SUPPLEMENTAL INFORMATION

Endogenous retroviruses can trigger premature transcriptional termination at a distance – Li et al.

Identification of ERVs using transposon junction assay

Out of 565,265 barcoded 454 pyrosequencing traces generated, 200,713 traces contain at least 50 nt of transposon-derived sequence and at least 15 nt of flanking genomic sequence, and mapped to the reference genome.

Validation of ERVs in B6 reference genome by targeted sequencing

To check if there is detection bias in identifying ERVs using the transposon junction assay, at a chromosomal level, we counted ERV elements (i.e. IAPLTR1, IAPLTR2 and IAPEY2) in 10 Mbp bins for both reference (expected) and sequencing output (observed). Using Fisher's exact test (alpha = 0.05), we found that zero out of 255 bins showed a distribution distinct from the reference distribution of IAPLTR1 elements. Similarly, out of 255 bins, only 2 (IAPLTR2) or 5 (IAPEY2) bins showed distributions distinct from the reference. Further, the observed numbers of ERVK elements in bins are significantly correlated well to the numbers of expected ERVK elements ($p<2.2E-16$). Pearson's correlation coefficients are IAPLTR1=0.75, IAPLTR2=0.78, IAPEY2=0.60.

Comparison between transposon integrants in reference genome vs. other mouse lineages.

The proportions of various classes of polymorphic transposons that are present in the reference genome are very similar to those identified in the previously unsequenced Celera strains (**Supp. Fig. 2**). Notably, the relative proportions of polymorphic SINE retrotransposons present in the reference genome vs. in the previously unsequenced strains are the most discrepant of all transposon classes (**Supp**. **Fig. 2**). This discrepancy could be due to technical aspects of alignment of WGS traces to the SINEs' target sites: polymorphic SINE integrants would be "skipped over" when present in the reference sequence, while by contrast their precise genomic

junctions and mobilized internal sequences would be required within the sequence traces when they are present in the previously unsequenced strains.

Analysis of other WGS traces, predicting a polymorphic transposon integrant present in a previously unsequenced genome (that is the traces' source) but absent from the reference genome, is more complicated. The reason why full-length sequences for such integrants usually cannot be determined from WGS trace alignments is that such sequence traces on average are only 799 nucleotides (nt) long, so they can extend only partially into longer polymorphic integrants (Akagi et al. 2008). The WGS traces generally were not derived from BAC clones, which would facilitate unique identification of repetitive elements because of the limited genomic sequences that are represented. Moreover, although traces were filtered for overall sequence quality, we observed that in many cases, their quality deteriorated at one end, with patches of sequence data with Phred scores below 20 (see Methods).

To address these issues about the particular categories of WGS trace alignments, where one end aligned well to the reference genome while the other was unaligned, we used RepeatMasker (Smit et al. 2009) to identify particular classes of mouse transposon sequences within them. We tested the hypothesis that a significant fraction of such WGS traces could be comprised of polymorphic transposon integrants, so they could be used to map junctions between transposon variants and unique genomic sequences. Since we previously observed that most polymorphic integrants in the B6 reference genome are caused by endogenous retrotransposition by LINE (L1), SINE and ERV-K LTR retrotransposons, we used RepeatMasker to identify corresponding WGS traces containing these sequences. These repeat sequences within the traces were then masked temporarily, while the remaining, non-repetitive sequences of the WGS traces were re-aligned to the reference genome using BLAT. The resulting alignments were used to categorize and map polymorphic integrants in the unknown strains.

SUPPLEMENTAL METHODS

Identification of ERVs using transposon junction assay

To develop the transposon junction assay, we designed forward PCR primers to anneal within young, highly conserved ERV integrant sequences. To amplify young ERVs, i.e. IAPLTR1 elements, we used a relatively non-

degenerate primer no. 1, i.e. DES3356, 5'-AGATTYTGGTCTGTGGTGTTCTTC-3', which anneals approximately 31 nt from the 3' end of the LTR. To amplify a wider range of ERV IAP subfamilies including older IAPLTR2 and IAPEY2 elements with more divergent LTRs, we used a more degenerate primer no. 2, DES3355, 5'- CTTTGCCGCAGAAGRWTCYG-3', positioned about 47 nt from the 3' end. Reverse primers are based on adaptor sequences ligated to genomic restriction endonuclease sites such as Sau3AI, adjacent to the genomic ERV integrant. See **Supp. Table 1**.

Resulting sequence traces were aligned to the mm8 mouse reference genome and analyzed for indel polymorphism status, using modifications of our sequence alignment pipeline (Akagi et al. 2008). A key advantage of this procedure, in comparison with other mapping procedures (Ewing and Kazazian 2010; Huang et al. 2010), is that it directly gives precise genomic sequence junctions for insertions. Similar methods have been described recently (Iskow et al. 2010; Witherspoon et al. 2010).

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. Prior to 454 sequencing, we multiplexed samples using unique identifying barcodes of ten nucleotides. Resulting sequence reads containing a perfectly matching barcode or a barcode with a single nucleotide mismatch were categorized according to sample inputs. After trimming barcodes from 454 reads, we confirmed the presence of distal ERV sequences, to identify *bona fide* genomic integrant junctions. Long terminal repeat sequences of mouse IAP retrotransposons from RepBase (Jurka et al. 2005) were used as query sequences for Cross_match to search the reads (parameter: minscore=35). Reads with long terminal repeat sequences immediately adjacent to the barcode (i.e within 5 nucleotides, nt in the sense direction) indicated positive PCR amplicons. A majority of reads contained valid IAP long terminal repeats (>70% for degenerate primer and >90 % for non-degenerate primer; data not shown).

Traces containing 10 nt or longer flanking genomic sequences were selected for further genome mapping. To identify ERV insertion sites, we aligned untrimmed/trimmed reads against mouse genome assembly (UCSC mm8 assembly). To re-map reference ERVs, i.e. those elements already annotated in the B6 genome, we aligned IAP untrimmed traces (containing both ERV LTR and flanking genomic sequence) to the reference assembly

using BLAT. If untrimmed reads aligned with identity >90%, coverage >90%, and score difference to next best hit >10 points, we called such reads as mapped to unique ERVs present in the reference genome. To identify previously unsequenced IAP retrotransposons not annotated in the reference assembly, we trimmed the ERV LTR sequences from unmapped reads and mapped the remaining sequence traces to the reference genome using BLAT. We used similar criteria for unique alignment cutoff (identity >90%, coverage >60%, score difference to next best hit >10 points).

Unique loci then were identified by clustering reads that each aligned within 20 nt to the same chromosomal loci in the same orientation. The absence of ERV insertions in the reference genome immediately adjacent to called integrants was confirmed by comparison of RepeatMasker output for the reference genome assembly.

The transposon junction assay was used to identify previously unsequenced ERV elements in six diverse mouse lineages, i.e. A/J, B6, CAST, MOLF, SPRET and WSB. Out of 565,265 barcoded 454 pyrosequencing traces generated, 200,713 traces contain at least 50 nt of transposon-derived sequence and at least 15 nt of flanking genomic sequence, and mapped to the reference genome.

Identification of mouse ERVs from conventional Celera sequences

We downloaded 26 million Celera whole genome shotgun sequence reads (Mural et al. 2002) from the NCBI TraceDB archive (http://www.ncbi.nlm.nih.gov/Traces). These conventional Sanger sequence reads average approximately 800 nt in length. They were obtained from "Celera" mouse strains, i.e. B6, A/J, DBA/2J, 129S1/SvImJ and 129X1/SvJ (Akagi et al. 2008). To identify both non-polymorphic ERV insertions and previously unknown, polymorphic ERV insertions, these reads were mapped to the B6 mouse reference genome assembly (UCSC mm8 build,(Kuhn et al. 2009)) using GMAP(Wu and Watanabe 2005). Resulting read alignments were categorized into different groups, distinguishing between insertions in the B6 reference genome vs. in the other strains. Reads from the first category aligned to the reference with two anchoring high scoring pairs (each length >200 nt, identity >90%) flanking one gap (between 100 nt and 10 kb), as we previously described (Akagi et al. 2008). Reads in the second category had one well-aligned end (size >200 nt, identity

>90%) and the other end not aligned (non-aligned fragment size >100 nt). To identify insertions in alternative strains, we analyzed these 2.1 million partially aligned traces further. Low phred quality scores were associated with the non-aligned region of 1.9 million of these reads. The rest of the traces included 50% or more repetitive elements in the non-aligned region or occasionally in the aligned region. To identify the recent movement of retrotransposons, we trimmed retrotransposon sequences (SINEs, LINEs, and LTRs including ERVs) from these traces using RepeatMasker (Smit et al. 2009) and remapped these trimmed traces to the mouse reference assembly using BLAT (Kent 2002). If the trimmed traces were mapped with identity>90% and coverage>90%, those that mapped within +/- 20 nt of each other in the same orientation were merged into unique insertion clusters. We identified >30,000 retrotransposon-related indel polymorphisms: 12,305 insertions in the B6 reference genome that are not present in at least one of the other Celera strains, and 18,594 insertions in at least one of the four other Celera strains that are not present in the reference B6 genome (**Supp. Fig. 2**).

Northern blots and RT-PCR assays

For Northern blot hybridization, total RNAs isolated from various mouse tissues were electrophoresced in agarose gels under standard conditions, transferred to charged nylon membranes (GE Amersham) and hybridized with radiolabeled DNA probes, essentially as described previously (Li et al. 1999). Membranes were washed and exposed to film for autoradiography. Probes at the 5' and 3' ends, respectively, of *Slc15a2* transcripts were generated by amplifying mouse brain cDNA using forward primer (exon 1) DES2848 (5' AAATGAGTCCAAGGAAACGCTC 3') and reverse primer (exon 6) DES2849 (5' GGTAGAAGACCGAGAAGTATC 3'), and forward primer (exon 13) DES2850 (5' TTTCTGGTCC TTGTCTTCATCCC 3') and reverse primer (exon 18) DES2851 (5' CTGAGAGGAGCATTGGCATC 3'). After agarose gel electrophoresis, PCR-generated probes were purified using a gel purification kit (Qiagen). Probes were labeled with alpha-P^{32} dCTP using the Megaprimer DNA labeling system (Amersham) and were purified using ProbeQuant G-50 Micro columns (Amersham).

To synthesize first strand cDNAs for reverse transcriptase-mediated polymerase chain reaction (RT-PCR) assays, 10 microgram each of mouse total RNAs was primed for reverse transcription, using T7 anchored

oligo(dT)$_{24}$ DES2633 (5’ GGCCAGTGAATTGTAATACGACTCACTATAGGGAGGCGGTTTTTTTTTTTTTTTTTTTTTTT 3’) and SuperScript II Reverse Transcriptase (Invitrogen). Gene-specific primers for *Slc15a2*, *Rpo1-4* and *Spon-1* 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* prematurely 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 transformed to linear differences by calculating 2 ^ $\Delta\Delta C_T$. *Spon1* premature truncation was measured similarly.

Bioinformatics tools

Graphical representations of exon microarray analysis results were generated using Partek Genomics Suite (http://www.partek.com). Venn diagrams were drawn using VennMaster (Kestler et al. 2008).

Western blots

To quantify PEPT2 expression levels and isoforms in untreated mouse samples, mice were euthanized by CO_2 asphyxiation, tissues immediately were dissected (Pathology and Histology Laboratory, National Cancer Institute – Frederick, MD), and immediately minced and flash frozen in liquid nitrogen. Frozen tissues were homogenized in 2 ml of NP-40 lysis buffer (containing 50 mM Tris HCl pH 8.0, 150 mM NaCl, and 1% NP-40 with protease inhibitor cocktail; Roche Diagnostics Corp, Indianapolis, IN) using several strokes of a hand-held homogenizer (size 20; Kontes Glass Co, Vineland, New Jersey). Samples were then ultrasonicated with 5 strokes at 50% power strength (Model W-225R; Heat System-Ultrasonics Inc, Plainview, NY) and centrifuged at 12,000 g for 10 min at 4ºC. Protein concentrations were measured using the Pierce BCA Protein Assay kit (Thermo Scientific, Rockford, IL) and stored at -80ºC. Whole tissue protein extracts were denatured at 37ºC for 30 min in 1x sample buffer containing 6.25 mM Tris HCl, pH 6.8, 2% SDS, 10% glycerol, 0.1% bromophenol blue, and 0.1

M DTT. Samples were separated by 7.5% SDS-PAGE gel electrophoresis and transferred to a PVDF membrane (Millipore, Bedford, MA). For Western blotting against PEPT2, membranes were blocked with 5% skim milk in TBST buffer (150 mM NaCl, 10 mM Tris HCl pH 7.4, and 0.1% Tween 20), incubated with primary rabbit anti-mouse PEPT2 antisera raised against the COOH-terminal region, KQIPHIQGNMINLETKNTRL, amino acids 721-740 (1:5,000 dilution) (Hu et al. 2008). The filters were washed three times with TBST and then incubated with goat anti-rabbit IgG conjugated to horseradish peroxidase (1:3,000 dilution) (Bio-Rad, Hercules, CA). For β-actin, the membrane was blotted with a mouse monoclonal antibody (1:1,000 dilution) (Santa Cruz Biotechnology, Santa Cruz, CA) followed by the secondary goat anti-mouse IgG conjugated to horseradish peroxidase at 1:3,000 dilution (Santa Cruz Biotechnology, Santa Cruz, CA). The membranes were then washed five times in TBST, and bound antibody was detected with Immobilon Western chemiluminescent substrate (Millipore, Billerica, MA).

PEPT2 functional assay

To assay PEPT2 protein functional activity in different mouse strains and tissues, six B6 mice (3 females, 3 males) and five DBA/2J mice (3 females, 2 males) were anesthetized with pentobarbital sodium (40 mg/kg ip) and then administered, via tail vein injection, 100 µL GlySar solution containing 5 µCi of ^{14}C-GlySar (98 mCi/mmol, 0.1mCi/ml; Moravek, Brea, CA). The final dosage of GlySar was based on 20 g mice at 2.5 nmol/g body weight. To correct for tissue vascular space, 100 µl of 0.2 µCi ^{3}H-dextran-70,000 (265 mCi/g; American Radiolabeled Chemicals, Saint Louis, MO) was injected intravenously two min prior to euthanasia and tissue collections. Whole blood, combined choroid plexuses (from lateral and fourth ventricles), and lung were obtained 60 min after GlySar administration. A 0.3-ml aliquot of hyamine hydroxide was added to each sample and treated under standard protocol; 7 ml of CytoScint scintillation liquid was then added. Samples were stabilized for one week and counted on a scintillation counter. Corrected tissue concentrations of GlySar (nmol/g of wet tissue) were calculated as $C_{tiss} – DS \times C_b$, where C_{tiss} is the uncorrected tissue concentration of GlySar (nmol/g), DS is the dextran space (mL/g), and C_b is the GlySar blood concentration. (Ocheltree et al. 2005; Shen et al. 2007)

Since differences in tissue concentration could be due to concentration differences of GlySar in the perfusing blood, the data are reported as tissue-to-blood concentration ratio (i.e., normalized ratio).

SUPPLEMENTAL REFERENCES

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**(6): 869-880.

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**(19): 5800-5808.

Ewing AD, Kazazian HH, Jr. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**(9): 1262-1270.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**(7157): 1050-1053.

Hu Y, Smith DE, Ma K, Jappar D, Thomas W, Hillgren KM. 2008. Targeted disruption of peptide transporter Pept1 gene in mice significantly reduces dipeptide absorption in intestine. *Mol Pharm* **5**(6): 1122-1130.

Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T et al. 2010. Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**(7): 1171-1182.

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**(7): 1253-1261.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**(1-4): 462-467.

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**(7364): 289-294.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**(4): 656-664.

Kestler HA, Muller A, Kraus JM, Buchholz M, Gress TM, Liu H, Kane DW, Zeeberg BR, Weinstein JN. 2008. VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC Bioinformatics* **9**: 67.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M et al. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**(Database issue): D755-761.

Li J, Protopopov AI, Gizatullin RZ, Kiss C, Kashuba VI, Winberg G, Klein G, Zabarovsky ER. 1999. Identification of new tumor suppressor genes based on in vivo functional inactivation of a candidate gene. *FEBS Lett* **451**(3): 289-294.

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**(5573): 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**(1): 240-247.

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**(7): 1209-1216.

Smit AFA, Hubley R, Green P. 2009. RepeatMasker.

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.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(9): 1859-1875.

SUPPLEMENTAL FIGURE LEGENDS.

Supp. Fig. 1. **Transposon junction assay efficiently identifies most young ERV integrants.** (**A**) Schematic of assay. *Top*: full-length mouse ERV insertion containing LTRs (*blue horizontal arrows*) and flanked by target site duplications (TSD, *red circles*). Genomic DNA samples were digested with restriction endonucleases (*red, blue, green arrows*) such as DpnII, MseI or BfaI, and then MfeI (*yellow arrow*) to destroy internal ERV sequences that otherwise would be trivially sequenced. Adaptors including 454 primer and barcode sequences were ligated to resulting overhangs. Transposon junctions were amplified by PCR, using a degenerate primer specific for transposon sequences paired with an adaptor primer, and 454 sequencing was performed. *Bottom*: resulting sequence traces were aligned to the mm8 mouse reference assembly, confirming known ERV integrants (*left*) and identifying previously unsequenced indel polymorphisms (*right*) (Akagi et al. 2008). (**B**) Remapping reference ERV integrants. *Horizontal bars:* schematic of mouse genome represented in chromosomes 1-19, X and Y. We divided the mouse genome into 10 MB bins (total, 255 bins) and compared counts of young ERVK elements (categorized as IAPLTR1, IAPLTR2, and IAPEY2) in each bin for observed sequencing output (*histograms above, blue*) vs. expected reference elements (*histograms below, red*). Using Fisher's exact test (alpha=0.05), for IAPLTR1 elements, 0 out of 255 bins showed observed counts significantly different from the reference distribution. For IAPLTR2 elements, only 2 out of 255 bins, and for IAPEY2, 5 out of 255 bins showed observed counts significantly different from the reference counts. (**C**) A collection of ERV integrants identified by the transposon junction assay using 454 sequencing (*top*) was validated by PCR (*bottom*). (See **Supp. Table 2**.) *Top*: In each row, the name, chromosomal coordinates, number of 454 reads identified from the transposon junction assay per strain, presence (*1*) or absence (*0*) status in the reference genome assembly, and ERV subfamily classification are presented for ERV integrants tested by PCR. Genomic DNAs purified from B6, CAST and SPRET mice were amplified by PCR using site-specific primers. *Bottom*: Products from each tested ERV

integrant locus were electrophoresced in agarose gels. Occupied vs. empty target sites were identified by bands of appropriate sizes.

Supp. Fig. 2. **Distributions of retrotransposon integrants in various mouse lineage genomes.** (**A, B**) The proportions of various transposon family polymorphisms in distinct Celera mouse strains are represented in pie charts. The proportions are very similar regardless of which strain genome is considered as the reference genome. By aligning 26 million reads from Celera shotgun sequencing (Akagi et al. 2008), we identified (**A**) 12,305 polymorphic transposon insertions that are present in the C57BL6/J reference strain and absent from at least one of the other Celera strains (129S1/SVIMJ, 129X1/SVJ, A/J, and DBA/2J), and (**B**) 18,594 transposon insertions present in at least one of the four Celera strains and absent from the B6 reference. *Legend*: various transposon families as defined by RepBase. (**C**) Chromosomal distributions of reference and polymorphic LTR (ERV; *left, green histograms*) vs. L1 (*right, blue histograms*) integrants, displayed as the number of integrants per 500 kb for all autosomes and the X chromosome (*scale bars, bottom*). B6 reference genome (*dark histograms, x-axis*) and cumulative polymorphic insertions in other Celera strains (*light histograms, x-axis*) are plotted as histograms. *Center, G + C grayscale*: percentage of G+C nucleotide content in 500 kb chromosomal windows.

Supp. Fig. 3. **Significant changes in truncated and non-terminated *Slc15a2* transcripts in strains harboring ERV$_{Slc15a2}$.** (**A**) Loading controls for total RNAs in Northern blot (**Fig. 3**). *Top*: 28S; *bottom*: 18S rRNA bands for each sample as indicated. (**B**) Quantitative RT-PCR was performed using primers to detect non-terminated (*light gray*) and upstream (*dark gray*) *Slc15a2* transcripts in pooled total RNAs from brains of different mouse strains. Results were normalized using a standard curve, generated using dilutions of known concentrations of *Slc15a2* cDNA template, and are displayed as total pg *Slc15a2* transcripts (x 10^{-4}). Differences between upstream and non-terminated transcript levels were calculated to quantify prematurely truncated transcripts (*black*). When the difference was less than 0, we arbitrarily set it as 0 (i.e. A/J mice). *Error bars*: range of duplicates. Total RNAs from adults (day 72) from indicated mouse strains A/J (*red*) or B6 (*blue*) were assayed using Affymetrix mouse exon microarrays. Log-scale signal intensities (*y-axis*) are displayed for each exon in *Slc15a2* (schematic, top,

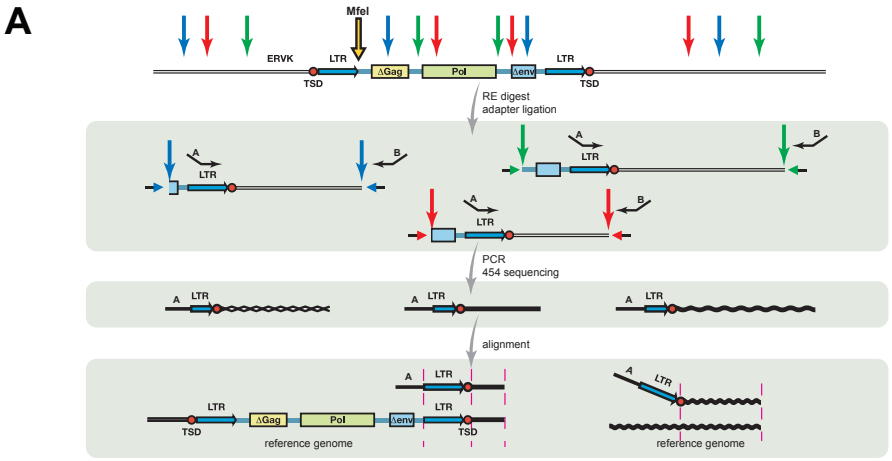
with probes in each exon, *x-axis*). (**C**) Total RNAs from adults (day 72) from indicated mouse strains A/J (*red*) or B6 (*blue*) were assayed using Affymetrix mouse exon microarrays. Log-scale signal intensities (*y-axis*) are displayed for each exon in *Slc15a2* (schematic, top, with probes in each exon, *x-axis*). (**D**) To assess for allele-specific expression differences in truncated or non-terminated transcripts, total RNAs from heterozygous CASTxC57 and C57xCAST F1 hybrid mice, and homozygous CAST mice were compared. Gene-specific, spliced transcripts were amplified by RT-PCR at *Slc15a2*, and either sequenced in bulk (*bottom left*) or cloned into Topo plasmid vector followed by sequencing of individual colonies (*bottom right*). The results indicate that both alleles are expressed in heterozygous mice, both in truncated and non-terminated transcripts. *Top*: schematics of genes indicating exon and primer locations, PCR amplicons *T* (truncated) and *N* (non-terminated), and presence or absence of ERV integrants. *Bottom*: sequence chromatograms. *Asterisks*: chromosomal and sequence location of SNP distinguishing between strains in *Slc15a2*, chr. 16: 36772531 (mm9). *Arrows*: orientation of reading frame relative to strand that was sequenced.

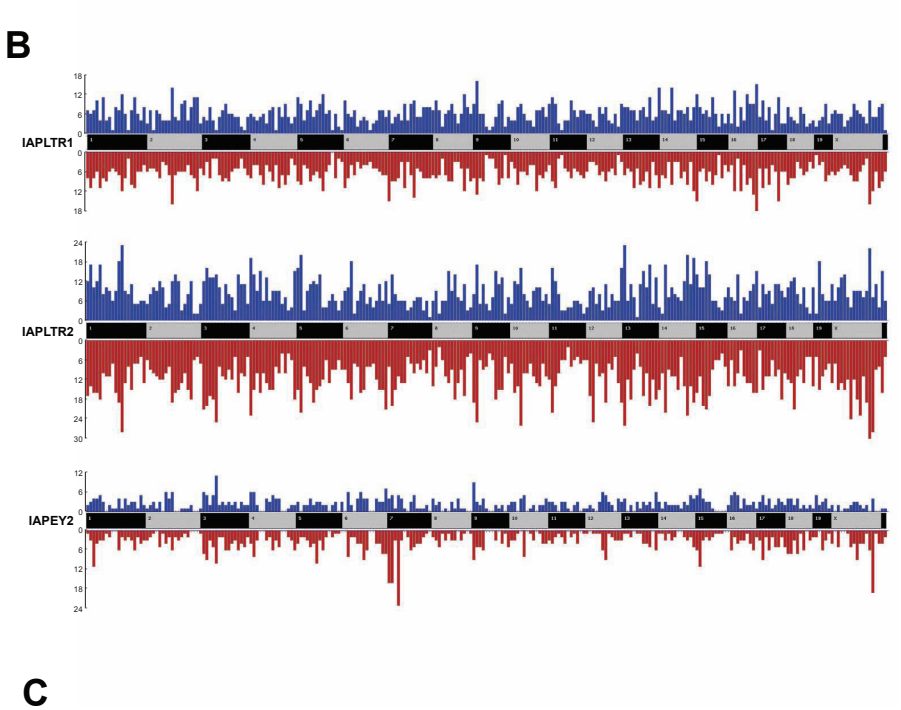
Supp. Fig. 4. **Verification of ERV$_{Slc15a2}$ status in BxD RI mice.** PCR assays for (*top*) left junction; (*middle*) right junction and (*bottom*) empty target sites at ERV$_{Slc15a2}$ integration site were performed using sequence specific PCR primers and genomic DNA samples from indicated mouse lineages. *Blue headers*: discrepant mouse lineages, whose DBA-like genotypes (ERV$_{Slc15a2}^{-/-}$) are consistent with DBA-like transcript expression levels and refute the B6-like genotypes determined using surrogate SNP at rs4173858 (**Fig. 5**).

Supp. Fig. 5. **Transcriptional disruption is strongly associated with ERV genotype at *Slc15a2*, independent of genetic background.** Individual F1 and F2 offspring of various crosses between B6 and CAST-derived mice were genotyped for the presence or absence of ERV$_{Slc15a2}$ (**Supp. Table 3**). Total RNAs were extracted from day 72 brains, and qRT-PCR assays were performed to quantify both prematurely truncated and non-terminated *Slc15a2* transcripts. To standardize input RNA levels, results were normalized to *Hprt* transcript levels. Individual results are shown in **Fig. 6D**. Results are presented as box plots for (**A**) ratio of truncated to non-terminated transcripts; (**B**) truncated; (**C**) non-terminated transcripts: *thick black line*, median; *red box*: 25th and 75th

percentiles; *whiskers*: standard adjacent values; and *small black dots*: outside values. Statistical significance for pairwise comparisons was adjusted using Bonferroni correction. *Brackets*: p-values for certain comparisons: (**A**) all pairwise comparisons are highly significantly different except for that between ERV+/- and ERV-/- ($p = 0.057$); (**B**) all comparisons are highly significantly different except for that between ERV+/- and ERV-/+ ($p = 1$), ERV+/+ and ERV-/+ ($p = 0.12$), and between ERV+/+ and ERV+/- ($p = 0.25$); (**C**) all comparisons are highly significantly different except for that between ERV-/+ and ERV-/- ($p = 1$).

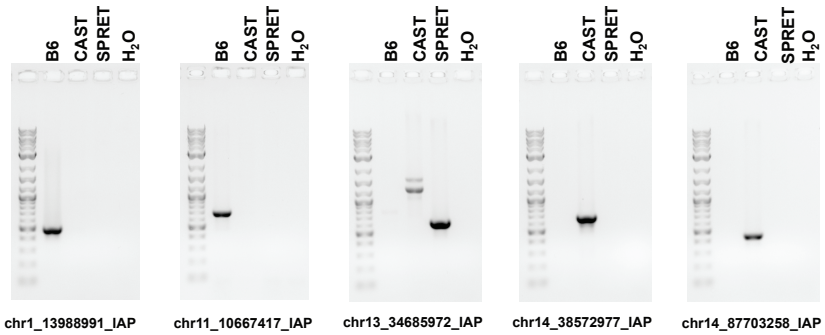
Supp. Fig. 6. **Disrupted expression of *Polr1a* and *Spon1*.** (**A**) Total RNAs from adult brains from indicated mouse strains A/J (*red*) or DBA/2J (*blue*) were assayed using Affymetrix mouse exon microarrays. Log-scale signal intensities (*y-axis*) are displayed for each exon in *Polr1a* (schematic, top, with probes in each exon, *x-axis*). (**B-C**) To assess for allele-specific or biallelic expression in various total RNAs were assessed in heterozygous CASTxC57 and C57xCAST F1 hybrid mice, and as a control in homozygous CAST mice. Gene-specific, spliced transcripts were amplified by RT-PCR at (**B**) *Polr1a* and (**C**) *Spon1* and sequenced in bulk, so both alleles can be visualized simultaneously. *Top*: schematics of genes indicating exon and primer locations, PCR amplicons *T* (terminated) and *N* (non-terminated), and presence or absence of ERV integrants. *Bottom*: sequence chromatograms. *Asterisks*: chromosomal locations of SNPs distinguishing between strains are: (**B**) *Polr1a* termination, chr. 6: 71904935 and non-terminated, chr. 6: 71913740; and (**C**) *Spon1*, chr. 7: 121022730. The results indicate that both alleles are expressed at approximately equivalent levels in heterozygous mice, both in prematurely terminated and in non-terminated transcripts.



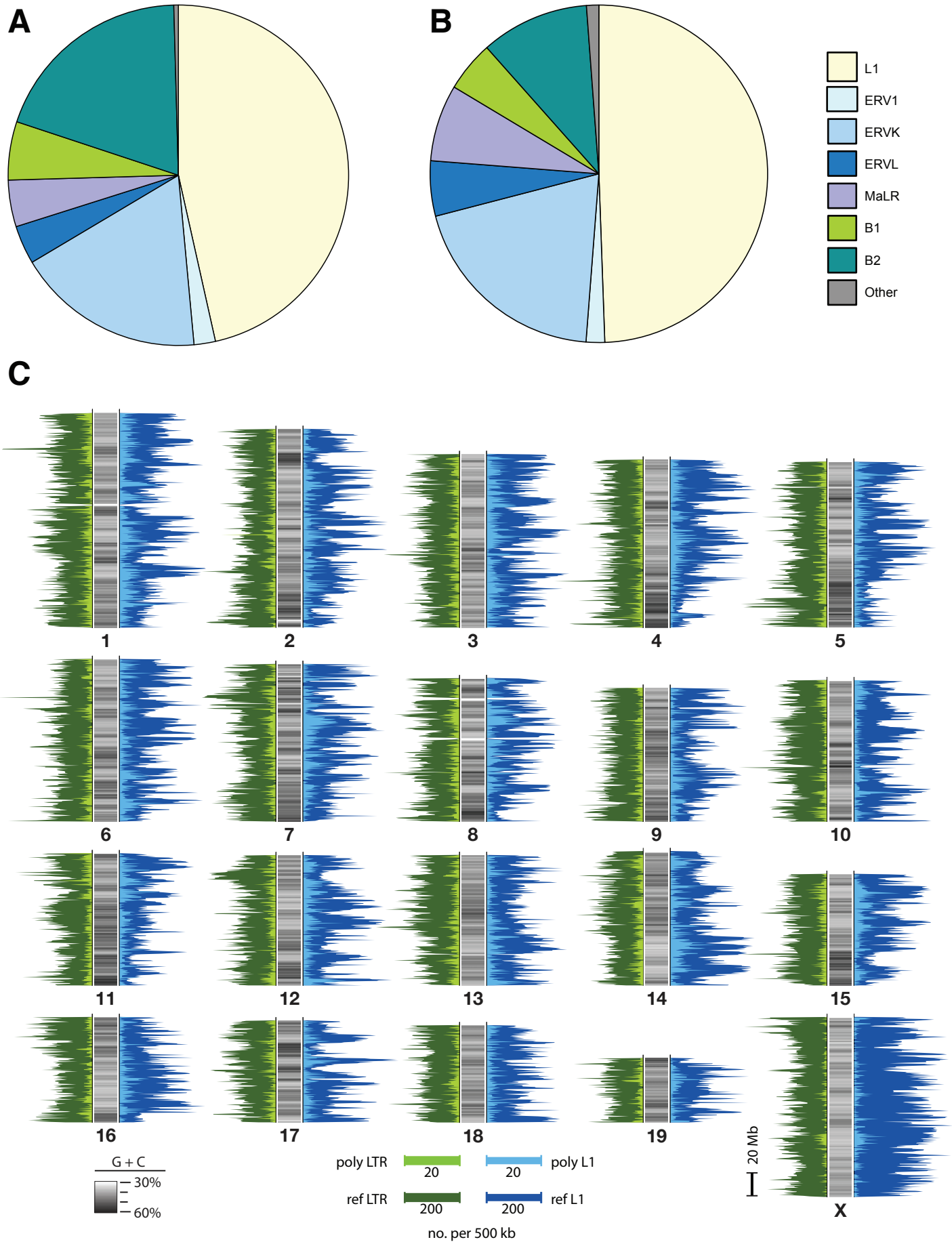


C

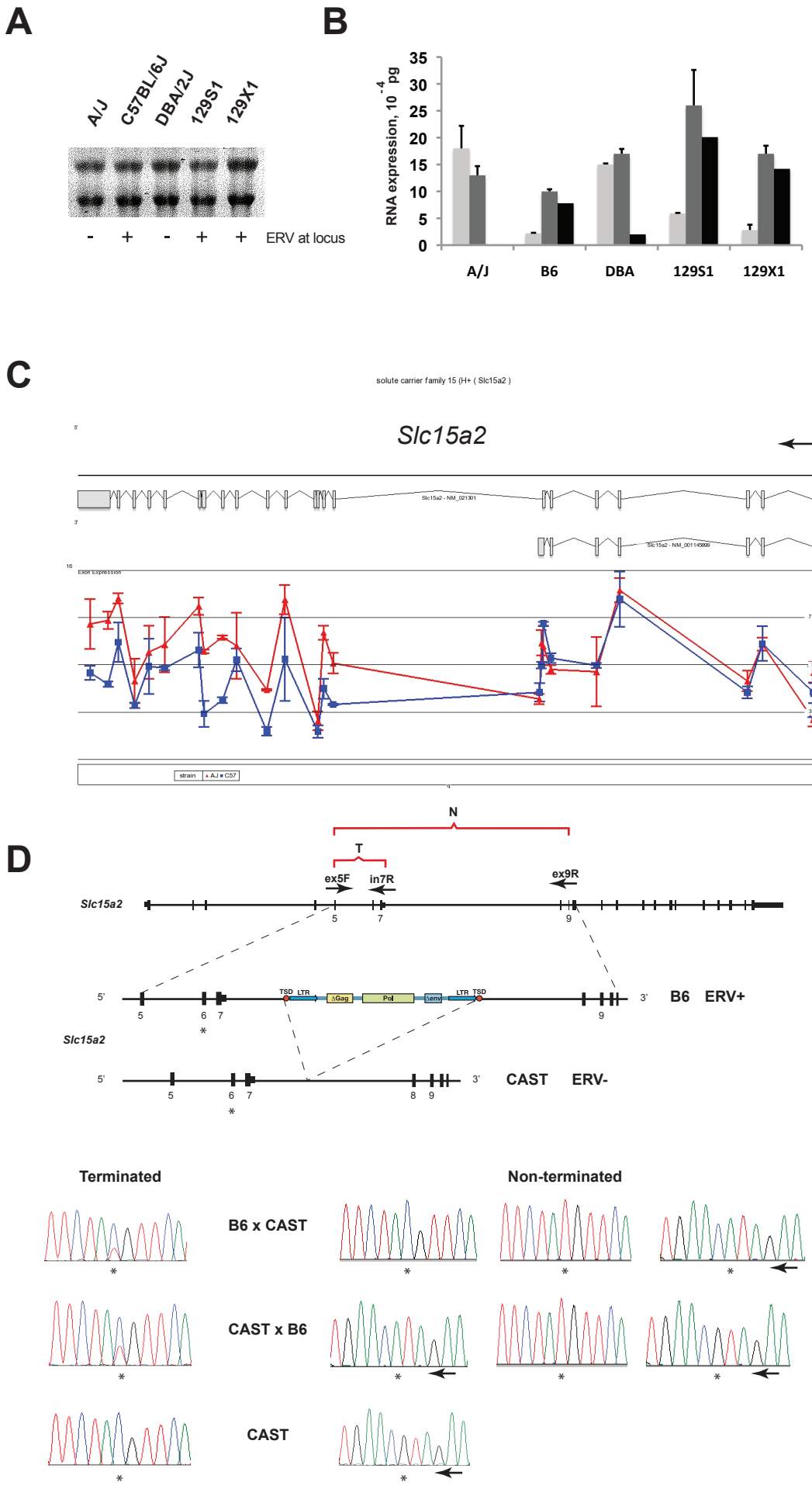
name	num_reads	addr	strand	start	stop	B6	CAST	SPRET	reference	ref_repeat	rel_ltr
chr1_13988991_IAP	1	13988991	+	13988991	13989227	1	0	0	1	IAPLTR1a_Mm	3LTR
chr11_10667417_IAP	1	10667417	+	10667417	10667615	1	0	0	1	IAPLTR1a_Mm	3LTR
chr13_34685972_IAP	128	34685972	-	34685781	34685975	0	0	128	0	ND	
chr14_38572977_IAP	11	38572977	+	38572977	38573078	0	11	0	0	ND	
chr14_87703258_IAP	5	87703258	-	87703185	87703258	0	5	0	0	ND	



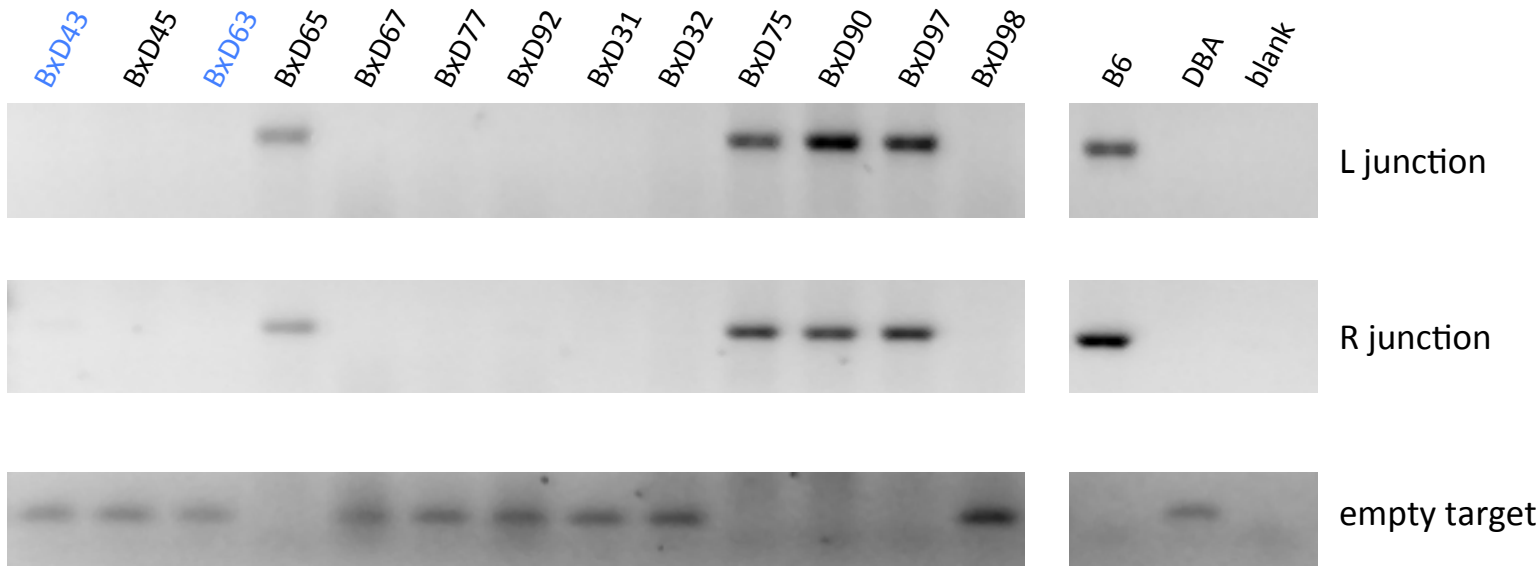
Supp. Fig. 1, Li et al.


A
B
L1
ERV1
ERVK
ERVL
MaLR
B1
B2
Other
C
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
G + C
30%
60%
poly LTR
20
20
poly L1
ref LTR
200
200
ref L1
no. per 500 kb
20 Mb

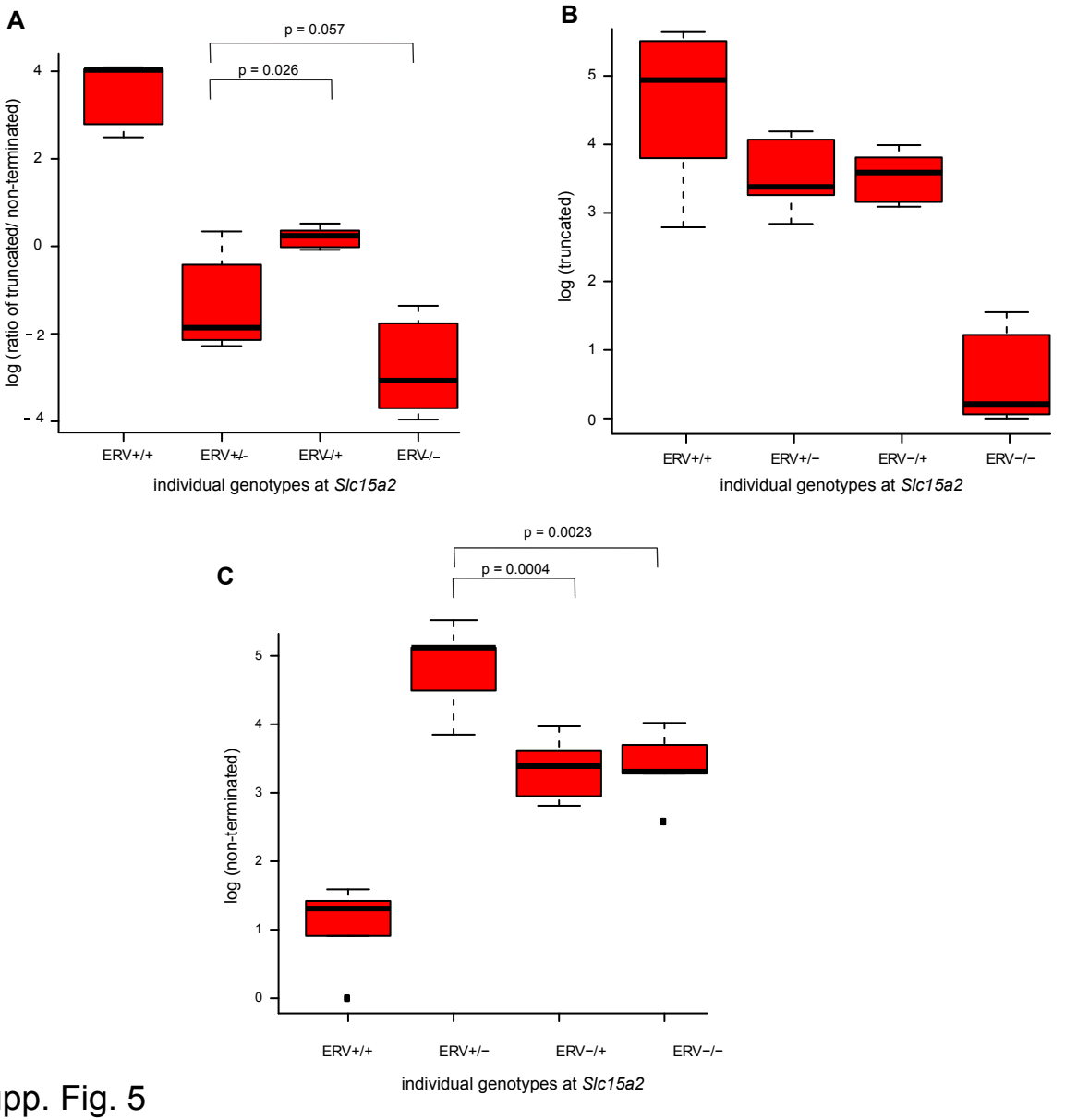
Supp. Fig. 2


A
A/J
C57BL/6J
DBA/2J
129S1
129X1
- + - + + ERV at locus
B
RNA expression, 10 -4 pg
35
30
25
20
15
10
5
0
A/J
B6
DBA
129S1
129X1
C
solute carrier family 15 (H+ (Slc15a2)
Slc15a2
Slc15a2 - NM_021301
Slc15a2 - NM_001145899
Exon Expression
strain A.J C57
D
N
T
ex5F
in7R
ex9R
Slc15a2
5
7
9
5'
TSD LTR
ΔGag
Pol
Δenv
LTR TSD
3'
B6 ERV+
5 6 7 9
*
Slc15a2
5'
3'
CAST ERV-
5 6 7 8 9
*
Terminated
Non-terminated
B6 x CAST
CAST x B6
CAST
*

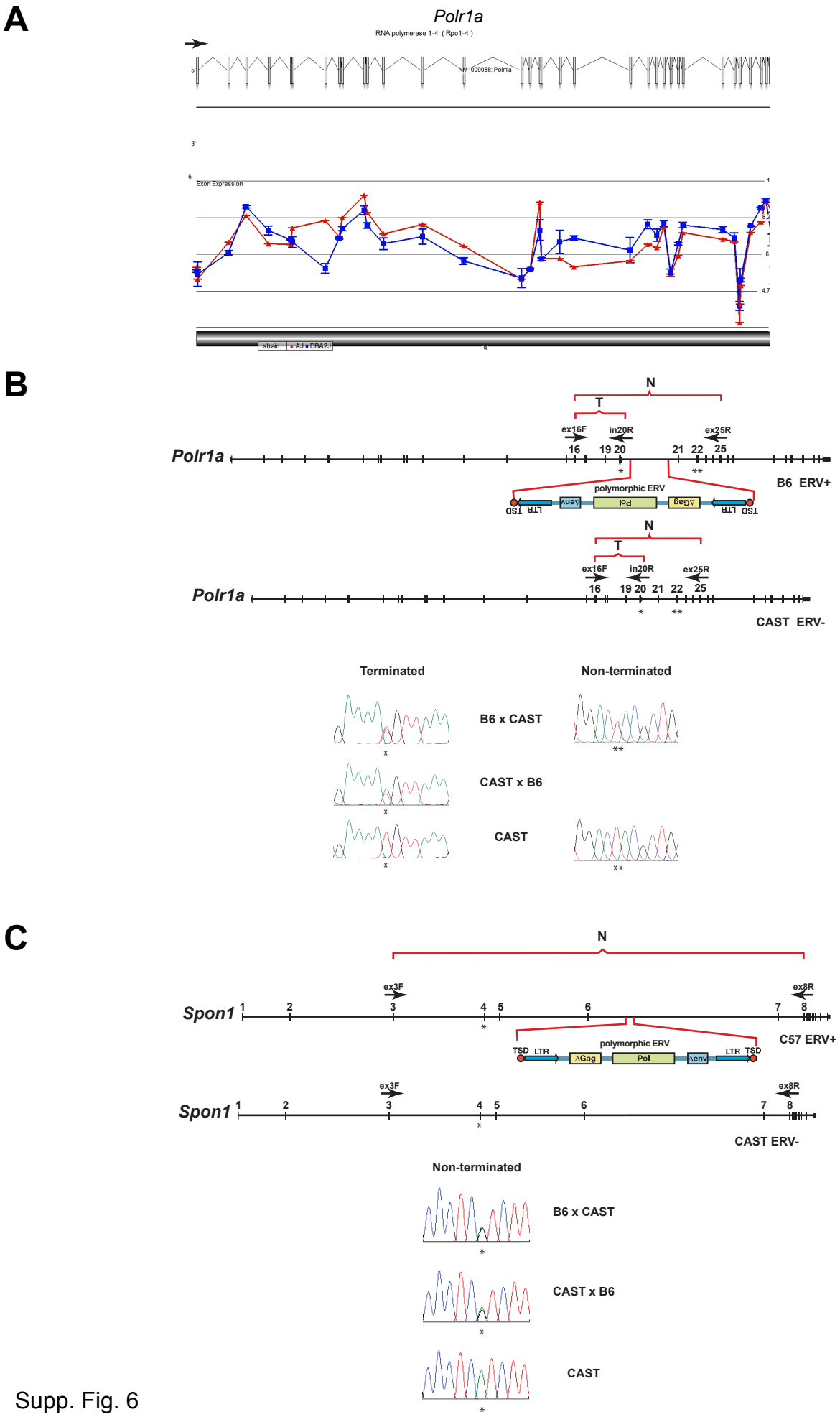
Supp. Fig. 3


BxD43
BxD45
BxD63
BxD65
BxD67
BxD77
BxD92
BxD31
BxD32
BxD75
BxD90
BxD97
BxD98
B6
DBA
blank
L junction
R junction
empty target

Supp. Fig. 4


A
p = 0.057
p = 0.026
log (ratio of truncated/ non-terminated)
4
2
0
−2
−4
ERV+/+
ERV+/−
ERV−/+
ERV−/−
individual genotypes at Slc15a2
B
log (truncated)
5
4
3
2
1
0
ERV+/+
ERV+/−
ERV−/+
ERV−/−
individual genotypes at Slc15a2
C
p = 0.0023
p = 0.0004
log (non-terminated)
5
4
3
2
1
0
ERV+/+
ERV+/−
ERV−/+
ERV−/−
individual genotypes at Slc15a2

Supp. Fig. 5


A
Polr1a
RNA polymerase 1-4 (Rpo1-4)
NM_009088: Polr1a
Exon Expression
strain
AJ
DBA2J
B
N
T
ex16F
in20R
ex25R
16
19 20
21 22 25
Polr1a
B6 ERV+
polymorphic ERV
LTR
TSD
Pol
ΔGag
Δenv
CAST ERV-
Terminated
Non-terminated
B6 x CAST
CAST x B6
CAST
C
Spon1
ex3F
ex8R
1 2 3 4 5 6 7 8
C57 ERV+
ΔGag
Pol
Δenv
Non-terminated
B6 x CAST
CAST x B6
CAST

Supp. Fig. 6

SUPPLEMENTAL TABLES

Supp. Table 1(A)

run	restriction enzyme(s)	internal ERV primer	454 reads count	reads with IAP count	reads with IAP percent	mapped reads count	mapped reads percent	loci count
1	DpnII	1	35,649	31,754	89.1%	28,144	88.6%	660
2	BfaI	1	22,176	14,187	64.0%	11,241	79.2%	672
3	BfaI/DpnII/MseI	2	33,050	25,014	75.7%	21,342	85.3%	4,131
4	BfaI/DpnII/MseI	2	57,961	41,206	71.1%	34,294	83.2%	4,600
	total		148,836	112,161	75.4%	95,021	84.7%	5,334

1(B)

ERV subfamily	no. mapped reads	no. ERVs in reference "able to be sequenced"	no. reference ERVs detected	% detected	fold coverage (mapped reads/ detected element)
IAPEY_LTR	1,382	438	104	23.7%	13.29
IAPEY2_LTR	8,524	852	623	73.1%	13.68
IAPEY3_LTR	162	443	50	11.3%	3.24
IAPLTR1	24,847	1,665	1,538	92.4%	16.16
IAPLTR2	31,298	2,967	2,321	78.2%	13.48
IAPLTR3	0	99	0	0.0%	0.00
IAPLTR4	6	323	6	1.9%	1.00
total	66,219	6,787	4,642	68.4%	14.27

1(C)

strain	no. barcoded reads	no. reads with transposon and flanking sequence	no. reads mapped to mm8 reference genome	mapping rate	no. ERVK loci (including single read loci)	fold coverage
A/J	85,932	60,865	47,290	77.7%	4,827	9.80
C57	92,011	66,220	55,636	84.0%	5,204	10.69
CAST	99,383	73,211	45,237	61.8%	3,650	12.39
MOLF	95,463	72,523	40,055	55.2%	3,867	10.36
SPRET	98,147	68,683	31,761	46.2%	2,910	10.91
WSB	94,329	67,455	46,822	69.4%	4,611	10.15

Supp. Table 1. **Optimizing identification of IAP ERV subfamily integrants in the reference genome.** (**A**) We compared identification rates of "known" ERV integrants in the reference genome, using various restriction enzymes and primers annealing within the ERV long terminal repeats (LTRs). In the first two trial runs (#1 and

2), we used single restriction enzymes and a non-degenerate primer 1, i.e. DES3356, 5'-AGATTYTGGTCTGTGGTGTTCTTC-3', which targets the long terminal repeat of young ERVs, i.e. IAPLTR1 elements. These two runs each identified less than 700 IAPLTR loci, in comparison with the B6 reference genome, which contains 1,770 IAPLTR1s. In runs #3 and 4, we prepared genomic DNA fragments separately with three restriction enzymes to reduce false negatives caused by restriction site bias. We also used degenerate primer 2, DES3355, 5'- CTTTGCCGCAGAAGRWTCYG-3' to amplify a wider range of ERV IAP subfamilies with various LTRs. Combining the third and fourth runs, 5,204 IAP loci were identified in the reference genome. (**B**) Using a combination of restriction enzymes for genomic DNA digestion and degenerate primers to detect various ERV IAP subfamilies by the transposon junction assay, we identified over 92% of possible young IAP_LTR1 elements in the reference B6 mouse genome. Overall fold coverage exceeded 14 sequence traces per detected ERV. See Supp. Methods for more detail about the primers used. (**C**) Thousands of ERV elements were identified in various diverse, previously unsequenced mouse strains at >10-fold sequence coverage. In most cases, a large majority of sequence reads from these diverse genomes could be mapped to the B6 reference genome. For example, out of 92,011 barcoded pyrosequencing traces generated from C57BL/6J mice with a single degenerate primer, 66,220 contained both transposon and flanking mappable genomic sequences, thereby allowing identification of both the ERV integrant's family and its insertion boundary for over 5,000 individual ERV loci.

Supp. Table 2

chr	coordinate	AJ	B6	CAST	MOLF	SPRET	WSB
1	6420461	2	22	0	3	0	0
2	3400785	17	58	0	32	0	0
3	146027772	0	10	0	0	0	6
4	14569445	0	18	0	26	0	0
5	20747238	49	220	0	0	0	42
6	4318314	0	95	0	0	0	0
9	85859472	0	150	0	0	0	22
10	32628688	58	224	0	0	0	31
15	3906390	30	34	0	0	0	13
16	15110111	73	273	0	0	0	27
17	52476255	5	9	0	0	0	3
18	55436483	0	26	0	0	0	18
19	32514954	47	55	0	0	0	61
X	121690432	0	25	0	0	0	0
3	72875729	0	0	8	8	0	0
4	53131105	0	0	72	0	0	0
5	70236080	0	0	0	16	0	25
6	47962182	0	0	23	24	0	0
7	76310444	10	0	0	0	0	7
8	25765694	13	0	0	0	0	1
9	104148403	0	0	0	0	30	0
10	16414015	0	0	92	0	0	0
11	12093540	0	0	0	29	0	0
12	53281304	58	0	0	0	0	0
17	36502445	0	0	74	0	0	54
18	59811943	36	0	0	0	0	22
19	37514345	23	0	0	0	0	3
X	118057415	0	0	27	0	0	0

	454	PCR	Sanger
	integrant	integrant	integrant
	empty	empty	empty
	empty	empty	integrant
	empty	N.D.	empty
	empty	N.D.	integrant

Supp. Table 2. **Identification and validation of ERV integrants in diverse mouse strains.** Using the transposon junction assay (**Supp. Fig. 1**), we mapped young polymorphic ERV integrants in various strains to chromosomal coordinates (relative to the reference B6 genome, mm8). Presented here are chromosomal coordinates for selected ERV integrants whose presence or absence was called by (*key, colors*) 454 sequencing, PCR validation and Sanger WGS sequencing (Keane et al. 2011). The results show that 454 reads from the transposon junction assay accurately call presence/absence status of tested ERV integrants. *Numbers*: counts of supporting 454 sequencing platform traces from the transposon junction assay, per strain. *N.D.*, not done due to PCR amplicon failure (likely due to genome polymorphisms).

Supp. Table 3(A)

chr: coord mm8	gene locus	strand	AJ	AKR/J	BALB/CJ	BALB/cByJ	BTBR T(+)	CAST/EiJ	C57BL/6J	C3H/HeJ	DBA/2J	FVB/NJ	KK/HiJ	MOLF/EiJ	NOD/LtJ	NZB/BlNJ	NZW/LacJ	129S1/SvImJ	129X/SvJ	PWD/PhJ	SPRET/EiJ	SWR/J	WSB/EiJ
chr1:58267894-58275017	Aox3l1 intron 20	AS	+	+	+	+	+		+	+		+	+					+	+				+
chr1:60320055-60320379	Cyp20a1 intron 8	AS	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	+
chr2:93625620-93632699	Ext2 5 prime	sense							+							+							
chr2:112349800-112354621	Aven intron 2	AS		+			+		+														+
chr3:123463783-123468157	Prss12 intron 2	AS		+			+		+		+	+	+		+		+	+	+			+	+
chr4:94305261-94311531	Tek intron 7	AS					+		+		+	+	+			+							
chr4:117681292-117688517	Jmjd2a 5 prime	sense							+													+	
chr4:120433792-120440897	Zfp69 intron2	sense							+				+	+						+			
chr6:71885312-71889893	Rpo1-4 intron 20	AS	+		+	+	+		+			+				+	+	+	+			+	
chr7:55814787-55821973	Nipa2 intron 1	AS							+														
chr7:56204435-56208793	p intron18	sense	+	+			+		+	+		+			+	+		+				+	
chr8:24016713-24023839	Slc20a2 intron5	sense							+														
chr9:121171076-121177181	Trak1 intron1	AS	+		+	+	+		+	+		+	+		+	+	+	+	+	+		+	
chr9:121255107-121262206	Trak1 intron 3	AS					+		+				+		+	+	+	+	+			+	
chr10:24375847-24381032	Enpp1 intron 1	AS							+		+												
chr11:6041513-6048633	Nudcd3 intron3	sense							+		+												
chr11:51281139-51285478	Col23a1 intron2	sense	+	+	+	+	+		+	+		+						+	+				
chr11:74356372-74361593	Garnl4 intron 2	AS							+			+											+
chr12:32613270-32617632	Prkar2b intron 2	AS							+		+	+	+		+	+	+					+	
chr12:57665867-57670188	Slc25a21 intron6	sense	+	+	+	+			+	+	+	+	+		+								
chr13:56722503-56729600	Smad5 intron 1	AS		+					+			+	+										
chr13:63166458-63173288	AK141931 intron5	AS		+					+	+													
chr13:63227285-63232431	AK141931 intron10	AS		+	+				+	+													
chr14:33212947-33218236	Mmrn2 intron 1	AS	+	+	+	+	+		+	+		+	+				+						
chr14:116061928-116068529	Gpc6 intron2	sense							+							+							
chr15:86140030-86145361	Tbc1d22a intron 8	AS	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+			+	+
chr16:32067824-32073088	Lrrc33 intron 2	AS							+									+					
chr16:36684294-36689592	Slc15a2 intron7	sense							+			+					+	+	+			+	
chr17:12127053-12131538	Map3k4 intron 1	AS		+	+	+	+		+			+			+	+	+	+	+				
chr19:42399286-42403622	Crtac1 intron 2	AS		+			+		+	+		+						+	+			+	
chrX:69194673-69201828	Nsdhl intron3	sense					+		+					+		+	+	+	+				
chrX:146918864-146925949	Phf8 intron 13	AS		+			+		+	+	+						+	+					

Supp. Table 3(B)

chr. coord.	rep_class	gene	AJ	AKR/J	BALB/cByJ	BALB/cJ	BTBR T(+)	CAST/EiJ	C57BL/6J	C3H/HeJ	DBA/2J	FVB/NJ	KK/HiJ	MOLF/EiJ	NOD/LtJ	NZB/BINJ	NZW/LacJ	129S1/SvImJ	129X/SvJ	PWD/PhJ	SPRET/EiJ	SWR/J	WSB/EiJ
chr1:9128701	IAPLTR1a_Mm	Sntg1	+		+	+						+	+	+							poly		+
chr2:79157470	IAPLTR2_Mm	Cerkl	+	+	+	+				+	+		+										
chr3:75705017	IAPLTR1_Mm	Serpini1									+				+								
chr4:123292951	IAPLTR1_Mm	Rhbdl2															+	+	+				
chr5:73814056	IAPLTR1_Mm	Dcun1d4	+		+	+		poly				+					+						
chr6:54884609	IAPLTR1a_Mm	Nod1	+	+	+	+	+			+	+	+	+		+	+	+	+	+		poly	+	
chr7:75733735	IAPLTR1a_Mm	Klhl25	+	+	+	+	+			+	+	+			+	+		+	+			+	+
chr8:84929640	IAPLTR2b	Inpp4b	+							+	+		+							poly		+	
chr9:50294821	IAPLTR2b	Bcdo2	+		+	+				+	+				+								
chr9:101947474	IAPLTR1_Mm	Ephb1	+	+	+	+				+	poly				+			+	+	+			
chr10:9355860	IAPLTR2_Mm	Samd5	+	+	+	+	+			+					+				+				+
chr11:75966043	IAPLTR1a_Mm	Vps53	+	+	+	+	+	poly		+	+	+	+		+	+		+	+	poly	poly	+	
chr11:107773658	IAPLTR1_Mm	Prkca	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+			+	
chr12:21436945	IAPLTR2_Mm	Ddef2	+	+	+	+																	
chr12:53469442	IAPLTR2_Mm	Arhgap5														+		+	+	+			
chr13:76986602	IAPLTR2_Mm	Mctp1	+	+	+	+	+	+		+	+	+	+		+	+	+	+	+			+	+
chr14:65228560	IAPLTR1_Mm	Ptk2b								+	+							+	+				
chr15:53531671	IAPLTR1_Mm	Samd12	+								+	+									poly		
chr16:34964522	IAPLTR1a_Mm	Ptplb								+	+												
chr17:24006885	IAPLTR1_Mm	Abca17	+	+			+			+	+		+		+	+	+	+	+			+	
chr19:22461274	IAPLTR1a_Mm	Trpm3	+	+	+	+	+			+	+	+						+	+			+	
chrX:73839228	IAPLTR1_Mm	Tbl1x	+	+	+	+				+	+	+	+		+	+	+	+	+			+	
Ercc3 intron3		Ercc3																+					
AK008696 intron2	AK008696					+			+				+	+	+	+	+	+	+				

Supp. Table 3(C)

name	gene	rep_class	AJ	AKR/J	BALB/cByJ	BALB/cJ	BTBRT(+)	CAST/EiJ	C57BL/6J	C3H/HeJ	DBA/2J	FVB/NJ	KK/HlJ	MOLF/EiJ	NOD/LtJ	NZB/BINJ	NZW/LacJ	129S1/SvImJ	129X/SvJ	PWD/PhJ	SPRET/EiJ	SWR/J	WSB/EiJ
chr1:37718431-37723143	Tsga10	IAPLTR1a_Mm	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+			+	+
chr2:146617817-146618285	Gm114	IAPLTR2_Mm	+	+	+	+	+		+	+	+		+	+		+	+	+	+	+		+	
chr3:9714241-9721394	Pag1	IAPLTR1a_Mm	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+			+	
chr4:86608399-86608809	Slc24a2	IAPLTR2a	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	+
chr5:5124695-5125152	Pftk1	IAPLTR2_Mm	+	+	+	+	+		+	+	+	+	+			+	+	+	+			+	+
chr6:122921491-122928070	Clec4a3	IAPLTR1a_Mm	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+			+	+
chr7:24286926-24292893	Zfp575	IAPLTR1_Mm	+	+	+	+	+		+		+	+	+			+	+	+	+			+	+
chr8:27337001-27344085	Kcnu1	IAPLTR1_Mm	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	
chr9:13504633-13508798	Mtmr2	IAPEY_LTR	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	
chr10:41640986-41641416	Armc2	IAPLTR2_Mm	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	+
chr11:54420003-54427121	Rapgef6	IAPLTR1a_Mm	+	+	+	+	+		+	+	+	+			+			+	+			+	+
chr12:81336111-81343260	Wdr22	IAPLTR1_Mm	+	+	+	+	+		+	+	+		+			+	+	+	+				+
chr13:8468880-8474087	Adarb2	IAPLTR1a_Mm	+		+	+			+	+	+	+						+	+			+	
chr14:9748652-9748988	Fhit	IAPEY2_LTR	+	+	+	+	+	+	+			+			+	+						+	+
chr15:53296582-53296682	Samd12	IAPLTR1_Mm	+	+		+	+		+	+	+	+	+		+	+	+	+	+			+	+
chr16:33766985-33772335	Itgb5	IAPLTR1_Mm	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	+
chr17:10941209-10948368	Park2	IAPLTR1_Mm	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
chr18:71755798-71756290	Dcc	IAPLTR2_Mm	+		+	+	+		+	+	+	+	+		+	+	+	+	+			+	
chr19:34197151-34204254	Ankrd22	IAPLTR1a_Mm	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	+
chrX:133035714-133042830	Il1rapl2	IAPLTR1_Mm	+	+	+	+	+		+	+	+	+	+		+	+	+	+	+			+	

Supp. Table 3. **Identification and validation of ERV integrants in diverse mouse strains.**

(**A-C**) Using PCR assays developed for individual target sites, we determined the presence (*solid pink, + symbols*) or absence (*white*) of a collection of 76 intragenic integrants identified in (**A, C**) reference or (**B**) non-reference genomes at indicated chromosomal loci (*rows*) in 21 mouse strains and species (*columns*). Of the 21 genomic DNA samples assayed here, 16 strains overlap (*underlined*) with Perlegen SNP sets (A_J, AKR_J, BALB_cJ, BALB_cByJ, BTBRT, C3H_HeJ, C57BL_6J, CAST_EiJ, DBA_2J, FVB_NJ, KK_HlJ, MOLF_EiJ, NOD_LtJ, NZB_BlNJ, NZW_LacJ, 129S1_SvImJ, 129X1_SvJ, PWD_PhJ, SPRET_EiJ, SWR_J, WSB_EiJ) (Frazer et al. 2007). *Rep_class*, ERV subfamily defined using RepeatMasker classes; *poly*, flanking genomic polymorphisms precluded PCR assay so ERV presence/absence status is unknown. (**A**) In addition to validating their presence in the reference genome, we found these 32 genomic ERVs are completely missing from orthologous target sites in SPRET/EiJ and CAST/EiJ, and mostly absent from MOLF/EiJ species. (**B**) 24 non-reference ERV integrants (not in the B6 genome) were identified in other mouse lineages as indicated. The results validated the ERVs' absence

from orthologous loci in the reference genome (B6 DNA) as expected. (**C**) In addition to validating their presence in the reference genome (B6 DNA) as expected, we found these 20 genomic ERVs are mostly conserved at orthologous positions in diverse mouse lineages but still are mostly excluded from the most divergent strains MOLF/EiJ, PWD/PhJ and SPRET/EiJ. One ERV integrant within *Park2*, on chr. 17, is conserved in all strains assayed.

Supp. Table 4

ID no.	genotype at *Slc15a2*	dam	sire	generation
B6	ERV+/+	B6	B6	pool
2716	ERV+/+	B6	CASTxB6	F2
2717	ERV+/+	B6	CASTxB6	F2
2781	ERV+/+	CASTxB6	B6	F2
2903	ERV+/+	B6	B6xCAST	F2
2573	ERV+/-	B6	CAST	F1
2574	ERV+/-	B6	CAST	F1
2714	ERV+/-	B6	B6xCAST	F2
2719	ERV+/-	B6	CASTxB6	F2
2815	ERV+/-	CASTxB6	CAST	F2
2650	ERV-/+	CAST	CASTxB6	F2
2652	ERV-/+	CAST	CASTxB6	F2
2668	ERV-/+	CAST	B6	F1
2669	ERV-/+	CAST	B6	F1
2670	ERV-/+	CAST	B6	F1
2671	ERV-/+	CAST	B6	F1
2705	ERV-/+	CAST	B6xCAST	F2
2754	ERV-/+	CAST	CASTxB6	F2
2757	ERV-/+	CAST	CASTxB6	F2
2702	ERV-/-	CAST	B6xCAST	F2
2703	ERV-/-	CAST	B6xCAST	F2
2755	ERV-/-	CAST	CASTxB6	F2
2816	ERV-/-	CASTxB6	CAST	F2
CAST	ERV-/-	CAST	CAST	pool

Supp. Table 4. **ERV genotypes of individual mice assayed for *Slc15a2* expression.** Transcript levels in F1, F2 and wildtype mice were quantified using qRT-PCR (**Fig. 6D** and **Supp. Fig. 5**). Genotypes were assayed using locus-specific PCR primers to amplify the empty or occupied $ERV_{Slc15a2}$ target site. *Genotype at Slc15a2*, ERV genotypes are described using the convention *ERVx/y*, where *x* is the maternal haplotype and *y* is the paternal.

gene_name	gb_acc	chr.	gene_start	gene_stop	ERV_start	ERV_stop	rep_name	poly	dist.	orient.
1200014J11Rik	AK163780	11	72864064	72874857	72879658	72880061	IAPLTR2a	U	4801	sense
2010015L04Rik	AK008257	4	154252989	154259787	154265481	154265848	IAPLTR1_Mm	U	5694	sense
2010111I01Rik	AK008397	13	63161067	63225106	63227284	63232434	IAPLTR1_Mm	Y	2178	AS
2510009E07Rik	AK053253	16	21605836	21608592	21597627	21598173	IAPLTR2_Mm	U	7663	AS
3300002I08Rik	AK013537	2	150035691	150054185	150032643	150032972	IAPLTR2b	N	2719	AS
4930430F08Rik	AK085292	10	100016189	100018936	100008665	100013442	IAPLTR1a_Mm	U	2747	AS
6720457D02Rik	AK030471	13	62564232	62568213	62558170	62560723	IAPLTR2_Mm	N	3509	AS
Abcg3	AK037614	5	105214356	105222934	105211401	105211692	IAPEY3_LTR	U	2664	AS
Acly	AK078680	11	100311043	100312521	100310288	100310716	IAPLTR2a	U	327	AS
Adamts3	AK031900	5	90909430	90958533	90905920	90906342	IAPLTR2_Mm	Y	3088	AS
Agbl4	AK016502	4	109895663	111064942	111069681	111075088	IAPLTR1_Mm	Y	4739	sense
Akr1c14	AK086887	13	4058817	4060269	4065480	4069155	IAPLTR1_Mm	U	5211	AS
Angpt1	AK012027	15	42326300	42506987	42323062	42323394	IAPEY3_LTR	U	2906	sense
Arfgef2	AK087574	2	166496781	166500741	166503404	166508674	IAPLTR1_Mm	U	2663	AS
Asb3	AK038897	11	30854447	30929760	30930229	30930744	IAPLTR2_Mm	U	469	AS
Atp6v1h	AK081492	1	5073244	5089858	5097588	5099400	IAPLTR2b	N	7730	AS
Bmx	AK080038	X	159564486	159602297	159558699	159559144	IAPLTR2_Mm	U	5342	AS
Ccdc15	BC032924	9	37091081	37098059	37081040	37088148	IAPLTR1_Mm	U	2933	AS
Ccdc46	AK160317	11	108241366	108288883	108291117	108291609	IAPLTR2_Mm	Y	2234	sense
Cdk5rap1	AK136570	2	154052691	154064172	154045704	154051034	IAPLTR1_Mm	Y	1657	AS
Cenpq	AK088680	17	40393497	40396699	40389137	40389464	IAPEY_LTR	U	4033	sense
Chl1	AK082640	6	103476734	103610838	103613692	103614151	IAPLTR2_Mm	U	2854	sense
Cmah	AK145110	13	24424192	24478028	24478076	24478394	IAPLTR2b	U	48	AS
Cog6	AK048986	3	53083728	53105139	53082224	53082600	IAPLTR2_Mm	U	1128	AS
Col4a4	AK142553	1	82388196	82465748	82380063	82385013	IAPLTR1a_Mm	N	3183	AS
Ctps2	AK161564	X	158245342	158322634	158322738	158329862	IAPLTR1_Mm	U	104	AS
Cyp20a1	AK020848	1	60287868	60317253	60320054	60320379	IAPLTR2b	N	2801	AS
Dcp1b	AK216011	6	119141009	119166481	119168643	119169014	IAPLTR2a	U	2162	AS
Dctn4	AK089264	18	60651596	60653341	60653737	60659091	IAPLTR1_Mm	U	396	AS
Diap3	AK166225	14	85907485	85972890	85900107	85900426	IAPLTR2b	Y	7059	AS
Dnahc1	Z83815	14	30114986	30133504	30105973	30106433	IAPLTR2_Mm	U	8553	AS
Dnahc7b	AK143375	1	46089242	46123263	46125427	46125782	IAPLTR1_Mm	Y	2164	sense
Dph5	AK045548	3	115880197	115893206	115901130	115906444	IAPLTR1_Mm	U	7924	sense
Dsg2	BC020144	18	20701255	20725919	20727208	20727348	IAP-d	N	1289	AS
Dtd1	AK010822	2	144291456	144438366	144442330	144442744	IAPLTR2a	U	3964	sense
Dym	AK081242	18	75144145	75231773	75232462	75233045	IAPLTR2_Mm	U	689	AS
E130309F12Rik	AK053451	4	49080555	49323369	49329019	49329345	IAPEY3_LTR	U	5650	AS
Enpp6	AK014741	8	48567747	48582293	48582879	48588122	IAPLTR1_Mm	Y	586	AS
Epb4.1l2	AJ539144	10	25131002	25202494	25207244	25207663	IAPLTR2_Mm	N	4750	AS
Eps15	AK018619	4	108821843	108823907	108829603	108830465	IAPLTR2_Mm	U	5696	AS
Exoc6	AK082881	19	37615582	37674249	37682580	37689630	IAPLTR1a_Mm	Y	8331	AS
Fancd2	AK019136	6	113497480	113504496	113509418	113509834	IAPLTR2a	N	4922	AS
Galk2	AK053002	2	125557708	125638794	125648615	125653155	IAPEY_LTR	U	9821	AS
Galnt10	AK082774	11	57541868	57543106	57551061	57551408	IAPLTR2a	U	7955	sense
Gimap5	AK054194	6	48675884	48678212	48680492	48680874	IAPEY_LTR	N	2280	AS
Gm4979	AB470966	2	149909539	149931030	149931223	149931551	IAPLTR2b	U	193	AS
Golga3	AK171773	5	110417161	110426802	110435195	110440153	IAPLTR2a	Y	8393	AS
Gpsm1	AK192364	2	26147466	26149332	26155230	26160484	IAPLTR1_Mm	U	5898	AS
Iars2	AK043729	1	187012633	187030166	187004285	187007297	IAPLTR2_Mm	U	5336	AS
Ifi44	AK085407	3	151683078	151687333	151678041	151678507	IAPLTR2_Mm	U	4571	AS
Iqca	AK015512	1	91929185	91984572	91921096	91921535	IAPLTR2_Mm	U	7650	sense
Irak3	AK014783	10	119568035	119604647	119561872	119562209	IAPLTR3	U	5826	sense
Itgb3bp	AK088352	4	99319334	99321176	99310396	99310885	IAPLTR2_Mm	U	8449	AS
Katnal1	AK049620	5	149196074	149239349	149186846	149187178	IAPEY3_LTR	N	8896	sense
Klhl13	AK045102	X	22519152	22522022	22509895	22510337	IAPLTR2_Mm	U	8815	AS
Lama3	AK138740	18	12647734	12690582	12696874	12697194	IAPEY2_LTR	N	6292	AS
Letm1	AK185302	5	34078436	34079443	34076522	34076997	IAPLTR2_Mm	U	1439	AS
Mapk4	AK138531	18	74094101	74096823	74084587	74084957	IAPEY2_LTR	U	9144	AS
Me3	AK036221	7	89507725	89518865	89525384	89529803	IAPLTR1a_Mm	U	6519	AS
Me3	AK035780	7	89611922	89693625	89693986	89701186	IAPLTR1_Mm	Y	361	sense
Mllt10	BC069988	2	17982437	18043975	18053064	18053420	IAPLTR1_Mm	U	9089	AS
Myom1	AK041025	17	70991252	70997958	71002830	71003288	IAPLTR2_Mm	U	4872	AS
Ophn1	AK039304	X	94914917	95093744	94911223	94911608	IAPEY2_LTR	U	3309	AS
Orai2	AK195764	5	136446280	136455211	136440305	136444837	IAPLTR1a_Mm	U	1443	AS
Otoa	AK158280	7	120909532	120951916	120952564	120952901	IAPEY2_LTR	U	648	AS
Paqr3	BC050248	5	97342443	97351898	97341286	97341616	IAPEY3_LTR	U	827	AS
Parp4	AK054229	14	55529766	55601337	55608421	55608739	IAPLTR2b	U	7084	sense
Phc3	AK051918	3	31111900	31123989	31110438	31110951	IAPLTR2_Mm	U	949	sense
Phf8	AK040943	X	146861902	146913227	146918864	146925949	IAPLTR1_Mm	Y	5637	AS
Pkd1l2	AK131811	8	119895299	119910072	119888542	119893691	IAPLTR1a_Mm	U	1608	sense
Polr1a	AK031689	6	71838561	71884659	71884936	71890270	IAPLTR1_Mm	Y	277	AS
Poteg	AK015473	8	28913605	28930048	28930570	28937625	IAPLTR1_Mm	U	522	sense
Prrxl1	EU670677	14	31428416	31459362	31466759	31467209	IAPLTR2_Mm	U	7397	sense
Qrsl1	AK038801	10	43569328	43590158	43565506	43568369	IAPLTR2a	U	959	AS
Rab3gap1	AK035603	1	129696334	129709992	129719531	129724801	IAPLTR1_Mm	Y	9539	AS
Ralgapa1	AK153664	12	56680685	56739208	56678094	56678382	IAPEY3_LTR	U	2303	AS
Rhbdl2	AK162292	4	123312450	123314508	123318547	123318952	IAPLTR2_Mm	N	4039	AS
Rnf157	AK011693	11	116219936	116229122	116205495	116210800	IAPLTR1_Mm	Y	9136	AS
Sag	AY651760	1	89635062	89656471	89664063	89664494	IAPLTR2_Mm	N	7592	AS
Sema3d	AK134412	5	12389536	12514528	12523108	12527153	IAPLTR1a_Mm	U	8580	AS
Sgip1	AK203945	4	102350825	102368582	102373594	102380766	IAPLTR1_Mm	U	5012	AS
Sirt5	BC087898	13	43376484	43391648	43397766	43404850	IAPLTR1a_Mm	U	6118	sense
Slc15a2	BC018335	16	36691113	36704018	36684294	36689592	IAPLTR1_Mm	Y	1521	sense
Slc17a5	AK087395	9	78355574	78373711	78348561	78353688	IAPLTR1a_Mm	Y	1886	sense
Slc20a2	AK076380	8	24004323	24006845	24016713	24023839	IAPLTR1_Mm	Y	9868	sense
Slc25a46	BC020087	18	31724593	31752896	31723531	31723826	IAPEY3_LTR	U	767	sense
Slc38a1	AK050914	15	96412266	96470284	96408676	96409323	IAPEY3_LTR	U	2943	sense
Slc38a1	AK050914	15	96412266	96470284	96409340	96409939	IAPEY3	U	2327	sense
Slc38a1	AK050914	15	96412266	96470284	96411103	96411401	IAPEY3_LTR	U	865	sense
Slco6b1	AK016463	1	98777925	98827974	98774847	98775165	IAPLTR2b	U	2760	AS
Snx29	BC034114	16	11334285	11592913	11593592	11598886	IAPLTR1_Mm	Y	679	sense
Sypl2	AK005920	3	108353629	108354024	108346630	108346758	IAPEz	U	6871	AS
Tbc1d22a	AK178601	15	86119924	86139607	86140030	86145361	IAPLTR1_Mm	Y	423	AS
Tcfcp2	AK076444	15	100374099	100379808	100370544	100373739	IAPEY2_LTR	U	360	AS
Tmco3	AK177090	8	13303785	13308376	13317387	13317707	IAPLTR3	U	9011	sense
Tmed7	AK180182	18	46718702	46722573	46715367	46715766	IAPLTR2a	N	2936	AS
Trpm2	AK039405	10	77365918	77372926	77353859	77361014	IAPLTR1_Mm	U	4904	AS
Txndc11	AK156590	16	11027666	11048191	11019642	11020199	IAPLTR3	U	7467	AS
Uty	AK043154	Y	508989	582181	508356	508824	IAPLTR2_Mm	U	165	AS
Vps52	AK178157	17	33573188	33573472	33574867	33575206	IAPLTR1a_Mm	U	1395	AS
Whsc1	AK031931	5	34191611	34194139	34200255	34200686	IAPLTR2_Mm	N	6116	AS
Zfand3	EF437371	17	29862373	29862507	29869271	29869657	IAPLTR2a	N	6764	AS
Zfp407	AK132276	18	84414012	84546613	84410958	84411060	IAPEz	U	2952	sense

Supp. Table 5. **Identification of candidate transcripts truncated prematurely by ERV integrants.** A bioinformatics screen was conducted to identify candidate transcripts that may be disrupted by ERV integrants within 10 kb genomic distance from premature termination sites. This analysis focused on reference B6 genes, because currently available mouse transcriptome data are limited mostly to that lineage. Presented here are gene names; *gb_acc*, truncated transcript GenBank accession numbers; chromosomal coordinates for the truncated gene and the ERV; the ERV subfamily (defined by RepeatMasker); *poly*, ERV polymorphism status at orthologous position in four other Celera strains compared with presence in reference genome (*Y*, polymorphic; *N,* non-polymorphic; *U*, unknown due to partial sequence coverage at locus); *dist.*, genomic distance from ERV to prematurely terminated transcript end; and *orient.,* the orientation of the ERV relative to the gene's coding strand. *Slc15a2* (**Fig. 3 - 7**), *Polr1a* (**Fig. 8**) and *Cdk5rap1* (Druker et al. 2004) were identified in this screen. *Spon1* (**Fig. 8**) is not listed here, although its transcription is disrupted by an intronic ERV (Fig. 8), because the distance from the ERV to its premature termination site is > 10 kb, exceeding the arbitrary cutoff used here.