Corrected 13 April 2021. See full text.

RESEARCH ARTICLE SUMMARY

EVOLUTION

# Recurrent evolution of vertebrate transcription factors by transposase capture

Rachel L. Cosby, Julius Judd, Ruiling Zhang*, Alan Zhong*, Nathaniel Garry, Ellen J. Pritham, Cédric Feschotte†

**INTRODUCTION:** How novel protein architectures evolve remains poorly understood. The rearrangement of domains with preexisting functions into new composite architectures through exon shuffling is a powerful path to form genes encoding proteins with novel functionalities. Although exon shuffling is thought to account for the evolution of many protein structures, the source of new exons and splice sites as well as the mechanisms by which they become assimilated have been scarcely characterized. In this work, we investigate the contribution of DNA transposons to the formation of novel protein-coding genes through exon shuffling during vertebrate evolution.

**RATIONALE:** DNA transposons are widespread mobile elements encoding transposase proteins that promote their selfish replication in host genomes. Transposases typically contain DNA binding and catalytic nuclease domains, which may be repurposed for cellular functions. By inserting functional domains into new genomic contexts, transposase sequences can generate host-transposase fusion (HTF) genes through alternative splicing. Several genes with critical developmental functions, such as the *Pax* transcription factors, are thought to have been born through this process. However, the mechanism by which transposase domains are captured to generate HTFs, how common this process is, and the functions of most known HTF genes remain unclear.

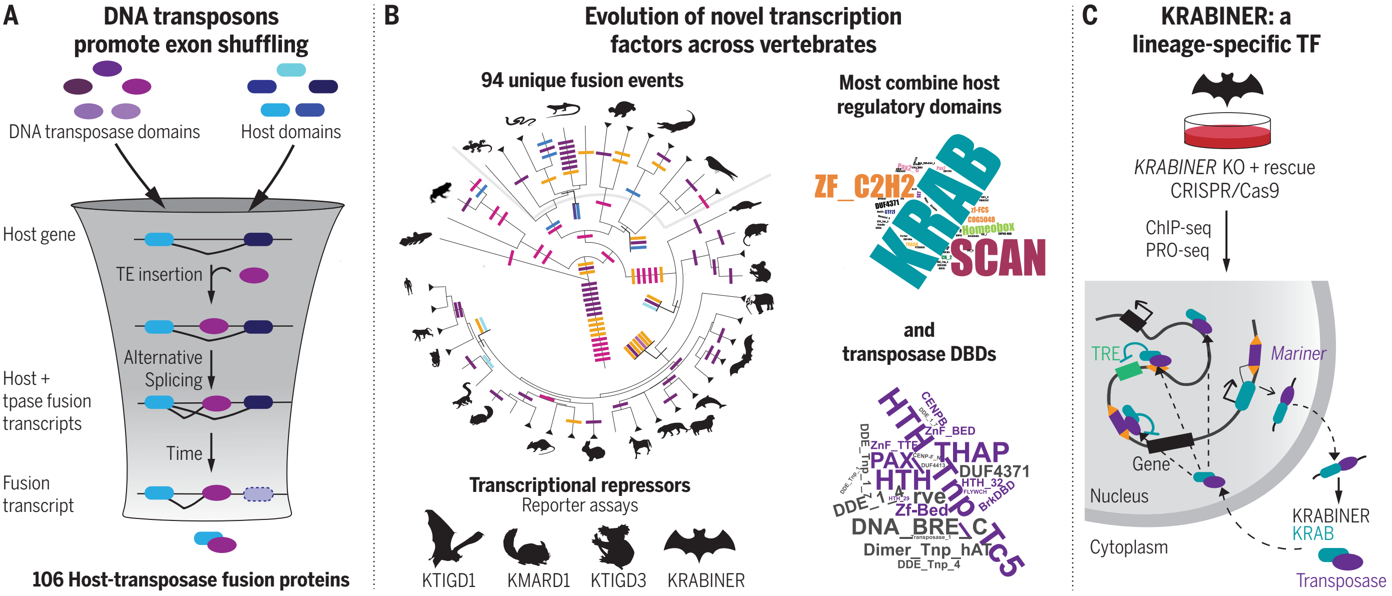
**RESULTS:** We used comparative genomics to survey all tetrapod genomes with available gene models (596) for putative HTFs. We identified 106 distinct HTFs derived from 94 independent fusion events over the course of ~300 million years of evolution. We found that most HTFs evolved through the alternative splicing of host domains to transposase proteins using splice sites provided by the transposon. The transposase domains of all HTFs analyzed (81) are evolving under purifying selection, which suggests that they have been maintained for organismal function. The domain composition of HTF proteins indicates that most of them consist of transposase DNA binding domains fused to host domains that are predicted to function in transcriptional and/or chromatin regulation, especially the repressive Krüppel-associated box (KRAB) domain (involved in ~30% of all HTFs), which suggests that many HTFs function as transcriptional regulators. Supporting this hypothesis, we show that four independently evolved KRAB-transposase fusion proteins repress gene expression in a sequence-specific manner in reporter assays. Furthermore, loss of function, rescue, and regulatory genomics experiments in bat cells revealed that the bat-specific KRABINER fusion protein binds hundreds of cognate transposons genome-wide and controls a large network of genes and cis-regulatory elements.

**CONCLUSION:** Our findings confirm that exon shuffling is a major evolutionary force generating genetic novelty. We provide evidence that DNA transposons promote exon shuffling by inserting transposase domains in new genomic contexts. This process provides a plausible path for the emergence of several ancient transcription factors with important developmental functions. By illustrating how a transcription factor and its dispersed binding sites can emerge simultaneously from a single transposon family, our results bolster the view that transposons are key players in the evolution of gene regulatory networks. ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: cf458@cornell.edu

Cite this article as R. L. Cosby *et al.*, *Science* **371**, eabc6405 (2021). DOI: 10.1126/science.abc6405

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abc6405



**Transposase capture contributes to the evolution of transcription factors by combining DNA transposase and host domains.** (**A**) Model for how transposase capture occurs. (**B**) Abundance and characteristics of identified HTFs. (**C**) Summary of KRABINER's role as a transcription factor (TF) in bat cells. TE, transposable element; tpase, transposase; DBDs, DNA binding domains; KO, knockout; ChIP-seq, chromatin immunoprecipitation sequencing; PRO-seq, precision run-on sequencing; TRE, transcribed regulatory element.

RESEARCH ARTICLE

EVOLUTION

# Recurrent evolution of vertebrate transcription factors by transposase capture

Rachel L. Cosby[1], Julius Judd[1], Ruiling Zhang[1]*, Alan Zhong[1]*, Nathaniel Garry[1],
Ellen J. Pritham[2], Cédric Feschotte[1]†

Genes with novel cellular functions may evolve through exon shuffling, which can assemble novel protein architectures. Here, we show that DNA transposons provide a recurrent supply of materials to assemble protein-coding genes through exon shuffling. We find that transposase domains have been captured—primarily via alternative splicing—to form fusion proteins at least 94 times independently over the course of ~350 million years of tetrapod evolution. We find an excess of transposase DNA binding domains fused to host regulatory domains, especially the Krüppel-associated box (KRAB) domain, and identify four independently evolved KRAB-transposase fusion proteins repressing gene expression in a sequence-specific fashion. The bat-specific KRABINER fusion protein binds its cognate transposons genome-wide and controls a network of genes and cis-regulatory elements. These results illustrate how a transcription factor and its binding sites can emerge.

Gene duplications contribute to the birth and functional diversification of many genes (1), including developmental regulators (2, 3). Although gene duplicates can evolve diverging developmental functions relative to their parental gene, their domain architectures and biochemical activities tend to remain the same (2, 3), and proteins with novel biochemical functions arising through gene duplication (4, 5) appear to be rare (6). Although completely new proteins can occasionally evolve de novo from previously noncoding sequences (7, 8), the most obvious path to forming proteins with new functionalities is the rearrangement of domains with preexisting functions into new composite architectures through exon shuffling. Exon shuffling occurs when new combinations of exons are assembled through RNA splicing, and it may have created new protein architectures in eukaryotic evolution (9). Although the process of exon shuffling may account for the evolution of many new protein architectures (10–12), the source of new exons and the mechanisms by which they become assimilated have been scarcely characterized. Here, we investigate the role of DNA transposons as a source of raw material for the birth of novel proteins through exon shuffling.

DNA transposons encode transposase proteins, which recognize and mobilize DNA through direct sequence-specific interaction with their cognate transposons (13). The canonical architecture of transposase proteins consists of a DNA binding domain and a catalytic nuclease domain. Both domains may be repurposed or domesticated for cellular function (13). Moreover, the mobility of DNA transposons may facilitate exon shuffling by inserting these functional domains into new genomic contexts, where they may be spliced to generate host-transposase fusion (HTF) genes.

Three genes born via transposase capture have been documented: one specific to placental mammals [*GTF2IRD2* (14)] and two specific to primates [*SETMAR* (15) and *PGBD3*-CSB (16)]. A similar scenario has been proposed to explain the origin of the paired DNA binding domain of the *Pax* family of transcription factors (TFs). However, because of the deep ancestry of *Pax* genes, which coincides with the emergence of metazoans, the precise steps by which these factors evolved has been obscured (13, 17). Further, the extent of transposase capture, the mechanisms facilitating it, and the functions of the resulting genes often remain unclear.

## Transposase capture is a recurrent mechanism for novel gene formation in tetrapods

To identify HTF genes, we surveyed all tetrapod gene annotations [using the National Center for Biotechnology Information Reference Sequence (NCBI Refseq) database; table S1] that are predicted to encode proteins with at least one domain of transposase origin (Pfam; table S2) fused in-frame to a host-derived protein sequence [using the Conserved Domain Architecture Retrieval Tool (CDART)] (18). We also required RNA sequencing (RNA-seq) evidence supporting all annotated exon-intron junctions (19). To trace the evolutionary origin of each HTF gene, we used a homology-based approach (BLASTn) to query all vertebrate genomes available in the NCBI Refseq database for syntenic orthologs and paralogs. An HTF was considered to be orthologous in two or more lineages if it fulfilled the following criteria: (i) had a hit containing both the transposase domain and the host domain in the same transcript (nr/nt database) or in the same orientation on the same contig (Refseq genomes database) and (ii) was located in a syntenic region of the genome, determined by the identity of flanking genes. This analysis yielded 106 individual HTF genes originating from 94 independent fusion events and 12 subsequent duplication events across the 596 species we examined (Fig. 1 and table S3).

Placing fusion genes onto the species phylogeny suggests that they have evolved continuously during evolution (Fig. 1). Some fusion events (11.6%) preceded the divergence of tetrapods [>350 million years (Ma) ago], whereas others, conserved across narrow species lineages (<5 species; 26.1%) or found in a single species (21%), arose more recently (Fig. 1 and table S3). Several species' lineages experienced multiple HTFs of recent origins, as in the green anole ($n = 6$), the Burmese python ($n = 3$), the tropical clawed frog ($n = 2$), and the vespertilionid bats ($n = 2$), which is consistent with recent episodes of DNA transposon activity in these lineages (Fig. 1) (20–24). Mammals generally have more HTF genes (mean = 40.8 ± 3.55) than other clades (reptiles, mean = 29.3 ± 1.53; amphibians, mean = 30 ± 2.65), which reflects apparent bursts of HTF evolution in mammalian (5.3% of all events), therian (3.2%), and eutherian (8.4%) ancestors. All known major eukaryotic DNA transposon superfamilies contribute to HTFs (Fig. 1), but Tc1/*mariner* (36.1%), hAT (23.4%), and P element/Kolobok (21.3%) transposases predominate, which mirrors the success of these superfamilies throughout tetrapod evolution (13, 25).

To validate that the transposase coding region of each HTF gene has evolved under functional constraint, we performed codon selection analysis on the transposase domain of each HTF shared by two or more species separated by >50 Ma of divergence. All tested HTFs ($n = 81$) display signatures of purifying selection on their transposase domains [ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site (dN/dS) < 1; $P < 0.05$, likelihood ratio test (LRT)] (table S3), which supports their domestication for organismal function.

To further assess the functional capacity of HTF genes, we used publicly available data to examine the RNA expression patterns across 54 tissues for 44 HTFs present in the human genome (GTEx portal). Each of the genes were expressed [transcripts per million (TPM) > 1] in at least one human tissue and could be

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850, USA. [2]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: cf458@cornell.edu

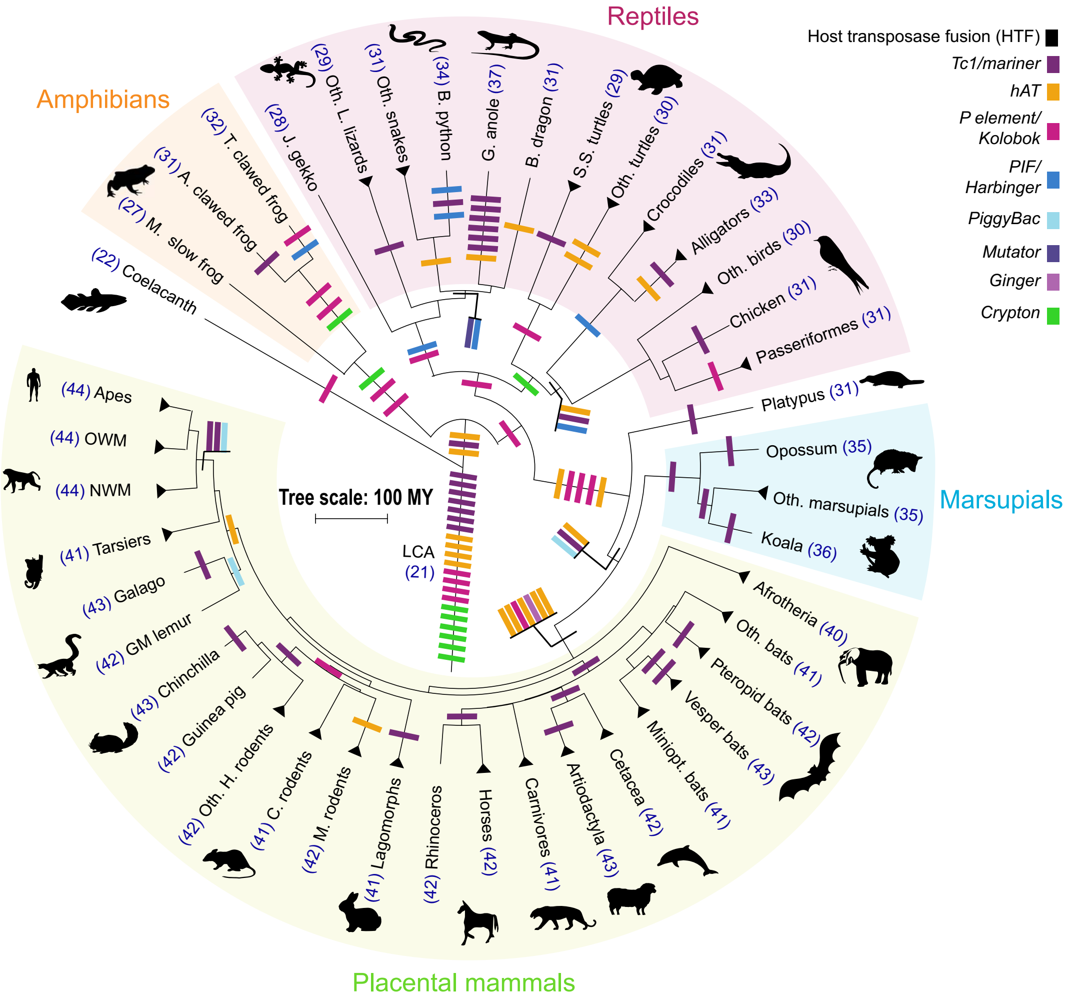


**Fig. 1. Gene birth by transposase capture in tetrapods.** Tetrapod phylogenetic tree with boxes representing HTF fusion genes. Colors indicate the transposase superfamily assimilated. Numbers in parentheses indicate the number of HTF genes identified in the specified lineage. OWM, Old World monkeys; NWM, New World monkeys; GM, gray mouse; Oth., other; H., hystricoid; C., castorid; M., muroid; Miniopt., miniopterid; Vesper, vespertilionid; S.S., soft-shelled; B., bearded dragon; G., green; B., Burmese python; L., lacertid; J., Japanese; T., tropical; A., African; M., mountain; LCA, last common ancestor; MY, million years.

classified into three categories: lowly expressed broadly (TPM > 1 in 80% of the tissues, $n$ = 23), highly expressed broadly (TPM > 10 in 80% of tissues, $n$ = 12), and tissue-restricted (TPM > 1 in <20% tissues, $n$ = 9) (fig. S1). These data suggest that most human HTF genes are broadly expressed and may function in a variety of contexts. Collectively, these data demonstrate that HTF has been a recurrent mechanism for the generation of novel cellular genes in tetrapod evolution, including at least 44 HTFs in humans.

### Transposase capture occurs through alternative splicing

To illuminate the mechanism by which transposase domains are captured to form new chimeric proteins, we examined the gene structure of HTFs. In all cases, the transposase-derived domains are encoded by exons distinct from the host domains, which suggests that transposase capture occurred through splicing events. To further delineate the process, we investigated in detail the birth of *KRABINER*, a recently evolved HTF in vespertilionid bats. *KRABINER* is predicted to encode a 447–amino acid protein consisting of an N-terminal Krüppel-associated box (KRAB) domain fused to a full-length *Mlmar1 mariner* DNA transposase (Fig. 2A). Using a combination of comparative genomics, polymerase chain reaction (PCR), and reverse transcription polymerase chain reaction (RT-PCR) (*19*), we inferred that *KRABINER* originated in the common ancestor of the nine vespertilionids examined, but after their divergence from miniopterids—~45 Ma ago—through the following steps: (i) *mariner* insertion into the last intron of *ZNF112*, a gene present in all eutherian mammals; (ii) alternative splicing to the upstream exons of *ZNF112* using a splice acceptor site preexisting in the ancestral *Mlmar1 mariner* transposon; and (iii) a unique single nucleotide deletion in the transposase coding sequence that generated a single open reading frame coding the chimeric protein (fig. S2 and Fig. 2B). This sequence of events is similar to the process that resulted in *SETMAR* (*15*) and *PGBD3-CSB* (*16*) originating in the primate lineage, and it suggests that DNA transposons have features that may facilitate their capture through alternative splicing. Alternatively, acquisition of a splice site by a DNA transposon may increase its ability to generate an HTF gene.

We surveyed all HTF gene models for evidence of alternative splicing and found unequivocal evidence for the coexistence of both fusion and parental gene transcripts for most of the young HTFs (18 of 33 HTFs <100 Ma old). As HTFs were retained within genomes over time, the ancestral parental transcripts were lost, and only the fusion transcript was generally detected (Fig. 2C and table S3). These findings suggest that most HTFs are born as alternatively spliced variants of an ancestral gene, but over time the HTF transcript often becomes the primary or sole transcript for that gene. Thus, alternative splicing may represent a mechanism for the assimilation of transposase domains by the host proteome.

The splice site enabling the capture of *KRABINER*'s transposase was provided by the ancestral *Mlmar1 mariner* transposon (fig. S2C). To investigate whether this is a more general mechanism of novel gene evolution, we selected eight additional HTFs derived from recent transposon families to trace the origin of the splice site used for transposase

Corrected 13 April 2021. See full text.

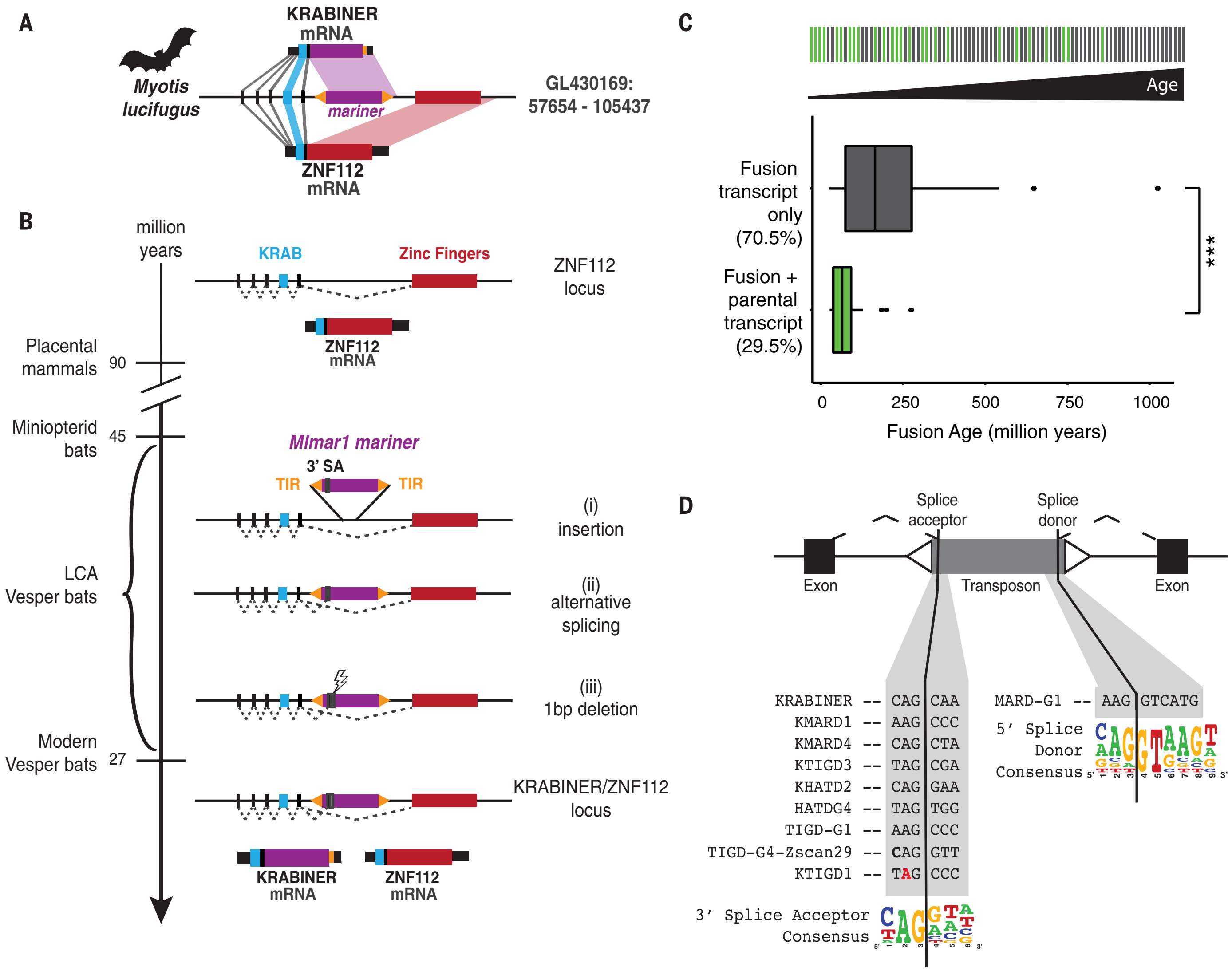


**Fig. 2. Transposase capture by alternative splicing.** (**A**) *ZNF112/KRABINER* locus in vespertilionid bats. (**B**) Steps required for *KRABINER* birth. (**C**) Age of fusion genes with (green) or without (gray) evidence for alternative splicing. Fusion age (bottom) is determined by the midpoint of the age range for each fusion, as described in table S3; the top shows a qualitative illustration of host transcript loss over time. (**D**) Summary of transposon splice site usage for nine HTFs, with canonical mammalian splice sites shown as a sequence logo. Red denotes nucleotides in the splice site that diverge from the transposon consensus sequence. SA, splice acceptor. ***$P < 0.01$; two-sample Wilcoxon test.

capture. In all cases, we found that the splice site was directly derived from the transposon sequence. We then generated a majority-rule consensus sequence for each family to approximate the ancestral transposon [data 1 (26)]. For six of eight HTFs, the splice site sequence was strictly identical to that of the consensus sequence; in the remaining two, the splice site differed from the consensus by a single substitution (Fig. 2D). Though we cannot exclude the possibility of independent splice-site acquisition, these results suggest that they preexisted in the ancestral transposon.

### Fusion of transposase DNA binding domains to KRAB is prevalent

To explore the cellular function of HTFs, we first characterized their protein domain architecture and composition (Fig. 3A and fig. S3). Among transposon-derived domains, DNA binding domains predominate (76.5%; fig. S3), although some HTFs also include catalytic or accessory transposase domains (fig. S3). Among host domains (i.e., not normally found in transposases), we identified 55 distinct conserved domains, most of which (76%) were involved in a single fusion event (Fig. 3A). Several of the host domains are predicted to function in transcriptional and/or chromatin regulation, such as the KRAB, SET, and SCAN domains (27–29). KRAB was the host domain most frequently fused to transposase: We inferred this domain to have been involved in 32 independent fusion events across the phylogeny, accounting for approximately one-third of all HTFs (Fig. 3A). KRAB domains are abundant in tetrapod genomes and most commonly found in KRAB–zinc finger proteins (KRAB-ZFPs), an exceptionally diverse family of TFs (>200 genes in most tetrapod genomes; 487 in humans) (30). Although it is possible that the frequency of KRAB-transposase fusions reflects the natural abundance of KRAB domains, the identification of two independent KRAB-transposase fusions in bird genomes, despite their paucity in KRAB-ZFPs (approximately eight genes per genome) (30), suggests that the combination of KRAB and transposase may be evolutionarily adaptive.

### KRAB-transposase fusions act as sequence-specific repressors of gene expression

Given the prevalence of KRAB-transposase fusions and the canonical function of KRAB

Corrected 13 April 2021. See full text.

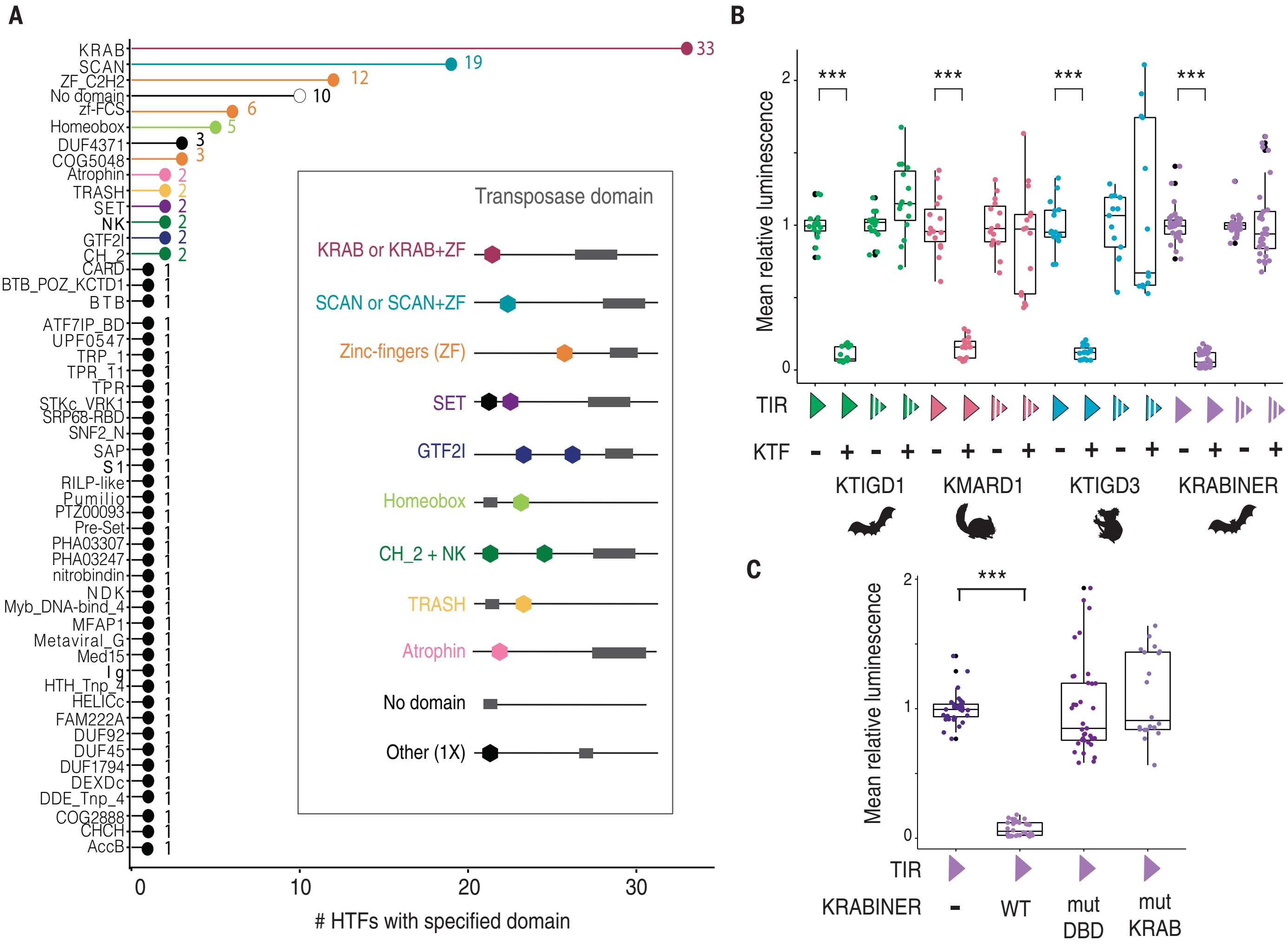


**Fig. 3. Biochemical activities of HTF proteins.** (**A**) Diverse host domains are fused to transposases. The *x* axis specifies the number of HTF genes a given domain is present in; some fusions contain more than one domain. The inset shows a representative domain architecture schematic for select HTFs. (**B**) KRAB-transposase fusions repress gene expression in a sequence-specific manner. (**C**) KRABINER requires both its KRAB and DBD domains to repress gene expression. The *y* axes in (B) and (C) correspond to mean luminescence relative to the empty vector control for each comparison ($n \geq 15$). KTF, KRAB-transposase fusion. Filled triangles indicate consensus TIRs, and interrupted triangles indicate scrambled TIRs. The plus and minus signs indicate the presence or absence, respectively, of each KTF. Adjusted ***$P < 0.001$; two-sample Wilcoxon test, Holm-Bonferroni correction.

domains in establishing silent chromatin when tethered to DNA (27), we next used these genes as a paradigm to test the hypothesis that transposase fusion creates novel sequence-specific transcriptional regulators. We selected four recently emerged KRAB-transposase fusions for which we had generated consensus sequences of their cognate transposon family, enabling us to identify their terminal inverted repeats (TIRs), which typically contain the transposase binding site (fig. S4). We cloned the consensus sequence of each TIR or a scrambled version upstream of a firefly luciferase reporter and measured luciferase expression in HEK293T cells in the presence or absence of a vector expressing the cognate HTF protein. Each KRAB-transposase fusion protein repressed luciferase expression in the presence of its cognate intact TIR but not the scrambled sequence (Fig. 3B). These results indicate that KRAB-transposase proteins can repress gene expression in a sequence-specific manner.

To test whether KRAB-transposase repression is dependent on KAP1 (TRIM28), the transcriptional corepressor often recruited by the KRAB domain (27), we repeated the reporter assays in HEK293T cells that lack *KAP1* (31). The results (fig. S5) show that repression by KMARD1 and KTIGD1 is dependent on KAP1, whereas KRABINER and KTIGD3 are only partially dependent on KAP1.

To further dissect the requirement of individual domains, we generated two mutant versions of KRABINER by altering residues predicted to compromise DNA binding activity (mutDBD) or the function of the KRAB domain (mutKRAB). To generate the DBD mutant, we exploited the similarity of KRABINER's *mariner* transposase to that of *Mos1* (23), a transposon from *Drosophila*. Electrophoretic mobility shift assays demonstrated that a single point mutation in the first helix-turn-helix motif of the *Mos1* transposase was sufficient to abolish binding to its TIR (32). We mutated the homologous site in KRABINER's DBD, as well as three additional residues that directly contact TIR DNA in the *Mos1* transpososome crystal structure (33) (fig. S4B). To generate the KRAB mutant, we introduced several point mutations altering conserved residues previously identified as critical for KRAB-mediated repression (34–37). Although the mutDBD and mutKRAB proteins were expressed at comparable levels to wild-type (WT) KRABINER (fig. S4C), both failed to repress reporter gene expression (Fig. 3C). Together, the results of these reporter assays support the hypothesis

Corrected 13 April 2021. See full text.

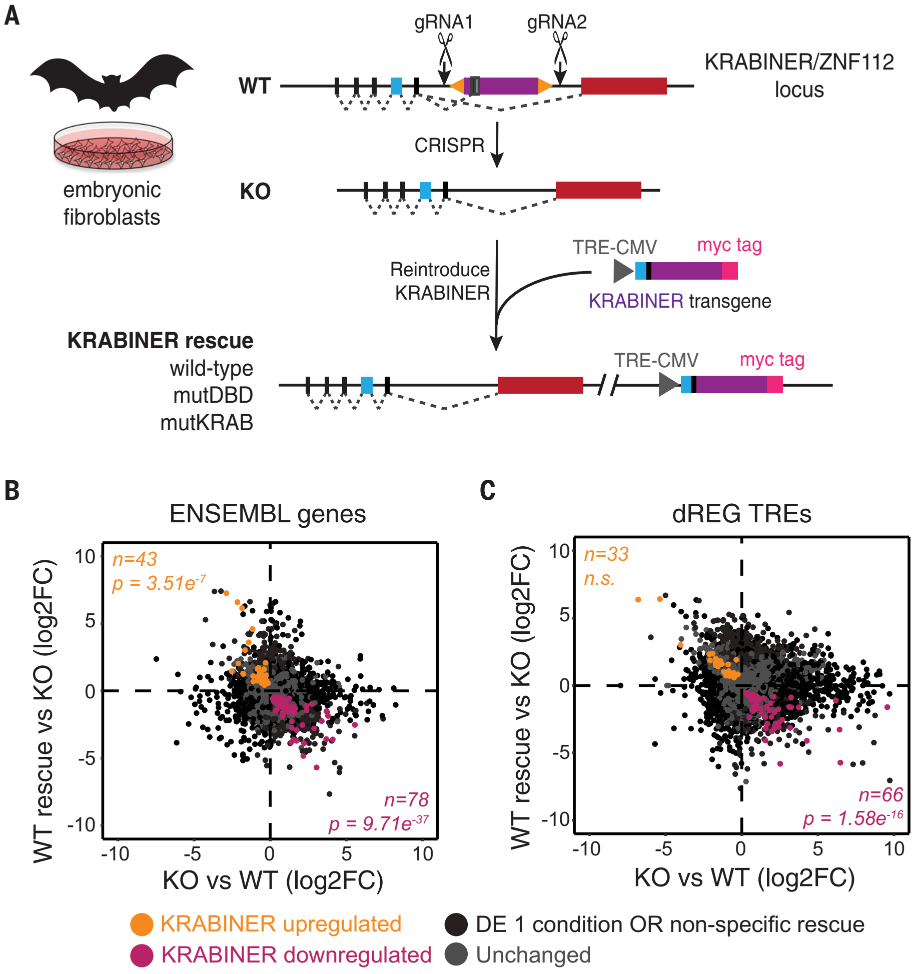


**Fig. 4. KRABINER regulates transcription of genes and TREs in bat cells.** (**A**) Strategy to generate KRABINER KO and rescue lines. TRE, tet responsive element; CMV, cytomegalovirus. (**B** and **C**) Summary of transcriptional changes of genes and TREs, respectively, upon loss and restoration of KRABINER. KRABINER-regulated genes (up- or down-regulated) change reciprocally between KO versus WT and WT KRABINER rescue versus KO comparisons. *P* values were calculated using a right-tailed hypergeometric test. DE 1 condition refers to differential transcription in either the KO versus WT or WT KRABINER versus KO comparison. Nonspecific refers to a gene rescued by WT KRABINER and one or both mutDBD and mutKRAB variants. Unchanged refers to genes or TREs with adjusted *P* > 0.05 (Wald test).

that KRAB-transposase fusions yield modular proteins functioning as sequence-specific transcriptional repressors.

### KRABINER regulates transcription in bat cells

To further test whether transposase capture gives birth to transcriptional regulators, we investigated the ability of KRABINER to modulate gene expression in embryonic fibroblasts of the bat *Myotis velifer*, where the gene is endogenously expressed (fig. S2). We used the CRISPR-Cas9 system to engineer a *KRABINER* knockout (KO) cell line with a pair of guide RNAs (gRNAs) designed to precisely delete the *mariner* transposon from the *ZNF112* locus, leaving the rest of the gene intact (Fig. 4A and fig. S6). We then used a *piggyBac* vector to deliver transgenes at ectopic chromosomal sites into the KO cell line to establish independent clonal lines reintroducing the WT *KRABINER* ($n$ = 4 cell lines), the predicted DNA binding mutant (mutDBD; $n$ = 3), or the predicted KRAB mutant (mutKRAB; $n$ = 3) (fig. S6). Each transgene was cloned under the control of a tetracycline-inducible promoter and contained a C-terminal *myc* tag to monitor protein expression (Fig. 4A and fig. S7). The noninduced condition showed leaky expression more closely recapitulating the level of WT *KRABINER* transcription (hereafter referred to as rescue), whereas transgene induction resulted in *KRABINER* overexpression (OE) relative to the parental cell line (fig. S8A).

To investigate whether KRABINER modulates transcription, we profiled *KRABINER* KO and WT cells with precision run-on followed by sequencing (PRO-seq), which provides a sensitive measurement of nascent transcription throughout the genome, including genes bodies and transcribed regulatory elements (TREs) such as promoters and enhancers ([38], [39]). By quantifying changes in gene body transcription, we identified 2644 genes differentially transcribed between WT and KO cells—1295 were up-regulated in KO (UP), 1349 were down-regulated (DOWN) (DESeq2; Wald test; adjusted $P$ < 0.05) ([40])—which suggests that KRABINER is capable of regulating genic transcription. To identify transcriptional changes which require both the DNA binding domain and KRAB activity of KRABINER, we also assessed transcriptional changes in the transgenic rescue lines. Of the 2644 altered genes, 121 genes (43 UP and 78 DOWN) had their transcription level consistently restored in WT transgenic lines but not in any of the mutant transgenic lines (Fig. 4B and table S4; overlap $P$ < 0.001; right-tailed hypergeometric test). A similar pattern was observed for TREs (identified using dREG) ([39]), with 3472 differentially transcribed TREs after loss of KRABINER, of which 99 were restored exclusively in the WT lines (33 UP and 66 DOWN; Fig. 4C and table S5) (overlap $P$ < 0.001 for down-regulated TREs; right-tailed hypergeometric test). A subset of these TREs are associated with restored gene body transcription (18% UP and 12% DOWN), whereas others are distal (>100 kb) to genes (18% UP and 33% DOWN) or associated with genes bodies that are not differentially transcribed (64% UP and 55% DOWN) (table S4). Although our reporter assays indicate that KRABINER can act as a strong repressor, our loss-of-function analyses in bat cells suggest that the protein exerts a range of transcriptional modulation on the bat genome.

In addition to transcriptional changes specific to the WT transgenic lines, there were several genes and TREs that were rescued by the WT transgene and either the mutDBD (100 genes and 79 TREs) or mutKRAB transgenes (73 genes and 65 TREs; fig. S9). Thus, although at some loci KRABINER's regulatory activity appears to require both its DNA binding and KRAB domains, its transcriptional effects on other loci only requires one of these domains. These mechanisms may also explain the ability of KRABINER to either activate or repress transcription in a locus-dependent fashion. Taken together, these data suggest that KRABINER contributes to the transcriptional regulation of a subset of genes and cis-regulatory elements in the examined embryonic cell line.

To investigate whether KRABINER regulates a discrete set of genes or a network of related genes, we performed gene ontology enrichment analysis (two-sided hypergeometric test, Bonferroni step-down correction) for all genes rescued by the WT transgene ($n$ = 121) as well

Corrected 13 April 2021. See full text.

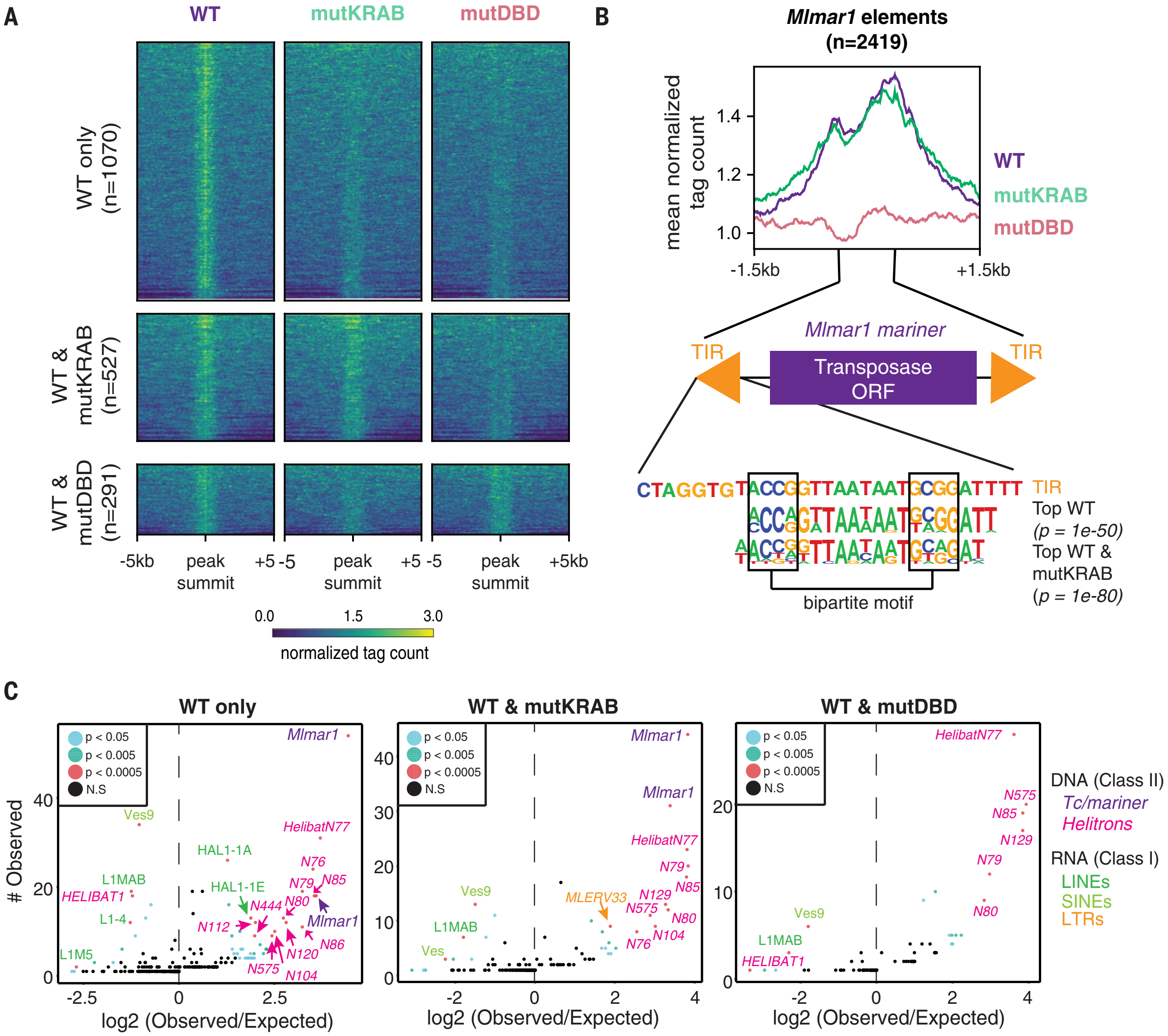


**Fig. 5. KRABINER binds to *mariner* TIRs in bat cells.** (**A**) Heatmaps summarizing merged, library-size, and input-normalized ChIP-seq coverage of each KRABINER variant centered on the summit of WT (top), WT-mutKRAB (middle), and WT-mutDBD (bottom) peak sets. (**B**) Metaplot summarizing normalized ChIP-seq coverage of each KRABINER variant over all genomic *Mlmar1* elements (top). The top enriched motif in the WT and the WT and mutKRAB peak sets is identical to the predicted bipartite binding motif within the *Mlmar1 mariner* TIR (bottom) (HOMER). (**C**) Enrichment of transposon families in WT only (left), WT-mutKRAB (center), and WT-mutDBD (right) peak sets. Observed refers to the number of overlaps between a TE family and a given peak set. Expected refers to the number of expected overlaps between a TE family and a given peak set after shuffling TE locations 1000 times. *P* values were determined using the binomial distribution with *n* = 1000 shuffles. WT is colored purple, mutDBD is pink, and mutKRAB is green. DBD, DNA binding domain; ORF, open reading frame; N.S., not significant; LINEs, long interspersed nuclear elements; SINEs, short interspersed nuclear elements; LTRs, long terminal repeats.

as additional target genes whose promoters were found to be differentially transcribed in the TRE analysis ($n$ = 57) (table S4; ClueGO) (*19*, *41*). This analysis revealed a significant enrichment for genes involved in negative regulation of cell migration [gene ontology (GO) ID: 0030336; adjusted $P = 2.1 \times 10^{-6}$] and in gastrulation (GO: 0007369; adjusted $P = 2.5 \times 10^{-5}$) as the most enriched terms (fig. S10A and table S5). Additional significant terms were linked to morphogenetic or developmental pathways, including positive regulation of the Wnt signaling pathway (GO: 2000096; adjusted $P = 5.2 \times 10^{-3}$), artery morphogenesis (GO: 0048844; adjusted $P = 2.5 \times 10^{-2}$), heart valve morphogenesis (GO: 0003179; adjusted $P = 8.8 \times 10^{-3}$), and neural crest differentiation (GO: 0014033; adjusted $P = 2.2 \times 10^{-2}$) (fig. S10A and table S5). A similar result was obtained for the down-regulated genes alone, which suggests that KRABINER's direct targets may be down-regulated (fig. S10B and table S5). These results suggest that, in the embryonic cell line examined, KRABINER regulates a set of genes enriched for developmental functions, as may be expected for a canonical TF.

### KRABINER binds to many genomic sites, with a preference for *mariner* TIRs

We next used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to determine whether and where KRABINER binds throughout the bat genome. We profiled the binding of myc-tagged KRABINER WT protein in the KO cell line background as well as that of mutKRAB and mutDBD mutant proteins 24 hours after induction of transgene expression ($n$ = 3 each) (fig. S11). With these samples, we called binding peaks using MACS2 (*42*) for each genotype relative to input and filtered

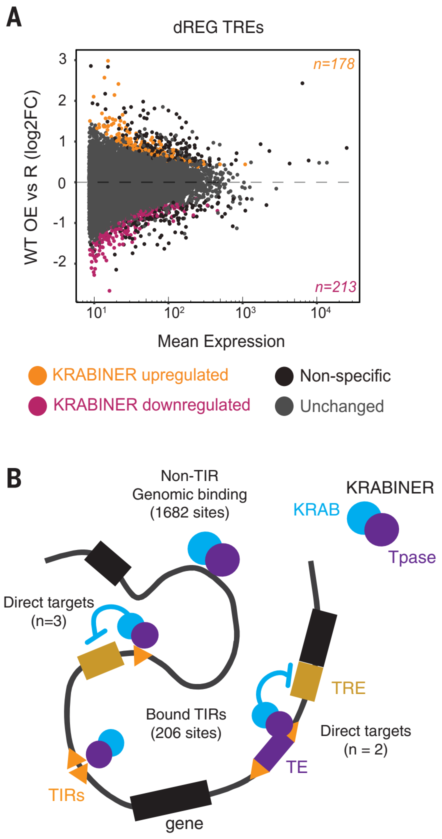


**Fig. 6. KRABINER regulates a network of genes and TREs in bat cells.** (**A**) MA plot summarizing changes in TRE transcription upon OE of WT KRABINER. Nonspecific (black) refers to changes in TRE transcription that are shared between OE of WT KRABINER and one or both mutant KRABINER variants. Unchanged (gray) refers to TREs with adjusted $P > 0.05$ (Wald test). (**B**) Proposed model for KRABINER's function as a TF in bats. KRABINER directly binds to *mariner* TIRs within the genome and leads to direct down-regulation of a subset of TREs. KRABINER also binds to other genomic regions and indirectly regulates a number of genes and TREs. R, rescue; Tpase, transposase.

out the peaks identified in all three genotypes (>50% reciprocal overlap), which are likely to represent spurious or nonspecific interactions (19). This analysis identified 1888 WT, 5702 mutKRAB, and 4264 mutDBD peaks, respectively. The higher number of peaks obtained with mutKRAB and mutDBD likely reflects the higher expression of those transgenes relative to the WT transgene (fig. S8). To identify genomic sites likely bound directly via KRABINER, we focused on the WT peaks and used the mutKRAB and mutDBD peaks as additional filters or background sets.

Of the 1888 WT KRABINER binding peaks, 56% ($n$ = 1070) were specific to the WT condition, which suggests that most of KRABINER's binding requires both functional domains (Fig. 5A). However, there were about twice as many peaks shared between the WT and mutKRAB conditions ($n$ = 572; 28%) than there were between the WT and mutDBD conditions ($n$ = 291; 15%) (>50% reciprocal overlap; Fig. 5A), which suggests that most of KRABINER's binding is dependent on its DBD. The set of peaks overlapping exclusively between the WT and mutDBD conditions likely represent indirect genomic interactions, possibly mediated through protein-protein interactions via the KRAB domain (43) or other region of the protein.

We then looked for sequence motifs that might explain KRABINER binding to its genomic sites (HOMER) (44). We extracted a 200–base pair (bp) window centered on the peak summit for all WT peaks and examined which de novo motifs were enriched relative to mutDBD peaks for the WT and WT-mutKRAB peak sets or to the mutKRAB peaks for the WT-mutDBD peak set. For the WT-only and WT-mutKRAB peaks, the top enriched motif (binomial test; $P = 1 \times 10^{-50}$ and $P = 1 \times 10^{-80}$) was identical and resembled a bipartite region within the *Mlmar1 mariner* TIR sequence that aligns with the binding sites mapped previously for the *Mos1* transposase (45) (Fig. 5B and fig. S12). Of the additional 12 and 17 enriched motifs predicted for the WT and WT-mutKRAB peak sets, respectively, all were derived from *Mlmar1* elements (fig. S12). We identified only one enriched motif in the WT-mutDBD peaks that bore no resemblance to *Mlmar1* TIRs or any known metazoan TF [data 6 (26)]. These data suggested that the WT and mutKRAB, but not mutDBD, proteins bind many *Mlmar1* elements dispersed throughout the genome. Consistent with this, *Mlmar1* elements are enriched in WT [log2 fold enriched (FE) = 4.43; $P = 2.2 \times 10^{-16}$] and WT-mutKRAB peaks (log2FE = 4.95; $P = 2.2 \times 10^{-16}$), but not the WT-mutDBD peaks [$P$ = 0.63; two-sided binomial test, 1000 bootstraps (19)] (Fig. 5C). In total, the WT KRABINER transgenic protein binds to 206 (8.5%) of all *Mlmar1* elements annotated in the bat genome assembly.

To determine where in the *Mlmar1* element KRABINER binds, we plotted the input-normalized ChIP-seq reads for each genotype over all *Mlmar1* transposons in the genome. We found that a fraction of these transposons was bound by the WT and mutKRAB proteins, but not the mutDBD protein, consistent with the peak-based approach (Fig. 5B and fig. S12). The ChIP-seq read coverage peaks within the TIR regions, and especially the 3′ TIR (Fig. 5B and fig. S12), which is consistent with the binding activity of other *mariner* transposases (32, 33, 46). Collectively, our ChIP-seq data demonstrate that KRABINER is capable of binding numerous genomic sites, with a preference for *Mlmar1* TIRs. Further, its ability to bind *Mlmar1* TIRs is dependent on its transposase DNA binding domain.

## KRABINER binding is associated with down-regulation of nearby TREs

To test whether KRABINER binding leads to transcriptional change, we next induced expression of the *KRABINER* transgenes and performed PRO-seq 24 hours after induction (OE)—conditions matching our ChIP-seq experiments. We then identified TREs differentially transcribed between the OE versus rescue conditions, which are of the same genotype and in principle differ only in the level of KRABINER expression (figs. S7 and S8). Because TREs represent discrete transcriptional units such as promoters and enhancers, we reasoned that KRABINER binding to or near these regions would more likely affect transcription than binding within a gene body. We identified 391 TREs (178 UP and 213 DOWN; Fig. 6A) that were differentially transcribed upon OE of the WT KRABINER protein but neither of the mutant proteins (mutDBD or mutKRAB). Additionally, several TREs were differentially expressed in the same direction upon OE of the WT protein and either mutDBD or mutKRAB (fig. S13), consistent with the hypothesis that a subset of KRABINER's transcriptional changes require only one of its functional domains.

To determine whether KRABINER binding is associated with differential TRE transcription, we first examined whether WT KRABINER ChIP-seq peaks were located near (<1 kb) differentially expressed TREs. Although the total number of KRABINER peaks located nearby differentially expressed TREs was small ($n$ = 6), it was a significant enrichment over the random expectation [permutation test; log2FE = 2.58; empirical $P$ = 0; 10,000 bootstraps (19)] (fig. S14). Notably, all were down-regulated TREs, consistent with the results of our reporter assays, which suggests that tethering KRABINER to DNA induces local transcription repression. Furthermore, five of six differentially expressed TREs were located within ~1 kb of at least one *Mlmar1* element, and the TIRs of these transposons were located near the summit of the KRABINER peaks. Finally, several of the TREs connected with KRABINER binding to nearby *Mlmar1* elements were also associated with changes in adjacent gene expression.

For example, a differentially expressed TRE is located in the promoter region of the bat ortholog of the DNA damage-recognition and repair-factor gene *XPA*, which is down-regulated (Wald test; log2FC = –0.69; adjusted $P$ = 0.0079) upon WT KRABINER OE and is located immediately adjacent to a *mariner* TIR bound by KRABINER (fig. S15). A similar pattern is seen for an intergenic TRE, located between the bat homolog of the family with

Corrected 13 April 2021. See full text.

sequence similarity 174 member B (*FAM174B*) and chromodomain helicase DNA binding protein (*CHD2*) genes (fig. S16). This region contains three distinct TREs, two of which are down-regulated upon WT KRABINER OE (Wald test; log2FC = –1.1; adjusted *P* = 0.04; and log2FC = –1.09; adjusted *P* = 0.006, respectively), and this change is associated with KRABINER binding to the *Mlmar1* TIRs immediately upstream of these TREs (fig. S16). Other regulated TREs include one located in the promoter region of the bat homolog of the small nuclear ribonucleoprotein polypeptide A′ (*SNRPA1*) gene (Wald test; log2FC = –0.84; adjusted *P* = 0.02), which is located downstream of four bound *Mlmar1* elements (fig. S17), and a distal TRE (Wald test; log2FC = –1.24; adjusted *P* = 0.01) upstream of the *nucleoporin 50* (*NUP50*) gene (fig. S18). Notably, each of the KRABINER–down-regulated TREs are near two or more bound TIR sequences, which suggests that KRABINER's effect on TREs may be strengthened by additional binding sites.

Collectively, our PRO-seq and ChIP-seq data demonstrate that KRABINER acts as a canonical TF and that some of its transcriptional regulatory activity in bat cells is accomplished by KRABINER binding to its cognate *mariner* TIRs. However, only a minority of genes regulated by KRABINER appear to be direct targets, which suggests that KRABINER is integrated within a complex transcriptional network (Fig. 6B).

## Discussion

Although gene birth through duplication has been extensively documented, how novel protein architectures and biological functions are born has remained poorly characterized. Here, we validate that exon shuffling is a major evolutionary force generating genetic novelty (*9*), and we provide evidence that DNA transposons fuel the process not only by supplying protein domains to assemble new protein architectures, but also, in many cases, by introducing the splice sites that enable the fusion process. Although these events must be relatively rare on an evolutionary time scale, the mobility of DNA transposons likely increases the probability of generating a functional gene via exon shuffling by introducing genetic material into new contexts. We also derived first principles of how transposase-mediated exon shuffling occurs, providing a foundation for the identification of HTF genes in other lineages.

Transposase-mediated exon shuffling offers a plausible mechanism for the birth of known developmental regulatory proteins, such as Pax6, which controls eye development and patterning across animals (*47*). Another example is *POGZ*, a gene expressed predominantly in the brain and associated with autism and intellectual disability when mutated in humans (*48*, *49*). Although these are examples of HTFs with relatively deep evolutionary origins and likely serving broadly conserved functions, our functional analysis of *KRABINER*—a bat-specific protein with transcriptional modulatory activities—suggests that the process of gene birth via transposase capture has been a continuous source of regulatory innovation.

Many studies have implicated transposable elements in the dispersion of TF binding sites and cis-regulatory elements that have rewired gene regulatory networks during evolution (*50*–*55*). However, these studies do not explain how a new regulatory network, including its associated regulatory proteins, initially evolves. The data presented here offer a plausible path by which a new trans-regulatory protein and its cis-binding sites in the genome simultaneously emerge from the same transposon family (*56*). Historically this model has been difficult to test because previously recognized transposase-derived TFs (such as *Pax*) and their network of regulated genes evolved hundreds of millions of years ago, which would have obscured the transposon origin of their genomic binding sites. Our study of KRABINER, a recently evolved HTF, together with studies of two other young HTF genes evolved during primate evolution—*SETMAR* [40 to 58 Ma old (*15*, *57*)] and *PGBD3-CSB* [>40 Ma old (*16*, *58*, *59*)]—have captured the early steps by which a new cis-regulatory circuit, including both coding and noncoding components, can emerge from a transposon family.

Although we focused on the tetrapod lineage in this study, we propose that the principles and implications of transposase capture revealed herein extend beyond vertebrates. Transposases are ancient and possibly the most abundant and ubiquitous genes in nature (*60*), and a variety of host domains that regulate transcription exist in all branches of the tree of life. It is easy to envision how these sequences have provided the raw material for the assemblage of endless combinations of transposase-host fusion proteins throughout evolution.

## Materials and methods

### Identifying and characterizing transposase fusion genes

To identify HTF genes, we used a domain-centric approach [CDART (*18*)] to identify NCBI Refseq (*61*) tetrapod gene annotations (table S1) that contained a transposase domain (table S2), had two or more exons (to exclude standalone transposases), and had RNA-seq support for all annotated introns or were identified in de novo RNA-seq data [data 2 and data 3 (*26*)]. We further characterized each HTF by its domain structure, originating gene and transposon, and the timing of gene birth (*19*). We identified orthologs of each HTF gene, where present, using a combination of homology (BLASTn) and synteny (*19*). We further assigned each HTF an age in units of million years on the basis of the estimated evolutionary divergence [Timetree (*62*)] between all species that have the HTF. Finally, we tested the transposase domains of all HTFs conserved in two or more species that diverged >50 Ma ago for evidence of purifying selection [phylogenetic analysis by maximum likelihood (PAML) (*63*)], with significance determined by comparing the estimated model with a model assuming neutral evolution (LRT; *P* < 0.05; chi-square distribution) (*19*).

### Determining mechanism for HTF gene birth

We first stratified HTF genes into two classes: genes born via splicing or via gene duplication. Of those born via splicing, we considered gene models containing both the original host transcript and the fusion transcript to have originated via alternative splicing. For a subset of HTFs for which we were able to reconstruct the consensus sequence of the originating transposon (*KTIGD1*, *KMARD1*, *KTIGD3*, *KRABINER*, *KMARD4*, *KHATD2*, *TIGD-G1*, *TIGD-G4-Zscan29*, and *MARD-G1*), we also inferred whether the splice site was present in the transposon and, if so, whether it was also present in the consensus sequence (*19*).

### Tracing the birth and evolution of KRABINER

We first determined the timing of the *Mlmar1 mariner* insertion into the *ZNF112* locus using both homology-based approaches (BLAST) against publicly available bat genomes and PCR amplification followed by sequencing of the *mariner* insertion in an additional seven bat species (*19*). We then performed RT-PCR on cDNA from cell lines derived from three bat species (*M. velifer*, *Myotis lucifugus*, and *Eptesicus fuscus*) using primers designed to either amplify the key splice junction (exons 4 to 5 and 5 to 6) or to amplify the full length (exons 1 and 6) KRABINER (fig. S2 and table S6).

### Luciferase assays

To assess the ability of four KRAB-transposase fusion (KTF) genes (*KRABINER*, *KMARD1*, *KTIGD1*, and *KTIGD3*), we performed luciferase reporter assays. For KRABINER, we also included DNA binding domain or KRAB mutant variants, which were generated on the basis of preexisting studies of a closely related transposon (*32*, *33*) or studies that identified residues critical for KRAB domain function (*34*–*37*), respectively. For each HTF, we transfected either WT or KAP1 KO HEK293T cells (*31*) with three plasmids: a KRAB-transposase fusion expression vector or empty vector, a firefly luciferase expression vector with the cognate consensus or scrambled TIR sequence located immediately upstream of the promoter, and a renilla luciferase expression vector as an internal control. We lysed cells 48 hours after transfection and split the lysate into five wells of a 96-well plate ($n$ = 5 technical replicates). We then measured firefly and renilla luminescence by means of a plate reader and normalized each value as follows to obtain a final relative luminescence per replicate: [(KTF-lum$_{\text{firefly}}$ / KTF-lum$_{\text{renilla}}$) / mean(empty-lum$_{\text{firefly}}$ / empty-lum$_{\text{renilla}}$)] (19). We repeated each experiment a minimum of three times and considered differences in mean relative luminescence significant at adjusted $P$ < 0.05 (pairwise Wilcoxon test with Bonferroni multiple testing correction).

### Generating and validating KRABINER KO and rescue cell lines

We generated a *KRABINER* KO cell line using the CRISPR-Cas9 system as previously described (64). We used a pair of gRNAs [gRNA 1 (GL430169:87458-87477) - CATTTAGTTTCAGCCTCTCATGG; gRNA 2 (GL430169:89264-89286) - TAATACGTAAGCTGCTGTGTGGG] that flank the *Mlmar1 mariner* insertion at the *ZNF112* locus to precisely delete the *mariner* element (fig. S6A), leaving the parental gene intact (19). After clonal expansion, we identified a single cell line with a homozygous deletion that we verified through Sanger sequencing. We further verified absence of *KRABINER* transcription by RT-PCR (fig. S6B and table S6).

To rescue KRABINER expression, we used the *piggyBac* system to introduce transgenes encoding inducible forms of either WT, DNA binding mutant (mutDBD), or KRAB mutant (mutKRAB) variants of KRABINER into our KO cell line (19). Transduced cells were selected through puromycin treatment (1.5 μg/ml) for 1 week, and then clonally expanded ($n$ = 3 minimum). We verified insertion of the transgene using PCR followed by sequencing (fig. S7B and table S6) and expression of the transgene at the RNA level through RT-PCR (fig. S7C and table S6) and PRO-seq (fig. S8A) and at the protein level through Western-blot and immunofluorescence assays (fig. S7) (19).

### PRO-seq analysis

We performed PRO-seq on KRABINER WT ($n$ = 2), KO ($n$ = 2), and rescue cell lines (minimum $n$ = 3). Rescue cell lines were treated with 1 μg/ml doxycycline (induced, OE) or not (noninduced, rescue) 24 hours before PRO-seq. PRO-seq libraries (20 million cells per genotype per treatment, >90% viability) were prepared as previously described (19, 38, 65, 66). Libraries were sequenced on an Illumina NextSeq 500 platform with 37 bp by 37 bp chemistry (table S7), and raw reads for all samples are accessible at SRP256595. We processed the resulting PRO-seq data using a pipeline available at Zenodo (19, 67). To quantify gene body transcription, we defined gene body regions as the transcription start site (TSS) plus 500 bp to the transcription end site (TES) and the TSS minus 500 bp to the TES for + and – strand Myoluc2 ENSEMBL gene annotations, respectively. We called TREs for each sample using dREG (39) and merged [bedtools merge (68)] to generate a comprehensive TRE set. We quantified read counts in both gene bodies and TREs at single nucleotide resolution using a custom script and bedtools map (68). TRE annotations and bigWig coverage files for each sample are available at GSE148789.

We performed differential transcription analysis for the gene bodies and TREs separately using DESeq2 (40). We performed three comparisons: KRABINER KO versus WT; WT, mutDBD, or mutKRAB rescue versus KO; and WT, mutDBD, or mutKRAB OE versus rescue. We considered a gene or TRE to be regulated by KRABINER if it exhibited significant (adjusted $P$ < 0.05; Wald test) reciprocal changes in the KO versus WT and the WT rescue versus KO comparison or if it was differentially transcribed in the OE versus rescue comparison (adjusted $P$ < 0.05; Wald test). Raw expression counts for genes and TREs as well as DESeq2 outputs for all comparisons are available at GSE148789. We also performed gene ontology enrichment for KRABINER-regulated genes (table S4) (19).

### ChIP-seq analysis

We performed ChIP-seq on KRABINER rescue cell lines (minimum $n$ = 3; 20 million cells each; >90% viability) 24 hours after treatment with 1 μg/ml doxycycline to induce transgene expression (19). We prepared libraries for each immunoprecipitation (IP) and input (3.33% total sonicated chromatin for each sample, pooled across genotype) using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) according to manufacturer instructions with four cycles of PCR amplification. Libraries were sequenced on the HiSeq 4000 platform in PE150 mode (Novogene Corporation Inc., Sacramento, CA). Reads for all samples are available at SRP256596.

We then quality processed reads, mapped them to the Myoluc2.0 assembly, and called KRABINER binding peaks for each variant (WT, mutDBD, and mutKRAB) relative to matched input samples using MACS2 (19, 42). We removed peaks called in all three genotypes, which are likely to be spurious, and further subset the WT peaks into three categories: those specific to the WT transgene and those shared between either the WT and mutDBD or between the WT and mutKRAB transgenes (19). To annotate the peaks, we identified sequence motifs enriched in each of the WT peak categories relative to a background set (all mutDBD peaks for WT only and shared WT-mutKRAB or all mutKRAB peaks for the shared WT-mutDBD peaks) [HOMER, data 4 to data 6 (26)] (19, 44). We then determined which transposable elements (TEs) were enriched for KRABINER binding, as previously described (69), and considered a TE family to be enriched if it overlapped more than expected by chance (binomial $P$ < 0.05; 1000 shuffles). We further determined whether KRABINER binding peaks were enriched in or near differentially transcribed genes or TREs using a pseudorandom shuffling method to calculate empirical $P$ values (significant at $P$ < 0.05; 10,000 shuffles) (19). All raw and normalized bigWig files [DeepTools, (19, 70)] and peak files are available at GSE148789.

## REFERENCES AND NOTES

1. S. Ohno, *Evolution by Gene Duplication* (Springer, 1970).
2. F. H. Ruddle *et al*., Evolution of *Hox* genes. *Annu. Rev. Genet.* **28**, 423–442 (1994). doi: 10.1146/annurev.ge.28.120194.002231; pmid: 7893134
3. M. Bouchard, A. Schleiffer, F. Eisenhaber, M. Busslinger, *Pax Genes: Evolution and Function* (Wiley, 2008).
4. C. Deng, C. H. C. Cheng, H. Ye, X. He, L. Chen, Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21593–21598 (2010). doi: 10.1073/pnas.1007883107; pmid: 21115821
5. V. J. Lynch, Inventing an arsenal: Adaptive evolution and neofunctionalization of snake venom phospholipase $A_2$ genes. *BMC Evol. Biol.* **7**, 2 (2007). doi: 10.1186/1471-2148-7-2; pmid: 17233905
6. H. Innan, Population genetic models of duplicated genes. *Genetica* **137**, 19–37 (2009). doi: 10.1007/s10709-009-9355-1; pmid: 19266289
7. S. B. Van Oss, A.-R. Carvunis, De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019). doi: 10.1371/journal.pgen.1008160; pmid: 31120894
8. J. Luis Villanueva-Cañas *et al*., New Genes and Functional Innovation in Mammals. *Genome Biol. Evol.* **9**, 1886–1900 (2017). doi: 10.1093/gbe/evx136; pmid: 28854603
9. W. Gilbert, Why genes in pieces? *Nature* **271**, 501 (1978). doi: 10.1038/271501a0; pmid: 622185
10. M. Long, C. Rosenberg, W. Gilbert, Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 12495–12499 (1995). doi: 10.1073/pnas.92.26.12495; pmid: 8618928
11. L. Patthy, Modular assembly of genes and the evolution of new functions. *Genetica* **118**, 217–231 (2003). doi: 10.1023/A:1024182432483; pmid: 12868611
12. M. Long, N. W. VanKuren, S. Chen, M. D. Vibranovski, New gene evolution: Little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013). doi: 10.1146/annurev-genet-111212-133301; pmid: 24050177
13. C. Feschotte, E. J. Pritham, DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007). doi: 10.1146/annurev.genet.40.110405.090448; pmid: 18076328
14. H. J. Tipney *et al*., Isolation and characterisation of *GTF2IRD2*, a novel fusion gene and member of the TFII-I family of transcription factors, deleted in Williams–Beuren syndrome. *Eur. J. Hum. Genet.* **12**, 551–560 (2004). doi: 10.1038/sj.ejhg.5201174; pmid: 15100712
15. R. Cordaux, S. Udit, M. A. Batzer, C. Feschotte, Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8101–8106 (2006). doi: 10.1073/pnas.0601161103; pmid: 16672366
16. J. C. Newman, A. D. Bailey, H.-Y. Fan, T. Pavelitz, A. M. Weiner, An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLOS Genet.* **4**, e1000031 (2008). doi: 10.1371/journal.pgen.1000031; pmid: 18369450
17. R. Breitling, J.-K. Gerber, Origin of the paired domain. *Dev. Genes Evol.* **210**, 644–650 (2000). doi: 10.1007/s004270000106; pmid: 11151303
18. L. Y. Geer, M. Domrachev, D. J. Lipman, S. H. Bryant, CDART: Protein homology by domain architecture. *Genome Res.* **12**, 1619–1623 (2002). doi: 10.1101/gr.278202; pmid: 12368255
19. See supplementary materials.

20. J. Alföldi *et al*., The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011). doi: 10.1038/nature10390; pmid: 21881562
21. T. A. Castoe *et al*., The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20645–20650 (2013). doi: 10.1073/pnas.1314475110; pmid: 24297902
22. T. Mitros *et al*., A chromosome-scale genome assembly and dense genetic map for *Xenopus tropicalis*. *Dev. Biol.* **452**, 8–20 (2019). doi: 10.1016/j.ydbio.2019.03.015; pmid: 30980799
23. D. A. Ray *et al*., Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res.* **18**, 717–728 (2008). doi: 10.1101/gr.071886.107; pmid: 18340040
24. E. J. Pritham, C. Feschotte, Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1895–1900 (2007). doi: 10.1073/pnas.0609601104; pmid: 17261799
25. C. G. Sotero-Caio, R. N. Platt 2nd, A. Suh, D. A. Ray, Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.* **9**, 161–177 (2017). doi: 10.1093/gbe/evw264; pmid: 28158585
26. R. Cosby, J. Judd, R. Zhang, A. Zhong, N. Garry, E. Pritham, C. Feschotte, Cosby_et_al_2020_Supplemental_Data [Data Set], Zenodo (2020); http://doi.org/10.5281/zenodo.4060329.
27. M. Bruno, M. Mahgoub, T. S. Macfarlan, The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annu. Rev. Genet.* **53**, 393–416 (2019). doi: 10.1146/annurev-genet-112618-043717; pmid: 31518518
28. H.-M. Herz, A. Garruss, A. Shilatifard, SET for life: Biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem. Sci.* **38**, 621–639 (2013). doi: 10.1016/j.tibs.2013.09.004; pmid: 24148750
29. L. C. Edelstein, T. Collins, The SCAN domain family of zinc finger transcription factors. *Gene* **359**, 1–17 (2005). doi: 10.1016/j.gene.2005.06.022; pmid: 16139965
30. M. Imbeault, P.-Y. Helleboid, D. Trono, KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017). doi: 10.1038/nature21683; pmid: 28273063
31. C. H. Tie *et al*., KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. *EMBO Rep.* **19**, 45000 (2018). doi: 10.15252/embr.201745000; pmid: 30061100
32. L. Zhang, A. Dawson, D. J. Finnegan, DNA-binding activity and subunit interaction of the mariner transposase. *Nucleic Acids Res.* **29**, 3566–3575 (2001). doi: 10.1093/nar/29.17.3566; pmid: 11522826
33. J. M. Richardson, S. D. Colloms, D. J. Finnegan, M. D. Walkinshaw, Molecular architecture of the Mos1 paired-end complex: The structural basis of DNA transposition in a eukaryote. *Cell* **138**, 1096–1108 (2009). doi: 10.1016/j.cell.2009.07.012; pmid: 19766564
34. J. F. Margolin *et al*., Krüppel-associated boxes are potent transcriptional repression domains. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4509–4513 (1994). doi: 10.1073/pnas.91.10.4509; pmid: 8183939
35. R. Witzgall, E. O'Leary, A. Leaf, D. Onaldi, J. V. Bonventre, The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4514–4518 (1994). doi: 10.1073/pnas.91.10.4514; pmid: 8183940
36. J. R. Friedman *et al*., KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes Dev.* **10**, 2067–2078 (1996). doi: 10.1101/gad.10.16.2067; pmid: 8769649
37. K. E. Murphy *et al*., The Transcriptional Repressive Activity of KRAB Zinc Finger Proteins Does Not Correlate with Their Ability to Recruit TRIM28. *PLOS ONE* **11**, e0163555 (2016). doi: 10.1371/journal.pone.0163555; pmid: 27658112
38. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013). doi: 10.1126/science.1229386; pmid: 23430654
39. Z. Wang, T. Chu, L. A. Choate, C. G. Danko, Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* **29**, 293–303 (2019). doi: 10.1101/gr.238279.118; pmid: 30573452
40. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi: 10.1186/s13059-014-0550-8; pmid: 25516281
41. G. Bindea *et al*., ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009). doi: 10.1093/bioinformatics/btp101; pmid: 19237447
42. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
43. P.-Y. Helleboid *et al*., The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J.* **38**, e101220 (2019). doi: 10.15252/embj.2018101220; pmid: 31403225
44. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: 10.1016/j.molcel.2010.05.004; pmid: 20513432
45. J. M. Richardson *et al*., Mechanism of Mos1 transposition: Insights from structural analysis. *EMBO J.* **25**, 1324–1334 (2006). doi: 10.1038/sj.emboj.7601018; pmid: 16511570
46. G. Yang, D. H. Nagel, C. Feschotte, C. N. Hancock, S. R. Wessler, Tuned for transposition: Molecular determinants underlying the hyperactivity of a *Stowaway* MITE. *Science* **325**, 1391–1394 (2009). doi: 10.1126/science.1175688; pmid: 19745152
47. Z. Kozmik, Pax genes in eye development and evolution. *Curr. Opin. Genet. Dev.* **15**, 430–438 (2005). doi: 10.1016/j.gde.2005.05.001; pmid: 15950457
48. H. A. F. Stessman *et al*., Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. *Am. J. Hum. Genet.* **98**, 541–552 (2016). doi: 10.1016/j.ajhg.2016.02.004; pmid: 26942287
49. J. White *et al*., POGZ truncating alleles cause syndromic intellectual disability. *Genome Med.* **8**, 3 (2016). doi: 10.1186/s13073-015-0253-0; pmid: 26739615
50. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017). doi: 10.1038/nrg.2016.139; pmid: 27867194
51. D. R. Fuentes, T. Swigut, J. Wysocka, Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018). doi: 10.7554/eLife.35989; pmid: 30070637
52. J. Pontis *et al*., Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019). doi: 10.1016/j.stem.2019.03.012; pmid: 31006620
53. V. Sundaram, J. Wysocka, Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Phil. Trans. R. Soc. B* **375**, 20190347 (2020). doi: 10.1098/rstb.2019.0347; pmid: 32075564
54. V. J. Lynch *et al*., Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015). doi: 10.1016/j.celrep.2014.12.052; pmid: 25640180
55. M. Trizzino, A. Kapusta, C. D. Brown, Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* **19**, 468 (2018). doi: 10.1186/s12864-018-4850-3; pmid: 29914366
56. C. Feschotte, Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008). doi: 10.1038/nrg2337; pmid: 18368054
57. M. Tellier, R. Chalmers, Human SETMAR is a DNA sequence-specific histone-methylase with a broad effect on the transcriptome. *Nucleic Acids Res.* **47**, 122–133 (2019). doi: 10.1093/nar/gky937; pmid: 30329085
58. A. D. Bailey *et al*., The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. *DNA Repair* **11**, 488–501 (2012). doi: 10.1016/j.dnarep.2012.02.004; pmid: 22483866
59. L. T. Gray, K. K. Fong, T. Pavelitz, A. M. Weiner, Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLOS Genet.* **8**, e1002972 (2012). doi: 10.1371/journal.pgen.1002972; pmid: 23028371
60. R. K. Aziz, M. Breitbart, R. A. Edwards, Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010). doi: 10.1093/nar/gkq140; pmid: 20215432
61. N. A. O'Leary *et al*., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016). doi: 10.1093/nar/gkv1189; pmid: 26553804
62. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017). doi: 10.1093/molbev/msx116; pmid: 28387841
63. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi: 10.1093/molbev/msm088; pmid: 17483113
64. F. A. Ran *et al*., Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013). doi: 10.1038/nprot.2013.143; pmid: 24157548
65. D. B. Mahat *et al*., Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016). doi: 10.1038/nprot.2016.086; pmid: 27442863
66. J. Judd *et al*., A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). bioRxiv 2020.05.18.102277 [Preprint]. 19 May 2020. https://doi.org/10.1101/2020.05.18.102277.
67. JAJ256, JAJ256/PROseq_alignment.sh: PRO-seq alignment pipeline used for Recurrent evolution of vertebrate transcription factors by transposase capture, version 1.0, Zenodo (2020); http://doi.org/10.5281/zenodo.4019173.
68. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
69. A. Kapusta *et al*., Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLOS Genet.* **9**, e1003470 (2013). doi: 10.1371/journal.pgen.1003470; pmid: 23637635
70. F. Ramírez *et al*., deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975

## ACKNOWLEDGMENTS

We acknowledge D. Ray, W. Wright, N. Elde, J. Lis, H. Rowe, J. Wysocka, and T. Macfarlan for providing reagents. We also thank members of the Feschotte laboratory for helpful discussion. **Funding:** This work was supported by R35GM122550 from the National Institutes of Health to C.F. J.J. was supported by NHGRI fellowship F31HG010820. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **Author contributions:** R.L.C., E.J.P., and C.F. conceived of and designed the study. R.L.C. and R.Z. identified and analyzed HTF genes. R.L.C. and A.Z. performed luciferase reporter assays. R.L.C. generated KRABINER KO cell lines, and R.L.C. and N.G. generated the KRABINER rescue cell lines. J.J. generated and processed PRO-seq and ChIP-seq data. R.L.C. analyzed all data and generated figures. R.L.C. wrote the paper, and C.F. reviewed and edited the paper. All authors reviewed the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All processed sequencing data are deposited in the Gene Expression Omnibus (GSE148789), and raw data are deposited in the Sequence Read Archive (PRO-seq: SRP256595; ChIP-seq: SRP256596). Supplemental sequence and motif data (data 1 to data 6) are available on Zenodo (*26*), and code for processing the PRO-seq data is available on Zenodo (*67*). All cell lines and reagents are available on request.

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6531/eabc6405/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S7
References (*71*–*100*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

8 May 2020; accepted 18 December 2020
10.1126/science.abc6405

Science

# Recurrent evolution of vertebrate transcription factors by transposase capture

Rachel L. Cosby, Julius Judd, Ruiling Zhang, Alan Zhong, Nathaniel Garry, Ellen J. Pritham and Cédric Feschotte

*Science* **371** (6531), eabc6405.
DOI: 10.1126/science.abc6405

**A recipe for new genes**

Most lineages contain evolutionarily novel genes, but their origin is not always clear. Cosby *et al.* investigated the origin of families of lineage-specific vertebrate genes (see the Perspective by Wacholder and Carvunis). Fusion between transposable elements (TEs) and host gene exons, once incorporated into the host genome, could generate new functional genes. Examination of *KARABINER*, a bat gene that arose through this process, shows how the retention of part of the TE within this gene allows the transcribed protein to bind throughout the genome and act as a transcriptional regulator. Thus, TEs interacting within their host genome provide the raw material to generate new combinations of functional domains that can be selected upon and incorporated within the hierarchical cellular network.

*Science*, this issue p. eabc6405; see also p. 779

*Science* (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.