genotypically imbalanced expression" (GTIE), respectively. In light of the low number of samples for TTFs, genes with borderline significance in this test ($0.05 < P < 0.07$) were classified as "potentially" imprinted or "GTIE" and classified as separate categories indicated with lowercase letters (*e.g.*, *mat*, *pat*, *gt1*, and *gt2*). Genes skewing significantly more often than expected under the null model yet lacking significant association with either parental or genetic origin (two-tailed $P > 0.07$, Barnard's exact test) and skewing toward each allele at least once across experiments were classified as "randomly" skewed. A small number of genes lacking significant association because they were detected in only a small number of samples yet skewing only toward one allele were classified as "undetermined" (38 in liver samples, 434 in TTFs). To provide estimates for false discovery rates (FDRs) in each category (balanced, imprinted, random, and GTIE), one randomized data set was carried through the classification

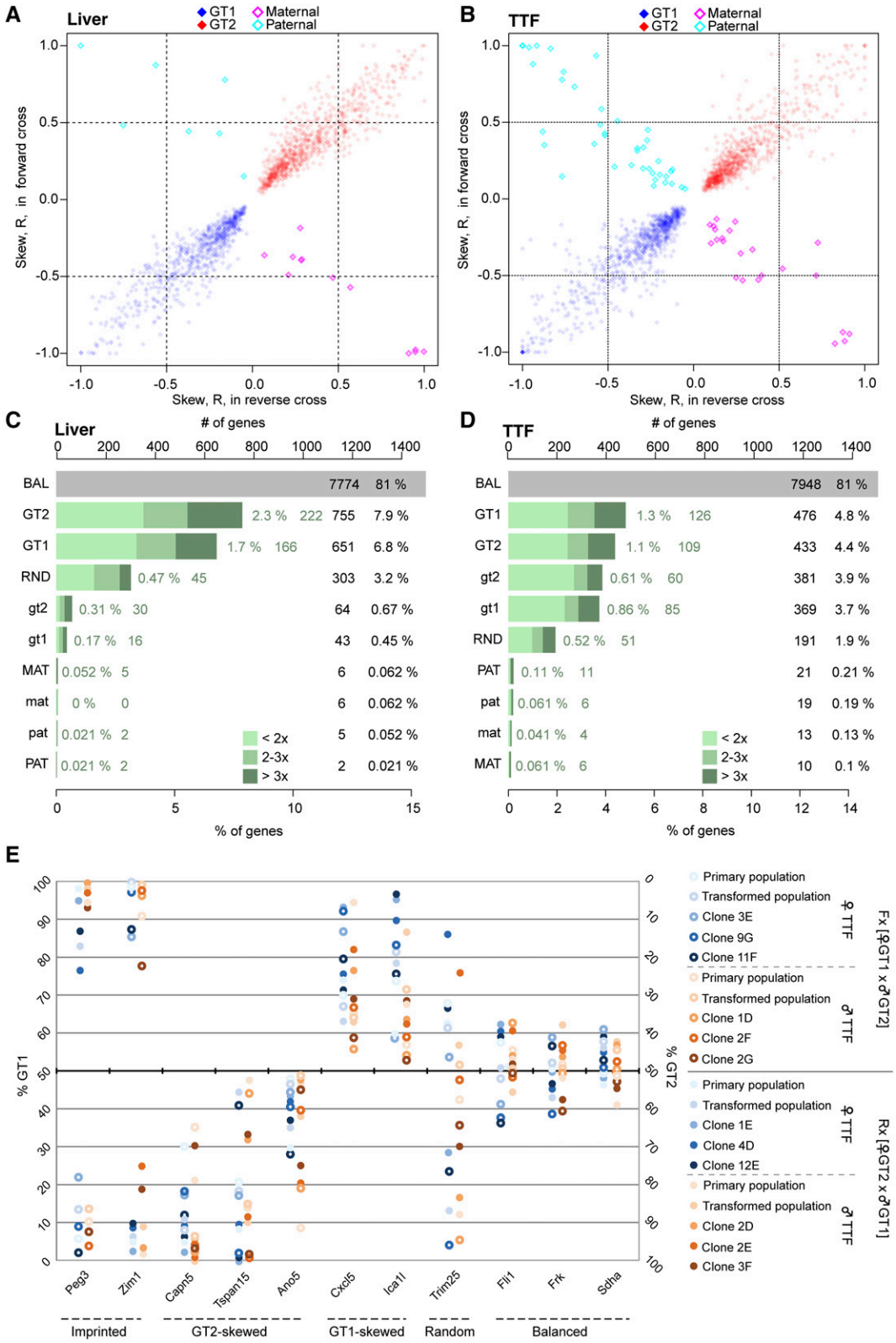

Figure 1 One-fifth of autosomal genes are preferentially expressed from one allele. (A and B) Distribution of median skew values *R* from −1 to +1, or 100–0% *CAST/EiJ* origin (GT1) [where read ratio *R* = (genotype 1 − genotype 2)/(genotype 1 + genotype 2)] for GTIE (red, blue) and imprinted genes (open symbols, magenta, cyan) in forward and reverse crosses of liver (A) and TTF (B) samples. (C and D) Counts of balanced (BAL) and skewed autosomal genes shaded by fold difference between alleles for liver (C) and TTF (D) samples. Imprinted (MAT/PAT) and GTIE (GT1/GT2) categories are capitalized ($P < 0.05$, two-tailed Barnard's exact test) or lowercased ($0.05 < P < 0.07$), and randomly skewed (RND). Expressed X, Y, and mitochondrial genes (C and D: 208 and 340, respectively), genes skewed in only a single cross (C and D: 38 and 434), and genes not expressed or lacking variant

(FDR < 0.02 for all the categories described earlier in both liver and TTF samples). A full list of genes showing allelic imbalance in liver and TTFs can be found in File S2 and File S3, respectively.

Feature and enrichment analyses of imbalanced genes

Imprinted and randomly skewed genes were compared to collections of known imprinted (mousebook.org) and recently identified random monoallelic genes (Zwemer *et al.* 2012; Eckersley-Maslin *et al.* 2014; Gendrel *et al.* 2014). Significance of enrichment in the corresponding categories was determined under a hypergeometric distribution using the union of balanced, imprinted, random, and GTIE genes as a background set. Equivalent testing was performed to assess the significance of overlap with specific Ensembl gene categories (biotypes), ribosomal protein genes, and category identity between liver and TTF tissues. Similarly, gene ontology (GO) annotations were queried using the same background set of expressed and assessable genes, and GO-term enrichment *P*-values were determined by hypergeometric test (adjusted $P < 0.05$). For comparison of quantitative features [fragments per kilobase per million (FPKM), allelic read support, gene and exon length, and density of sequence variation over exonic, intronic, upstream, and downstream intervals] between categories (balanced, imprinted, random, and GTIE), the Mann-Whitney *U*-test was used, and significance values were denoted as described in the figure legends.

RNA-DNA FISH

TTFs from *CAST/EiJ*, *129S1/SvImJ*, or F_1 hybrid mice were cytospun onto glass slides, extracted with CSKT buffer (100 mM NaCl, 300 mM sucrose, 10 mM PIPES, 3 mM $MgCl_2$, and 0.5% Triton X-100, pH 6.8) and fixed in 4% paraformaldehyde. DNA probes were generated from fosmid or BAC clones (*Apbb1ip*: WIBR1-0890H01; *Pon3*: WIBR1-1690D10; *Ywhag*: WIBR1-2316B24; *Spred2*: RP24-294M15; *Egfr*: RP23-263C13; *Grb10*: RP24-345M19; Children's Hospital Oakland Research Institute, Oakland, CA) and labeled by nick translation using Cy3-conjugated dUTP. For detection of RNA, ~50 ng of probe and ~1 μg of mouse Cot-1 DNA in 10% dextran sulfate, 50% formamide, and 2× SSC (300 mM NaCl and 30 mM sodium citrate, pH 7.4) were denatured at 95° prior to application to slides and hybridization overnight at 42°. Slides then were washed with 50% formamide and 2× SSC and mounted, and images were captured on a Nikon Eclipse 90i microscope workstation with Volocity software (Improvision, Coventry, England). For detection of DNA, slides were pretreated with RNase A (400 μg/mL in PBS) at 37° for 40 min and denatured in 70% formamide and 2× SSC at 80° for 10 min. Methods and probe design for the allele-specific DNA FISH experiments using Oligopaint FISH probe sets are described elsewhere (Beliveau *et al.* 2015).

Allele-specific qualitative reverse transcriptase polymerase chain reaction (qRT-PCR)

RNA was isolated from cells using TRIzol Reagent (Invitrogen, Carlsbad, CA), from which cDNA was generated with SuperScript III Reverse Transcriptase (Invitrogen) and Oligo (dT)15 Primer (Promega, Madison, WI). For qRT-PCR, 250 nM each of universal and either GT1 (*CAST/EiJ*)– or GT2 (*C57BL6/J* or *129S1/SvJm*)–specific primer and SYBR Green Supermix (Bio-Rad, Hercules, CA) were used. The allele-specific primers were designed to target genetic variants according to the method of TaqMAMA (Glaab and Skopek 1999) or further optimized for allelic discrimination as described previously (Li *et al.* 2004). Expression levels were compared against the opposite allele using the formula fold difference = 2^Ct(GT2 − GT1) and corrected for primer bias and differential primer efficiencies by comparison with amplification of pure parental or hybrid genomic DNA (see File S10 for primer information).

Results and Discussion

To investigate how genotypic variation may contribute to gene expression on a transcriptome-wide scale, we analyzed RNA-Seq data from primary and clonal populations of cells representing developmentally distinct tissues—specifically, liver cells (Goncalves *et al.* 2012) and TTFs. Using parental and hybrid mice from reciprocal crosses, we categorized genes by frequency and direction of allelic imbalance and validated our findings using orthogonal methods. We began by designing a statistical framework to take advantage of the count-based nature of sequencing data while controlling for overdispersion based on empirical estimates derived from inbred parental samples (Figure S1, A–C). Imbalanced gene calls were based on aggregate read counts across each transcript as in Deveale *et al.* (2012) and, for most genes, represented at least two exons and over five variants, with excellent agreement between individual variants (Figure S1, D–I). To classify allelic imbalance, we tested whether significant overrepresentation of one allele ($P < 0.05$, cumulative binomial) was detected reproducibly across samples. We estimated the significance of reproducing allelic imbalance against multiply permuted data sets generated by randomly distributing reads between two alleles. Association of allelic imbalance with genetic or parental origin was assessed using 2 × 2 contingency tables (Barnard's exact test). Reproducibly imbalanced genes lacking association were classified as random, provided that each allele was the preferred allele at least once. Not all randomly imbalanced genes

information (C and D: 24,255 and 23,475) were excluded; 34,110 genes in total (Ensembl release 66). (E) Relative allelic expression levels (from 0 to 100% GT1, left axis, or GT2, right axis) based on allele-specific qRT-PCR for the indicated genes (*Capn5*, *Tspan15*, and *Ano5*: GT2-skewed; *Cxcl5* and *Ica1l*: GT1-skewed; *Peg3* and *Zim1*: maternally, paternally imprinted; *Trim25*: random; *Fli*, *Frk*, and *Sdha*: balanced). Primary and transformed populations and three independent clones for both sexes in forward (Fx) and reverse (Rx) crosses are as indicated in the legends.

represent instances of RMAE, however, because this distinction would have required applying an arbitrary cutoff to define RMAE. Instead, we report all randomly imbalanced genes irrespective of the magnitude of the imbalance, as we do for GTIE and imprinted genes. Our approach detected allelic imbalance across a wide range of expression and treated each sample in each cross independently to conservatively estimate the reproducibility and direction of allelic imbalance (Figure S2, A and B).

We observed that GTIE was more frequent than imprinting in both liver and TTFs (ascending *vs.* descending diagonal, respectively) (Figure 1, A and B), with allelic skew [read ratio $R = (\mathrm{GT2} - \mathrm{GT1})/\mathrm{total}$] stretching across the full range from 100% *CAST/EiJ* [genotype 1 (GT1), in blue, $R = -1$] to 0% [genotype 2 (GT2), in red, $R = 1$]. Because of the lower number of F_1 hybrid samples in TTFs, we applied a relaxed cutoff ($P < 0.07$, two-tailed Barnard's exact test) to detect additional candidate imprinted and GTIE genes (denoted by lowercase labels in Figure 1, C and D). Overall, ~20% of genes appeared to be imbalanced, almost half of which showed greater than twofold difference between the two alleles (Figure 1, C and D). Approximately 5% of all expressed genes were candidates for monoallelic expression (greater than threefold difference between alleles) in both liver (488) and TTFs (458). Overall, GTIE was associated with higher, not lower, expression than was seen with balanced genes (Figure S3A), indicating that the detected allelic imbalance is not restricted to lowly expressed genes. In addition, we excluded genes that were more likely to be subject to technical noise because of very low read numbers (see Supporting Materials and Methods, File S1). Although median expression decreased somewhat with increasing magnitude of imbalance (Figure S3B), the range of GTIE expression levels still overlapped the range for balanced genes.

Because RMAE is known to be a clonal event, detection of randomly imbalanced genes not only in clonal TTFs but also in primary liver cells was unexpected. However, allelic read support for randomly imbalanced genes in liver samples, but not TTF clones, dropped off with increasing allelic fold difference compared to GTIE genes (Figure S3C). In addition, randomly imbalanced genes in liver were detected in a lower fraction of samples and were depleted in monoallelic genes (greater than threefold difference) compared to TTFs (Figure S3D). These observations are consistent with the possibility that biopsies from solid tissue can capture spatially linked subsets of cells that preserve some clonal mosaicism present in the animal. Interestingly, randomly imbalanced genes in TTFs were lower in length-normalized expression (FPKM) than balanced and GTIE genes (Figure S3, A and B). This difference could be attributed entirely to their significantly longer transcripts and gene bodies (Figure S3E), an observation that may prove useful in elucidating the mechanisms of RMAE.

To verify imbalanced expression of gene candidates, we performed allele-specific qRT-PCR using complementary DNA (cDNA) reverse-transcribed from RNA of hybrid clonal cell lines representing forward and reverse mouse crosses (Figure 1E). In addition, we tested primary and immortalized cells of males and females to rule out potential transformation artifacts and sex-specific phenomena, respectively. GTIE genes (*Capn5*, *Tspan15*, and *Ano5*: GT2-skewed; *Cxcl5* and *Ica1l*: GT1-skewed) showed significant (greater than twofold) skewing in the direction observed in RNA-Seq across the majority of samples, whereas imprinted genes (*Peg3*: paternally expressed; *Zim1*: maternally expressed) showed appropriate parent-specific expression patterns, and randomly skewed (*Trim25*) and balanced (*Fli1*, *Frk*, and *Sdha*) control genes were centered around 50:50 expression. Importantly, skewing of imbalanced genes was observed repeatedly, even in heterogeneous primary and immortalized populations, demonstrating that allelic imbalance cannot be attributed solely to clonal expansion but is reproducible in direction and magnitude of skewing at the population level.

Genes imbalanced owing to parental origin were expectedly enriched in known imprinted genes ($P < 6 \times 10^{-17}$, $P < 4 \times 10^{-19}$, hypergeometric test, liver and TTFs, respectively) (Figure 2, A and B). Hierarchical clustering of imprinted genes by the numerical score we assigned to allelic skew (R from -1 to 1) correctly separated samples by forward and reverse cross and genes by maternal and paternal expression. Surprisingly, expression of some known imprinted genes was associated with genotype rather than parental origin. These included genes that cluster within larger imprinted domains (Ono *et al.* 2003) but show tissue-specific imprinting, such as *Ddc* (Menheniott *et al.* 2008) (next to *Grb10* in Figure 2C) in liver and *Pon2*, *Pon3*, and *Ppp1r9a* (Schulz *et al.* 2006) (next to *Sgce* and *Peg10* in Figure 2D) in TTFs. Moreover, despite expected paternal expression of *Peg10* and *Sgce* in this cluster, recently identified imprinting candidate *Casd1* (Babak *et al.* 2008) showed balanced expression (Figure 2D), and neighboring *Asb4* (Mizuno *et al.* 2002) and *Tfpi2* (Monk *et al.* 2008) were randomly imbalanced rather than imprinted, though intriguingly in anti-correlated fashion with each other (Figure 2B). In sum, these examples support the notion that even within a single imprinted cluster, genotype can modify penetrance of parent-of-origin effects and reveal that several genes imprinted in the placenta (Proudhon and Bourc'his 2010) are subject to GTIE in adult tissues. Among our novel imprinted candidates (*e.g.*, *Vcp*, *Fam46b*, and neighboring *Trnp1*), allelic imbalance was attenuated compared to most known imprinted genes, with the exception of *H13* (Figure 2, E, F and G). It is plausible that partial and tissue-specific imprinting patterns are more difficult to detect than complete and widespread parent-of-origin effects, raising the question of how many partially imprinted genes exist but have gone undetected and how they may be regulated. Likewise, our collection of randomly imbalanced genes includes instances of RMAE, raising the question of whether arbitrary numerical cutoffs for defining monoallelic expression should be applied to define the phenomenon of RMAE or whether attenuated allelic imbalance, as observed for many imprinted

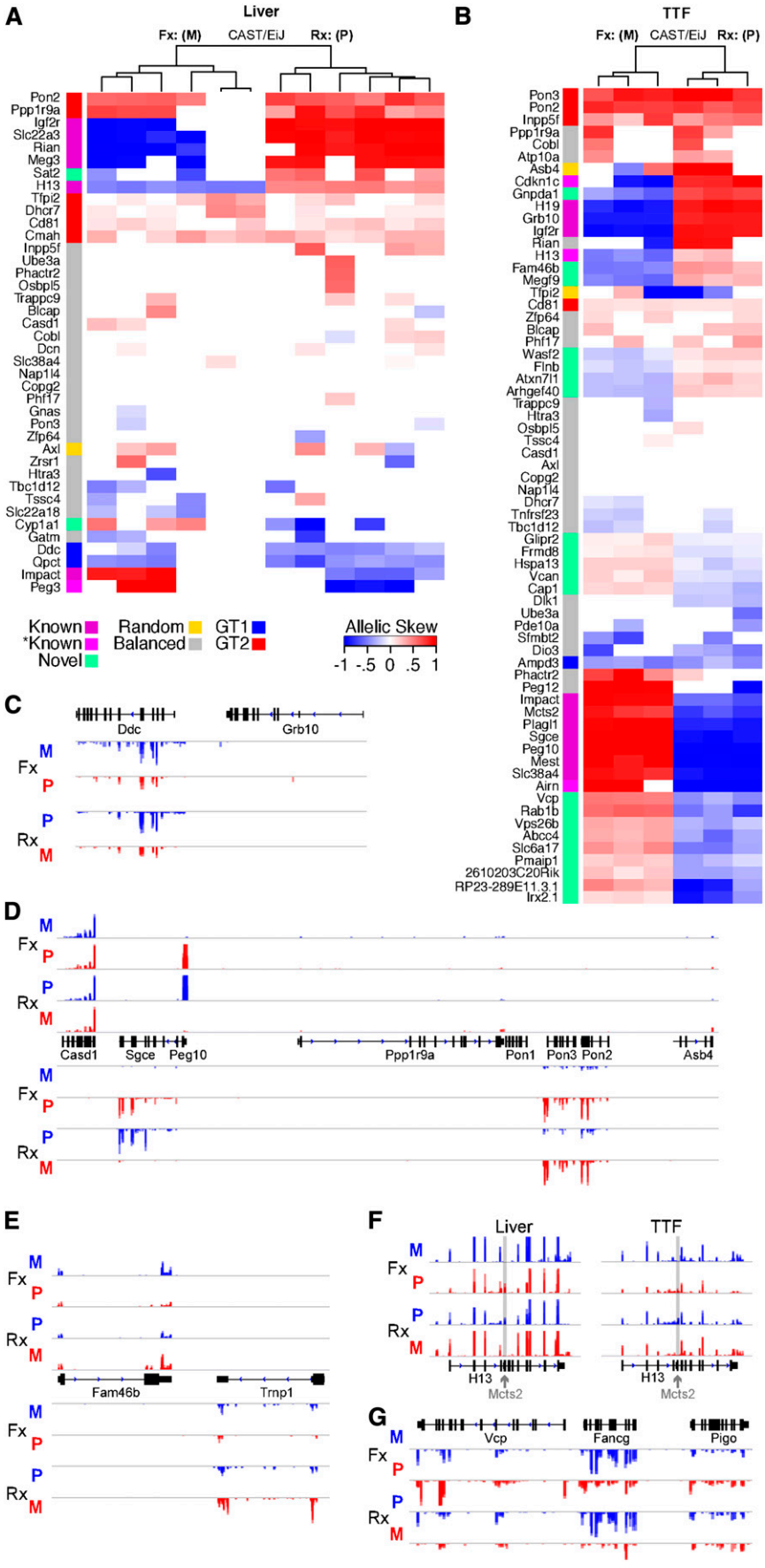

Figure 2 Genotype can affect some genes in imprinted clusters. (A and B) Allelic skew values [blue-red scale from −1 to 1 or 100–0% *CAST/EiJ* origin (GT1), respectively] of previously known and novel imprinted genes in forward (Fx, *CAST/EiJ* maternal, M) and reverse crosses (Rx, *CAST/EiJ* paternal, P) of individual liver (A) and TTF (B) samples. Color-coded panels next to gene names label previously known imprinted genes (www.mousebook.org) as either confirmed herein (dark/light purple) or scoring as GTIE (blue: GT1; red: GT2), randomly skewed (gold), or balanced (gray). Novel imprinted gene candidates (bright green, $P < 0.05$, two-tailed Barnard's exact test) and known imprinted genes ($P < 0.05$ or *, $0.05 < P < 0.07$) cluster liver (A) and TTF samples (B) by parental origin (into forward and reverse crosses) and genes by maternal (top) and paternal (bottom) expression. Previously known imprinted genes (6 of 8 and 10 of 31 in liver and TTFs, respectively) are expectedly enriched (hypergeometric test, $P < 6 \times 10^{-17}$ in liver, $P < 4 \times 10^{-19}$ in TTFs). (C and D) GTIE of known imprinted genes in liver (C) and TTFs (D). Allele-specific GT1 (blue) and GT2 (red) coverage in forward (Fx) and reverse (Rx) crosses (M, maternal; P, paternal). Multiple samples combined for each overlay track (+ strand above genes, − strand below). (E–G) Examples of incomplete penetrance of parental origin in imprinted candidates *Fam46b* and *Trnp1* (E) and *Vcp* (G) compared to incomplete penetrance of known imprinted gene *H13* in both liver and TTFs (F).

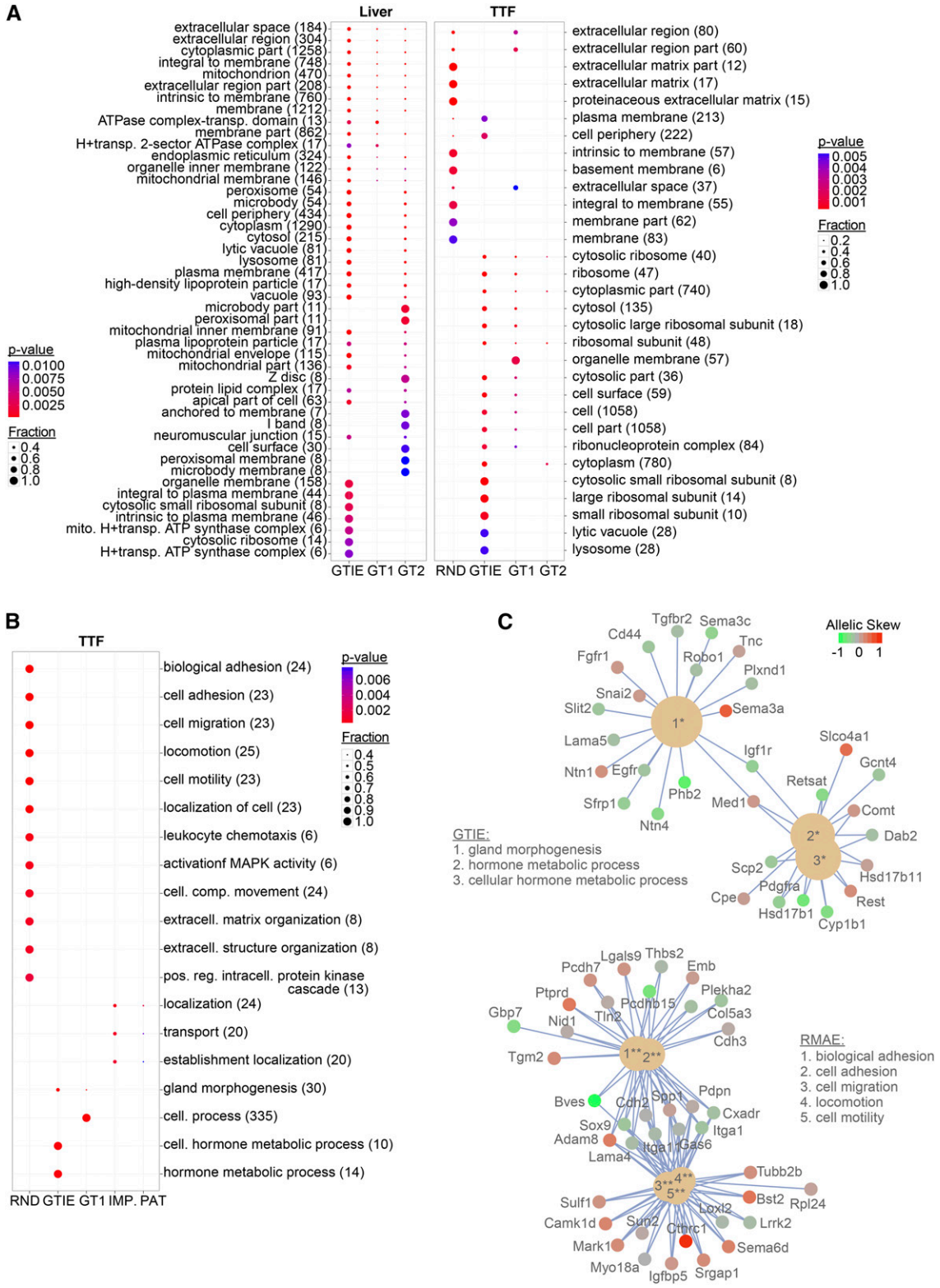

Figure 3 Randomly skewed and GTIE genes are enriched in distinct cellular component and biological process annotations. (A and B) Enriched cellular component (A) GO terms in liver and TTFs and biological processes (B) in TTFs. Bubble color denotes significance of enrichment, bubble size (see side legend) indicates the fraction of genes linked to a given term out of all genes across allelic imbalance classes (GTIE; either GT1 or GT2; RND, random). (C) Examples of enriched biological processes (brown circles, * and **, adjusted $P < 0.05$ and $P < 0.01$, hypergeometric test) for randomly skewed and GTIE genes in TTF samples. Median allelic skew for each gene is color-coded (−1, green, to +1, red, for 100–0% *CAST/EiJ*, respectively).

genes (*e.g.*, *H13*) (Figure 2F)(Goncalves *et al.* 2012), should be included, as was done here.

We next addressed whether imbalanced genes were enriched in specific transcript classes or annotations. RMAE and monoallelic GTIE candidates were modestly enriched in antisense, pseudogene, and long intergenic noncoding RNA (lincRNA) classes, but their numbers were small (<10 for any category). Analysis of GO terms revealed that GTIE and RMAE genes are enriched in distinct cellular component and biological process annotations (Figure 3, A and B, and File S4, File S5, File S6, File S7, File S8, and File S9). In agreement with previous reports (Chess 2012; Gendrel *et al.* 2014), randomly skewed genes in TTFs, for example, were enriched in cell surface and adhesion/migration categories (Figure 3, A–C), where monoallelic representation may serve to increase cell heterogeneity. Interestingly, GTIE genes in TTFs also were enriched in gland morphogenesis and hormone metabolism terms, which may reflect inherent biological differences between the parental strains used in this experiment (*M.m. musculus vs. M.m. castaneus*). In addition, we found significant enrichment of ribosomal protein (RP) genes (Figure S4, A and B) among GTIE genes in both liver ($P < 5 \times 10^{-3}$) and TTFs ($P < 4 \times 10^{-9}$, hypergeometric test). This observation is intriguing because of the known monoallelic state of ribosomal RNA genes (Schlesinger *et al.* 2009) and is reminiscent of nucleolar dominance in plants (Ge *et al.* 2013). Although we cannot entirely rule out that allelic calls were less reliable in these genes because of the high number of RP pseudogenes in mammalian genomes (Zhang *et al.* 2002), we see this explanation as less likely given that only uniquely mapping reads were used for both variant mapping and RNA-Seq analyses, and moreover, RP genes were enriched even when considering intronic variants in allelic imbalance calls since RP pseudogenes in rats and mice lack introns (Wool *et al.* 1995; Zhang *et al.* 2004). Future study of GTIE in RP genes may yield insight into the elusive mechanism behind their coordinated expression (Hu and Li 2007).

The genetic variants that give rise to allelic differences in RNA abundance may affect either the act of transcription or the transcript itself and may act in a tissue-specific or tissue-spanning fashion. To determine what fraction of genes comprises each of the latter two categories, we compared allelic calls across tissues in the 7465 genes that were expressed in both liver and TTFs (77.4 and 72.5% of all genes expressed in these samples). We found that 5129 genes (<69%) were balanced in both tissues, 1916 genes (~26%) were imbalanced in one tissue, and 420 genes (~6%) were imbalanced in both tissues (Figure 4, A–C). Considering that we find 31% of genes to skew in at least one of two tissues sampled from a single mouse cross (*M.m. musculus vs. M.m. castaneus*), it is likely that an even larger number of genes will show allelic imbalance across additional tissues and/or genetic backgrounds. Importantly, of genes classified as GTIE in both liver and TTFs, 79% preferred the same allele in both tissues (Figure 4, A and B), switching only occasionally (21% of GTIE genes) (Figure 4, A and B). This enrichment was highly significant (Figure 4C). We investigated whether the primary variants acting on these genes were more likely to be found in close proximity (*e.g.*, promoters, splicing motifs, and untranslated regions) rather than in distal, possibly tissue-specific enhancers. While identification of causal variants is beyond the scope of this analysis, we asked whether imbalanced genes were generally more divergent than balanced genes (Figure 4D and Figure S5, A and B). Indeed, variant density in GTIE genes was significantly higher than in balanced genes (Figure 4D, *P*-values on plot, and Figure S5, A and B, significance plotted in bottom panels, Mann-Whitney *U*-test) across coding (exonic) and noncoding sequences, and this difference appeared to be sensitive to distance (least different at ±10 kb). Importantly, the significance of difference in medians between balanced (gray) and GTIE genes (blue, red) was consistently greater in the gene body than in up/downstream regions but significant ($P < 0.05$, Mann Whitney *U*-test) in both. Of note, the magnitude of allelic imbalance generally lacked correlation with variant density (Figure S6, A and B), indicating that while a greater variant density may increase detection of allelic imbalance, it does not appear to increase its magnitude. These results suggest that the overall likelihood, not the magnitude, of GTIE for a given gene is proportional to the level of divergence present in its two alleles. How individual variants (coding *vs.* noncoding, proximal *vs.* distal) contribute to the observed allelic imbalance, however, will require taking into account additional epigenomic and transcriptomic information. For example, cognate sites for DNA or RNA binding proteins, such as transcription or RNA splicing/processing factors, may be especially helpful in this regard because their occupancy is likely to alter nascent transcription or RNA stability and to be sensitive to single-nucleotide changes.

Because most GTIE genes in TTFs are balanced in liver and vice versa (Figure 4, A and B), it is plausible that their allelic imbalance is at least in part determined by variants in tissue-specific enhancers. Conversely, genes that maintain direction of allelic imbalance across tissues may be subject to proximal variants that exert their effect on either cotranscriptional processes or the transcript itself and thus may override effects of distal enhancer-associated variants. We therefore asked how the distribution of proximal variants differed between genes that maintained GTIE across tissues (245, green in top panels of Figure 4E) and genes that showed GTIE in only a single tissue (1623, purple in top panels of Figure 4E) or genes balanced in both tissues (5129, gray in top panels of Figure 4E). Indeed, genes that reproduce the direction of allelic imbalance across tissues were more divergent (Figure 4E, bottom panels, Mann-Whitney *U*-test) than genes imbalanced in a single tissue (orange lines in bottom panels of Figure 4E) or genes balanced in both tissues (green lines in bottom panels of Figure 4E). However, the significance of the former comparison (orange lines above dotted line in bottom panels of Figure

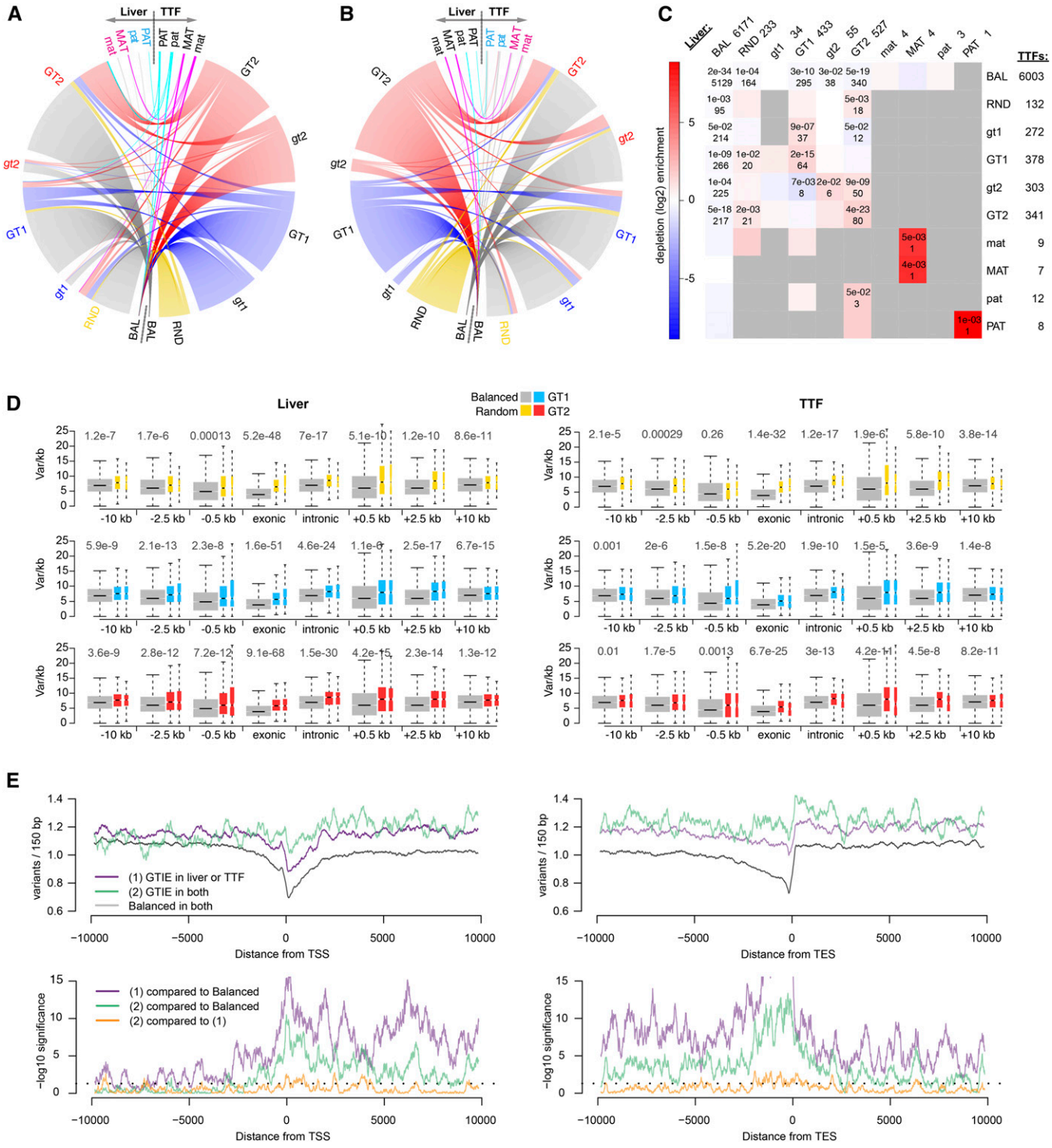

Figure 4 GTIE genes expressed across tissues frequently maintain direction of skew. (A and B) Comparison of skewed gene categories across tissues. Bar plots, reporting the number of genes in a given category, are arranged around a circle. Bar sizes are proportional to the number of genes in a given category except for balanced genes. Membership in a given category in TTFs can be traced across the circle to the corresponding categories in liver (A) and vice versa (B). (A) GTIE genes in TTFs (GT1/GT2, blue/red in right half-circle) typically recapitulate skew direction in liver (left half-circle) unless expressed in a balanced fashion (BAL, gray) and vice versa. (B) Call in liver genes (left half-circle); in TTFs (right half-circle). Randomly skewed genes (RND, gold) replicate poorly across tissues. (C) Maintenance of direction for GTIE genes across tissues is statistically significant (hypergeometric test, *P*-values indicated in quilt plot). Colors denote enrichment (red) and depletion (blue) as seen in scale on left. (D) Variant density in skewed genes (gold, blue, red) is significantly higher than in balanced (gray) genes (*P*-values above, Mann-Whitney *U*-test) in exonic, intronic, and proximal noncoding sequence (±0.5, 2.5, and 10 kb) for liver (left) and TTFs (right). Median indicated by black horizontal mark; box width based on number of genes in each category. Thinner colored box in each comparison contains only genes skewing more than threefold. Immediate promoter sequence (−0.5 kb) is more divergent

4E) was restricted to sequence that is part of the transcript proper (*i.e.*, UTRs as opposed to, *e.g.*, the promoter). This observation suggests that variants in transcript termini have a strong influence on allelic imbalance detected across tissues, likely because they become part of the actual transcript and can affect processes downstream of transcription (*e.g.*, splicing, polyadenylation, export, and turnover).

We next asked whether the GTIE observed by RNA-Seq, which provides a population average across millions of cells, is recapitulated at the single-cell level. In principle, overrepresentation of one allele in a population of cells can result from imbalanced expression within each cell or from a disproportionate number of cells each expressing one preferred allele. To parse out these possibilities, we performed sequential nascent RNA-DNA FISH on several candidate genes in hybrid TTF clones (Figure 5A). A known imprinted gene (*Grb10*) presented with one RNA spot in almost all cells, as expected. Predicted GTIE genes (*Apbb1ip*, *Egfr*, and *Pon3*) presented with one RNA spot in most cells but also exhibited a smaller fraction of cells with two spots. Conversely, predicted balanced genes (*Spred2* and *Ywhag*) presented with two spots in most cells, with *Spred2* also exhibiting a smaller fraction with one spot ($n > 200$). These data, taken together with our allele-specific RNA-Seq and qRT-PCR results, suggest that within a clonal population of cells, GTIE may stem in part from a difference in transcriptional firing rates between the two alleles, with a single allele being preferred in most cells at any given time.

To trace the genotypic origin of transcripts at the single-cell level, we repeated the RNA-DNA FISH using a novel allele-specific Oligopaint DNA FISH probe technology (Beliveau *et al.* 2012, 2015), with probe sets designed to target variants specific to either GT1 or GT2 (Figure 5B). By tiling an ~2.5-Mb region around the genes of interest with a pair of differentially labeled Oligopaint probe sets, with one being specific for GT1 and the other for GT2, this approach allowed us to discern whether the nascent RNA signal colocalized with the GT1 or GT2 allele. As expected, imprinted gene *Grb10* was expressed exclusively from the GT1 allele (pink bar, left panels, in Figure 5B) in almost all cells derived from a forward mouse cross (Fx) but from the GT2 allele (green bar, right panels, in Figure 5B) in cells derived from the reverse cross (Rx). Conversely, GTIE gene *Egfr* was expressed primarily from the GT1 allele (1.5- to 1.7-fold compared to GT2, in excellent agreement with the RNA-Seq estimate of 1.5- to 1.8-fold) in both crosses. However, *Spred2* was expressed equally between both alleles, with expression being biallelic in most of cells, or expressed at equal frequency from either allele in the minority of cells with only one signal. Thus, the combination of RNA FISH and allele-specific DNA FISH examining single-cell dynamics corroborates the conclusions drawn from RNA-Seq analysis of cell populations.

In principle, allelic imbalance due to genotype (GTIE) observed in hybrids could be imagined to take three possible forms in isogenic cell lines of the respective genotypes: (1) biallelic expression in cells of one parental strain and much reduced expression in cells of the other, (2) biallelic expression in cells of both parental strains, or (3) stable or stochastic monoallelic expression in cells of both parental strains. In scenario 1, one of the parental genomes carries a nonfunctional allele, causing hybrids to receive only one functional allele from the other parental genome. In scenario 2, nonequivalence between alleles only manifests in the hybrid (in the presence of divergent variants). In scenario 3, genes subject to RMAE or stochastic monoallelic expression in isogenic cells gain a genotypic preference in hybrids (similar to the *Xce* effect on XCI choice) (Migeon 1998). As seen in Figure 5C (GT1, GT2), GTIE genes *Apbb1ip*, *Egfr*, and *Pon3* remained largely monoallelic in isogenic cells of either parental strain, as did imprinted gene *Grb10*, whereas balanced genes *Spred2* and *Ywhag* remained mostly biallelic. We conclude that, at least for the genes tested here, the observed GTIE in hybrid cells may reflect random or stochastic monoallelic expression in cells derived from isogenic strains. We propose that reported stochastic and uncoordinated firing of single alleles (Raj *et al.* 2006; Dar *et al.* 2012) observed in isogenic cells (Figure 5C) may be a prerequisite for the GTIE observed in F_1 hybrids (Figure 5A), possibly because the presence of divergent variants in F_1 cells may determine which allele fires more frequently. However, it is worth noting that the presence of these variants also appears to exacerbate the apparent "competition" between alleles, as seen by the increase in occurrence of monoallelic cells for each GTIE gene among hybrid *vs.* isogenic backgrounds (compare cell counts in Figure 5, A and C).

In summary, we have determined the extent of allelic imbalance in hybrids as a function of divergence between parental genomes and have provided a classification of genes imbalanced by genetic variation and epigenetic mechanisms. Although the notion that genetic variation can modify epigenetic phenomena has been appreciated in unique cases (*e.g.*, *Xce* in XCI) (Rastan and Cattanach 1983), our findings extend this to include instances of genomic imprinting. These findings are mirrored in very recent work (published while

in GTIE (GT1 in blue, GT2 in red) and randomly skewed genes (gold). (E) Tissue-spanning allelic imbalance is associated with increased genetic divergence in transcribed sequence. Averaged binned density of genetic variants that distinguish parental genomes is plotted across ±10 kb of sequence centered on either the TSS (left) or TES (right). Variants across balanced genes (gray), GTIE genes in liver, or TTFs (line 1, purple) and GTIE genes in both tissues (line 2, green) were binned at approximately nucleosome resolution (150 bp) and assessed for difference in medians. Top panels show variant density; bottom panels denote significance of difference between these groups (with dotted lines representing $P = 0.05$, Mann-Whitney *U*-test). Both lines 1 and 2 are compared to balanced genes and are identified by their respective colors (purple, green) in the bottom panel. Orange lines plot the difference in median variant density comparing GTIE genes in liver or TTFs (1) and GTIE genes maintaining their preferred allele in both tissues (2).

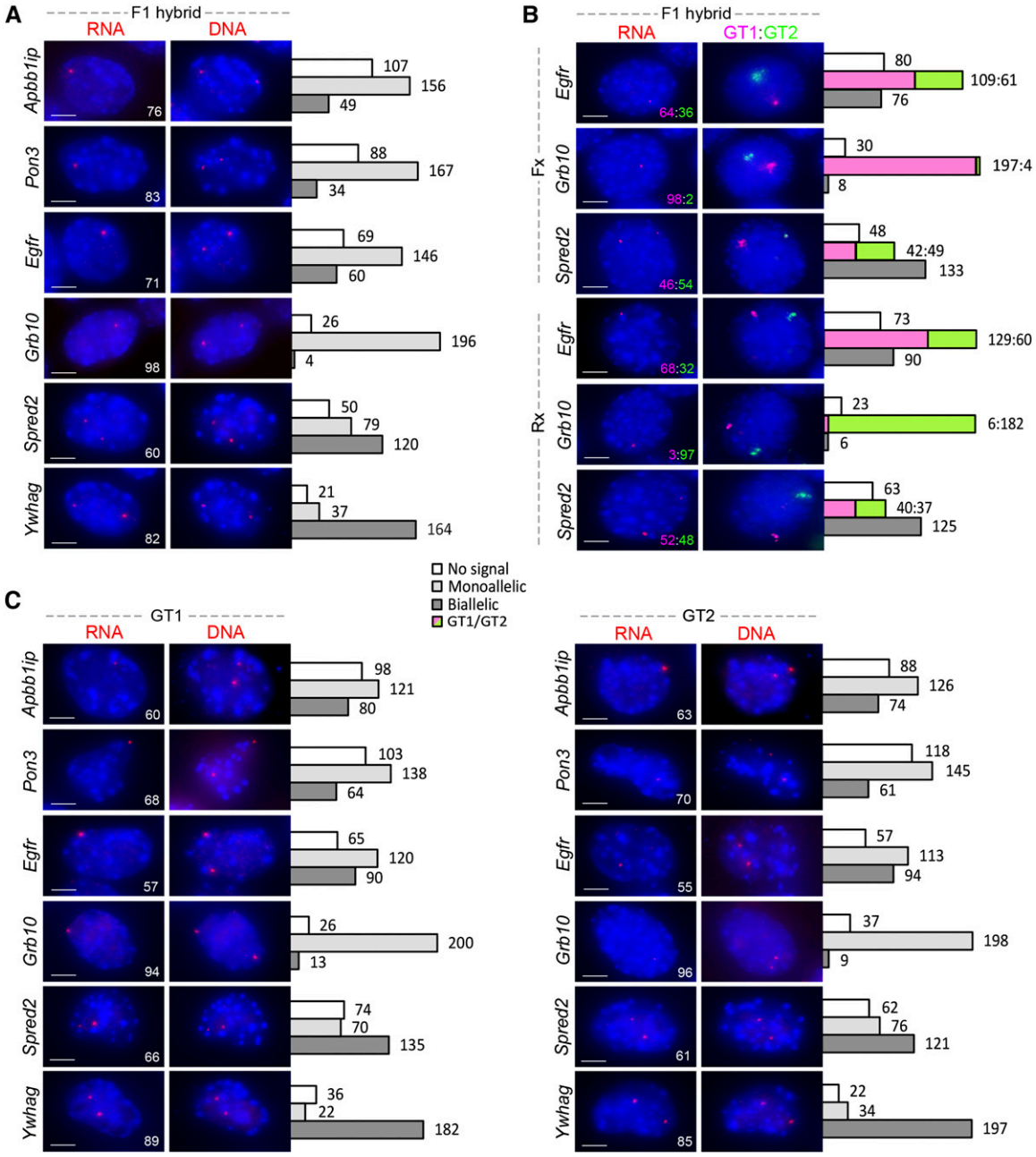

Figure 5 RNA-DNA FISH demonstrates monoallelic expression in individual cells of hybrid and parental origins. (A) Sequential RNA-DNA FISH for *Apbb1ip*, *Egfr*, and *Pon3* (GTIE); *Grb10* (imprinted); and *Spred2* and *Ywhag* (balanced) in clonal TTFs from F_1 hybrid animals. Diploid nuclei with RNA signal ($n > 200$) were scored for each sample, with representative images shown. Inset values list percentage of cells with RNA signal presenting with the number of spots (either one or two) shown in the corresponding image. Adjacent bar graphs list total counts for cells displaying no signal or monoallelic or biallelic expression. All genes presented with two spots in subsequent DNA FISH, verifying probe specificity and 2*n* ploidy. RNA spots always colocalized with DNA spots. Scale bar, 5 μm. (B) As in A but repeated using allele-specific Oligopaint probe sets for DNA FISH in TTFs from F_1 hybrid animals of both forward (Fx) and reverse (Rx) crosses. Adjacent bar graphs now display additional allelic resolution obtained for monoallelic population (GT1, magenta; GT2, green). Inset values list proportion of RNA signal coming from GT1 *vs.* GT2 alleles. (C) As in A but repeated using clonal TTFs from isogenic parental (GT1, GT2) lines.

this paper was under review): Crowley *et al.* (2015) present evidence for *cis*-regulatory variation affecting imprinting patterns, including many cases of incomplete imprinting. Furthermore, their work supports our assertion above that the majority of mammalian genes may exhibit allelic imbalance depending on available cell types and genotypes.

In addition, our results provide the surprising insight that genes subject to genotype-based skewing identified in hybrids can already be imbalanced in isogenic parental cells. This confluence of genetics and epigenetics may reflect a competition between two alleles for limiting *trans*-acting factors or nuclear niches that results in RMAE or stochastic

monoallelic expression in isogenic cells but manifests as a genotype-dependent choice in the presence of divergent alleles (one strong, one weak). Alternatively, these genes may be subject to concurrent *cis*- and *trans*-acting regulatory variation. Goncalves *et al.* (2012) hypothesize that genes that show genotype-based allelic imbalance in the hybrid yet no differential expression between the two parental isogenic lines (as in Figure 5C) reflect an epistatic relationship between *cis*- and *trans*-acting regulatory variation that is uncovered only in hybrid cells as a result of sharing of *trans*-acting factors. In other words, variation that changes the activity or abundance of a given transcription factor may have resulted in selection of compensatory *cis*-regulatory mutations in its corresponding target genes. In the hybrid, however, both sets of *trans*-acting factors act on these *cis*-regulatory elements, leading to loss of *cis*-driven compensation and emergence of differential allelic expression. These two models are not mutually exclusive, and both remain untested for now.

The relevance of allelic imbalance likely extends to all outbred diploid organisms, including humans, especially in view of loss of heterozygosity and haploinsufficiency (Savova *et al.* 2013). Although the parental genomes of most humans are less divergent than those of the two mouse lines used in this study, the level of heterozygosity is likely predictive of the fraction of genes showing allelic imbalance, and by extrapolation, these genes may be expected to number in the hundreds. In fact, recent human single-cell analyses revealed ~6% (35 of 568) of assessable genes to undergo nonstochastic allelic imbalance, and human epigenomes show widespread allelic bias (8–14% of genes in any given H1 embryonic stem cell lineage) (Dixon *et al.* 2015). Interestingly, RMAE genes in humans have recently been shown to feature a distinctive chromatin signature consistent with allelic differences in transcription (Nag *et al.* 2013). GTIE and RMAE may play an important role in the penetrance of recessive and dominant traits in development and may contribute to human diseases, for which a large number of genome-wide association studies incriminate noncoding genetic variants (Zhang *et al.* 2014).

Acknowledgments

We thank A. Saltzman for critical reading of the manuscript, M. Tolstorukov for valuable discussions, and all laboratory members for helpful suggestions. We acknowledge H. Sunwoo for FISH protocols and W. Press for assistance with handling of mice. This work was supported by grants from the German Research Foundation (DFG) and the MGH Fund for Medical Discovery to S.F.P., NIH training grant 2T32GM007226-37A1 to D.C., NIH grant R01-GM61936 and the NIGMS Director's Pioneer Award to C-t.W., and NIH grant R01-GM090278 to J.T.L. J.T.L. is an investigator at the Howard Hughes Medical Institute. The authors declare no conflicts of interest.

Author contributions: D.C. generated cell lines and performed allele-specific qRT-PCR and FISH experiments. S.F.P. performed bioinformatics and statistical analyses of allele-specific RNA-Seq and DNA-Seq data. R.I.S. performed preliminary bioinformatics analyses of published data sets (Pinter *et al.* 2012; Yildirim *et al.* 2011). B.P. and E.Y. contributed cell lines. B.J.B. and C-t.W. contributed allele-specific FISH methodology and probes. S.F.P., D.C., and J.T.L conceived of and designed the study and wrote the manuscript. J.T.L. obtained funding and supervised the research.

Literature Cited

Babak, T., B. Deveale, C. Armour, C. Raymond, M. A. Cleary *et al.*, 2008 Global survey of genomic imprinting by transcriptome sequencing. Curr. Biol. 18: 1735–1741.

Bartolomei, M. S., and A. C. Ferguson-Smith, 2011 Mammalian genomic imprinting. Cold Spring Harb. Perspect. Biol. 3: a002592.

Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman *et al.*, 2014 Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24: 14–24.

Beliveau, B. J., A. N. Boettiger, M. S. Avendaño, R. Jungmann, R. B. McCole *et al.*, 2015 Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. Nat. Commun. DOI: 10.1038/ncomms8147.

Beliveau, B. J., E. F. Joyce, N. Apostolopoulos, F. Yilmaz, C. Y. Fonseka *et al.*, 2012 Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. USA 109: 21301–21306.

Borel, C., P. G. Ferreira, F. Santoni, O. Delaneau, A. Fort *et al.*, 2015 Biased allelic expression in human primary fibroblast single cells. Am. J. Hum. Genet. 96: 70–80.

Brown, M., M. McCormack, K. G. Zinn, M. P. Farrell, I. Bikel *et al.*, 1986 A recombinant murine retrovirus for simian virus 40 large T cDNA transforms mouse fibroblasts to anchorage-independent growth. J. Virol. 60: 290–293.

Chess, A., 2012 Mechanisms and consequences of widespread random monoallelic expression. Nat. Rev. Genet. 13: 421–428.

Cheung, V. G., R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins *et al.*, 2010 Polymorphic *cis*- and *trans*-regulation of human gene expression. PLoS Biol. 8: e1000480.

Crowley, J. J., V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci *et al.*, 2015 Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. 47: 353–360.

Dar, R. D., B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum *et al.*, 2012 Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc. Natl. Acad. Sci. USA 109: 17454–17459.

Deng, Q., D. Ramsköld, B. Reinius, and R. Sandberg, 2014 Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343: 193–196.

DeVeale, B., D. van der Kooy, and T. Babak, 2012 Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS Genet. 8: e1002600.

Disteche, C. M., 2012 Dosage compensation of the sex chromosomes. Annu. Rev. Genet. 46: 537–560.

Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget *et al.*, 2015 Chromatin architecture reorganization during stem cell differentiation. Nature 518: 331–336.

Eckersley-Maslin, M. A., D. Thybert, J. H. Bergmann, J. C. Marioni, P. Flicek *et al.*, 2014 Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev. Cell 28: 351–365.

Ferguson-Smith, A. C., 2011 Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. 12: 565–575.
Ge, X. H., L. Ding, and Z. Y. Li, 2013 Nucleolar dominance and different genome behaviors in hybrids and allopolyploids. Plant Cell Rep. 32: 1661–1673.
Gendrel, A.-V., M. Attia, C.-J. Chen, P. Diabangouaya, N. Servant *et al.*, 2014 Developmental dynamics and disease potential of random monoallelic gene expression. Dev. Cell 28: 366–380.
Gimelbrant, A., J. N. Hutchinson, B. R. Thompson, and A. Chess, 2007 Widespread monoallelic expression on human autosomes. Science 318: 1136–1140.
Glaab, W. E., and T. R. Skopek, 1999 A novel assay for allelic discrimination that combines the fluorogenic 5′ nuclease polymerase chain reaction (TaqMan) and mismatch amplification mutation assay. Mutat. Res. 430: 1–12.
Goncalves, A., S. Leigh-Brown, D. Thybert, K. Stefflova, E. Turro *et al.*, 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. 22: 2376–2384.
Heap, G. A., J. H. M. Yang, K. Downes, B. C. Healy, K. a. Hunt *et al.*, 2010 Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum. Mol. Genet. 19: 122–134.
Hu, H., and X. Li, 2007 Transcriptional regulation in eukaryotic ribosomal protein genes. Genomics 90: 421–423.
Lagarrigue, S., L. Martin, F. Hormozdiari, P.-F. Roux, C. Pan *et al.*, 2013 Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics 195: 1157–1166.
Lee, J. T., and M. S. Bartolomei, 2013 X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*. 152: 1308–1323
Li, B., I. Kadura, D. J. Fu, and D. E. Watson, 2004 Genotyping with TaqMAMA. Genomics 83: 311–320.
Marinov, G. K., B. a. Williams, K. McCue, G. P. Schroth, J. Gertz *et al.*, 2014 From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 24: 496–510.
Menheniott, T. R., K. Woodfine, R. Schulz, A. J. Wood, D. Monk *et al.*, 2008 Genomic imprinting of Dopa decarboxylase in heart and reciprocal allelic expression with neighboring Grb10. Mol. Cell. Biol. 28: 386–396.
Migeon, B. R., 1998 Non-random X chromosome inactivation in mammalian cells. Cytogenet. Cell Genet. 80: 142–148.
Mizuno, Y., Y. Sotomaru, Y. Katsuzawa, T. Kono, M. Meguro *et al.*, 2002 Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. Biochem. Biophys. Res. Commun. 290: 1499–1505.
Monk, D., A. Wagschal, P. Arnaud, P. S. Muller, L. Parker-Katiraee *et al.*, 2008 Comparative analysis of human chromosome 7q21 and mouse proximal chromosome 6 reveals a placental-specific imprinted gene, TFPI2/Tfpi2, which requires EHMT2 and EED for allelic-silencing. Genome Res. 18: 1270–1281.
Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle *et al.*, 2010 Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.
Nag, A., V. Savova, H.-L. Fung, A. Miron, G.-C. Yuan *et al.*, 2013 Chromatin signature of widespread monoallelic expression. eLife 2: e01256.
Ono, R., H. Shiura, H. Aburatani, T. Kohda, T. Kaneko-Ishino *et al.*, 2003 Identification of a large novel imprinted gene cluster on mouse proximal chromosome 6. Genome Res. 13: 1696–1705.
Payer, B., and J. T. Lee, 2008 X chromosome dosage compensation: how mammals keep the balance. Annu. Rev. Genet. 42: 733–772.
Payer, B., M. Rosenberg, M. Yamaji, Y. Yabuta, M. Koyanagi-Aoi *et al.*, 2013 Tsix RNA and the germline factor, PRDM14, link X reactivation and stem cell reprogramming. Mol. Cell 52: 805–818.
Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt *et al.*, 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.
Pinter, S. F., R. I. Sadreyev, E. Yildirim, Y. Jeon, T. K. Ohsumi *et al.*, 2012 Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. Genome Res. 22: 1864–1876.
Prickett, A. R., and R. J. Oakey, 2012 A survey of tissue-specific genomic imprinting in mammals. Mol. Genet. Genomics 287: 621–630.
Proudhon, C., and D. Bourc'his, 2010 Identification and resolution of artifacts in the interpretation of imprinted gene expression. Brief. Funct. Genomics 9: 374–384.
Raj, A., C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, 2006 Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 4: e309.
Raj, A., and A. van Oudenaarden, 2008 Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135: 216–226.
Rastan, S., and B. M. Cattanach, 1983 Interaction between the Xce locus and imprinting of the paternal X chromosome in mouse yolk-sac endoderm. Nature 303: 635–637.
Savova, V., S. Vigneau, and A. A. Gimelbrant, 2013 Autosomal monoallelic expression: genetics of epigenetic diversity? Curr. Opin. Genet. Dev. 23: 642–648.
Schlesinger, S., S. Selig, Y. Bergman, and H. Cedar, 2009 Allelic inactivation of rDNA loci. Genes Dev. 23: 2437–2447.
Schulz, R., T. R. Menheniott, K. Woodfine, A. J. Wood, J. D. Choi *et al.*, 2006 Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies. Nucleic Acids Res. 34: e88.
Starmer, J., and T. Magnuson, 2009 A new model for random X chromosome inactivation. Development 136: 1–10.
Wool, I. G., Y. L. Chan, and A. Gluck, 1995 Structure and evolution of mammalian ribosomal proteins. Biochem. Cell Biol. 73: 933–947.
Wutz, A., and J. Gribnau, 2007 X inactivation Xplained. Curr. Opin. Genet. Dev. 17: 387–393.
Yang, F., T. Babak, J. Shendure, and C. M. Disteche, 2010 Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Res. 20: 614–622.
Yildirim, E., R. I. Sadreyev, S. F. Pinter, and J. T. Lee, 2011 X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. Nat. Struct. Mol. Biol. 19: 56–61.
Zhang, X., S. D. Bailey, and M. Lupien, 2014 Laying a solid foundation for Manhattan—'setting the functional basis for the post-GWAS era'. Trends Genet. 30: 140–149.
Zhang, Z., N. Carriero, and M. Gerstein, 2004 Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 20: 62–67.
Zhang, Z., P. Harrison, and M. Gerstein, 2002 Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res. 12: 1466–1482.
Zwemer, L. M., A. Zak, B. R. Thompson, A. Kirby, M. J. Daly *et al.*, 2012 Autosomal monoallelic expression in the mouse. Genome Biol. 13: R10.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176263/-/DC1

Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome

Stefan F. Pinter, David Colognori, Brian J. Beliveau, Ruslan I. Sadreyev, Bernhard Payer, Eda Yildirim, Chao-ting Wu, and Jeannie T. Lee

DOI: 10.1534/genetics.115.176263

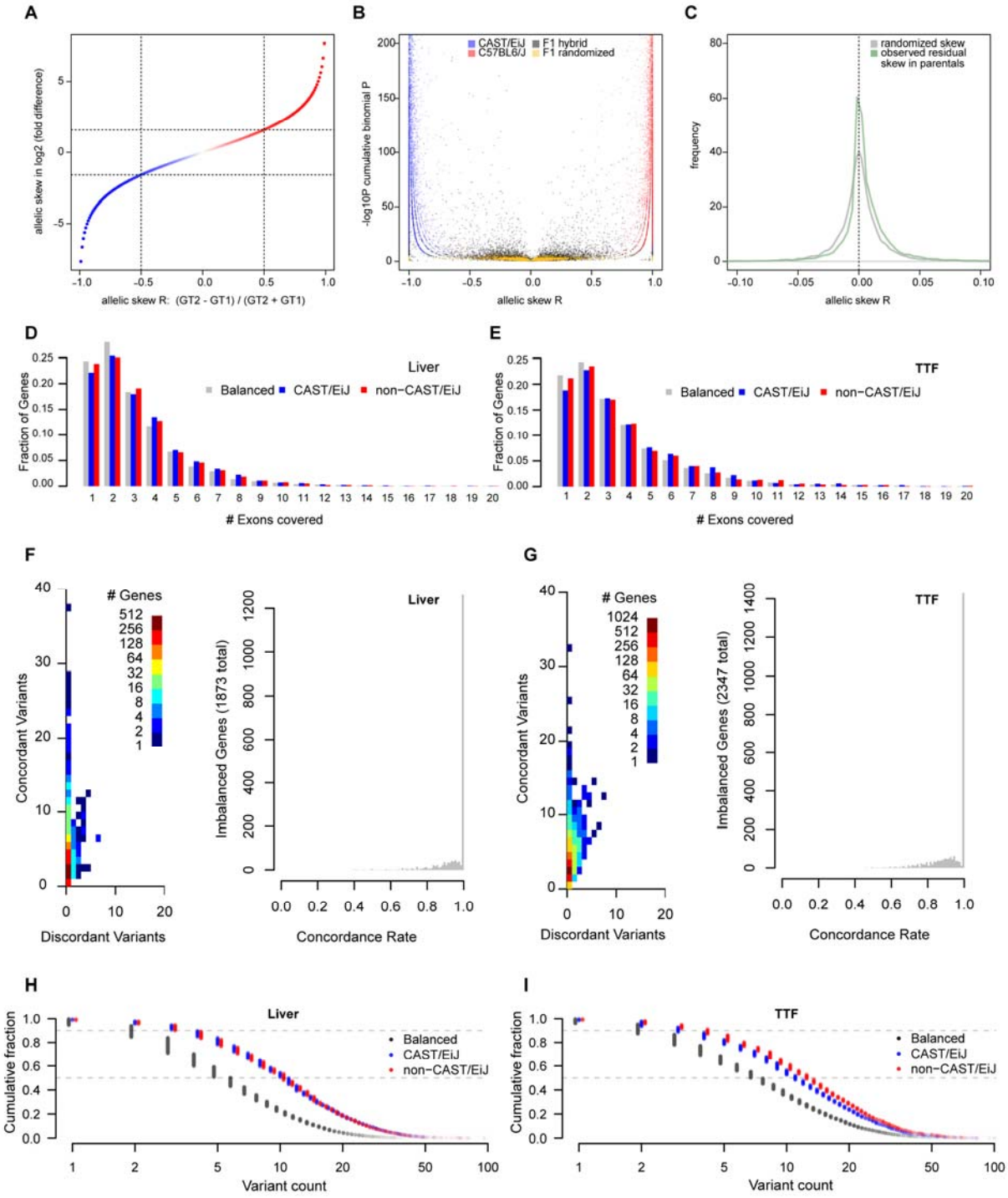

Figure S1 Distribution of Allelic Imbalance in Genes, Exons and Variants. (A) Conversion plot of log2 transformed fold difference between alleles (*CAST/EiJ*, blue, other in red) over allelic skew value R, ranging from -1 for 100% *CAST/EiJ* (blue) to 0% *CAST/EiJ* (red). Dashed lines indicate 3-fold difference (or R = +/- 0.5) between alleles. (B) Sample raw data from parental controls (*CAST/EiJ* in blue, *C57BL6/J* in red), a representative hybrid F1 sample (grey), and the same sample after shuffling exonic-SNP overlapping reads randomly between alleles (gold). Negative log10 transformed cumulative binomial P value (assuming $P_0 = 0.5$) is plotted over the degree of allelic skew (R). Parental samples fall expectedly on the sides with highly significant p-values compared to an actual hybrid F1 sample. Genes skewing to the non-represented allele in parental controls are excluded from further analysis (486 in liver, 1008 in TTF samples). Randomized F1 data fall expectedly flatter than actual F1 hybrid data, yet produce significant allelic skew ($p < 0.05$, cumulative binomial probability) at a certain rate, which is estimated from 100-fold permuted datasets using actual total read numbers for each sample and each gene. (C) Distribution of skewing in a randomized sample (grey, $P_0 = 0.5$) is compared to residual skewing observed in parental samples (green, estimated by calculating allelic skew in *CAST/EiJ* and *C57BL6/J* parents and summing the two). Ideally, for any given gene, the sum of parental allelic skew values should add up to 0, or some value within the distribution obtained by allelic read randomization. Residual skew values obtained from matched parental samples are used to correct the expected P_0 to

account for overdispersion. (D, E) Distribution of exons per gene covered as fraction of all genes in category (Balanced, *CAST/EiJ* and non-*CAST/EiJ*) in liver (D) and TTFs (E). In both tissues > 75% of genes contained at least 2 informative exons. (F, G) Left panels show bivariate distributions with variant counts per gene split into variant calls matching whole gene calls ("concordant", y-axis) and opposing variant calls ("discordant", x-axis). Number of genes encoded in log2–scaled color key. Genes scoring with 0 dis/concordant variants have sufficient variant coverage for a call in aggregate (gene-level) but no individual variant was covered sufficiently for a call (minimum of 5 reads). Right panels show the distribution of concordance rates (concordant calls / all variant calls). Variant calls of imbalanced genes strongly match gene calls in the vast majority of genes in both liver and TTF samples. (H, I) Cumulative fractions of covered variants per gene are shown for all genes in category (Balanced, *CAST/EiJ* and non-*CAST/EiJ*) in liver (H) and TTFs (I). Dashed grey lines indicate 90th percentile and medians. Covered variants include all variants with a minimum of 3 reads. In both liver and TTFs, >90% of imbalanced featured a minimum of 3 variants.

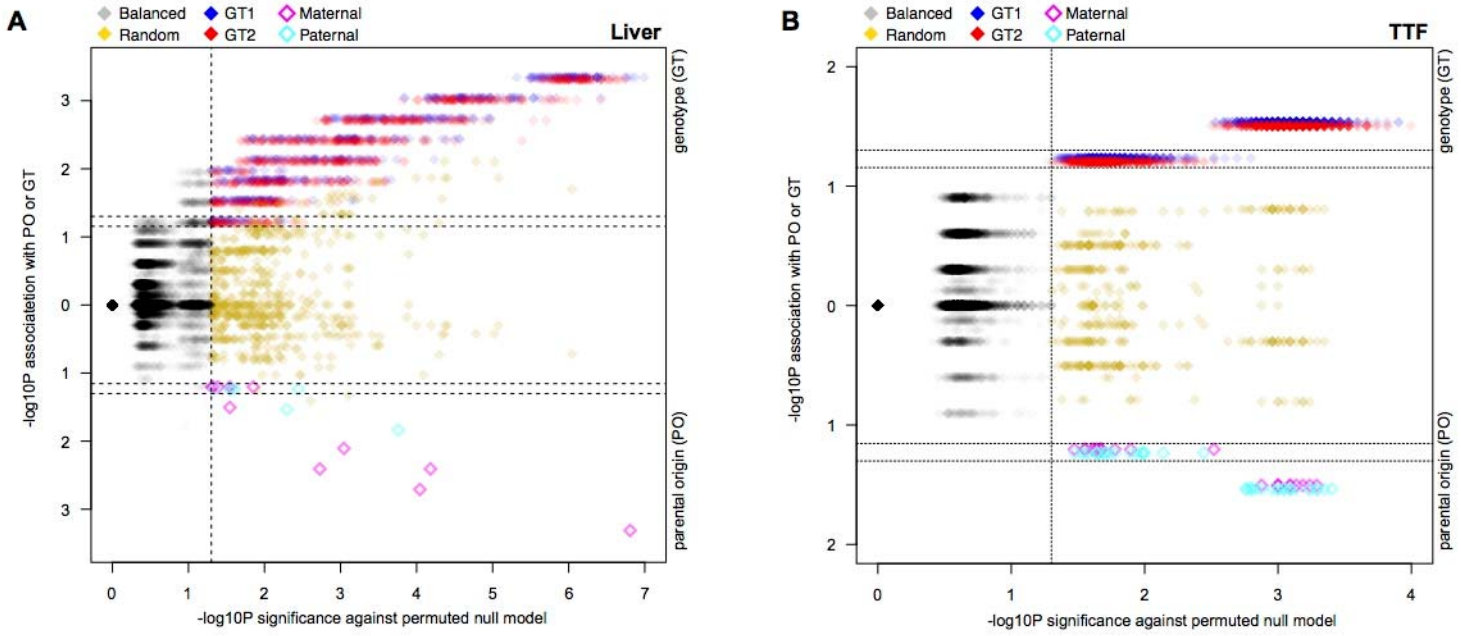

Figure S2 Classification of allelic imbalance. (A, B) Balanced (grey) and skewed genes with imprinted (open symbols, magenta and cyan for maternal and paternal, respectively), GTIE (blue GT1, red GT2) or random (gold) direction resolve by significance of GTIE or parental origin (y-axis) over significance against shuffled datasets (x-axis) in liver (A) and TTF samples (B). Points marginally offset on y-axis to facilitate viewing.

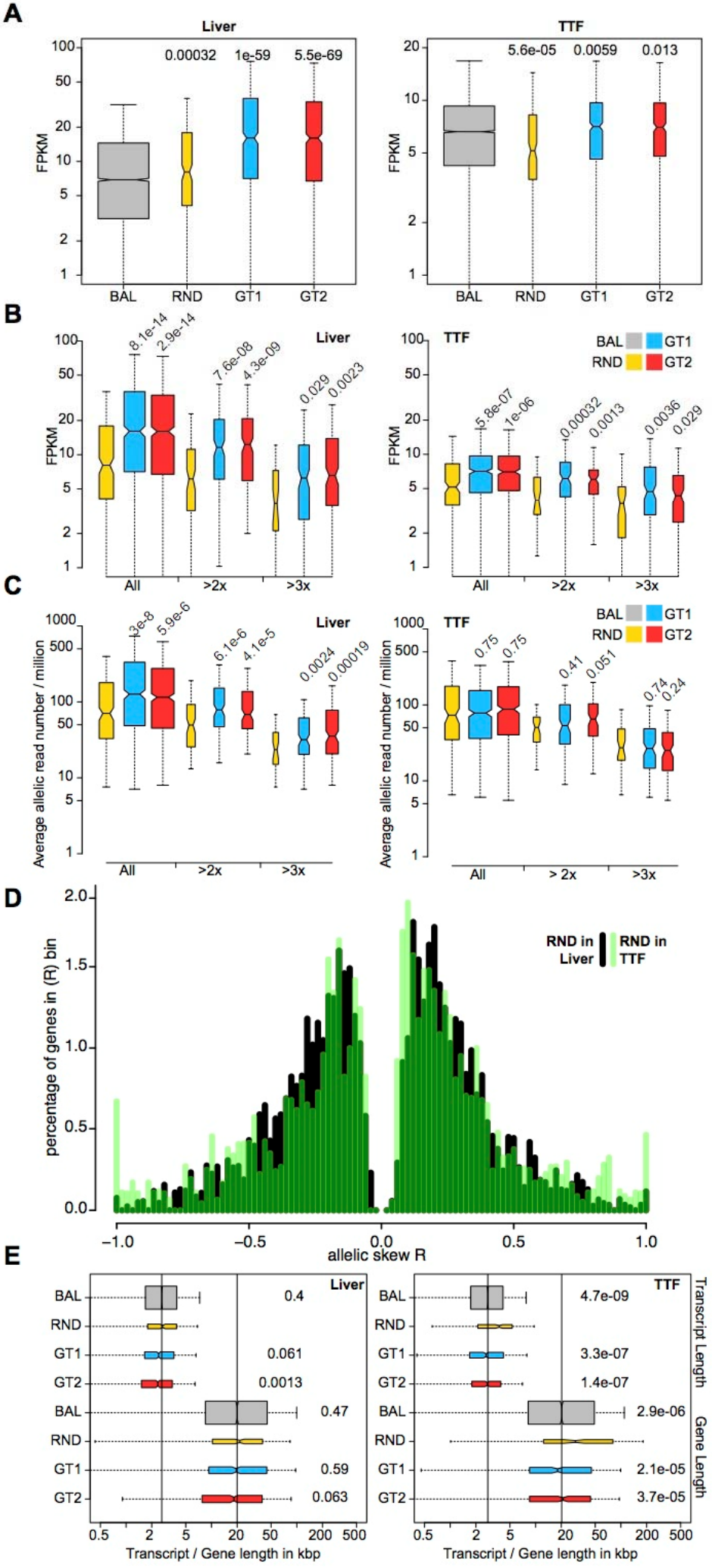
A
Liver
TTF
FPKM
0.00032
1e-59
5.5e-69
5.6e-05
0.0059
0.013
BAL
RND
GT1
GT2
B
8.1e-14
2.9e-14
7.6e-08
4.3e-09
0.029
0.0023
5.8e-07
1e-06
0.00032
0.0013
0.0036
0.029
All
>2x
>3x
C
Average allelic read number / million
3e-8
5.9e-6
6.1e-6
4.1e-5
0.0024
0.00019
0.75
0.75
0.41
0.051
0.74
0.24
> 2x
D
percentage of genes in (R) bin
RND in Liver
RND in TTF
allelic skew R
E
0.4
0.061
0.0013
0.47
0.59
0.063
4.7e-09
3.3e-07
1.4e-07
2.9e-06
2.1e-05
3.7e-05
Transcript Length
Gene Length
Transcript / Gene length in kbp

Figure S3 Skewed genes differ in expression and transcript / gene length from balanced genes. (A, B) FPKM for liver and TTF genes classified as balanced (grey), random (gold), or GTIE (blue, red for GT1, GT2 respectively). Shown are all genes in category (A) or split into all, >2-fold and > 3-fold difference between alleles (B). Significance of shift in median (Mann-Whitney U test) to balanced genes (A) or randomly skewed genes (B) is indicated above boxplots. Error bars denote 1.5-fold interquartile range. Genes with a greater fold-difference between alleles tend to have lower FPKMs although still significantly greater than balanced genes (except for randomly skewed genes in TTFs). Randomly skewed genes tend to be lower in expression than GTIE genes in both liver and TTF samples. (C) Randomly skewed genes have significantly lower read support (average number of allelic reads per million) than GTIE genes (Mann-Whitney U test, see p-values above boxplots) in primary liver but not clonal TTFs. (D) Randomly skewed genes in liver cells are depleted for monoallelic expression compared to randomly skewed genes from TTFs. Histogram showing percentage of genes binned by allelic skew R (black in liver, green in TTFs). (E) Transcript (top) and gene lengths (bottom) reveal that lower expression of randomly skewed genes in TTFs (though not in liver) is due to a significantly greater transcript length of random compared to GTIE genes (Mann-Whitney U-test, p-values indicated next to boxplots). This observation holds for gene length as well.

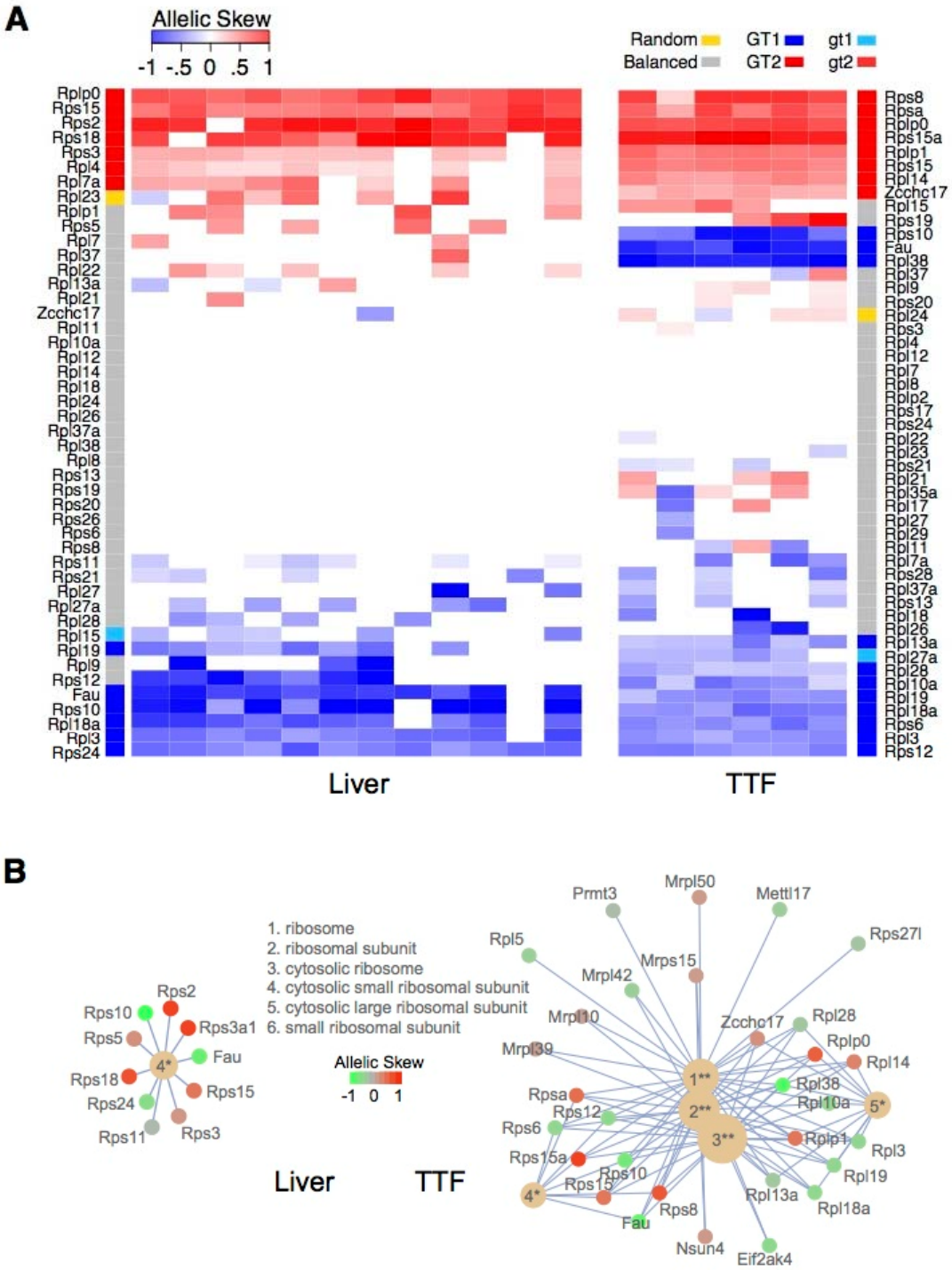

Figure S4 GTIE in ribosomal protein genes. (A) Allelic skew values of ribosomal protein genes in liver (left) and TTF (right) samples. Color-coded panel labels direction of skew, either GT1 (blues), GT2 (reds), randomly skewed (gold) or balanced expression (grey). Enrichment of ribosomal proteins is significant in both tissues (liver: $p < 5 \times 10^{-3}$, TTFs: $p < 4 \times 10^{-9}$, hypergeometric test). (B) Skewed ribosomal protein genes linked GO terms (cellular component ontology) for ribosomal proteins (* and **, adjusted $p < 0.05$ and $p < 0.01$, hypergeometric test) for GTIE genes in liver (left) and TTF samples (right). Median allelic skew for each gene is color-coded (-1, green to +1, red).

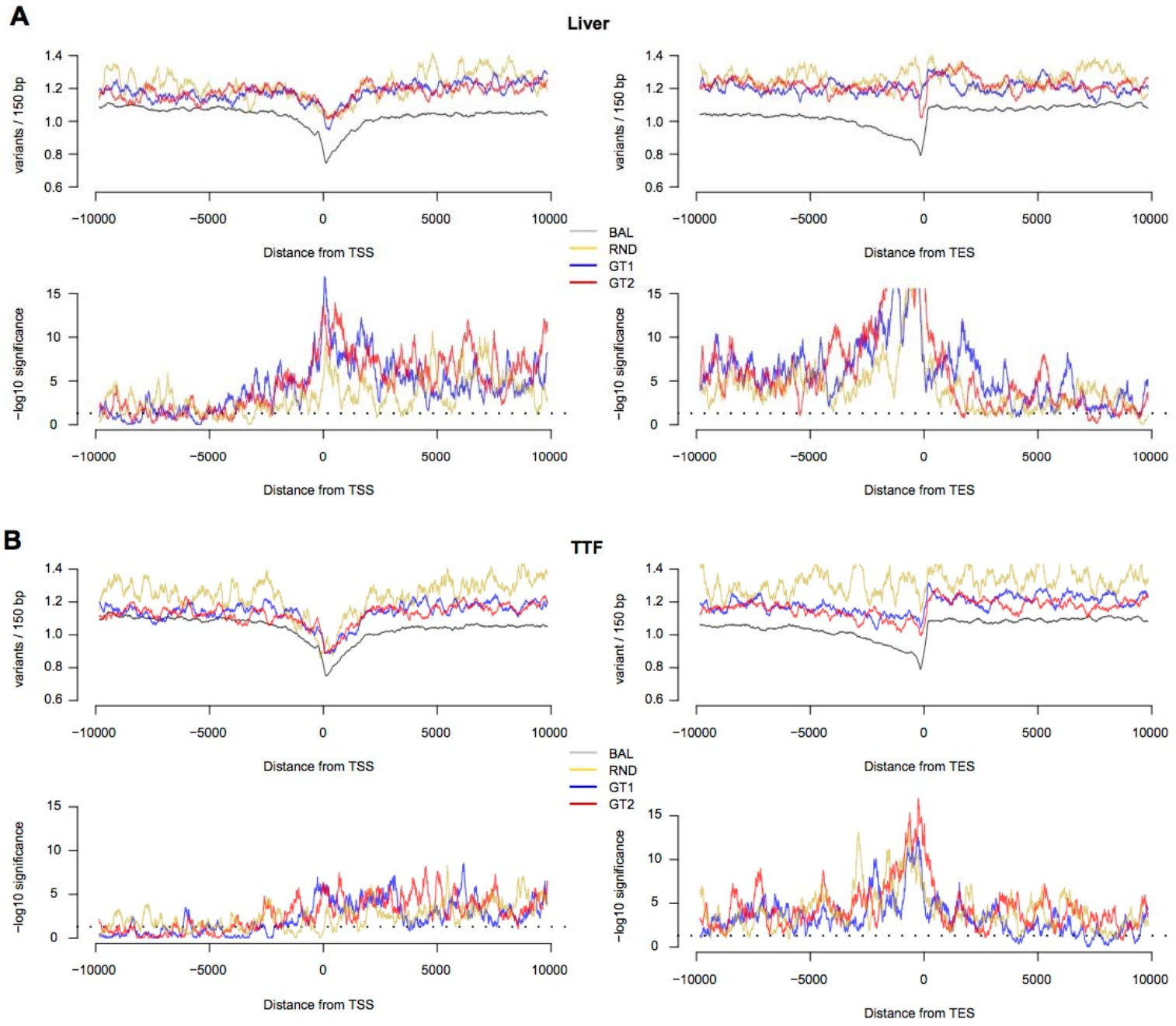

Figure S5 Allelic imbalance is associated with increased genetic divergence. (A, B) Averaged binned density of variants that distinguish parental genomes is plotted across +/- 10 kb of sequence centered on either the TSS (left) or TES (right). Variants across random (gold) and GTIE (blue, red) genes were binned at approximately nucleosome resolution (150 bp) and assessed for difference in medians compared to balanced genes (grey). Top panels (A/B, liver/TTFs) show variant density, bottom panels denote significance (with dotted lines representing $p < 0.05$, Mann-Whitney U test).

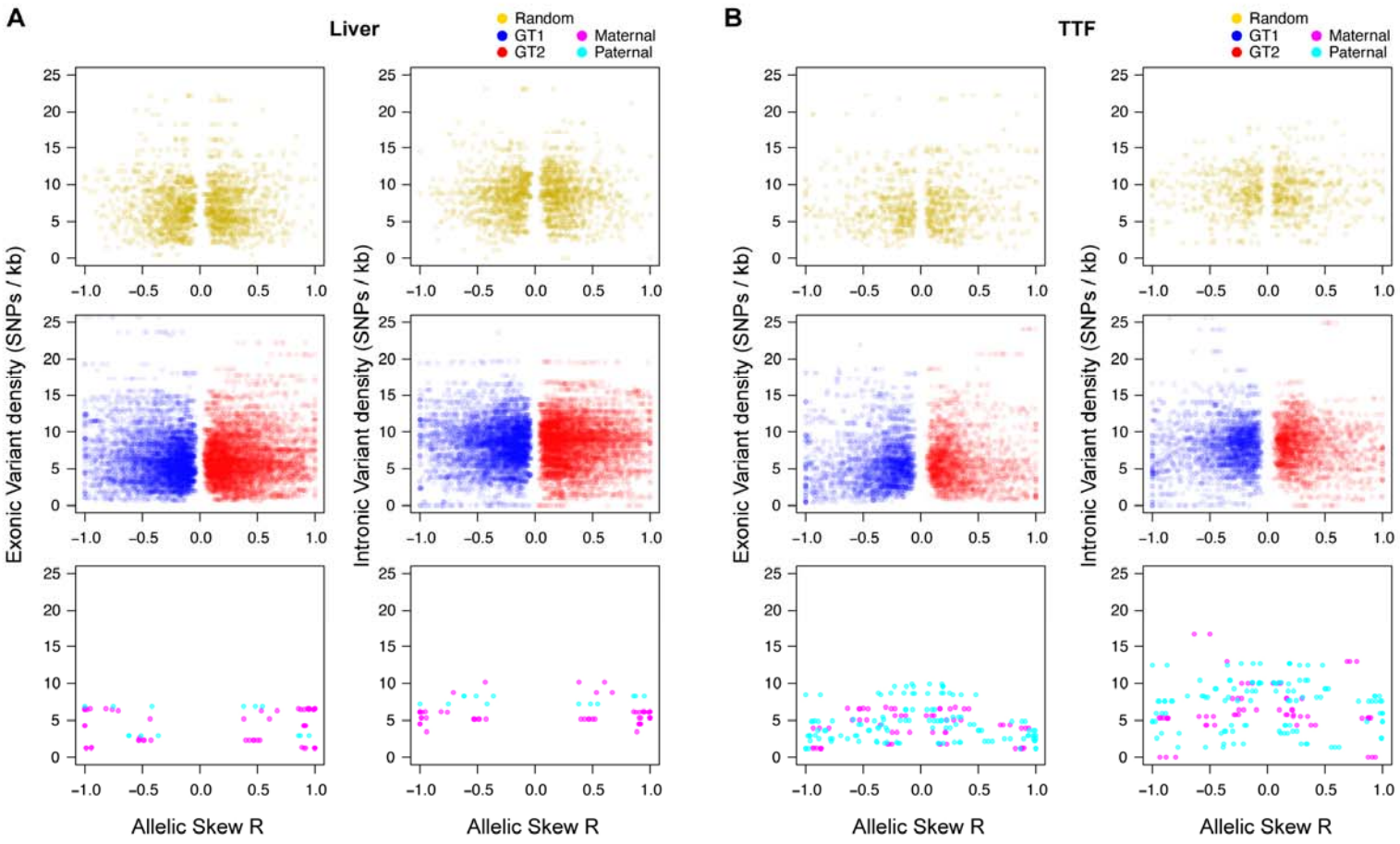

Figure S6 Magnitude of skewing is not predicted by overall exonic or intronic variant density. (A, B) Average exonic and intronic variant density per kb plotted above allelic skew R for liver (A) and TTF samples (B). There is no apparent correlation between variant density and allelic skewing for random (gold), GTIE (blue, red for GT1, GT2 respectively) or imprinted (magenta, cyan for maternal, paternal respectively) genes in either liver or TTF samples.

Table S1 RNA-seq datasets analyzed

Sample	Study	Tissue	Type	Genotype	Read pairs
ERR120684	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	32,075,256
ERR120686	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	30,192,493
ERR120702	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	33,152,941
ERR120704	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	29,979,442
ERR185942	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	30,373,020
ERR185943	GONCALVES *et al.* 2012	Liver	primary	*C57BL/6J*	32,907,241
ERR120692	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	33,511,533
ERR120694	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	32,950,724
ERR120698	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	29,951,804
ERR185946	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	27,222,930
ERR185947	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	29,485,779
ERR185948	GONCALVES *et al.* 2012	Liver	primary	*CAST/EiJ*	25,918,629
ERR120672	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	28,333,765
ERR120678	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	12,137,349
ERR120696	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	31,814,198
ERR120700	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	32,712,218
ERR185940	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	44,587,843
ERR185941	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	12,353,269
ERR185944	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	33,063,510
ERR185945	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	32,221,296
ERR185949	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	27,973,105
ERR185950	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	29,362,228
ERR185951	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	19,967,614
ERR185952	GONCALVES *et al.* 2012	Liver	primary	*F1 hybrid*	24,502,787
dSP27b	This study	TTF	immortal	*CAST/EiJ*	62,569,552
dSP27a	This study	TTF	immortal	*129S1*	85,645,055
dSP27c	This study	TTF	immortal	*129S1*	60,837,075
dSP23a	This study	TTF	clone 1E	*F1 hybrid*	54,285,349
dSP23b	This study	TTF	clone 4D	*F1 hybrid*	64,597,830
dSP23c	This study	TTF	clone 12E	*F1 hybrid*	87,823,179
dSP24a	This study	TTF	clone 3E	*F1 hybrid*	59,240,562
dSP24b	This study	TTF	clone 9G	*F1 hybrid*	61,779,313
dSP24c	This study	TTF	clone 11F	*F1 hybrid*	66,198,059

File S1

Supporting Materials and Methods

Mouse crosses and derivation of cell lines

To generate the hybrid cell lines, mice of *Mus castaneus* (*CAST/EiJ*) and *Mus musculus* (*129S1/SvImJ*) origins were crossed or obtained from (PAYER *et al.* 2013). Primary mouse embryonic fibroblasts (MEFs) or tail tip fibroblasts (TTFs) were prepared from F1 embryos collected at embryonic day 13.5 or from 2-5 day old pups, respectively. The sex and strain of each parent as well as the sex of the offspring from which cells were isolated were varied to obtain each of the possible combinations listed in Figure 1E. MEFs and TTFs were later immortalized by SV-40 T antigen (BROWN *et al.* 1986), from which the resulting heterogeneous population of cells was saved and used for analysis or further subcloned by limiting dilution to obtain independent clones.

Sequencing, alignment, and allele-specific transcript assembly

Total RNA from TTFs was harvested using the mirVana RNA extraction kit (Ambion), of which 500 ng was used for polyA selection of mRNA. 50 ng of the resulting mRNA was submitted for automated directional RNA-seq library generation on the Apollo 324 system (IntegenX) at the Biopolymers facility (Harvard Medical School) according to the manufacturer's instructions. 1 µg of genomic DNA was used for *M. musculus* parental genomic DNA library construction, and prepared in a similar automated fashion. Sequencing was performed on a HiSeq2000 instrument (Illumina) according to the manufacturer's instructions, yielding ~300 million paired-end 50 nt reads, approximating to ~10.7x coverage of the genome (mm9 build) after discarding PCR duplicates and multiply mapping reads (Novoalign V3.00.03). Unreferenced single-nucleotide polymorphisms (SNPs) in this parental genome were identified using samtools mpileup with extended base alignment quality calculation. High-quality SNPs were placed in the *C57BL/6J* reference genome while ambiguous calls were masked to 'N'. The two reconstructed parental genomes differ in 18,696,209 positions, of which 18,055,048 variants (96.6%) were obtained from the *CAST/EiJ* variant collection of the mouse genomes project (KEANE *et al.* 2011).

Parental and F1 hybrid RNA libraries were sequenced to obtain on average ~60 million paired-end 50 nt reads per sample (Table S1). Read pairs were aligned to both parental genomes allowing for a maximum of 3 mismatched or gapped bases per read (Tophat v2.0.8), followed by removal of PCR duplicates and multiply mapping reads. Read pairs considered allele-specific aligned a.) to only one parental genome, or b.) with both ends to one but only a single end to the other, or c.) with fewer nucleotide edits to one parental genome than the other (mismatches, deletions or insertions). Allele-specific pairs and pairs aligning equally well to both genomes were used for transcript assembly and quantitation (Cufflinks v2.1.1) using the Ensembl reference (release 66). For transcript assembly tRNA and rRNA genes were masked and fragment bias correction was applied, otherwise default parameters were used.

Directional, paired-end RNA-seq data generated from poly-A selected RNA in parental and hybrid F1 liver samples was acquired from a previous study examining *cis-trans* regulation of gene expression in the liver (GONCALVES *et al.* 2012) (ERP001401,

European Nucleotide Archive). Alignment and transcript assembly was performed as described above. One parental sample (ERS134271) was dropped from the analysis due to RNA contamination evident in a significant presence of the non-represented allele.

Future allele-specific RNA-seq studies may benefit from performing analyses of simulated RNA-seq data (BUSBY *et al.* 2013; GRANT *et al.* 2011) to properly estimate required sequencing depth *a priori*. Such an approach may be particularly helpful when the estimate can be based on a known gene with pre-determined level of minimum assessable expression (SIMS *et al.* 2014) and a small number of reporting variants.

Identification/classification of skewed genes

Exonic SNPs for each annotated transcript were queried for allele-specific counts and aggregated across the transcript as in (DEVEALE *et al.* 2012). For gene-level analyses the subset of top-scoring transcripts (by Cufflinks assembly isoscore) was chosen to represent Ensembl genes. To systematically identify genes preferentially expressed from one allele over the other as well as genes exhibiting balanced expression within a given sample, a number of qualifying criteria were applied to each gene. Transcripts lacking sufficient expression, assembly support and showing only spurious allelic coverage were excluded by requiring: a.) a non-zero FPKM, b.) at least 1 SNP covered by 3 allele-specific reads or more, and c.) a sufficient number of allele-specific reads across all exonic SNPs to achieve a power of 0.5 for rejecting the null hypothesis of $P_0 = 0.5$ at a significance level of $p < 0.05$ (binomial test) given a minimum fold-difference between alleles. For this analysis, a 3-fold difference was arbitrarily chosen as a cutoff, therefore requiring a minimum of 13 allele-specific reads overlapping the exonic SNPs of a given transcript. For each assessable gene, the skew (R) and the cumulative probability (p) under a binomial distribution were calculated, first under the assumption of equal probability for mapping to either allele ($P_0 = 0.5$). Allelic read counts are termed C_{REF} and C_{CAST} for non-*CAST/EiJ* and *CAST/EiJ* genomes respectively.

(1) $$R = \frac{C_{REF} - C_{CAST}}{C_{REF} + C_{CAST}}$$ (2) $$\mathrm{p}(X \leq C_{min}) = \sum_{i=0}^{C_{min}} \binom{C_{sum}}{i} {P_0}^i (1 - P_0)^{C_{sum} - i}$$

, where $C_{min} = min(C_{REF}, C_{CAST})$, and $C_{sum} = C_{REF} + C_{CAST}$

A number of genes (486 in liver samples, 1008 in TTFs, mostly pseudo –or predicted genes) skewing significantly in the direction of the non-represented allele in the parental controls were excluded from further analysis. To control for overdispersion frequently encountered in count-based biological data, the true null probability for each gene was then estimated empirically by aggregating the observed skew values of each gene in the parental controls, where sufficient parental data was available. Skew and cumulative binomial probability in the F1 hybrid samples were then calculated after taking into account the observed residual skew (R_{obs}) and adjusted null probability estimated from parental controls.

(3) $$R_{obs} = \overline{R}_{REF} + \overline{R}_{CAST}$$ (4) $$R_{F1} = R + R_{obs}$$ (5) $$F_{obs} = \frac{1 + R_{obs}}{1 - R_{obs}}$$

, where $R_{REF/CAST}$ are the parental skew values for a given gene averaged over all parental controls, R_{F1} is the adjusted skew value of the gene in the F1 hybrid sample, and F_{obs} is the observed residual skew R_{obs} across the two parents transformed to fold-difference. The gene-specific parentally adjusted null probability P_{CAST} is then expressed as the probability of observing fewer (or more) than the expected count of *CAST/EiJ* alleles, such that P_{CAST} (or conversely, $P_{REF} = 1 - P_{CAST}$) can now be substituted for P_0 in equation (2).

(6) $$P_{CAST} = (1 + F_{obs})^{-1}$$

To determine the significance of detecting significantly ($p < 0.05$, cumulative binomial probability) skewed genes reproducibly across a number of samples, each allele-specific read was first randomly redistributed between the two alleles for each dataset to build a null model based on actual sample sizes (number of allelic reads for each gene in each experiment). In this analysis, 100 permutations were found to be sufficient for adequate precision. The significance of observing a given gene skewing multiple times in the set of experiments was then estimated using the rate of observing the same gene skew *k* times in *m* permuted (*n = 100*) experiments.

(7) $$P_0 = 1 - \frac{k}{n \times m}$$ (8) $$\mathrm{p}(X \leq \mathrm{nm} - \mathrm{k}) = \sum_{i=0}^{nm-k} \binom{nm}{i} {P_0}^i (1 - P_0)^{nm-i}$$

Genes failing to reproduce a significant number ($p < 0.05$, binomial test) of skewed observations in either cross were classified as `balanced`. Genes skewing significantly more often across experiments than expected under the null model were then assessed in 2x2 contingency tables (forward/reverse cross vs. genotypes) to determine whether allelic skew associates with parental or genetic origin. In the latter case, switching the alleles in the reverse cross allowed application of the equivalent test as the 2x2 contingency table for parental origin. Contingency tables contained the number of skewed observations in each category, such that the sum of all squares equals the number of samples in which a skewed call was made and is therefore no greater than the total number of F1 hybrid experiments (12 in liver cells, 6 in TTFs). Genes associating significantly (two-tailed $p < 0.05$, Barnard's exact test) and unanimously with parental or genetic origin were classified as `imprinted` or showing `genotypically imbalanced expression` (GTIE), respectively. In view of the low number of samples in TTFs, genes with borderline significance in this test ($0.05 < p < 0.07$) were classified as `potentially` imprinted or GTIE. Genes lacking significant association with either parental or genetic origin (two-tailed $p > 0.07$, Barnard's exact test) yet skewing towards each allele at least once across experiments were classified as `randomly` skewed. A small number of genes lacking significant association due to being detected in only a small number of samples, yet skewing only towards one allele were classified as `undetermined` (38 in liver samples, 434 in TTFs).

To provide estimates for false discovery rates (FDRs) in each category (balanced, imprinted, random, GTIE) one randomly distributed dataset was carried through the classification (FDR < 0.02 for all of the categories described above in both liver and TTF samples).

Feature and enrichment analyses of imbalanced genes

Imprinted and randomly skewed genes were compared to collections of known imprinted (mousebook.org) and recently identified random monoallelic genes (ECKERSLEY-MASLIN *et al.* 2014; GENDREL *et al.* 2014; ZWEMER *et al.* 2012). Significance of enrichment in the corresponding categories was determined under a hypergeometric distribution, using the union of balanced, imprinted, random and GTIE genes as a background set. Equivalent testing was performed to assess the significance of overlap with specific Ensembl gene categories (biotypes), ribosomal protein genes as well as category identity between liver and TTF tissues.

To identify enrichment in functional annotation categories (GO, gene ontologies) genes from each of the skew categories were queried for GO terms using the clusterProfiler bioconductor package (YU *et al.* 2012) with the identical gene background set as above. GO-term enrichment p-values determined by hypergeometric test were adjusted for multiple hypothesis testing using the Benjamini & Hochberg method and only significant GO terms (adjusted $p < 0.05$) are shown.

For comparison of quantitative features (FPKM, allelic read support, gene and exon length, and density of sequence variation over exonic, intronic, upstream and downstream intervals) between categories (balanced, imprinted, random and GTIE) the Mann-Whitney U-test was used and significance values denoted as described in figure legends.

RNA-DNA FISH

Approximately 10^5 TTFs from *CAST/EiJ*, *129S1/SvImJ*, or hybrid mice were cytospun onto glass slides, rinsed with PBS, treated with ice cold CSKT buffer (100 mM NaCl, 300 mM sucrose, 10 mM PIPES, 3 mM $MgCl_2$, 0.5% Triton X-100, pH 6.8) for 10 min, and fixed in 4% paraformaldehyde for 10 min. Prior to hybridization, cells were dehydrated through a 70%, 80%, 90%, 100% ethanol series and allowed to air dry briefly. DNA probes were generated from fosmid or BAC clones (Children's Hospital Oakland Research Institute; *Apbb1ip*: WIBR1-0890H01, *Pon3*: WIBR1-1690D10, *Ywhag*: WIBR1-2316B24, *Spred2*: RP24-294M15, *Egfr*: RP23-263C13, *Grb10*: RP24-345M19) and labeled by nick translation using Cy3-conjugated dUTP. For detection, ~50 ng probe and ~1 ug mouse Cot-1 DNA in 10% dextran sulfate, 50% formamide, and 2x SSC (300 mM NaCl, 30 mM sodium citrate, pH 7.4) were denatured at 95°C for 10 min and allowed to pre-anneal at 42°C for 30 min. RNA FISH was performed first: probe mixture was applied to dehydrated slides and coverslips sealed with rubber cement and placed in a humid chamber for hybridization overnight at 42°C. The next day, slides were washed with 50% formamide, 2x SSC for 20 min, 2x SSC for 5 min, 2x SSC with Hoechst stain for 5 min, and with 2x SSC again for 5 min. Slides were mounted and images captured with positions recorded on a Nikon Eclipse 90i microscope workstation with Volocity software (Improvision). Slides were then rinsed in PBS with 0.2% Tween-20, dehydrated again, treated with RNase A (400 ug/mL in PBS) at 37°C for 40 min to remove RNA signals, and denatured for DNA FISH in 70% formamide, 2x SSC at 80°C for 10 min. Dehydration, hybridization, and washes were repeated as for RNA FISH with the exception of higher stringency 0.2x SSC in the final 3 washes. Slides were remounted and reimaged at recorded positions. Methods and probe design for the allele-specific DNA FISH experiments using Oligopaint probe sets will be described elsewhere (BELIVEAU *et al*. 2015).

Allele-specific qRT-PCR

RNA was isolated from cells using TRIzol Reagent (Invitrogen), from which cDNA was generated with SuperScript III reverse transcriptase (Invitrogen) and oligo(dT)15 primer (Promega). For qPCR, 20-µl reactions were run in technical duplicate on 96-well plates using 250 nM each of universal and either GT1 (*CAST/EiJ*)- or GT2 (*C57BL6/J* or *129S1/SvJm*)-specific primer and SYBR Green supermix (Bio-Rad). The PCR program consisted of 45 cycles of: 95°C, 15 sec; 60°C, 30 sec; 72°C, 30 sec. Primers targeting SNPs were designed with at least 2 nt differences in the 3' terminal 4 nt or according to the method of TaqMAMA (GLAAB and SKOPEK 1999). Specifically, we aimed to target exonic regions containing ≥2-nt difference between alleles within a 4-nt span. For genes in which this does not occur or does not easily lend itself to primer design, single mismatches were intentionally placed at the nucleotide directly 5' to individual SNPs to achieve higher sensitivity. The identity of the mismatched base was chosen to optimize allelic discrimination, as previously described (LI *et al.* 2004). Expression levels were compared against that of the opposite allele using the formula: Fold Difference = 2^Ct(GT2-GT1). The data were corrected for primer bias and differential primer efficiencies by comparison to amplification of pure parental or hybrid genomic DNA. Six independent clones (3 males, 3 females) for the forward and reverse crosses were tested. Analyses were also conducted on primary and transformed populations for the forward and reverse crosses to rule out clonal effects and transformation artifacts. Each plate was tested twice (technical replicates), and a second pair of allele-specific primers used to validate select genes in order to rule out primer-specific artifacts. See File S10 for primer information.

File S2

List of genes showing allelic imbalance in liver

Coordinates (chromosome, start, end, strand), FPKM, Call (GT1, gt1, GT2, gt2, MAT, mat, PAT, pat, RND), and skew (R) values in the forward (f) and reverse (r) crosses listed accordingly.

File S3

List of genes showing allelic imbalance in TTFs

Coordinates (chromosome, start, end, strand), FPKM, Call (GT1, gt1, GT2, gt2, MAT, mat, PAT, pat, RND), and skew (R) values in the forward (f) and reverse (r) crosses listed accordingly.

File S4

List of GO categories "Biological Process" enriched in genes with allelic imbalance in liver

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S5

List of GO categories "Cellular Component" enriched in genes with allelic imbalance in liver

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S6

List of GO categories "Molecular Function" enriched in genes with allelic imbalance in liver

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S7

List of GO categories "Biological Process" enriched in genes with allelic imbalance in TTFs

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S8

List of GO categories "Cellular Component" enriched in genes with allelic imbalance in TTFs

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S9

List of GO categories "Molecular Function" enriched in genes with allelic imbalance in TTFs

Cluster (GTIE, GT1, GT2, IMP, MAT, PAT, RND), ID of GO category (ID), Description of GO category (Description), Ratio of genes in cluster matching GO category over all genes in cluster (GeneRatio), Ratio of all genes scored matching GO category over all genes scored (BgRatio), P value (pvalue), P value adjusted for multiple comparisons (p.adjust), False discovery rate (qvalue).

File S10

Allele-specific qPCR primer information

Primer sequences are listed alongside the genetic variants used to discriminate between alleles. Each primer's ability to distinguish between alleles was assessed (see "primer efficiency") using both *CAST/EiJ* (*cas*) and *129S1/SvImJ* (*mus*) genomic DNA as amplification template. Any amplification bias toward one allele or the other was assessed (see "primer bias") using F1 hybrid genomic DNA as template.

File S11

Supporting Literature Cited

BELIVEAU, B. J., A. N. BOETTIGER, R. JUNGMANN, M. S. AVENDAÑO, R. B. MCCOLE, E. F. JOYCE, C. KIM-KISELAK, F. BANTIGNIES, C. Y. FONSEKA, J. ERCEG, M. A. HANNAN, H. G. HOANG, D. COLOGNORI, J. T. LEE, W. M. SHIH, P. YIN, X. ZHUANG, C.-T. WU, Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. Nat. Commun. DOI: 10.1038/ncomms8147.

BROWN, M., M. MCCORMACK, K. G. ZINN, M. P. FARRELL, I. BIKEL *et al.*, 1986 A recombinant murine retrovirus for simian virus 40 large T cDNA transforms mouse fibroblasts to anchorage-independent growth. Journal of virology **60:** 290-293.

BUSBY, M. A., C. STEWART, C. A. MILLER, K. R. GRZEDA and G. T. MARTH, 2013 Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics **29:** 656-657.

DEVEALE, B., D. VAN DER KOOY and T. BABAK, 2012 Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS genetics **8:** e1002600.

ECKERSLEY-MASLIN, M. A., D. THYBERT, J. H. BERGMANN, J. C. MARIONI, P. FLICEK *et al.*, 2014 Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation. Developmental cell **28:** 351-365.

GENDREL, A.-V., M. ATTIA, C.-J. CHEN, P. DIABANGOUAYA, N. SERVANT *et al.*, 2014 Developmental dynamics and disease potential of random monoallelic gene expression. Developmental cell **28:** 366-380.

GLAAB, W. E., and T. R. SKOPEK, 1999 A novel assay for allelic discrimination that combines the fluorogenic 5' nuclease polymerase chain reaction (TaqMan) and mismatch amplification mutation assay. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis **430:** 1-12.

GONCALVES, A., S. LEIGH-BROWN, D. THYBERT, K. STEFFLOVA, E. TURRO *et al.*, 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome research **22:** 2376-2384.

GRANT, G. R., M. H. FARKAS, A. D. PIZARRO, N. F. LAHENS, J. SCHUG *et al.*, 2011 Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics **27:** 2518-2528.

KEANE, T. M., L. GOODSTADT, P. DANECEK, M. A. WHITE, K. WONG *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature **477:** 289-294.

LI, B., I. KADURA, D. J. FU and D. E. WATSON, 2004 Genotyping with TaqMAMA. Genomics **83:** 311-320.

PAYER, B., M. ROSENBERG, M. YAMAJI, Y. YABUTA, M. KOYANAGI-AOI *et al.*, 2013 Tsix RNA and the germline factor, PRDM14, link X reactivation and stem cell reprogramming. Molecular cell **52:** 805-818.

SIMS, D., I. SUDBERY, N. E. ILOTT, A. HEGER and C. P. PONTING, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet **15:** 121-132.

YU, G., L.-G. WANG, Y. HAN and Q.-Y. HE, 2012 clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters, pp. 284-287 in *OMICS: A Journal of Integrative Biology*.

ZWEMER, L. M., A. ZAK, B. R. THOMPSON, A. KIRBY, M. J. DALY *et al.*, 2012 Autosomal monoallelic expression in the mouse. Genome biology **13:** R10.