Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance

Jingfeng Li,[1] Keiko Akagi,[1] Yongjun Hu,[2] Anna L. Trivett,[3] Christopher J.W. Hlynialuk,[1] Deborah A. Swing,[4] Natalia Volfovsky,[5] Tamara C. Morgan,[6] Yelena Golubeva,[6] Robert M. Stephens,[5] David E. Smith,[2] and David E. Symer[1,7,8]

[1]*Human Cancer Genetics Program and Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio 43210, USA;* [2]*Department of Pharmaceutical Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA;* [3]*Laboratory of Molecular Immunoregulation and* [4]*Mouse Cancer Genetics Program, National Cancer Institute, Frederick, Maryland 21702, USA;* [5]*Advanced Biomedical Computing Center, Information Systems Program and* [6]*Histotechnology Laboratory, SAIC-Frederick, Inc., National Cancer Institute, Frederick, Maryland 21702, USA;* [7]*Department of Internal Medicine and Department of Biomedical Informatics, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio 43210, USA*

Endogenous retrotransposons have caused extensive genomic variation within mammalian species, but the functional implications of such mobilization are mostly unknown. We mapped thousands of endogenous retrovirus (ERV) germline integrants in highly divergent, previously unsequenced mouse lineages, facilitating a comparison of gene expression in the presence or absence of local insertions. Polymorphic ERVs occur relatively infrequently in gene introns and are particularly depleted from genes involved in embryogenesis or that are highly expressed in embryonic stem cells. Their genomic distribution implies ongoing negative selection due to deleterious effects on gene expression and function. A polymorphic, intronic ERV at *Slc15a2* triggers up to 49-fold increases in premature transcriptional termination and up to 39-fold reductions in full-length transcripts in adult mouse tissues, thereby disrupting protein expression and functional activity. Prematurely truncated transcripts also occur at *Polr1a*, *Spon1*, and up to ~5% of other genes when intronic ERV polymorphisms are present. Analysis of expression quantitative trait loci (eQTLs) in recombinant BxD mouse strains demonstrated very strong genetic associations between the polymorphic ERV in *cis* and disrupted transcript levels. Premature polyadenylation is triggered at genomic distances up to >12.5 kb upstream of the ERV, both in *cis* and between alleles. The parent of origin of the ERV is associated with variable expression of nonterminated transcripts and differential DNA methylation at its 5′-long terminal repeat. This study defines an unexpectedly strong functional impact of ERVs in disrupting gene transcription at a distance and demonstrates that ongoing retrotransposition can contribute significantly to natural phenotypic diversity.

[Supplemental material is available for this article.]

The laboratory mouse is the premier model organism, facilitating comparative studies of human diseases, development, and natural variation. Numerous distinct mouse lineages manifest phenotypic differences such as various coat colors and differential susceptibilities to diseases (Wade and Daly 2005). The molecular basis for natural phenotypic variation or allele-specific expression differences remains unclear in most cases, although recent studies have associated differential gene expression with various forms of structural variation in mouse and human genomes (Yan et al. 2002; Adams et al. 2005; Yang et al. 2007; She et al. 2008; Cahan et al. 2009; Schlattl et al. 2011; Yalcin et al. 2011).

Transposons are a potentially major but relatively unexamined determinant of such allele-specific, transcriptional variation. They are strong candidates for regulating or disrupting gene expression since they comprise almost half of the mouse genome and certain elements are still actively mobilized. Approximately 10% of naturally occurring mutations in the mouse have been attributed to insertional mutagenesis of coding sequences due to endogenous retrotransposition (Waterston et al. 2002). We recently showed that thousands of polymorphic retrotransposon integrants of various active classes are present in the C57BL/6J (hereafter abbreviated as B6) reference genome but absent from one or more other classical inbred mouse lineages (i.e., A/J; DBA/2J, DBA; 129S1/SvImJ, 129S1; and 129X1/SvJ, 129X1) (Akagi et al. 2008). Of these, new endogenous retrovirus (ERV)-K integrants may be particularly capable of altering transcriptomes in diverse tissues (van de Lagemaat et al. 2003; Medstrand et al. 2005). Members of various ERV families make up ~10% of the mouse genome. While most such genomic elements are ancient and are comprised of solo long terminal repeats (LTRs), the ERV-K family includes a significant number of young full-length elements flanked by virtually identical LTRs. Of these, intracisternal A-particle (IAP) retrotransposons are still capable of autonomous mobilization and are transcriptionally activated by genome-wide cytosine demethylation (Walsh et al. 1998), contributing to tumor formation (Howard et al. 2007). Approximately 1000 of these elements contain intact open reading frames (ORFs). They have long been known to be active and polymorphic in various mouse lineages and tumors (Shen-Ong and Cole 1982; Lueders et al. 1993; Zhang et al. 2008). To simplify nomenclature, we refer to these IAP retrotransposons as ERVs.

[8]Corresponding author.
E-mail david.symer@osumc.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.130740.111.

22:870–884 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

Well-characterized retrotransposon integrants that alter gene expression and mediate phenotypic variability include the *dilute* and *hairless* coat color mutations (Copeland et al. 1983; Stoye et al. 1988). In these cases, intronic murine leukemia virus (MLV) insertions cause aberrant splicing of overlapping gene transcripts. MLV sequences are incorporated directly at the 3′ ends of disrupted transcripts, which are then prematurely terminated (Seperack et al. 1995; Cachon-Gonzalez et al. 1999). In contrast, ERV (IAP) integrants upstream of or within the *A* (i.e., agouti) and *Axin1* (i.e., axin 1) genes inserted active, heterologous promoters in the resulting agouti viable yellow (A^{vy}) and axin fused ($Axin1^{Fu}$) alleles, resulting in epigenetically regulated, variable initiation of downstream fusion transcripts (Morgan et al. 1999; Whitelaw and Martin 2001). Full-length ERV integrants also can affect neighboring gene transcription by direct incorporation of polyadenylation signal sequences and/or binding sites for transcription factors (van de Lagemaat et al. 2003; Medstrand et al. 2005), as was observed recently for *Adamts13* (i.e., a disintegrin-like and metallopeptidase [reprolysin type] with thrombospondin type 1 motif, 13), a von Willebrand factor-cleaving protease disrupted by an intronic ERV integrant (Banno et al. 2004; Zhou et al. 2007). However, until now, other effects by ERV polymorphisms have not been characterized. To associate variable gene expression levels with the presence or absence of local ERV insertions, we mapped ERVs in several highly divergent mouse lineages. We then identified and characterized numerous cases in which gene transcripts are strongly and differentially disrupted at a distance by nearby ERV polymorphisms.

Results

Identification of polymorphic ERVs in diverse mouse strains by transposon junction assay

To study possible effects of ERV integrants on neighboring gene expression levels, first we mapped such integrants in previously unsequenced, diverse mouse strains, without prior knowledge of their location or polymorphism status. Recently, various methods to find ERV insertions, both polymorphic and nonpolymorphic, have been described (Horie et al. 2007; Akagi et al. 2008; Takabatake et al. 2008; Zhang et al. 2008; Qin et al. 2010; Ray et al. 2011). We developed and optimized a sensitive, high-resolution genomic mapping assay using PCR and 454 Life Sciences (Roche) sequencing, which we call the transposon junction assay (Supplemental Fig. 1; Pornthanakasem and Mutirangura 2004; Iskow et al. 2010; Witherspoon et al. 2010). Using transposon sequence-specific and degenerate primers for genomic PCR amplification, we targeted members of certain young ERV families including IAPLTR1, IAPLTR2, and IAPEY2 elements (Kapitonov and Jurka 2008) since they are anticipated to be polymorphic (Qin et al. 2010) in diverse mouse lineages. Based on their features observed in the reference B6 genome, the IAPLTR1 integrants are most likely to be full length and to have identical LTRs (data not shown), consistent with their status as the youngest ERV insertions (Qin et al. 2010).

We optimized the transposon junction assay using various combinations of restriction enzymes and degenerate primers (Supplemental Table 1). This method reidentified 1538 out of 1665 (92.4%) of the youngest mappable ERV (IAPLTR1) integrants in the reference B6 genome at ~14-fold sequencing coverage. To validate these results, we compared chromosomal distributions of identified ERVs with previously annotated reference elements (Supplemental Fig. 1). Overall, the correlation between local densities of reference versus resequenced transposon integrants is excellent, particularly for IAPLTR1 elements ($p < 2.2 \times 10^{-16}$). Pearson's correlation coefficients are 0.75 (IAPLTR1), 0.78 (IAPLTR2), and 0.60 (IAPEY2). Thus no significant global bias was detected in identifying ERVK elements by targeted resequencing.

We used this assay to define the genomic locations of young ERVs in six diverse mouse lineages, i.e., A/J, B6, CAST/EiJ (CAST), MOLF/EiJ (MOLF), SPRET/EiJ (SPRET), and WSB/EiJ (WSB); 25,069 integrants were identified (Supplemental Table 1). The chromosomal integration sites of IAPLTR1 integrants are almost entirely different in comparing the highly divergent, wild mouse lineages (SPRET, CAST, and MOLF), because only four out of several thousand IAPLTR1 integrants are present at orthologous loci (Fig. 1). This result strongly suggests that the rare, shared integrants are

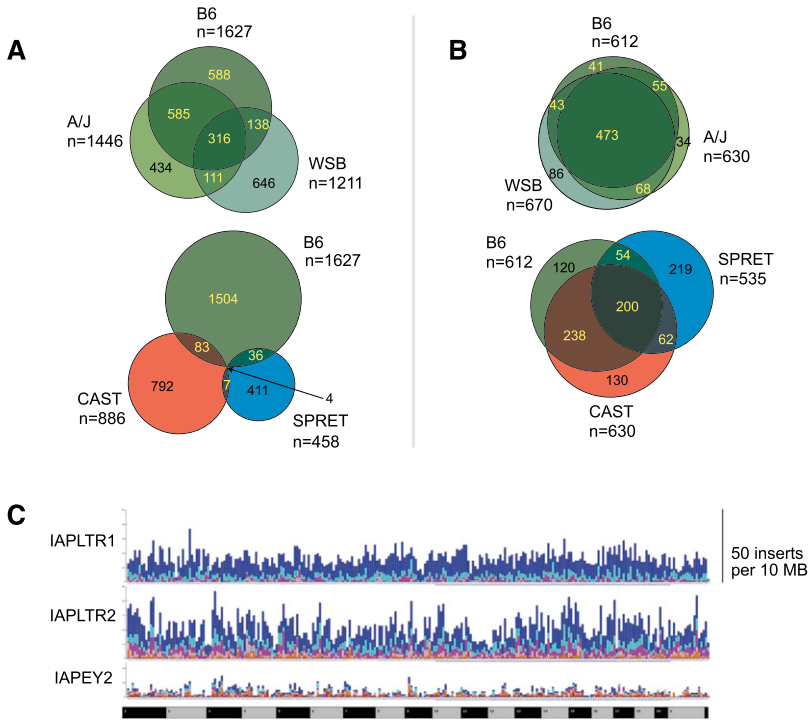

Figure 1. Genomic variation due to ERVs in diverse mouse strains. (*A*,*B*) Venn diagrams indicating counts (*n*) of shared versus distinct ERV elements at individual integration sites (orthologous locations) in previously unsequenced mouse strains. The youngest IAPLTR1 (*A*) and older IAPEY2 (*B*) elements were compared at genomic insertion sites in related B6, A/J, and WSB (*top*) and in divergent B6, CAST, and SPRET (*bottom*) mouse strains. MOLF integrants are not presented here. Only four of several thousand youngest IAP integrants occur at orthologous loci in the most divergent strains (*lower left*). (*C*) Genome-wide distributions of ERV polymorphisms in diverse mouse lineages. Histograms display the numbers of strains containing polymorphic ERVs at orthologous loci within 10-MB genomic bins for (*top*) IAPLTR1; (*middle*) IAPLTR2; and (*bottom*) IAPEY2 elements. (Dark blue) Integrants present in one strain; (light blue) two strains; (purple) three; (light pink) four; (orange) five; (red) all six. (*Bottom*) Mouse chromosomes 1–19, X, and Y (alternating shading).

identical by descent (Salem et al. 2005; Ray et al. 2011) and reveals extensive, lineage-specific retrotransposition. As expected, they are conserved at higher frequencies at orthologous loci in the related classical lines (B6, A/J, and WSB). Additionally, our bioinformatics analysis identified thousands of additional, previously unreported ERV polymorphisms that are present in one or more of the nonreference "Celera strains," i.e., A/J, DBA, 129S1, and 129X1 mice (Akagi et al. 2008) but absent from the B6 reference genome. Very similar proportions of polymorphic retrotransposon families including ERVs are present or absent in these strains, regardless of the genome chosen for comparison (Supplemental Fig. 2).

To compare the chromosomal distributions of polymorphic versus nonpolymorphic ERVs in various lineages, we plotted the number of elements counted in 10-Mb genomic intervals in single versus additional strains (Fig. 1C). IAPLTR1 integrants are mostly present in only one of the six diverse strains studied here; they are highly polymorphic. When their counts are summed up across the different strains, the polymorphic integrants are quite uniformly distributed across the genome, without large hotspot or desert regions in the chromosomes. In contrast, older IAPLTR2 and particularly IAPEY2 integrants tend to be more nonpolymorphic in multiple diverse strains. Both reference and polymorphic ERV integrants are more uniformly distributed across the genome than are reference and polymorphic L1 retrotransposons (Supplemental Fig. 2; Akagi et al. 2008).

We validated a collection of ERV integrants identified here, by amplifying occupied or empty genomic target sites in up to 21 diverse mouse lineages using PCR (Supplemental Fig. 1; Supplemental Tables 2, 3). The results demonstrate that both the transposon junction assay and our analysis of Celera sequence traces (Akagi et al. 2008) accurately determine the presence of integrants, and confirm that ERV insertions are extremely polymorphic (Zhang et al. 2008). Recent whole-genome shotgun (WGS) sequencing of 17 mouse strains has facilitated identification of thousands more retrotransposon polymorphisms (Keane et al. 2011; Yalcin et al. 2011). We compared ERVs mapped by the transposon junction assay, PCR validation, and WGS predictions. Out of 140 verifiable integrants called by the transposon junction assay, 135 (>96%) were validated both by PCR and by WGS sequencing (Supplemental Table 2). In four of the five discrepant cases, our transposon junction assay and confirming PCR indicated empty target sites, but WGS demonstrated the presence of ERV integrants. These cases indicate that the genomic DNA samples used in our assays versus WGS sequencing are likely to include bona fide sequence differences of unknown cause.

Genomic distribution of polymorphic ERVs suggests deleterious impact on gene expression

We assessed the genomic locations of young ERV polymorphisms relative to annotated genes, since their existing distribution would reflect insertion preferences and/or losses of deleterious elements. All classes of young ERV polymorphisms occur in gene introns at lower densities than expected from a simulated pattern of insertions due solely to chance (Fig. 2; Table 1). They are even more strongly depleted from particular genes involved in embryogenesis and/or highly expressed in embryonic stem (ES) cells (Fig. 2; Mikkelsen et al. 2007). The oldest ERVs mapped here, IAPEY2 elements, occur at even lower densities in intragenic locations than younger IAPs such as LTR1 and LTR2. Of the ERVs within genes, ~72%–83% are oriented antiparallel to the sense (coding) strand of the genes, rather than the 50/50 orientation frequency expected if

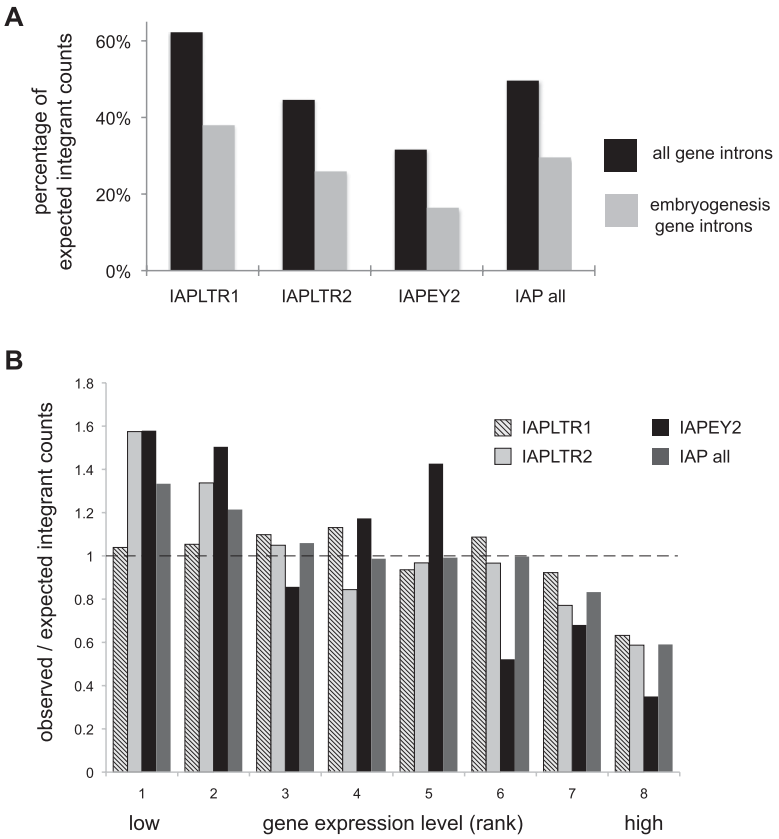

Figure 2. Young ERVs are excluded from introns, particularly from embryogenesis and highly expressed genes. (*A*) "Observed" ERV integrant counts are plotted as percentages of "expected" counts within all gene introns (black histograms) or embryogenesis genes (gray). Genomic locations of various classes of ERVs (*x*-axis) were identified in diverse mouse lineages. Expected counts were determined by random simulation of 2 million insertion sites across the reference genome. By chance, ~35% of ERV insertions would be expected to fall within RefSeq gene introns, and ~2.7% of all insertions would fall within embryogenesis genes, defined by the Mouse Genome Informatics database (http://www.informatics.jax.org). This normalization corrects for gene lengths. Percentages <100% signify relative exclusion of certain ERV subtypes from particular gene categories. (*B*) Based on their expression levels in mouse ES cells measured by microarrays (Mikkelsen et al. 2007), genes were binned into eight groups ranked from 1 (lowest expression) to 8 (highest), each with roughly equivalent numbers of genes expressed at comparable levels. Ratios of the observed numbers of genes containing intronic ERV integrants versus the expected number of genes identified by random simulation are presented (Brady et al. 2009) for different classes of ERV integrants (key, *upper right*). (Dashed line) Ratio = 1 signifies equivalence between observed and expected counts; ratios < 1 signify relative exclusion of ERV integrants from particular groups of genes.

Table 1. Young ERVs occur at low densities in gene introns

	Number of elements	Number of intronic elements	% intronic	Number of sense oriented, intronic	% sense oriented, intronic
Reference ERV class					
IAPLTR1	4410	961	21.8%	256	26.6%
IAPLTR2	5772	901	15.6%	235	26.1%
IAPEY2	1112	123	11.1%	21	17.1%
All ref IAP	11,661	2026	17.4%	518	25.6%
Simulation	2,000,000	700,262	35.0%	349,896	50.0%
Nonreference ERV class					
IAPLTR1	2789	622	22.3%	164	26.4%
IAPLTR2	3494	585	16.7%	161	27.5%
IAPEY2	501	58	11.6%	15	25.9%
All non-ref IAP	7002	1291	18.4%	342	26.5%
Simulation	2,000,000	700,262	35.0%	349,896	50.0%

Summary of ERV elements identified in (*top*) B6 reference genome and (*bottom*) from combined analysis by transposon junction assay and bioinformatics analysis of A/J, DBA, 129S1, 129X1, WSB, SPRET, CAST, and MOLF mouse lineages. (Simulation) As a control, we simulated ERV insertions randomly throughout the reference genome. The results show strong biases against intronic insertions, particularly of the older IAPEY2 elements, and against elements positioned in the same orientation as gene open reading frames (i.e., sense orientation).

integration and retention of ERVs were due to chance (Table 1). This orientation bias has been observed previously (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2006; Zhang et al. 2008). It plausibly could reflect patterns of de novo integration, but new ERV integrants in mouse, human, and chicken do not display such orientation bias (Dewannieux et al. 2004; Barr et al. 2005; Brady et al. 2009). The relative lack both of genomic ERV integrants within transcribed genes and of sense-oriented intragenic integrants presumably reflects strong ongoing purifying selection against them, due to their putative deleterious consequences on gene expression, structure, and function. Genes that lack such ERV integrants across all lineages are likely to be functionally essential in early development and viability of the organism. Alternatively, current patterns of insertions could reflect de novo integration preferences, but this possibility is refuted by the finding that older ERV integrants occur at even lower densities in gene introns. Together, these results strongly suggest that ERV integrants can exert deleterious effects on gene expression and function, and therefore the remaining extant integrants are relatively absent from gene introns.

An intronic ERV in *Slc15a2* disrupts transcription, protein expression, and function

The identification of thousands of polymorphic, intronic ERVs in diverse mouse strains provided a unique opportunity to assess their roles in transcriptional and functional variation, by comparing gene expression and functions in strains with and without such individual integrants. We first screened a collection of genes containing intronic ERVs by using reverse transcriptase–mediated polymerase chain reaction (RT-PCR) to identify fusion transcripts initiated from upstream polymorphic ERV promoters (Wheelan et al. 2005). We identified spliced downstream fusion transcripts initiated from intronic ERVs in *Slc15a2* (i.e., solute carrier family 15 [H^+/peptide transporter], member 2) and *Polr1a* (i.e., polymerase [RNA] I polypeptide A) among others, expressed in a tissue-specific manner (data not shown). We designate such polymorphic integrants as $ERV_{Slc15a2}$, ERV_{Polr1a}, etc.

To assess whether additional *Slc15a2* transcriptional variants are associated with the presence or absence of $ERV_{Slc15a2}$, we probed RNA blots using cDNA probes generated from both 5′ and 3′ ends of conventional, full-length *Slc15a2* transcripts (Fig. 3). Contrary to our initial RT-PCR results, the ERV-*Slc15a2* downstream fusion transcript is not abundantly expressed, as shown by Northern blot using a 3′ probe (Fig. 3A). However, both 5′ and 3′ probes demonstrated that the presence of $ERV_{Slc15a2}$ in intron 7 is strongly associated with premature transcriptional termination, resulting in up to 39-fold reductions in full-length 4-kb transcripts and concomitant increases of up to 13-fold or more of prematurely truncated 1.2-kb transcripts (Fig. 3A). These very significant changes in transcript structures and levels were observed in several tissues including brain and kidney. Notably, the prematurely truncated transcripts are expressed strongly only in strains harboring $ERV_{Slc15a2}$, confirmed by quantitative reverse transcriptase–mediated PCR (qRT-PCR) and expression microarray assays (Supplemental Fig. 3).

Slc15a2 encodes PEPT2, a well-studied transporter of peptide-like molecules that is expressed in mouse brain, choroid plexus, kidney, lung, breast, and eye. PEPT2 has significant biological importance since it transports pharmacologic agents such as beta-lactam antibiotics in the kidney and brain (Brandsch et al. 2008), protects against 5-aminolevulinic acid neurotoxicity (Hu et al. 2007), and reduces the analgesic effect of L-kyotorphin (Jiang et al. 2009). While *Slc15a2* experimental knockout mice are viable and fertile, their physiological transport of certain oligopeptides is disrupted (Shen et al. 2003; Smith et al. 2011).

The ERV-associated transcriptional disruption results in approximately threefold to ninefold reductions in protein expression in each individual B6 mouse tested, when compared with DBA/2J individuals that lack $ERV_{Slc15a2}$ (Fig. 3B). Resulting PEPT2 functional peptide transport activity is also significantly reduced in B6 brain and lung compared with DBA/2J tissues (Fig. 3C). Thus, significant functional variation between the mouse strains is strongly associated with the presence or absence of an intronic ERV integrant within this genetic locus.

The truncated *Slc15a2* transcript also was detected in strains lacking $ERV_{Slc15a2}$, albeit at very low levels, upon prolonged exposure of Northern blots (Fig. 4A). This prematurely truncated transcript does not terminate inside $ERV_{Slc15a2}$ itself and includes no sequences templated by that ERV, in contrast to transcripts terminated within other intragenic transposable elements (Zhou et al. 2007; Zhang et al. 2008; Li et al. 2010). Instead, 3′ rapid cloning of cDNA ends (rapid amplification of cDNA ends, RACE) experiments using adult brain total RNAs demonstrated that truncated *Slc15a2* transcripts stop ~1.5 kb upstream of $ERV_{Slc15a2}$, distal to the splice donor site at the 3′ end of exon 7 (GenBank accession numbers JF495121–JF495122) (Fig. 4B). Their 3′ ends precisely match transcripts previously identified in mammary tissue, tumor, and in day 16 embryos from B6 and FVB mouse strains (GenBank accession numbers NM_001145899, BC018335, AK018393, and BC051199). Notably, two weak predicted polyadenylation signal sequences (Beaudoing and Gautheret 2001)

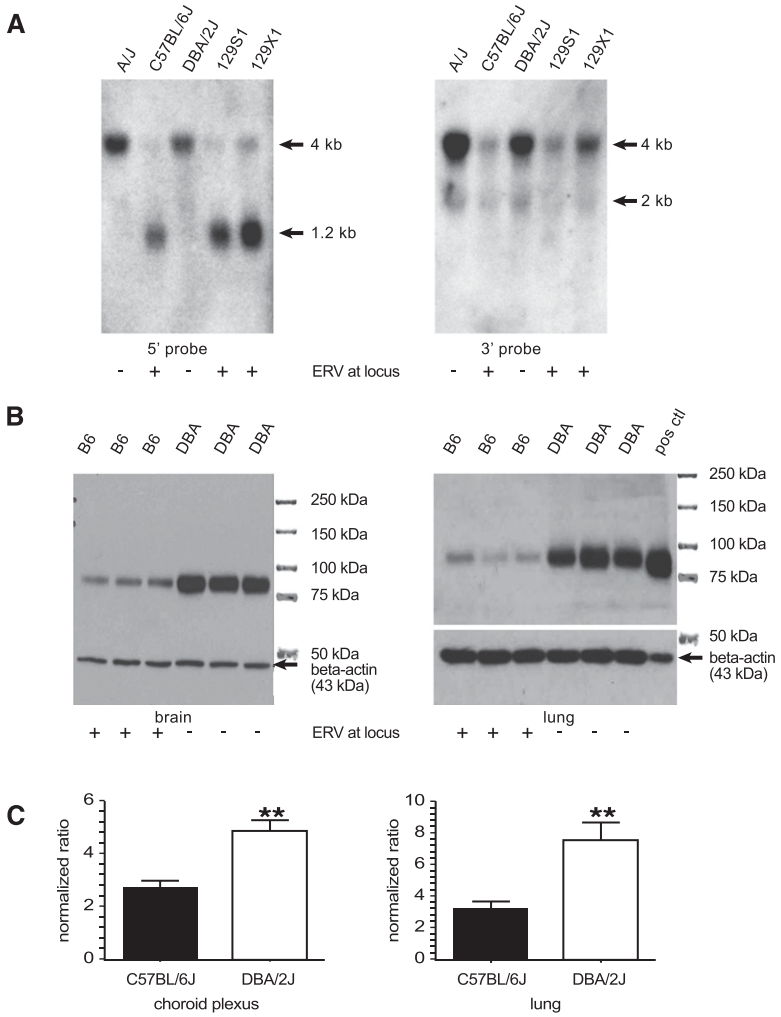

Figure 3. An intronic ERV polymorphism disrupts *Slc15a2* expression and function. (*A*) Northern blots. Equivalent amounts (10 mcg each) of total RNAs from brains pooled from several individuals from the indicated lineages were electrophoresed. Northern blots were probed with 5′ (*left*) and 3′ (*right*) probes from *Slc15a2*. (*Left*) Truncated transcripts (1.2 kb, arrow) correlate with the presence of a polymorphic ERV in B6, 129S1, and 129X1 strains but absent from the others. The full-length (nonterminated, 4 kb) *Slc15a2* transcript is expressed robustly in the absence of the ERV integrant in A/J and DBA mice. (*Right*) No appreciable downstream fusion transcript (2 kb) was detected, although it was identified by qRT-PCR (data not shown). Loading controls are shown in Supplemental Figure 3A. (*B*) Western blots. Protein extracts from individual brains (*left*) and lungs (*right*) from B6 and DBA mice were electrophoresced and probed for PEPT2 using protein-specific antiserum. (C) Functional assay in vivo. Accumulation of radiolabeled Gly-Sar dipeptide substrate was measured in choroid plexus and lung from B6 versus DBA mouse lineages, indicating significantly different PEPT2 functional activities (asterisks).

we could assess relationships between SNPs genome-wide and variable *Slc15a2* transcription by considering both truncated and full-length transcripts as eQTLs. The results demonstrate a very strong association between the $ERV_{Slc15a2}$-positive haplotype (as approximated by the closest informative SNP, rs4173858) and differential *Slc15a2* expression, i.e., both truncated and full-length transcripts (Fig. 5). Almost all BxD RI lines that are $ERV_{Slc15a2}$-positive express significantly more truncated *Slc15a2* transcript and significantly less full-length transcripts (Fig. 5B, bottom, cf. probe sets 1, 2, and 3). A few discrepant BxD lineages have SNP genotypes that appear to contradict the *Slc15a2* expression levels. These apparent discrepancies each were resolved by checking the absence/presence status of $ERV_{Slc15a2}$ (Supplemental Fig. 4; data not shown), rather than the adjacent SNP surrogate. Thus the ERV genotypes are all strongly correlated with the expression levels measured in each BxD RI strain.

We resequenced 1 kb upstream of and downstream from the premature termination site in multiple mouse strains (data not shown), disclosing only a single, previously identified, nonsynonymous SNP within exon 6 that does not correlate either with differential *Slc15a2* expression or with the polymorphic $ERV_{Slc15a2}$ integrant. Moreover, we compared 335 kb of adjacent genomic sequences in B6 versus DBA/2J wild-type genomes, thereby identifying 42 SNPs and seven small indel polymorphisms. None of these variants, other than $ERV_{Slc15a2}$ itself, are located inside of known coding genes; they each are upstream, downstream, or within gene introns, or within noncoding genes. None are classified as deleterious. Thus we conclude that $ERV_{Slc15a2}$ itself is the genetic determinant of variable transcription of *Slc15a2* in *cis*.

occur in intron 7, i.e., 5′-GATAAA and 5′-ATTAAA, immediately downstream from exon 7 and and upstream of the premature, added poly(A) tails (Fig. 4B).

To strengthen the genetic association between $ERV_{Slc15a2}$ and transcriptional disruption further, we analyzed *Slc15a2* expression data collected in kidneys from 53 distinct recombinant inbred (RI) mouse strains. These mice were derived from intercrosses of B6 and DBA/2J lines (B6 × DBA intercrossed mouse, BxD), resulting in a panel of highly recombinant mice with homozygosity at virtually every genetic locus, facilitating the identification of the genetic determinants of expression quantitative trait loci (eQTLs) (Chesler et al. 2005). Since B6 and DBA wild-type mice do and do not contain the $ERV_{Slc15a2}$ integrant, respectively,

Effects of the heterozygous ERV's parent of origin

To assess possible consequences of $ERV_{Slc15a2}$ heterozygosity on *Slc15a2* expression, we reciprocally crossed homozygous strains with (B6) and without (CAST/EiJ, CAST) this ERV, respectively. These intercrosses resulted in F_1 offspring with $ERV_{Slc15a2}$ integrants having either parent of origin. Nonterminated (i.e., presumably full-length) *Slc15a2* transcripts are significantly reduced in heterozygous CAST × B6 F_1 offspring with paternally derived $ERV_{Slc15a2}$ (Fig. 6A,B). This reduction is comparable to that observed in homozygous B6 mice, rather than an intermediate level of expression as predicted from heterozygosity of the ERV. Thus, terminated *Slc15a2* transcript expression is strongly associated

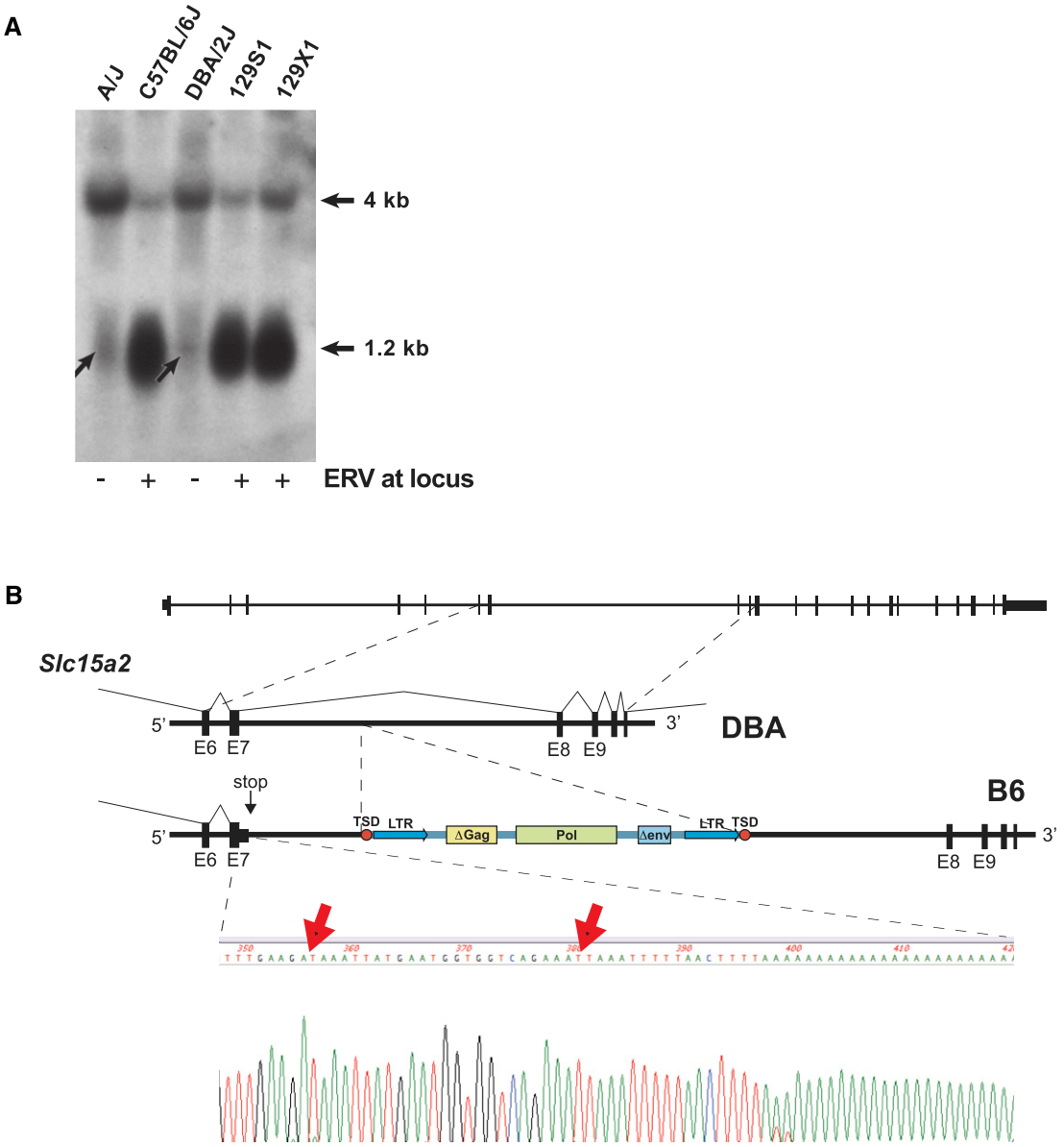

Figure 4. Transcriptional termination occurs at pre-existing signal upstream of ERV. (*A*) Prematurely terminated transcripts are present at low levels (*arrows*) in kidneys of mouse strains lacking the polymorphic ERV, indicating that the premature transcriptional polyadenylation signal exists both in strains that have or lack the ERV, and is not templated by the ERV per se. (*B*, *top*) Schematic of the *Slc15a2* locus showing site of premature transcriptional termination ~1.5 kb upstream of the intronic ERV polymorphism present in the B6 strain, in intron 7. (*Bottom*) Sequence trace from 3′-RACE experiment, demonstrating that the 3′ end of the prematurely truncated transcript is polyadenylated ~1.5 kb upstream of the ERV and contains no ERV-templated sequences per se. (Red arrows) Weak pre-existing polyadenylation signals (i.e., 5′-GATAAA and ATTAAA) are present in the intron, immediately upstream of the added poly(A) tail. GenBank accession numbers JF495121–JF495122.

with the introduced intronic ERV; one (ERV^+) allele can affect expression from the other (ERV^-). In contrast, F_1 offspring with the maternally derived $ERV_{Slc15a2}$ allele exhibit robust expression of nonterminated *Slc15a2* transcripts (Fig. 6A,B). In both crosses, we observed expression of both alleles at approximately equivalent levels (Supplemental Figs. 3 and 5). Thus the parent of origin of the $ERV_{Slc15a2}$ polymorphism affects the expression levels of nonterminated *Slc15a2* transcripts in the offspring, and transcriptional disruption can occur between alleles.

In contrast to differential expression of full-length transcripts, the prematurely truncated 1.2-kb transcript is detected at approximately equivalent, high levels in all mice that contain the ERV, much more than in strains lacking it (Fig. 6A,C). Notably, in some cases, the reduced expression of full-length transcripts is not correlated inversely with increased expression of prematurely truncated transcripts.

We sought to compare *Slc15a2* expression levels in individual, age-matched mice with the same $ERV_{Slc15a2}$ genotypes but derived from different genetic backgrounds. Thus we set up additional genetic crosses of wild-type and F_1 mice on both B6 and CAST genetic backgrounds, resulting in individual F_1 and F_2 offspring with all possible homozygous or heterozygous $ERV_{Slc15a2}$ geno-

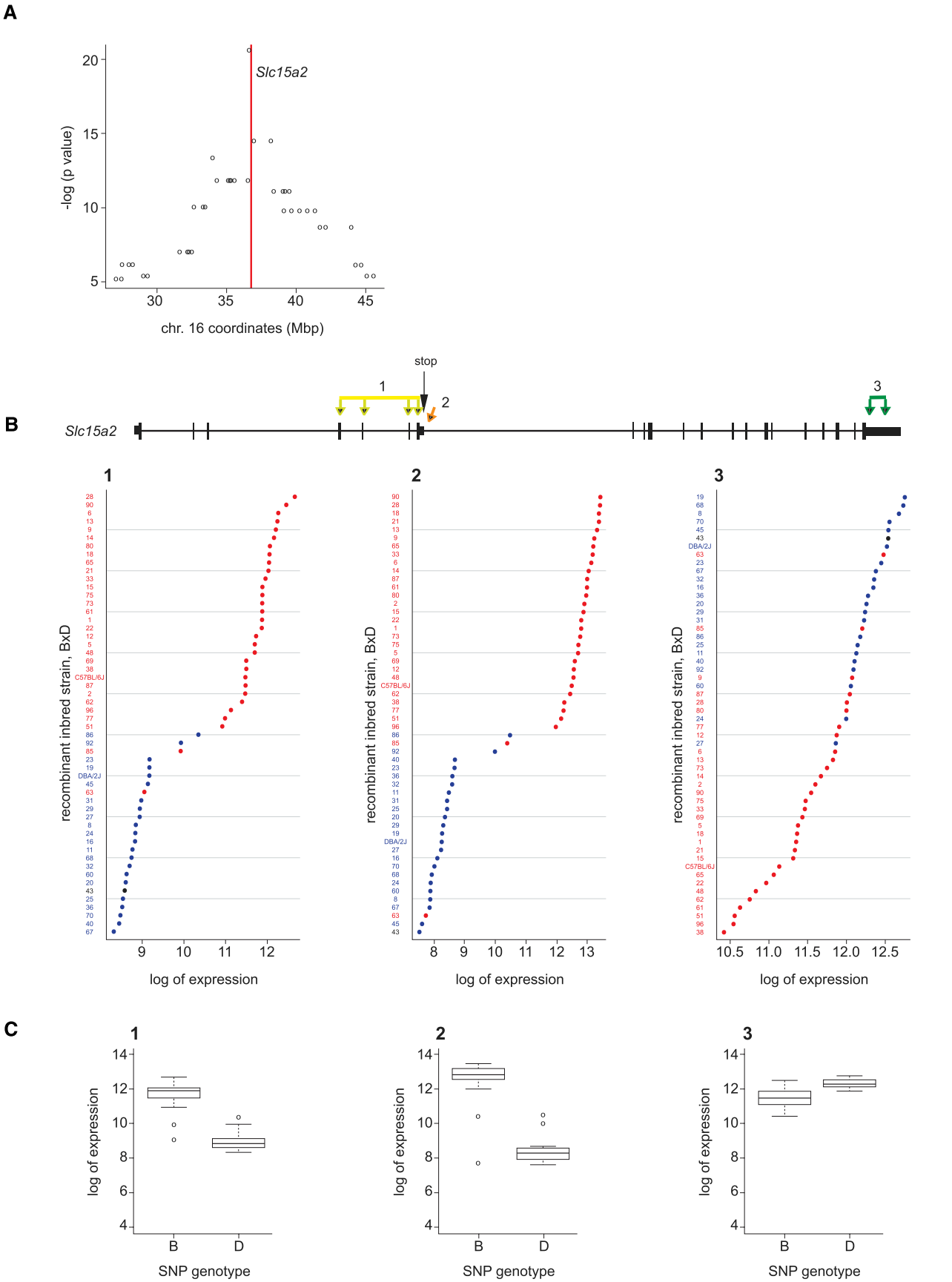

Figure 5. Strong genetic associations between transcriptional disruption and ERV$_{Slc15a2}$ status in *cis*. (*A*) eQTL permutation analysis indicates a very strong association between a SNP (rs4173858) genotype, which serves as a surrogate for ERV$_{Slc15a2}$ ~ 137 kb distant, and expression of the *Slc15a2* truncated transcript in mouse recombinant inbred BxD strain kidneys. (Red line) The chromosomal position of *Slc15a2*; (*y*-axis) *P*-values were calculated for the association between each SNP at the indicated chromosomal coordinates and truncated *Slc15a2* transcript levels. (*B, top*) Schematic of Affymetrix microarray probe sets detecting (1, 2) truncated or (3) full-length transcripts. (*Bottom*) Individual expression data (*x*-axis, log scale) measured by microarray probe sets (1–3) for each recombinant inbred BxD strain with indicated SNP genotypes: (red) B6; (blue) DBA; (black) heterozygous or indeterminate. (C) Box plots showing log of transcript expression versus genotypes: (B) B6; (D) DBA/2J. Error bars indicate SD. *P*-values for expression differences between genotypes B and D were calculated using a *t*-test: probe 1 = 1.80×10^{-22}; probe 2 = 5.53×10^{-23}, and probe 3 = 4.58×10^{-10}.

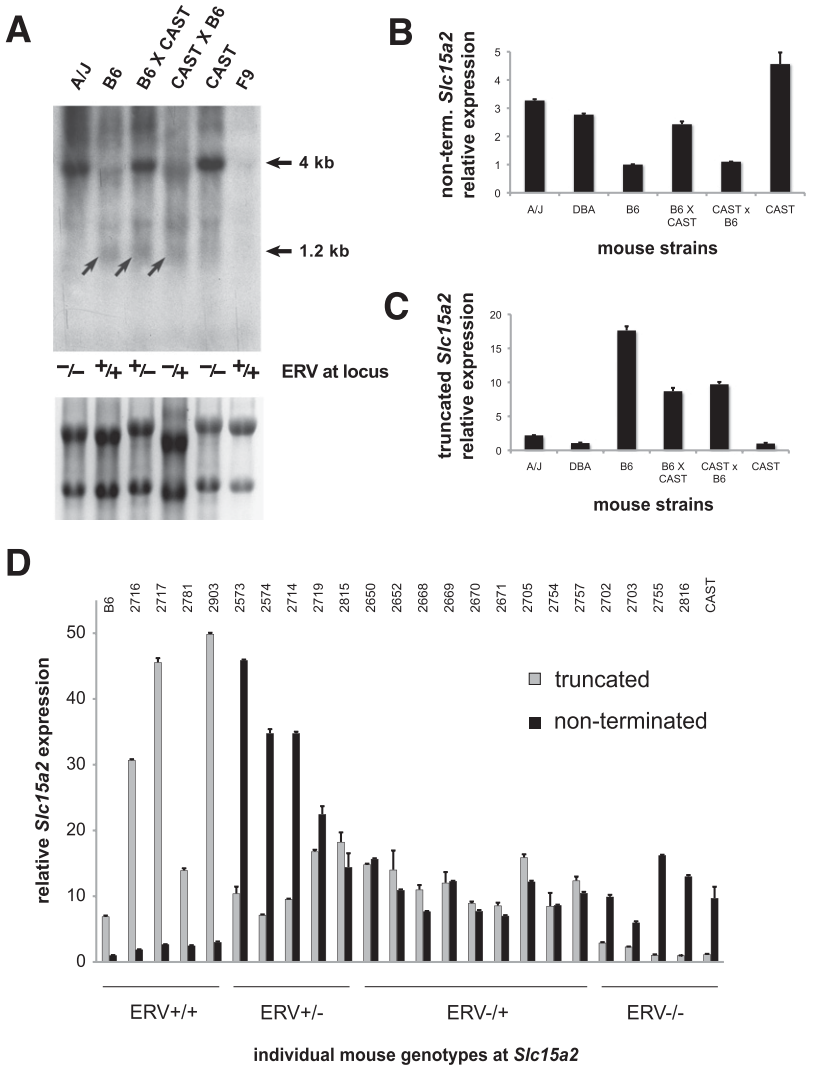

Figure 6. Transcriptional termination occurs between alleles in F_1 and F_2 mice. (*A*, *top*) Northern blot demonstrating differential reduction in full-length transcripts in brains from CAST × B6 but not B6 × CAST F_1 hybrid with heterozygous ERV integrants. In contrast, truncated transcripts (arrows) are detected in both lineages. (*Bottom*) Loading control showing 28S and 18S rRNA. Comparable amounts (10 mcg) of total RNA were loaded in each lane. (*B*) Quantitative RT-PCR assay for full-length transcripts (extending past exon 7) in brains from various mouse strains. Results are expressed as the fold change in levels relative to the sample with the lowest concentration. (*C*) Quantitative RT-PCR assay for the 3′ end of prematurely truncated transcripts shows that their expression is boosted specifically in strains containing $ERV_{Slc15a2}$. (*D*) Quantitative RT-PCR assays for full-length and prematurely terminated transcripts (each in duplicate or triplicate) in individual mice with indicated genotypes. Results were normalized to *Hprt* (i.e., hypoxanthine guanine phosphoribosyl transferase) transcript expression. (Error bars) Range of data. Numbers at *top* are identifiers for individual mice (Supplemental Table 3).

types where the allelic parents of origin are known unambiguously. We quantified both nonterminated and truncated *Slc15a2* transcripts in individual whole-brain extracts using qRT-PCR (Fig. 6D; Supplemental Fig. 5; Supplemental Table 4). Consistent with the results presented above (Fig. 3; Supplemental Fig. 3), nonterminated transcripts are significantly reduced in the presence of $ERV_{Slc15a2}$, up to ~16-fold, compared with its absence. Nonterminated transcript levels also are significantly lower in individuals with the paternally derived $ERV_{Slc15a2}$, compared with its maternal inheritance. Prematurely truncated transcripts are expressed robustly whenever the ERV is present and are increased further when $ERV_{Slc15a2}$ is present in homozygosity, i.e., up to ~49-fold (Fig. 6D). The results also demonstrate relatively modest variability between individuals with the same ERV genotype at *Slc15a2*, regardless of their diverse ancestries. Thus, neither ancestral exposures to the ERV nor unlinked, distant genetic modifiers alter these ERV-mediated effects substantially.

Since the nonterminated transcript levels reflect the parent of origin of the intronic ERV, we were prompted to assess DNA methylation at $ERV_{Slc15a2}$ in various heterozygous and homozygous mice. We observed differential methylation that is associated with the ERV's parent of origin (Fig. 7). Its 5′ long terminal repeat (LTR), closer to upstream *Slc15a2* exon 7, is relatively hypomethylated in B6 × CAST F1 mice, with only ~50% CpGs methylated. The 5′ LTR is more densely methylated in B6 (74%) and particularly in CAST × B6 F_1 (91%) mice. In contrast, the 3′ LTR is densely methylated (95%–100%) in all lineages tested (Fig. 7). Increased methylation at the 5′ LTR is associated with decreased levels of the nontruncated transcripts.

Disruption of other genes by polymorphic, intronic ERVs

We asked whether intronic ERVs in other, independent genes could disrupt their expression similarly. Using RT-PCR, we identified prematurely truncated transcripts at *Polr1a* and *Spon1* (i.e., spondin 1). The differential expression of truncated transcripts again correlates precisely with the presence or absence of intronic ERVs acting at a distance (Fig. 8; Supplemental Fig. 6; truncated *Polr1a* transcript, GenBank accession number AK087773.1). The polymorphic ERVs are oriented either parallel or antiparallel, respectively, relative to the genes' reading frames, indicating that transcriptional termination can be triggered independent of the ERV's orientation. While downstream fusion transcripts are robustly expressed in the case of ERV_{Polr1a} (data not shown), such expression is not necessary for premature truncation of the overlapping gene (as demonstrated at *Slc15a2*) (Fig. 3).

Polr1a nonterminated transcripts are significantly reduced with paternally derived ERV_{Polr1a} when compared with maternally derived ERV_{Polr1a} (Fig. 8D). This is similar to the association between reduced expression of nonterminated (full-length) *Slc15a2* and the paternally derived $ERV_{Slc15a2}$. Moreover, truncated transcripts are expressed at approximately equivalent levels from both alleles in offspring from both reciprocal crosses (Supplemental Fig. 6). Nonterminated transcripts of *Spon1* display biallelic expression, regardless of the parent of origin of ERV_{Spon1}.

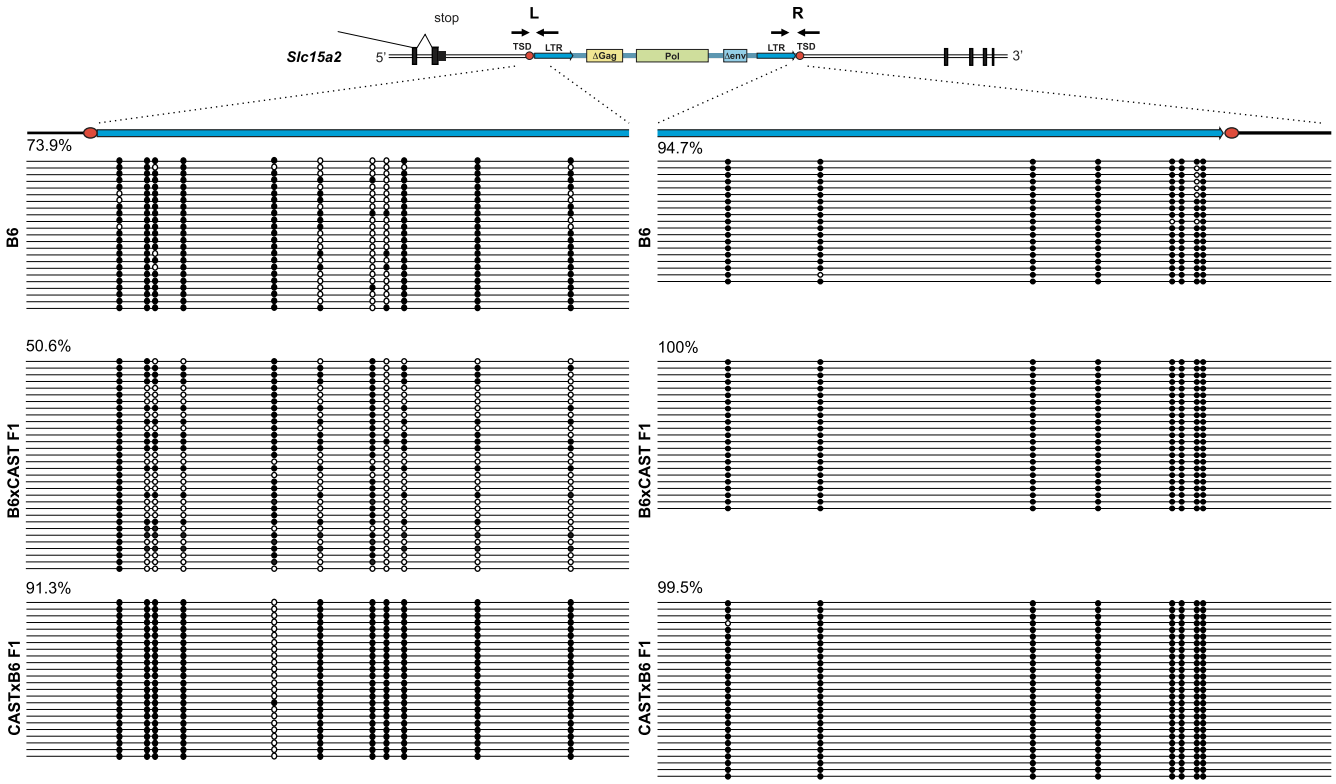

Figure 7. Differential methylation at ERV$_{Slc15a2}$ reflects its parent of origin. DNA methylation at left (L) 5′ and right (R) 3′ LTRs of ERV$_{Slc15a2}$ was assessed using bisulfite sequencing of genomic DNA purified from brains of indicated mouse lineages. (*Top*, schematic) In amplicon L, primers DES2652 and DES4883 yielded a 272-nt genomic DNA fragment to assess the methylation status of 11 CpG dinucleotides (circles) presented for multiple cloned alleles (horizontal lines). In amplicon R, primers DES4881 and DES2649 yielded a 304-nt fragment to assess eight CpGs. (Filled circles) Methylated cytosine; (open) unmethylated. (*Upper left* corner of each panel) Percentages of cytosines that are methylated.

We surveyed the mouse transcriptome for additional candidate genes whose expression may be disrupted by intronic ERVs <10 kb away from upstream exons. In addition to reidentifying premature transcriptional termination at *Slc15a2*, *Polr1a*, and *Cdk5rap1* (i.e., *Cabp*, CDK5 regulatory subunit associated protein 1) (Druker et al. 2004), this bioinformatics screen identified more than 100 independent genes including non-RefSeq transcripts (Table 2; Supplemental Table 5). Adding the prematurely truncated transcript at *Spon1*, where full-length transcription is disrupted by an intronic ERV at a genomic distance exceeding 12.5 kb (Fig. 8), we anticipate that many more prematurely truncated transcripts will be associated with adjacent ERVs in future studies of distinct mouse tissues, developmental stages, and nonreference mouse strains.

Discussion

By developing the transposon junction assay with targeted deep sequencing, we mapped thousands of young ERVs in highly divergent mouse strains. The resulting catalog of ERV polymorphisms facilitated the identification of particular transcripts whose differential expression in the highly divergent mouse lineages could be attributed to them. Integrants that are identical (i.e., present at orthologous loci) across such widely divergent lineages represent ancestral retrotransposition events that are identical by descent (Salem et al. 2005; Ray et al. 2011), while the youngest integrants are likely to be lineage-specific and are highly polymorphic (Fig. 1; Zhang et al. 2008; Qin et al. 2010). The ERV polymorphisms occur mostly in extragenic chromosomal regions and are at particularly low densities within embryogenesis genes and genes that are highly expressed in ES cells (Fig. 2). When present, they are oriented mostly antiparallel to the genes' reading frames, extending previous studies to previously unsequenced, highly divergent lineages (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2006; Zhang et al. 2008) and suggesting that they can disrupt gene expression and function. This distribution of polymorphic ERVs also suggests that remaining intronic integrants have survived purifying selection in the diverse mouse lineages over evolutionary time, because older elements are particularly depleted from intragenic sites.

We characterized several genes whose usual expression is disrupted profoundly by polymorphic ERV integrants acting at a distance, both in *cis* and between alleles. We conclude that the ERVs themselves are the genetic determinants of transcriptional disruption occurring at a distance in *cis*, because of several observations: (1) There is a strong and consistent association, across many mouse strains, between the presence of high levels of prematurely truncated transcripts and the presence of downstream, polymorphic, intronic ERVs. (2) Multiple independent genes at several different chromosomal positions exhibit similar effects. (3) Expression quantitative trait locus (eQTL) analysis in BxD recombinant strains established very strong genetic associations between the ERV-containing genotype and disrupted transcript isoforms (Fig. 5). (4) There are no other polymorphic genomic

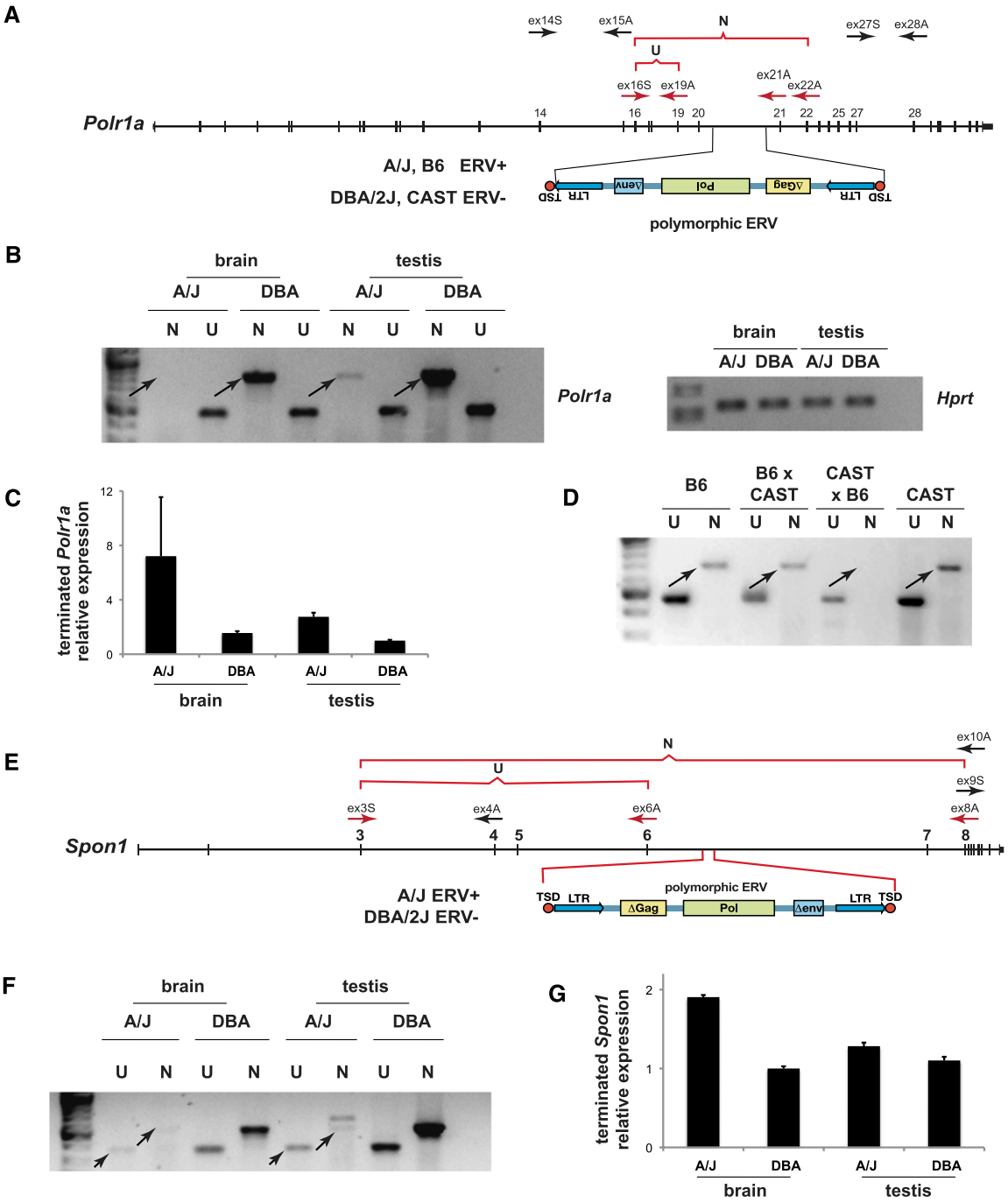

Figure 8. Disruption of additional genes by polymorphic, intronic ERVs in either orientation. (*A*) Genome structure of *Polr1a* containing a polymorphic AS ERV in intron 20, present in A/J and B6 and absent from DBA/2J and CAST mice. Various PCR primers are shown; (ex) exon number; (S) sense; (A) antisense. (Red arrows and brackets) cDNA amplicons; (U) upstream; (N) nonterminated, i.e., full-length. (*B*) Premature *Polr1a* termination occurs in brain and testis of A/J but not DBA mice. RT-PCR assays measured expression of upstream (U) versus nonterminated (N) transcripts, using ex16S and ex19A versus ex16S and ex22A primers, respectively. (Arrows) Differentially expressed, nonterminated transcripts. (*Right*) Loading control for spliced *Hprt* transcript assayed by RT-PCR. (*C*) Quantitative RT-PCR assay measuring relative differences between upstream and downstream *Polr1a* transcript levels, i.e., prematurely terminated transcripts. (Error bars) Range of duplicates. (*D*) Parent-of-origin effect on nonterminated *Polr1a* transcript levels in heterozygous mice. RT-PCR assays measured expression of upstream (U) vs. nonterminated (N) transcripts, using ex16S and ex19A versus ex16S and ex21A primers, respectively. (Arrows) Differentially expressed, nonterminated transcripts. See Supplemental Figure 6. (*E*) Genomic structure of *Spon1* containing a polymorphic ERV in intron 6, present in A/J but not DBA mice. (*F*) Premature *Spon1* termination in brain and testis of A/J but not DBA mice. (Arrows) Differentially expressed, upstream (U) and nonterminated (N) transcripts shown by RT-PCR assays. Both upstream and particularly full-length *Spon1* transcripts are reduced in A/J mice (based on similar input RNA levels vs. DBA mice). (*G*) Quantitative RT-PCR assay measuring relative differences between upstream and downstream *Spon1* transcript levels, i.e., prematurely terminated transcripts. (Error bars) Range of duplicates.

Table 2. Identification of candidate transcripts truncated prematurely by ERV integrants

Gene name		
1200014J11Rik	*Dtd1*	*Phf8*
2010015L04Rik	*Dym*	*Pkd1l2*
2010111I01Rik	*E130309F12Rik*	*Polr1a*
2510009E07Rik	*Enpp6*	*Poteg*
3300002I08Rik	*Epb4.1l2*	*Prrxl1*
4930430F08Rik	*Eps15*	*Qrsl1*
6720457D02Rik	*Exoc6*	*Rab3gap1*
Abcg3	*Fancd2*	*Ralgapa1*
Acly	*Galk2*	*Rhbdl2*
Adamts3	*Galnt10*	*Rnf157*
Agbl4	*Gimap5*	*Sag*
Akr1c14	*Gm4979*	*Sema3d*
Angpt1	*Golga3*	*Sgip1*
Arfgef2	*Gpsm1*	*Sirt5*
Asb3	*Iars2*	*Slc15a2*
Atp6v1h	*Ifi44*	*Slc17a5*
Bmx	*Iqca*	*Slc20a2*
Ccdc15	*Irak3*	*Slc25a46*
Ccdc46	*Itgb3bp*	*Slc38a1*
Cdk5rap1	*Katnal1*	*Slco6b1*
Cenpq	*Klhl13*	*Snx29*
Chl1	*Lama3*	*Sypl2*
Cmah	*Letm1*	*Tbc1d22a*
Cog6	*Mapk4*	*Tcfcp2*
Col4a4	*Me3*	*Tmco3*
Ctps2	*Me3*	*Tmed7*
Cyp20a1	*Mllt10*	*Trpm2*
Dcp1b	*Myom1*	*Txndc11*
Dctn4	*Ophn1*	*Uty*
Diap3	*Orai2*	*Vps52*
Dnahc1	*Otoa*	*Whsc1*
Dnahc7b	*Paqr3*	*Zfand3*
Dph5	*Parp4*	*Zfp407*
Dsg2	*Phc3*	

Listed here are candidate genes that may be disrupted by ERV (IAP) integrants found within 10 kb genomic distance downstream from premature termination sites. This analysis focused on reference B6 genes because currently available mouse transcriptome data are limited mostly to this strain. *Slc15a2* (Figs. 3–7), *Polr1a* (Fig. 8), and *Cdk5rap1* (Druker et al. 2004) were identified in this screen. *Spon1* is not listed here, although its transcription is disrupted by an intronic ERV (Fig. 8), because we limited this screen to identify only candidate ERVs <10 kb from the premature termination site.

features in *cis* that plausibly or consistently explain the observed expression differences.

The prematurely terminated transcripts identified here read past canonical splice donor sites and appear to use pre-existing intronic polyadenylation signal sequences that otherwise are not used routinely (Fig. 4B). We hypothesize that the polymorphic ERVs dramatically alter the use of these splicing and termination signals, which are nonpolymorphic regardless of the integrants' presence or absence (Figs. 3B, 4A). Such alternative transcriptional processing can occur coordinately (Wang et al. 2008), as exemplified by lamins expressed from *Lmnb2* (Furukawa and Hotta 1993) and *LMNA* (Lin and Worman 1993). Resulting transcripts containing intronic sequences at their 3′ ends have been termed "composite exons" (Yan and Marr 2005).

Similar transcriptional termination occurring at a distance has been attributed to an intronic ERV in *Cdk5rap1* (Druker et al. 2004). To our knowledge, the prematurely truncated *Cdk5rap1* transcripts are the only previously reported case of transcriptional disruption occurring upstream of such an ERV integrant. However, significant variability in levels of both downstream and prematurely terminated upstream *Cdk5rap1* transcripts was described between individual animals. Thus, the *Cdk5rap1* locus was described as a "metastable epiallele," comparable to highly variable expression of A^{vy} and $Axin1^{Fu}$ (Morgan et al. 1999; Whitelaw and Martin 2001). In contrast, we did not observe such a high degree of inter-individual variability at *Slc15a2* (Figs. 3, 6; Supplemental Fig. 5).

Nonterminated (i.e., full-length) transcript levels at *Slc15a2* and *Polr1a* reflect the parent of origin of heterozygous ERVs (Figs. 6, 8; Supplemental Figs. 3, 5, 6). When the intronic, heterozygous ERV was derived from the father, expression of nonterminated transcripts is reduced significantly from both alleles. In contrast, when the heterozygous ERV was maternally derived, full-length transcripts are expressed robustly, i.e., at levels similar to those in the ERV's absence. Regardless of the ERV's parent of origin and their overall expression, the nonterminated transcripts are expressed from both allelic templates at approximately equivalent levels (Supplemental Figs. 3, 6). Previously, transposons have been implicated as targets for establishment of imprinting and differentially methylated regions at particular loci (Suzuki et al. 2007), although a detailed molecular mechanism was not described. We found that the 5′ LTR of heterozygous $ERV_{Slc15a2}$ is differentially methylated, depending on its parent of origin (Fig. 7). When inherited from the father, the 5′ LTR is densely methylated, whereas when it is maternally inherited its methylation is reduced. This differential epigenetic control appears to be associated with differential levels of nonterminated transcripts. Such silencing epigenetic marks may mediate this parent-of-origin effect possibly by affecting transcriptional processivity past the ERV (Rebollo et al. 2011), although this does not directly explain the effects between alleles that we observed. The ERV's 3′ LTR is consistently methylated, regardless of its parent of origin (Fig. 7).

We did not observe a parent-of-origin effect in expression levels of prematurely truncated transcripts. Their expression appears to be boosted whenever the ERV is present. Moreover, the expression of full-length and truncated transcripts is not always inversely correlated; they do not sum up to a constant level of upstream initiation and transcription (Fig. 6D; Supplemental Fig. 5). This implies that transcriptional initiation, prolongation, splicing, and premature termination may not be coordinately regulated in some cases, and instead may undergo independent, complex patterns of regulation.

Several other distinct cases of transcriptional regulation and disruption illustrate multiple potential effects by ERVs on gene expression. An inverse association was reported between differential DNA methylation of the *Cdk5rap1* intronic ERV's 5′ LTR and premature termination of *Cdk5rap1* transcripts (Druker et al. 2004). Additionally, nonterminated *Cdk5rap1* transcripts are consistently expressed, independent of the terminated transcripts' variable expressivity. In axin fused mice, variable expression of downstream *Axin* transcripts has been associated with differential methylation of the 5′ LTR of the intronic ERV in the Ax^{Fu} allele, and either parent can transmit the epigenetic state (Rakyan et al. 2003). In contrast, in the A^{vy} mouse, variable expressivity of nonagouti is related to differential methylation of the 5′ LTR of an upstream ERV, and the epigenetic state is maternally inherited (Morgan et al. 1999). Transcripts from the imprinted mouse gene *H13* recently were found to undergo alternative polyadenylation that is regulated epigenetically by differential methylation of an internal promoter, albeit not in an ERV (Wood et al. 2008). Thus these various distinct expression patterns contrast with the transcriptional disruption described here.

Recently, bioinformatics analysis of the human transcriptome correlated antisense (AS) transcription with alternative splicing of overlapping sense-strand transcripts (Morrissy et al. 2011). The aberrantly spliced and terminated transcripts at *Cdk5rap1* (Druker et al. 2004) were attributed tentatively to possible AS transcription initiated from the intronic ERV promoter. However, evidence for AS transcription was not demonstrated, and the underlying molecular mechanism remains unknown. Such a model for transcriptional interference, i.e., collisions of bidirectional RNA polymerase complexes (Eszterhas et al. 2002), plausibly could explain *Cdk5rap1* disruption (Druker et al. 2004). However, transcriptional interference would not explain transcript expression differences between alleles as described here (Figs. 6, 8; Supplemental Figs. 3, 5, and 6). Alternatively, diffusible AS transcripts could act at long genomic distances, both in *cis* and between alleles. AS transcripts could affect host gene splicing by blocking U1 snRNA base-pairing with pre-mRNAs (Kaida et al. 2010). ERV-mediated alterations in gene "punctuation," where the polymorphic ERV could alter gene looping or interactions between homologous alleles by disrupting long-range interactions between upstream gene promoters and various downstream terminator sites (Tan-Wong et al. 2008), could provide another explanation for transcriptional disruption. Intragenic ERV integrants could introduce targets for heterochromatin formation that could disrupt full-length transcription in *cis*. Other possibilities also are plausible (Wilusz and Spector 2010).

We identified about 100 intronic ERV candidates that may trigger premature transcriptional termination at a distance (Table 2), out of approximately 1025 genes displaying evidence for premature termination. We speculate that other types of ERVs (i.e., ETn/ MusD elements) (Zhang et al. 2008), retrotransposons, or repetitive elements similarly could trigger transcriptional truncation in at least some of the remaining ~90% of genes lacking such intronic ERV candidates. We are addressing this interesting question currently.

Extrapolating from the number of intronic ERV polymorphisms identified in diverse mouse lineages, we estimate that up to ~10% of all genes containing intronic ERVs exhibit transcriptional disruption mediated by the integrants acting at a distance. This calculation may underestimate the full extent of ERV-mediated transcriptional disruption, since comprehensive transcipt expression data are lacking from various tissues and developmental time points from the divergent strains studied here. On the other hand, most of these candidates have not been validated by molecular assays. We did not detect a significant difference in the relative orientation of intronic ERVs that appear to trigger premature truncation (i.e., ~30% AS compared with overlapping genes), when compared with all intronic ERVs (i.e., ~23% AS, $p = 0.102$). We postulate that thousands of other intronic ERVs present in different lineages (Table 1) are unlikely to disrupt overlapping gene expression and function in this way, because such genes presumably lack the pre-existing, weak polyadenylation or alternative splicing signals that could be boosted by them. De novo intronic ERV integrants that strongly affect gene transcription, particularly of essential genes, would be expected to be highly deleterious, explaining their relative exclusion from embryogenesis genes and genes highly expressed in ES cells (Fig. 2). This conclusion is consistent with the demonstration that fusion transcripts are initiated from ERV LTR promoters in oocytes and in early embryogenesis (Peaston et al. 2004).

Our results strongly suggest that genome-wide studies based solely on SNP genotyping may miss important determinants of transcriptional variation and functional diversity. Comprehensive knowledge of all forms of structural variation within and between individuals, including indel polymorphisms caused by actively mobilized repetitive elements such as ERVs, will be critically important to understand the molecular basis for phenotypic variation (Li et al. 2010; Keane et al. 2011; Yalcin et al. 2011). Although ERVs appear to be inactive in humans, ~10% of the genome is comprised of such elements, suggesting that similar transcriptional disruption could be mediated by their promoter activities. While too numerous and diverse to describe here, other transposon families and retroposed elements continue to be actively mobilized in both the mouse and human genomes, thereby also introducing promoter activities, new polyadenylation signal sequences in *cis* (Li et al. 2010), new splicing sites, and targets for epigenetic regulation (Macfarlan et al. 2011; Monk et al. 2011). Further characterization of the molecular causes and consequences of transcriptional variation caused by genomic ERVs and other families of transposons and, in particular, the detailed mechanisms for premature transcriptional polyadenylation triggered at a distance undoubtedly will be promising areas for further study.

Methods

Mouse colony and genomic DNA

Mice were maintained and euthanized according to approved Institutional Animal Care and Use Committee protocols (National Cancer Institute, Frederick, MD; and Ohio State University, Columbus, OH). Mouse strains and purified genomic DNA were purchased from the Jackson Laboratory (Bar Harbor, ME).

Bioinformatics tools and statistical analysis

Alignments of pyrosequencing reads to the B6 reference genome assembly were performed using GMAP (Wu and Watanabe 2005; Akagi et al. 2008), BLAT, and BLAST. Results were parsed using custom Perl scripts with BioPerl modules. Statistical analyses were performed using SPSS (http://www.spss.com) or R software as described. Analysis of mouse ES cell expression data was based on public data sets in the GEO repository under accession number GSE8024 (Mikkelsen et al. 2007). Genes annotated as embryogenesis genes were identified from the Mouse Genome Informatics database (http://www.informatics.jax.org).

Further details about our assessment of possible bias in detecting ERVs using the transposon junction assay and our procedures for identification of ERVs in four "Celera strains" are provided in the Supplemental Material. To assess their chromosomal distributions (Supplemental Fig. 2), we counted retrotransposons in 500-kb bins genome-wide. Reference distributions of retrotransposons for each class (L1, ERV, and SINE) were obtained from the UCSC mm8 mouse reference assembly. Similarly, polymorphic retrotransposon distributions were determined by counting both unique insertions in reference and insertions in alternative strains we identified from four strains by Celera shotgun sequencing (Mural et al. 2002).

To identify candidate genes with prematurely truncated transcripts (Table 2), we compared chromosomal coordinates of ~20,180 RefSeq reference genes (UCSC Genome Browser) with those from the Known Gene track in the UCSC database. We compared annotated gene symbols in cases in which the genomic template for a Known Gene transcript is >20 kb shorter than that for a corresponding RefSeq gene transcript. In approximately 1025 cases in which the assigned NCBI gene ID numbers are

identical, we called such a Known Gene transcript a truncated variant of the full-length RefSeq gene. In such cases, we scanned within 10 kb downstream from the 3′ end of the truncated transcript for ERV elements (i.e., all IAP subtypes as defined by RepeatMasker).

Identification of ERVs using transposon junction assay

To map previously unsequenced ERV integrants in divergent mouse lineages, we developed a new high-throughput assay using nested PCR (Pornthanakasem and Mutirangura 2004) to amplify genomic sequences containing 3′ junctions of transposon integrants, followed by deep 454 sequencing (Supplemental Fig. 1). Forward PCR primers were designed to anneal within young, highly conserved ERV integrant sequences; details are provided in the Supplemental Material. Resulting sequence traces were aligned to the mm8 mouse reference genome and analyzed for indel polymorphism status, using modifications of our sequence alignment pipeline described in the Supplemental Material (Akagi et al. 2008). To identify previously unsequenced ERV integrants, sequencing reads from these PCR amplicons were mapped to the reference mouse genome assembly in three steps: preprocessing of reads, mapping of reads, and clustering of overlapping reads defining discrete insertion sites. We used this assay to identify previously unsequenced ERV elements in six diverse mouse lineages, i.e., A/J, B6, CAST, MOLF, SPRET, and WSB.

RNA isolation

To preserve high-quality total RNAs for downstream transcriptome analysis, we collected tissues from both sexes of inbred mouse strains at day 72. RNAs were collected from strains B6, 129S1, 129X1, DBA/2J, A/J, CAST, SPRET, MOLF, WSB, and intercrossed F_1 hybrid offspring B6 × CAST, and CAST × B6, respectively. Trimmed tissues were immediately immersed in RNA Later (Ambion) and either snap-frozen at −80°C or transferred to TRIzol (Invitrogen), homogenized, and frozen. The quality of total RNAs was determined using a model 2100 Agilent bioanalyzer where >95% of the samples had RIN scores >9. RNA specimens isolated from the same strain, tissue, gender, etc. were pooled from at least five individuals, unless noted.

Northern blots and RT-PCR assays

Total RNAs from indicated mouse tissues were electrophoresed in agarose gels under standard conditions, transferred to charged nylon membranes (GE Amersham), and hybridized with radiolabeled DNA probes at the 5′ and 3′ ends, respectively, of *Slc15a2* transcripts. Membranes were washed and exposed to film for autoradiography. To synthesize first-strand cDNAs for reverse transcriptase–mediated polymerase chain reaction (RT-PCR) assays, 10 μg each of mouse total RNAs was primed for reverse transcription, using T7-anchored oligo$(dT)_{24}$ and SuperScript II Reverse Transcriptase (Invitrogen). Gene-specific primers for *Slc15a2*, *Polr1a*, and *Spon1* were used to amplify resulting first-strand cDNAs. Products were assessed by agarose gel electrophoresis. Quantitative RT-PCR was performed using these cDNAs and Power SYBR Green PCR master mix (ABI) on a StepOnePlus instrument (ABI). To quantify relative expression of *Polr1a* truncated transcripts (Fig. 8), we measured upstream and downstream transcript levels, calculated the difference in PCR cycle numbers $\Delta\Delta C_T$ = (ex14S–ex15A) − (ex27S–ex28A), and then calculated linear differences as $2^{\Delta\Delta C_T}$. *Spon1* premature truncation was measured similarly. Further details are in the Supplemental Material.

RACE

5′- and 3′-RACE analyses were performed using the 5/3 RACE Kit, second generation (Roche Applied Science), the FirstChoice RLM-RACE kit (Ambion), and primers for *Slc15a2* (DES2622, 5′-CTTC TGACAAGCACTCTGGAG-3′) and *Polr1a* (DES4410, 5′-TGGTCT CACCCTTTCTGTAACG-3′) according to the kit manufacturers' protocols.

Western blots and PEPT2 functional assay

PEPT2 protein expression in tissues from individual mice was assayed by Western blots as described in the Supplemental Material. To assay PEPT2 protein functional activity, six B6 mice (three females, three males) and five DBA/2J mice (three females, two males) were injected via tail vein injection with 100 μL of GlySar solution containing 5 μCi of ^{14}C-GlySar (98 mCi/mmol, 0.1mCi/mL; Moravek). Tissue concentrations of GlySar (nanomoles per gram of wet tissue) were calculated as described in the Supplemental Material (Ocheltree et al. 2005; Shen et al. 2007).

Expression quantitative trait locus analysis

Slc15a2 expression data from kidneys of 53 BxD RI mouse strains were obtained from the Gene Network (http://www.genenetwork.org). Transcript levels were measured using the Affymetrix M430v2 platform (database access code MA_M2_0806_R), which includes three *Slc15a2*-specific probe sets. Two of these probe sets detect the 3′ end of truncated transcripts (1424730_a_at, 1447808_s_at), and the other detects the 3′ end of the full-length transcript (14171600_at) (Fig. 5B). To assess local strain genotypes B (B6) and D (DBA), 72 informative SNPs within 10 Mb on either side of *Slc15a2* were identified from the Gene Network. Expression levels for each genotype B and D were determined, and *P*-values were calculated using a *t*-test with multiple test correction according to the Holm method. For all three *Slc15a2* probe sets, the maximal −log(*P*-value) occurred at the SNP rs4173858, the closest informative SNP to *Slc15a2* in *cis*. Additional genomic sequence variants including SNPs and small indel polymorphisms flanking *Slc15a2* were identified in B6 and DBA/2J strains using Sanger Institute mouse genome sequencing data (http://www.sanger.ac.uk/resources/mouse/genomes/; SNP 20110125 release REL1101 and indel20100713 release REL1007).

Data access

Sequences as indicated were assigned Genbank accession numbers JF495121–JF495122. All ERVs identified here are accessible via our MouseIndelDB website at http://variation.osu.edu/ (Akagi et al. 2010).

Acknowledgments

We thank Drs. Albert de la Chapelle, Neal Copeland, and Maura Gillison for helpful comments; Robert Williams, Lu Lu, and Jesse F. Ingels (University of Tennessee) for help with the Gene Network and providing several BxD RI mouse strain genomic DNA samples; Holly Morris, Rob Koogle, and Sherry Rausch (National Cancer Institute) for superb maintenance of our mouse colony; Xiaolin Wu and Hongling Liao (SAIC Frederick) for assistance with exon microarray experiments; Clive Evans (Virginia Bioinformatics Institute) for 454 sequencing; and Richard Frederickson (SAIC Frederick) and Anthony Baker (OSU) for graphical illustration. This project was supported by the Intramural Research Program, Center for Cancer Research, National Cancer Institute, National Institutes

of Health (to J.L., K.A., A.L.T., D.A.S., T.C.M., D.E.Sy.); NIH research grant R01-GM035498 (to Y.H. and D.E.Sm.); contract no. HHSN261200800001E by the National Cancer Institute to SAIC, Inc. (to N.V., T.C.M., Y.G., and R.M.S.); and by The Ohio State University Comprehensive Cancer Center (to J.L., K.A., C.J.W.H., and D.E.Sy.). We thank the Ohio Supercomputer Center for providing computational resources (grant PAS0425-2 to K.A. and D.E.Sy.). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. NCI-Frederick and OSU are accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International and follow the U.S. Public Health Service Policy for the Care and Use of Laboratory Animals. Animal care was provided in accordance with the procedures outlined in the "Guide for Care and Use of Laboratory Animals" (National Research Council, 1996, National Academy Press, Washington, DC). Mouse studies were performed following protocols approved by the Animal Care and Use Committee, NCI Frederick or by the Institutional Animal Care and Use Committee, OSU.

References

Adams DJ, Dermitzakis ET, Cox T, Smith J, Davies R, Banerjee R, Bonfield J, Mullikin JC, Chung YJ, Rogers J, et al. 2005. Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat Genet* **37:** 532–536.

Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE. 2008. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res* **18:** 869–880.

Akagi K, Stephens RM, Li J, Evdokimov E, Kuehn MR, Volfovsky N, Symer DE. 2010. MouseIndelDB: A database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic Acids Res* **38:** D600–D606.

Banno F, Kaminaka K, Soejima K, Kokame K, Miyata T. 2004. Identification of strain-specific variants of mouse *Adamts13* gene encoding von Willebrand factor-cleaving protease. *J Biol Chem* **279:** 30896–30903.

Barr SD, Leipzig J, Shinn P, Ecker JR, Bushman FD. 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J Virol* **79:** 12035–12044.

Beaudoing E, Gautheret D. 2001. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* **11:** 1520–1526.

Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, Bushman FD. 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev* **23:** 633–642.

Brandsch M, Knutter I, Bosse-Doenecke E. 2008. Pharmaceutical and pharmacological importance of peptide transporters. *J Pharm Pharmacol* **60:** 543–585.

Cachon-Gonzalez MB, San-Jose I, Cano A, Vega JA, Garcia N, Freeman T, Schimmang T, Stoye JP. 1999. The *hairless* gene of the mouse: Relationship of phenotypic effects with expression profile and genotype. *Dev Dyn* **216:** 113–126.

Cahan P, Li Y, Izumi M, Graubert TA. 2009. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* **41:** 430–437.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37:** 233–242.

Copeland NG, Hutchison KW, Jenkins NA. 1983. Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. *Cell* **33:** 379–387.

Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T. 2004. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat Genet* **36:** 534–539.

Druker R, Bruxner TJ, Lehrbach NJ, Whitelaw E. 2004. Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Res* **32:** 5800–5808.

Eszterhas SK, Bouhassira EE, Martin DI, Fiering S. 2002. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Mol Cell Biol* **22:** 469–479.

Furukawa K, Hotta Y. 1993. cDNA cloning of a germ cell specific lamin B3 from mouse spermatocytes and analysis of its function by ectopic expression in somatic cells. *EMBO J* **12:** 97–106.

Horie K, Saito ES, Keng VW, Ikeda R, Ishihara H, Takeda J. 2007. Retrotransposons influence the mouse transcriptome: implication for the divergence of genetic traits. *Genetics* **176:** 815–827.

Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. 2007. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene* **27:** 404–408.

Hu Y, Shen H, Keep RF, Smith DE. 2007. Peptide transporter 2 (PEPT2) expression in brain protects against 5-aminolevulinic acid neurotoxicity. *J Neurochem* **103:** 2058–2065.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141:** 1253–1261.

Jiang H, Hu Y, Keep RF, Smith DE. 2009. Enhanced antinociceptive response to intracerebroventricular kyotorphin in Pept2 null mice. *J Neurochem* **109:** 1536–1543.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468:** 664–668.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* **9:** 411–414.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294.

Li Z, Mulligan MK, Wang X, Miles MF, Lu L, Williams RW. 2010. A transposon in *Comt* generates mRNA variants and causes widespread expression and behavioral differences among mice. *PLoS ONE* **5:** e12181. doi: 10.1371/journal.pone.0012181.

Lin F, Worman HJ. 1993. Structural organization of the human gene encoding nuclear lamin A and nuclear lamin C. *J Biol Chem* **268:** 16321–16326.

Lueders KK, Frankel WN, Mietz JA, Kuff EL. 1993. Genomic mapping of intracisternal A-particle proviral elements. *Mamm Genome* **4:** 69–77.

Macfarlan TS, Gifford WD, Agarwal S, Driscoll S, Lettieri K, Wang J, Andrews SE, Franco L, Rosenfeld MG, Ren B, et al. 2011. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev* **25:** 594–607.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. *Genome Res* **12:** 1483–1495.

Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL. 2005. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* **110:** 342–352.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Monk D, Arnaud P, Frost JM, Wood AJ, Cowley M, Martin-Trujillo A, Guillaumet-Adkins A, Iglesias Platas I, Camprubi C, Bourc'his D, et al. 2011. Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes. *Nucleic Acids Res* **39:** 4577–4586.

Morgan HD, Sutherland HG, Martin DI, Whitelaw E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23:** 314–318.

Morrissy AS, Griffith M, Marra MA. 2011. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res* **21:** 1203–1212.

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Ocheltree SM, Shen H, Hu Y, Keep RF, Smith DE. 2005. Role and relevance of peptide transporter 2 (PEPT2) in the kidney and choroid plexus: In vivo studies with glycylsarcosine in wild-type and PEPT2 knockout mice. *J Pharmacol Exp Ther* **315:** 240–247.

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7:** 597–606.

Pornthanakasem W, Mutirangura A. 2004. LINE-1 insertion dimorphisms identification by PCR. *BioTechniques* **37:** 750–752.

Qin C, Wang Z, Shang J, Bekkari K, Liu R, Pacchione S, McNulty KA, Ng A, Barnum JE, Storer RD. 2010. Intracisternal A particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Mol Carcinog* **49:** 54–67.

Rakyan VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KV, Whitelaw E. 2003. Transgenerational inheritance of epigenetic states at the murine *Axin^{Fu}* allele occurs after maternal and paternal transmission. *Proc Natl Acad Sci* **100:** 2538–2543.

Ray A, Rahbari R, Badge RM. 2011. IAP display: A simple method to identify mouse strain specific IAP insertions. *Mol Biotechnol* **47:** 243–252.
Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, et al. 2011. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet* **7:** e1002301. doi: 10.1371/journal.pgen.1002301.
Salem AH, Ray DA, Batzer MA. 2005. Identity by descent and DNA sequence variation of human SINE and LINE elements. *Cytogenet Genome Res* **108:** 63–72.
Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. 2011. Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res* **21:** 2004–2013.
Seperack PK, Mercer JA, Strobel MC, Copeland NG, Jenkins NA. 1995. Retroviral sequences located within an intron of the dilute gene alter dilute expression in a tissue-specific manner. *EMBO J* **14:** 2326–2332.
She X, Cheng Z, Zollner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. *Nat Genet* **40:** 909–914.
Shen H, Smith DE, Keep RF, Xiang J, Brosius FC III. 2003. Targeted disruption of the PEPT2 gene markedly reduces dipeptide uptake in choroid plexus. *J Biol Chem* **278:** 4786–4791.
Shen H, Ocheltree SM, Hu Y, Keep RF, Smith DE. 2007. Impact of genetic knockout of PEPT2 on cefadroxil pharmacokinetics, renal tubular reabsorption, and brain penetration in mice. *Drug Metab Dispos* **35:** 1209–1216.
Shen-Ong GL, Cole MD. 1982. Differing populations of intracisternal A-particle genes in myeloma tumors and mouse subspecies. *J Virol* **42:** 411–421.
Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657–663.
Smith DE, Hu Y, Shen H, Nagaraja TN, Fenstermacher JD, Keep RF. 2011. Distribution of glycylsarcosine and cefadroxil among cerebrospinal fluid, choroid plexus, and brain parenchyma after intracerebroventricular injection is markedly different between wild-type and Pept2 null mice. *J Cereb Blood Flow Metab* **31:** 250–261.
Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM. 1988. Role of endogenous retroviruses as mutagens: The *hairless* mutation of mice. *Cell* **54:** 383–391.
Suzuki S, Ono R, Narita T, Pask AJ, Shaw G, Wang C, Kohda T, Alsop AE, Marshall Graves JA, Kohara Y, et al. 2007. Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet* **3:** e55. doi: 10.1371/journal.pgen.0030055.
Takabatake T, Ishihara H, Ohmachi Y, Tanaka I, Nakamura MM, Fujikawa K, Hirouchi T, Kakinuma S, Shimada Y, Oghiso Y, et al. 2008. Microarray-based global mapping of integration sites for the retrotransposon, intracisternal A-particle, in the mouse genome. *Nucleic Acids Res* **36:** e59. doi: 10.1093/nar/gkn235.
Tan-Wong SM, French JD, Proudfoot NJ, Brown MA. 2008. Dynamic interactions between the promoter and terminator regions of the mammalian *BRCA1* gene. *Proc Natl Acad Sci* **105:** 5160–5165.
van de Lagemaat LN, Landry JR, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19:** 530–536.
van de Lagemaat LN, Medstrand P, Mager DL. 2006. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol* **7:** R86. doi: 10.1186/gb-2006-7-9-r86.
Wade CM, Daly MJ. 2005. Genetic variation in laboratory mice. *Nat Genet* **37:** 1175–1180.
Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20:** 116–117.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.
Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wheelan SJ, Aizawa Y, Han JS, Boeke JD. 2005. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15:** 1073–1078.
Whitelaw E, Martin DI. 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat Genet* **27:** 361–365.
Wilusz JE, Spector DL. 2010. An unexpected ending: Noncanonical 3′ end processing mechanisms. *RNA* **16:** 259–266.
Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. 2010. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11:** 410. doi: 10.1186/1471-2164-11-410.
Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D, Oakey RJ. 2008. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev* **22:** 1141–1146.
Wu TD, Watanabe CK. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21:** 1859–1875.
Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellaker C, Goodstadt L, Nicod J, Bhomra A, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature* **477:** 326–329.
Yan J, Marr TG. 2005. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15:** 369–375.
Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.
Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. *Nat Genet* **39:** 1100–1107.
Zhang Y, Maksakova IA, Gagnier L, van de Lagemaat LN, Mager DL. 2008. Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genet* **4:** e1000007. doi: 10.1371/journal.pgen.1000007.
Zhou W, Bouhassira EE, Tsai HM. 2007. An IAP retrotransposon in the mouse ADAMTS13 gene creates ADAMTS13 variant proteins that are less effective in cleaving von Willebrand factor multimers. *Blood* **110:** 886–893.

Received August 16, 2011; accepted in revised form February 9, 2012.