# Widespread contribution of transposable elements to the innovation of gene regulatory networks

Vasavi Sundaram,[1,4] Yong Cheng,[2,4] Zhihai Ma,[2] Daofeng Li,[1] Xiaoyun Xing,[1] Peter Edge,[3] Michael P. Snyder,[2] and Ting Wang[1]

[1]*Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri 63108, USA;* [2]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [3]*Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, Minnesota 55455, USA*

**Transposable elements (TEs) have been shown to contain functional binding sites for certain transcription factors (TFs). However, the extent to which TEs contribute to the evolution of TF binding sites is not well known. We comprehensively mapped binding sites for 26 pairs of orthologous TFs in two pairs of human and mouse cell lines (representing two cell lineages), along with epigenomic profiles, including DNA methylation and six histone modifications. Overall, we found that 20% of binding sites were embedded within TEs. This number varied across different TFs, ranging from 2% to 40%. We further identified 710 TF–TE relationships in which genomic copies of a TE subfamily contributed a significant number of binding peaks for a TF, and we found that LTR elements dominated these relationships in human. Importantly, TE-derived binding peaks were strongly associated with open and active chromatin signatures, including reduced DNA methylation and increased enhancer-associated histone marks. On average, 66% of TE-derived binding events were cell type-specific with a cell type-specific epigenetic landscape. Most of the binding sites contributed by TEs were species-specific, but we also identified binding sites conserved between human and mouse, the functional relevance of which was supported by a signature of purifying selection on DNA sequences of these TEs. Interestingly, several TFs had significantly expanded binding site landscapes only in one species, which were linked to species-specific gene functions, suggesting that TEs are an important driving force for regulatory innovation. Taken together, our data suggest that TEs have significantly and continuously shaped gene regulatory networks during mammalian evolution.**

**[Supplemental material is available for this article.]**

A large portion of eukaryotic genomes is derived from transposable elements (TEs) (Adams 2000; International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002). TEs have been described as parasitic or junk DNA (Doolittle and Sapienza 1980). However, there is mounting evidence for their significant evolutionary contribution to the wiring of gene regulatory networks (Wang et al. 2007; Bourque et al. 2008; Feschotte 2008; Kunarso et al. 2010; Xie et al. 2010; Lynch et al. 2011; Rebollo et al. 2012b; Jacques et al. 2013), a theory rooted in Barbara McClintock's discovery that TEs can control gene expression (McClintock 1950, 1956; Britten and Davidson 1969; Feschotte 2008). TEs have the potential to affect phenotypes by driving coding, regulatory, and chromosomal structural changes that provide dynamism to genomes (TE-thrust hypothesis) (Oliver and Greene 2011). In spite of the enormous potential that these sequences have to affect gene expression and species, the functional potential of TEs was undercharacterized in the early days of genomics.

More recently, a series of genomics studies have demonstrated the role of TEs in establishing and rewiring gene regulatory networks (Feschotte and Gilbert 2012; Rebollo et al. 2012b). In particular, several studies found TEs to contain functional binding sites for transcription factors (TFs), including TP53, POU5F1, NANOG, and CTCF (Jordan et al. 2003; Bejerano et al. 2006; Wang et al. 2007; Polavarapu et al. 2008; Roman et al. 2008; Sasaki et al. 2008; Bourque 2009; Kunarso et al. 2010; Pi et al. 2010; Schmidt et al. 2012; Chuong et al. 2013; de Souza et al. 2013). Additional studies suggested that TEs can be epigenetically modified in a tissue-specific manner, thus providing potential tissue-specific regulatory elements (Jordan et al. 2003; Xie et al. 2013). Although most of these studies focused on single TFs or single biological systems (i.e., cell type or species), the broad range of observations encouraged the thought that the mechanism of TEs spreading TF binding sites is a general mechanism of regulatory network evolution and can impact many different TFs. What is the extent to which TEs have contributed to the ongoing evolution of gene regulation, and how have these TE-derived TF binding sites evolved? These thoughts motivated us to comprehensively evaluate the evolutionary contribution of TEs to the target sites of a large variety of TFs in the context of specific cellular epigenetic landscape, in search for further support of the theories proposed by Barbara McClintock, Roy J. Britten, and Eric H. Davidson (McClintock 1950, 1956; Britten and Davidson 1969).

In this study, we systematically compared genome-wide binding of TFs encoded by orthologous genes in human and mouse with respect to their relationship to TEs. We generated genome-wide binding profile data for 26 pairs of orthologous TFs in two pairs of cell lines in human and mouse, as well as comprehensive

[4]**These authors contributed equally to this work.**
**Corresponding authors: mpsnyder@stanford.edu, twang@genetics.wustl.edu**
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.168872.113.

epigenomic profiles for these cell lines, including six histone modification marks and DNA methylation. We confirmed that TEs contributed binding sites for TFs, but in a highly TF-specific and TE subfamily-specific manner. The vast majority of TE-derived binding sites were species-specific. These included sites derived from primate- or rodent-specific TE subfamilies, as well as species-specific sites contained within TE subfamilies that were shared between human and mouse, underscoring rapid binding site creation and turnover mediated via TE activity. In addition, we also discovered conserved TF binding within TE fragments shared between human and mouse. We found that TEs containing TF binding peaks strongly enriched for TF binding motifs, suggesting that sequence features of TEs predispose specific TEs for evolving binding sites for specific TFs. Binding of TFs to TEs was found to be strongly associated with the epigenetic status of the TEs, which, despite long being considered suppressed by epigenetic mechanisms, displayed strong enhancer and other active chromatin signatures, a large fraction of which were also cell type-specific. Taken together, our study provides by far the most comprehensive investigation of the interactions between TFs and TEs in human and mouse.

## Results

### Up to 40% transcription factor binding sites were derived from transposable elements

As part of the ENCODE (The ENCODE Project Consortium 2012) and Mouse ENCODE (The Mouse ENCODE Consortium et al. 2014) effort to annotate functional elements of the human and mouse genomes, we profiled the genome-wide occupancy for 26 pairs of TFs that are encoded by orthologous genes in human and mouse leukemia cell lines (K562 and MEL) and lymphoblast cell lines (GM12878 and CH12) using ChIP-seq technology (Johnson et al. 2007). The TFs analyzed here represent TFs that interact with DNA through specific consensus sequences (58%), chromatin-modifying and remodeling factors (23%), general transcriptional machinery (19%), and RNA polymerase II (POL2; POLR2A). A brief description of these TFs can be found in Table 1. Different TFs have dramatically different numbers of binding peaks (see Methods; Supplemental Table 1; Table 2), ranging from 162 peaks (KAT2A in GM12878) to 66,051 peaks (CTCF in K562). In total, we defined 695,042 binding peaks in the human genome and 679,820 binding peaks in the mouse genome that were associated with at least one of the 26 TFs binding in at least one cell type. When consolidated, these binding peaks comprised 94.25 Mb (3%) of the human genome, and 81.33 Mb (2.96%) of the mouse genome. Size distributions of these peaks were plotted in Supplemental Figure 1, and the numbers of binding peaks for each TF in each species were included in Table 2.

We defined a binding peak (defined by uniquely mapping reads) to be derived from a TE if the center of the peak falls within a TE annotated by RepeatMasker (see Methods; Smit et al. 1996-2010). The distribution of peak scores for peaks in TEs and non-TE regions was found to be largely the same (Supplemental Fig. 2). We found 135,422 peaks (19%) and 140,058 peaks (20%) were derived from TEs for human and mouse, respectively. Interestingly, the fractions and numbers of TE-derived binding peaks span a large range across different TFs (Fig. 1A; Supplemental Fig. 3, respectively), in human and mouse. On the lower end, 93 of 4655 peaks (2%) of KAT2A in the mouse genome were derived from TEs; on the higher end, 27,851 of 69,331 peaks (40%) of CTCF binding sites in the mouse genome were derived from TEs. Numbers of TE-derived binding peaks for each TF were included in Table 2. These

**Table 1. List of transcription factors (TFs) analyzed in this study along with a description of the TF and the cells in which their binding was assayed**

| TF names | Leukemia | | Lymphoblast | | Description | TF category |
|---|---|---|---|---|---|---|
| | K562 | MEL | GM12878 | CH12 | | |
| BHLHE40 | ✓ | ✓ | ✓ | ✓ | Basic helix-loop-helix family, member e40 | Sequence-specific transcription factor |
| CHD1 | ✓ | ✓ | ✓ | ✓ | Chromodomain helicase DNA binding protein 1 | Chromatin modifying and remodeling factor |
| CHD2 | ✓ | ✓ | ✓ | ✓ | Chromodomain helicase DNA binding protein 2 | Chromatin modifying and remodeling factor |
| CTCF | ✓ | ✓ | ✓ | ✓ | CCCTC-binding factor (zinc finger protein) | Sequence-specific transcription factor |
| E2F4 | ✓ | ✓ | ✓ | ✓ | E2F transcription factor 4 | Sequence-specific transcription factor |
| EP300 | ✓ | ✓ | ✓ | ✓ | E1A binding protein p300 | General transcription factor |
| ETS1 | ✓ | ✓ | ✓ | ✓ | v-ets erythroblastosis virus E26 oncogene homolog 1 (avian) | Sequence-specific transcription factor |
| GATA1 | ✓ | ✓ | | | GATA binding protein 1 (globin transcription factor 1) | Sequence-specific transcription factor |
| JUND | ✓ | ✓ | ✓ | ✓ | jun D proto-oncogene | Sequence-specific transcription factor |
| KAT2A | | | ✓ | ✓ | K (lysine) acetyltransferase 2A | Chromatin modifying and remodeling factor |
| MAFK | ✓ | ✓ | | | v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian) | Sequence-specific transcription factor |
| MAX | ✓ | ✓ | ✓ | ✓ | MYC associated factor X | Sequence-specific transcription factor |
| MAZ | ✓ | ✓ | ✓ | ✓ | MYC-associated zinc finger protein (purine-binding transcription factor) | Sequence-specific transcription factor |
| MXI1 | ✓ | ✓ | ✓ | ✓ | MAX interactor 1, dimerization protein | Sequence-specific transcription factor |
| MYC | ✓ | ✓ | ✓ | ✓ | v-myc myelocytomatosis viral oncogene homolog (avian) | Sequence-specific transcription factor |
| PAX5 | | | ✓ | ✓ | Paired box 5 | Sequence-specific transcription factor |
| POLR2A | ✓ | ✓ | ✓ | ✓ | Polymerase (RNA) II (DNA directed) polypeptide A | General transcription factor |
| RAD21 | ✓ | ✓ | ✓ | ✓ | RAD21 homolog (*S. pombe*) | Chromatin modifying and remodeling factor |
| RCOR1 | ✓ | ✓ | ✓ | ✓ | REST corepressor 1 | Chromatin modifying and remodeling factor |
| RDBP | ✓ | ✓ | | | Negative elongation factor complex member E | General transcription factor |
| SIN3A | ✓ | ✓ | ✓ | ✓ | SIN3 transcription regulator homolog A (yeast) | General transcription factor |
| SMC3 | ✓ | ✓ | ✓ | ✓ | Structural maintenance of chromosomes 3 | Chromatin modifying and remodeling factor |
| TAL1 | ✓ | ✓ | | | T-cell acute lymphocytic leukemia 1 | Sequence-specific transcription factor |
| TBP | ✓ | ✓ | ✓ | ✓ | TATA box binding protein (TBP) | General transcription factor |
| UBTF | ✓ | ✓ | | | Upstream binding transcription factor, RNA polymerase I | Sequence-specific transcription factor |
| USF2 | ✓ | ✓ | ✓ | ✓ | Upstream transcription factor 2, c-fos interacting | Sequence-specific transcription factor |

**Table 2. Number of TF binding peaks in the analyzed data sets and the corresponding number of TE-derived TF binding peaks**

| | HUMAN | | MOUSE | |
|---|---|---|---|---|
| TFs | Number of peaks | Number of TE-derived peaks | Number of peaks | Number of TE-derived peaks |
| BHLHE40 | 30,718 | 5639 | 43,807 | 6685 |
| CHD1 | 12,859 | 1468 | 9735 | 1401 |
| CHD2 | 18,640 | 2797 | 21,704 | 2717 |
| CTCF | 80,305 | 18,349 | 69,331 | 27,851 |
| E2F4 | 9179 | 334 | 791 | 31 |
| EP300 | 43,660 | 11,924 | 59,966 | 11,766 |
| ETS1 | 4120 | 788 | 42,756 | 6570 |
| GATA1 | 4069 | 1044 | 42,094 | 9173 |
| JUND | 45,969 | 11,562 | 8553 | 1353 |
| KAT2A | 162 | 13 | 4655 | 93 |
| MAFK | 19,309 | 5732 | 1829 | 383 |
| MAX | 51,547 | 9911 | 38,118 | 6081 |
| MAZ | 38,774 | 4132 | 10,266 | 593 |
| MXI1 | 19,368 | 1952 | 40,046 | 5488 |
| MYC | 25,106 | 3591 | 38,502 | 7864 |
| PAX5 | 30,673 | 6916 | 187 | 22 |
| POLR2A | 37,215 | 6096 | 29,478 | 3358 |
| RAD21 | 51,078 | 9809 | 56,093 | 18,477 |
| RCOR1 | 36,953 | 10,385 | 13,693 | 2493 |
| RDBP | 440 | 46 | 13,988 | 921 |
| SIN3A | 16,794 | 1151 | 25,382 | 2514 |
| SMC3 | 35,204 | 5489 | 40,234 | 11,431 |
| TAL1 | 26,210 | 8949 | 17,874 | 4225 |
| TBP | 25,228 | 3894 | 36,366 | 6681 |
| UBTF | 13,613 | 884 | 5063 | 177 |
| USF2 | 10,192 | 2567 | 9309 | 1710 |

data suggest that TEs indeed have widely contributed DNA elements to gene regulatory networks as binding sites for TFs, but the degree of contribution is highly TF-specific.

Individual TFs differed greatly in the number of binding peaks they had in the whole genome; thus, it could be expected that the more binding peaks a TF had, there would be proportionally more peaks overlapping with TEs. Unexpectedly, we found that for individual TFs, the percentage of TE-derived peaks positively scaled with the total number of peaks (Fig. 1B). TFs with higher numbers of total peaks not only had more peaks derived from TEs, but also larger fractions of peaks derived from TEs. These data are consistent with the hypothesis that TEs may have contributed in growing large repertoires of target sites for certain TFs (Wang et al. 2007; Feschotte 2008; Bourque 2009).

## Specific TE classes, families, and subfamilies were enriched for binding peaks

We next asked if different types of TEs had contributed similarly or differently to binding peaks of different TFs. Mammalian TEs are classified hierarchically into classes (LINE, SINE, LTR, and DNA elements), families (ERVK, ERVL, L1, L2, etc., constituting 56 families in human and 52 families in mouse), and many more subfamilies (928 in human and 790 in mouse). Similar to the previous analyses, we used RepeatMasker (Smit et al. 1996-2010) annotations and classifications of TEs to identify which class, family, and subfamily each of the TE-derived TF binding peaks belonged to.

Binding peaks of these TFs did not distribute uniformly across different TE classes, families, and subfamilies (Fig. 2A). At the class level, LTR elements were overrepresented in contributing binding peaks for human. LTR elements made up 19% human TEs, but they contained 39% of TE-derived binding peaks for human. This result agrees with previous studies that showed the participation of LTR-elements in the regulation of primate genes (Samuelson et al. 1990; Medstrand et al. 2001; Dunn et al. 2003; Feschotte 2008; Cohen et al. 2009). In contrast, SINE elements contributed more binding peaks than expected for mouse, such that 20% of TE-sequence space occupied by SINE elements harbored 54% TE-derived binding peaks. ERV/LTR sequences have been proposed to have a greater propensity to be co-opted for regulatory functions (Feschotte and Gilbert 2012; Rebollo et al. 2012b); therefore, it was not surprising that they were more enriched for containing TF binding sites in human. Alternatively, the observed enrichment of human LTR and mouse SINE elements might reflect these elements' lineage-specific abundance and mappability.

We further examined whether any TE subfamily contributed a significant amount of binding peaks for specific TFs. Genomic copies of the same TE subfamily are generally phylogenetically linked and result from the rapid deposition and expansion of the same transposable element. This evolutionary event has been hypothesized to be a major driving force for quickly creating a large number of target sites for a TF, if the TE contained a DNA motif that could be directly used by the TF or easily converted to binding sites of the TF (Feschotte 2008). As a measure of enrichment, we computed a log-odds ratio-based score for all pairwise relationships between TE subfamilies (a total of 928 for human and 790 for mouse) and TFs (26 for both human and mouse). The score estimates the log-odds ratio between the observed number of TF binding peaks overlapping specific TE subfamilies and the number of binding peaks expected by chance (see Methods). At a log-odds ratio cutoff of 1.5, which represents roughly a threefold enrichment of observing binding peaks within a specific TE subfamily over genome-wide binding peak background, we were able to define 710 pairwise relationships between specific TE subfamilies and specific TFs (527 for human and 183 for mouse) (Fig. 2B; Supplemental Fig. 4; Supplemental Table 2A,B). This number increased to 1031 when we lowered the cutoff to a log-odds ratio of 1, representing a twofold enrichment. Of the 710 pairwise relationships, 566 were involved with TE subfamilies of the LTR class (439 for human and 127 for mouse), further confirming the significant contribution of LTR elements to transcriptional regulation (Samuelson et al. 1990; Medstrand et al. 2001; Dunn et al. 2003; Feschotte 2008; Cohen et al. 2009). For some TFs, e.g., E2F4, KAT2A, and UBTF, no TE subfamilies enriched for their binding peaks (see Methods for the thresholds used to define enriched TE subfamilies), suggesting TEs were not a major source of binding sites for these factors.

Several TFs have colocalized binding peaks. Previous work has shown that CTCF colocalizes with proteins of the cohesin complex (RAD21 and SMC3) (Wendt et al. 2008; Nitzsche et al. 2011), and we observed the same but slightly weaker trend on TE-derived binding

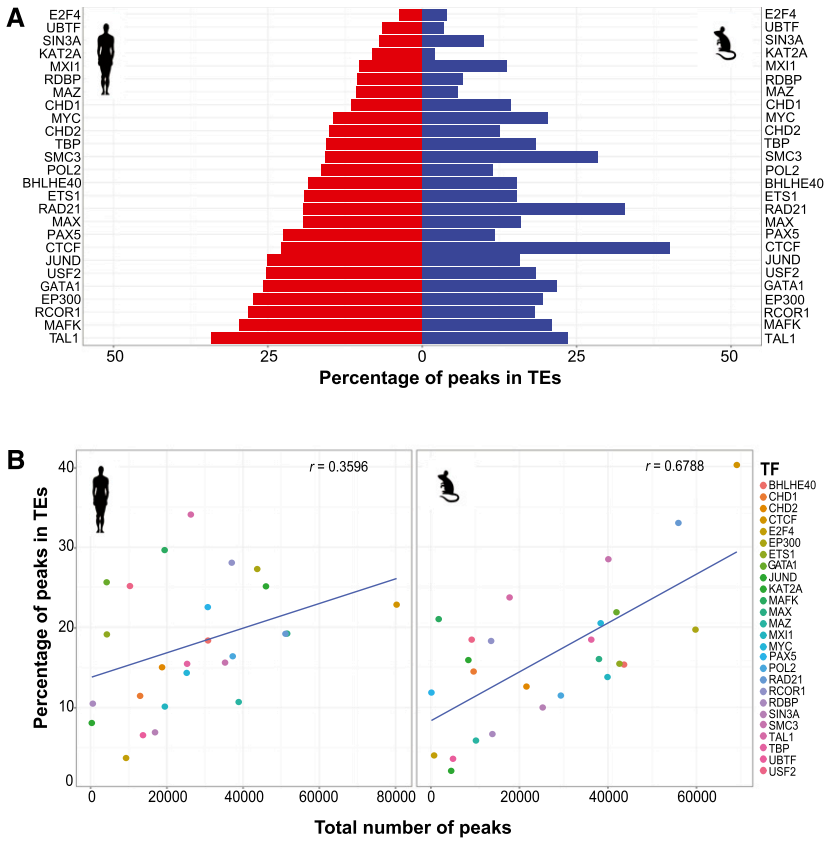


**Figure 1.** Different human and mouse TFs had different numbers and fractions of their binding derived from TEs. (*A*) Percentage of TF binding peaks that occurred in TEs, in human (*left* panel, red) and mouse (*right* panel, blue). (*B*) Correlation between the number of TF binding peaks in the genome and the percentage of the TF binding peaks in TEs in human (*left* panel) and mouse (*right* panel). Correlation between these two variables was measured with Pearson's correlation coefficient (*r*).

peaks (Supplemental Fig. 5). We also confirmed the previous finding that CTCF binding sites had a rodent-specific expansion via TEs (Supplemental Table 2B; Schmidt et al. 2012). In addition, we identified an independent, TE-mediated expansion in the human lineage (Supplemental Table 2A), driven by primate-specific TE subfamilies including *LTR13*, *LTR35*, *LTR60*, *LTR2C*, *HERVH48-int*, and *HUERS-P2-int* as well as conserved TE-derived binding sites that were shared between human and mouse (see below). CTCF and the cohesin complex (RAD21 and SMC3) are known for their role in mediating long-range interactions in the genome (Merkenschlager and Odom 2013). The cobinding of these factors is thought to structurally be involved in regulating gene expression. Our findings highlight a close relationship between TEs and the evolution of genome organization bridged by a pleiotropic factor.

Furthermore, we found that EP300, a general enhancer cofactor, colocalized with TE-derived binding events of several other TFs. Of 11,854 TE-derived EP300 peaks in human, 9204 were bound to TFs other than CTCF, RAD21, and SMC3, and 8805 of 11,735 EP300 peaks in mouse were bound to TFs other than CTCF, RAD21, and SMC3 (Supplemental Fig. 6). The majority of EP300 peaks in human and in mouse colocalized with other TFs, suggesting EP300 might be marking TE-derived binding sites as active enhancers. In contrast, a much smaller portion of the EP300 and cohesin-associated binding sites overlapped (1252 in human and 597 in mouse). This is in agreement with the distinct functions of CTCF and EP300, despite both being considered as pleiotropic factors. CTCF is known for its role in shaping chromatin domains (Phillips and Corces 2009) and orchestrating enhancer-promoter looping (Merkenschlager and Odom 2013), whereas EP300 is known for its role in marking active enhancers (Visel et al. 2009). In contrast to EP300, a majority of the cohesin-associated binding sites did not overlap with the other TFs studied here, suggesting a distinct role for CTCF and cohesion-related binding sites from binding sites of general TFs. A large fraction of binding sites of other TFs did not associate with EP300 or cohesin-related factors (37,249 in human and 23,258 in mouse), highlighting the importance of analyzing TF-specific data in addition to data of general factors (i.e., EP300). Additionally, this lack of association of TFs with EP300 could suggest that a large portion of these TE-derived TF binding peaks are not functional as transcriptional regulatory sites, possibly due to the failure to recruit necessary cofactors, or they might have evolved functions other than enhancers.

## Detecting TF–TE relationship by including nonuniquely mapped reads

TEs often appear as repetitive sequences in the genome, and they are notoriously difficult to map using short-read sequencing technology. Reads that map to multiple locations (defined as multireads), including different copies of the same TE, are typically discarded. Since our analysis until now has focused on binding peaks based by uniquely mapped reads (defined as unique reads), the TF–TE relationships we identified likely represented a lower bound. To address this issue, we recently developed a method to harness sequencing reads that map to multiple TE genomic copies by associating multireads to a TE subfamily (Xie et al. 2013). Here we adapted the method to detect TE subfamilies enriched for sequencing reads from any TF ChIP-seq experiment by using all reads, including nonuniquely mapped reads (see Methods).

Overall, we found that the TF–TE relationships obtained by using enrichment of both unique and multireads recapitulate some of the TF–TE relationships obtained by using enrichment of binding peaks, which were defined using unique reads only (Supplemental Fig. 7). TF–TE interactions detected based on enriched binding peaks were strongly supported by a high enrichment score calculated based on all reads (Fig. 3A). Of the 710 TF–TE relationships defined by enrichment of TF binding peaks, 148 were strongly supported by enrichment of sequencing reads. In addition, we also identified 215 new TF–TE relationships that enriched for sequencing reads but did not pass the threshold we used for binding peaks (Supplemental Table 3A,B). As expected, these TEs usually had lower mappability (defined as the level of uniqueness of sequences in a genome sequence assembly (Fig. 3B; Kent et al. 2002). For example, we identified that *RLTR4* enriched for POL2 sequencing reads but not POL2 peaks. *RLTR4* (LTRs of Moloney

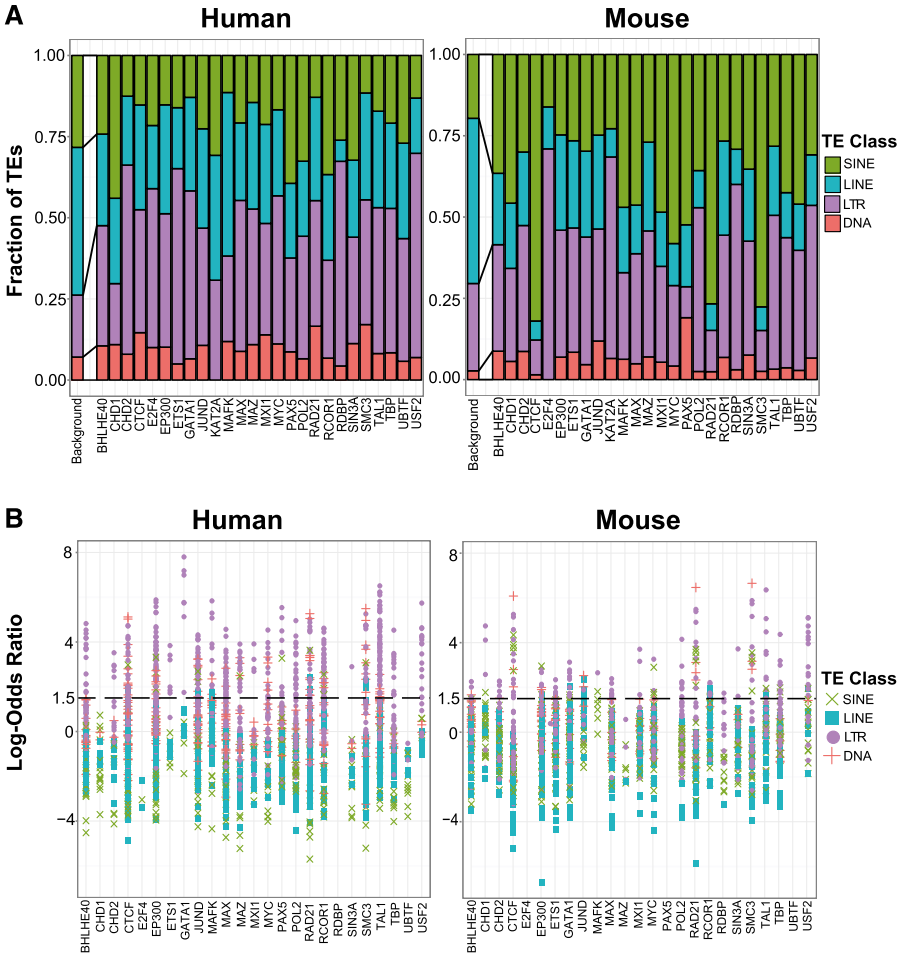


**Figure 2.** Specific TE classes and subfamilies enriched for TF binding peaks. (*A*) Proportion of the TF binding peaks that occurred in each TE class in human (*left* panel) and mouse (*right* panel). For comparison, "Background" represents the proportion of the genome (in bp) that each TE class constitutes. (*B*) Enrichment of TF binding peaks in TE subfamilies. We used a log-odds ratio (see Methods) for the definition of enrichment and estimated this for each TE subfamily in human (*left* panel), and mouse (*right* panel). Dots with the same color and shape represent TE subfamilies that belong to the same TE class. Several TFs lack data points, which represents no enrichment of binding peaks in TE subfamilies because they had ≤10 binding peaks occurring within the subfamily (see Methods).

Leukemia virus) is a rodent-specific ERV1 family TE and has 218 copies across the mouse genome, averaging 484 bp per copy. Based on their mappability scores, only 4.6% of *RLTR4* genomic sequences can be mapped by 36-bp reads (7.34% for 75-bp reads) (Fig. 3C). None of the copies overlapped with any binding peaks of POL2. However, in the MEL cell line, *RLTR4* was associated with 16 times more POL2 ChIP-seq reads than input control, indicating that at least some copies of *RLTR4* were associated with POL2 activity and might function as regulatory elements (Fig. 3D).

## Evolutionary dynamics of binding peaks derived from TEs

Human and mouse split ~75 million years ago, and it was estimated that most of their TEs are specific for either the primate or the rodent lineage (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002). Here, we determined the fractions of TE-derived binding events that are shared by both species and that are species-specific. To do this, we reciprocally mapped TE-derived binding peaks from one species to the other and evaluated whether orthologous regions existed in the other species; if so, how were the orthologous regions annotated with respect to TE and with respect to TF binding (see Methods)? We found that the majority of binding peaks derived from TEs were specific to the human lineage or mouse lineage, but we also found some peaks were conserved between human and mouse (Fig. 4A).

Overall, 3226 (2%) human TE-derived peaks and 1411 (1%) mouse TE-derived peaks were mapped syntenically to a binding peak of the same TF in the other genome; of these, 748 human-mouse peak pairs were annotated as the same TE, whereas the remaining were often annotated as TE in one species but not in the other species (Fig. 4A). Since the substitution rate in the mouse lineage was twice that of the human lineage after the primate-rodent split, we expect that TE sequences of ancient subfamilies (prior to the human-mouse split) might not be detectible in mouse. However, this could also reflect a limitation in annotating TE sequences at the genome-wide level, because a further examination of sequences not annotated as TEs revealed that they shared sequence identity with the TE sequences from the other species (Supplemental Fig. 8), suggesting that these human-mouse peak pairs represent TE-derived events of the same evolutionary origin. These TE sequences might have converted into TF binding sites independently in human and mouse; alternatively, a more parsimonious explanation is that they became TF binding sites before primate and rodent split. Importantly, these TE sequences exhibited a signature of purifying selection (Fig. 4B), underscoring their potentially conserved function.

Different TFs had a different number of conserved TE-derived binding events. We found that certain TFs, such as CTCF, RAD21, and SMC3, had many more conserved TE-derived binding events. Of the 3226 human TE-derived binding peaks whose occupancy was conserved in mouse, 762, 544, and 401 corresponded to the binding of CTCF, RAD21, and SMC3, respectively. Similarly, of the 1411 mouse TE-derived binding peaks whose occupancy was conserved in human, 301, 257, and 152 corresponded to the binding of CTCF, RAD21, and SMC3, respectively. Most of the TE subfamilies that encode conserved binding sites of these TFs were DNA and LTR elements shared by human and mouse. The major contributors of the conserved binding sites were *MER91B*, *LTR41*, *LTR41B*, *MER20* (Lynch et al. 2011), and *MER20B*. Here we present two examples of conserved binding sites derived from TEs that landed in the ancestral genome of human and mouse (Fig. 4C,D), evidence that TEs contributed TF binding sites before the primate-rodent split.

In contrast, the majority of TE-derived binding peaks—132,197 (98%) of human and 138,649 (99%) of mouse—were species-specific (Fig. 4A). These binding peaks can be further separated into two classes based on the phylogenetic relationship of

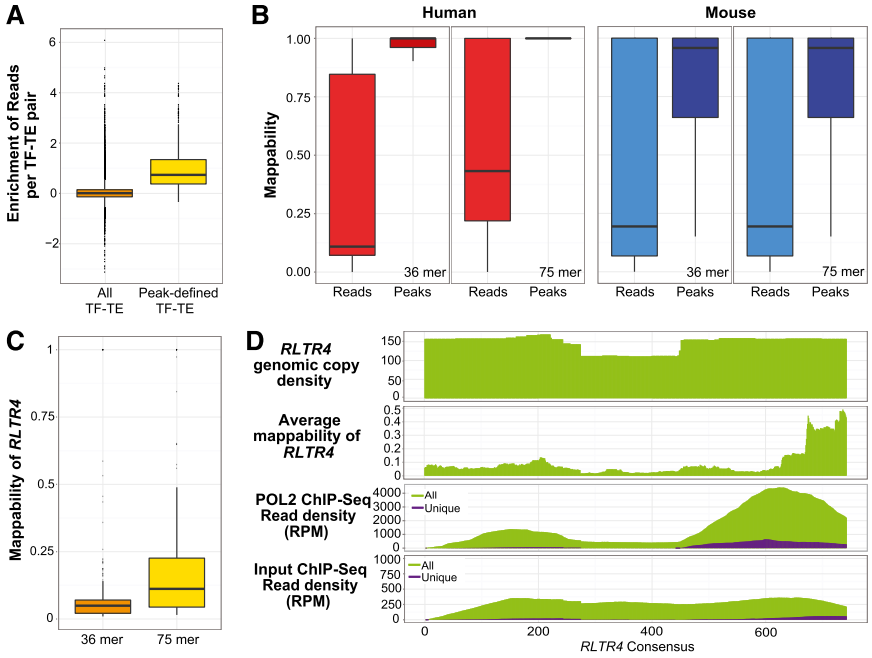


**Figure 3.** Including nonuniquely mapped reads (multireads) captured additional binding events on TEs. (*A*) Comparison between the enrichment of reads (see Methods) for all TF–TE pairs and TF–TE pairs that were enriched for interactions defined by peaks (see Methods). (*B*) Distribution of mappability scores (for 36-mer and 75-mer, respectively) of TE subfamilies. (Peaks) TE subfamilies enriched for TF binding peaks; (reads) TE subfamilies enriched for ChIP-seq reads (including both unique and nonunique reads). We calculated the mappability score (level of sequence uniqueness) for each genomic copy of a particular TE subfamily using mappability tracks downloaded from the UCSC Genome Browser, where 1.0 = 100% mapped uniquely, and 0 = 0% mapped uniquely. (*C*) Distribution of the mappability scores for the genomic copies of the *RLTR4* subfamily in mouse by using 36-mer and 75-mer sequence reads. (*D*) Comparison of the ChIP-seq signal using unique and all reads in *RLTR4*. The *first* panel shows genomic coverage of *RLTR4* copies on the *RLTR4* consensus sequence. The *second* panel shows the average mappability (75-mer) score (mappability file was downloaded from the UCSC Genome Browser). The *third* panel shows the accumulation of POL2 ChIP-seq reads over *RLTR4* consensus, with purple representing unique reads, and green representing both unique and non-unique reads. The *fourth* panel shows the accumulation of ChIP input reads over the *RLTR4* consensus sequence.

the TE subfamilies: subfamilies that are common to human and mouse and subfamilies that are specific to either the primate or the rodent lineage. More than half the species-specific TF binding peaks were derived from TE subfamilies that are specific to either the primate or rodent lineage. From the annotation of TEs in Repbase (Jurka et al. 2005), human and mouse share ~450 TE subfamilies, and these TE subfamilies likely entered the genome of the common ancestor before the primate-rodent split. When we examined the 710 pairwise TF–TE relationships representing TE subfamilies that enriched for TF binding peaks (defined earlier), we found 266 TE subfamilies in human and 77 in mouse that were shared by human and mouse. Of these shared TE subfamilies in human and mouse, 41 were coenriched, representing eight TFs. The "coenrichment" pattern suggested that the ancestral TE might have played a role in spreading binding sites of TFs even before the primate-rodent split. These shared TE subfamilies that enriched for TF binding peaks were responsible for 14,778 human peaks (867 of which were conserved) and 2345 mouse peaks (387 of which were conserved).

The data also revealed that, even for TF binding peaks derived from TE subfamilies that were shared by human and mouse, the majority of them were not conserved. Of 95,682 TE-derived binding peaks from shared TE subfamilies in human, 2946 were conserved. Similarly, of 28,287 TE-derived binding peaks from shared subfamilies in mouse, 1208 were conserved. There are several possible explanations for this. First, although the TE subfamily is shared between human and mouse, the transpositions may be lineage specific after the primate-rodent split. Therefore, the initial element might be shared in human and mouse, but the various instances in each genome will be different. A second mutually exclusive explanation is that shared TE subfamilies have accumulated lineage-specific substitutions and mutations in the shared TE fragments, thereby removing or rendering them unrecognizable in the other species.

These TE-derived, species-specific TF binding peaks could have contributed to species-specific functions for the TFs. As we noted earlier, some TFs had TE-driven expansion of binding peaks in either human or mouse. The expansion might allow the TF to exploit a bigger target gene reservoir and evolve new regulatory functions. For example, human acquired 11,562 new TE-derived binding peaks for JUND. We examined differences in functional annotation of genes near these binding peaks using GREAT (McLean et al. 2010). As expected, peaks common to human and mouse were associated with key functions, including apoptosis and interleukin signaling pathways. In contrast, the human-specific, TE-derived binding peaks of JUND were associated with functions including lipopolysaccharide-mediated signaling pathway and macrophage differentiation. These observations are consistent with previous studies showing differences in metabolism (Ames et al. 1993; Demetrius 2005) and in the immune system (Mestas and Hughes 2004) between human and mouse. Results for other TFs are summarized in Supplemental Table 4A and B.

## TEs that contributed TF binding peaks contained TF binding motifs

We next examined sequence features of TEs with respect to specific TF binding motifs. Specifically, we asked whether TE sequences that underlie binding peaks contained binding motifs of the binding TF. Not all the TFs being analyzed in this study have published binding motifs; therefore, we designed a strategy to evaluate motif enrichment in an unbiased fashion (see Methods). In brief, for each TF, we de novo predicted motifs from sequences of its binding peaks defined by ChIP-seq assay after removing all repetitive sequences; this usually resulted in several top ranking motifs. We reasoned that for orthologous, sequence-specific TFs, their binding specificity should be conserved between human and mouse (Cheng et al. 2014); therefore, we selected a specific motif only if the prediction made on human agreed with the prediction made on mouse. By these criteria, we were able to define specific motifs for 19 pairs of TFs (Supplemental Table 5). We then examined TE sequences that were part of a binding peak for enrichment (see Methods for enrichment calculation) of the corresponding

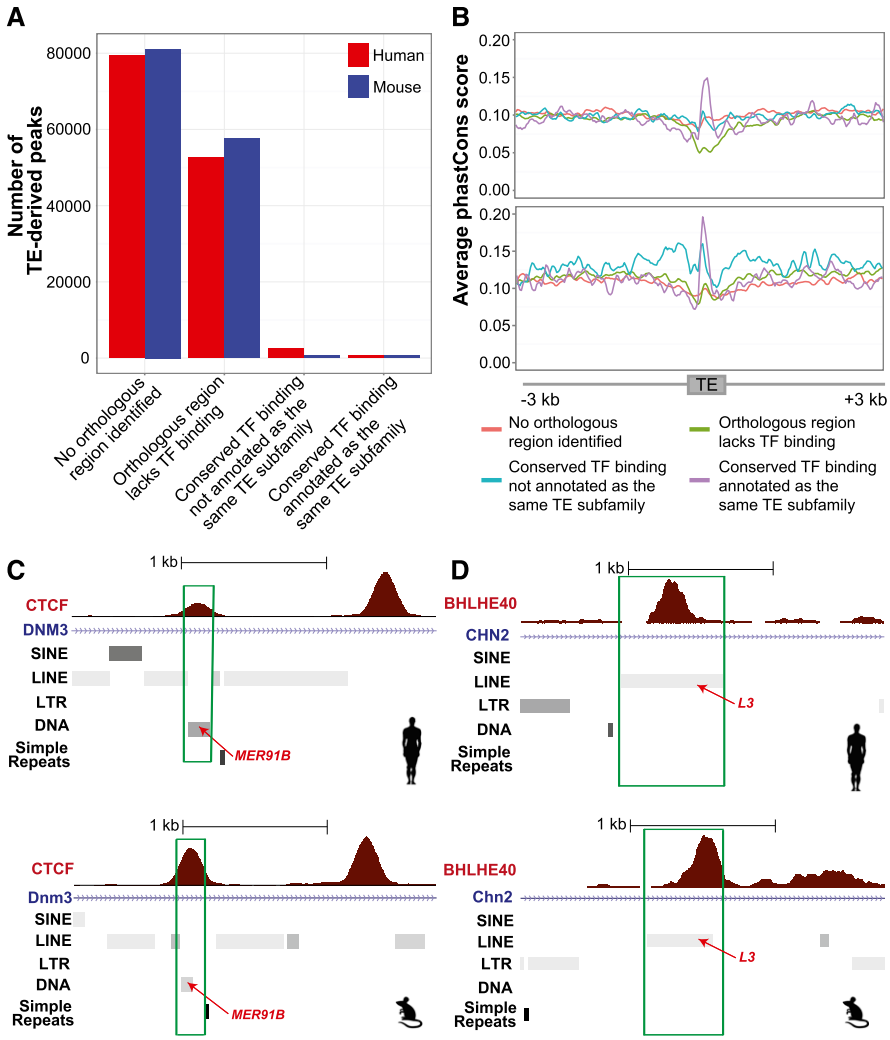


**Figure 4.** The majority of TE-derived binding peaks were species-specific, whereas the shared ones exhibited sequence conservation. (*A*) Bar plot showing the numbers of binding events in various categories of conservation. We classified TE-derived binding events in each species into four categories: (1) no orthologous region identifiable for the TF binding peak; (2) orthologous region lacks TF binding; (3) conserved TF binding, not annotated as the same TE subfamily in both species; and (4) conserved TF binding, annotated as the same TE subfamily in both species. (*B*) Distribution of phastCons scores (see Methods) in 6-kb windows centered on the TE sequences in the four categories defined above. The signal was averaged across 25-bp bins and smoothened for this plot. TE-derived binding peaks that could be mapped to binding peaks derived from the same TE subfamily in the other species exhibited increased sequence constraint over the background. (*C*,*D*) UCSC Genome Browser images of TE-derived TF binding peaks, whose occupancy was conserved between human (*upper* panel) and mouse (*lower* panel). The browser images correspond to (*C*) CTCF binding encoded on *MER91B* fragments in human and mouse, and (*D*) BHLHE40 binding encoded on *L3* fragments in human and mouse.

motif. Indeed, we observed significant enrichment of motif sites within these TE sequences, although the prediction of the motif was made on TE-free sequences (Fig. 5A). This result suggests that TEs that contributed TF binding peaks shared similar sequence features as non-TE sequences that contributed TF binding peaks, at least at the level of enriched motifs.

A similar conclusion can be reached by examining motif enrichment within TE subfamilies that enriched for binding peaks (Fig. 5B), such that a majority of the TF–TE relationships were supported by a high motif enrichment score. However, having a sequence motif or enriching for sequence motif did not perfectly predict binding. For example, we predicted that *MER91B* deposited CTCF binding sites in the mouse genome (Fig. 5C). In some genomic fragments, the region where predicted CTCF binding motif resided was deleted or modified, and as a result, these fragments did not correlate with CTCF binding. Of the 215 *MER91B* elements, 31 were predicted to have a CTCF motif and 36 were bound in mouse (24 of which had a CTCF motif) (Fig. 5C). The occurrence of CTCF binding without a predicted motif could be a result of the colocalization of CTCF with proteins of the cohesin complex (Nitzsche et al. 2011). Similarly, of 259 *LTR18A* elements, 72 were predicted to have a MAFK binding site, and 28 were bound by MAFK (26 of which had the MAFK motif) (Fig. 5D). Interestingly, when we examined the *LTR18A* fragments that lacked MAFK motifs and peaks, we identified mutations in the binding site (Supplemental Fig. 9). These mutations occurred in high-information positions of the MAFK binding site, suggesting that it might interfere with the binding of the MAFK. We also identified 2026 TE subfamilies enriched for having predicted binding sites for one of the TFs in the study (Supplemental Fig. 10) but that did not exhibit any binding peaks. These results suggest that specific sequences within TEs are important or required for TF binding, but the sequences alone do not guarantee binding, at least in the cell types we assayed.

## TEs contributing binding peaks had active epigenetic signatures

TEs are thought to be silenced in somatic cells by epigenetic mechanisms, including DNA methylation, to suppress mutational insertions caused by TE transposition and TE-mediated changes of gene expression (Morgan et al. 1999; Bird 2002; Slotkin and Martienssen 2007). However, recent studies have shown that some TEs exhibit a regulated epigenetic status and could serve as tissue-specific enhancers (Ekram et al. 2012; Rebollo et al. 2012a; Xie et al. 2013). Thus, we examined the epigenetic signatures of TE-derived TF binding peaks. We profiled six histone modifications (i.e., H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3) assayed by ChIP-seq, and DNA methylation, assayed by two complementary technologies: MeDIP-seq, and MRE-seq (Maunakea et al. 2010; Stevens et al. 2013; Zhang et al. 2013).

We discovered two distinct epigenetic signatures for TEs that contributed TF binding peaks. TEs that contributed binding peaks for 23 of 26 pairs of TFs had a clear epigenetic signature associated with active regulatory sequences (promoters and enhancers), in-

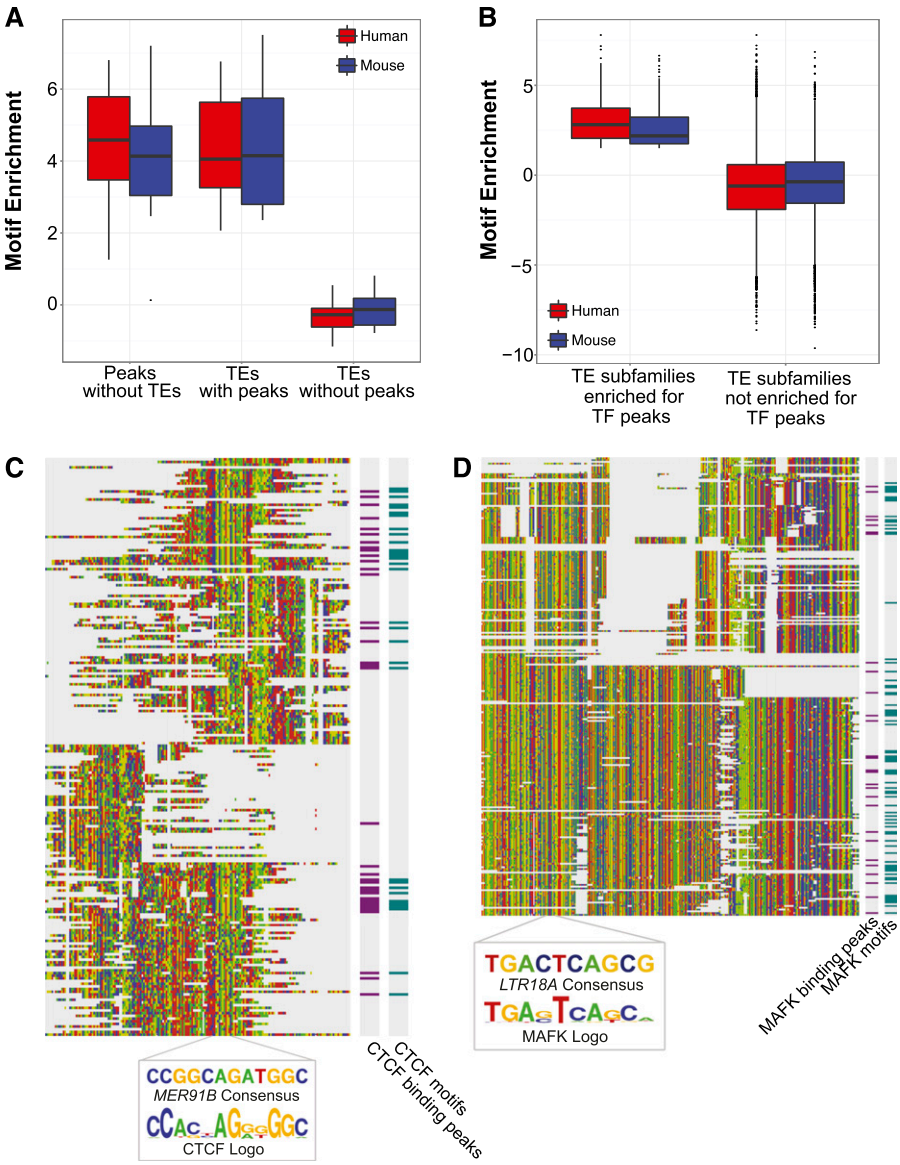


**Figure 5.** Binding motifs of TFs were enriched in TE-derived binding peaks. (*A*) Distribution of motif enrichment scores (log-odds ratio) (see Methods) in non-TE, TF binding peaks (training set for de novo motif prediction), TE-derived TF binding peaks (test set), and TEs without peaks (control). (*B*) Distribution of motif enrichment scores in TE subfamilies that were enriched in TF binding peaks compared with all other TE subfamilies. (*C*) Multiple sequence alignment of *MER91B* genomic copies ($n$ = 215) from the mouse genomes and the *MER91B* consensus sequence (*bottom* row of the alignment). *Beside* the alignment are indications of genomic copies that had CTCF binding peaks (purple) and CTCF motifs (green). (*D*) Similarly, multiple sequence alignment of *LTR18A* genomic copies ($n$ = 259) from the human genome along with the *LTR18A* consensus sequence (*bottom* row of the multiple-sequence alignment). Annotated on the *right* are indications of genomic copies that had MAFK binding peaks and MAFK motifs. Nucleotides in alignments are color-coded: (A) green; (C) blue; (G) yellow; (T) red.

cluding increased H3K27ac, H3K4me1, H3K4me3, and reduced DNA methylation (Fig. 6A). This TE-associated epigenetic signature was qualitatively similar to the signature associated with the non-TE sequences underlying binding peaks (Fig. 6B). The lower signal in the epigenetic profile of TEs that contributed TF binding peaks, compared with the peaks that did not associate with TEs, is likely a result of the reduced mappability of TEs (Supplemental Fig. 11). In contrast, TEs that did not contribute TF binding peaks exhibited epigenetic status consistent with being in a silenced chromatin state (Fig. 6C). TEs from which 23 (of 26) TFs derived their binding sites had a signature of increased H3K27ac, H3K4me1, H3K4me3 and reduced DNA methylation (Supplemental Fig. 12). CTCF, RAD21, and SMC3 (cohesin-associated factors) had a distinct epigenetic profile from other TFs (Supplemental Fig. 13).

Taken together, our data confirmed that a small fraction of TEs exhibited an active epigenetic signature of DNA regulatory elements, at least in the cell types we examined. The signature was shared between binding peaks contributed by TE sequences and by non-TE sequences, suggesting their common regulatory potential. Importantly, the phenomenon seemed to be conserved between human and mouse, despite the very different TF binding landscape across the two genomes.

## Epigenetic profiles of TEs were associated with cell type-specific TF binding

Our data also confirmed that cell type-specific TF–TE associations were strongly associated with a cell type-specific epigenetic landscape. Overall, we observed a slightly larger fraction of TE-derived TF binding peaks in leukemia cell lines (human: 19%; mouse: 22% of total TF binding peaks) than in lymphoblast cell lines (human: 15%; mouse: 19%) (Supplemental Fig. 14). For individual TFs, the range of cell type-specific TE-derived peaks was from nine (for E2F4 in MEL) to up to 22,141 (for CTCF in MEL). For human, there were 29,117 (64% of 45,252 TE-derived peaks in GM12878) TE-derived binding peaks that were lymphoblast specific, and 68,560 (81% of 84,695 TE-derived peaks in K562) TE-derived binding peaks that were leukemia specific. Similarly, for mouse, there were 48,292 (60% of 80,065 TE-derived peaks in CH12) TE-derived binding peaks that were lymphoblast specific, and 47,172 (60% of 78,945 TE-derived peaks in MEL) that were leukemia specific (Fig. 7A). On average, 72% and 60% of TE-derived TF binding peaks were cell-type specific in human and mouse, respectively. Epigenetic profiling of these cell type-specific TE-derived binding peaks in both cell types exhibited clear cell-type specific epigenetic signatures (Fig. 7B). The cell type-specific TE-derived binding peaks were associated with active chromatin, including increased H3K27ac, H3K4me1, H3K4me3, and reduced DNA methylation. The same TE fragments were depleted for these epigenetic marks in the other cell line, highlighting distinct cell-type specificity in their regulatory potential.

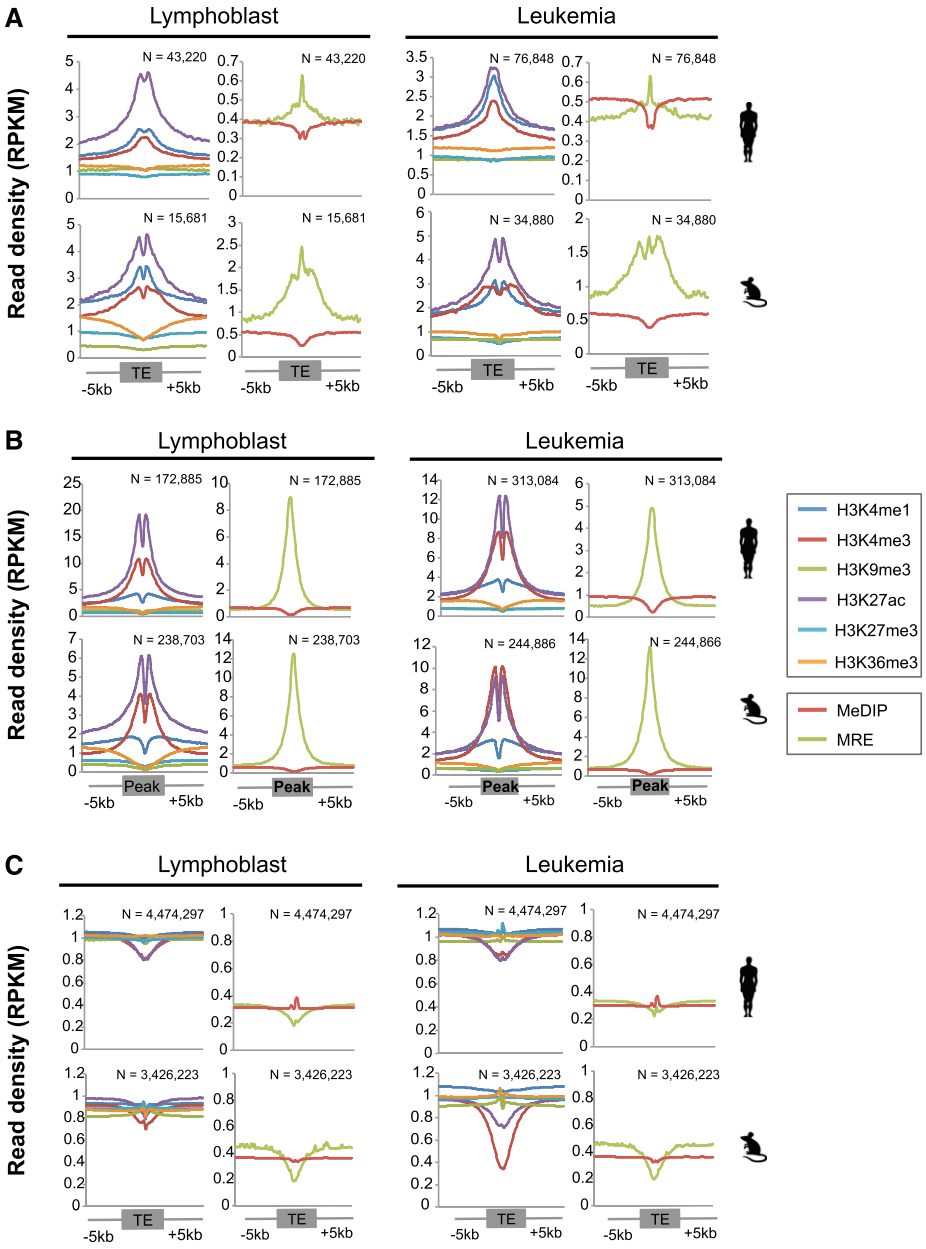


**Figure 6.** Epigenetic profile of TE-derived TF binding peaks in lymphoblast and leukemia cell lines in human and mouse. All figures represent the average signal density at 50-bp resolution over a 10-kb window centered on the genomic regions of interest (i.e., TEs or peaks). *Left* panels display profiles of histone modification marks, and the *right* panels display DNA methylation data. (*A*) Epigenetic profile of TE-derived TF binding peaks, which represents TEs that contained binding peaks for any one TF in human (*upper* panels) and mouse (*lower* panels). (*B*) Epigenetic profile of TF binding peaks that do not overlap any TE. (*C*) Epigenetic profile of TEs that do not overlap any TF binding peak.

Intriguingly, 3643 TE-derived binding peaks were not only conserved between human and mouse but also conserved in their cell type specificity (Fig. 7C). Of the cell type-specific peaks in human, 1271 lymphoblast-specific and 1076 leukemia-specific TE-derived peaks were mapped to syntenic binding peaks in the corresponding mouse cells. Similarly in mouse, of the cell-specific peaks, 693 lymphoblast-specific and 603 leukemia-specific TE-derived peaks were mapped to syntenic binding peaks in the corresponding human cells. Taken together, this suggests that TEs can impact the regulatory landscape of cells in a cell type-specific manner, and these might be TEs that have similarly affected human and mouse cells.

## Discussion

Transposable elements (TEs) are no longer discarded in genomic research. Instead, mounting evidence suggests that they have played important roles in the evolution of gene regulation. The idea that TEs could carry and deposit binding sites for TFs is not new. In fact, this mechanism was proposed by Barbara McClintock (McClintock 1950, 1956) when she first discovered TEs in maize, and has since been refined by generations of scientists (Britten and Davidson 1969; Feschotte 2008). Modern genomic studies (Wang et al. 2007; Bourque et al. 2008; Roman et al. 2008; Bourque 2009; Kunarso et al. 2010;

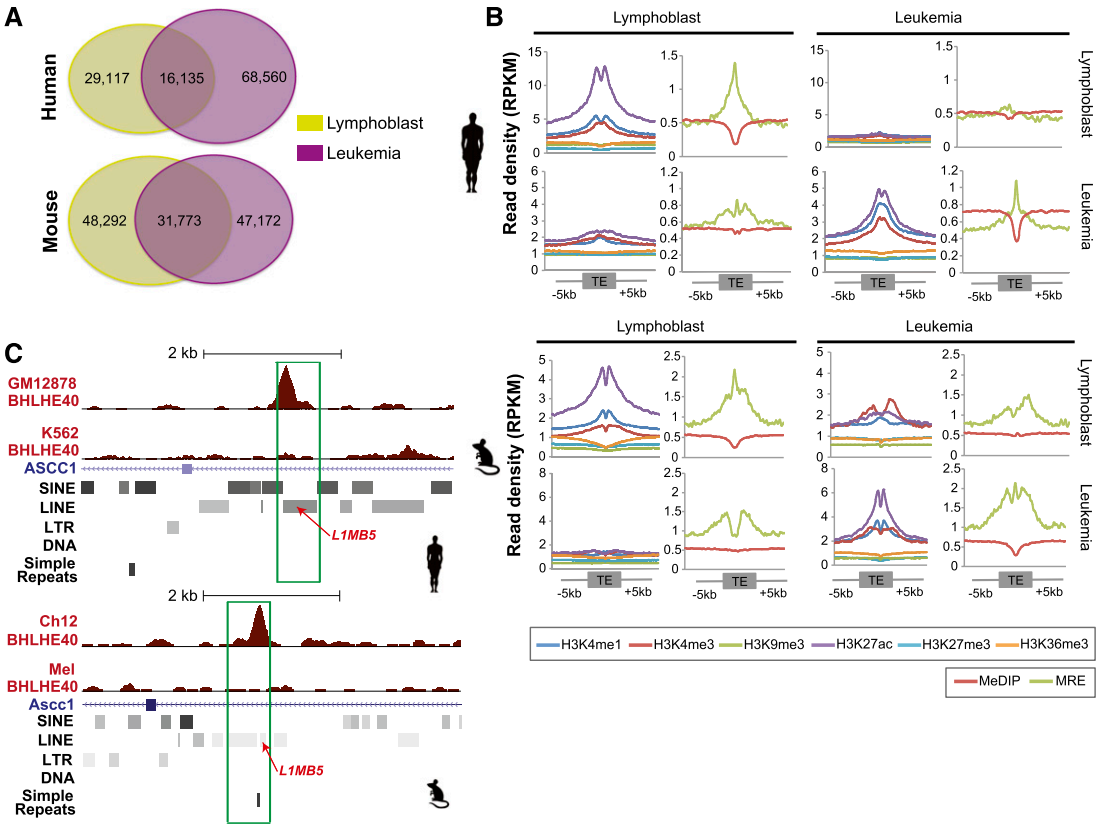


**Figure 7.** TE-derived, cell type-specific TF binding peaks. (*A*) Venn diagrams represent the numbers of cell type-specific TE-derived binding events in human (*top* panel) and mouse (*bottom* panel). The numbers represent the total number of instances in which a TE-derived TF binding peak for a particular TF was found in one cell type but not in the other. (*B*) Comparison of the epigenetic profiles of TEs that contained TF binding peaks. These represent TEs that contained a TF binding peak for any TF in one cell line but not in the other. For comparison, we plotted the average epigenetic profile of TEs that contributed cell type-specific peaks in both cell types. (*C*) UCSC Genome Browser view of a lymphoblast-specific TE-derived (*L1MB5*) BHLHE40 binding peak that was conserved between human and mouse.

Pi et al. 2010; Lynch et al. 2011; Rebollo et al. 2012b; Schmidt et al. 2012; Chuong et al. 2013; Jacques et al. 2013; Kapusta et al. 2013; Xie et al. 2013) provided more evidence for this mechanism.

Certain lineage-specific TEs were shown to encode binding sites for several TFs. For example, some primate-specific ERVs contained >30% of the binding sites for the tumor suppressor protein, TP53 (Wang et al. 2007). Similarly, ~20% of POU5F1 and NANOG binding sites were contributed by lineage-specific TEs in humans and mice (Kunarso et al. 2010), and a rodent-specific expansion of CTCF binding sites was also connected to retrotransposons (Schmidt et al. 2012). Interestingly, Kunarso et al. (2010) also showed that the lineage-specific TEs wired new genes into the human pluripotency transcriptional program. Likewise, a eutherian-specific TE, *MER20*, was suggested to have wired ~13% of pregnancy-related genes involved in signaling pathways related to implantation into endometrial stromal cells during the evolution of pregnancy in placental mammals (Lynch et al. 2011). Other than TF binding, recent reports have shown that TEs contribute DNase I hypersensitivity sites (Jacques et al. 2013), and may also contribute to the evolution and expression of lncRNA (Kapusta et al. 2013). Additionally, hypomethylated TEs have been shown to associate with tissue-specific enhancers (Xie et al. 2013). Taken together, TEs form an effective model for rewiring gene regulatory networks. The large portion of mammalian genomes that TEs represent is thought to have provided raw material for the evolution of *cis*-regulatory elements (Feschotte 2008), possibly via binding site turnover or via spreading of binding sites when the TE transposes (Feschotte 2008).

We have conducted here by far the most comprehensive study of interactions between TFs and TEs. In summary, our study made the following major discoveries. We found that TEs have contributed on average ~20% of TF binding sites in two representative cell lines in human and mouse. Of the 26 TFs we analyzed here, the extent to which TEs contribute to TF binding peaks exhibited TF-specific differences. Although certain TE subfamilies were shared between human and mouse genomes, a very small portion of peaks were conserved between human and mouse; most of the TE-derived binding peaks were species-specific. Importantly, our data confirmed that epigenetic regulation of TEs might be much more dynamic than previously thought (Morgan et al. 1999; Bird 2002; Slotkin and Martienssen 2007; Ekram et al. 2012; Rebollo et al. 2012a; Xie et al. 2013). TEs that were bound by TFs also enriched for enhancer epigenetic marks, such as increased H3K4me1, H3K27ac, and reduced DNA methylation, and often in a cell type-specific manner. Binding site motifs were strongly associated with TF binding but they did not perfectly predict binding. In this regard, TE-derived sequences behave no different from non-TE genomic sequences. The interplay between sequence features, TF binding, and epigenetic modification of TE sequences can only be elucidated with additional experimentation. Taken together, our results support the model of Britten and Davidson (1969) of TEs contributing to the

evolution of TF binding sites and potentially rewiring gene regulatory networks.

TE-derived TF binding peaks in human and mouse shared general characteristics such as their epigenetic profiles, but differed greatly in their genomic distribution such that majority of these peaks were species-specific. Interestingly, we identified distinct functional enrichment of genes associated with species-specific binding peaks, suggesting that TEs might have contributed to the evolution of species-specific regulatory functions and perhaps contributed to the phenotypic differences between species. A fundamental question that needs to be addressed next is how many of these TE-derived TF binding peaks are biologically functional. The binding events we report here indicate a biochemical activity of TF-DNA association, but whether these TE-derived TF binding peaks can influence expression of genes remains to be investigated. Alternatively, these TE sequences could function in nonconventional ways. For example, they might not directly result in a transcription read out, but could provide a buffer of extra binding sites to trap transcription factors or serve as a "landing pad" to allow transcription factors to quickly attach to and scan DNA. TEs clearly provided materials for evolving new binding sites and represent an efficient mechanism for rapid TF binding site turnover. We note that a functionally conserved binding site of a TF that resides in unique genomic sequences could also be derived from a TE, but the event may be difficult to identify as TE-derived if either the sequence context of the binding site has degenerated, or the event is simply too ancient to detect using sequence comparison, which more easily detects younger TEs (de Koning et al. 2011). Therefore, the number of TE-derived TF binding sites we reported here is likely a lower bound. Because these reported TE-derived binding sites are generally quite young, they may also be transient in the context of evolutionary time. The majority of them may be functionally neutral and disappear as the species continue to evolve, but a select few might stand the test of evolution if they convey fitness advantage for the species. In this regard, perhaps at least one function of TE-derived binding sites is to provide material from which regulatory innovation can be evolved.

## Methods

### ChIP-seq

Chromatin immunoprecipitation was carried out as previously described (Landt et al. 2012). Cultured cells for biological replicates were grown in separate batches and at separate times. Briefly, $5 \times 10^7$ cells were grown to a density of 0.6–0.8 $\times$ $10^6$/mL and then cross-linked in 1% formaldehyde for 10 min at room temperature. Nuclear lysates were sonicated using a Branson 250 Sonifier (power setting 7, 100% duty cycle for $12 \times 20^{-s}$ intervals), such that the chromatin fragments ranged from 50 to 2000 bp. Protein-DNA complexes were captured on Protein A/G agarose beads (Millipore #16-156/16-266) and eluted in 1% SDS TE buffer at 65°C. Following cross-link reversal and purification, the ChIP DNA sequencing libraries were prepared as described before (Kasowski et al. 2010) and sequenced on an Illumina Genome Analyzer II. All the data sets had reads that were 36 bp long.

### Data sources and data processing

TF ChIP-seq data sets for the 26 TFs in the two human-mouse pair of cell lines were processed by a uniform processing pipeline (Landt et al. 2012) and obtained from Cheng et al. (2014) via the consortia's Data Coordination Center. Reads were mapped by BWA (Li and Durbin 2009), and only reads that can be mapped to exactly one location in the genome were retained. We used the SPP peak caller (Kharchenko et al. 2008) to identify and score (rank) potential occupancy peaks. For obtaining optimal thresholds, we used the Irreproducible Discovery Rate (IDR) framework (https://sites.google.com/site/anshulkundaje/projects/idr) to determine high confidence occupancy events by leveraging the reproducibility and rank consistency of the identified peaks across replicate experiments of a data set. As recommended by the ENCODE working group, a cutoff of 0.2 was used with IDR.

We analyzed the extent of TE-mediated expansion of TF binding sites, using merged peaks from the two cell lines in human, and mouse. To merge data sets, we used the mergeBed function with default parameters, from the BEDTools package (Quinlan and Hall 2010). We also analyzed the extent of TE-derived TF binding specifically in each cell. For this, we merged the data sets from various institutions that exist for each cell line into one data set.

To determine the epigenetic state of our regions of interest, we used ChIP-seq data sets for six histone marks: H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. We used the density of the aligned reads, available in the Downloads section of the ENCODE data hub. Each data set had replicates that were averaged for the final analyses. We also assayed the DNA methylation levels in all four cell lines (see below).

### Enrichment of TF ChIP-seq binding peaks in TEs

We used enrichment calculation to evaluate the extent of TF binding peaks that were derived from TE. To identify TE-derived TF binding peaks, we required that the centers of the peaks overlapped with TE fragments, which were annotated using RepeatMasker (Smit et al. 1996-2010), in the human (hg19), and mouse (mm9) assembly (files were downloaded from the UCSC Genome Browser). We used the intersectBed tool (with default parameters) from the BEDTools package (Quinlan and Hall 2010) to calculate the intersection. Enrichment of TF binding peaks in TEs was defined as

$$LOR_{i,j} = \log_2 \left( \frac{\text{Number of TF '}i\text{' peak centers in TE subfamily '}j\text{'} / \text{Length of TE subfamily '}j\text{'(kb)}}{\text{Number of TF '}i\text{' peaks in the genome} / \text{Genome length(kb)}} \right)$$

To identify TE subfamilies that were enriched for TF binding peaks, we used a threshold of 1.5 to identify subfamilies, which represents approximately a threefold enrichment. To overcome TF–TE candidates that had high enrichment values resulting from very few TE instances, we required that (1) the number of genomic fragments in a TE subfamily should be greater than 30, and (2) the number of peaks overlapping the fragments of a TE subfamily should be greater than 10.

### Enrichment of ChIP-seq reads for TF binding data in TEs

In an attempt to overcome issues of poor mappability of TEs, we adapted our recently published repeat-alignment pipeline (RAP) (Xie et al. 2013). The adapted pipeline uses all (including nonunique) sequencing reads and maps them to the TE consensus. Each TE subfamily has a consensus sequence that was used for the genome-wide annotation of TEs, curated by RepeatMasker (Smit et al. 1996-2010). This provides a normalized signal (RPKM) for each TE subfamily, based on the alignment to the TE consensus.

$$\text{RPKM} = \frac{\text{Number of reads}(R)\text{that map to a TE consensus sequence}^* 10^9}{\text{Length of the TE consensus sequence}(K)^* \quad \text{Total number of mapped reads in the data set } (M)}$$

We did the RPKM calculation for each TE subfamily using both ChIP-seq reads and input reads. The enrichments were calculated by the following equation:

$$LOR_{i,j} = \log_2\left(\frac{\text{RPKM of TF '}i\text{'reads that map to TE subfamily '}j\text{'}}{\text{RPKM of input reads that map to TE subfamily '}j\text{'}}\right).$$

### Occupancy conservation of TE-derived TF binding sites

To identify conserved TF binding, we used a one-to-one nucleotide mapper called bnMapper (O Denas, R Sandstorm, Y Cheng, K Beal, J Herrero, RC Hardison, and J Taylor, in prep.) (https://bitbucket.org/james_taylor/bx-python/wiki/bnMapper) and mapped ChIP-seq binding peaks between human and mouse and vice versa. The mapping strategy in this tool is bijective, which means that genomic regions from one species are mapped to only one region in the other species. Therefore, the reverse mapping of a mapped nucleotide will return the original nucleotide. The tool ignores mapped regions that span multiple blocks of different chains or map to multiple chromosomes. In order to unambiguously map features from the human genome to the mouse genome and vice versa, we used a reciprocal-best chain, as the human-mouse alignment provided by UCSC (based on BLASTZ pairwise alignment) is not symmetric. The reciprocal-best chain was created using a netting procedure and chaining only the first layer to make the original human-mouse alignment reciprocal. The reciprocal chain files can be downloaded at http://bx.mathcs.emory.edu/~odenas/mapper_comparisons/UCSC/UCSC_reciprocal.

Once we identified the orthologous regions of TF binding peaks in one species, we overlapped the orthologous regions with the binding peaks of the same TF in the other species to see if the orthologous region was actually bound by the same TF (i.e., conserved occupancy). For this, we used the intersectBed tool (Quinlan and Hall 2010) and required that at least half the peak region overlapped the mapped orthologous region. Using these occupancy-conserved regions, we determined whether or not the TE-derived TF binding peaks were conserved. If occupancy-conserved regions overlapped the same TE (i.e., peak center overlapping the TE) in both species, we called the binding event a shared, TE-derived TF binding event. We also identified several cases in which the occupancy was conserved, but the TE annotation was different or missing. Additionally, there were many TE-derived TF binding events that were not conserved (defined by the lack of binding in the other species, or being unmappable).

### Sequence identity of TE-derived TF binding events

To determine the sequence identity of occupancy-conserved TE-derived TF binding peaks, we used the chain files (described above) to identify alignable regions of the genomes. Using this, we measured the sequence identity between the pairs of alignable regions. For comparison, we randomly picked 1000 TEs from RepeatMasker-annotated TEs (Smit et al. 1996-2010), in the human (hg19) and mouse (mm9) genome assemblies.

### Sequence conservation of TE-derived TF binding peaks (phastCons)

We used phastCons scores (Siepel et al. 2005) to examine the sequence constraint on the TEs that had shared and species-specific TF binding events derived from it. For conserved binding events (i.e., cases in which the binding event is encoded on TE subfamilies either annotated as the same in both species, or where the annotation is not the same in both species), we used the TE sequences that the TF binding peak was derived from. For non-TE-based conserved binding events, representing unmappable (i.e., no orthologous region identifiable) and unoccupied (i.e., the orthologous region lacked TF binding) binding events, we used the peaks regions. We downloaded the vertebrate phastCons data from the UCSC Genome Browser for human (phastCons46way) and mouse (phastCons30way). We defined 6-kb regions, centered on each TE or peak, and profiled the phastCons score over the region. We then averaged the scores across various genomic regions in each category of occupancy conservation.

### De novo prediction of TF binding motifs

Because several TFs being analyzed here do not have known TF binding motifs, we used the HOMER software (Heinz et al. 2010) for de novo binding site prediction from TF binding peaks. We ran HOMER on unique ChIP-seq binding regions (we excluded TEs and repetitive sequences) for each TF in each species. Once we identified binding motifs for each TF in the human and mouse genome, we took advantage of the orthologous data to select binding motifs. From the top five ranked HOMER motifs, we selected the highest ranked motif that was the same between the two species (Supplemental Table 5). With this criterion, we were able to identify binding motifs for 19 of the 26 TFs. Since we trained the de novo motif predictor on sequences lacking repetitive sequences, we tested the prediction on TE sequences that we knew had binding peaks overlapping it; we found that as expected, the training and testing data sets were both enriched for the motif (Fig. 5A).

### Enrichment of TF binding motifs in TEs

To measure the enrichment of TF binding motifs in TEs, we first scanned the human and mouse genomes with the de novo predicted motifs, using FIMO (Grant et al. 2011). We then overlapped the TF binding motifs in the genome with RepeatMasker-annotated TE fragments (Smit et al. 1996-2010), using the intersectBed tool (Quinlan and Hall 2010) to identify TE-derived TF binding sites. Enrichment of TF binding motifs in TEs was defined as

$$LOR_{i,j} = \log_2\left(\frac{\text{Number of TF '}i\text{'binding motifs in TE subfamily '}j\text{'}/\text{Length of TE subfamily '}j\text{'(kb)}}{\text{Number of TF '}i\text{'binding motifs in the genome}/\text{Genome length(kb)}}\right)$$

To compare the enrichment of motifs in TEs with the enrichment of peaks in TEs, we required (as mentioned earlier) that an enriched TE subfamily had greater than 30 genomic copies.

### Sequence alignment of TE subfamilies enriched for TF peaks and motifs

To evaluate the motif conservation in TEs, we first downloaded the sequences of the TE subfamily and the subfamily's consensus from Repbase (Jurka et al. 2005). We then chose TE subfamilies that were enriched for TF binding motifs and TF binding peaks, and aligned the sequences of the TE fragments using Clustal Omega (Sievers et al. 2011). We further removed columns in the alignment that contributed gaps in the consensus sequence.

### Determining the epigenetic state of TEs

To evaluate the epigenetic profiles of TEs that encode TF binding events, we used histone data sets for the human-mouse pairs of lymphoblast and leukemia cell lines, which are available

from the Data Coordination Center (DCC) of the ENCODE and Mouse ENCODE Consortia. The data sets represented assays for six histone marks: H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. To profile the histone signal on TEs, or peaks, we chose a 10-kb region, centered region of interest (TE or TF binding peak) and calculated the normalized read density in 50-bp bins. To overlap the epigenetic data sets with the regions of interest, we used the intersectBed tool (with default parameters) from the BEDTools package (Quinlan and Hall 2010). We averaged the signal from all regions and the replicates and plotted the enrichment of the histone signal over the input data set.

### Assaying the methylation state of the TEs

To profile the DNA methylation patterns in the human-mouse pair of lymphoblast and leukemia cell lines, we performed two complementary assays, MeDIP-seq and MRE-seq, on the cells, as described earlier (Maunakea et al. 2010). We aligned the sequencing reads from these assays back to the human (hg19) and mouse (mm9) genome assemblies, using BWA (Li and Durbin 2009). We overlapped the signal from each of these assays with the 10-kb regions of interest (as described above) and estimated the average assay signal.

## Data access

All data from this study have been submitted to the Mouse ENCODE Data Coordination Center (DCC; http://www.mouseencode.org). All data sets used in this study along with their NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) accession numbers or ENCODE DCC data set identifiers are listed in Supplemental Table 6.

## Acknowledgments

We thank collaborators of the ENCODE and Mouse ENCODE Consortia who have generated and processed data that were used in this project. V.S. is supported in part by the Cancer Biology Pathway, Washington University. M.P.S. is supported by Mouse ENCODE Consortium grant 3RC2HG005602. T.W. is supported by the Basil O'Connor Starter Scholar Research Award 5-FY10-491 from the March of Dimes Foundation, the Edward Jr. Mallinckrodt Foundation, American Cancer Society grant RSG-14-049-01-DMC, and NIH grants 5U01ES017154, R01HG007354, R01HG007175, and R01ES024992.

## References

Adams MD. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.
Ames BN, Shigenaga MK, Hagen TM. 1993. Oxidants, antioxidants, and the degenerative diseases of aging. *Proc Natl Acad Sci* **90:** 7915–7922.
Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87–90.
Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16:** 6–21.
Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* **19:** 607–612.
Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18:** 1752–1762.
Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165:** 349–357.
Cheng Y, Ma Z, Kim BH, Cayting P, Boyle AP, Wu W, Sundaram V, Xing X, Li J, Euskirchen G, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* doi: 10.1038/nature13985.
Chuong EB, Rumi MA, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45:** 325–329.
Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448:** 105–114.
de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7:** e1002384.
de Souza FSJ, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? *Mol Biol Evol* **30:** 1239–1251.
Demetrius L. 2005. Of mice and men. *EMBO Rep* **6:** 39–44.
Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603.
Dunn CA, Medstrand P, Mager DL. 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human β1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci* **100:** 12841–12846.
Ekram MB, Kang K, Kim H, Kim J. 2012. Retrotransposons as a major source of epigenetic variations in the mammalian genome. *Epigenetics* **7:** 370–382.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9:** 397–405.
Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13:** 283–296.
Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Jacques PÉ, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* **9:** e1003504.
Johnson DS, Mortazavi A, Myers RM. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **80:** 1497–1502.
Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19:** 68–72.
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462–467.
Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470.
Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328:** 232–235.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996–1006.
Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26:** 1351–1359.
Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42:** 631–634.
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22:** 1813–1831.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.
Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43:** 1154–1159.
Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253–257.
McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci* **36:** 344–355.
McClintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21:** 197–216.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.
Medstrand P, Landry JR, Mager DL. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **276:** 1896–1903.
Merkenschlager M, Odom DT. 2013. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152:** 1285–1297.
Mestas J, Hughes CCW. 2004. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172:** 2731–2738.
Morgan HD, Sutherland HG, Martin DI, Whitelaw E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23:** 314–318.
The Mouse ENCODE Consortium, Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* doi: 10.1038/nature13992.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H, de Vries I, Ding L, Huebner N, Mann M, et al. 2011. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* **6:** e19470.
Oliver KR, Greene WK. 2011. Mobile DNA and the TE-thrust hypothesis: supporting evidence from the primates. *Mob DNA* **2:** 8.
Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137:** 1194–1211.
Pi W, Zhu X, Wu M, Wang Y, Fulzele S, Eroglu A, Ling J, Tuan D. 2010. Long-range function of an intergenic retrotransposon. *Proc Natl Acad Sci* **107:** 12992–12997.
Polavarapu N, Mariño-Ramírez L, Landsman D, McDonald JF, Jordan IK. 2008. Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics* **9:** 226.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.
Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager DL. 2012a. Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biol* **13:** R89.
Rebollo R, Romanish MT, Mager DL. 2012b. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46:** 21–42.
Roman AC, Benitez DA, Carvajal-Gonzalez JM, Fernandez-Salguero PM. 2008. Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression *in vivo*. *Proc Natl Acad Sci* **105:** 1632–1637.
Samuelson LC, Wiebauer K, Snow CM, Meisler MH. 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol Cell Biol* **10:** 2513–2520.
Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci* **105:** 4220–4225.
Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148:** 335–348.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7:** 539.
Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8:** 272–285.
Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.
Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, Ligon KL, Hirst M, Marra MA, Costello JF, et al. 2013. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res* **23:** 1541–1553.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457:** 854–858.
Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci* **104:** 18613–18618.
Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451:** 796–801.
Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. 2010. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res* **20:** 804–815.
Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* **45:** 836–841.
Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, Cheng JB, Li D, Stevens M, Lee HJ, et al. 2013. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* **23:** 1522–1540.

*Received October 27, 2013; accepted in revised form April 18, 2014.*