Science
AAAS

Supplementary Materials for

Regulatory Evolution of Innate Immunity through Co-option of Endogenous Retroviruses

Edward B. Chuong, Nels C. Elde, and Cédric Feschotte
correspondence to: nelde@genetics.utah.edu (N.C.E), cedric@genetics.utah.edu (C.F.)

This PDF file includes:

Materials and Methods

Dataset GEO accession numbers

The genomic data analyzed in this study were obtained from publicly available datasets. ChIP-Seq datasets were obtained from GSE31477 (IRF1 and STAT1 for IFNG-stimulated K562 cells, STAT1 for IFNG-stimulated HeLa cells), GSE43036 (IRF1, STAT1, and H3K27ac for IFNG-stimulated CD14+ primary macrophages), and GSE33913 (STAT1 for IFNG-stimulated mouse primary bone marrow-derived macrophages). RNA-Seq datasets were obtained from GSE66809. Individual sample library accession codes are available in Table S2 (RNA-Seq) and S5 (ChIP-Seq).

ChIP-Seq analysis

Illumina FASTQ reads were downloaded from the NCBI Gene Expression Omnibus (GEO) and aligned to the human (hg19) or mouse (mm9) genomes using BWA SW v0.7.5a (*35*). Reads were filtered to remove low-quality reads and reads mapping to multiple locations (alignment quality score > 10). Alignments corresponding to replicate samples were merged, and ChIP-Seq peaks were called at an FDR < 0.05 using MACS2 v2.1.0 (*36*), with all ChIP libraries normalized against respective input control libraries. Normalized profiles corresponding to read coverage per 1 million reads were used for heatmap and metaprofile visualization, which were generated using the deepTools v1.5.8.1 (*37*). The CRG 36 bp alignability track was used to inspect genome sequence uniqueness. Complete peak datasets (hg19 or mm9) are available in Table S5.

Transposable element analysis

All TE analysis was performed using the Repeatmasker annotations downloaded from http://www.repeatmasker.org **(**Repeat Library 20140131). Nucleotide sequence alignments were performed using MUSCLE v3.8.31(*38*).

Identifying overrepresented TE families. For each set of ChIP-Seq peak summits, the expected representation of each TE family was determined by shuffling peaks across the genome. To account for non-random TE integration biases, peaks were shuffled while maintaining the same overall distribution of regions relative to their distances to host genes (directly overlapping a gene transcriptional start site, exon, intron, 10-kb gene proximal region, 10-100 kb distal region or >100 kb intergenic region). Each dataset was shuffled 10,000 times, and the mean number of overlaps from these shuffled datasets was used to determine the expected fraction of TE overlaps in a binomial test. Multiple testing correction was performed by multiplying the *P* value by the number of TE families tested (*N*=1,172). TE families that were enriched with a stringent $P < 10^{-5}$ were further filtered to retain only families where at least 25 copies were bound and the number of bound copies exceeded a 3-fold enrichment over the background expectation.

Enrichment near IFNG-stimulated genes. Correlation analysis between bound ERVs and ISGs was performed using CD14+ data because both ChIP-Seq and RNA-Seq were available for the same cell type and conditions. Repeatmasker-annotated ERVs that overlapped either STAT1 or IRF1 ChIP-Seq peak summits in CD14+ macrophages were merged to form a set of 7,098 STAT1/IRF1 bound ERVs (listed in Table S2). A set of 2,000 ISGs were detected using a matched dataset (Table S2, see RNA-Seq section below), and the absolute distance to the nearest

ISG was determined for all 7,098 ERVs. These distances were grouped by 10 kb bin sizes in Fig 1B. The expected background was determined by randomly sampling an equal number (7,098) of the remaining 720,997 annotated ERVs that were not bound by either TF, and determining the number of elements located within each 10 kb bin relative to the nearest ISG transcriptional start site. Sampling was repeated 10,000 times and the mean number of elements was used as the expected value for comparison to the observed count of STAT1/IRF1-bound ERVs. Statistical significance was determined for the first 10 kb bin, by binomial test. The analysis was repeated using several gene sets: 1) 1,733 genes significantly downregulated by IFNG based on the same RNA-Seq dataset, 2) 2,000 randomly selected genes that were expressed in the same dataset but not differentially regulated, and 3) 2,000 randomly selected genes from the human genome (Fig S2). Functional enrichment based on proximal gene annotations was determined using GREAT v3.0 default enrichment settings (http://great.stanford.edu) (*22*).

TE motif analysis. Binding motif position-weight matrices for STAT1 (MA0137.3 for Gamma Activated Site/GAS motif, MA0517.1 for the Interferon Stimulated Response Element/ISRE motif) were obtained from the JASPAR CORE database (*39*). Motif occurrences within TE consensus sequences or the human genome were identified using FIMO v4.10.0 (*40*) using a p-value cutoff of $1x10^{-4}$. For the heatmap visualization of motif presence in Fig 1D, HeLa-bound copies were used because they represented the largest dataset, only repeat instances that were >50% intact by length were retained, and repeat 5' start locations were recalculated based on Repeatmasker annotations.

Repeat age estimation. Species divergence times were based on (*34*). Repeat ages were estimated by dividing the percent divergence of extant copies from the consensus sequence by the species neutral substitution rate. Substitution rates (mutations/yr) used were as follows: anthropoidea $2.2x10^{-9}$, carnivora $3.8x10^{-9}$, artiodactyls $3.0x10^{-9}$ from (*41*); lemuriformes, $3.0x10^{-9}$, vespertillionidae $2.7x10^{-9}$ from (*42*). Evolutionary lineage placements based on repeat divergence times were independently verified by analysis of presence or absence of MER41-like relatives, using a script by A. Kapusta (Feschotte lab, https://github.com/4ureliek/TEorthology). Briefly, for all mammalian genomes in which extant MER41-like insertions were annotated by Repeatmasker, all insertions not nested in another (masked) repeat were extracted including 5' and 3' flanking 100 nt of flanking sequences, and queried against 23 mammalian genomes representing all major eutherian taxa using BLASTN with an E cutoff of $<1x10^{-50}$. Hits are filtered based on the presence of both 5' and 3' flanking sequence, which is used as a proxy to determine if matches in other species represented syntenic/orthologous elements. Both the total number and number of potentially orthologous matches are recorded (summary output in Table S4). At $E<1x10^{-50}$, lineage-specific MER41 elements do not align to any other family (e.g. primate-specific MER41 elements do not align to MER41-like elements of any other lineage). Lineages with 0 hits across all contained species (e.g. rodents, afrotheria) are inferred to not possess MER41-like elements.

RNA-Seq analysis

Raw FASTQ files corresponding to mock treated and IFNG-treated (2 biological replicates each) were aligned to the human genome (hg19) using the spliced junction mapper HISAT v0.1.6 (*43*), and filtered to remove reads that mapped to multiple locations. Transcript assembly, including assembly of unannotated genes, was performed using StringTie v1.0.4 (*44*). Differential expression analysis of both protein-coding (RefSeq) and annotated/de novo non-coding transcripts was performed using Cuffdiff v2.2.1 (*45*). Significantly upregulated loci at the gene level (i.e. merging all isoforms per gene) with an FDR < 0.05 were considered ISGs. Loci were collapsed to their transcriptional start sites to determine relative distance from ERVs.

Note: For the association analysis between STAT1/IRF1-bound ERVs and ISGs, the binding site and expression datasets were generated in two different studies (ChIP-Seq (*17*), RNA-Seq (*21*)). These studies were performed by the same laboratory, and cell culture conditions and treatments were consistent between the studies (human donor-derived CD14 macrophages cultured using the same conditions, then either mock-treated or treated with 100 U/ml IFNG for 24 hrs). In experiment used for RNA-Seq, both unstimulated and IFNG-stimulated samples were further treated using 10 ng/ml Pam3CSK4 for 4 hours before harvesting, which activates the TLR2-pathway. Because both mock and IFNG-stimulated cells were treated, genes identified from this dataset still represent CD14+ macrophage ISGs, though a fraction of these ISGs may be dependent on TLR2 activation, which was not performed in the ChIP-Seq study.

Cell culture and treatments

Human HeLa F2 cells and human primary fibroblast cells were gifts from A. Geballe (Fred Hutchinson Cancer Research Center, Seattle, WA). Other primate primary fibroblasts were obtained from the Coriell Cell Repositories (Camden, NJ): chimpanzee, PR00226; rhesus macaque, AG06252; white-fronted marmoset, PR00789. All cells were maintained in High Glucose DMEM F12 media (HyClone, Logan, UT) with 100 ug/ml penicillin-streptomycin, 2 mM L-glutamine, and 10% heat-inactivated FBS. All IFNG treatments were performed using 1000 U/ml recombinant human IFNG (cat #11500-2, PBL assay science, Piscataway, NJ) for 24 hrs. Human recombinant IFNG was used to stimulate primary fibroblasts from other primates. All transfections were performed using FuGENE HD (Promega, Madison, WI) following manufacturer's instructions.

Luciferase reporter assays

LTR enhancer sequences were either cloned from genomic DNA (MER41.AIM2) or synthesized as Gene Fragments (Integrated DNA Technologies, Coralville, IA) and cloned into pGLuc Mini-TK 2 Gaussia enhancer reporter plasmids upstream of the minimal promoter. Reporter constructs were co-transfected with a pTK-CLuc constitutive Cypridina vector (ratio of 100 ng Gaussia plasmid: 5 ng Cypridina control plasmid) into HeLa cells. After 24 hrs, media was replaced and cells were either mock treated or stimulated with 1000 U/ml IFNG. 24 hrs following stimulation, cell culture supernatants were assayed for secreted reporter activity using the BioLux *Gaussia* and *Cypridina* Luciferase Assay systems (New England Biolabs, Ipswich, MA), and all reported values were normalized to Cypridina co-transfection controls. All experiments were performed with 3 biological replicates per condition in a 96-well plate format. Sequences are available in Table S2. Results are representative of at least 3 independent experiments. Barplots are presented as mean +/- s.d.

Gene expression analysis

Confluent cells were either mock treated or stimulated with 1000 U/ml IFNG for 24 hrs, then total RNA was extracted from cells using the RNA MiniPrep kit (Zymo, Irvine, CA). 100 ng RNA was prepared for quantitative real-time PCR reactions using the Power SYBR® Green

RNA-to-CT™ 1-Step Kit (Life Technologies) and run on the 7900HT Fast Real-Time PCR System (Applied Biosystems, Waltham, MA). Transcript CT Values were normalized against transcript levels of *CTCF.* Primers were all designed against primate-conserved sequences in the coding sequences of each gene (Table S6). Experiments were performed with 2 biological replicates per condition in a 12-well plate format. Results are representative of at least 3 independent experiments. Barplots are presented as mean +/- s.d.

Statistical analyses

Statistical significance for both qPCR and luciferase reporter assays was assessed using a two-tailed unpaired Student's *t*-test with a threshold of $p < 0.05$.

Virus infections

Cells were seeded at $5x10^5$ cells per well of a 6-well plate. After 24 hrs, cells were transfected either with an empty control vector (pEGFP-N1) or a pcDNA3-hAIM2-T7 *AIM2* rescue plasmid (gift from Emad Alnemri, Addgene #51538). 24 hrs following transfection, cells were washed twice with PBS then infected with vaccinia virus (MOI = 5.0) in 1 ml serum free media. Both cells and supernatants were harvested for Western blot analysis 24 hrs after infection.

Generation and analysis of CRISPR-Cas9 mutants

For each MER41 element (associated with *AIM2, APOL1, IFI6, SECTM1*), two gRNA sequences were designed to generate a specific internal deletion encompassing the predicted STAT1 binding sites. All gRNA sequences, including the 20 bp seed and 3 bp PAM (NGG) sequence, were verified to be unique targets in the human genome using BLAT against the hg19 genome assembly. gRNA oligos were ordered from IDT and cloned into pSpCas9(BB)-2A-Puro vectors (gift from Feng Zhang, Addgene #62988) following a published protocol (*46*). HeLa cells were co-transfected with both gRNA constructs and placed under puromycin selection for 4 days. Clonal lines were isolated using limited dilution cloning, and 100-200 clones were screened for homozygous deletions by PCR using both flanking and internal primer pairs at the expected deletion site (Fig S8). To determine deletion breakpoint sequences, PCR products flanking each breakpoint were TA-cloned into PCR2.1 TOPO vectors (Life Technologies, Carlsbad, CA) and transformed into DH5a cells, and at least 30 individual colonies were sequenced using Sanger sequencing (Genewiz, South Plainfield, NJ).

Western blot analysis

Cell lysates were prepared with RIPA buffer and resuspended in 2x Laemelli buffer supplemented with 5% beta-mercaptoethanol. Cell culture supernatants were concentrated using methanol-chloroform precipitation and resuspended in 2x Laemelli buffer. Antibodies used were as follows: AIM2 (cat #D5X7K, Cell Signaling Technologies, Danvers, MA), Caspase-1 (cat #sc-515, Santa Cruz Biotechnology, Santa Cruz, CA), and Actin (cat #612656, BD Biosciences, Franklin Lakes, NJ). Results are representative of at least 3 independent experiments.

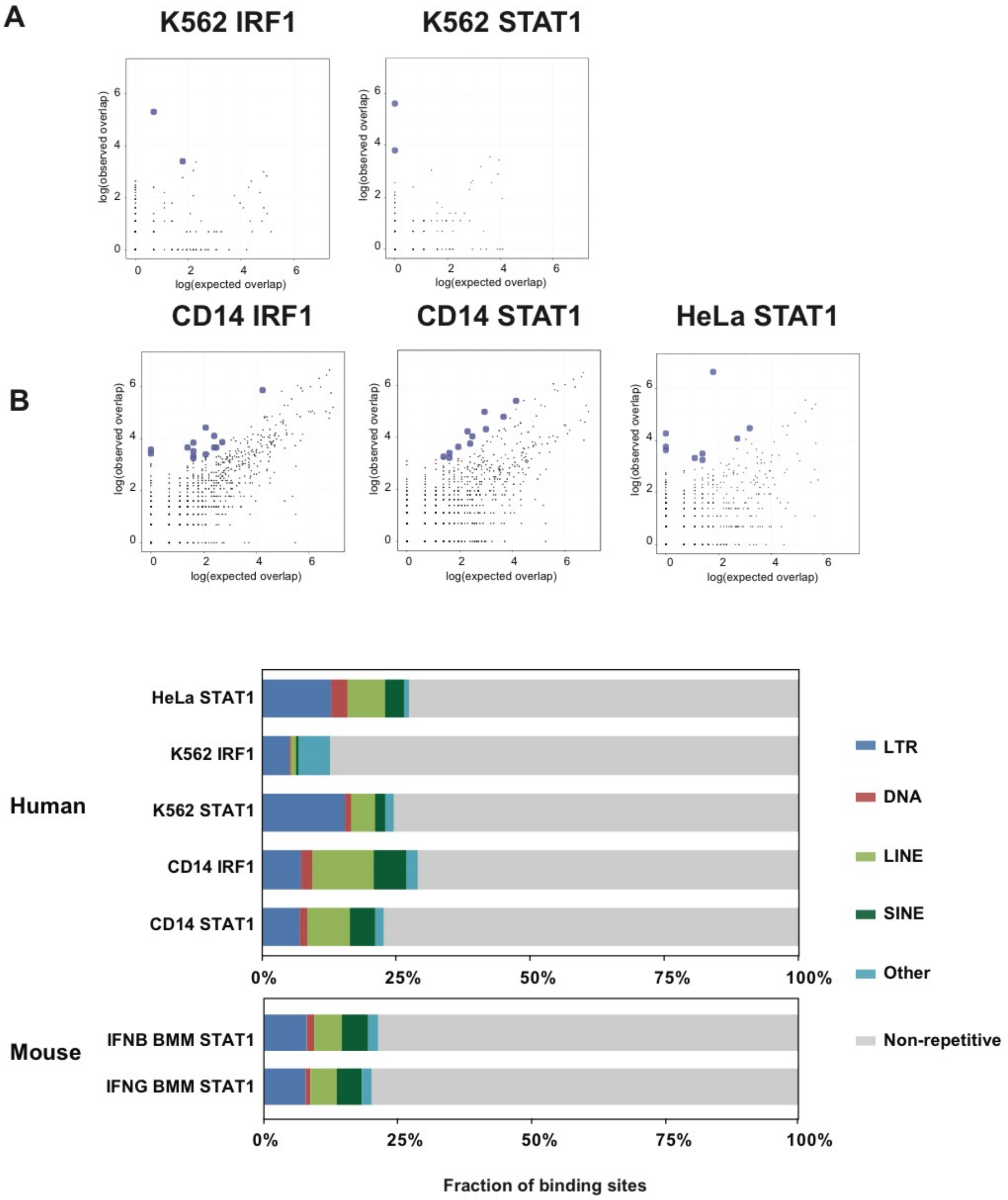


Fig. S1.

A) Scatter plots of TE families (each dot) plotted against the log-normal observed versus expected number of intersections for each ChIP-Seq dataset. Significantly enriched families are shown in blue. See Table S1 for repeat-annotated analysis. B) Fraction of binding sites contained within each major class of TE.

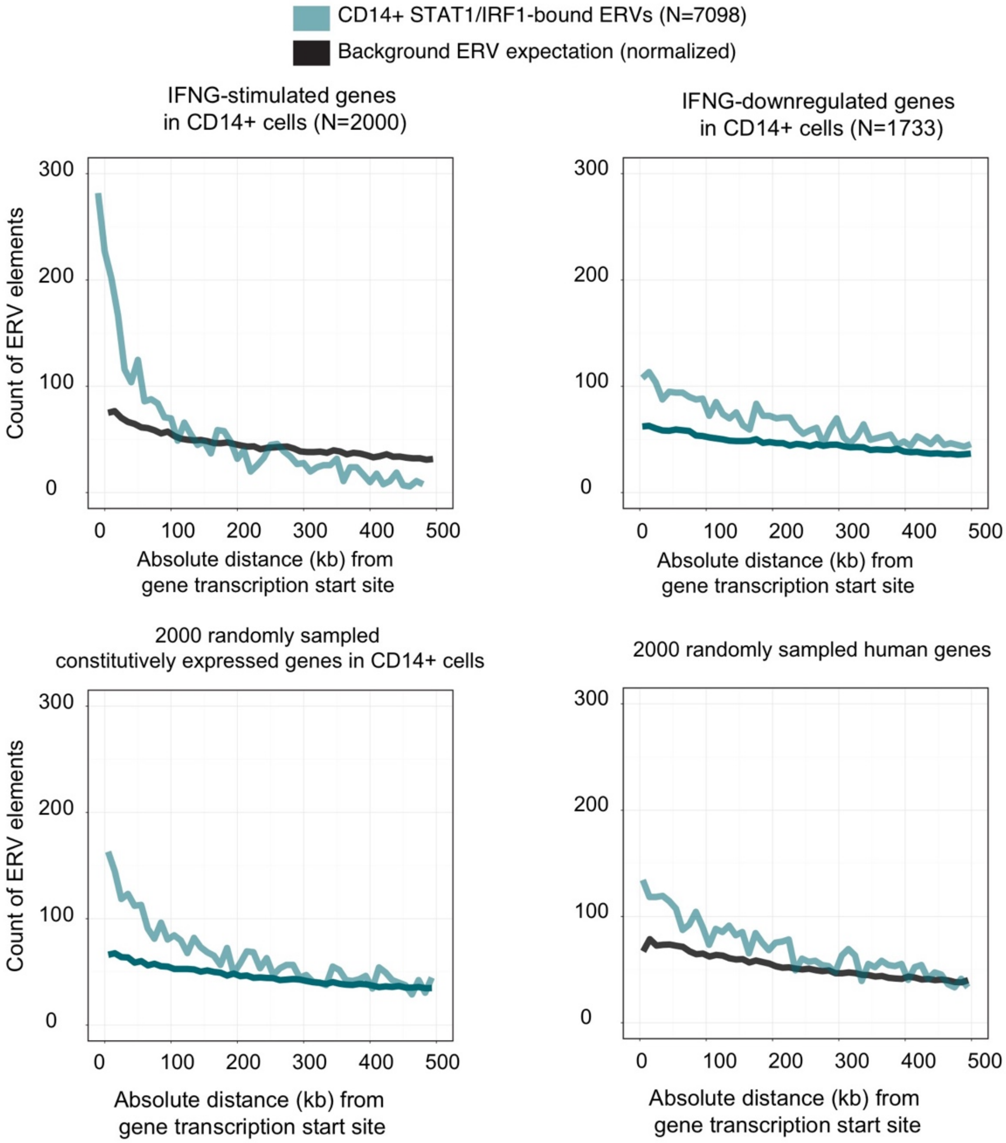


Fig. S2.

Analysis shown in Fig 1B (main text) repeated using other sets of genes for comparison. Each graph depicts frequency distributions of ERV absolute distances to the nearest gene within the particular gene set. 7,098 CD14+ STAT1/IRF1-bound ERVs are compared to a normalized background expectation based on all other unbound ERVs (see Materials and Methods section).

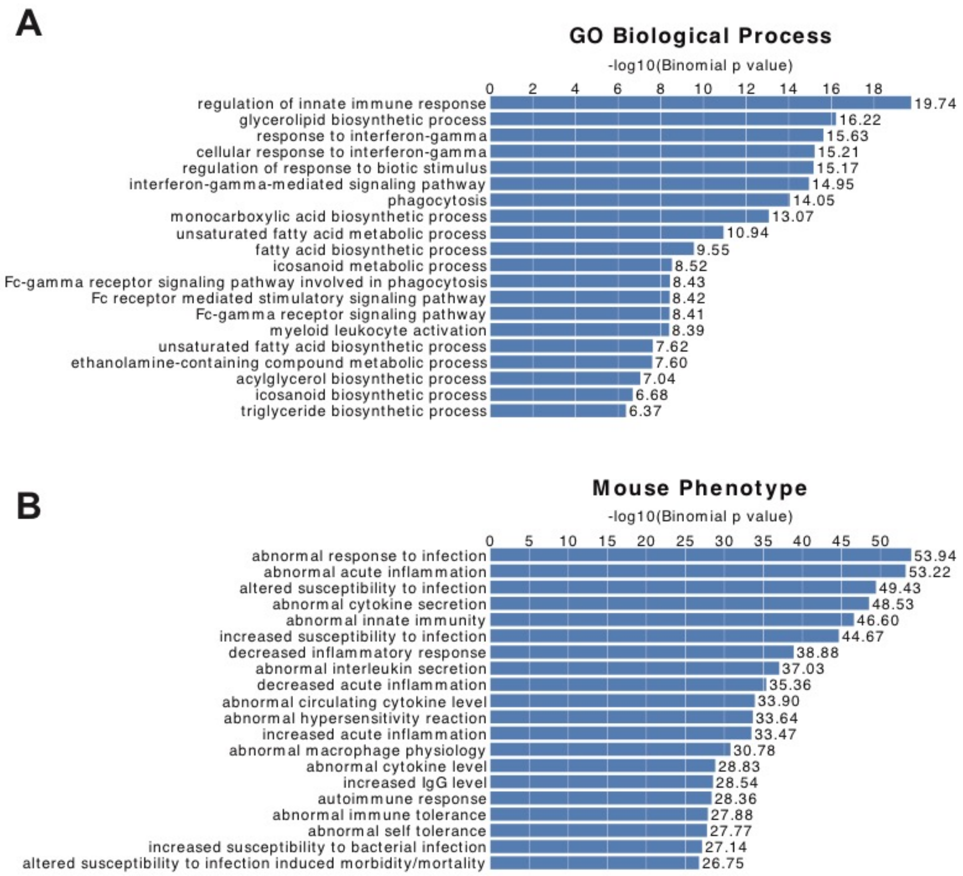


Fig. S3.

IFNG-inducible ERV binding sites are enriched near immunity genes. 7,098 ERVs bound by either IRF1 or STAT1 in IFNG-stimulated CD14+ macrophages were analyzed using the GREAT tool with default settings (http://great.stanford.edu). All displayed categories were significant by both binomial and hypergeometric tests. A) genes with characterized roles in biological processes. B) genes with known knockout phenotypes in mouse.

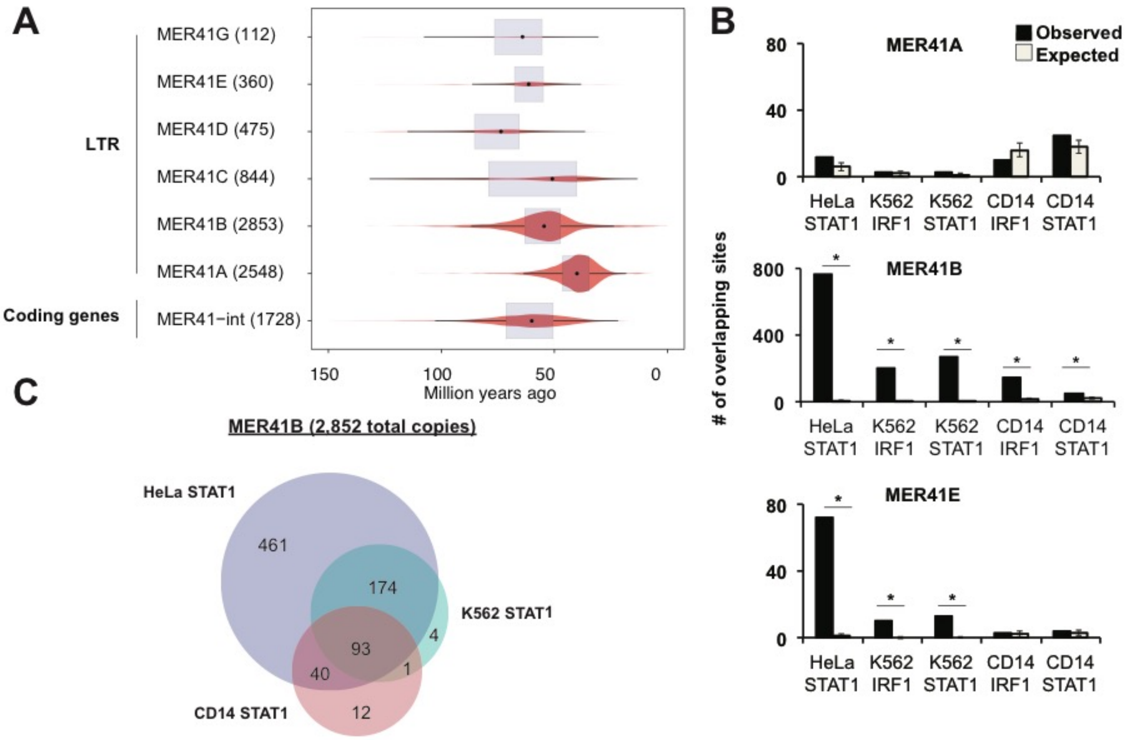
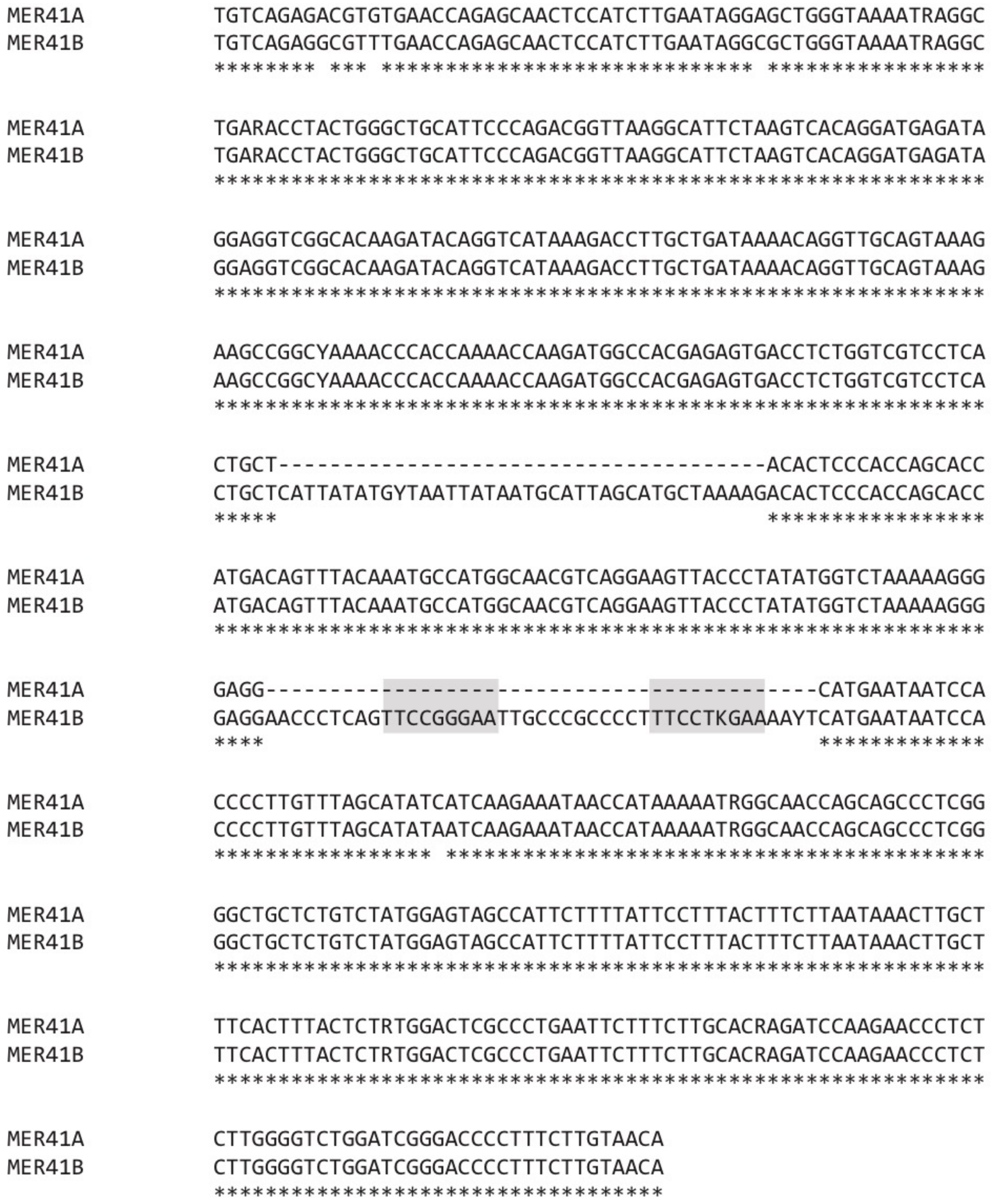


Fig. S4.

Dispersal of IFNG-inducible STAT1 binding sites by MER41 subfamilies. A) Violin plots, scaled to family size (in parenthesis), depicting estimated age distribution of MER41 subfamilies (Methods). Boxplot range indicates 25th and 75th percentiles and inner circle depicts median value. B) Bar plots for each subfamily comparing observed (black) and expected (grey) intersections with each ChIP-Seq dataset analyzed. Expected values are plotted as the mean ± s.d. of intersections from 10,000 shuffled datasets (see Methods). * Bonferroni $P < 0.005$, Binomial test. C) Venn diagram depicting cell-type overlap of individual MER41B elements bound by STAT1.



```
MER41A        TGTCAGAGACGTGTGAACCAGAGCAACTCCATCTTGAATAGGAGCTGGGTAAAATRAGGC
MER41B        TGTCAGAGGCGTTTGAACCAGAGCAACTCCATCTTGAATAGGCGCTGGGTAAAATRAGGC
              ******** *** ***************************** *****************

MER41A        TGARACCTACTGGGCTGCATTCCCAGACGGTTAAGGCATTCTAAGTCACAGGATGAGATA
MER41B        TGARACCTACTGGGCTGCATTCCCAGACGGTTAAGGCATTCTAAGTCACAGGATGAGATA
              ************************************************************

MER41A        GGAGGTCGGCACAAGATACAGGTCATAAAGACCTTGCTGATAAAACAGGTTGCAGTAAAG
MER41B        GGAGGTCGGCACAAGATACAGGTCATAAAGACCTTGCTGATAAAACAGGTTGCAGTAAAG
              ************************************************************

MER41A        AAGCCGGCYAAAACCCACCAAAACCAAGATGGCCACGAGAGTGACCTCTGGTCGTCCTCA
MER41B        AAGCCGGCYAAAACCCACCAAAACCAAGATGGCCACGAGAGTGACCTCTGGTCGTCCTCA
              ************************************************************

MER41A        CTGCT--------------------------------------ACACTCCCACCAGCACC
MER41B        CTGCTCATTATATGYTAATTATAATGCATTAGCATGCTAAAAGACACTCCCACCAGCACC
              *****                                      *****************

MER41A        ATGACAGTTTACAAATGCCATGGCAACGTCAGGAAGTTACCCTATATGGTCTAAAAAGGG
MER41B        ATGACAGTTTACAAATGCCATGGCAACGTCAGGAAGTTACCCTATATGGTCTAAAAAGGG
              ************************************************************

MER41A        GAGG-------------------------------------------CATGAATAATCCA
MER41B        GAGGAACCCTCAGTTCCGGGAATTGCCCGCCCCTTTCCTKGAAAAYTCATGAATAATCCA
              ****                                           *************

MER41A        CCCCTTGTTTAGCATATCATCAAGAAATAACCATAAAAATRGGCAACCAGCAGCCCTCGG
MER41B        CCCCTTGTTTAGCATATAATCAAGAAATAACCATAAAAATRGGCAACCAGCAGCCCTCGG
              ***************** ******************************************

MER41A        GGCTGCTCTGTCTATGGAGTAGCCATTCTTTTATTCCTTTACTTTCTTAATAAACTTGCT
MER41B        GGCTGCTCTGTCTATGGAGTAGCCATTCTTTTATTCCTTTACTTTCTTAATAAACTTGCT
              ************************************************************

MER41A        TTCACTTTACTCTRTGGACTCGCCCTGAATTCTTTCTTGCACRAGATCCAAGAACCCTCT
MER41B        TTCACTTTACTCTRTGGACTCGCCCTGAATTCTTTCTTGCACRAGATCCAAGAACCCTCT
              ************************************************************

MER41A        CTTGGGGTCTGGATCGGGACCCCTTTCTTGTAACA
MER41B        CTTGGGGTCTGGATCGGGACCCCTTTCTTGTAACA
              ***********************************
```

Fig. S5.

Pairwise alignment of the MER41A and MER41B LTR consensus sequences. GAS STAT1 binding sites are highlighted in gray.

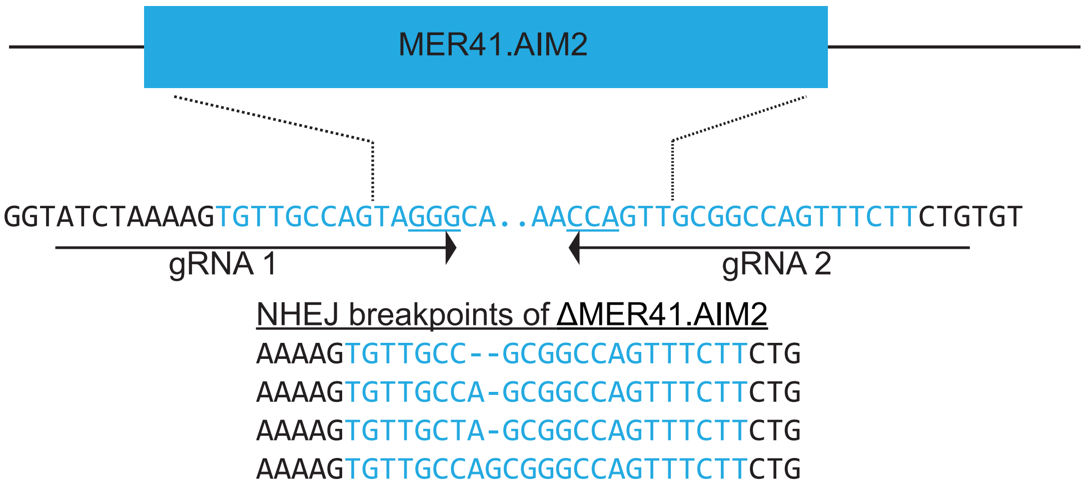
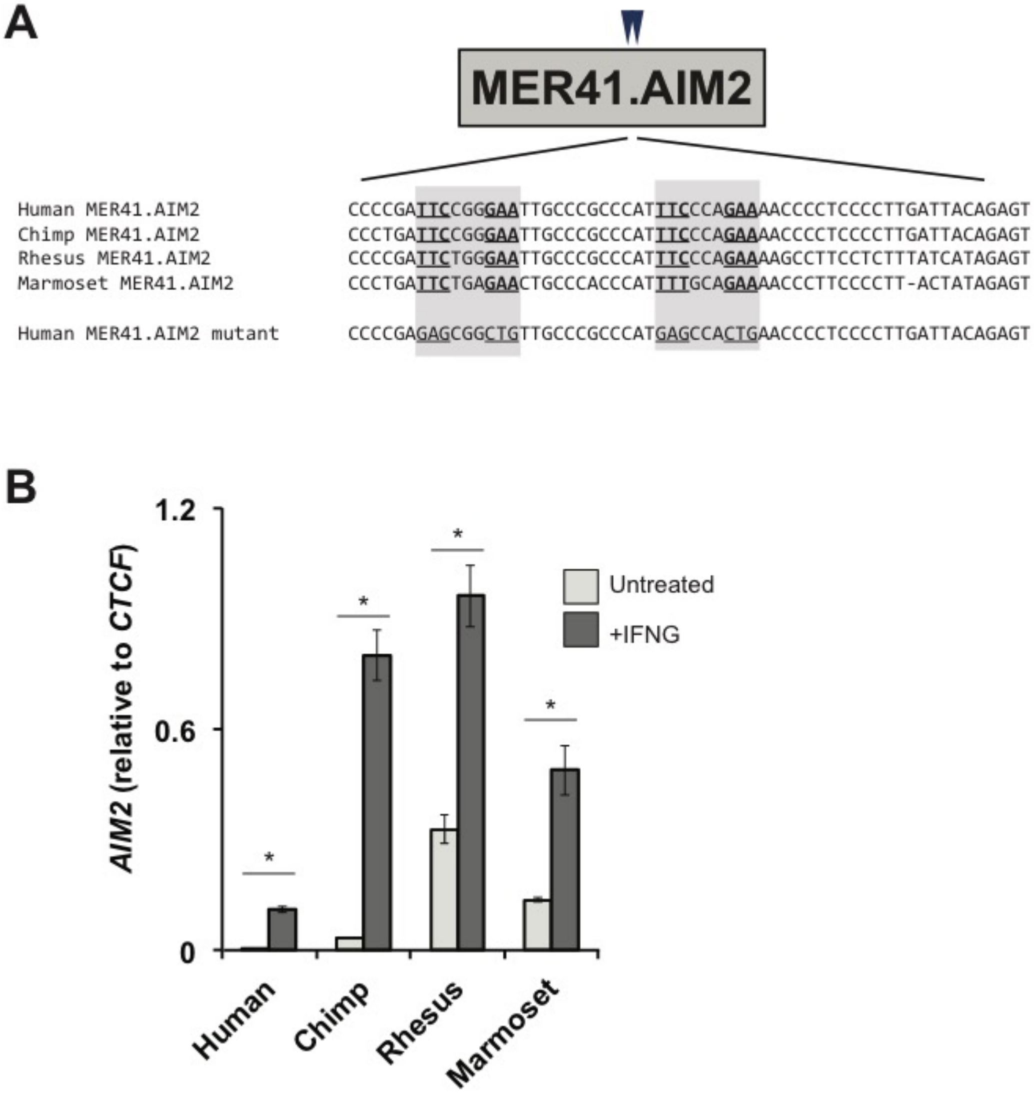

MER41.AIM2
GGTATCTAAAAGTGTTGCCAGTAGGGCA..AACCAGTTGCGGCCAGTTTCTTCTGTGT
gRNA 1
gRNA 2
NHEJ breakpoints of ΔMER41.AIM2
AAAAGTGTTGCC--GCGGCCAGTTTCTTCTG
AAAAGTGTTGCCA-GCGGCCAGTTTCTTCTG
AAAAGTGTTGCTA-GCGGCCAGTTTCTTCTG
AAAAGTGTTGCCAGCGGGCCAGTTTCTTCTG

Fig. S6.

Schematic depicting CRISPR-Cas9 strategy used to delete MER41.AIM2. In our final MER41.AIM2 deletion mutant, four distinct breakpoints were recovered via Sanger sequencing (HeLa cells harbor four copies of the AIM2 locus (*47*)), indicating a homozygous deletion. Arrows indicate gRNA orientation and PAM (NGG) sequences are underlined. Sequences masked as MER41 are in blue. Arrows indicate gRNA orientation and PAM (NGG) sequences are underlined. Sequences masked as MER41 are in blue.



```
Human MER41.AIM2           CCCCGATTCCGGGAATTGCCCGCCCATTTCCCAGAAAACCCCTCCCCTTGATTACAGAGT
Chimp MER41.AIM2           CCCTGATTCCGGGAATTGCCCGCCCATTTCCCAGAAAACCCCTCCCCTTGATTACAGAGT
Rhesus MER41.AIM2          CCCCGATTCTGGGAATTGCCCGCCCATTTCCCAGAAAAGCCTTCCTCTTTATCATAGAGT
Marmoset MER41.AIM2        CCCTGATTCTGAGAACTGCCCACCCATTTTGCAGAAAACCCTTCCCCTT-ACTATAGAGT

Human MER41.AIM2 mutant    CCCCGAGAGCGGCTGTTGCCCGCCCATGAGCCACTGAACCCCTCCCCTTGATTACAGAGT
```

Fig. S7.

MER41.AIM2 is a primate-specific IFNG-inducible enhancer of AIM2. A) Sequence alignment of MER41.AIM2 sequences tested in luciferase reporter assays, encompassing the tandem GAS STAT1 binding motifs. B) qPCR of AIM2 primate orthologs in response to IFNG stimulation, performed in a panel of primate primary fibroblasts. * $p < 0.05$, Student's t-test.

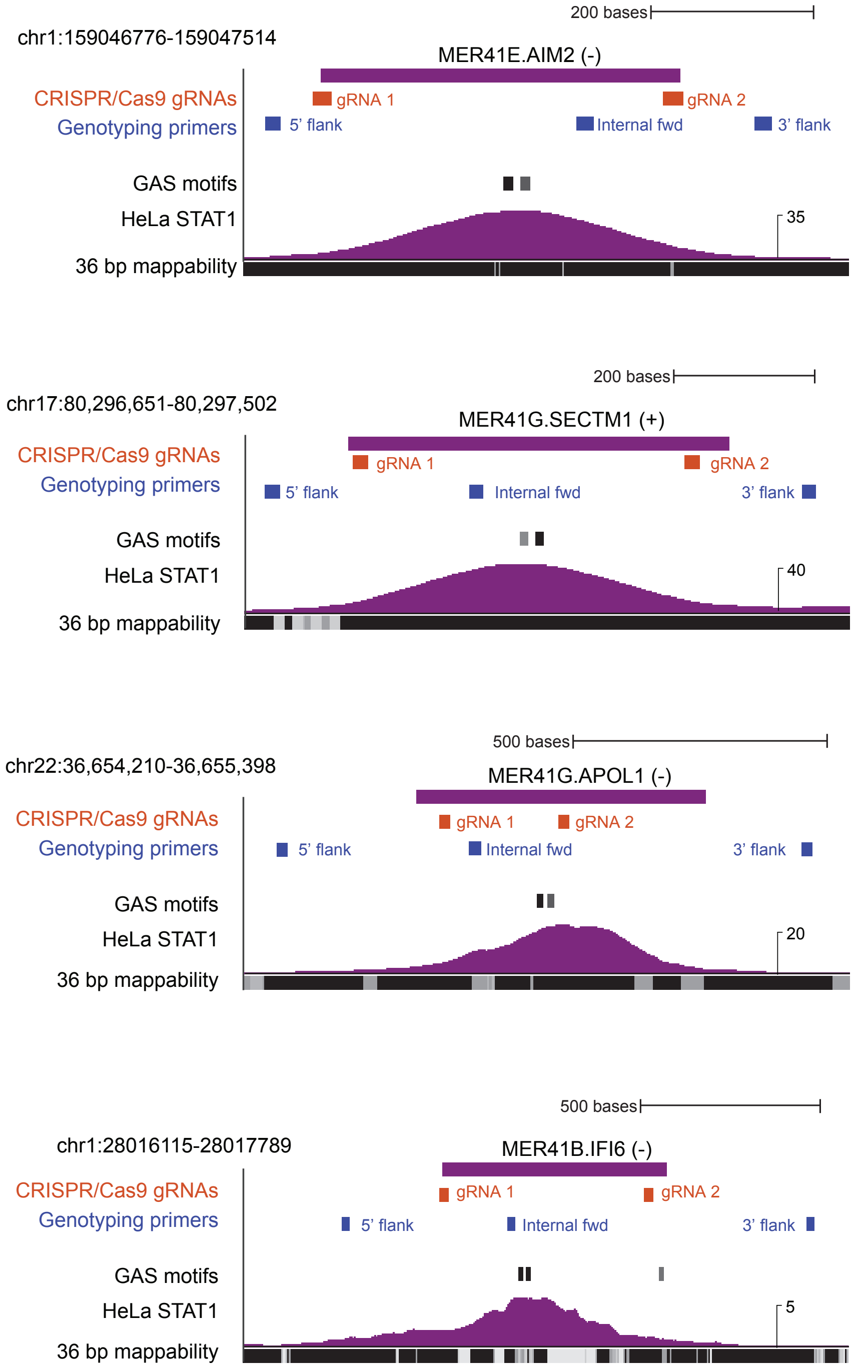


Fig. S8.

Genome browser schematic of primers and guide RNAs used for CRISPR-Cas9 genome editing and genotyping.

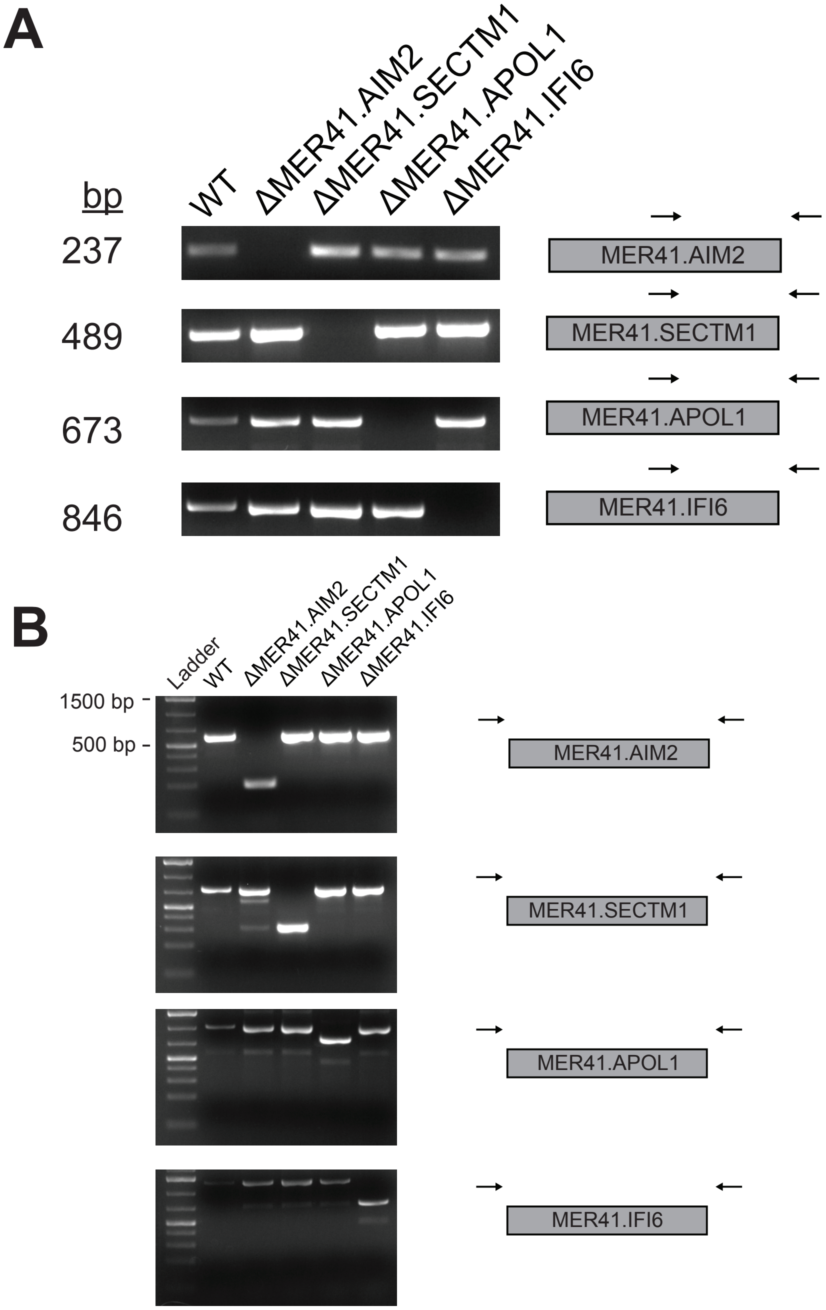
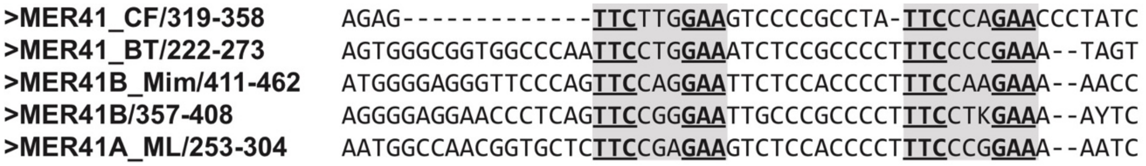


Fig. S9.

PCR validation on wild-type or mutant genomic DNA, using primers either flanking (A) or internal to (B) the expected deletion. Primer locations are depicted in Fig S7 and Table S6.



```
>MER41_CF/319-358      AGAG-------------TTCTTGGAAGTCCCCGCCTA-TTCCCAGAACCCTATC
>MER41_BT/222-273      AGTGGGCGGTGGCCCAATTCCTGGAAATCTCCGCCCCTTTCCCCGAAA--TAGT
>MER41B_Mim/411-462    ATGGGGAGGGTTCCCAGTTCCAGGAATTCTCCACCCCTTTCCAAGAAA--AACC
>MER41B/357-408        AGGGGAGGAACCCTCAGTTCCGGGAATTGCCCGCCCCTTTCCTKGAAA--AYTC
>MER41A_ML/253-304     AATGGCCAACGGTGCTCTTCCGAGAAGTCTCCACCCCTTTCCCGGAAA--AATC
```

Fig. S10.

MER41-like ERVs independently amplified in other lineages. Alignment of consensus sequences from selected MER41-like relatives that exhibit conservation of the GAS (TTCNNNGAA) STAT1 binding motifs. MER41_CF (dog Repeatmasker annotation) is present in all carnivores, MER41_BT (cow) in artiodactyls, MER41B_Mim (mouse lemur) in lemuriformes, MER41B (human) in anthropoid primates, and MER41A_ML (little brown bat) in vesper bats.

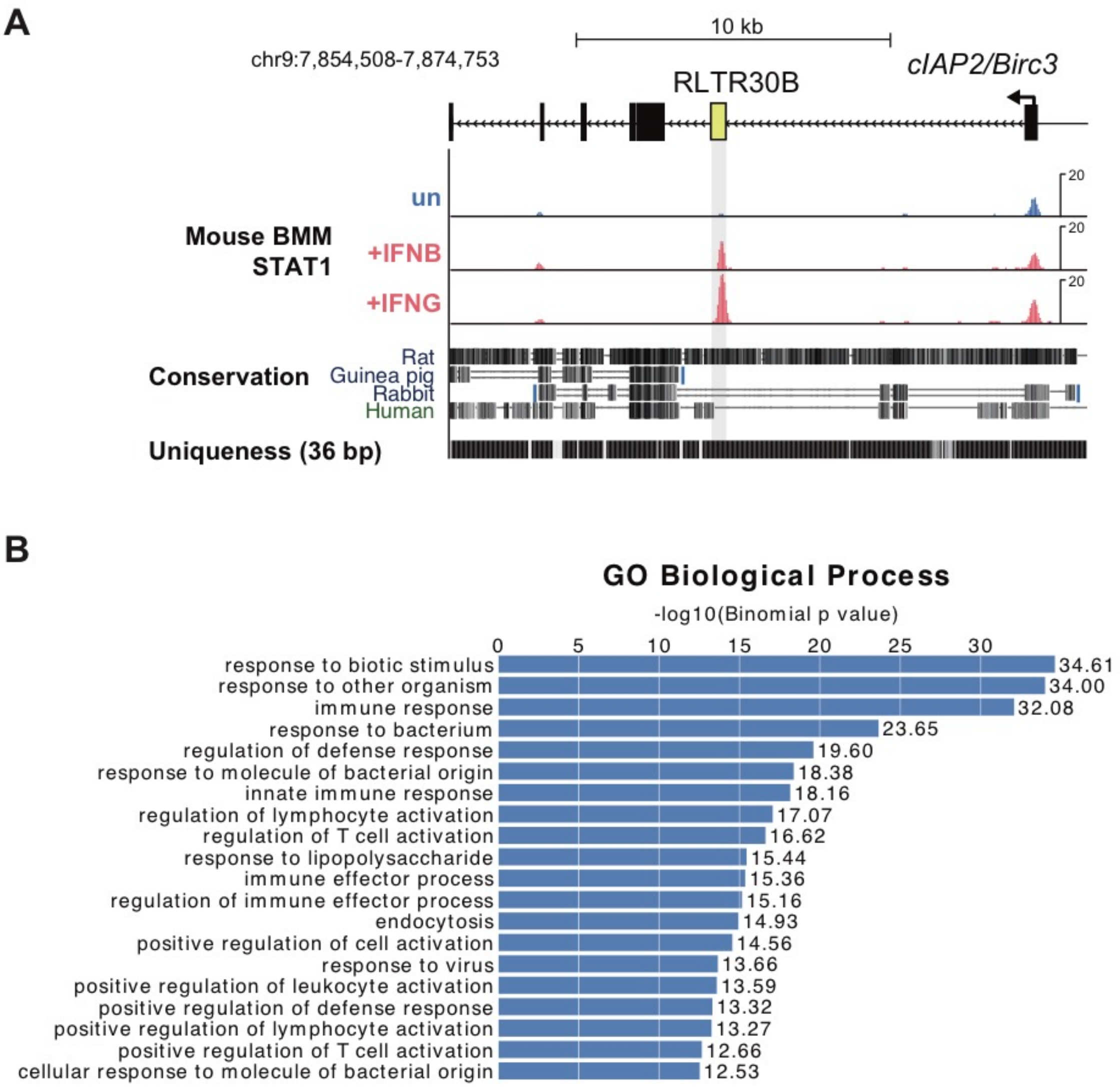


Fig. S11.

A) Genome browser view of an RLTR30B element inducibly bound by STAT1 in response to either IFNB or IFNG in mouse BMM. cIAP2 is an inhibitor of apoptosis with reported roles in innate immunity (*48*), and has previously been proposed to be regulated by retrotransposons (*49*). B) GREAT analysis of 3,336 ERVs bound by STAT1 in IFNG-stimulated mouse macrophages.

Supplementary Tables

Additional Data table S1 (separate file)

TE enrichment analyses for human and mouse ChIP-Seq datasets.

Additional Data table S2 (separate file)

Analysis of STAT1/IRF1-bound ERVs and IFNG-stimulated genes in primary CD14+ macrophages.

Additional Data table S3 (separate file)

List of MER41 and RLTR30B repeats bound by TFs and their closest gene.

Additional Data table S4 (separate file)

BLAST analysis of presence/absence of MER41-like elements across mammalian species representative of each major eutherian lineage.

Additional Data table S5 (separate file)

Uniformly analyzed ChIP-Seq peaks used in this study.

Additional Data table S6 (separate file)

Sequences used in this study (CRISPR-Cas9 gRNAs, qPCR and genotyping primers, synthesized LTR promoters).