Supplemental Material

Title:

Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes

Running title:

Gene remodeling by LTRs

Authors:

Vedran Franke[2,6], Sravya Ganesh[1], Rosa Karlic[2], Radek Malik[1], Josef Pasulka[1], Filip Horvat[2], Maja Kuzman[2], Helena Fulka[1], Marketa Cernohorska[1], Jana Urbanova[1], Eliska Svobodova[1], Jun Ma[5], Yutaka Suzuki[3], Fugaku Aoki[4], Richard M. Schultz[5,7], Kristian Vlahovicek[2], and Petr Svoboda[1]

Keywords:

LTR, retrotransposon, oocyte, zygote, lncRNA, gene expression

Affiliations:

[1] Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Videnska 1083, 142 20 Prague 4, Czech Republic

[2] Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, Zagreb, Croatia

[3] Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

[4] Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

[5] Department of Biology, University of Pennsylvania, Philadelphia, 19104 USA

[6] Current address: Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Germany

[7] Current address: Department of Anatomy, Physiology and Cell Biology, School of Veterinary Medicine, University of California, Davis, CA 95616, USA

Correspondence to:

Petr Svoboda, Institute of Molecular Genetics ASCR, Videnska 1083, 142 20 Prague 4, Czech Republic, tel. # +420 241063147, e-mail: svobodap@img.cas.cz

Kristian Vlahovicek, Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, Zagreb University, Horvatovac 102a, Zagreb, Croatia, tel. # +385 1 4606306, e-mail: kristian@bioinfo.hr

Content

Supplemental Tables S1-S7 are provided as separate files:

Supplemental_Table_S1.xls	Class III (ERVL) RepeatMasker data
Supplemental_Table_S2.xls	Annotation of class III (ERVL) LTR exaptations in *Mus musculus*
Supplemental_Table_S3.xls	NGS data collection
Supplemental_Table_S4.xls	Primer table
Supplemental_Table_S5.xls	Annotation of class III (ERVL) LTR exaptations in *Mesocricetus auratus*
Supplemental_Table_S6.xls	Annotation of class III (ERVL) LTR exaptations in *Homo sapiens*
Supplemental_Table_S7.xls	Annotation of class III (ERVL) LTR exaptations in *Bos taurus*
Supplemental_Table_S8.xls	Recycled mouse pseudogenes

Supplemental File

Supplemental_File_S1.rar	archive of R scripts used to generate data presented in figures

Supplemental Methods

Bioinformatic analyses

Below is detailed description of bioinformatics analyses. Relevant R scripts are provided in the file archive Supplemental_File_S1.rar.

ERVL LTR sequence family classification

To produce family classification and examine RepeatMasker reliability of ERVL LTR classification and similarity among different LTR groups, we analyzed MaLR and MT2 LTR sequences using random forest walk. We first randomly sampled 200 instances per each of the LTR classes defined from the RepeatMasker reference sequences as follows:

```
comboClasses <- c(

"MLT1" = c("MLT1A", "MLT1A0", "MLT1A1", "MLT1B", "MLT1C", "MLT1D", "MLT1E", "MLT1E1",
"MLT1E1A", "MLT1E2", "MLT1E3", "MLT1F", "MLT1F1", "MLT1F2", "MLT1G", "MLT1G1",
"MLT1G3", "MLT1H", "MLT1H1", "MLT1H2", "MLT1I", "MLT1J", "MLT1J1", "MLT1J2", "MLT1K",
"MLT1L", "MLT1M, "MLT1N2", "MLT1O"),
"MLT2" = c("MLT2B1", "MLT2B2", "MLT2B3", "MLT2B4", "MLT2B5", "MLT2C1", "MLT2C2",
"MLT2D", "MLT2E", "MLT2F"),
"MT2"   = c("MT2_Mm"),
"MT2A"  = c("MT2A"),
"MT2B"  = c("MT2B", "MT2B1", "MT2B2"),
"MT2C"  = c("MT2C_Mm"),
"MTA"   = c("MTA_Mm"),
"MTB"   = c("MTB","MTB_Mm"),
"MTC"   = c("MTC"),
"MTD"   = c("MTD"),
"MTE"   = c("MTEa", "MTEb"),
"MTE2"  = c("MTE2a", "MTE2b"),
"ORR1A" = c("ORR1A0", "ORR1A1", "ORR1A2", "ORR1A3", "ORR1A4"),
"ORR1B" = c("ORR1B1", "ORR1B2"),
"ORR1C" = c("ORR1C1", "ORR1C2"),
"ORR1D" = c("ORR1D1", "ORR1D2"),
"ORR1E" = c("ORR1E"),
"ORR1F" = c("ORR1F"),
"ORR1G" = c("ORR1G")
```

We collected full-length or nearly full-length LTR sequences (for each group defined as the length of the RepeatMasker prototype sequence ±5%). At this scale, ERVL LTR sequence divergence was not optimal for comparing all ERVL LTR groups using multiple sequence alignment. Therefore, the 3,800 selected LTRs were described and compared based on hexamer frequencies. The hexamer frequencies were calculated for each sequence and used as a basis for classification validation with random forest classifier. The confusion matrix yielded by the analysis was visualized as a heatmap where rows are observed classes and columns predicted, i.e. each cell in a row reveals for a particular LTR type the number of instances where random forest walk analysis classified it as a different (off-diagonal) or original (diagonal) LTR type.

Classification of MTA, ORR1A0 and MT2 LTR insertions in the mouse genome

Data for analysis of LTR insertions were downloaded from RepeatMasker Viz. mm10 track in the UCSC Genome Browser (http://genome.ucsc.edu/, (Kent et al. 2002)) on 27/05/2016. For each of the repeats (MT2 = MT2_Mm, MTA = MTA_Mm, and ORR1A0), RepeatMasker Viz. inserts ("joined elements") were sorted into four categories: solo LTR, full element, pseudoelement, and not classified.

Solo LTRs form upon homologous recombination, which recombines out the internal sequence. Accordingly, as solo LTR were classified all RepeatMasker inserts that consisted of a single annotated LTR sequence that were in the range of the consensus length ± 5%. Full elements consist of two complete LTRs flanking an internal sequence (int). Full elements were classified insertions, which 1) had 3' and 5' LTR fragments of full length, 2) all remaining fragments in the insertion were derived from int fragments, and 3) the total length of the insert (2xLTR+int) was in the range of the insert consensus length ± 5%.

Pseudoelements are created by LINE-1 retrotransposition factors, which directly reverse transcribe and integrate an element's transcript (i.e., without restoring full LTR sequences). Therefore, pseudoelements have distinctly modified LTRs. Their 5' LTRs of lack the promoter sequence whereas 3' LTRs are truncated after the polyadenylation signal, which is followed by $(A)_n$ motif. Disruption of reverse transcription may result in partial integration, in which case the insert has 3' pseudoelement features but truncated at the 5' end. We filtered candidate pseudoelements by manually reviewing all RepeatMasker Viz. inserts composed of an int sequence (of any length) followed by a corresponding truncated 3' LTR (316-376 bp for MTA, 266-326 bp for ORR1A0, and 413-473 bp for MT2). We included in pseudoelement categories all instances where an element was properly truncated after a predicted the poly(A) signal, which was followed by an $(A)_n$ motif.

"No classified" inserts category then includes all remaining inserts that did not fit any of the three categories mentioned above (18% of MTA inserts, 26% or ORR1A0 inserts, and 42% of MT2 inserts). These inserts represent various fragments scattered throughout the genome. In case of MT2, this category includes an unknown number of full inserts that were eroded by insertions of other sequences such that RepeatMasker Viz. did not recognize them as a single insert.

Nucleotide substitution rate analysis

Each set of the analyzed 19 ERVL groups consisted of 200 randomly selected sequences. We first performed a multiple alignment of each set of 200 sequences and extracted a matrix of substitution rate values; a substitution rate between two aligned sequences was defined as a fraction of mismatches, all insertions/deletions were omitted. Hierarchical clustering by the average was applied for the matrix and only the closest distances for each sequence were extracted and plotted as the final boxplot.

MaLR and MT2 sequence divergence analysis.

The relationship among MaLR and MT2 LTR sequences was visualized using a dimensionality reduction method - t-SNE (van der Maaten and Hinton 2008) on their corresponding k-mer content. The rationale for this strategy were difficulties with multiple sequence alignment of a large LTR sample containing ancestral and derived LTR subfamilies in MLT, ORR1, MT, and MT2 families. While Dfam database (http://dfam.org/, (Hubley et al. 2016) reveals significant similarity across the MLT, ORR1, MT, and MT2 families, especially for the ancestral subfamilies found across rodents (e.g., MTE has significant similarities to all other MTs as well as to ORR1A-G, MLT2A-D, MT2A and B), producing a common alignment of sequences from all MaLR and MT2 LTRs is virtually impossible. Therefore we opted for an alignment-free method and analyzed MaLR and MT2 sequence divergence through phylogenetic foot-printing. For each LTR, the frequency of each 6-mer was counted in a window sliding by 1 nucleotide. Reverse complement 6-mers were counted together. The 6-mer content for each LTR was normalized relative to the geometric mean in the following way: log2(kmer frequency/geometric mean (kmer frequency)). The dimensionality of the 6-mer space was reduced to 3 dimensions using t-SNE, with perplexity parameter set to 30.

Analysis of MaLTR and MT2 LTRs splice site sequence logo

Sequences of 20 nt around functional splicing sites in co-opted LTRs (Table S2) were aligned using Clustal Omega (Sievers et al. 2011) with 100 combined guide tree/HMM iterations to get a better alignments (--iter=100). Sequence logos were produced using WebLogo 2.8 software (Crooks et al. 2004)

Phylogenetic analysis of MT LTR sequences

Five thousand randomly selected LTRs and 773 5'exon LTRs from MT family were aligned using ClustalO software, version 1.2.3, with the default parameters. The tree was constructed using the FastTree software, version 2.1.9, (with the following parameters: -gamma -nt -gtr) that constructs an approximate maximum likelihood tree from a large number of sequences.

Mapping of NGS data

All data in the report are based on the mm10/NCBI38 genome version, which includes our previously published data (Abe et al. 2015; Karlic et al. 2017) that were mapped onto mm9/NCBI37 mouse genome version. Here, they were remapped onto the mm10/NCBI38 genome version using the STAR mapper (Dobin et al. 2013) and the genome index was constructed with the addition of the mm10 Ensembl gene annotation, downloaded on 09/02/2016 from the Ensembl database.

All remaining Illumina RNA-seq data were analyzed as follows. First, we filtered and removed the adapters from using the Trimmomatic software (Bolger et al. 2014), with the following parameters:

```
ILLUMINACLIP: TruSeq2-PE.fa:2:30:10 TRAILING:20 MINLEN:36.
```

The filtered reads were mapped onto the corresponding genomes using the STAR mapper (Dobin et al. 2013):

```
STAR –_GenomeIndex--readFilesIn $file1 $file2 --runThreadN 20 --genomeLoad LoadAndKeep
--outFilterMultimapNmax 10 --outFileNamePrefix $filename --outReadsUnmapped Fastx --
outFilterMismatchNoverLmax 0.2 --outSAMstrandField intronMotif --sjdbScore 2
```

The following genome version were used for mapping the data: mouse - mm10/NCBI38, human - hg19, cow – bosTau7 , hamster - GCF_000349665.1_MesAur1.0.

The genome index was constructed with the addition of the gene annotation from each species: mouse - mm10/GRCm38.83 downloaded from the Ensembl database, human -hg19/GRCh37.75 annotation downloaded from the Ensembl database, cow -bostTau7 RefSeq annotation downloaded from the UCSC database, and hamster - GCF_000349665.1_MesAur1.0 annotation downloaded from the RefSeq database.

Mapping of SOLiD RNA-seq reads on the mouse genome

50SE SOLiD data (Park et al. 2013) were mapped on the mm9/NCBI37 genome version with the TopHat aligner (version 1.3.2) (Trapnell et al. 2009) as in the original publication. The most stringent set of parameters was used for mapping:

```
--initial-read-mismatches 3 --segment-mismatches 2 --segment-length 25
```

Data visualization in UCSC Genome Browser

Data were visualized in the UCSC Genome Browser by constructing bigWig tracks using the UCSC tools (Kent et al. 2010).

Annotation of retrotransposon co-options in mammals

For annotation of retrotransposon exaptation in mice, we collected existing transcript annotations from the Ensembl mm10 database, and combined them with lncRNA transcript models generated with Scripture (Guttman et al. 2010) from NGS data from mouse oocytes and early embryos (Smallwood et al. 2011; Abe et al. 2015; Veselovska et al. 2015; Karlic et al. 2017). Additionally, we generated an early embryonic transcriptome using Stringtie (Pertea et al. 2015). LTR sequence exaptation was classified into four categories with a home-made script according to the LTR sequence overlap with exons. In case of non-LTR retrotransposons, the entire sequence was used. Retrotransposon annotation was based on the RepeatMasker track provided in the UCSC mm10 database for the UCSC Genome Browser.

I - 5’ exon contribution (retrotransposon-derived promoter, transcription start site, and/or splice

II -internal exon contribution (retrotransposon-derived splice donor and/or acceptor),

III - 3’ exon contribution (retrotransposon-derived splice acceptor and/or polyA site)

IV - transcript tagging where a transcript contains a retrotransposon sequence that does not contribute to mRNA formation.

Categories I-III were further divided based on whether a retrotransposon made a full or partial contribution to a given exon type – a full contribution designates that the retrotransposon overlaps both exon borders (i.e. both the TSS and the splicing donor). Classification criteria included NGS support by at least ten sequence reads from the overlapping exon in at least one of the NGS samples. For classes I-III, there was an additional requirement that the retrotransposon overlaps at least two spliced reads in at least two samples.

The human and bovine annotations of retrotransposon co-options were made analogically for human hg19 and bovine bosTau7 genome assemblies and oocyte and early expression data (Xue et al. 2013; Graf et al. 2014).

For the golden hamster (*Mesocricetus auratus*) retrotransposon co-option annotation, the current GCF_000349665.1_MesAur1.0 genome assembly was obtained from RefSeq- http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/genomes/refseq (PMID: 24259432). Both the RepeatMasker transposon annotation and the putative gene models were also obtained from RefSeq. We additionally augmented the golden hamster transcriptome by constructing a maternal transcriptome assembly using Stringtie.

For each organism we constructed a maternal and early embryonic transcriptome using Stringtie. A transcriptome was constructed for each individual stage of development, and the resulting transcriptomes were then merged, for each organism respectively. Stringtie was used with the following parameters: `-G $gtf -f 0.05 -a 3 -M 0.25` where $gtf are the known genes for the corresponding organism. Chromosomes M and Y were omitted from the assembly process. Individual transcriptomes were then merged using Stringtie merge command. Gene annotation files used for the transcriptome assembly and annotation are the following: mouse -mm10/GRCm38.83 downloaded from the Ensembl database, human: hg19/GRCh37.75 annotation downloaded from the Ensembl database, cow -bostTau7 RefSeq annotation downloaded from the UCSC database, hamster - GCF_000349665.1_MesAur1.0 annotation downloaded from the RefSeq database

Germline LTR expression heatmap

To estimate expression of LTRs in different tissues, we mapped the reads from polyA NGS datasets from different stages of early development (Table S3) to the mm10 genome using the STAR aligner and counted the number of reads, which overlapped LTR sequences. To minimize the influence of highly expressed genes, only LTRs that did not overlap any protein coding genes annotated in GENCODE Version M10 were included in the analysis. Repeats were divided into 15 MaLR and 4 MT2 repeat families as described above. A sequencing read mapped to more than one genomic sequence belonging to the same LTR family was counted only once. LTR's abundance in polyA transcriptome was calculated as reads per million mapped reads (RPM), where the number of uniquely mapped reads was used as the total library size. For experiments with multiple replicates the RPM was calculated as the mean RPM value over all replicates. For LINE-1 and IAP, all RepeatMasker annotated sequences that did not overlap any protein-coding gene were included.

Analysis of cumulative RNA expression around ERVL inserts

For the analyses of cumulative RNA expression, we used one hundred of each MuERV-L elements, MT2 solo LTRs or ORR1A0 solo LTRs with highest expression (in FPKM units). Next, we defined flanking regions of 150 kb in each direction, in which we masked annotated genes and repeat elements (knownGene and RepeatMasker tables from UCSC). Next, FPKM values of the remaining regions were calculated and multiple filtering conditions were examined to filter out local peaks generating background noise. Upon examining different cut-off values up to 10 FPKMs, we filtered out from the genomic flanks regions with FPKM > 1. This value was chosen to make a compromise between eliminating the noise while retaining signal of transcription d ownstream of inserts. It should be noted that this cut-off is an equivalent of FPKM ~4-5 of common ribozero NGS because our NGS was performed on total RNA where 75 (MII) -86% (2-cell) of the NGS data were from rRNA, which results in a proportionally lower FPKM (Abe et al. 2015). For each filtered region one-base resolution coverage was computed for GV, 1-cell, 2-cell, 2-cell with aphicolidin and 4-cell stages. Coverage values were then aligned with end coordinates of each element and summed up. Finally, the summed coverage was binned into 10 kb bins and combined FPKM values were calculated for each bin; the coverage in one bin was summed and divided by bin length in kb and library size in millions of reads.

The statistical significance of MuERV-L elements scanning was calculated by comparing binned (100 kb) cumulative profiles of the 100 top most expressed MERV-L elements (data used for Fig. 7B). The distribution of expression in each downstream transcribed bin was compared with the expression of the corresponding upstream bin using a one sided *Mann*-*Whitney U* test.

Supplemental Material - *D6Ertd527e* transcript models

Mus musculus

>transcript_A|UCSC_mm10:chr6:87102324-87102786,87110877-
87113003:+|longest_predict_pep=464|cpat_prob=0.9998125761148actgaagcaggaggtgactattacccaggagaatctg
gctttctgtggaagaggagcagctgcaagctggctactaaagtcctaggtgttggagctccctcgcagcagccagcacaggccccctccttcctctt
tgtgctcgacccttgcatgttatggatcagaggtttctttggactgactttgggaggagagcccaggacaccagctgcagtgtcctctgactgccgt
gtatctgacacctgactgacacttcctaaggctttggagatcaagagctgaaggcaagaccgagagcctcttgcacgtatttttctcagtctttgtt
ctcaccacgcttttttttttttttctggcttctagtccttttggcatccggagaatagctgaagtcagggaggctgggacagttcccagtccacag
tccattagggacccacaaagaATGTTGATGACTCTGA**GT**AGCCGCCGCCGCTGCAGCAGCAGCAGTCGGAGCAGCCTCAGCAGCCGCAGCAGTGACA
CCAGCACCAGCAGTGACACCAGCAGTGACACCAGCACCAGCACTAGCACCAGCACCAGCACCAGCCACAGCAACAACAGCAGCAGCAACAGCAGCAG
AAAACCTAGTAACAAAGGCAGCAGCAGCCTCTCGAGCAGCAGCAGCAACAGCAGCAGCAAACCTAGTGACACAGACAGCAACAGCAGCAGCATCTCT
TGCAGCAGCAACAGCCCTAGTAACACAGACAGCAGCAGCCTCTCTAGCAGCAGCAGCAACAGCAGCAGACCTAGTAACACAGGCAGCAGCAGCCTCT
CTAGCAGCAGCAGCAACAGCAGCAGACCTAGTAACACAGGCAGCAGTAGCAGCAGCAGCAGCAACAGCAGCAACATCAGCAACAGCAGCATCAGACC
TAGTAACAGGGGCAGCATCAGCAACTATGACAACAGCAGCAACAGCAGCAGTCCCCAACCCTCTAGTGGCAACATCAGCAACAGAAGACCTAGTAAC
ACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGTAACACAGTCAACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACA
CCACCACCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGACCTAGTAACAGGGGCAGCATCAGCAACTACAGCAACAGCAGCCTCAGACCTAGTAA
CAGGGGCAACATCAGCAACTATGACAACAGCAGCACCAGACCTAGTAACAGGGCCAACATCAGCAACTATGGCAACAGCAGCAATAGCAGCAGTCCC
CAACCCTCTAGCAGCAAAATCAGCAACAGAAGACCTAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGAAACACAGTCA
ACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACACCACCAGCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGATCTAGTAACACAGA
CAGCAGCAGCAGCCCTGGCAACAGCCACACCAGTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTAGCAGCAGCCGCGGCAACAGTGGC
CCCAGACCCAGTAAAACAGGCAGCATAAGCAGCCAAAGCAACAGCGGCCCCAGATCTAGTAACACAGACAGCAGCAGCAGCAGCAGCCCCGGCAATA
TCAAAACCAGCAGCAGCAGCAGCAGCAGTAACAGCAGCAGCAGCAGCCACAGCAGCTACAGCTGCAGCAGCAGCAGCCACAGCAGCAGCAGCAGCCG
CAGCCATAGCAGCAGCCACAGTCACAGCAGCCACAGTCGCACACCATGGGAATGAtttcctaagggtatactgctgagttccttcagtctttcaggc
cacatgtacttcgtaaaagattcctgtctgagcatcttccccagtggcctgtgaggatgagctcatcctttgaacttggcaggcgagaagcgttccg
atttggttcagcacaagcacagaaaggtgattgttaagtagaaattatccagtcagcaatcagacagagtcagtcctttcttccccagaagatgctg
gccaacttctggagaagtttctgctggtgggtttgcctaggcccatatttcctggctttcaacagaggctctgagatttggttagctttgagagctt
ctatggaggaaagggggcccaagagacatagactaatgcaaaaacagaaatgtaggatgaaggaaaccaaatcagaccagttactagagttcaagg
aatggagaagcaaggtttagaggggaaggagttggccggcagctgggatgagggtgtgcctgtcctgctgttattaggtgggtgtggggaccaagca
aatgacaattaaagccgtgggccagttctacacaggaactagaacaagaaagcaaacagctgggcagttatggggctggatgctcagggaaaggggc
ctagggtttaagccaaaccaaaccccaaacaaacacccattgaattggatcgatatcaaagcttttcctgtacatttttttattattgtttgtttat
ttcaataaagatttctttatttttatatgtc

>transcript_B|UCSC_mm10:chr6:87103673-87104142,87110877-
87113003:+|longest_predict_pep=227|cpat_prob=0.9086602726158ctctaatcttgcttgtcacctgttgcttgctctgcag
ggagctgccatgctgggtgctccacagactttaaaaccgagctaggaagggtgctgagtagcaagtgcttgcccagccacctagggccttagactca
gtcctgagtgcaaaactaaccagccaaccagccagtcaaccaaccagccaatcaaccaaccagccaaccaaccaaccaaccaaccaaccaaccaacc
agccagtcagccagccagctagccagtcaaccagccaaccaaccaaccagctagccagccagccagccagccagccagccagccagccaaatactga
ctttcagtttccaaacttgggatagttgtcacacaaatctatagatccctcacaggatgggaaccatatctactgattacccactctgaaaacttct
ggaaaaaacctgggatttccctaagaaattaccacctttaaaaa**gt**agccgccgccgctgcagcagcagcagtcggagcagcctcagcagccgcagc
agtgacaccagcaccagcagtgacaccagcagtgacaccagcaccagcactagcaccagcaccagcaccagccacagcaacaacagcagcagcaaca
gcagcagaaaacctagtaacaaaggcagcagcagcctctcgagcagcagcagcaacagcagcagcaaacctagtgacacagacagcaacagcagcag
catctcttgcagcagcaacagccctagtaacacagacagcagcagcctctctagcagcagcagcaacagcagcagacctagtaacacaggcagcagc
agcctctctagcagcagcagcaacagcagcagacctagtaacacaggcagcagtagcagcagcagcagcaacagcagcaacatcagcaacagcagca
tcagacctagtaacaggggcagcatcagcaactATGACAACAGCAGCAACAGCAGCAGTCCCCAACCCTCTAGTGGCAACATCAGCAACAGAAGACC
TAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGTAACACAGTCAACATCAGCAACTATAGCAACAGTGGCCCCAGACCC
AGTAACACCACCACCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGACCTAGTAACAGGGGCAGCATCAGCAACTACAGCAACAGCAGCCTCAGAC
CTAGTAACAGGGGCAACATCAGCAACTATGACAACAGCAGCACCAGACCTAGTAACAGGGCCAACATCAGCAACTATGGCAACAGCAGCAATAGCAG
CAGTCCCCAACCCTCTAGCAGCAAAATCAGCAACAGAAGACCTAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGAAAC
ACAGTCAACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACACCACCAGCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGATCTAGTA
ACACAGACAGCAGCAGCAGCCCTGGCAACAGCCACACCAGTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTAGCAGCAGCCGCGGCAA
CAGTGGCCCCAGACCCAGTAAAACAGGCAGCATAAgcagccaaagcaacagcggccccagatctagtaacacagacagcagcagcagcagcagcccc
ggcaatatcaaaaccagcagcagcagcagcagcagtaacagcagcagcagcagccacagcagctacagctgcagcagcagcagccacagcagcagca
gcagccgcagccatagcagcagccacagtcacagcagccacagtcgcacaccatgggaatgatttcctaagggtatactgctgagttccttcagtct
ttcaggccacatgtacttcgtaaaagattcctgtctgagcatcttccccagtggcctgtgaggatgagctcatcctttgaacttggcaggcgagaag
cgttccgatttggttcagcacaagcacagaaaggtgattgttaagtagaaattatccagtcagcaatcagacagagtcagtcctttcttccccagaa
gatgctggccaacttctggagaagtttctgctggtgggtttgcctaggcccatatttcctggctttcaacagaggctctgagatttggttagctttg
agagcttctatggaggaaagggggcccaagagacatagactaatgcaaaaacagaaatgtaggatgaaggaaaccaaatcagaccagttactagag
ttcaaggaatggagaagcaaggtttagaggggaaggagttggccggcagctgggatgagggtgtgcctgtcctgctgttattaggtgggtgtgggga
ccaagcaaatgacaattaaagccgtgggccagttctacacaggaactagaacaagaaagcaaacagctgggcagttatggggctggatgctcaggga
aaggggcctagggtttaagccaaaccaaaccccaaacaaacacccattgaattggatcgatatcaaagcttttcctgtacatttttttattattgtt
tgtttatttcaataaagatttctttatttttatatgtc

>transcript_C|UCSC_mm10:chr6:87104746-87104842,87110877-
87113003:+|longest_predict_pep=464|cpat_prob=0.99981913110056ggagcaagcctgtaacaagttcctccgtgggctctg
cattggttcctgcctctaggaacctcctgccctgacttctctggATGATGGACTACAAGT**TT**AGCCGCCGCCGCTGCAGCAGCAGCAGTCGGAGCAG
CCTCAGCAGCCGCAGCAGTGACACCAGCACCAGCAGTGACACCAGCAGTGACACCAGCACCAGCACTAGCACCAGCACCAGCACCAGCCACAGCAAC
AACAGCAGCAGCAACAGCAGCAGAAAACCTAGTAACAAAGGCAGCAGCAGCCTCTCGAGCAGCAGCAGCAACAGCAGCAGCAAACCTAGTGACACAG
ACAGCAACAGCAGCAGCATCTCTTGCAGCAGCAACAGCCCTAGTAACACAGACAGCAGCAGCCTCTCTAGCAGCAGCAGCAACAGCAGCAGACCTAG
TAACACAGGCAGCAGCAGCCTCTCTAGCAGCAGCAGCAACAGCAGCAGACCTAGTAACACAGGCAGCAGTAGCAGCAGCAGCAGCAACAGCAGCAAC
ATCAGCAACAGCAGCATCAGACCTAGTAACAGGGGCAGCATCAGCAACTATGACAACAGCAGCAACAGCAGCAGTCCCCAACCCTCTAGTGGCAACA

TCAGCAACAGAAGACCTAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGTAACACAGTCAACATCAGCAACTATAGCAA
CAGTGGCCCCAGACCCAGTAACACCACCACCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGACCTAGTAACAGGGGCAGCATCAGCAACTACAGC
AACAGCAGCCTCAGACCTAGTAACAGGGGCAACATCAGCAACTATGACAACAGCAGCACCAGACCTAGTAACAGGGCCAACATCAGCAACTATGGCA
ACAGCAGCAATAGCAGCAGTCCCCAACCCTCTAGCAGCAAAATCAGCAACAGAAGACCTAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGG
CCCCAGACCTGGAAACACAGTCAACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACACCACCAGCAGCAGCAACAGCCAAAGCAACAGC
AGCCCCAGATCTAGTAACACAGACAGCAGCAGCAGCCCTGGCAACAGCCACACCAGTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTA
GCAGCAGCCGCGGCAACAGTGGCCCCAGACCCAGTAAAACAGGCAGCATAAGCAGCCAAAGCAACAGCGGCCCCAGATCTAGTAACACAGACAGCAG
CAGCAGCAGCAGCCCCGGCAATATCAAAACCAGCAGCAGCAGCAGCAGCAGTAACAGCAGCAGCAGCAGCCACAGCAGCTACAGCTGCAGCAGCAGC
AGCCACAGCAGCAGCAGCAGCCGCAGCCATAGCAGCAGCCACAGTCACAGCAGCCACAGTCGCACACCATGGGAATGAtttcctaagggtatactgc
tgagttccttcagtctttcaggccacatgtacttcgtaaaagattcctgtctgagcatcttccccagtggcctgtgaggatgagctcatcctttgaa
cttggcaggcgagaagcgttccgatttggttcagcacaagcacagaaaggtgattgttaagtagaaattatccagtcagcaatcagacagagtcagt
cctttcttccccagaagatgctggccaacttctggagaagtttctgctggtgggtttgcctaggcccatatttcctggctttcaacagaggctctga
gatttggttagctttgagagcttctatggaggaaagggggcccaagagacatagactaatgcaaaaacagaaatgtaggatgaaggaaaccaaatc
agaccagttactagagttcaaggaatggagaagcaaggtttagaggggaaggagttggccggcagctgggatgagggtgtgcctgtcctgctgttat
taggtgggtgtggggaccaagcaaatgacaattaaagccgtgggccagttctacacaggaactagaacaagaaagcaaacagctgggcagttatggg
gctggatgctcagggaaaggggcctagggtttaagccaaaccaaaccccaaacaaacacccattgaattggatcgatatcaaagcttttcctgtaca
tttttttattattgtttgtttatttcaataaagatttctttatttttatatgtc

>transcript_D|UCSC_mm10:chr6:87105249-87105613,87110877-
87113003:+|longest_predict_pep=474|cpat_prob=0.99984326129421tgttgccattttgcccgtgtttttggtgatgtcagt
tgtgacatcactggagctgtggactcggggtgtctccttgggcttcttgcgacagtgggtgctggcactgtggtggtggtggtggtgtgccaagaca
ccagctgaagagctgggggcacagaatggctgagtaccatcgtagggagacctgctgctcaaggcagtgcatagcagtgttgcctccagactccggc
ttgggagatagctgtttggattcccttatgcttctctgaccttccctcacagaacggtcctctaactccgtagcacgcccaccctggaATGCATCCT
CCCACCTTTCCCCTTCAGTGTACCCTCTATGGGACAG**<u>GT</u>**AGCCGCCGCCGCTGCAGCAGCAGCAGTCGGAGCAGCCTCAGCAGCCGCAGCAGTGACA
CCAGCACCAGCAGTGACACCAGCAGTGACACCAGCACCAGCACTAGCACCAGCACCAGCACCAGCCACAGCAACAACAGCAGCAGCAACAGCAGCAG
AAAACCTAGTAACAAAGGCAGCAGCAGCCTCTCGAGCAGCAGCAGCAACAGCAGCAGCAAACCTAGTGACACAGACAGCAACAGCAGCAGCATCTCT
TGCAGCAGCAACAGCCCTAGTAACACAGACAGCAGCAGCCTCTCTAGCAGCAGCAGCAACAGCAGCAGACCTAGTAACACAGGCAGCAGCAGCCTCT
CTAGCAGCAGCAGCAACAGCAGCAGACCTAGTAACACAGGCAGCAGTAGCAGCAGCAGCAGCAACAGCAGCAACATCAGCAACAGCAGCATCAGACC
TAGTAACAGGGGCAGCATCAGCAACTATGACAACAGCAGCAACAGCAGCAGTCCCCAACCCTCTAGTGGCAACATCAGCAACAGAAGACCTAGTAAC
ACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGTAACACAGTCAACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACA
CCACCACCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGACCTAGTAACAGGGGCAGCATCAGCAACTACAGCAACAGCAGCCTCAGACCTAGTAA
CAGGGGCAACATCAGCAACTATGACAACAGCAGCACCAGACCTAGTAACAGGGCCAACATCAGCAACTATGGCAACAGCAGCAATAGCAGCAGTCCC
CAACCCTCTAGCAGCAAAATCAGCAACAGAAGACCTAGTAACACAGGCAGCAGCAGCAACCAGGTTAACAGCGGCCCCAGACCTGGAAACACAGTCA
ACATCAGCAACTATAGCAACAGTGGCCCCAGACCCAGTAACACCACCAGCAGCAGCAACAGCCAAAGCAACAGCAGCCCCAGATCTAGTAACACAGA
CAGCAGCAGCAGCCCTGGCAACAGCCACACCAGTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTAGCAGCAGCCGCGGCAACAGTGGC
CCCAGACCCAGTAAAACAGGCAGCATAAGCAGCCAAAGCAACAGCGGCCCCAGATCTAGTAACACAGACAGCAGCAGCAGCAGCAGCCCCGGCAATA
TCAAAACCAGCAGCAGCAGCAGCAGCAGTAACAGCAGCAGCAGCAGCCACAGCAGCTACAGCTGCAGCAGCAGCAGCCACAGCAGCAGCAGCAGCCG
CAGCCATAGCAGCAGCCACAGTCACAGCAGCCACAGTCGCACACCATGGGAATGAtttcctaagggtatactgctgagttccttcagtctttcaggc
cacatgtacttcgtaaaagattcctgtctgagcatcttccccagtggcctgtgaggatgagctcatcctttgaacttggcaggcgagaagcgttccg
atttggttcagcacaagcacagaaaggtgattgttaagtagaaattatccagtcagcaatcagacagagtcagtcctttcttccccagaagatgctg
gccaacttctggagaagtttctgctggtgggtttgcctaggcccatatttcctggctttcaacagaggctctgagatttggttagctttgagagctt
ctatggaggaaagggggcccaagagacatagactaatgcaaaaacagaaatgtaggatgaaggaaaccaaatcagaccagttactagagttcaagg
aatggagaagcaaggtttagaggggaaggagttggccggcagctgggatgagggtgtgcctgtcctgctgttattaggtgggtgtggggaccaagca
aatgacaattaaagccgtgggccagttctacacaggaactagaacaagaaagcaaacagctgggcagttatggggctggatgctcagggaaaggggc
ctagggtttaagccaaaccaaaccccaaacaaacacccattgaattggatcgatatcaaagcttttcctgtacatttttttattattgtttgtttat
ttcaataaagatttctttatttttatatgtct

Rattus norvegicus

>transcript_A|UCSC_Rnorvegicus.rn6:chr4:118914603-118915055,118920120-
118921088:+|longest_predict_pep=63|cpat_prob=0.15486075797917actggagcaggaggtgaggattcaccgggagaatct
ggctttctgtgaagaggagcagctggcagttggctcctttagtccttaggtgccgaggctcttgcagcagtcggctcggcctcctcttccctgcttc
gtgctgcacccttgcgtgttatttaccagtggtttccttggactgactttgggaggagagcccaggacgttggctgcagggtcctccagctaggggc
gcctgacatctgactgtgtcttcctaaggctttggagatcaagggccgaaggcaagagtgggagcctcttgtgtgtattttcctcagtctttgttct
caccatgctttttttctggcttctagtccttttggcgtcaggagagtagctgaagtcagggaggctggcacggttcccaggccacagcctagttagg
gattcacaaagaatgttggtgactctga**<u>gt</u>**agctgccagtaccagcaccagcaccagcagtagccatagcaacagcagccgcagcaATGACACAGAT
GCTGGGACCACCAGCAACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCCTCGGCCACCACAGCTGCTGCAGCC
ATAGCTGCACTCCACAGAAATGATTCCCAAAGGCATACTGCTGAGCTCTTTCAGTCTTTTCAGACCACATGTACTTTGTAAaaggttcctgcccgag
catcttccccagtggcctgcgagaatgagctcatcctttgaacttggcaggcaagaagcattccgatctggttcagcacaagtacagaaaggtgatt
gttaagtagaaattatccagtcagcaatcagacagggccagtcctttcttcccccagaagatgctggccaacttctggagaagtttctgcaggaggg
tttgcctagacccatacatcctggcttccaacagaggctatgagatctggttagctttgagggcttctgtggaggacaggggaacccaggagagaca
gactaatgtgaaaacagaaatgtaggatgaaggaaactaaatcagaccagttactaaagtccaaggaatggaaaaactgggtttagagagaacaagt
cggccagcaactgggatgagggtgtgcctgtctcactgttattaggtgggtgtggggaccaagtaaatgacacttaaggccttgggccggttctaca
caggaactggaacaagaaagcaaacagctgggcagttatggggctgaatgctcagggaaaggggcctagggttcaagcaaaaccaaaccaaactgaa
atcacccattgacgtgacctcaaaagcttttcctggacatttttattgttattgtttatgttctttccccagtggggaaaagtcttttttttttca
ataaagatccctttatttttatgtgtat

>transcript_B|UCSC_Rnorvegicus.rn6:chr4:118915749-118916268,118920120-
118921088:+|longest_predict_pep=63|cpat_prob=0.1547023337591ctctcatcttgccttgtcaccttctgctgctctgcag
ggagctgccatgctaggagctccaaggcctttaagctgggttgggaatgggctgagtacctacttgtccagcattcctaaggtattgggactcagtc
aaacaaacaaacaaacaaacacacaacaacaaagcagaaaaatttttcaaacccggtaaatgacataacagagaagtcctcatcaccagtgtagaaa
tgagcctttccatgccccagttggaaagggagcaataaatctgaccatgccccagcttacagactgactccacccactggagaccagagagctcacc

cagctactccaatgcagactcagcctgctggagccgaggaaaaataaatactgacttccagcttccaaacttgggatggttgtcacacagatctata
ggtcactcacagggtgggtgccacgtatgctaattcctcactcctatattttggaaaaactgagatttccctaagaattactacctttaaaaa**gt**a
gctgccagtaccagcaccagcaccagcagtagccatagcaacagcagccgcagcaATGACACAGATGCTGGGACCACCAGCAACAGCAGCAGCAGCA
GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCCTCGGCCACCACAGCTGCTGCAGCCATAGCTGCACTCCACAGAAATGATTCCCAAA
GGCATACTGCTGAGCTCTTTCAGTCTTTTCAGACCACATGtACTTTGTAAaaggttcctgcccgagcatcttccccagtggcctgcgagaatgagct
catcctttgaacttggcaggcaagaagcattccgatctggttcagcacaagtacagaaaggtgattgttaagtagaaattatccagtcagcaatcag
acagggccagtcctttcttcccccagaagatgctggccaacttctggagaagtttctgcaggagggtttgcctagacccatacatcctggcttccaa
cagaggctatgagatctggttagctttgagggcttctgtggaggacaggggaacccaggagagacagactaatgtgaaaacagaaatgtaggatgaa
ggaaactaaatcagaccagttactaaagtccaaggaatggaaaaactgggtttagagagaacaagtcggccagcaactgggatgagggtgtgcctgt
ctcactgttattaggtgggtgtggggaccaagtaaatgacacttaaggccttgggccggttctacacaggaactggaacaagaaagcaaacagctgg
gcagttatggggctgaatgctcagggaaaggggcctagggttcaagcaaaaccaaaccaaactgaaatcacccattgacgtgacctcaaaagctttt
cctggacatttttattgttattgtttatgttctttccccagtgggggaaaagtcttttttttttcaataaagatccctttattttatgtgtat

>transcript_C|UCSC_Rnorvegicus.rn6:chr4:118916975-118917072,118920120-
118921088:+|longest_predict_pep=63|cpat_prob=0.15570238174178ggaggcaagcctgtaagaagtttctccatgggctct
gcattggttcctgtctccaagtgcctcctgtcctgacttctctggatgatggactacaagttagctgccagtaccagcaccagcaccagcagtagc
catagcaacagcagccgcagcaATGACACAGATGCTGGGACCACCAGCAACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA
GCAGCAGCCTCGGCCACCACAGCTGCTGCAGCCATAGCTGCACTCCACAGAAATGATTCCCAAAGGCATACTGCTGAGCTCTTTCAGTCTTTTCAGA
CCACATGTACTTTGTAAaaggttcctgcccgagcatcttccccagtggcctgcgagaatgagctcatcctttgaacttggcaggcaagaagcattcc
gatctggttcagcacaagtacagaaaggtgattgttaagtagaaattatccagtcagcaatcagacagggccagtcctttcttcccccagaagatgc
tggccaacttctggagaagtttctgcaggagggtttgcctagacccatacatcctggcttccaacagaggctatgagatctggttagctttgagggc
ttctgtggaggacaggggaacccaggagagacagactaatgtgaaaacagaaatgtaggatgaaggaaactaaatcagaccagttactaaagtccaa
ggaatggaaaaactgggtttagagagaacaagtcggccagcaactgggatgagggtgtgcctgtctcactgttattaggtgggtgtggggaccaagt
aaatgacacttaaggccttgggccggttctacacaggaactggaacaagaaagcaaacagctgggcagttatggggctgaatgctcagggaaagggg
cctagggttcaagcaaaaccaaaccaaactgaaatcacccattgacgtgacctcaaaagcttttcctggacatttttattgttattgtttatgttct
ttccccagtgggggaaaagtcttttttttttcaataaagatccctttattttatgtgtat

>transcript_D|UCSC_Rnorvegicus.rn6:chr4:118917499-118917845,118920120-
118921088:+|longest_predict_pep=63|cpat_prob=0.15511166990572cgctgtcatttgcccctgctcttggtgatgtcagt
tttgacatcaccgaagctgtggtcttagggtgtctccctgggcttcctgagacagtaggcgctggcactgtggcggtgctgtgcggagacgacagct
gcagagccgggagcacagaatggctgagtgacttagcagagagaccggctgctcaaggcaggggacagcaatgtcctctccagactcctgcttggga
gacagctgtttgggttcccttatacttctctgagcttccctcacagagtggtggtccaactctacagcacggtcacaccctctcatcttcctttcag
agtgccctctgtgggacaa**gt**agctgccagtaccagcaccagcaccagcagtagccatagcaacagcagccgcagcaATGACACAGATGCTGGGACC
ACCAGCAACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCCTCGGCCACCACAGCTGCTGCAGCCATAGCTGCA
CTCCACAGAAATGATTCCCAAAGGCATACTGCTGAGCTCTTTCAGTCTTTTCAGACCACATGTACTTTGTAAaaggttcctgcccgagcatcttccc
cagtggcctgcgagaatgagctcatcctttgaacttggcaggcaagaagcattccgatctggttcagcacaagtacagaaaggtgattgttaagtag
aaattatccagtcagcaatcagacagggccagtcctttcttcccccagaagatgctggccaacttctggagaagtttctgcaggagggtttgcctag
acccatacatcctggcttccaacagaggctatgagatctggttagctttgagggcttctgtggaggacaggggaacccaggagagacagactaatgt
gaaaacagaaatgtaggatgaaggaaactaaatcagaccagttactaaagtccaaggaatggaaaaactgggtttagagagaacaagtcggccagca
actgggatgagggtgtgcctgtctcactgttattaggtgggtgtggggaccaagtaaatgacacttaaggccttgggccggttctacacaggaactg
gaacaagaaagcaaacagctgggcagttatggggctgaatgctcagggaaaggggcctagggttcaagcaaaaccaaaccaaactgaaatcacccat
tgacgtgacctcaaaagcttttcctggacatttttattgttattgtttatgttctttccccagtgggggaaaagtctttt

Mesocricetus auratus

>transcript_A|UCSC_mesAur:NW_004801630.1:3721991-3722454,3729562-
3730999:+|longest_predict_pep=162|cpat_prob=0.872829886678
actagagcaggaagtgaccattccctgggaaagtctgacttcttgggaagaggagcagcaggaagcctgctccagactgtggggcctaaagttagag
gctgggagcactcctaagcagccagctcaggcccctcttccccgatttgtggggtcacttgtatgctcttcttggactggggtttaggagaaccca
ggacactggctgctgtggcctccagctagggggcatctgacatctgattgcaactttgcctcttcctaaggctttggagagcaagggctgaaggcag
gactgggcctttctgtccccccccccccatctttctcatgaggacttggttctcacagcgctttctgggcttctactccttttggcaccaggagact
agctgaagtcagggaggctggggcagttcccagtccacagcccattagggacccacaaagaATGTCAGTGACTGA**GC**AGCCGCAGCTGAAGCCTCAG
CATTGTCACCAACATCACCACCAGCAACACCAGCAACAGCCGCAGCACCAGCAACACCAGCAACACCAGCAACACCAGCAACACCAGCAACAGCCGC
AGCACCAGCATCAGCCAAACAGCAGCAGCAGCAGCAGCAGTAGTAGTAGTAGCAGCAGCAGCAGTAGCAGCAGCAGCAGCAGTTGCTGCAGCAGGCG
TACTGCCAGTGGCAGCAGCCCCCCCACCAACCATAACACCACCAACCATAACACCGCCAACTGTCACAGCAACACCGCCAACCACAGCACCGCCAAC
AGTCACAGCAACACCGCCAACCACAGCACCGCCAACTGTCACAGCAACACCGCCAACCACAGCACCGCCAACAGTCACAGCAACGCCACTAGCAGCT
GTGGCATCACCACCAGCAGTGGCATTACCAGCAGGAGCCGCGGCTTCCGCAGCCCTCCATAGggatgatttctcaagagtccactgctgcgtttcct
gagtcttacaggccacgtgtgcttcttttataaaagattccttgcctgagcgtctccccagcgcccagcgaagatgagctcatcctttgaacttggc
aggcaagaagcattccgatctggttcagtacaggcacagaaaggtgattgttaaacacaagttatccagtcagcaatcaaaagagccagtcctttct
tccccagaagatgctggccaacttctggagaagtttctgctggcagtgtttgcctagacccacagatcttggcttcaatcagaggctgtaagagctg
gttaattttgagggcctctatggaggaaaggggacccaagagaagcagactaacatgaaaacagaaatgtaggacagaggaaaccaaatcagacaa
gttgctggaacccagagaatggaaaagcagagtttggaggagagcaaggcgctcagcgactggaatgagggtgtgcctatcctgcttaggtgggggt
ggggactgagcagatgagagtgaagggcttgggccagttctacaggatcagagcaagaaagcaaacagctggacaaatatggggctgaatgcttagt
gaaaggggcccagggttacaaacaaaagcacccatcgaattgaatggaaagatgtcaaaaaagcttttcctgtaattttttgagttgttatttgtttt
tccagtggggaaagctttctctctctttttttaaagattgctttatttttatgtgtatgagtgttttgcctgcatgtagtataagtacagtgcctgg
ttcctgcagaggccaggagaggcatcagatcccctggagctggagttatggatggctgtgagcttctgtatgggtgctgggagctaaacccaggtcc
tctgaaagaataacaaatgctcttagtggtggagccatctctccatccccccaaataaa

>transcript_B|UCSC_mesAur:NW_004801630.1:3723018-3723134,3729562-
3730999:+:|longest_predict_pep=36|cpat_prob=0.019787929297679gcatccttccctgacacaccagagaaggcccatgcc
ccaaacctatctctccacattctggagaatgtctgtaaggggaaaattaatgacacactgaaggatacatctggataga**gc**agccgcagctgaagc
ctcagcattgtcaccaacatcaccaccagcaacaccagcaacagccgcagcaccagcaacaccagcaacaccagcaacaccagcaacaccagcaaca

gccgcagcaccagcatcagccaaacagcagcagcagcagcagcagtagtagtagtagcagcagcagcagtagcagcagcagcagcagttgctgcagc
aggcgtactgccagtggcagcagcccccccaccaaccataacaccaccaaccataacaccgccaactgtcacagcaacaccgccaaccacagcaccg
ccaacagtcacagcaacaccgccaaccacagcaccgccaactgtcacagcaacaccgccaaccacagcaccgccaacagtcacagcaacgccactag
cagctgtggcatcaccaccagcagtggcattaccagcaggagccgcggcttccgcagccctccatagggatgatttctcaagagtccactgctgcgt
ttcctgagtcttacaggccacgtgtgcttcttttataaaagattccttgcctgagcgtctccccagcgcccagcgaagatgagctcatcctttgaac
ttggcaggcaagaagcattccgatctggttcagtacaggcacagaaaggtgattgttaaacacaagttatccagtcagcaatcaaaagagccagtcc
tttcttccccagaagatgctggccaacttctggagaagtttctgctggcagtgtttgcctagacccacagatcttggcttcaatcagaggctgtaag
agctggttaattttgagggcctctatggaggaaaggggacccaagagaagcagactaacatgaaaacagaaatgtaggacagaggaaaccaaatca
gacaagttgctggaacccagagaatggaaaagcagagtttggaggagagcaaggcgctcagcgactggaatgagggtgtgcctatcctgcttaggtg
ggggtggggactgagcagatgagagtgaagggcttgggccagttctacaggatcagagcaagaaagcaaacagctggacaaatatggggctgaatgc
ttagtgaaaggggcccagggttacaaacaaaagcacccatcgaattgaatggaaagatgtcaaaaaagcttttcctgtaattttgagttgttattt
gtttttccagtggggaaagctttctctctctttttttaaagattgctttatttttATGTGTATGAGTGTTTTGCCTGCATGTAGTATAAGTACAGTG
CCTGGTTCCTGCAGAGGCCAGGAGAGGCATCAGATCCCCTGGAGCTGGAGTTATGGATGGCTGTGAgcttctgtatgggtgctgggagctaaaccca
ggtcctctgaaagaataacaaatgctcttagtggtggagccatctctccatcccccaaataaa

>transcript_C|UCSC_mesAur:NW_004801630.1:3723707-3723836,3729562-
3730999:+|longest_predict_orf=36|cpat_prob=0.019783372356702ccagcgggtggtcctggatggccgagcaagccaggag
gaaagtcagtaagcagcactcctccatggcttctgcattggttcctgcctccaggttcctgccctgacttccctggatgatggactataaat**tc**agc
cgcagctgaagcctcagcattgtcaccaacatcaccaccagcaacaccagcaacagccgcagcaccagcaacaccagcaacaccagcaacaccagca
acaccagcaacagccgcagcaccagcatcagccaaacagcagcagcagcagcagcagtagtagtagtagcagcagcagcagtagcagcagcagcagc
agttgctgcagcaggcgtactgccagtggcagcagcccccccaccaaccataacaccaccaaccataacaccgccaactgtcacagcaacaccgcca
accacagcaccgccaacagtcacagcaacaccgccaaccacagcaccgccaactgtcacagcaacaccgccaaccacagcaccgccaacagtcacag
caacgccactagcagctgtggcatcaccaccagcagtggcattaccagcaggagccgcggcttccgcagccctccatagggatgatttctcaagagt
ccactgctgcgtttcctgagtcttacaggccacgtgtgcttcttttataaaagattccttgcctgagcgtctccccagcgcccagcgaagatgagct
catcctttgaacttggcaggcaagaagcattccgatctggttcagtacaggcacagaaaggtgattgttaaacacaagttatccagtcagcaatcaa
aagagccagtcctttcttccccagaagatgctggccaacttctggagaagtttctgctggcagtgtttgcctagacccacagatcttggcttcaatc
agaggctgtaagagctggttaattttgagggcctctatggaggaaaggggacccaagagaagcagactaacatgaaaacagaaatgtaggacagag
gaaaccaaatcagacaagttgctggaacccagagaatggaaaagcagagtttggaggagagcaaggcgctcagcgactggaatgagggtgtgcctat
cctgcttaggtgggggtggggactgagcagatgagagtgaagggcttgggccagttctacaggatcagagcaagaaagcaaacagctggacaaatat
ggggctgaatgcttagtgaaaggggcccagggttacaaacaaaagcacccatcgaattgaatggaaagatgtcaaaaaagcttttcctgtaatttt
gagttgttatttgtttttccagtggggaaagctttctctctctttttttaaagattgctttatttttATGTGTATGAGTGTTTTGCCTGCATGTAGT
ATAAGTACAGTGCCTGGTTCCTGCAGAGGCCAGGAGAGGCATCAGATCCCCTGGAGCTGGAGTTATGGATGGCTGTGAgcttctgtatgggtgctgg
gagctaaacccaggtcctctgaaagaataacaaatgctcttagtggtggagccatctctccatcccccaaataaa

>transcript_D|UCSC_mesAur:NW_004801630.1:3724215-3724559,3729562-
3730999:+|longest_predict_orf=72|cpat_prob=0.23109411264975cattgtcattttgctcaggttgttagtgatgtcaacaa
tgacatcactgaggctgtagactctgtgtctctccaggcttcttgagacagcagatgctggcactgtggtgttggaggtagctgctggtagtggctg
ggctccagacgcatttctagggagactggctgttcaaggcagtggccagccacctattacctcctgcttgaaagctgtttaggttcccatttatttc
tctgaacttccctcacaaaatggccttctaaagctatggcaagttcactctggaatgcATGACTGGGGAACCCATCTCATCTTTCACTTCAGTTTAC
CCTCTAATGGGACAA**GC**AGCCGCAGCTGAAGCCTCAGCATTGTCACCAACATCACCACCAGCAACACCAGCAACAGCCGCAGCACCAGCAACACCAG
CAACACCAGCAACACCAGCAACACCAGCAACAGCCGCAGCACCAGCATCAGCCAAACAGCAGCAGCAGCAGCAGCAGTAGtagtagtagcagcagca
gcagtagcagcagcagcagcagttgctgcagcaggcgtactgccagtggcagcagcccccccaccaaccataacaccaccaaccataacaccgccaa
ctgtcacagcaacaccgccaaccacagcaccgccaacagtcacagcaacaccgccaaccacagcaccgccaactgtcacagcaacaccgccaaccac
agcaccgccaacagtcacagcaacgccactagcagctgtggcatcaccaccagcagtggcattaccagcaggagccgcggcttccgcagccctccat
agggatgatttctcaagagtccactgctgcgtttcctgagtcttacaggccacgtgtgcttcttttataaaagattccttgcctgagcgtctcccca
gcgcccagcgaagatgagctcatcctttgaacttggcaggcaagaagcattccgatctggttcagtacaggcacagaaaggtgattgttaaacacaa
gttatccagtcagcaatcaaaagagccagtcctttcttccccagaagatgctggccaacttctggagaagtttctgctggcagtgtttgcctagacc
cacagatcttggcttcaatcagaggctgtaagagctggttaattttgagggcctctatggaggaaaggggacccaagagaagcagactaacatgaa
aacagaaatgtaggacagaggaaaccaaatcagacaagttgctggaacccagagaatggaaaagcagagtttggaggagagcaaggcgctcagcgac
tggaatgagggtgtgcctatcctgcttaggtgggggtggggactgagcagatgagagtgaagggcttgggccagttctacaggatcagagcaagaaa
gcaaacagctggacaaatatggggctgaatgcttagtgaaaggggcccagggttacaaacaaaagcacccatcgaattgaatggaaagatgtcaaaa
aagcttttcctgtaattttgagttgttatttgtttttccagtggggaaagctttctctctctttttttaaagattgctttatttttatgtgtatga
gtgttttgcctgcatgtagtataagtacagtgcctggttcctgcagaggccaggagaggcatcagatcccctggagctggagttatggatggctgtg
agcttctgtatgggtgctgggagctaaacccaggtcctctgaaagaataacaaatgctcttagtggtggagccatctctccatcccccaaataaa

Supplemental References

Abe K, Yamamoto R, Franke V, Cao M, Suzuki Y, Suzuki MG, Vlahovicek K, Svoboda P, Schultz RM, Aoki F. 2015. The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *The EMBO journal* **34**: 1523-1537.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome research* **14**: 1188-1190.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Fabre PH, Hautier L, Dimitrov D, Douzery EJ. 2012. A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC evolutionary biology* **12**: 88.

Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. 2014. Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 4139-4144.

Graham T, Boissinot S. 2006. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *Journal of biomedicine & biotechnology* **2006**: 75327.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology* **28**: 503-510.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Molecular biology and evolution* **32**: 835-845.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic acids research* **44**: D81-89.

Karlic R, Ganesh S, Franke V, Svobodova E, Urbanova J, Suzuki Y, Aoki F, Vlahovicek K, Svoboda P. 2017. Long non-coding RNA exchange during oocyte-to-embryo transition in mice. *DNA Research* doi:10.1093/dnares/dsw058.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome research* **12**: 996-1006.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Orostica KY, Verdugo RA. 2016. chromPlot: visualization of genomic data in chromosomal context. *Bioinformatics*: btw137 %@ 1367-4803.

Park SJ, Komata M, Inoue F, Yamada K, Nakai K, Ohsugi M, Shirahige K. 2013. Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. *Genes & development* **27**: 2736-2748.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**: 290-295.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G. 2011. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature genetics* **43**: 811-814.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *The Journal of Machine Learning Research* **9**: 85.

Veselovska L, Smallwood SA, Saadeh H, Stewart KR, Krueger F, Maupetit-Mehouas S, Arnaud P, Tomizawa S, Andrews S, Kelsey G. 2015. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome biology* **16**: 209.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE et al. 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**: 593-597.

Zeng F, Baldwin DA, Schultz RM. 2004. Transcript profiling during preimplantation mouse development. *Developmental biology* **272**: 483-496.

Author Contributions

Study Conception and Design: PS, VF, SG, KV

Acquisition of Bioinformatic Data: VF, RK, JP, FH, MF, KV, SG, ES, JU, FA, YS

Acquisition of Molecular Biology Data: RM, ES, HF, MC, JU, SG, RJ, JM

Analysis and Interpretation of Data: VF, SG, RK, JP, FH, MF, KV, ES, JU

Manuscript Preparation: PS, RMS, KV, VF, RM

The project was led by PS and co-supervised by KV

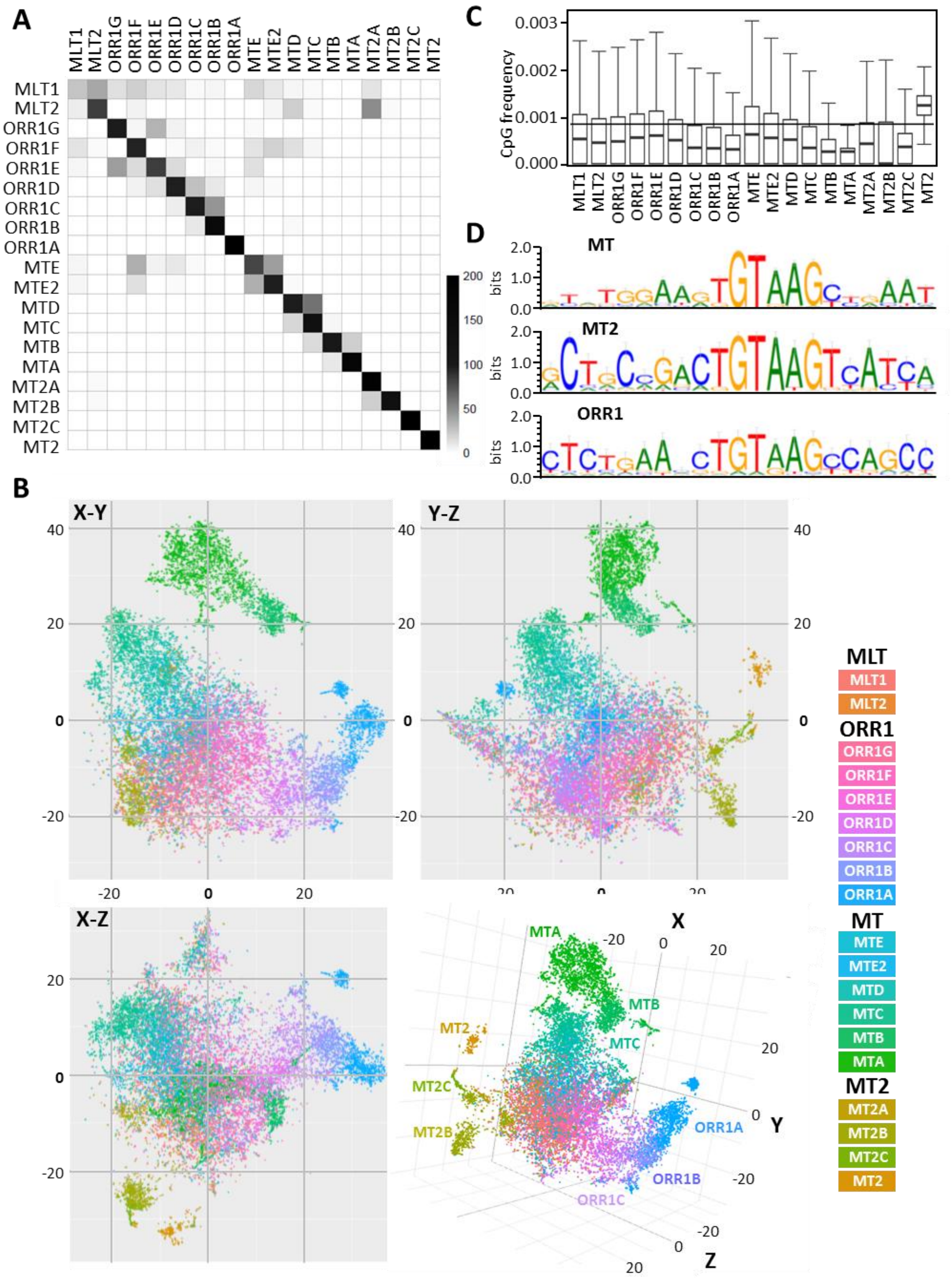

Figure S1. Supplementary data for MaLR and MT2 LTR analysis. (A) Murine ERVL LTR classification by random forest walk. Shown is the confusion matrix where rows are observed classes and columns predicted classes. The gray-scale indicates the number of LTRs; 200 full-length LTRs were randomly collected for each class from RepeatMasker annotation. (B) Sequence divergence of MaLR and MT2 LTRs. Shown is a map of LTR sequences projected on a 3D space using t-SNE. Each point represents one LTR colored by a corresponding LTR subfamily. The map reproduces known relationships within the ERVL group with older subfamilies (MLT1, MLT2, ORR1D-G, MTD- E, MT2A clustered in the center, and younger subfamilies branching to the periphery in different directions. Specifically labeled are younger LTR subfamilies (MTA-C, ORR1A-C, MT2B, MT2C, and MT2), which emerged in the mouse lineage upon the split from the hamster lineage. (C) CpG frequencies in selected LTRs subfamilies. The horizontal line depicts the mouse genome average CpG frequency (0.0083). (D) Sequence logos of splice donors in selected LTR families. For the combined logo in Fig. 1F, highly abundant MTA and MTB LTR sequences were not used because their inclusion made the logo almost the same as that of the MT family.

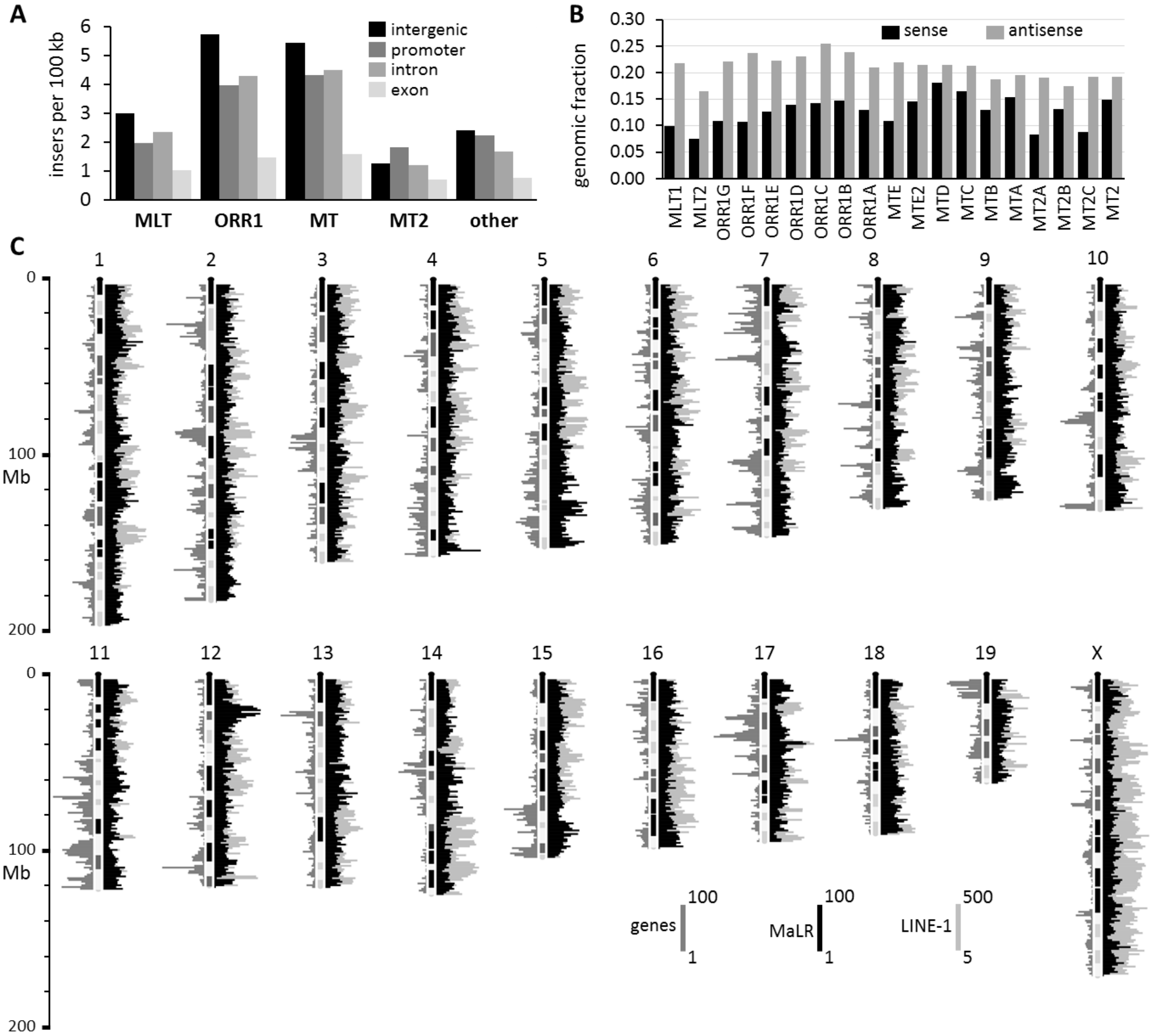

Figure S2. Genome distribution of selected LTRs. (A) Frequency of insertions of LTR families in different genome regions. (B) Frequency of sense and antisense oriented LTR-bearing insertions of ERVL subfamilies. (C).MaLR, LINE-1, and gene distribution along mouse chromosomes. LINE-1 elements are enriched in AT-rich, gene-poor regions, and sex chromosomes (Graham and Boissinot 2006). The chromosomal idiograms were created using the R package chromPlot (Orostica and Verdugo 2016). The cytogenetic bands were taken from the UCSC Genome Browser. The genomic data on both sides of each chromosome were calculated as histograms representing numbers of genes or element inserts in 1Mb bins. The UCSC genes, and RepeatMasker viz annotation of MaLR and LINE-1 insertions in the mm10 genome annotation were used for calculations. The colored scale bars indicate numbers of insertions in 1Mb bins for genes and MaLR and LINE-1 elements. One of the apparently enriched loci is chr.2 A1.2-.3, which harbors two complex segmental duplication domains.

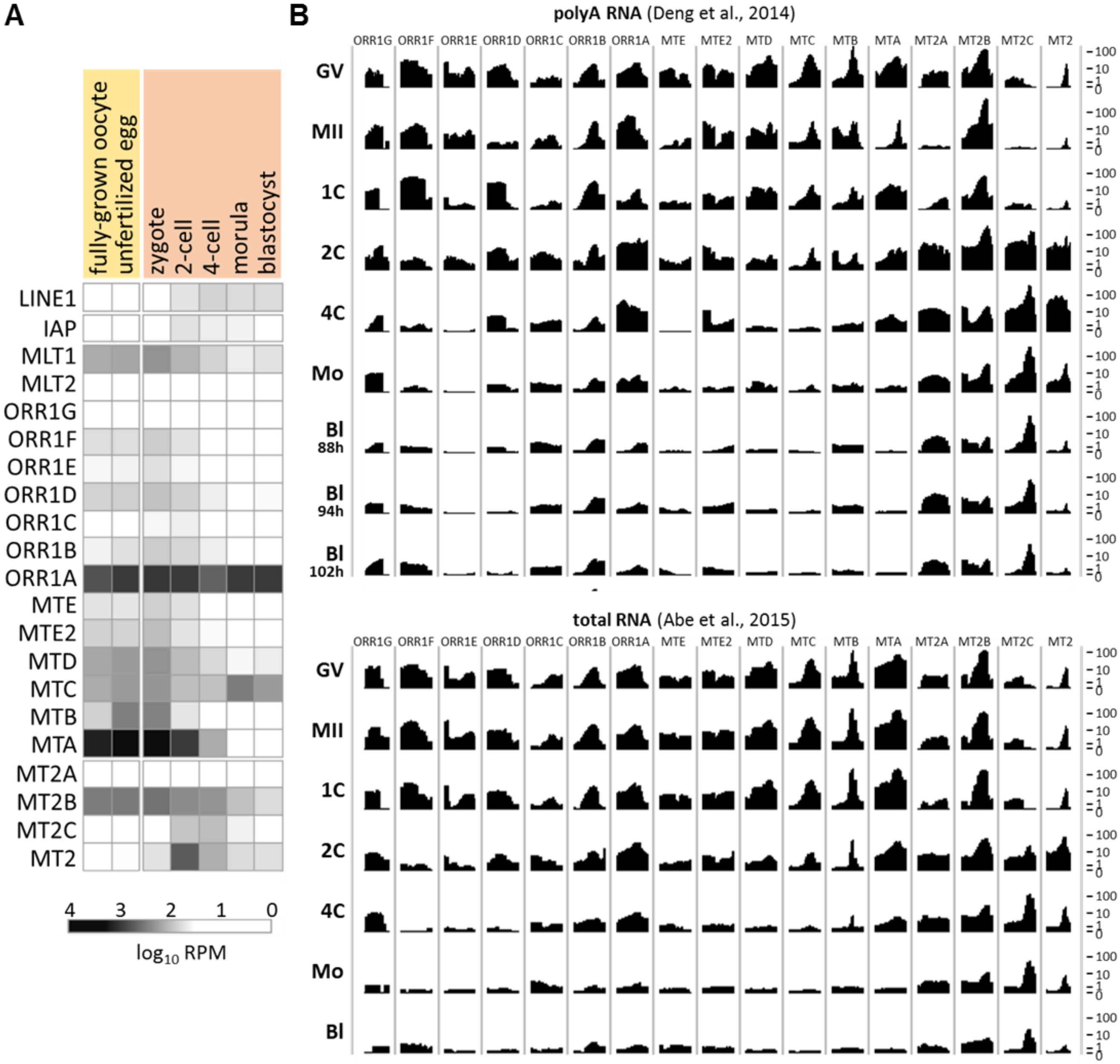

Figure S3. Developmental expression profiles of MaLR and MT2 LTRs. (A) Display of ERVL LTR RNA abundance in oocytes and during early development. The heatmap shows $\log_{10}$ FPKM of LTRs from different MaLR and MT2 subfamilies in our previous NGS of total RNA E-MTAB-2950 (MII oocyte -blastocyst), E-MTAB-4775 (GV oocyte). Sample abbreviations: GV, full-grown GV oocyte; MII, metaphase II oocyte; 1C, 1-cell stage (fertilized egg); 2C, 2-cell stage; 4C, 4-cell stage; Mo, morula; Bl, blastocyst. (B) Average levels of transcripts derived from the selected LTR subfamilies and stages calculated from polyA RNA (GSE45719, (Abe et al. 2015)) and total RNA (E-MTAB-2950, (Abe et al. 2015)) next generation sequencing experiments. The y scale shows log(RPKM) of the co-opted elements (Table S2) calculated for consecutive 5% bins along the LTR sequence.

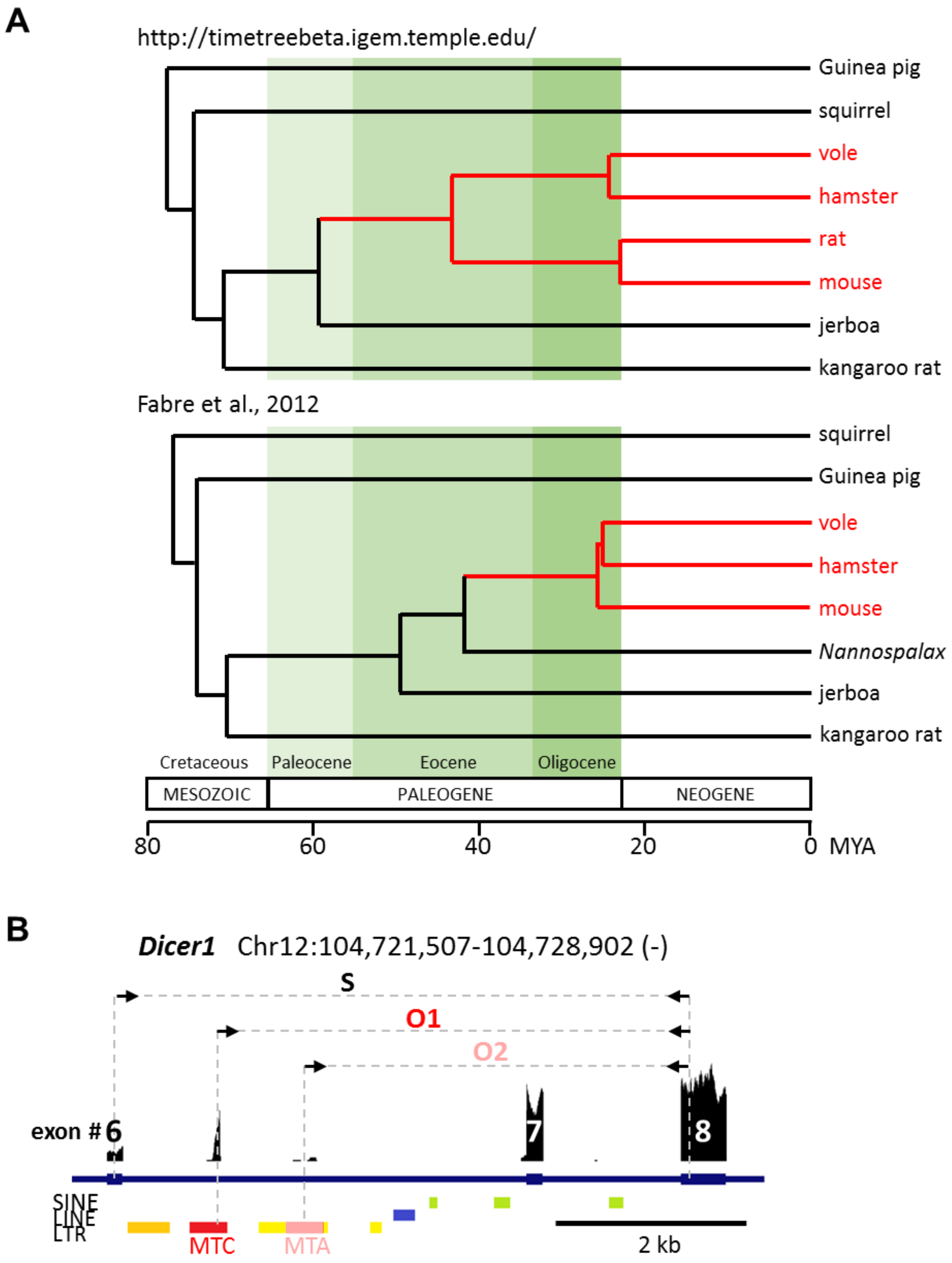

Figure S4. Supplementary data for *Dicer1* analysis. (A) Phylogenetic analysis of *Dicer1*^{O} isoform. Highlighted in red in two independently produced phylogenetic trees is the common ancestry of the MTC LTR insertion, which is responsible for *Dicer1*^{O} expression. The upper tree was generated from rodent species with sequenced genomes in the TimeTree public knowledge-base (http://timetreebeta.igem.temple.edu/, (Hedges et al. 2015)). The lower tree was adopted from Fabre et al. (Fabre et al. 2012). Both trees were aligned according to the time-scale, which was accompanying them. (B) Schematic depiction of mouse *Dicer1* gene fragment (chr12:104,721,507-104,728,902) carrying MT LTRs controlling expression of *Dicer1*^{O} transcript isoforms and approximate positions of primers used for qPCR analyses to detect specific *Dicer1* isoforms (O1 and O2 = MT-driven oocyte-specific isoforms, S =-somatic full-length isoform). Above the scheme of the gene is displayed relative density of NGS data from fully-grown GV oocytes (E-MTAB-4775).

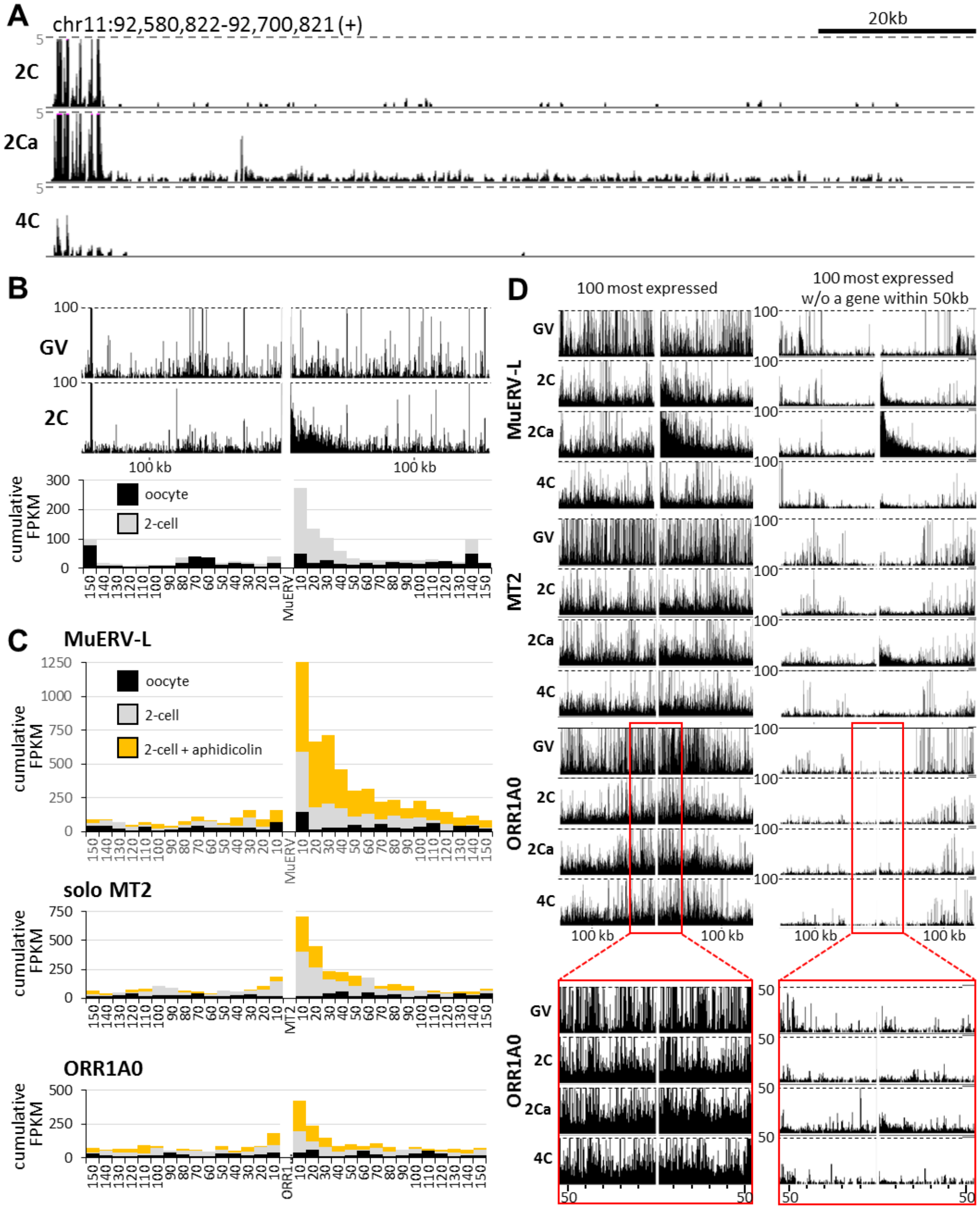

Figure S5. Transcription downstream of MT2 and ORR1A0 LTRs during ZGA A) Transcription downstream of a selected MuERV-L during ZGA and the effect of inhibiting replication in 2-cell embryos on RNA abundance. Shown is a UCSC Genome Browser snapshot of an MuERV-L insertion that becomes transcriptionally active during ZGA. The grey horizontal lines indicate maximum CPM values of displayed tracks and set to five to visualize low levels of transcripts extending into the genomic flank. Developmental stages are indicated on the left. (B) Downstream transcription can be observed in an independent next generation sequencing dataset. Shown is a cumulative display of 150 kb genomic flanks around hundred MuERV-L elements generated as in Fig. 7B from SOLiD total RNA sequencing (DRA001066 (Park et al. 2013)). The upper panel shows a cumulative display of NGS data from GV oocytes and 2-cell embryos, the graph below calculation of cumulative FPKM values for 10kb bins. (C) Cumulative display of transcription in 150 kb genomic flanks around hundred MuERV-L elements, and MT2 and ORR1A0 solo LTRs, which are most expressed during ZGA. The signal (cumulative FPKM) is displayed for 10 kb bins for oocytes, 2-cell embryos and 2-cell embryos treated with aphidicolin. (D) Cumulative expression displays for full-length MuERV-L MT2 LTRs, solo MT2 LTRs, and ORR1A0 LTRs. The left column show cumulative displays of 100 most expressed LTR loci without any filtering, the right column shows a cumulative display for 100 most expressed intergenic LTR loci without an annotated gene sequence within 50kb from the element. Dashed horizontal lines indicate cumulative CPM values.

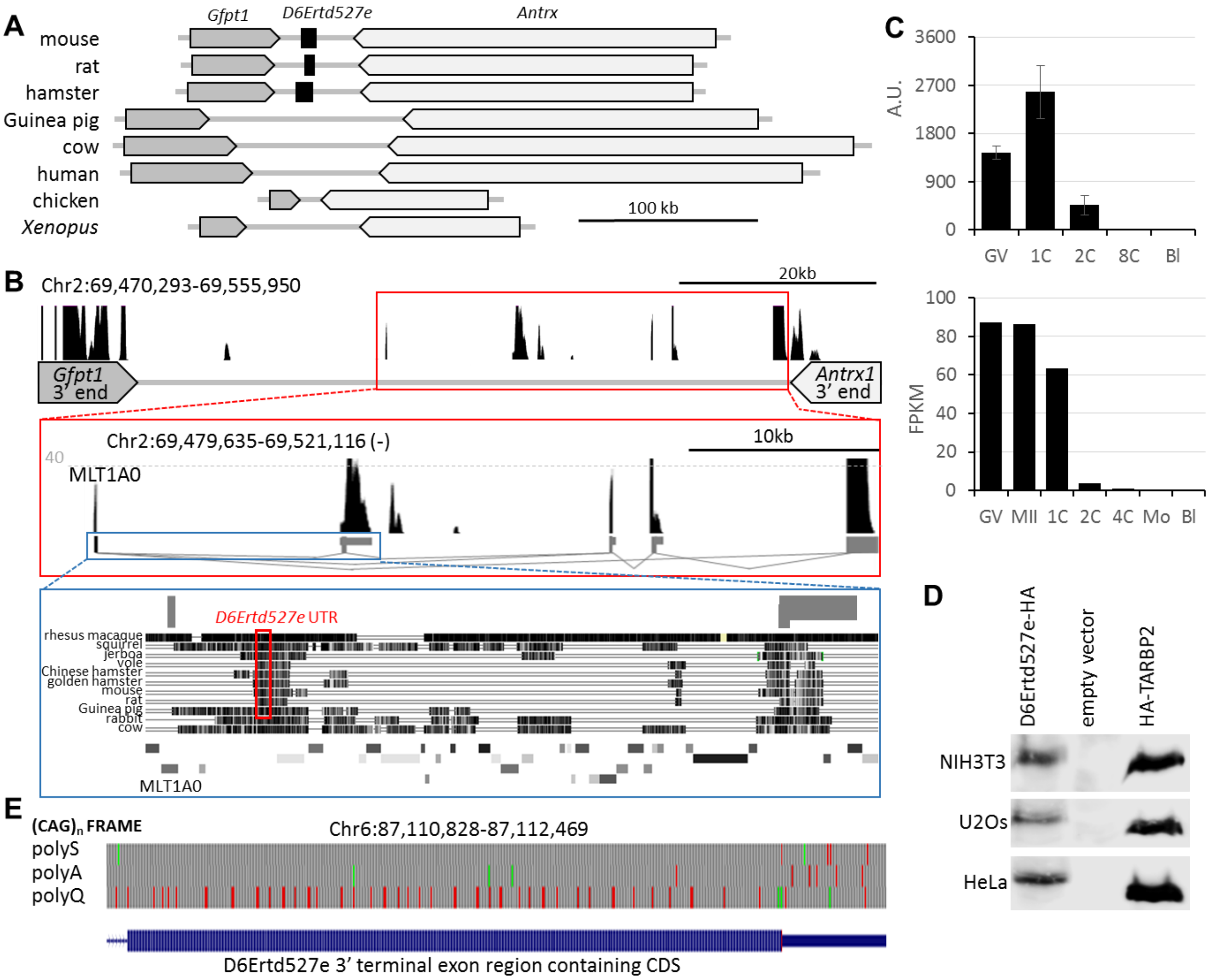

Figure S6. Supplementary data for *D6Ertd527e* analysis. (A) Schematic display of synteny of the *D6Ertd527e* locus in vertebrates. Gray rectangular arrows indicate relative size and transcriptional orientation of *Gfpt1* and *Antrx* genes flanking the *D6Ertd527e* locus. *D6Ertd527e* gene is depicted as a black rectangle. (B) Overview of the human syntenic region, which hosts an MLT1A0 insertion controlling expression of lncRNA. Position of the conserved 3'UTR of *D6Ertd527e* is indicated by the red rectangle in the lower panel. The central panel corresponds to the panel displayed in Fig. 7B. (C) *D6Ertd527e* is expressed maternally. The left graph depicts normalized microarray hybridization signal for *D6Ertd527e* probe 1433484_at on the Affymetrix MOE 430.2 array (Zeng et al. 2004), the right graph displays quantification of *D6Ertd527e* in our NGS profile of early development (Abe et al. 2015; Karlic et al. 2017). (D) Immunoblot analysis of three different cell lines transfected with *D6Ertd527e* or *Tarbp2* vectors expressing HA-tagged protein, which were visualized with an anti-HA antibody. (E) Initiation and stop codons in polyS, polyA and polyQ open reading frames in the last exon of *D6Ertd527e*. Initiation and stop codons are depicted as vertical green and red lines, respectively. Position of the last exon is shown below the ORF diagram.