Differential selective regimes identify categories of transposable elements with regulatory function

Yari Cerruti[1], Daniela Fusco[1], Mahdis Jahanbin[1], Davide Marnetto[1,*], Paolo Provero[1,2,*]

[1]Department of Neurosciences "Rita Levi Montalcini," University of Turin, Turin, Italy
[2]Center for Omics Sciences, IRCCS Ospedale San Raffaele, Milan, Italy

*corresponding authors: davide.marnetto@unito.it, paolo.provero@unito.it

Abstract

Transposable elements (TEs) are powerful drivers of genome evolution, in part through their ability to rapidly rewire the host regulatory network by creating transcription factor binding sites that can potentially be turned to the host's advantage in a process called exaptation. We use methods of comparative genomics and population genetics to identify the TE categories that have most contributed to this process. For each TE category, we compare the copies overlapping regulatory elements to all other copies in terms of measures of selection. Using phyloP scores, we find a large number of TE categories showing evidence of conservation specifically when overlapping regulatory elements. The TE categories identified by such analysis are involved in the regulation of gene expression, in the heritability of complex traits, and in specific development-related functions. The same analysis conducted using Tajima's D, thus exploring more recent selection, shows similar results, although with much lower statistical power. These results suggest that TE categories are heterogeneous in their ability to rewire the human regulatory network in a way that is useful and thus exapted by the host genome; and that the analysis of selective pressure is a powerful tool to identify the TE categories playing the most important roles in gene regulation.

Introduction

Transposable elements (TEs) have been shown to provide raw material for the rapid evolution of genomes, and specifically of gene regulation, by creating quickly dispersing genetic elements potentially exploitable by the host as binding sites for DNA-binding factors involved in regulating transcription and three-dimensional genome conformation (see [1–3] for recent reviews), in a phenomenon called TE exaptation. Human regulatory network rewiring by TE exaptation has been recently shown to be relevant to a variety of biological processes, including in particular

immune response [4–7], germ cell development [8,9], longevity [10], pluripotency [11], and 3D genome conformation [12].

TE exaptation can be used by the host not only to rewire the regulatory network by providing new regulatory targets of transcription factors (TFs), but also to increase the robustness of the existing regulatory network by providing additional, redundant binding sites near existing ones [13–15]. This fact suggests that methods based on the analysis of selective pressure on TE sequences might usefully complement those based on genetic perturbations, as the latter might be prone to false negatives when the role of the exapted TEs is to provide redundancy. Analytical approaches based on selective pressure have the added advantage of bypassing the need to identify the tissue, cell type, or biological condition in which the role of the exapted TEs is exerted.

In this work we propose the use of selection assessment as a tool to identify TE categories associated with significant rewiring of the human regulatory network. We reasoned that a TE category that has been significantly exapted to provide regulatory sequence to the host should show a stronger deviation from neutrality in those copies that overlap a regulatory region (referred to as “regulatory TEs" in the following) compared with the copies that do not (“non-regulatory TEs”). We use two assessments of deviation from neutrality: the phyloP score, which by measuring inter-species conservation assesses selective pressures acting on a longer time scale; and Tajima’s D, based on the allele frequency spectrum in contemporary human populations, and suitable for the investigation of recent selection. By comparing, for each TE category, regulatory vs non-regulatory copies, we identify those classes whose exaptation as regulatory elements has contributed the most to the evolutionary rewiring of the human regulatory network.

Results

Differences in long-term selection regime between regulatory and non-regulatory TEs

The exaptation of TEs to rewire the human regulatory network has been taking place along the whole evolutionary history of mammals [1–3,16]. To identify TE categories which have undergone selection on time scales comparable to that of mammalian evolution, we analyzed sequence conservation in eutherian mammals, as measured by the phyloP score derived from the recent multiple alignment of 241 species obtained by the Zoonomia project [17].

We considered a total of six types of regulatory elements: the 5 types of candidate cis-regulatory elements (cCREs) defined by ENCODE [18] (TSS-overlapping with promoter-like signatures (PLS); TSS-proximal with enhancer-like signatures (pELS); TSS-distal with enhancer-like signatures (dELS); not TSS-overlapping and with high DNase and H3K4me3 signals only

(DNase–H3K4me3); not TSS-overlapping and with high DNase and CTCF signals only (CTCF-only)); and the set of all sequences labeled as strong enhancers in at least one ENCODE cell type by ChromHMM [19].

We use the term “TE category” to indicate an item belonging to any level of the RepeatMasker hierarchical classification of TEs (“repName”, “repFamily”, or “repClass”). For each TE category and each regulatory element type we divided the TE copies into those that do and those that do not overlap a regulatory element, and proceeded to evaluate the corresponding difference in phyloP score (Δphylop).

To provide a statistical assessment of these differences we divided the genome into 59 segments of roughly equal size; the smaller chromosomes (chr19 - chr22) form their own segment, while the larger ones were divided into 2 to 5 segments. PhyloP values were then averaged across the base pairs covered by all the relevant TE copies lying into each segment and the 59 values thus obtained were used as replicates for statistical analysis. This approach provides a set of independent observations with comparable sequence length for each TE category, accounting for possible lack of independence of nearby TE copies. As controls, we also generated, for each TE category, a set of randomized copies by circular randomization of the genomic regions covered by the real copies.

The total number of TE categories analyzed and the number of those reaching statistical significance (paired Wilcoxon test across the 59 segments; FDR < 0.05 computed considering all regulatory types together) is shown in Table 1. For the vast majority of significant classes (but not all) Δphylop has the expected sign (higher phyloP for regulatory TEs). Fig. 1 shows the comparison of phyloP values between regulatory and non-regulatory TEs for the six types of regulatory regions.

	total	significant	P reg > P nonreg
ChromHMM	169	105	103
dELS	185	100	93
pELS	36	28	28
DNase-H3K4me3	17	4	1
CTCF-only	31	3	1
PLS	8	2	1

Tab. 1. Number of TE categories tested for significant differences in phyloP scores between regulatory and non-regulatory TEs for each type of regulatory element, and number of classes

with significant difference (paired Wilcoxon test on 59 genomic segments, FDR < 0.05). The last column indicates the number of significant cases in which the difference goes in the expected direction, namely higher phyloP scores for regulatory TEs.

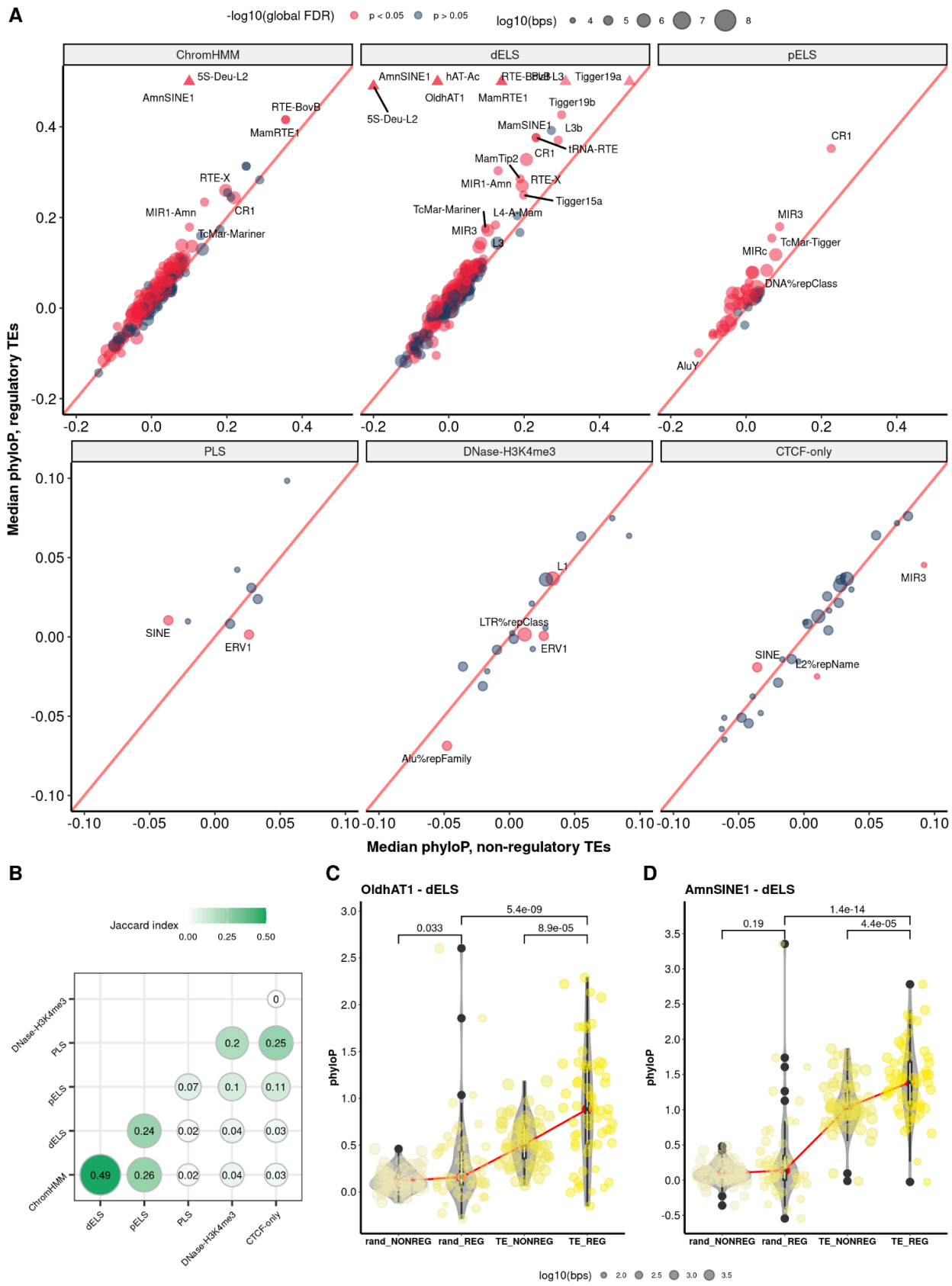
A
-log10(global FDR)
p < 0.05
p > 0.05
log10(bps)
ChromHMM
dELS
pELS
PLS
DNase-H3K4me3
CTCF-only
Median phyloP, regulatory TEs
Median phyloP, non-regulatory TEs
5S-Deu-L2
AmnSINE1
RTE-BovB
MamRTE1
RTE-X
MIR1-Amn
CR1
TcMar-Mariner
hAT-Ac
OldhAT1
Tigger19a
Tigger19b
MamSINE1
L3b
tRNA-RTE
MamTip2
Tigger15a
L4-A-Mam
MIR3
L3
MIRc
TcMar-Tigger
DNA%repClass
AluY
SINE
ERV1
L1
LTR%repClass
Alu%repFamily
L2%repName
B
Jaccard index
DNase-H3K4me3
PLS
pELS
dELS
ChromHMM
C
OldhAT1 - dELS
phyloP
rand_NONREG
rand_REG
TE_NONREG
TE_REG
D
AmnSINE1 - dELS
log10(bps)

Fig 1: A: Comparison of the phyloP score of regulatory vs. non-regulatory TEs for each TE category in six types of regulatory elements. The dots represent the median value of the phyloP score across the 59 genomic segments, their size reflecting the number of base pairs on which the score for regulatory TEs was computed, and the color the statistical significance. Note that the scale of the two axes is different in the two rows, the regulatory types in the top row showing generally higher phyloP scores than those in the bottom row. Triangles indicate out-of-scale significant TEs. Double labels assigned to single points indicate that lower-hierarchy categories coincide with higher categories in the RepeatMasker classification of TEs (i.e. 5S-Deu-L2 ≅ AmnSINE1). B: Jaccard index of the overlap between TE categories with significant ∆phyloP in the six types of regulatory regions considered. C,D: Violin plot of the phyloP scores for regulatory and non-regulatory TEs (TE_REG, TE_NONREG) and their randomized counterparts (rand_REG, rand_NONREG) for oldhAT1 (C) and AmnSINE1 (D) in dELS.

The top 10 significant categories by ∆phyloP are shown in Table 2. Complete results are available in Suppl. Tab. 1).

type	TE class	level	phyloP reg	phyloP non-reg	∆phyloP	FDR
dELS	OldhAT1	repName	0.8836	0.5065	0.3771	6.17e-05
dELS	AmnSINE1	repName	1.393	1.032	0.3611	7.154e-06
dELS	MamRTE1	repName	0.5206	0.3422	0.1784	4.233e-07
dELS	MamSINE1	repName	0.3762	0.2318	0.1445	1.115e-05
dELS	Tigger19a	repName	0.5757	0.4433	0.1324	0.003289
dELS	Tigger19b	repName	0.4266	0.2997	0.1269	0.001762
pELS	CR1	repFamily	0.3522	0.226	0.1261	0.0003255
dELS	MamTip2	repName	0.2852	0.1891	0.09607	3.803e-05
pELS	MIR3	repName	0.1795	0.08915	0.09038	5.29e-07
pELS	MIRc	repName	0.1544	0.06836	0.086	6.519e-06

Tab. 2. TE categories showing significant (FDR < 0.05) difference in phyloP score between regulatory and non-regulatory TEs. The top 10 classes in terms of ∆phyloP are shown. The RepeatMasker hierarchical level of each TE category is indicated in the third column. Some redundancy due to coinciding categories at different RepeatMasker hierarchical levels has been

removed, and TE categories appearing in more than one regulatory element type are shown only once.

Widespread exaptation of AmnSINE1 elements has been associated with the emergence of mammalian-specific traits, specifically in brain development [20,21]. Both AmnSINE1 and MamSINE1 were found enriched in active regulatory regions in human development [22], and to be involved in the rewiring of the NF-κB regulome [7], while MIR3 and MIRc were shown to contribute to gene regulation in decidual stromal cells in mammals [23] and the regulome of estrogen receptor alpha [24].

Among other significant results not included in the table (see Suppl. Tab. 1) we noted the presence of 11 out of the 14 TE subfamilies shown in [6] to be epigenetically activated upon bacterial infection (MER41B, MER44B, MLT1F/H/I/K/L, THE1B/C, and Tigger3a/b - all of them being found under significant selection when overlapping dELS and/or chromHMM regulatory regions). Some significant TE categories are also likely to be involved in immune response through the rewiring of the NF-κB regulome [7], including again MER41B and MER44B. The ERV1 family, which was shown to be associated with TP53 binding sites [25], was found to be under selection when overlapping ChromHMM and dELS regulatory regions.

We also analyzed the phyloP scores of circularly randomized TEs, divided into regulatory and non-regulatory. To compare real with randomized TEs we used linear regression to model the phyloP score as a function of two binary variables and their interaction: one of the variables tags regulatory vs non-regulatory status, and the other real vs randomized status. In such a model, a significant interaction term indicates that the effect of regulatory status on selection is different between TEs and the rest of the genome, represented by the randomized TE copies.

PhyloP scores showed for most TE categories/regulatory type comparisons (304 out of 446) a significant (FDR < 0.05) interaction effect between being regulatory and being a TE. In all such cases, the interaction term was negative, implying that the effect of being regulatory on the phyloP score is less pronounced for TEs than for the rest of the genome. However some of the top significant results shown in Tab. 2, including those shown in Fig.1C and D have non-significant interaction terms: For these TE categories selection on regulatory copies, measured by the difference in their phyloP score compared with non-regulatory copies, is comparable to selection acting on non-TE-derived regulatory regions.

Functional roles of regulatory TEs

If ΔphyloP is a measure of the role played by each TE category in shaping the human regulatory network, we expect it to be associated with independent assessments of functional relevance. We analyzed three such functionality measures, namely the effect of TE categories on gene expression (evaluated from the number of genes with eQTLs residing inside the regulatory copies of the TE); the relevance of TE categories to the heritability of complex traits; and the functional characterization of the closest genes, considered as putative TE regulatory targets.

To obtain independent assessments of these functional measures, suitable for statistical analysis, these analyses were conducted on dELS regulatory regions only, and limited to the finest layer of TE categories, that we refer to as “TE species”, namely those belonging to the “repName” level of the RepeatMasker hierarchy.

∆phyloP is positively correlated with influence on complex traits and gene expression

A standard method to quantify the influence of a set of genomic regions on a complex trait is to evaluate the fraction of this trait's heritability that is explained by such genomic regions, and specifically whether the heritability portion is enriched with respect to the genomic fraction covered. Such heritability enrichment for several TE categories and traits was evaluated in [26], accounting for the expected depletion of functional annotations overlapping with TEs, and subsequently meta-analyzing the results for different traits to give a single estimate for each TE category. To test the hypothesis that such functionally-weighted enrichment should come mostly from regulatory TEs, we compared it with our selection-based measure of functionality, i.e. ∆phyloP, finding indeed a positive and significant correlation (Spearman’s rho=0.18, p = 0.031), see Fig. 2A. Reassuringly, when limited to the TE species with a significant ∆phyloP, such correlation is markedly stronger (Spearman’s rho=0.509, p = 1.3e-05).

The proposed molecular function through which these TE categories exert their influence on complex traits is by regulating the expression of nearby genes. We therefore asked if ∆phyloP was also correlated with a measure of regulatory activity. For each TE species we intersected all copies annotated as regulatory with expression quantitative loci (eQTL) from GTEx, thus identifying the associated eGenes. Following the same rationale as above, we computed an enrichment against the expected eGene count, defined as a ratio over the count of eGenes associated with randomized TEs generated in the first section. Again the correlation with ∆phyloP is strongly positive both when considering all TE species (Spearman rho =0.254 , p =0.00139) and only those with a significant ∆phyloP (Spearman rho = 0.564, p = 2.3e-07), see Fig. 2B. Complete data are available in Suppl. Table 2. These independent lines of evidence confirm that ∆phyloP is a strong predictor of the functional activity of individual TE species.

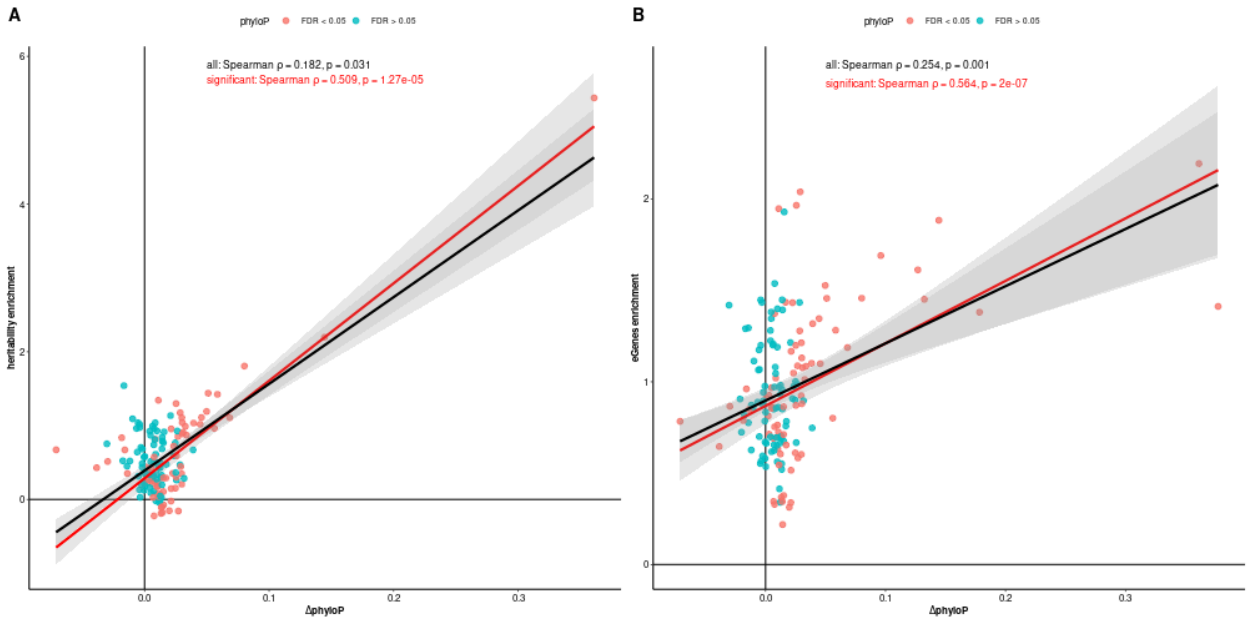

Fig. 2: A: correlation between ΔphyloP and enrichment of heritability of complex traits for TE species estimated by Hormozdiari et al. [26] *with significant (red) and non-significant (blue) ΔPhyloP in dELS regulatory regions. B: correlation between ΔphyloP and enrichment in number of eGenes associated with regulatory TEs.*

ΔphyloP is associated with enrichment of putative targets in specific development-related gene functions

For each TE species, we considered as putative regulatory targets the closest protein-coding gene to each regulatory copy. We then analyzed the Gene Ontology (biological process, GOBP) enrichment of each list of putative targets, and for each GOBP term enriched for at least 10 TE species we compared the ΔphyloP of enriched vs. non-enriched TE species. Complete results are in Suppl. Tab. 3 In this way we identified the biological processes associated with the regulatory targets of TE species subjected to strong selective pressure.

Out of 269 GOBP terms that could be tested, 6 (Tab. 3) showed a significant difference in ΔphyloP between enriched and non-enriched TE species (FDR < 0.05, Mann-Whitney U test). In all 6 cases the enriched TE species showed higher ΔphyloP than non-enriched species. Several TE species with high ΔphyloP were associated to all 6 GOBP terms, while some were specific of a few or a single term. The relationship between significant GOBP terms and associated TE species is represented in Fig. 3.

GOBP	⊿PhyloP enriched	⊿PhyloP other	FDR	TEs
muscle tissue development	0.03269	0.01132	0.03836	L2 MamSINE1 MamTip2 MER5B MLT1J MLT1L OldhAT1
blood circulation	0.04977	0.01149	0.03836	MamSINE1 MamTip2 MER102b MER5B OldhAT1
artery development	0.04465	0.01162	0.03836	MLT1H OldhAT1
forebrain development	0.03005	0.01127	0.03836	L2 MamSINE1 MamTip2 MER5B MLT1H MLT1K OldhAT1
camera-type eye morphogenesis	0.03461	0.01162	0.03996	L2 MamSINE1 MLT1H
kidney development	0.03021	0.01132	0.0412	L2 MamSINE1 MamTip2 MER33 MER5B MLT1H MLT1H1 MLT1J MLT1K OldhAT1 Tigger19a

Tab. 3. Gene Ontology Biological Process (GOBP) terms associated with significantly higher ⊿phyloP of the associated TE species. The last column shows the TE species whose regulatory targets are enriched in each GOBP term. Several TE species (AmnSINE1, L2a/b/c, L3, L3b, MamRTE1, MIR3, and MIRc) are associated with all these terms, and are not included in the last column.

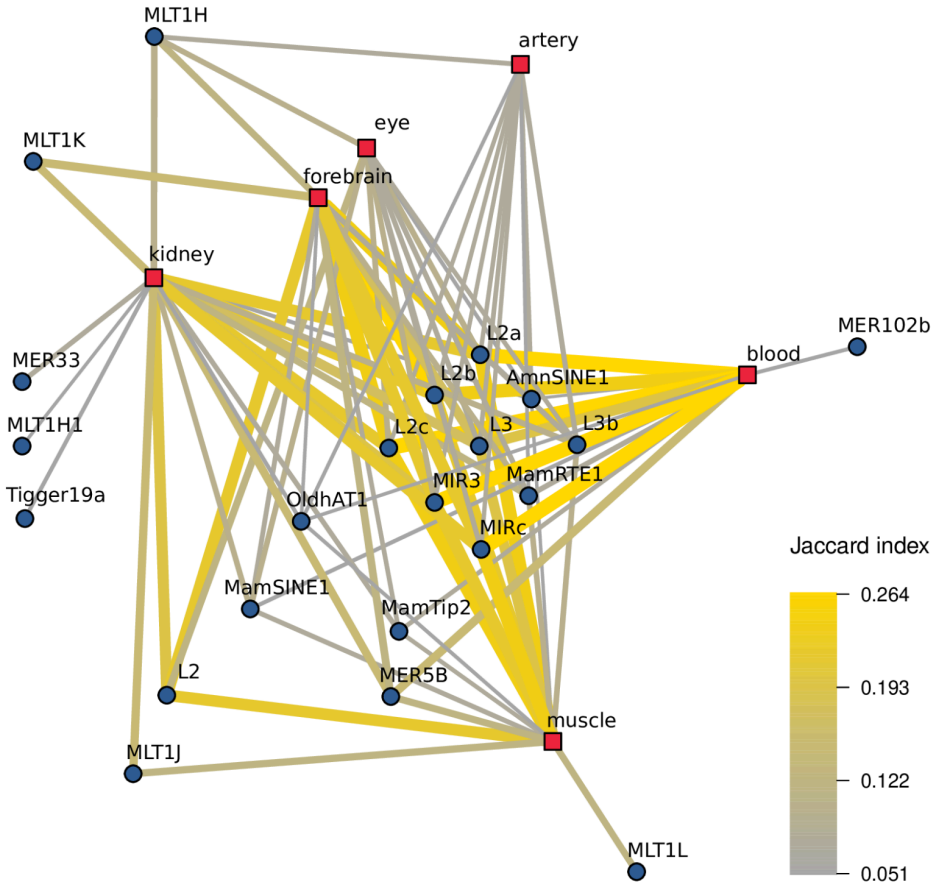

Fig. 3: Network of TE-GOBP associations. Blue nodes represent TE species whose targets are enriched in at least one of the 6 significant GOBP terms, while red nodes are the GOBP terms. Edges represent significant enrichments (exact Fisher test FDR<0.01). The edge color and thickness is scaled for Jaccard index between the genes associated to the TE species and the GOBP term: the lowest index included here is 0.051 (between AmnSINE1 and "blood circulation" GOBP term), while the highest is 0.264 (between L2c and "blood circulation"). Jaccard indices are computed over all genes that are associated to at least one of the 6 GOBP terms and one of the TE species shown. Artery: "artery development"; muscle: "muscle tissue development"; eye: "camera-type eye morphogenesis"; forebrain: "forebrain development"; blood: "blood circulation"; kidney: "kidney development".

Differences in human-specific selection regime between regulatory and non-regulatory TEs

To investigate whether the differences in selection regimes across the evolutionary history of mammals are still detectable within our species, we adopted Tajima's D [27] as a measure of non-neutrality. Tajima's D relies on the allele frequency spectrum of genetic variants overlapping a genomic region, considering an approximately uniform population at demographic equilibrium, which we epitomized by selecting European individuals from The 1000 Genomes project [28].

The calculation and interpretation of Tajima's D computed on TEs required some specific precautions (see Methods for details): First we had to discard repetitive sequences with extremely low mappability, including instead variants in flanking regions to enhance the signals of recent selection; then we discarded altogether the TE categories with an extremely low number of variants overlapping the regulatory copies. Finally, to avoid a large disparity in variant count between regulatory and non-regulatory TEs, we estimated the non-regulatory D value by subsampling the TE copies to match the regulatory ones. For each of the 59 genomic segments defined for the phyloP analysis, the non-regulatory D was estimated then as the average across 10 random subsampling iterations. Complete results are found in Suppl. Tab. 4.

As shown in Fig. 4, Tajima's D differences (ΔD hereafter) were significant for only 4 TE categories/regulatory type pairs out of 260 tested, namely: SINE and Alu in pELS; and L1 and LINE in chromHMM. Notably these are all large TE classes, suggesting that only by considering a large number of variants the signal captured by Tajima's D would be visible as non-neutrality for exapted TEs. Although not significant for finer TE categories such as families and species, ΔD tends to be negative overall (Fig. 4), which is expected if regulatory TEs are under stronger selective regimes than their non-regulatory counterparts. These results suggest that the comparison of Tajima's D values between regulatory and regulatory TEs lacks statistical power to detect selection in all but the largest TE classes.

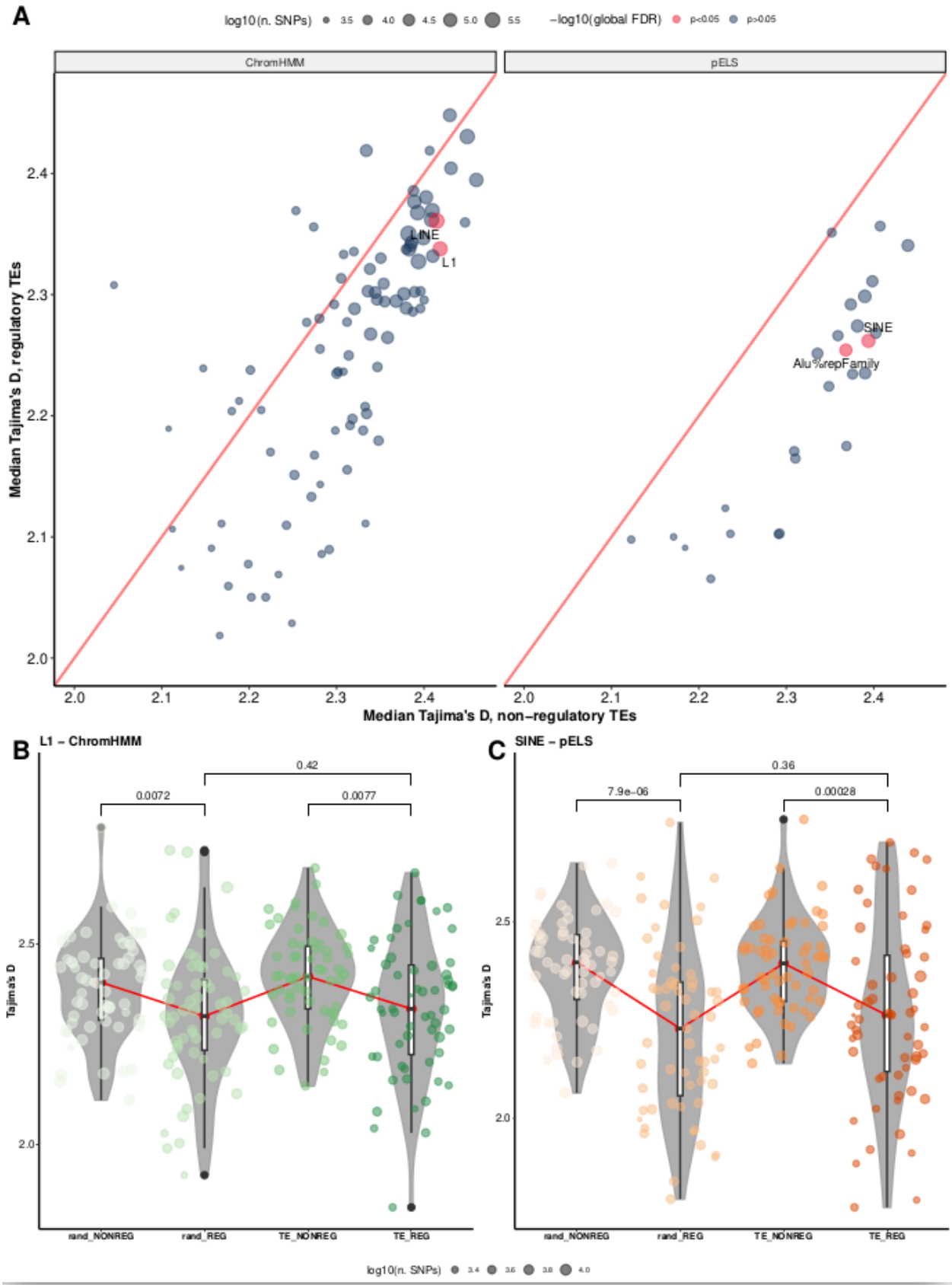
A
log10(n. SNPs)
3.5
4.0
4.5
5.0
5.5
−log10(global FDR)
p<0.05
p>0.05
ChromHMM
pELS
LINE
L1
SINE
Alu%repFamily
Median Tajima's D, regulatory TEs
Median Tajima's D, non−regulatory TEs
2.0
2.1
2.2
2.3
2.4
B
L1 − ChromHMM
0.42
0.0072
0.0077
Tajima's D
2.5
2.0
rand_NONREG
rand_REG
TE_NONREG
TE_REG
C
SINE − pELS
0.36
7.9e−06
0.00028
log10(n. SNPs)
3.4
3.6
3.8
4.0

Fig. 4. A: Comparison of the Tajima's D values of regulatory vs. non-regulatory TEs for each TE category in two types of regulatory elements. Other regulatory element types did not reveal any significant TE category. The dots represent the median value of the Tajima's D across the 59 genomic segments, their size reflecting the number of base pairs on which the D value for regulatory TEs was computed, and the color the statistical significance. B,C: Violin plot of the D values for regulatory and non-regulatory TEs (TE_REG, TE_NONREG) and their randomized counterparts (rand_REG, rand_NONREG) for L1 in ChromHMM (B) and SINE in pELS (C).

Discussion

We have shown that comparing the selective regimes acting on regulatory and non-regulatory copies of each TE category allows the identification of those categories that have undergone the most relevant exaptation by the host genome, contributing to the rewiring of its regulatory network. The use of the phyloP score as a measure of long-term selective regime allowed the identification of many such exapted categories, some of which had been previously characterized in the literature. Several independent measures of functional relevance significantly correlated with ΔPhyloP, and such correlations tended to be stronger when only statistically significant ΔPhyloP values were considered, providing an external validation to both the choice of ΔPhyloP as a measure of selective regime and to the statistical method used to identify significant TE categories.

A similar analysis using Tajima's D as a measure of ongoing selection gave results in general agreement with those based on phyloP, but the number of exapted TE categories found was markedly smaller. Future research using different measures of ongoing selection might clarify whether this is due to lower selection acting on regulatory TEs on shorter evolutionary time scales, or simply to lack of power of the statistical procedure.

The main strengths of the method we propose are (a) the fact that selection is likely to act also on regulatory TEs whose function is to provide redundancy to existing regulatory networks, and could thus escape detection by methods based on genetic perturbations; and (b) the fact that methods based on selection regimes do not require the identification of the specific tissue, cell type, or biological process in which the TEs actually perform their regulatory function. It must be noted, however, that our use of ENCODE-based cCREs still implies some degree of tissue and cell-type dependence, since cCRES were derived from assays run on a very large, but not exhaustive, set of biological samples.

Such near-independence from tissue and cell type can also be seen as a limitation of our approach, which does not provide any information on the biology associated with the identified TEs. This can be investigated *a posteriori,* for example with a functional enrichment analysis of the putative regulatory targets of the identified TE categories. We found that high levels of selection are associated with specific developmental functions, including kidney, forebrain, eye, muscle, and artery development.

In summary, our results provide further support to the relevance of TE exaptation in the rewiring of the human regulatory network and introduces an evolutionary approach to the identification of the TE categories that played the most important role in this process.

Methods

Regulatory elements and TE data preprocessing

Bed files containing genome-wide coordinates of repetitive elements were retrieved from the latest UCSC Repeatmasker [29] GRCh38 annotation. TE categories mapping on chromosomes X, Y and MT or with cumulative genome-wide length smaller than 100 kbps were excluded from further analysis. TEs mapping within the Human Leukocyte Antigen (HLA) region of chromosome 6 were also excluded, owing to the extreme variability of the HLA region which could skew SNP-based selection statistics.

In order to quantify the effect of selection separately on TE categories irrespectively of their hierarchical organization contained in the Repeatmasker classification, repeats classified as repName (the lowest level, high resolution hierarchy with specific TE classification), repFamily (intermediate level hierarchy and resolution) and repClass (highest level hierarchy, low resolution) were investigated separately. Approximately 15 million TEs classified in 876 unique categories were analyzed.

Bed files containing genome-wide annotated functional regions were retrieved from the UCSC Genome Browser tracks: Encyclopedia of DNA Elements (ENCODE) Chromatin State Segmentation by HMM (Broad ChromHMM) [19] and Candidate Cis-Regulatory Elements (cCRE) [18]. The ChromHMM bed files from all 9 ChromHMM human cell lines were lifted from GRCh37 to GRCh38 with the UCSC LiftOver tool [30], and elements bearing strong enhancer signatures (annotated as “4_Strong_Enhancer” and “”5_Strong_Enhancer”) in any of the 9 cell lines were retained.

cCREs are classified in five types that were analyzed separately: *dELS* and *pELS* (enhancer-like signatures exhibiting high DNase and H3K27ac signal and low H3K4me3 signal, further split into *distal* and *proximal* according to the distance to their closest GENCODE-annotated TSS, with a 2000kbps cutoff)*, PLS* (promoter-like signatures with high DNase and H3K4me3 signals and falling within 200bps of a TSS)*, CTCF-only* (putative CTCF binding sites with high DNase and CTCF signals and low H3K4me3 and H3K27ac signals)*, DNase-H3K4me3* (high H3K4me3 but low H3K27ac signals and not falling within 200 bp of a TSS).

TE sets definitions

The overlaps between TEs and regulatory elements were computed with Bioconductor's Genomic Ranges R package. TE instances were annotated as *regulatory* when at least one base of a regulatory element was found to overlap a TE, or *non-regulatory* otherwise.

In order to provide a negative control for the following statistical testing, a circular randomization of TEs which survived the preprocessing was performed with the BioConductor RegioneR R package [31]. New TE positions were generated by a random displacement on each chromosome of the original TE positions in such a way that inter-repeat distances and chromosome annotation were preserved. Randomized sets were then annotated as *regulatory* or *non-regulatory* following the same procedure used for TEs.

For the Tajima's D analysis, the resulting ranges were filtered with the 1000 Genomes Project Accessibility mask with the "strict" criterion (which allows the removal of regions deemed inaccessible by the sequencing technologies used and covers 76.9% of non-N bases in autosomes) in order to only consider regions with robust base calling.

TE copies were then assigned to 59 genomic segments of approximately equal size (average size 48.7 Mbps, s.d. 4.8 Mbps). Chromosomes from 19 to 22 were considered each an independent segment, while the larger chromosomes were divided into 2 to 5 segments. The final result of the procedure consists in four sets of regions defined in each of 59 genomic segments: real regulatory, real non-regulatory, randomized regulatory, and randomized non-regulatory TEs.

PhyloP computation

Genome-wide phyloP scores were retrieved from the Cactus 241-way UCSC Genome Browser track [17]. In order to obtain phyloP averages separately for each set of regions and each genomic segment, nucleotide-based phyloP scores were summed over all covered bases belonging to each specific set and then divided by the respective number of covered bases.

Tajima's D computation

Tajima's D [27] compares the intra-species variability at genomic regions of interest (observed nucleotide diversity) to a null model that assumes neutral evolution (expected nucleotide diversity). We extracted variation data from 1000 Genomes Project Phase3 [28], retaining common (minor allele frequency > 0.01), biallelic SNPs from European (EUR superpopulation) samples (n=503). Rare variants were excluded because the presence of singletons and rare disease-associated variants could skew Tajima's D's computation. Only individuals of European ancestry were selected since varying population dynamics in different superpopulations could lead to inconsistent D values and nullify any discernible effect.

Since Tajima's D is sensitive to several properties of the SNP set, we applied some further precautions.

First, TE coordinates were extended with 1kbps enlargements upstream and downstream their original boundaries to account for the propagation effect exerted by linkage disequilibrium on selection [32]. Second, the resulting ranges were then filtered with the 1000 Genomes Project Accessibility mask with the "strict" criterion in order to only consider regions with robust SNP calling. Third, given the imbalance in the number of SNPs falling within *regulatory* or *non-regulatory* TEs (for which we observed an approximate ratio of 1/10) we performed 10 random samplings of *non-regulatory* TEs, each time with a sample size matching that of *regulatory* set of the corresponding TE category. These size-matched samples were each partitioned in 59 genomic segments as above, then for each TE category and genomic segment we reported as *non-regulatory* Tajima's D the average across D computed on each of the 10 samples.

Non-selected, neutrally evolving regions in an ideal population are expected to show D values close to 0, whereas sequences subjected to selective pressure should present negative D values, which can be interpreted as purifying or positive selection, thus as non-neutrality. Nevertheless, The human population on which our computation is based inevitably violates Hardy-Weinberg assumptions due to demographic factors, thus absolute Tajima's D values cannot be interpreted. Therefore, we expected non-regulatory TE sets to show D values approximately similar to the D genomic average and greater than regulatory sets of the same TE category.

Statistical analysis

Significant differences in phyloP scores or Tajima's D values between regulatory and non-regulatory TEs were evaluated by means of a paired two-sided Wilcoxon rank sum test on the 59 values associated with the genomic segments. The effect sizes of the comparisons were calculated as differences in Tajima's D or phyloP medians across segments. P-values were corrected with the Benjamini-Hochberg correction and a False Discovery Rate (FDR) < 0.05 was deemed significant.

In order to compare the effect of selection acting on real TEs to that acting on random genomic regions matched in length and genomic distribution, phyloP scores and Tajima's D were modeled as linearly dependent on regulatory status (overlapping vs. non-overlapping a regulatory element), TE status (real TE vs. randomized TE) and their interaction, with the following model:

$$D = \beta_0 + \beta_{TE\ status} \cdot x_{TE\ status} + \beta_{TE\ overlap} \cdot x_{TE\ overlap} + \beta_{interaction} \cdot x_{interaction} + \epsilon$$

Assessment of enrichment in heritability and eGenes

These analyses were run by restricting to TE species ("RepName" in the RepeatMasker classification), and to dELS regulatory elements, to ensure the independence among the TE sets examined. We exploited the estimates of heritability enrichments for 853 TE categories, estimated by Hormozdiari et al. [26] by meta-analyzing 41 independent traits, and extracted enrichments for 141 TE species. We correlated our ΔphyloP with the column "Enr" in Supplementary Data 2 of [26], which represents the expected heritability enrichment computed conditional on a set of 75 functional annotations from the baseline-LD (version 1.1) model [33].

To compute the eGene enrichment, we relied on eQTLs from all 49 tissues from GTEx [34]. By intersecting eQTL positions with each *regulatory* TE, both real and randomized, and then counting associated eGenes, we obtained the total number of eGenes for each TE species. Enrichments in expression genes are computed as the ratio between the number of eGenes associated with regulatory TEs and the same number associated with regulatory randomized TEs of the same species.

Gene Ontology enrichment analysis

Similarly to the section above, the GO enrichment analysis was performed exclusively on TE species ("repName"-annotated) overlapping dELS elements. For each TE species we defined a list of putative regulatory targets consisting of the closest protein coding gene to each regulatory copy. TE copy-gene associations were computed with the ChIPpeakAnno R package and default parameters. Gene Ontology Biological Process enrichments were thus computed separately for each TE species with respect to a background represented by all genes associated to all TEs, with a cutoff on FDR of 0.01 . To identify biological processes associated with strong TE conservation, we considered the GO terms associated with at least 10 TE species, and compared the median phyloP values of the associated species with that of all other TE species using a two-sided Mann-Whitney U-test. Multiple testing was accounted for with the Benjamini-Hochberg procedure and GO terms with FDR < 0.05 were considered significant.

Acknowledgements

Daniela Fusco is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma. Davide Marnetto and Yari Cerruti were supported in part by the University of Turin through Grant For Internationalization 2022. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on July 40 2022 . We are grateful to Mathilde André, Elena Grassi, Giorgia Modenini, and Roberta Zeloni for insightful discussions.

References

1. Senft, A.D., and Macfarlan, T.S. (2021). Transposable elements shape the evolution of mammalian development. Nat. Rev. Genet. 10.1038/s41576-021-00385-1.
2. Sundaram, V., and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *375*, 20190347. 10.1098/rstb.2019.0347.
3. Lawson, H.A., Liang, Y., and Wang, T. (2023). Transposable elements in mammalian chromatin organization. Nat. Rev. Genet. 10.1038/s41576-023-00609-6.
4. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. 10.1126/science.aad5497.
5. Srinivasachar Badarinarayan, S., Shcherbakova, I., Langer, S., Koepke, L., Preising, A., Hotter, D., Kirchhoff, F., Sparrer, K.M.J., Schotta, G., and Sauter, D. (2020). HIV-1 infection activates endogenous retroviral promoters regulating antiviral gene expression. Nucleic Acids Res. *48*, 10890–10908. 10.1093/nar/gkaa832.
6. Bogdan, L., Barreiro, L., and Bourque, G. (2019). Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. BioRxiv. 10.1101/707612.
7. Wang, L., Alizada, A., Rathnakumar, K., Khyzha, N., Taylor, T., Campitelli, L.F., Patel, Z.M., Antounians, L., Hughes, T., Roy, S., et al. (2022). Multi-species analysis of inflammatory response elements reveals ancient and lineage-specific contributions of transposable elements to NF-κB binding. Preprint at bioRxiv, 10.1101/2022.10.25.513724 10.1101/2022.10.25.513724.
8. Xiang, X., Tao, Y., DiRusso, J., Hsu, F.-M., Zhang, J., Xue, Z., Pontis, J., Trono, D., Liu, W., and Clark, A.T. (2022). Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. Nat. Commun. *13*, 463. 10.1038/s41467-022-28105-1.
9. Ito, J., Seita, Y., Kojima, S., Parrish, N.F., Sasaki, K., and Sato, K. (2022). A hominoid-specific endogenous retrovirus may have rewired the gene regulatory network shared between primordial germ cells and naïve pluripotent cells. PLoS Genet. *18*, e1009846. 10.1371/journal.pgen.1009846.
10. Tejada-Martinez, D., Avelar, R.A., Lopes, I., Zhang, B., Novoa, G., de Magalhães, J.P., and Trizzino, M. (2021). Positive selection and enhancer evolution shaped lifespan and body mass in great apes. Mol. Biol. Evol., msab369. 10.1093/molbev/msab369.
11. Hirata, M., Ichiyanagi, T., Katoh, H., Hashimoto, T., Suzuki, H., Nitta, H., Kawase, M., Nakai, R., Imamura, M., and Ichiyanagi, K. (2022). Sequence divergence and retrotransposon insertion underlie interspecific epigenetic differences in primates. Mol. Biol. Evol., msac208. 10.1093/molbev/msac208.
12. Choudhary, M.N., Quaid, K., Xing, X., Schmidt, H., and Wang, T. (2022). Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes and gene regulation. Preprint at bioRxiv, 10.1101/2022.02.01.475239 10.1101/2022.02.01.475239.
13. Du, A.Y., Chobirko, J.D., Zhuo, X., Feschotte, C., and Wang, T. (2023). Regulatory Transposable Elements in the Encyclopedia of DNA Elements. Preprint at bioRxiv, 10.1101/2023.09.05.556380 10.1101/2023.09.05.556380.
14. Diehl, A.G., Ouyang, N., and Boyle, A.P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nat. Commun. *11*, 1796. 10.1038/s41467-020-15520-5.
15. Choudhary, M.N., Friedman, R.Z., Wang, J.T., Jang, H.S., Zhuo, X., and Wang, T. (2020).

Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol. *21*, 16. 10.1186/s13059-019-1916-8.
16. Marnetto, D., Mantica, F., Molineris, I., Grassi, E., Pesando, I., and Provero, P. (2018). Evolutionary rewiring of human regulatory networks by waves of genome expansion. Am. J. Hum. Genet. *102*, 207–218. 10.1016/j.ajhg.2017.12.014.
17. Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. Nature *587*, 240–245. 10.1038/s41586-020-2876-6.
18. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. 10.1038/s41586-020-2493-4.
19. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. 10.1038/nmeth.1906.
20. Okada, N., Sasaki, T., Shimogori, T., and Nishihara, H. (2010). Emergence of mammals by emergency: exaptation. Genes Cells Devoted Mol. Cell. Mech. *15*, 801–812. 10.1111/j.1365-2443.2010.01429.x.
21. 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. U. S. A. *105*, 4220–4225. 10.1073/pnas.0709398105.
22. Pehrsson, E.C., Choudhary, M.N.K., Sundaram, V., and Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. Nat. Commun. *10*, 5640. 10.1038/s41467-019-13555-x.
23. Mika, K., and Lynch, V.J. (2022). Transposable Elements Continuously Remodel the Regulatory Landscape, Transcriptome, and Function of Decidual Stromal Cells. Genome Biol. Evol. *14*, evac164. 10.1093/gbe/evac164.
24. Testori, A., Caizzi, L., Cutrupi, S., Friard, O., De Bortoli, M., Cora', D., and Caselle, M. (2012). The role of Transposable Elements in shaping the combinatorial interaction of Transcription Factors. BMC Genomics *13*, 400. 10.1186/1471-2164-13-400.
25. Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. *18*, 1752–1762. 10.1101/gr.080663.108.
26. Hormozdiari, F., van de Geijn, B., Nasser, J., Weissbrod, O., Gazal, S., Ju, C.J.-T., Connor, L.O., Hujoel, M.L.A., Engreitz, J., Hormozdiari, F., et al. (2019). Functional disease architectures reveal unique biological role of transposable elements. Nat. Commun. *10*, 4054. 10.1038/s41467-019-11957-5.
27. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595. 10.1093/genetics/123.3.585.
28. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. 10.1038/nature15393.
29. Smit, A., Hubley, R., and Green, P. (1996). RepeatMasker Open-3.0.
30. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590-598. 10.1093/nar/gkj144.
31. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinforma. Oxf. Engl. *32*, 289–291. 10.1093/bioinformatics/btv562.
32. Nielsen, R. (2005). Molecular signatures of natural selection. Annu. Rev. Genet. *39*,

197–218. 10.1146/annurev.genet.39.073003.112420.
33. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., et al. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427. 10.1038/ng.3954.
34. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. 10.1126/science.aaz1776.

Legends for Supplementary Tables

Supplementary Table 1

Comparison of phyloP values between regulatory and non-regulatory TEs. The median phyloP values are computed across the 59 genomic segments; the P-value is obtained with a paired Wilcoxon test on the same segments, and the FDR with the Benjamini-Hochberg procedure.

Supplementary Table 2

Enrichment in eGenes in regulatory TEs (dELS type) compared with randomized regulatory TEs.

Supplementary Table 3

Gene Ontology (Biological Process) enrichment of the putative targets of regulatory TEs (dELS type).

Supplementary Table 4

Comparison of Tajima's D values between regulatory and non-regulatory TEs. The median D values are computed across the 59 genomic segments; the P-value is obtained with a paired Wilcoxon test on the same segments, and the FDR with the Benjamini-Hochberg procedure.