Evolutionary arms race between transposable elements and human genes: telomere-to-telomere genome comprehensive analysis identifies young L1 clusters in the interferon-alpha domain

Author: Daniil Nikitin[1*]

Affiliation:
Institute of Molecular Biology, National Academy of Science of the Republic of Armenia

*Correspondence: danya.nikitin.orel@gmail.com

ABSTRACT

Transposable elements (TEs) have played a central role in major evolutionary transitions across the human lineage, from eukaryogenesis to the emergence of the eutherian placenta, and are currently reactivated in cancer and autoimmune diseases. The availability of the complete telomere-to-telomere (T2T) human genome assembly enables comprehensive investigation of TE contributions to gene regulation. Using a 10 kb window in the T2T genome, we performed comprehensive mapping of 3,709,429 human TEs to 28,738 genes with random background correction and assessed the enrichment and functional associations of six TE classes and 44 families. We identified a 220 kb interferon-alpha genomic domain enriched with evolutionarily young L1 elements, suggesting a recent evolutionary arms race influencing innate immune responses. Distinct TE classes exhibited specific functional associations: SVA elements were enriched near genes involved in transcription termination; Alu elements were linked to RNA processing and splicing; MIR elements were associated with genes involved in zinc, copper, and cadmium detoxification; LINE elements were enriched near genes related to lipid metabolism and olfactory perception; and LTR elements were potentially associated with potassium ion channel function. This proximity-based analysis provides a foundational framework for evaluating the functional impact of transposable elements on human gene regulation and their role in driving regulatory innovation.

1. INTRODUCTION

The release of the first complete, gapless human genome assembly (T2T-CHM13) by the Telomere-to-Telomere consortium has fundamentally shifted the paradigms of evolutionary genomics (1). For over two decades, genomic analysis relied on references such as GRCh38, which contained hundreds of megabases of unresolved sequences, primarily concentrated in highly repetitive regions like centromeres, subtelomeres, and acrocentric short arms. These "dark regions" of the genome are disproportionately enriched with transposable elements (TEs), the selfish genetic entities that comprise more than half of the human DNA (2). The transition from GRCh38 to T2T-CHM13 involved the resolution of approximately 238 million base pairs of formerly unmasked DNA, which revealed a significantly higher repeat content than previously estimated (1). This resolved sequence is dominated by tandemly arrayed repeats, such as alpha satellites in centromeric regions, but also contains a vast number of TEs nested within complex genomic structures (2).

TEs are increasingly recognized as drivers of regulatory innovation during the major evolutionary transitions from eukaryogenesis to eutherian placenta and human neocortex (Nikitin, 2026). Moreover, TEs proliferate in a wave pattern within the human genome, and they are in a continuous evolutionary arms race against the host defense systems (4). During this process TEs insert their transcription factor binding sites (TFBS) in the vicinity of host genes and alter their expression (5,6).

There are dozens of studies of this epigenetic impact which rely on different methodology but are close in the main approach, namely proximity mapping: TEs mapped in the vicinity of human genes are likely to impact their expression (7–11). Despite the same principle, different methodologies are difficult to be compared: genomic proximity windows vary from 4 to 20 kb around a gene, the statistical frameworks and epigenomic modalities differ significantly. Moreover, all these studies are based either on hg19 or hg38 human genome assemblies. The unified approach utilizing the most up to date T2T human genome assembly could significantly improve the overall understanding of TEs-host genome evolutionary arms race and its impact on regulatory innovation and human health and disease.

Here we have taken the most widely used proximity window of 10 kb around human transcription start sites (TSS) of 28,738 unique human T2T-annotated genes and we have built a high-resolution functional map of TE enrichment at the level of TE classes and families, taking 3,709,429 individual elements into analysis. We show that SVA elements are enriched near genes responsible for transcription termination, Alu elements are co-associated with RNA processing and splicing genes, MIR repeats relate to genes responsible for zinc, copper and cadmium detoxification, LINE elements in general could impact lipid metabolism and sensory perception of smell, and LTRs are potentially connected with potassium ion channels. We showed that the 220 kb interferon alpha domain is uniquely enriched with young, low-divergence L1 elements indicating the recent example of evolutionary arms race shaping innate immune response. Furthermore, embryogenesis and nervous system genes were relatively depleted with TE insertions constituting evolutionary conservative processes. Finally, synaptic transmission and nervous system development were enriched with evolutionary ancient TEs.

This comprehensive proximity analysis serves as a critical baseline for understanding the functional impact of TEs, as most TE-host interactions, including enhancer exaptation and promoter birth, occur within these 10 kb proximal windows.

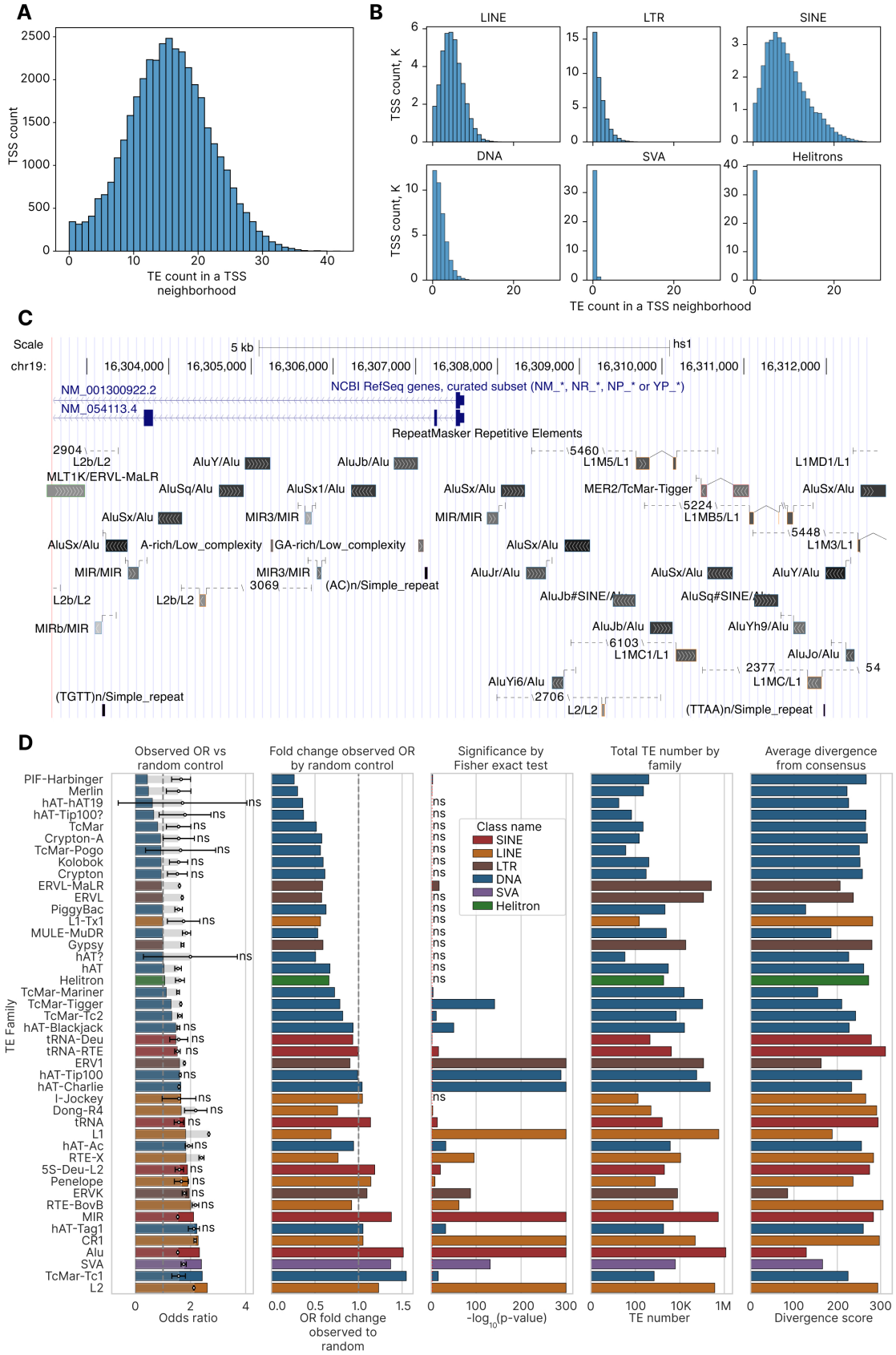
2. RESULTS

2.1.Mapping and enrichment of TEs on gene TSS 10 kb neighborhoods at the level of major classes

The TE coordinates (3,709,429 total entries) were mapped on gene TSS 10 kb neighborhoods (5 kb upstream and 5 kb downstream of each TSS) for 28,738 unique human genes, with 38,704 total unique TSS due to the possibility of multiple TSS per gene giving rise to isoform transcripts. TSS for isoforms were considered as distinct unique entities. The mapping showed that only 0.89% of unique TSS (343 ones) do not intersect with TEs at the proximity level of 10 kb and can be considered as free from TE-mediated regulatory impact at this level (Figure 1A).

On average, each gene harbored 15.05 TEs in its TSS neighborhood (4.39 LINEs, 1.32 LTRs, 7.82 SINEs, 1.49 DNA elements, 0.03 SVA elements and 0.004 Helitrons, Figure 1B). The gene with maximum number of TEs was CIB3 (calcium and integrin binding protein 3), having 42 TEs of various classes (17 LINEs, 23 SINEs, 1 LTR and 1 DNA element, Figure 1C). Lists of TEs mapped on each gene TSS, their divergence and classification, can be found in Supplementary File 1. Most genes had in their TSS neighborhoods at least one TE for the four major classes: 96.9% of all TSS had SINEs, 95.1% of TSS harbored LINEs, 68.5% had DNA elements and 58.7% had LTRs (Figure 1B).

Figure 1. Enrichment of TEs in the 10 kb vicinity of human TSS at the level of classes and families. (A) Distribution of TSS by TE count per TSS. (B) Distributions of TSS by TE count of individual classes. (C) UCSC Genome Browser visualization of CIB3, a gene having the highest TE count in its 10 kb neighborhood. (D) TE families sorted by their degree of enrichment near genes TSS. The leftmost (first) vertical bar plot with bars colored by TE families shows Fisher exact test odds ratio (OR), faint grey bars depict the mean OR of 500 random shuffling iteration, error bars denoting their standard deviations. “ns” marks indicate those families whose empirical p-value (quantile of observed OR in the distribution of 500 random OR values) was higher than 0.05 in either direction after FDR correction for multiple hypothesis testing. The second vertical bar plot shows fold change of the observed OR to mean of the random Ors distribution. The third vertical bar plot visualizes negative decimal logarithm of Fisher exact test p-value (FDR-corrected). The fourth and the fifth vertical bar plots show total TE number and average divergence (substitutions per 1000 base pairs) by family, respectively.


A
TSS count
TE count in a TSS neighborhood
B
LINE
LTR
SINE
DNA
SVA
Helitrons
TSS count, K
TE count in a TSS neighborhood
C
Scale
5 kb
hs1
chr19:
16,304,000
16,305,000
16,306,000
16,307,000
16,308,000
16,309,000
16,310,000
16,311,000
16,312,000
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*)
NM_001300922.2
NM_054113.4
RepeatMasker Repetitive Elements
L2b/L2
AluY/Alu
AluJb/Alu
L1M5/L1
L1MD1/L1
MLT1K/ERVL-MaLR
AluSq/Alu
AluSx1/Alu
AluSx/Alu
MER2/TcMar-Tigger
AluSx/Alu
MIR3/MIR
MIR/MIR
L1MB5/L1
A-rich/Low_complexity
GA-rich/Low_complexity
L1M3/L1
AluJr/Alu
MIRb/MIR
(AC)n/Simple_repeat
AluJb#SINE/Alu
AluSq#SINE/Alu
AluYh9/Alu
L1MC1/L1
AluJo/Alu
AluYi6/Alu
L1MC/L1
(TGTT)n/Simple_repeat
L2/L2
(TTAA)n/Simple_repeat
D
Observed OR vs random control
Fold change observed OR by random control
Significance by Fisher exact test
Total TE number by family
Average divergence from consensus
TE Family
PIF-Harbinger
Merlin
hAT-hAT19
hAT-Tip100?
TcMar
Crypton-A
TcMar-Pogo
Kolobok
Crypton
ERVL-MaLR
ERVL
PiggyBac
L1-Tx1
MULE-MuDR
Gypsy
hAT?
hAT
Helitron
TcMar-Mariner
TcMar-Tigger
TcMar-Tc2
hAT-Blackjack
tRNA-Deu
tRNA-RTE
ERV1
hAT-Tip100
hAT-Charlie
I-Jockey
Dong-R4
tRNA
L1
hAT-Ac
RTE-X
5S-Deu-L2
Penelope
ERVK
RTE-BovB
MIR
hAT-Tag1
CR1
Alu
SVA
TcMar-Tc1
L2
ns
Class name
SINE
LINE
LTR
DNA
SVA
Helitron
Odds ratio
OR fold change observed to random
-log10(p-value)
TE number
Divergence score

In order to test the significance of genes TSS neighborhoods by TEs of certain classes and families, we generated 500 random permutations of all human TEs from T2T RepeatMasker (1,706,485 SINEs, 1,005,214 LINEs, 531,410 LTRs, 458,177 DNA elements, 6,274 SVAs, and 1,869 Helitrons) and intersected them with the same set of gene TSS, counting number of TEs of a given group. Significance of enrichment or deficiency of any TE group (classes and families) was assessed at the two levels: firstly, by odds ratio (OR) according to Fisher exact test, and then the empirical p-value derived from the 500 random permutations to account for artificially higher probability of intersection for longer elements. Table 1 shows assessment of TE enrichment in genes TSS 10 kb neighborhoods by class. SINEs and SVA elements were enriched in the TSS proximity by a factor of 1.468 and 1.368, respectively, whereas LINEs, LTRs and DNA elements were depleted in the TSS neighborhoods by a factor of 0.877, 0.667 and 0.938, respectively. In general, the observed OR for all TEs was 1.94 (1.78 the random one), which was showing an enrichment by a factor of 1.097 (empirical p-value = 0.004).

Table 1. Enrichment of TE classes in gene TSS neighborhoods.

Class name	TE count in TSS	TE count total	Odds ratio (OR)	Fisher p-value	Adjusted Fisher p-value	Mean of random OR	SD of random OR	Observed to random OR fold change	Empirical p-value	Adjusted empirical p-value
LINE	169930	1005214	2.13	$<10^{-200}$	$<10^{-200}$	2.43	0.009	0.877	0.004	0.004
LTR	51103	531410	1.11	$6.3*10^{-112}$	$7.6*10^{-112}$	1.67	0.010	0.667	0.004	0.004
SINE	302480	1706485	2.25	$<10^{-200}$	$<10^{-200}$	1.53	0.005	1.468	0.004	0.004
DNA	57684	458177	1.51	$<10^{-200}$	$<10^{-200}$	1.61	0.010	0.938	0.004	0.004
SVA	1170	6274	2.40	$9.3*10^{-133}$	$1.4*10^{-132}$	1.75	0.094	1.368	0.004	0.004
Helitrons	173	1869	1.07	0.41	0.41	1.61	0.163	0.661	0.004	0.004

2.2.Mapping and enrichment of TEs on gene TSS 10 kb neighborhoods at the level of families

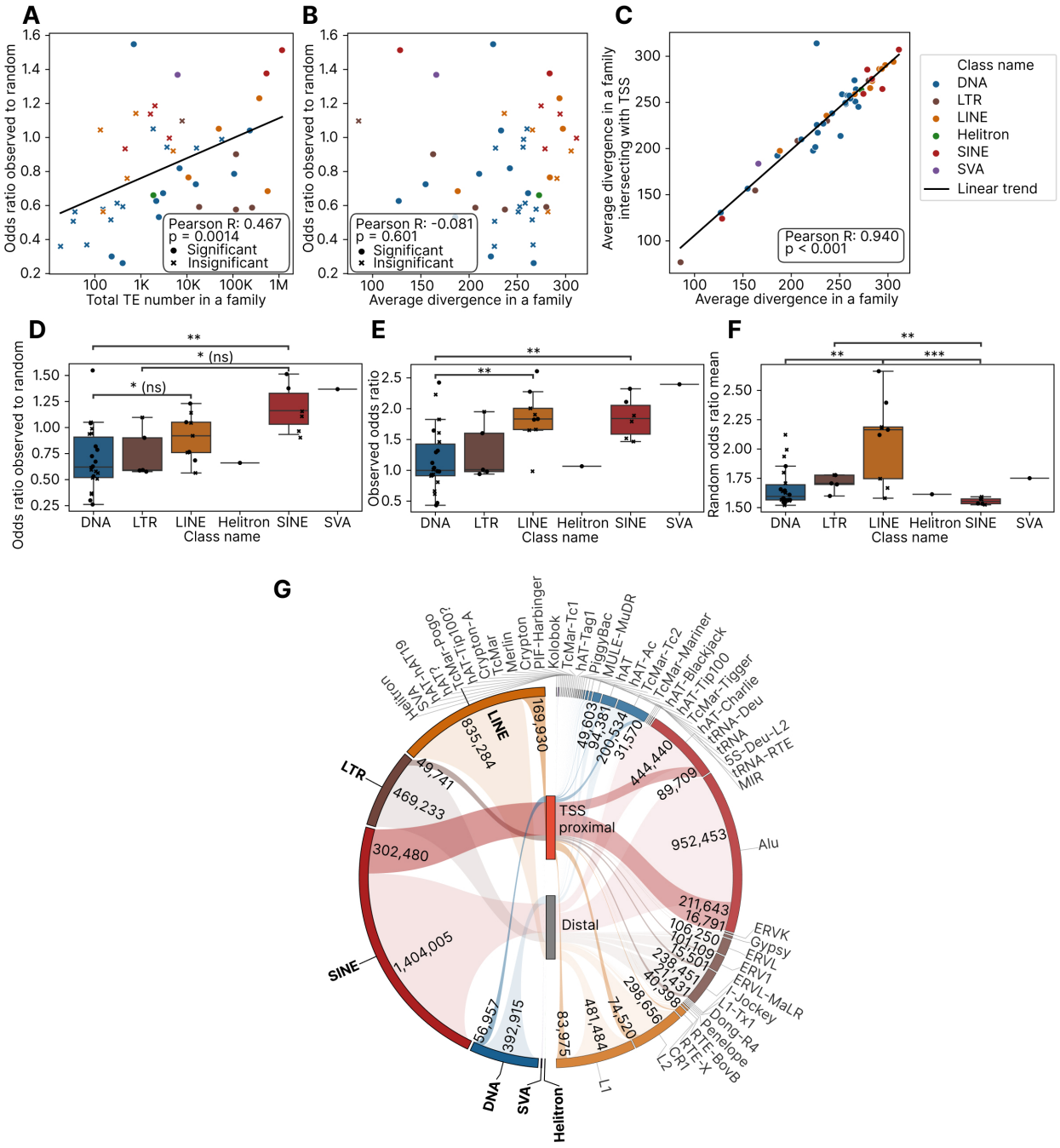
At the level of families, a more complicated picture was observed: DNA families were mainly depleted in the vicinity of genes TSS (compared to the random OR), whereas LINEs and LTRs were less depleted or even enriched near TSS (Figure 1D, Supplementary File 2). Only 7 out of 44 TE families were significantly enriched according to both tests (the Fisher exact and the permutation-based one): hAT-Charlie (DNA, 1.041 enrichment measured fold change of the observed to random OR), MIR (SINE, 1.377), CR1 (LINE, 1.051), Alu (SINE, 1.513), SVA elements (1.368), TcMar-Tc1 (DNA, 1.548) and L2 (LINE, 1.230). In contrast, 9 families were significantly depleted in the vicinity of TSS: PIF-Harbinger (DNA, 0.262 fold change of the observed to random OR), Merlin (DNA, 0.301), ERVL-MaLR (LTR, 0.587), TcMar-Mariner (DNA, 0.725), TcMar-Tigger (DNA, 0.786), TcMar-Tc2 (0.819), ERV1 (LTR, 0.902), L1 (LINE, 0.684), RTE-X (LINE, 0.765).

Families with lower number of members were tending to have lower fold change of the observed to random OR (Figure 2A), whereas there was no apparent relationship between the fold change (level of enrichment) and average divergence within a family (Figure 2B). Average divergence in all TEs by families was strongly correlated with average divergence in those TEs that are intersecting with TSS neighborhoods (Figure 2C), with a single exception of a DNA family

hAT-hAT19 which had 18 members in total, one of them appearing in the vicinity of gene TSS, and this single element had divergence 1.39 higher than the average all elements in a family.

Additionally, DNA elements were significantly depleted within the TSS neighborhoods compared to SINEs (Figure 2D, all the rest pairwise comparisons were non-significant after the FDR correction). The observed OR itself (without correction for random permutations) was significantly lower for DNA elements compared to both LINEs and SINEs (Figure 2E), whereas the highest number of significant differences was found by the random background OR: between DNA elements and LINEs, between LINEs and SINEs and between LTRs and SINEs (Figure 2F). These differences reflect the variable length of TEs by classes and highlight the importance of permutation-based random correction instead of the conventional Fisher exact test-based approaches. An integrative map of TEs that have been mapped to the TSS proximal regions and the distal ones at the level of classes and families is shown in Figure 2G.

Figure 2. Further investigation of TE enrichment in the vicinity of TSS at the level of families. (A) Scatter plot of observed to random OR by total number of TEs in a family. (B) Scatter plot of observed to random OR by average divergence in a TE family. (C) Scatter plot of average divergence of TSS neighborhood intersecting TEs by average divergence of all TEs in a family. (D) Box plot of observed to random OR by TE families between TE classes. (E) Box plot of observed OR by TE families between TE classes. (F) Box plot of random OR by TE families between TE classes. (G) Circular plot showing numbers of TE classes and families mapped in TSS proximal and distal regions.
For all group comparisons, significance is assessed by the Mann–Whitney test and FDR-corrected: ns, $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $0.0001 < p < 0.001$; ****, $p < 0.0001$.



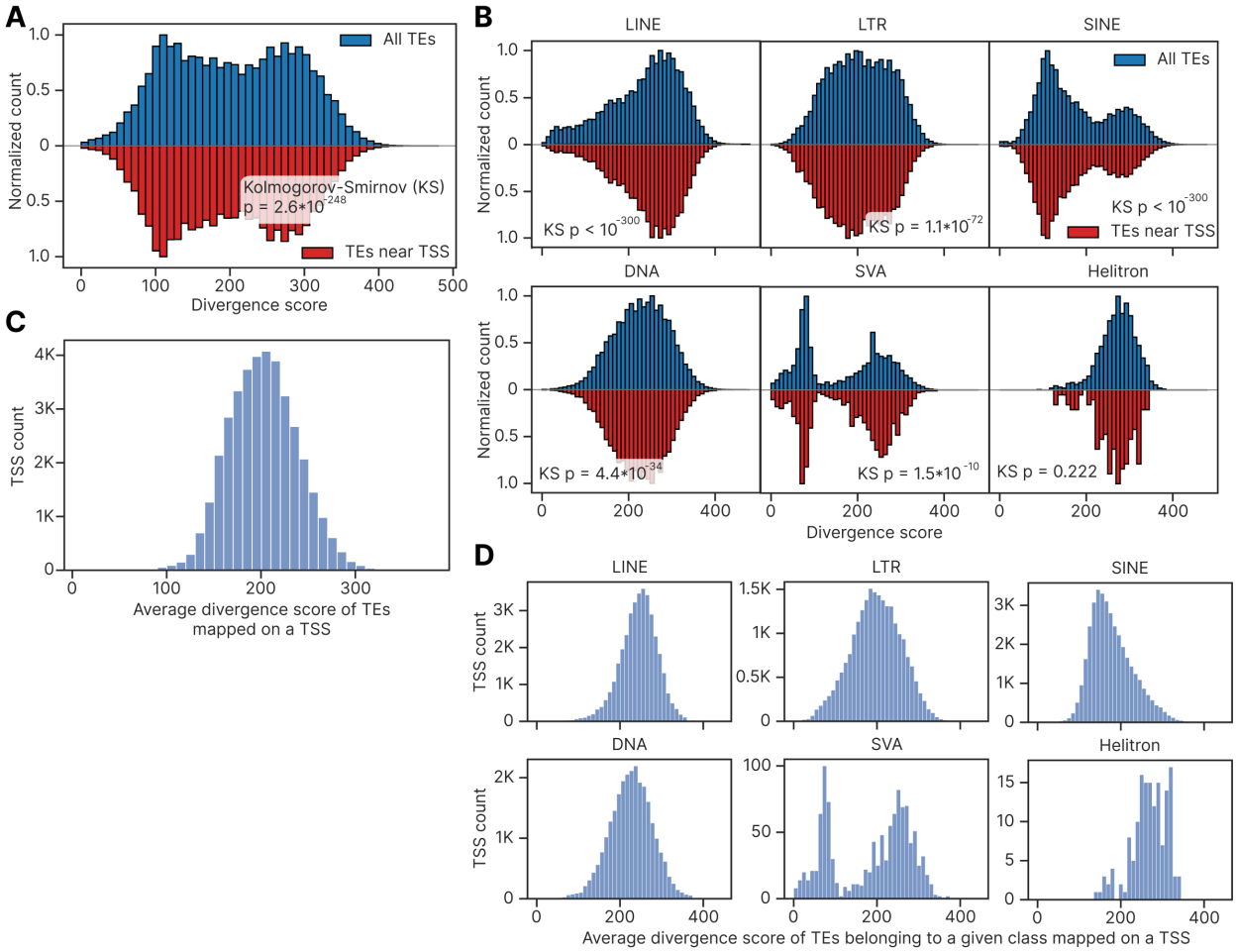
2.3.Evolutionary age and length of TSS-proximal and distal TEs

To further understand evolutionary dynamics of TEs insertions near host genes, we compared distributions of divergence between all TEs and those that mapped on the TSS neighborhoods (Figure 3A). The two distributions for all TEs were visually indistinguishable with characteristic bimodal shape observed earlier, although statistically significant differences were highlighted by Kolmogorov-Smirnov test. The same pattern was observed for individual TE classes (Figure 3B),

with two peaks found in SINEs and SVA elements. Conversely, divergence of TEs that were intersecting with gene TSS neighborhoods, averaged over all TEs at the level of individual genes, followed a unimodal pattern both for all classes and for individual classes (Figure 3C, 3D, respectively). For all classes, the peak divergence averaged among genes was 200-210 (Figure 3C), whereas individual TEs without averaging over genes showed two peaks 110-120 and 270-280, regardless of intersecting with gene TSS (Figure 3A). The same averaging effect was observed for SINEs: two peaks of 90-100 and 290-300 (Figure 3B) converged to the single peak of 150 (Figure 3D), indicating that there was no preference of integrating near any group of genes for SINE elements of different evolutionary age. For SVA elements the bimodal pattern was observed in the case of TSS averaging (Figure 3D) due to the low number of TSS with SVA elements in their vicinity (1,140 genes, 2.9%).
Comparison of TE length distributions between all TEs and those that mapped on TSS neighborhoods resulted in the similar absence of difference between the two TE groups, either for all TEs or for their individual classes (Supplementary Figure 1A, 1B).

We also compared divergence distributions between all and TSS-proximal TEs by individual families (Supplementary Figure 2). Applying Mann-Whitney test with FDR correction resulted in 21 out of 44 families showing significant difference, albeit only 4 families showed magnitude of absolute difference above 10%: TcMar-Tc1, ERVK and tRNA SINEs demonstrated more than 10% higher average divergence for all insertions compared to the TSS-proximal ones (10.3%, 10.1%, 10.1% respectively), and SVA elements had 10.6% higher divergence in insertions that overlapped with 10 kb TSS neighborhoods.
The analogous comparison by length resulted in 13 significantly different cases out of 44 TE families (Supplementary Figure 3), with 4 families showing higher than 10% increase of average length near genes (hAT-Charlie with 10.7%, ERV1 with 17.6%, ERVL with 19.8%, TcMar with 53.0%), and a single family, hAT, showing 11.9% decrease of average length for TSS-proximal insertions. The TcMar family had 223 total and 16 TSS-proximal members in the T2T genome, reflecting likely random nature of such a high difference.

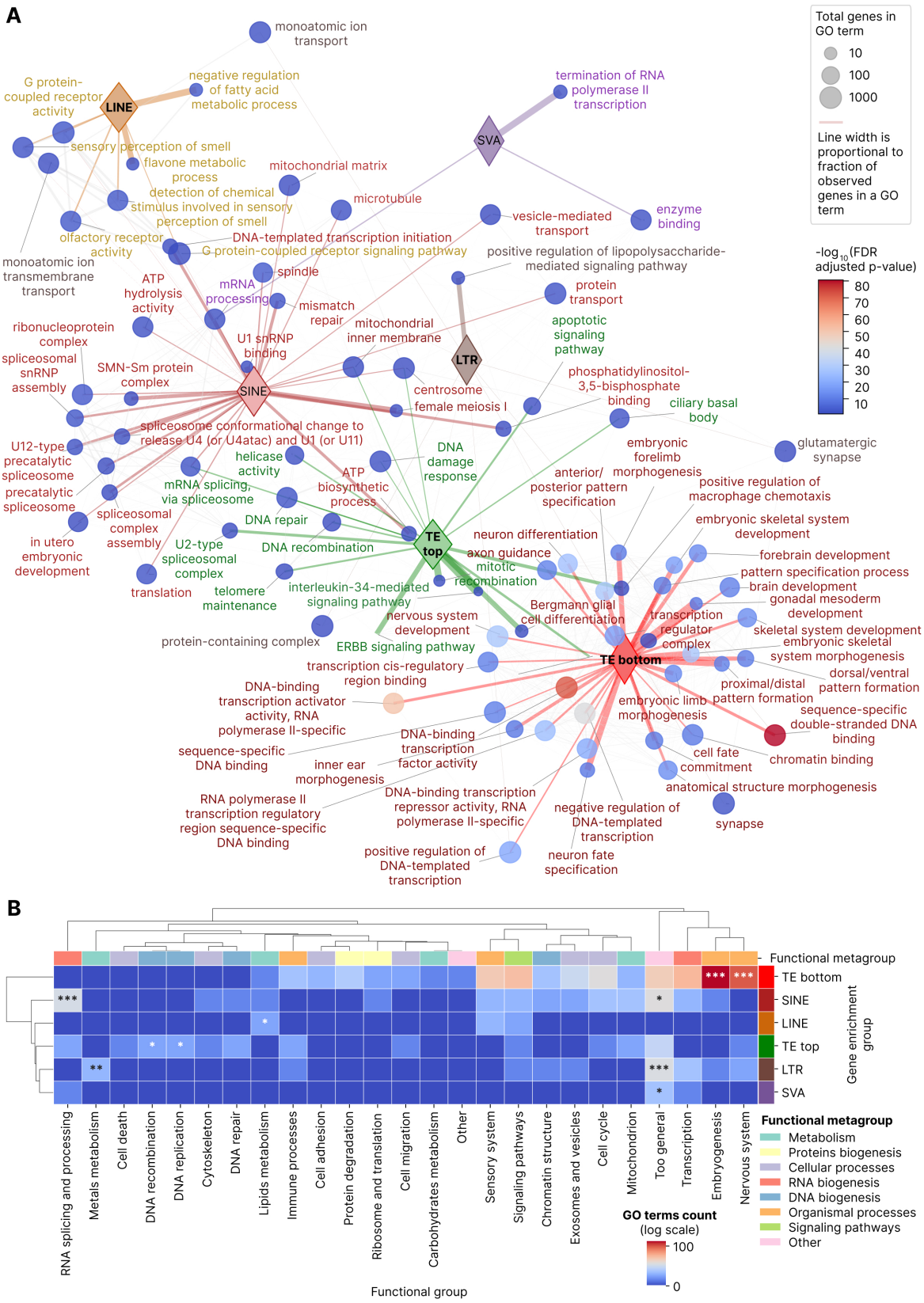
Figure 3. Evolutionary age comparison of all and TSS-proximal TEs in general and by classes. (A) Ridge plot of all and TSS-proximal TEs by divergence score. (B) Ridge plots of all and TSS-proximal TEs by divergence score at the level of individual classes. (C) Average divergence score distribution of TSS-proximal TEs, averaged by TSS. (D) Average divergence score distribution of TSS-proximal TEs at the level of individual classes.



2.4.Functional characterization of genes enriched by TE classes by count

To study whether TEs of different classes are nonrandomly accumulating near genes of certain functions, for each TE class we selected 5% of human genes whose TSS 10 kb neighborhoods contained highest number of TE elements of a given class. There were 1436 genes for each of the major classes (LTR, LINE, SINE, DNA elements, Supplementary Figure 4A, Supplementary File 3), genes having at least 5, 10, 18 and 5 elements of these classes have been taken, respectively (Supplementary Figure 5A). Also, we extracted 962 genes with SVA elements and 130 genes with Helitrons (all genes with SVA elements and Helitrons were taken). Finally, we added 1436 genes with highest (starting with 27 till 42 elements per TSS) and the lowest (starting with 4 till 0 elements per TSS) TEs count of all classes per TSS (Supplementary Figure 5A). Because 28,738 unique genes had 38,704 TSS coordinates, and TE counts were measured on TSS whereas GO analysis is possible with unique gene names only, genes with maximum TE count could have TSS with lower TE count than the minimal thresholds, as indicated in Supplementary Figure 5A. Since SINEs are the most numerous and enriched near genes TE class, they had the highest number of top genes which are shared with the top genes by all TEs (Supplementary Figure 4A), and lowest number of top genes that are unique to SINEs (612 compared to 1074 for DNA elements, 1133 for LTRs and 1031 for LINEs).

Figure 4. Functional analysis of genes whose TSS are enriched or depleted with TEs of different classes. (A) Connection map of GO terms extracted for top 5% of TSS enriched with LINEs, SINEs, LTRs in their proximity and all TSS with SVA elements, as well as top 5% of TSS enriched or depleted with TEs of any class. Connection line width is proportional to fraction of shared genes, color of each node denotes a GO term enrichment p-value (FDR-corrected), node size shows number of genes in a GO term. Top 30 terms by enrichment p-value were selected for each group, having FDR corrected p-value below 0.1. GO terms with more than 1000 genes were excluded to avoid too general terms. (B) Heatmap showing GO terms number by functional group and gene enrichment group (TE top, TE bottom and TE classes). Stars indicate FDR-corrected Fisher enrichment p-value of a given functional group in each gene enrichment group compared to other enrichment groups.


A
Total genes in GO term
10
100
1000
Line width is proportional to fraction of observed genes in a GO term
-log10(FDR adjusted p-value)
80
70
60
50
40
30
20
10
monoatomic ion transport
G protein-coupled receptor activity
LINE
negative regulation of fatty acid metabolic process
termination of RNA polymerase II transcription
SVA
sensory perception of smell
flavone metabolic process
mitochondrial matrix
detection of chemical stimulus involved in sensory perception of smell
microtubule
vesicle-mediated transport
enzyme binding
olfactory receptor activity
DNA-templated transcription initiation
G protein-coupled receptor signaling pathway
monoatomic ion transmembrane transport
positive regulation of lipopolysaccharide-mediated signaling pathway
spindle
ATP hydrolysis activity
mRNA processing
protein transport
mismatch repair
ribonucleoprotein complex
U1 snRNP binding
mitochondrial inner membrane
apoptotic signaling pathway
spliceosomal snRNP assembly
SMN-Sm protein complex
SINE
LTR
phosphatidylinositol-3,5-bisphosphate binding
centrosome
female meiosis I
ciliary basal body
spliceosome conformational change to release U4 (or U4atac) and U1 (or U11)
glutamatergic synapse
U12-type precatalytic spliceosome
helicase activity
DNA damage response
embryonic forelimb morphogenesis
precatalytic spliceosome
mRNA splicing, via spliceosome
ATP biosynthetic process
anterior/posterior pattern specification
positive regulation of macrophage chemotaxis
spliceosomal complex assembly
DNA repair
TE top
neuron differentiation
embryonic skeletal system development
in utero embryonic development
U2-type spliceosomal complex
DNA recombination
axon guidance
forebrain development
translation
telomere maintenance
interleukin-34-mediated signaling pathway
mitotic recombination
pattern specification process
brain development
Bergmann glial cell differentiation
gonadal mesoderm development
nervous system development
transcription regulator complex
skeletal system development
protein-containing complex
ERBB signaling pathway
embryonic skeletal system morphogenesis
TE bottom
transcription cis-regulatory region binding
dorsal/ventral pattern formation
DNA-binding transcription activator activity, RNA polymerase II-specific
proximal/distal pattern formation
embryonic limb morphogenesis
sequence-specific double-stranded DNA binding
DNA-binding transcription factor activity
sequence-specific DNA binding
cell fate commitment
chromatin binding
inner ear morphogenesis
anatomical structure morphogenesis
DNA-binding transcription repressor activity, RNA polymerase II-specific
RNA polymerase II transcription regulatory region sequence-specific DNA binding
negative regulation of DNA-templated transcription
synapse
positive regulation of DNA-templated transcription
neuron fate specification
B
Functional metagroup
TE bottom
SINE
LINE
TE top
LTR
SVA
Gene enrichment group
RNA splicing and processing
Metals metabolism
Cell death
DNA recombination
DNA replication
Cytoskeleton
DNA repair
Lipids metabolism
Immune processes
Cell adhesion
Protein degradation
Ribosome and translation
Cell migration
Carbohydrates metabolism
Other
Sensory system
Signaling pathways
Chromatin structure
Exosomes and vesicles
Cell cycle
Mitochondrion
Too general
Transcription
Embryogenesis
Nervous system
Functional group
Functional metagroup
Metabolism
Proteins biogenesis
Cellular processes
RNA biogenesis
DNA biogenesis
Organismal processes
Signaling pathways
Other
GO terms count
(log scale)
100
0

Each gene set was tested against the canonical Gene Ontology (GO) Biological Process, Molecular Function and Cellular Component databases and significant terms were extracted (FDR threshold of 0.1 was applied, Supplementary File 4). No significant enrichments were found for DNA elements and Helitrons. The integrative network visualization of top 30 the most significant terms per each of the remaining group (Figure 4A) showed that embryogenesis processes were the major ones among the 5% genes depleted with TEs of any class (with transcription being the second major category), whereas RNA splicing, DNA repair, telomere, apoptosis, IL-34 and ERBB signaling were among top 5% genes with highest TE count. Moreover, the three all TEs depleted terms with the lowest FDR-corrected p-values in the entire set (p-values in the range 10^{-40} – 10^{-80}) related to transcriptional activators. SINE-specific processes were the closest to the top TE count processes by the overlapping genes, sharing splicing and DNA repair as SINEs constitute the majority of TEs mapping in the vicinity of TSS (Figure 2G). The rest classes with significant GO terms based on genes with highest TE count led to olfactory receptor activity, flavone metabolism and regulation of fatty acids metabolism (LINEs), glutaminergic synapse, lipopolysaccharide mediated signaling and ion transport (LTRs) and RNA polymerase II termination (SVA elements).

The latter term genes that had SVA elements in their vicinity were POLR2A (core catalytic subunit of RNA polymerase II) and genes SSU72L1, SSU72L2, SSU72L3, SSU72L4 and SSU72L5 – protein phosphatases acting on RNA polymerase II C-terminal domain, whose 10 kb TSS neighborhoods are located in a 116 kb region of chromosome 11 (4293265 - 4409432) and have 3 copies of SVA B subfamily.

The GO terms were manually (with Gemini pro assistance) classified into 25 major groups according to the previous studies (5,12,13) and compared using Fisher exact test (Figure 4B), testing enrichment of a given molecular process in a given TE group (classes, TE top and TE bottom) versus the same molecular process in the rest TE groups. While there was no systematic co-clustering of molecular processes metagroups, RNA splicing and processing processes were specifically enriched in SINEs adjacent genes, metals metabolism was associated with LTRs, DNA replication and recombination related to all TEs, lipid metabolism was enriched in LINEs adjacent genes. Embryogenesis and nervous system processes genes were preferentially depleted with TE inserts of any class in their vicinity.

2.5.Functional characterization of genes enriched by TE classes by TE evolutionary age

The next part of the analysis was dedicated to study associations of evolutionary young and the most ancient TEs by classes with the host genes. For each of the major TE families (LTR, LINE, SINE, DNA) and all TEs we calculated average divergence score of all TEs mapped in the 10 kb vicinity of each TSS (Figures 3C, 3D). Then we selected top and bottom 5% of TSS with highest and lowest divergence, considering all TSS with TEs of a given group mapped. We extracted 1425 genes of top and bottom divergence for all TEs, 1396 genes for SINEs, 1372 for LINEs, 1015 for DNA elements and 905 for LTR elements (Supplementary Figure 4B, Supplementary File 5). The upper limits for lowest divergence were at the level of 10-17% depending on the TE class, and the lower limits for highest divergence were at the level of 28-31% (Supplementary Figure 5B).

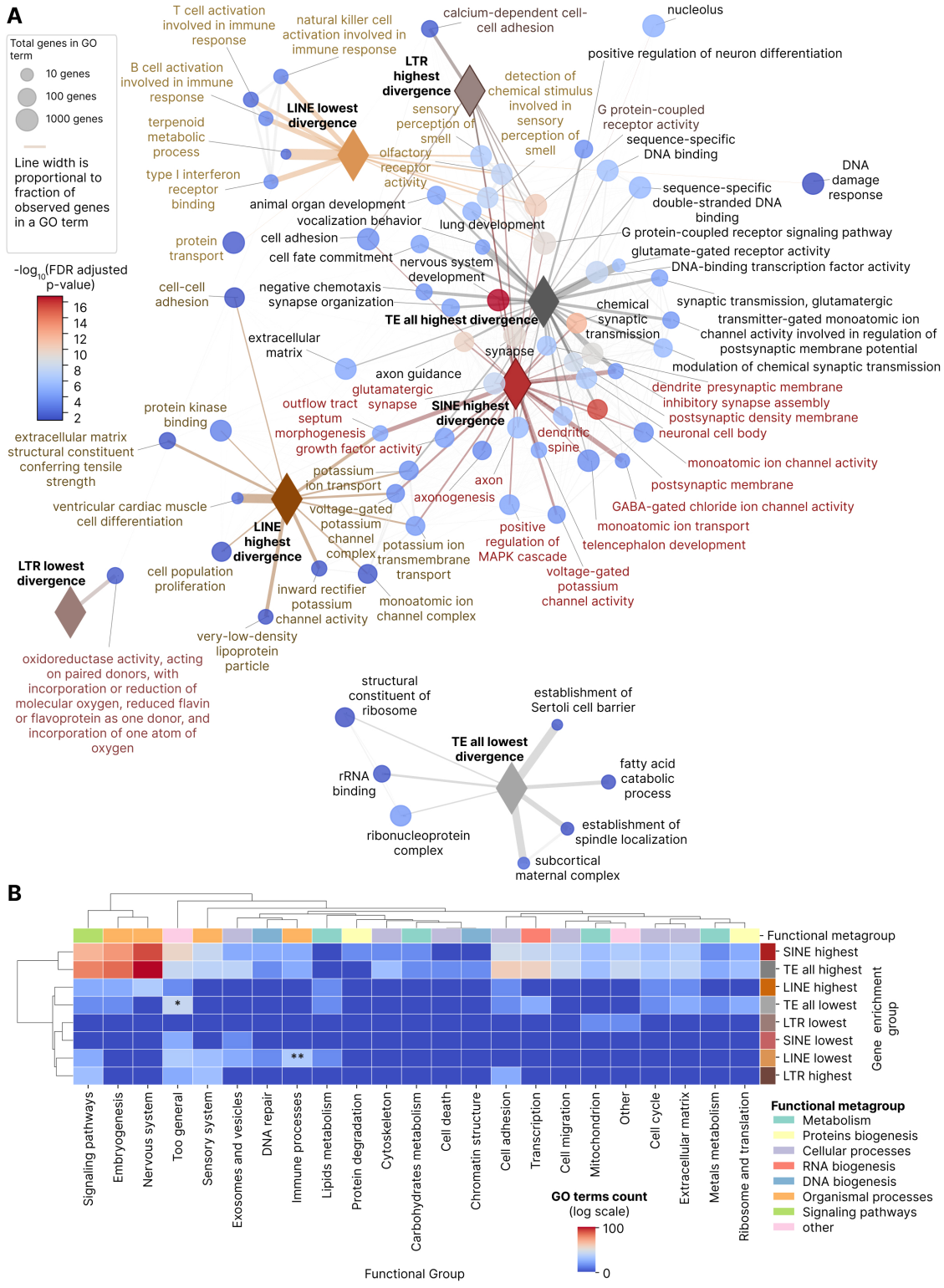
We visualized top 30 GO terms of each TE-divergence groups (Supplementary File 6), filtered by no more than 1000 genes per term to avoid too general classification (Figure 5A). The group of all TEs with low divergence was isolated from all other groups, returning terms about rRNA binding, spermatogenesis, mitotic spindle localization, subcortical maternal complex and fatty acids catabolism. LTR elements of lowest divergence demonstrated a single term of flavin-based oxidoreductases. Groups of high divergence (all TEs, LINEs, SINEs and LTRs) we co-clustered, with all TEs and LTRs of high divergence sharing olfactory receptors with LINEs of lowest divergence. LINEs of highest divergence were inserted near genes of voltage-gated potassium channel complex, cell adhesion, differentiation and proliferation, as well as genes of very-low-density lipoprotein components. SINEs and all TEs of highest divergence shared dozens of GO terms connected with nervous system, especially ion channels and synapses, with smaller number of embryogenesis and transcription GO terms, specific for all TEs of highest divergence. Top 3 GO terms with the most significant enrichment were nervous system development, postsynaptic membrane and synaptic chemical transition (FDR corrected p-value 10^{-16} – 10^{-11}), all three connected with all TEs of highest divergence and SINEs of highest divergence. LTR elements of highest divergence had also a specific term of calcium-dependent cell adhesion. LINE elements of lowest divergence demonstrated T, B and NK cell activation and type I interferon receptor binding, as well as terpenoid metabolism. These immune system GO terms have been sharing the same core interferon gene set, namely IFNA10, IFNA16, IFNA17, IFNA21, IFNA4, IFNA6, IFNA7, IFNW1, whose TSS neighborhoods located in the interferon alpha domain of chromosome 9 (coordinates 21150692 to 21370055, 220 kb region) and having average divergence of intersecting LINE elements at the level of 95 – 161.7 (Supplementary File 1, Supplementary Figure 6).

Finally, DNA elements of highest and lowest divergence, and SINEs of lowest divergence, revealed no significant GO enrichments reflecting potential absence of any functional preference of insertions, preferential mutational decay or negative possible selection, which is expected for DNA elements but surprising for young SINEs.

We then compared enrichment of different functional groups of GO terms in low or high divergence TEs of different classes (Figure 5B). While SINEs and all TE elements showed high similarity between their processes (because SINEs are quantitatively predominant and enriched near genes), only immune processes were enriched for LINEs of lowest divergence groups, indicating possible molecular or evolutionary mechanism behind this pattern. No biologically reasonable co-clustering of TE groups or functional metagroups was found.

Figure 5. Functional analysis of genes whose TSS are enriched with TEs of different classes with highest and lowest divergence as an estimator of evolutionary age. (A) Connection map of GO terms extracted for 5% of TSS with highest and lowest average divergence of LINEs, SINEs, LTRs, DNA elements and all TEs. Connection line width is proportional to fraction of shared genes, color of each node denotes a GO term enrichment p-value (FDR-corrected), node size shows number of genes in a GO term. Top 30 terms by enrichment p-value were selected for each group, having FDR corrected p-value below 0.1. GO terms with more than 1000 genes were excluded to avoid too general terms. (B) Heatmap showing GO terms number by functional group and gene enrichment group: highest and lowest divergence of LINEs, SINEs, LTRs, DNA

elements and all TEs. Stars indicate FDR-corrected Fisher enrichment p-value of a given functional group in each gene enrichment group compared to other enrichment groups.



2.6.Functional characterization of genes enriched by TE families by count

To analyze an enrichment of individual TE families near functional gene, we ordered TE families by number of TSS having at least one TE of a given family and then visualized the TE count distributions in the TSS vicinity (Supplementary Figure 7). 34977 (90.4%) of all TSS had at least one Alu copy in their 10 kb neighborhood (up to 29 copies per TSS), 30695 (79.3%, up to 19 copies) had MIR elements, 30184 (78.0%, again up to 19 copies) had L1 elements, 30376 (78.5%, again up to 17 copies) had L2 elements and 18731 (48.4%, again up to 19 copies) had hAT-Charlie elements in their vicinity, whereas all the rest families were occupying less proximal regions.

For TE families that had more than 5% of all unique genes in their vicinity (1436 and more) we selected 1436 genes having highest count of elements of a given family in their vicinity, whereas for the less numerous and/or enriched families we took all the potentially affected genes (Supplementary File 7). Network of these gene sets, and their overlapping's is shown in Supplementary Figure 8A (no filtering of the graph edges by Jaccard index). Overlapping's were very weak as it can be expected for families with random and independent insertions, with highest Jaccard index of genes with two families being 8% (Supplementary Figure 8A).

GO analysis for each family resulted in only 14 families having statistically significant (FDR-adjusted p-value < 0.1) terms (out of 44 total families): 4 LINE families (L1, L2, CR1, Dong-R4), 5 LTR ones (ERVL, ERVK, ERV1, ERVL-MaLR and Gypsy), 2 SINE families (Alu and MIR elements), SVA elements and 2 DNA families (hAT-Charlie and hAT-Tip100) (Figure 6A, Supplementary File 8). 3 of these families (hAT-Tip100, Dong-R4 and L2) had only 1 significant GO term per family, reflecting likely a random nature of these enrichments.

Visualization of top 30 GO terms by family in Figure 6A showed a high degree of functional distinction between processes by families. L1 elements, as it was previously shown for LINEs, were connected with olfactory receptors, fatty acids and flavone metabolism, Dong-R4 was connected with axonal transport of mitochondrion (3 out of 15 genes), L2 had non-informative protein binding, CR1 elements had surprising connections with neurotransmitter processes (postsynaptic receptor internalization and response to cocaine), ventricular septum development and endopeptidase inhibitor activity.

LTR families demonstrated 44 unique GO terms compared to 25 ones in LTRs as a class (Figure 5A). ERVL elements were connected with bitter taste receptors and with keratins, ERVK were associated with fatty acids metabolic process, ERV1 elements showed the most diverse set of GO terms and were inserted adjacent to genes of xenobiotics metabolism, fatty aldehyde dehydrogenases, succinyl-CoA breaking down, suppression of endosymbionts, arylsulfatase, zinc ion binding, ubiquitin ligase and sphingolipid metabolism. Gypsy elements were associated with translation activators, and ERVL-MaLR had weak connection with extracellular exosomes.

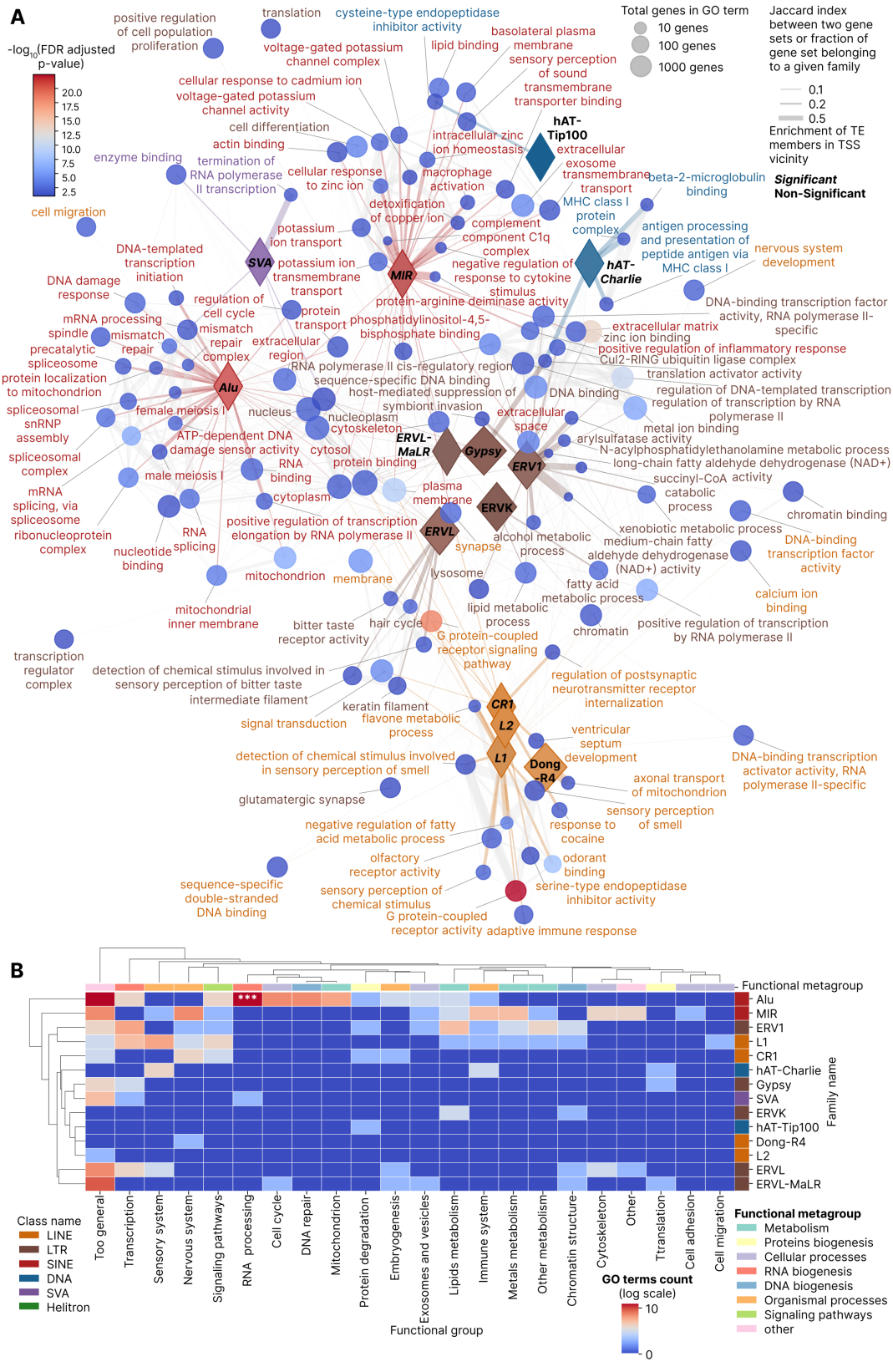
Alu and MIR insertions-adjacent GO terms were remarkably different (although both families enriched near genes, but Alu ones were significantly younger than the MIR ones), with Alu-

specific ones connected with RNA splicing, DNA repair, meiosis, cell cycle and transcription initiation, which is very similar to the general SINEs patterns. Contrastingly, MIR elements were inserted in the vicinity of genes of voltage-gated potassium ion channels, phosphatidylinositol-4,5-bisphosphate binding, arginine deiminases, cellular response to cadmium, copper and zinc ions, macrophage activation, exosomes, complement, sensory perception of sound and negative regulation of cytokine signaling.

The only two DNA elements that showed non-random enrichment of adjacent genes GO terms were hAT-Tip100 and hAT-Charlie. The former was associated with a single significant GO term of cysteine-type endopeptidase inhibitor activity, the latter related to MHC class I via three GO terms (Figure 6A).

Finally, SVA elements, as previously for the class-level analysis (Figure 5A), related to termination of RNA polymerase II transcription.

Figure 6. Functional analysis of genes with TSS enriched by TE inserts by family. (A) Connection map of GO terms extracted for 5% of TSS with number of TE insertions by family. Connection line width is proportional to fraction of shared genes, color of each node denotes a GO term enrichment p-value (FDR-corrected), node size shows number of genes in a GO term. Top 30 terms by enrichment p-value were selected for each group, having FDR corrected p-value below 0.1. (B) Heatmap showing GO terms number by functional group and gene enrichment group (TE family). Stars indicate FDR-corrected Fisher enrichment p-value of a given functional group in each gene enrichment group compared to other enrichment groups.


A
-log10(FDR adjusted p-value)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Total genes in GO term
10 genes
100 genes
1000 genes
Jaccard index between two gene sets or fraction of gene set belonging to a given family
0.1
0.2
0.5
Enrichment of TE members in TSS vicinity
Significant
Non-Significant
positive regulation of cell population proliferation
translation
cysteine-type endopeptidase inhibitor activity
basolateral plasma membrane
lipid binding
voltage-gated potassium channel complex
sensory perception of sound
transmembrane transporter binding
cellular response to cadmium ion
voltage-gated potassium channel activity
cell differentiation
actin binding
intracellular zinc ion homeostasis
hAT-Tip100
extracellular exosome
enzyme binding
termination of RNA polymerase II transcription
cellular response to zinc ion
macrophage activation
transmembrane transport
beta-2-microglobulin binding
cell migration
detoxification of copper ion
MHC class I protein complex
potassium ion transport
complement component C1q complex
antigen processing and presentation of peptide antigen via MHC class I
nervous system development
DNA-templated transcription initiation
SVA
potassium ion transmembrane transport
MIR
negative regulation of response to cytokine stimulus
hAT-Charlie
DNA damage response
protein-arginine deiminase activity
DNA-binding transcription factor activity, RNA polymerase II-specific
regulation of cell cycle
mRNA processing
protein transport
phosphatidylinositol-4,5-bisphosphate binding
extracellular matrix
zinc ion binding
mismatch repair
spindle
extracellular region
positive regulation of inflammatory response
precatalytic spliceosome
RNA polymerase II cis-regulatory region
Cul2-RING ubiquitin ligase complex
protein localization to mitochondrion
Alu
sequence-specific DNA binding
translation activator activity
host-mediated suppression of symbiont invasion
DNA binding
regulation of DNA-templated transcription
spliceosomal snRNP assembly
female meiosis I
nucleus
nucleoplasm
extracellular space
regulation of transcription by RNA polymerase II
cytoskeleton
metal ion binding
ATP-dependent DNA damage sensor activity
ERVL-MaLR
Gypsy
arylsulfatase activity
N-acylphosphatidylethanolamine metabolic process
spliceosomal complex
cytosol
ERV1
long-chain fatty aldehyde dehydrogenase (NAD+)
RNA binding
protein binding
male meiosis I
succinyl-CoA activity
mRNA splicing, via spliceosome
plasma membrane
catabolic process
ERVK
cytoplasm
chromatin binding
positive regulation of transcription elongation by RNA polymerase II
ERVL
xenobiotic metabolic process
alcohol metabolic process
medium-chain fatty aldehyde dehydrogenase (NAD+) activity
DNA-binding transcription factor activity
ribonucleoprotein complex
RNA splicing
synapse
nucleotide binding
lysosome
mitochondrion
membrane
fatty acid metabolic process
calcium ion binding
mitochondrial inner membrane
lipid metabolic process
bitter taste receptor activity
hair cycle
chromatin
positive regulation of transcription by RNA polymerase II
G protein-coupled receptor signaling pathway
transcription regulator complex
detection of chemical stimulus involved in sensory perception of bitter taste
regulation of postsynaptic neurotransmitter receptor internalization
intermediate filament
keratin filament
CR1
signal transduction
flavone metabolic process
L2
ventricular septum development
L1
Dong-R4
DNA-binding transcription activator activity, RNA polymerase II-specific
detection of chemical stimulus involved in sensory perception of smell
axonal transport of mitochondrion
glutamatergic synapse
sensory perception of smell
response to cocaine
negative regulation of fatty acid metabolic process
olfactory receptor activity
odorant binding
sequence-specific double-stranded DNA binding
sensory perception of chemical stimulus
serine-type endopeptidase inhibitor activity
G protein-coupled receptor activity
adaptive immune response
B
***
Functional metagroup
Alu
MIR
ERV1
L1
CR1
hAT-Charlie
Gypsy
SVA
ERVK
hAT-Tip100
Dong-R4
L2
ERVL
ERVL-MaLR
Family name
Too general
Transcription
Sensory system
Nervous system
Signaling pathways
RNA processing
Cell cycle
DNA repair
Mitochondrion
Protein degradation
Embryogenesis
Exosomes and vesicles
Lipids metabolism
Immune system
Metals metabolism
Other metabolism
Chromatin structure
Cytoskeleton
Other
Ttranslation
Cell adhesion
Cell migration
Functional group
Class name
LINE
LTR
SINE
DNA
SVA
Helitron
GO terms count
(log scale)
10
0
Functional metagroup
Metabolism
Proteins biogenesis
Cellular processes
RNA biogenesis
DNA biogenesis
Organismal processes
Signaling pathways
other

We classified all the family-level GO terms into the main functional groups, following the same methodology as for the class-level analysis (Figure 6B). Among 22 functional groups (including the “Too general” and “Other” groups), cell adhesion and cell migration were found only once (in MIR and L1 elements, respectively). Protein biogenesis processes (protein degradation and translation) were the next least present group, found in 4 and 3 cases, respectively. In contrast, too general terms, transcription and nervous system were the most frequent ones (46, 18 and 13 instances, respectively). Among all the 22 functional groups and 14 families with GO terms, there was no co-clustering by large-scale functional metagroups (Figure 6B) and only the RNA processing group was significantly overrepresented in Alu elements according to the FDR-adjusted Fisher exact test of a process enrichment in any given family versus all the rest ones.

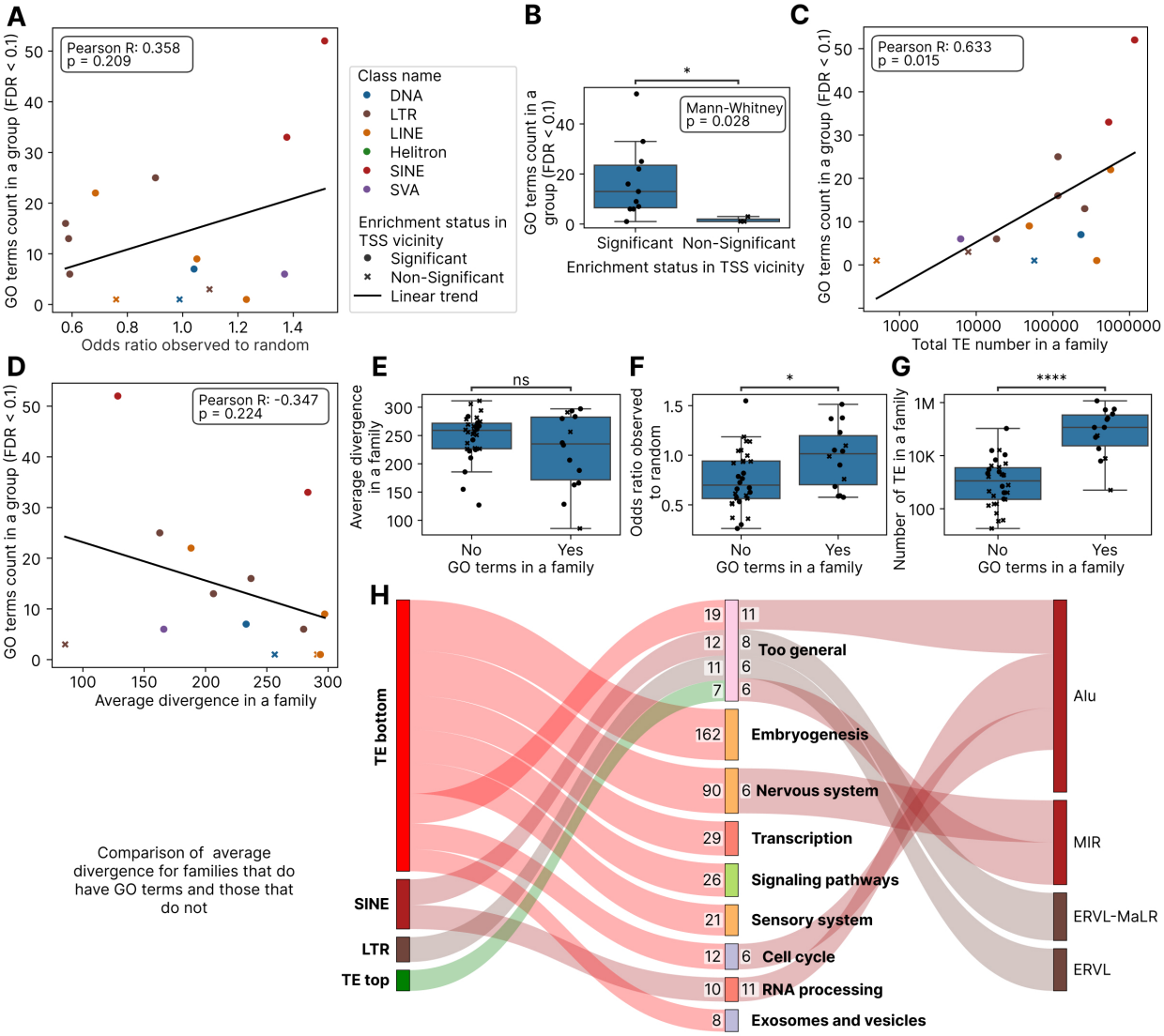
We also compared family-level GO terms with the respective class-level GO terms using a heatmap (Supplementary Figure 8B), applying again the same FDR-corrected Fisher exact test for each functional group in each TE group, where TE group could be a TE family, TE class or top and bottom groups by all TEs. There were 26 functional groups extracted in this analysis (Supplementary File 8). Both RNA processing and DNA repair terms were enriched for Alu elements, whereas metals metabolism and other (specific) terms were enriched for MIR TEs. ERV1 elements had statistically significantly overrepresented lipids and other metabolism groups, whereas ERVL-MaLR elements had general terms enriched, and ERVK endogenous retroviruses are also co-localized with lipid metabolism genes as in the case of ERV1 elements.

We than sought to elucidate the main factors determining the TE families functional impact measured as number of GO terms per family. We found no significant correlation with enrichment level of TEs in the genes TSS vicinity (Figure 7A), but TE families whose enrichment near TSS was significant against the randomized background check showed also higher number of GO terms (Figure 7B). Expectedly, GO terms number strongly depended on total number of elements in a family (Figure 7C, Pearson $r = 0.633$, p-value = 0.015). Additionally, there was no significant correlation of GO terms count with average divergence in a subfamily (Figure 7D, 7E), suggesting that TE families of different evolutionary ages can non-randomly associate with defined functional groups of human genes. Finally, TE families that showed non-zero number of GO terms had significantly higher enrichment level (Figure 7F) and total copy number (Figure 7G).

The overall comparison of functional groups enrichment at the level of classes (with total TE top and bottom groups) with the enrichment at the level of families is depicted in a filtered view in Figure 7H (connection ribbons in the Sankey plot are filtered by >= 5 GO terms per ribbon) and in an unfiltered representation in Supplementary Figure 5C. Cell cycle and RNA processing could be affected by Alu insertions, whereas MIR elements were enriched in the vicinity of nervous system genes. The major groups such as embryogenesis, transcription, nervous system, sensory system and signaling pathways are present mainly in the TE bottom group and neither of TE classes nor families provide more enriched GO terms than the TE bottom group does (Supplementary Figure 5B, these groups show an isolated clustering pattern). In contrast, the groups such as DNA repair, RNA processing, metals metabolism, other (specific) metabolism, cytoskeleton, cell death, DNA recombination and DNA replication do not appear in the TE

bottom group and, despite the highly variable total GO terms count, can be considered as TE-enriched. All the rest functional groups showed inconsistent pattern.

Figure 7. Analysis of main factors impacting TE functional associations with TSS at the level of TE families. (A) Scatter plot of significant GO terms count by OR observed to random ratio. (B) Box plot comparing TE families with significant and non-significant enrichment near TSS by significant GO terms count. (C) Scatter plot of significant GO terms count by total TE number in a family. (D) Scatter plot of significant GO terms count by average divergence in a family. (E) Box plot comparing TE families with and without significant GO terms by average divergence. (F) Box plot comparing TE families with and without significant GO terms by OR observed to random fold change. (G) Box plot comparing TE families with and without significant GO terms by TE count. (H) Sankey plot visualization of GO term groups found in TE classes (left) and families (right). Connecting ribbons were filtered by at least 5 GO terms. This filtering was applied to the visualization only.



3. DISCUSSION

In the present paper, we performed an integrated analysis of human TEs co-insertion with human genes functional groups, based on the recent T2T genome assembly at the level of TE classes and families. Currently dozens of research groups analyzed functional impact of human TEs on the host genes (8,10,11,14,15), all with different methodology and in different cellular and epigenetics contexts, setting the stage for large-scale comparing and review papers to understand the available evidence and put it in a single cohesive network. The current article is meant to establish a baseline for these studies by the co-mapping proximity analysis, since most the epigenetics-based studies are relying on the same proximity principle. Moreover, the availability of the complete human genome assembly (1) allows to do such an analysis with an unprecedented depth and precision, compared to the currently available literature.

3.1. 10 kb TSS neighborhood as an optimal window size based on public literature and the enrichment behavior

In the present study we used a co-mapping window of 10 kb, 5 kb upstream and 5 kb downstream, to select the TEs that are inserted in a vicinity of a given human TSS. This window or the comparable length windows were used previously in a series of functional epigenomics TE studies. A study of IFN-inducible enhancers being spread in human genome by LTR elements utilized a 10 kb window to assess LTRs enrichment by binomial test (9), whereas we used Fisher exact test for the same purpose. A landmark study by (10) that showed enrichment of LINEs and LTRs near duplicated genes and SINEs and DNA elements near singleton genes, utilized two proximity windows: 4 kb and 20 kb ones. In addition, one of the first studies of TE functional and epigenomic impact (11) relied on 5 kb upstream and downstream of coding exons as gene-proximal region for transcription factor (TF) - TE binding colocalization. Another landmark genomic proximity-based investigation of LTR elements functional impact (8) utilized a shifted 6 kb window: the upstream 5 kb and downstream 1 kb for each TSS were selected. Finally, a series of TF and chromatin modifications studies of retroelements performed by our consortium (5,6,12,13,16,17) relied on the same 10 kb window as applied here.

The current analysis validates this approach by highlighting the importance of permutation-based random corrections to account for the variable length of different TE classes. Because LINE elements (averaging 6 kb in full length) occupy a much larger genomic footprint than Alu elements (300 bp), they are statistically more likely to intersect a 10 kb window by chance (18). Conventional statistical methods, such as the Fisher exact test, treat each element as a point-like entity (19), which can lead to the artificial underestimation of the regulatory impact of shorter elements or the overestimation of longer ones. The use of 500 random permutations allows for the derivation of an empirical p-value that correctly identifies biological enrichment over length-driven stochasticity.

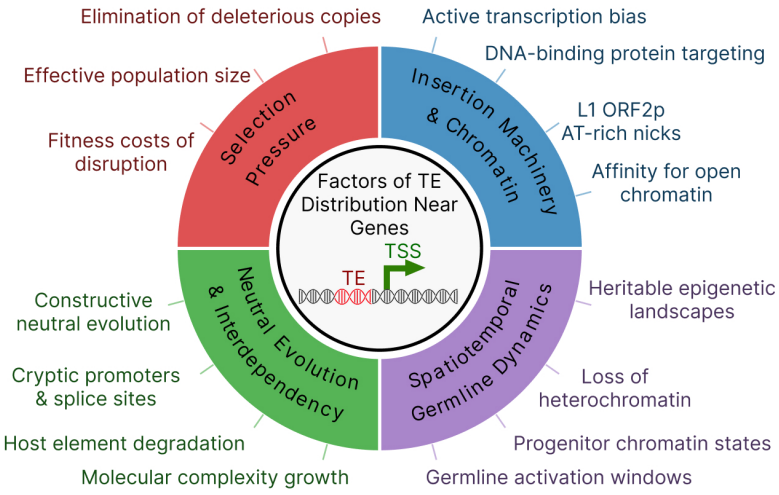
3.2.Enrichment of TE families and classes near human genes TSS

As a necessary preliminary analysis, we studied degree of enrichment of TE classes and families in the 10 kb neighborhood of human gene TSS. The results gained allow us to compare public evidence about human TE proliferation strategies, insertion machineries and preferences, as well as evolutionary conflicts, with the trends reported here based on the complete human genome assembly.

The degree of enrichment or deficiency of TE groups near human genes can be determined by the following factors (Figure 8):

1. TE insertion machinery affinity to open chromatin states, early replicating domains, DNA-binding proteins or actively transcribing genes, for L1 ORF2p which LINEs, SINEs and SVAs are relying upon (20–22). Also, L1 ORF2p prefers AT rich sequences for DNA nicks which further impacts retrotransposons distribution (23).
2. Selection pressure eliminating too deleterious copies, governed by the effective population size at the time of insertion (24,25).
3. Neutral evolutionary mechanisms of complexity and interdependency growth, such as constructive neutral evolution (CNE), rendering the novel insertions functionally indispensable (26,27). If a TE inserts near a gene, it can carry a weak, cryptic promoter or splice site, and this insertion is neutral or nearly neutral. A mutation then could degrade the host's original regulatory element, rendering the host gene functionally dependent on the TE (28). Such a mechanism of TE-driven molecular complexity growth was shown for XIST in macaques (29).
4. Spatiotemporal dynamics of TE activation during the germline development and transcriptional and chromatin states during this stage (30,31).

Figure 8. Schematic representation of mechanisms impacting distribution of TEs with respect to TSS.



The enrichment of SINEs and SVAs near TSS (1.468- and 1.368-fold enrichment over the random expectation) is a consequence of their successful exploitation of the L1 insertion machinery's preference for open chromatin (21,22), a preference that is biologically realized during the hypomethylated window of primordial germ cell (PGC) development (31), specifically for evolutionarily young and active TE families: L1HS (human-specific LINE-1), SVA (subfamilies E and F), and HERV-K (LTR) according to a recent preprint study (30). Conversely, the depletion of LINEs (0.877-fold change versus the random control) is a signature of purifying selection acting against the deleterious effects of large insertions in gene-regulatory domains, such as ectopic recombination and transcriptional interference (32), a force that is partially mitigated but not abolished by the historically low effective population size of humans. On the other hand, an Alu or SVA insertion in a promoter region is less likely to structurally

disrupt the nucleosome landscape or terminate transcription than a 6 kb L1 (33). While they may introduce TFBS, the immediate fitness cost is often low (34).

Our results of SVA elements being enriched near TSS are connected also to the fact that they resist complete repression during the reprogramming window (30) and essential for ZGA by acting as regulatory hubs (SVA D subfamily) (35)
Moreover, SVA elements are the youngest TE class in humans with active evolutionary arms race against the host genome (36). They contain binding sites for key pluripotency factors such as OCT4 and YY1 (37), which creates a potential positive feedback loop: an SVA inserts near a pluripotency gene in a PGC and binds OCT4, which in turn boosts SVA transcription, leading to more transposition. The SVA may act as an enhancer for the nearby gene via the CNE mechanism (38) and SVA elements density near the germline-expressed genes gradually amplifies. Finally, SVA elements contain GC-rich regulatory sequences (VNTRs) that facilitate their retention in gene-rich regions via CNE or exaptation (39).

DNA transposons (e.g., *TcMar*, *hAT*) utilize a "cut-and-paste" transposase mechanism distinct from the L1 machinery. They have been extinct in the human lineage for millions of years (40). Their depletion near TSSs (0.938 of the random expectation) likely reflects both a lack of targeting to these specific open chromatin regions (compared to retrotransposons) and the long-term action of purifying selection over deep evolutionary time.

LTR elements, derived from ancient endogenous retroviruses, are also heavily depleted near TSS (0.667-fold change against the random background). Most LTRs in the human genome are "solo" LTRs that remain after host-mediated recombination between the flanking LTRs of a provirus (41). Because LTRs are potent promoters, their proximity to host genes is a major risk for "onco-exaptation", where the illicit activation of a viral promoter drives the expression of an adjacent oncogene (42). Consequently, LTRs are typically sequestered in heterochromatic regions or subject to strict epigenetic silencing (43). For the full-length 6-8 kb endogenous retroviruses, their insertion in the TSS proximity could be deleterious disrupting promoter-enhancer interactions and perturbing the host authentic TFBS (44).

At the level of TE families, our results reveal a more complex regulatory landscape. Only 7 out of 44 families are significantly enriched near TSS: hAT-Charlie, MIR, CR1, Alu, SVA, TcMar-Tc1 and L2. The enrichment of the L2 family (1.230-fold change) is particularly notable because L2 elements have undergone extensive mutational decay in the last 100 million years, becoming a dominant LINEs in monotremes but surpassing their dominance to L1 in eutherians (45). Eventually they lost their autonomous propagation ability and became "domesticated" as TSS for microRNAs in the host genes 3′ UTRs, particularly in the brain where they are a major source of microRNAs (46,47). They are also coopted as enhancers in tissue-shared compared to tissue-specific genes, highly enriched in H3K27ac and H3K4me1 marks, emphasizing their transition from genetic parasites to functional regulatory modules via exaptation or CNE (48). Our findings suggest that such functional interdependency could save L2 elements from deletion in the gene-proximal regulatory active genomic regions.

The CR1 family (Chicken repeats, LINE) also shows significant enrichment (1.051-fold change), with surprising functional connections to neurotransmitter processes and postsynaptic receptor

internalization. This suggests that specific waves of LINE insertions, long before the dominance of L1, contributed to the foundational regulatory networks of the amniotes brain and these insertions could be remnants of those events (45).

Whereas SINEs showed 1.468-fold enrichment near TSS, their major families, Alu and MIR, exhibited 1.513- and 1.377-fold enrichment relative to random OR. Alus are considered "proto enhancers" that evolve into functional regulatory elements over time (49). They are specifically enriched in active chromatin marks such as H3K36me3 in the colon and brain (50), and they frequently accumulate TFBS that rewire host networks (51). MIR elements were shown here as strongly associated with nervous system genes and voltage-gated potassium channels. MIRs are uniquely involved in the naïve pluripotent state, where they are co-opted by ESRRB to build networks of enhancers and super-enhancers regulating pluripotency (52).

Finally, we observed no connection of family enrichment near TSS with their divergence for any TE class including LINEs, albeit the comparison of newly integrated, polymorphic and human specific L1 insertions (53) suggested that L1 are integrating in gene rich but are persisting in a genome for long term in gene poor regions. Nevertheless, our current analysis is concentrated on the evolutionary ancient LINEs with peak divergence of 270 corresponding to more than 100 million years of evolutionary age (7).

3.3.Functional groups enriched and deficient in TE insertions by classes count, divergence and families

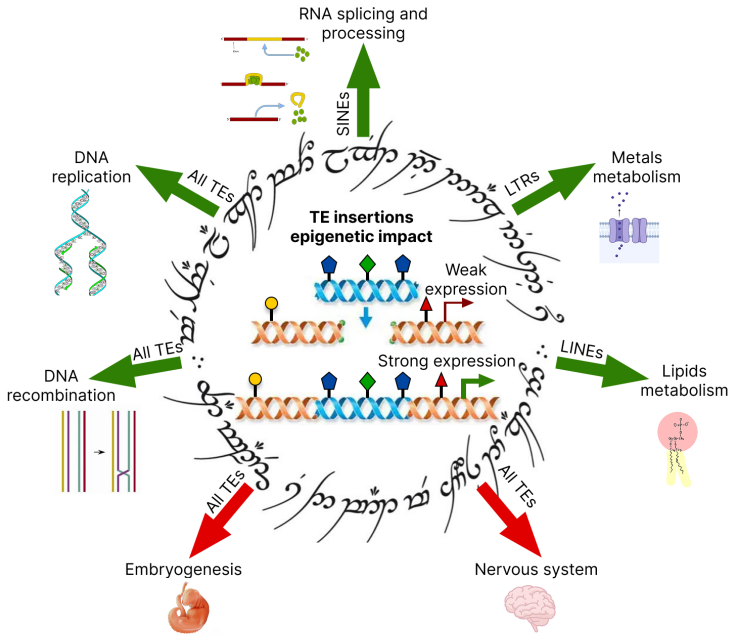
The analysis reported here shows that 99.11% of unique TSS are within 10 kb of at least one TE, suggesting that nearly every human gene is potentially subject to TE-mediated regulatory influence. This finding is particularly important when compared to prior databases like TE-TSS, which identified 5,768 TE-derived TSSs - roughly 25% of the human promoter landscape - using older assemblies (54). The shift from 25% "TE-derived" to 99.11% "TE-proximal" indicates that while TEs may not always serve as the primary TSS, their ubiquity in the immediate neighborhood allows them to act as auxiliary regulators, insulators, or chromatin modifiers (40).

On average, according to our results, each human gene harbors 15.05 TEs in its TSS neighborhood, with SINEs being the most frequent (7.82 per gene), followed by LINEs (4.39), DNA elements (1.49) and LTRs (1.32). The identification of *CIB3* as a genomic outlier, containing 42 TEs of various classes, points toward the existence of extreme repeat accumulation zones. *CIB3* (calcium and integrin binding protein 3) is integral to the mechano-electrical transduction (MET) apparatus in the inner ear, forming heteromeric complexes with TMC1 and TMC2 to stabilize cation channels (55). The high repeat density surrounding such a specialized sensory gene raises questions about whether these TEs provide modular regulatory controls for high-precision environmental sensing or if the locus simply resides in a genomic region with relaxed purifying selection.

In general, the functional characterization of genes based on their TE content reveals a clear segregation of biological processes (Figure 4B). This "Ring of Power" functional network

illustrates how the host genome manages the risks (deleterious mutations) and rewards (power of innovation) of TE proximity (Figure 9).

Figure 9. Schematic representation of molecular processes groups that were found as significantly enriched (green arrows) or depleted (red arrows) with individual TE classes or all TEs. The full comparison including the insignificant groups is shown in Figure 4B.



Genes involved in embryogenesis, transcription, and nervous system development are consistently found in the "TE bottom" group - those with the lowest number of TE insertions. This result strongly supports the "robustness hypothesis", which posits that core biological pathways must be protected from the regulatory noise introduced by TEs and TEs are purged by negative selection in these pathways. Embryogenesis is a process that requires extreme transcriptional precision; even minor disruptions to the timing or level of gene expression can be lethal (35), albeit TEs largely contribute to zygotic genome activation (56). The rest two large groups, sensory system and signaling pathways, were present mainly in the TE bottom group and neither of TE classes nor families provided more enriched GO terms than the TE bottom group did (Figure 4B).

Conversely, genes enriched with TEs (the "TE top" group) are involved in RNA splicing, DNA repair, telomere maintenance, and immune signaling. These processes represent the "innovation laboratory" where TEs are actively co-opted to increase transcriptome and proteome complexity (57).

- **RNA splicing:** Alu elements were early shown to profoundly impact RNA splicing by introducing new splice sites (exonization), acting as alternative exon sinks, and forming double-stranded structures that alter exon inclusion (58). We now show that Alu copies are significantly enriched near the host genes related to splicing, which is a completely orthogonal mean of Alu-mediated impact on splicing.

- **DNA repair:** The enrichment of TEs (particularly Alu elements) near DNA repair genes (double strand breaks and mismatch repair) could suggest a synergistic relationship where the more robust responses are evolving to DNA lesions induced by insertions of these elements (59).
- **Olfactory receptors:** L1 elements are specifically enriched near olfactory receptor genes, which could allow rapid diversification of sensory perception of smell sensing through TE-mediated rearrangements, leading to genes duplication and subfunctionalization or gene losses (60,61).

Other groups: metals metabolism, other (specific) metabolism, cytoskeleton, cell death, DNA recombination and DNA replication did not appear in the TE bottom group and, despite the highly variable total GO terms count or non-significant Fisher exact test results, can be considered as TE-enriched.

At the level of TE classes, the previous landmark study by (14) utilized 20 kb and 2 kb window to show non-random association of SINEs with housekeeping genes having "broad" promoters with multiple TSS and LINEs with tissue-specific ones having "sharp" promoters with single TSS (62). Precisely, in that study SINE-enriched genes were associated with ribosome, translation, RNA processing, nucleolus and protein transport, whereas LINE elements were enriched in genes of olfactory receptors, retinol metabolism, epoxygenase P450 pathway and immunoglobulin domains. Our analysis partly confirms these findings (Figure 4A, 4B). Similarly, we show that RNA processing, ribosome and translation, transcription and general pathways are enriched in SINEs. But such highly cell-type specific processes as embryogenesis and sensory system were also present as a minor part of SINE elements. For LINEs we confirmed olfactory receptors, whereas lipid metabolism contained different terms: flavone metabolism, negative regulation of fatty acids metabolism. Also, for L1 elements we report TF activity and calcium ion binding. These partial differences can be explained by the methodological differences: (14) profiled TE abundance separately in Promoter, Intron, Downstream, 5′ UTR, CDS and 3′ UTR regions of each gene, then performed quantile normalization and hierarchical clustering, analyzing gene groups arising from this clustering. Our approach is simpler and more interpretable, albeit the clustering can be more robust, so the comparison of both methods on the same T2T RepeatMasker dataset could be helpful. It should be also noted that the fact that genes of different functional profiles have different average number of TSS (62) can bias our analysis and requires further investigation.

At the level of families, Alu elements are of particular interest. In our analysis they were found near 90.4% of all human TSS, with up to 29 copies in a single 10 kb window. One of the first functional TE mapping study showed that Alus are preferentially inserted near metabolism, transport and signaling genes, whereas structural and information processing genes were depleted of Alu elements (63). Here we show the contrasting results that Alu are enriched near RNA processing genes, potentially affecting also genes of meiosis, DNA damage response, inner mitochondrial membrane and transcription initiation and elongation (Figure 6A). These drastic differences can be explained by the fact that (63) used chromosome 21 and 22 in their analyses, whereas here the complete human genome assembly is used.

Alus enrichment near genes of RNA processing is corroborated by the recent evidence: Alu elements often provide cryptic splice sites that lead to exonization and exon skipping in primates

in a lineage specific manner (64). This process is regulated by hnRNP C (33) and the Ku70/80 heterodimer (the latter one according to the recent preprint articles (65,66)), which compete with splicing factors to prevent the deleterious over-inclusion of Alu sequences while allowing for adaptive variations in the proteome. Furthermore, the presence of Alus near DNA repair genes suggests their role as "editing inducer elements", where inverted Alu repeats could facilitate the A-to-I editing of host transcripts, thereby fine-tuning enzymatic function in a primate-specific manner (33).

The MIR (Mammalian-wide Interspersed Repeat) family, though less numerous than Alus, showed strong functional preference for the nervous system (Figure 6B). MIR elements are significantly enriched near genes for voltage-gated potassium channels, which are the largest and most diverse ion channel family in the human genome (67). Potassium channels are essential for returning the cell to a resting state after an action potential (68). The association of MIR elements with these channels suggests that MIR-derived enhancers may contribute to the complex rules of subcellular localization and firing frequency that differentiate mammalian neurons (69). Furthermore, MIR elements are linked to macrophage activation and the sensing of metal ions (Cd, Cu, Zn) (Figure 6A), indicating they could also provide regulatory modules for ancient immune and metabolic responses.

SVA elements were previously shown to impact host genes via premature termination based on their internal polyadenylation signal (70) and they are now shown to enrich near genes related to termination of RNA polymerase II transcription, precisely the core catalytic RNA polymerase II subunit and its C-terminal domain phosphatases (3 SVA B copies in the 116 kb region of chromosome 11). A previous genomic proximity-based study revealed SVA enrichment near zinc finger clusters (39), albeit the methodology was different: (39) used hg19 human genome and 1Mb bins for enrichment testing instead of 10 kb and T2T genome in the current analysis. SVA elements are also known to facilitate 3' transduction events when their internal polyadenylation signal is bypassed by Pol II (71). The proximity of SVA elements to the genes that regulate Pol II termination suggests a potential co-evolutionary loop where SVAs have been integrated into the feedback mechanisms that control transcriptional processivity. This could represent a mechanism of CNE, where the host genome becomes functionally dependent on the presence of a TE-derived regulatory element for the accurate execution of core transcriptional cycles (26). Despite recent studies have identified human-specific subfamilies, such as SVA_F1, which are active in the human population and frequently mobilize adjacent gene sequences through transduction (71) – the RNA polymerase termination genes are enriched with SVA_B subfamily, so their SVA regulatory impact is likely not a recent evolutionary innovation.

Among the TE classes, only LTR families all demonstrated functional associations with human molecular processes (Supplementary Figure 8A, Figure 6A). Interestingly, Gypsy elements which belong to the chromoviruses group targeting themselves into heterochromatin regions via chromodomain (72), are strongly co-associated with translation activator genes according to the current analysis (Figure 6A).

DNA transposons, which have been extinct in the human lineage for millions of years, are mostly functionally neutral (Figure 6). However, the hAT-Charlie family shows a significant association with the MHC class I pathway (Figure 6A), suggesting that ancient DNA transposon

insertions contributed to the organization of the vertebrate adaptive immune system before their eventual inactivation.

It is important to note than the fact that certain TE families, precisely 30 out of 44 ones (Helitrons, 20 DNA families, 4 SINE ones, 5 LINE families and no LTR ones) did not result in significant functional host gene groups by GO analysis, suggests that their presence could be largely neutral - a result of random drift and tolerance rather than adaptive recruitment.

At the level of lowest and highest divergence by TE classes, ancient high-divergence SINEs (particularly the MIR family) and high-divergence LINEs (CR1, L1, Dong-R4 family) were strongly associated with the nervous system according to our results (Figures 5-6). Genes with high average TE divergence are enriched for processes such as postsynaptic membrane organization, synaptic chemical transmission, and voltage-gated potassium channels (Figure 5A). This pattern suggests that the foundational architecture of the mammalian brain was shaped by ancestral waves of TE activity that have since been stabilized and domesticated into the host's regulatory framework (73). High-divergence SINEs are particularly prominent near ion channel genes, which regulate the electrical excitability of neurons (68). The stabilization of these elements over hundreds of millions of years suggests they have evolved from genetic parasites into essential cis-regulatory modules that buffer the transcriptional output of housekeeping genes or provide tissue-specific enhancer logic.

3.4.Interferon alpha region of low LINEs divergence

One of the most biologically significant findings that we can report is the identification of a 220 kb patch on chromosome 9 containing a cluster of interferon alpha genes (*IFNA10, IFNA16, IFNA17, IFNA21, IFNA4, IFNA6, IFNA7, IFNW1*) that is uniquely enriched with young, low-divergence L1 elements. There are 12 RefSeq curated genes in this area according to the T2T genome assembly, which is significantly higher than an average human genome value of 12.5 genes per 1 Mb (74).

The enrichment of young LINEs in this region suggests an ongoing evolutionary arms race. Type I interferons are the first responders to viral infection (75), and their signaling is known to be heavily influenced by TE-derived promoters, enhancers and TE exonization (76,77). Additionally, interferon I signaling is regulated by L1 elements repression via HUSH complex which detects L1-derived dsRNAs (78). The presence of low-divergence elements indicates that these insertions are relatively recent and may be contributing to the diversification of the human antiviral response. Moreover, Alu and L1 elements can accumulate in the pre-existing loci already occupied by copies of themselves, creating a feed-forward loops of the progressive expansion in limited genome regions (79). This finding has profound implications for understanding autoimmune diseases like systemic lupus erythematosus (SLE), which is characterized by a type I interferon signature (80). The de-repression of TEs in these regions, perhaps due to environmental stressors or aging-related heterochromatin loss as was shown earlier (81), can lead to the formation of double-stranded RNA (dsRNA) that triggers the cGAS-STING pathway (82) and eventually a constitutive immune response that leads to chronic inflammation (83).

3.5.Connection of TE enrichment with cancer

The recent evidence points out high transpositional activity in cancer, with more than 500 L1 insertions per tumor in bladder cancer according to a recent preprint (84). Moreover, large-scale TE de-repression is an emerging hallmark of cancer. Our current work defines the scope of human molecular processes that can be aberrantly activated in cancer by distinct TE classes and families and potentially targeted by anticancer therapy (85).
The contemporary anticancer treatment heavily relies on the single molecule based biomarkers (86,87) while there is an increasing need for development of the compound ones, reflecting tumor genome and microenvironment distinct states (88–90) as well as plasma exosomes impact (91,92). Among the promising ones is gene signatures approach which measures activation of multiple genes simultaneously (93–95). Genes identified as enriched by certain TE classes and families can be used to compose gene expression signatures and measured in the publicly available cancer cohorts such as TCGA (96). In turn, such an evolutionary approach could improve cancer treatments outcomes as cancer is primarily an evolutionary disease arising from the genomic conflict of interests (97).

4. MATERIALS AND METHODS

4.1.Human TEs

Coordinates and class/family annotations for human TEs were obtained from the RepeatMasker track (98) based on the T2T human genome assembly, using the Table Browser tool (Group: *Variation and Repeats*, Track: *RepeatMasker*, Table: *T2T RepeatMasker*). The T2T RepeatMasker annotation itself was derived from(99), and average divergence scores (number of substitutions per 1000 base pairs) were extracted from the same UCSC RepeatMasker track.
Each TE was categorized by RepeatMasker into hierarchical levels of classification (98):

- **Class**, representing the highest level, defined by the mechanism of transposition.
- **Family**, representing an intermediate level, grouping elements of shared evolutionary origin that typically exhibit similar structural features and transposition mechanisms (e.g., Alu and L1 elements).

In this study, we adopted the RepeatMasker classification as provided in the T2T genome assembly without modification, ensuring consistency with the widely accepted TE annotation framework.
The dataset comprised elements assigned to the following classes: LINE, SINE, LTR, SVA (annotated as “Retroposon” in the source table), Helitron (denoted as “RC” for rolling-circle elements), and DNA transposons. In total, the dataset included 3,709,429 entries: 1,706,485 SINEs, 1,005,214 LINEs, 531,410 LTRs, 458,177 DNA elements, 6,274 SVAs, and 1,869 Helitrons.
Although SVA elements are evolutionarily and functionally related to Alu elements (100), SINEs and SVAs were treated as distinct classes in accordance with RepeatMasker nomenclature. Likewise, Helitrons were considered as a separate TE class, despite their mechanistic classification as a subset of DNA transposons due to their rolling-circle replication mechanism (101).

4.2.Human genes and TSS

Human genes coordinates were downloaded from UCSC Genome Browser (102) via the *Table Browser* tool, using the January 2022 assembly (T2T CHM13v2.0/hs1). Data were extracted from the group *Genes and Gene Predictions*, track *NCBI RefSeq*, and table *RefSeq All*. Only curated gene entries were retained by filtering for accessions beginning with prefixes NM, NR, NP, or YP. Transcription start site (TSS) coordinates were defined as the leftmost exon start (5′ UTR exon) for genes on the positive strand and the rightmost exon end for genes on the negative strand. The 10 kb promoter regions (TSS neighborhoods) were then constructed by extending 5 kb upstream and downstream of each TSS.
HUGO gene symbols (column *geneName2*) were used as primary identifiers, yielding 28,738 unique genes. Transcript isoforms corresponding to the same gene but differing in TSS position were treated as distinct entities during epigenomic profiling and RE mapping. The complete set of TSS 10 kb neighborhoods used in the mapping is provided in Supplementary File 1 (38,704 entries including isoforms).

4.3.Mapping of TEs on gene TSS 10 kb neighborhoods

The mapping of TE coordinates with their divergence onto gene TSS 10 kb neighborhoods was done using bedtools (bedtools map utility, version 2.31.1) (103), any TEs that have been intersecting with the 10 kb interval for a given gene were taken into the analysis. This could lead to artificial underestimation of enrichment and regulatory impact of longer TE classes (LINEs and LTR elements), so the appropriate random permutation controls were applied as described later.

4.4.Random control of TEs enrichment

To control TE enrichment near genes against a random background, we performed 1000 random permutations of TEs coordinates by the bedtools shuffle command, selecting random state number (seed) from 1 to 1000. An empirical p-value was calculated as fraction of random odds ratios (OR) that were above or below the observed one, depending on whether the observed OR was below or above the median across the random 1000 OR values. Then the empirical p-value was lower clipped by 2 / (N + 1), where N was the permutations number (1000).
An enrichment score was calculated as a fold change of observed versus random ORs.

4.5.Statistical tests

Group comparisons were done by Mann-Whitney U-test (104) implemented in scipy python library (105), version 1.17.1. Multiple hypotheses were corrected by the False Discovery Rate approach (106) implemented in the python statsmodels library (107), version 0.14.6. OR and p-values for TE classes and families' enrichment in the TSS 10 kb of genes, as well as enrichments of GO terms in TE groups by functional group were calculated using Fisher exact test (108) in scipy. Linear correlation coefficients and their p-values were calculated by Pearson (109). Distributions were compared by Kolmogorov-Smirnov test (110) implemented in scipy. FDR-corrected p-value 0.05 was selected as a threshold value for Mann-Whitney, Fisher exact and Kolmogorov-Smirnov tests for multiple hypotheses.

4.6.Gene Ontology and other functional annotations

TSS for isoforms of the same gene were considered as distinct unique entities, with enrichment, divergence and epigenomic analyses performed independently for them. For the Gene Ontology (GO) analysis gene names were deduplicated among the TSS that were extracted as enriched by a certain criterium. GO analysis was performed using goatools python library (Klopfenstein et al. 2018), version 1.6.4 on a local database downloaded on Dec 31st, 2025. FDR-corrected Fisher exact test p-value 0.1 was selected as a threshold value for GO terms.

4.7.Visualization

Plots for this article were drawn using matplotlib version 3.10.8 (112), plotly version 6.7.0 (113), statannotations version 0.7.2 (114) and seaborn version 0.13.2 (115) libraries in Python 3.11. Supervenn plots were built using the supervenn python library (GitHub - gecko984/supervenn, n.d.), version 0.5.0. Network visualizations were built using network version 3.6.1 (Hagberg et al. 2008) and pyvis version 0.3.2 (118) libraries in python. UCSC Genome Browser tracks were visualized using the Genome Browser web portal (119).

4.8.Data analysis

Data analysis was performed in a Jupyter Notebook environment using pandas version 3.0.2 (Pandas team, n.d.) and numpy version 2.4.4 (121) for tabular data handling and analysis.

4.9.AI usage

Gemini PRO (122) was used for code refining, assistance with literature search and pre-classification of GO terms into large biological groups. Chat GPT was used for grammar corrections of the manuscript (123).

5. ETHICAL STATEMENT

This study represents a purely computational analysis based on publicly available genomic datasets. All data analyzed were obtained from the T2T genome assembly and the RepeatMasker track (2) accessible through public repositories. As no new biological material was collected and the study did not involve human participants, animal subjects, or experimental interventions, ethical approval and informed consent were not required. All data retrieval and analysis were conducted in accordance with established ethical standards for the use of secondary biological data.

6. CONFLICT OF INTEREST

The author declares no conflict of interest.

7. AUTHORS CONTRIBUTION

Daniil Nikitin: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology development, project administration, resource provision, software development, supervision, validation, visualization, and writing (original draft, review, and editing).

As the sole author, I fulfill all established criteria for authorship. I made substantial contributions to the conception and design of the study, as well as to the acquisition, analysis, and interpretation of the data. I drafted the manuscript and critically revised it for important intellectual content, approved the final version for publication, and agree to take full responsibility for all aspects of the work.

8. ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the faculty and staff of the Department of Molecular Biology at Moscow State University. The academic environment, resources, and support I received during my bachelor's and master's studies were instrumental in shaping my early scientific trajectory and fostering my commitment to bioinformatics and evolutionary biology. I am especially indebted to Professor Galina Belyakova, whose mentorship played a pivotal role in my preparation for the 2013 International Biology Olympiad in Bern, where I was awarded a gold medal. Her guidance, together with that of my other Olympiad mentors, was fundamental in cultivating my interest in molecular and evolutionary biology well before the completion of my formal university training.

I am also grateful to my colleagues, particularly my team lead, Katerina Nuzhdina, and my fellow bioinformaticians. Their collegial support and the intellectually stimulating environment they created have been a continuous source of inspiration and motivation. Their shared expertise and encouragement significantly contributed to my research progress. I would like to extend special thanks to my colleagues Polina Oshchepkova and Anastasiya Yudina for their invaluable intellectual contributions and steadfast emotional support throughout this work.

I am deeply thankful to my parents, Olga Nikitina and Michael Nikitin, for fostering an environment of curiosity, learning, and exploration. Their influence instilled in me a lasting fascination with the diversity of nature and the universe, which ultimately shaped my enduring interest in evolutionary biology, particularly genome evolution.

My deepest and most heartfelt appreciation goes to my wife, Irina Nikitina. Her unwavering support, encouragement, and belief in me transformed a collection of study efforts into the ambitious pursuit of a full-scale research program. Her presence provided stability and strength during challenging periods, including times of illness, injury, and the complexities of relocating to Armenia with our young sons.

I apologize to colleagues and researchers whose relevant work may not have been cited in this manuscript.

This research was conducted without specific funding from public, commercial, or non-profit funding agencies.

9. SUPPLEMENTARY MATERIAL

Supplementary Figure 1. Ridge plots for length distribution comparison between all (blue) and TSS neighborhoods mapped TEs (red) (A) for all classes and (B) for individual classes.
Supplementary Figure 2. Ridge plots for divergence distribution comparison between all (pale colors) and TSS neighborhoods mapped TEs (full colors) for individual families. The distributions are colored by TE class. For all comparisons significance is assessed by the Mann–Whitney test and FDR-corrected: ns, $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $0.0001 < p < 0.001$; ****, $p < 0.0001$.
Supplementary Figure 3. Ridge plots for length distribution comparison between all (pale colors) and TSS neighborhoods mapped TEs (full colors) for individual families. For each family, top and bottom 2.5% of points were clipped to ensure visual capture of the differences. The distributions are colored by TE class. For all comparisons significance is assessed by the Mann–Whitney test and FDR-corrected: ns, $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $0.0001 < p < 0.001$; ****, $p < 0.0001$.
Supplementary Figure 4. Supervenn plots for gene set intersections. Each gene set is divided into gene groups that differ by gene sets sharing genes from a given gene group. For each panel, colored rectangles in the main plot show these gene groups, top grey bar plot indicates number of set that each gene group shares, lower number show gene count in each gene group. Right side bar plot visualizes gene counts per gene set. (A) Intersections of gene sets: enriched with all TEs (TE top), depleted with all TEs (TE bottom) and enriched by each of TE classes. (B) Intersections of gene sets with all TEs, LINEs, LTRs, DNA elements and SINEs with highest and lowest divergence.
Supplementary Figure 5. TSS distributions by count (A) and divergence (B) of mapped TEs, with red bars denoting those TSS whose genes were taken into the Gene Ontology analysis. Pale blue histograms show all TEs. On both panels distributions are shown for individual classes and all TEs. For the divergence panel (B) both highest and lowest divergence groups are shown.
Supplementary Figure 6. UCSC Genome Browser visualization of genes and repeats in the interferon alpha domain of chromosome 9 with coordinates 21150692 to 21370055.
Supplementary Figure 7. Log-scaled distributions of TSS by TE number mapped on their 10 kb neighborhood. The distributions are plotted for individual TE families and colored by their class. Numbers on the right show TSS counts and percentages with non-zero TEs in their vicinity.
Supplementary Figure 8. Genes and molecular processes enriched with TE of distinct families in their 10 kb vicinity. (A) Intersections map of genes sets enriched with TEs by family. Circle size is log-proportional to number of genes in a set. Circles are colored by TE class, and color intensity denotes OR observed to random. TE families with significant enrichment near TSS are marked as bold. Connection line width is proportional to Jaccard index between a two gene sets. (B) Cluster map of GO terms number by functional groups and TE groups: TE top, TE bottom, the four classes with significant GO terms (LINEs, SINEs, SVA elements and LTRs) and TE families with significant GO terms. Colors on the cluster map side annotation denote TE groups and functional metagroups. For all clustermap cells significance is assessed by the Fisher exact test of a given TE group and a given functional group agains the same functional group in the

rest TE groups and FDR-corrected: ns, $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $0.0001 < p < 0.001$; ****, $p < 0.0001$. (C) Sankey plot of GO terms count comparison by groups between the large-scale TE groups (the top one, the bottom one and the TE classes) and TE families.

Supplementary File 1. Genomic coordinates of human TSSs and associated TEs. For each of the 38,704 TSSs, the corresponding gene name is provided, along with lists of overlapping TE classes, families, subfamilies, and their divergence values.

Supplementary File 2. Enrichment statistics of TE subfamilies within 10 kb regions surrounding TSSs.

Supplementary File 3. Genes exhibiting enrichment or depletion of TEs in their vicinity, categorized by major TE groups, including individual TE classes and all TEs combined.

Supplementary File 4. GO terms, associated genes, and functional group classifications based on TE enrichment categories (TE classes, all TEs enriched, and TE-depleted groups).

Supplementary File 5. Genes enriched in TE classes and in all TEs, stratified by high and low divergence levels.

Supplementary File 6. GO terms, associated genes, and functional group classifications for TE groups stratified by divergence (low vs. high), including both TE classes and all TEs.

Supplementary File 7. Genes enriched in specific TE families based on TE counts in their genomic vicinity.

Supplementary File 8. GO terms, associated genes, and functional group classifications for TE family-level analyses.

10. DATA AVAILABILITY

All code used for the comprehensive proximity mapping, statistical analysis, and GO functional networking is available in the GitHub repository: https://github.com/Nikit357/T2T_transposons_genes.

Tables of TE-gene intersections, including divergence and family-level enrichment statistics, have been deposited in the same GitHub repository.

11. REFERENCES

1. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al. The complete sequence of a human genome. Science (1979). 2022 Apr 1;376(6588):44–53. doi:10.1126/SCIENCE.ABJ6987;ISSUE:ISSUE:DOI PubMed PMID: 35357919.
2. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science (1979). 2022 Apr 1;376(6588). doi:10.1126/SCIENCE.ABK3112;PAGEGROUP:STRING:PUBLICATION PubMed PMID: 35357925.

3. Nikitin D. Retroelements-driven regulatory evolution of human genes and molecular processes: analysis of genome binding profiles of transcription factors and histone modifications [Internet]. doi:10.5281/ZENODO.19052416
4. Betancourt AJ, Wei KHC, Huang Y, Lee YCG. Causes and Consequences of Varying Transposable Element Activity: An Evolutionary Perspective. Annu Rev Genomics Hum Genet. 2024 Aug 27;25(1):1. doi:10.1146/ANNUREV-GENOM-120822-105708 PubMed PMID: 38603565.
5. Nikitin D, Kolosov N, Murzina A, Pats K, Zamyatin A, Tkachev V, et al. Retroelement-Linked H3K4me1 Histone Tags Uncover Regulatory Evolution Trends of Gene Enhancers and Feature Quickly Evolving Molecular Processes in Human Physiology. Cells 2019, Vol 8,. 2019 Oct 8;8(10). doi:10.3390/CELLS8101219 PubMed PMID: 31597351.
6. Nikitin D, Penzar D, Garazha A, Sorokin M, Tkachev V, Borisov N, et al. Profiling of Human Molecular Pathways Affected by Retrotransposons at the Level of Regulation by Transcription Factor Proteins. Front Immunol. 2018 Jan 30;9(JAN). doi:10.3389/FIMMU.2018.00030 PubMed PMID: 29441061.
7. Kosuge M, Ito J, Hamada M. Landscape of evolutionary arms races between transposable elements and KRAB-ZFP family. Sci Rep. 2024 Dec 1;14(1). doi:10.1038/S41598-024-73752-7 PubMed PMID: 39375372.
8. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet. 2017 Jul 1;13(7):e1006883. doi:10.1371/JOURNAL.PGEN.1006883 PubMed PMID: 28700586.
9. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016 Mar 4;351(6277):1083. doi:10.1126/SCIENCE.AAD5497 PubMed PMID: 26941318.
10. Correa M, Lerat E, Birmelé E, Samson F, Bouillon B, Normand K, et al. The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality. Genome Biol Evol. 2021 May 1;13(5). doi:10.1093/GBE/EVAB062 PubMed PMID: 33973013.
11. Bourque G, Leong B, Vega VB, Chen X, Yen LL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008 Nov;18(11):1752. doi:10.1101/GR.080663.108 PubMed PMID: 18682548.
12. Nikitin D, Garazha A, Sorokin M, Penzar D, Tkachev V, Markov A, et al. Retroelement-Linked Transcription Factor Binding Patterns Point to Quickly Developing Molecular Pathways in Human Evolution. Cells. 2019 Feb 6;8(2). doi:10.3390/CELLS8020130 PubMed PMID: 30736359.
13. Igolkina AA, Zinkevich A, Karandasheva KO, Popov AA, Selifanova M V., Nikolaeva D, et al. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. Cells. 2019 Sep 1;8(9). doi:10.3390/CELLS8091034 PubMed PMID: 31491936.
14. Lu JY, Shao W, Chang L, Yin Y, Li T, Zhang H, et al. Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. Cell Rep. 2020 Mar 10;30(10):3296. doi:10.1016/J.CELREP.2020.02.048 PubMed PMID: 32160538.
15. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014 Dec 1;24(12):1963–76. doi:10.1101/GR.168872.113 PubMed PMID: 25319995.

16. Nikitin D, Sorokin M, Tkachev V, Garazha A, Markov A, Buzdin A. RetroSpect, a New Method of Measuring Gene Regulatory Evolution Rates Using Co-mapping of Genomic Functional Features with Transposable Elements. Evolution, Origin of Life, Concepts and Methods. 2019 Oct 1;85–111. doi:10.1007/978-3-030-30363-1_5
17. Nikitin D. Joint analysis of human retroelements-linked histone modification profiles reveals quickly evolving molecular processes connected with cancer [Internet]. 2025 Sep 27. doi:10.1101/2025.09.24.677146
18. Levin HL, Lee SP, Anand A. The 5′ truncation of retrotransposon L1: a process of genome integrity. Genetics. 2025 Dec 10;231(4). doi:10.1093/GENETICS/IYAF202 PubMed PMID: 41159775.
19. Kanduri C, Bock C, Gundersen S, Hovig E, Sandve GK. Colocalization analyses of genomic elements: approaches, recommendations and challenges. Bioinformatics. 2018 May 1;35(9):1615. doi:10.1093/bioinformatics/bty835 PubMed PMID: 30307532.
20. Levin HL, Moran J V. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 2011 Sep;12(9):615. doi:10.1038/NRG3030 PubMed PMID: 21850042.
21. Cost GJ, Golding A, Schlissel MS, Boeke JD. Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res. 2001 Jan 15;29(2):573. doi:10.1093/NAR/29.2.573 PubMed PMID: 11139628.
22. Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, et al. Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. Cell. 2019 May 2;177(4):837-851.e28. doi:10.1016/j.cell.2019.02.050 PubMed PMID: 30955886.
23. Lavie L, Maldener E, Brouha B, Meese EU, Mayer J. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res. 2004 Nov;14(11):2253. doi:10.1101/GR.2745804 PubMed PMID: 15520289.
24. Rishishwar L, Wang L, Clayton EA, Mariño-Ramírez L, McDonald JF, Jordan IK. Population and clinical genetics of human transposable elements in the (post) genomic era. Mob Genet Elements. 2017 Jan 2;7(1):1. doi:10.1080/2159256X.2017.1280116 PubMed PMID: 28228978.
25. Marino A, Debaecker G, Fiston-Lavier AS, Haudry A, Nabholz B. Effective population size does not explain long-term variation in genome size and transposable element content in animals. Elife. 2024 Sep 11;13. doi:10.7554/ELIFE.100574.1
26. Muñoz-Gómez SA, Bilolikar G, Wideman JG, Geiler-Samerotte K. Constructive Neutral Evolution 20 Years Later. J Mol Evol. 2021 Apr 1;89(3):172. doi:10.1007/S00239-021-09996-Y PubMed PMID: 33604782.
27. Catherall-Ostler AM, Dixit T. The Constructive Neutral Evolution of Behaviour. Ecol Evol. 2025 Jul 1;15(7):e71736. doi:10.1002/ECE3.71736 PubMed PMID: 40641486.
28. Pirogov SA, Maksimenko OG, Georgiev PG. Transposable Elements in the Evolution of Gene Regulatory Networks. Russian Journal of Genetics 2019 55:1. 2019 Apr 15;55(1):24–34. doi:10.1134/S1022795419010113
29. Cazottes E, Alfeghaly C, Rognard C, Necsulea A, Loda A, Castel G, et al. Remodeling of XIST regulatory landscape during primate evolution. Sci Adv. 2026 Jan 16;12(3):eadw5839. doi:10.1126/SCIADV.ADW5839;ISSUE:ISSUE:DOI PubMed PMID: 41544163.

30. Dietmann S, Keogh MJ, Tang W, Magnusdottir E, Kobayashi T, Chinnery PF, et al. Transposable elements resistant to epigenetic resetting in the human germline are epigenetic hotspots for development and disease. bioRxiv. 2020 Mar 20;2020.03.19.998930. doi:10.1101/2020.03.19.998930
31. Maupetit-Mehouas S, Vaury C. Transposon Reactivation in the Germline May Be Useful for Both Transposons and Their Host Genomes. Cells. 2020 May 8;9(5):1172. doi:10.3390/CELLS9051172 PubMed PMID: 32397241.
32. Graham T, Boissinot S. The Genomic Distribution of L1 Elements: The Role of Insertion Bias and Natural Selection. J Biomed Biotechnol. 2006;2006:75327. doi:10.1155/JBB/2006/75327 PubMed PMID: 16877820.
33. Daniel C, Behm M, Öhman M. The role of Alu elements in the cis-regulation of RNA processing. Cell Mol Life Sci. 2015 Nov 1;72(21):4063. doi:10.1007/S00018-015-1990-3 PubMed PMID: 26223268.
34. Cordaux R, Lee J, Dinoso L, Batzer MA. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene. 2006 May 24;373(1–2):138–44. doi:10.1016/j.gene.2006.01.020 PubMed PMID: 16527433.
35. DiRusso JA, Clark AT. Transposable elements in early human embryo development and embryo models. Curr Opin Genet Dev. 2023 Aug 1;81:102086. doi:10.1016/J.GDE.2023.102086 PubMed PMID: 37441874.
36. Nikitin D. Transposable element–host genome evolutionary arms race revealed by multi-modal epigenomic profiling in a telomere-to-telomere human genome reference. bioRxiv. 2026 Mar 23;2026.03.19.712972. doi:10.64898/2026.03.19.712972
37. Barnada SM, Isopi A, Tejada-Martinez D, Goubert C, Patoori S, Pagliaroli L, et al. Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. PLoS Genet. 2022 Jun 15;18(6):e1010225. doi:10.1371/JOURNAL.PGEN.1010225 PubMed PMID: 35704668.
38. Muñoz-Gómez SA, Bilolikar G, Wideman JG, Geiler-Samerotte K. Constructive Neutral Evolution 20 Years Later. J Mol Evol. 2021;89(3):172–82. doi:10.1007/s00239-021-09996-y
39. Gianfrancesco O, Geary B, Savage AL, Billingsley KJ, Bubb VJ, Quinn JP. The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. International Journal of Molecular Sciences 2019, Vol 20,. 2019 Nov 27;20(23). doi:10.3390/IJMS20235977 PubMed PMID: 31783611.
40. Li TD, Toohill K, Modzelewski AJ. From Junk DNA to Genomic Treasure: Impacts of Transposable Element DNA, RNA, and Protein in Mammalian Development and Disease. Wiley Interdiscip Rev RNA. 2025 Jul 1;16(4):e70022. doi:10.1002/WRNA.70022 PubMed PMID: 40804709.
41. Kelsey MMG, Kalekar RL, Sedivy JM. TE-Seq: a transposable element annotation and RNA-Seq pipeline. Mobile DNA 2025 16:1. 2025 Nov 27;16(1):44-. doi:10.1186/S13100-025-00381-W
42. Wang Z, Ying Y, Wang M, Chen Q, Wang Y, Yu X, et al. Comprehensive identification of onco-exaptation events in bladder cancer cell lines revealed L1PA2-SYT1 as a prognosis-relevant event. iScience. 2023 Dec 15;26(12):108482. doi:10.1016/J.ISCI.2023.108482

43. Stamidis N, Żylicz JJ. RNA-mediated heterochromatin formation at repetitive elements in mammals. EMBO J. 2023 Apr 17;42(8):e111717. doi:10.15252/EMBJ.2022111717 PubMed PMID: 36847618.
44. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, Van De Lagemaat LN, Mager DL. Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ Line. PLoS Genet. 2006 Jan;2(1):e2. doi:10.1371/journal.pgen.0020002 PubMed PMID: 16440055.
45. Suh A, Churakov G, Ramakodi MP, Platt RN, Jurka J, Kojima KK, et al. Multiple Lineages of Ancient CR1 Retroposons Shaped the Early Genome Evolution of Amniotes. Genome Biol Evol. 2014 Dec 27;7(1):205. doi:10.1093/GBE/EVU256 PubMed PMID: 25503085.
46. Petri R, Brattås PL, Sharma Y, Jonsson ME, Pircs K, Bengzon J, et al. LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLoS Genet. 2019 Mar 1;15(3):e1008036. doi:10.1371/JOURNAL.PGEN.1008036 PubMed PMID: 30865625.
47. Zottel A, Šamec N, Kump A, Dall'olio LR, Dominkuš PP, Romih R, et al. Analysis of miR-9-5p, miR-124-3p, miR-21-5p, miR-138-5p, and miR-1-3p in Glioblastoma Cell Lines and Extracellular Vesicles. International Journal of Molecular Sciences 2020, Vol 21, Page 8491. 2020 Nov 11;21(22):8491. doi:10.3390/IJMS21228491 PubMed PMID: 33187334.
48. Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, et al. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. Genome Biology 2021 22:1. 2021 Feb 18;22(1):62-. doi:10.1186/S13059-021-02260-Y PubMed PMID: 33602314.
49. Su M, Han D, Boyd-Kirkup J, Yu X, Han JDJ. Evolution of Alu Elements toward Enhancers. Cell Rep. 2014 Apr 24;7(2):376–85. doi:10.1016/J.CELREP.2014.03.011 PubMed PMID: 24703844.
50. Hyacinthe J, Bourque G. Transposable elements impact the human regulatory landscape through cell type specific epigenomic associations. bioRxiv. 2024 Nov 19;2024.08.07.606967. doi:10.1101/2024.08.07.606967
51. Häsler J, Strub K. Alu elements as regulators of gene expression. Nucleic Acids Res. 2006 Nov;34(19):5491. doi:10.1093/NAR/GKL706 PubMed PMID: 17020921.
52. Cipta NO, Zeng Y, Wong KW, Zheng ZH, Yi Y, Warrier T, et al. Rewiring of SINE-MIR enhancer topology and Esrrb modulation in expanded and naive pluripotency. Genome Biology 2025 26:1. 2025 Apr 28;26(1):107-. doi:10.1186/S13059-025-03577-8 PubMed PMID: 40296153.
53. Chen D, Cremona MA, Qi Z, Mitra RD, Chiaromonte F, Makova KD. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. Mol Biol Evol. 2020 Dec 1;37(12):3576. doi:10.1093/MOLBEV/MSAA194 PubMed PMID: 32722770.
54. Gu X, Wang M, Zhang XO. TE-TSS: an integrated data resource of human and mouse transposable element (TE)-derived transcription start site (TSS). Nucleic Acids Res. 2024 Jan 5;52(D1):D322–33. doi:10.1093/NAR/GKAD1048 PubMed PMID: 37956335.
55. Liang X, Qiu X, Dionne G, Cunningham CL, Pucak ML, Peng G, et al. CIB2 and CIB3 are auxiliary subunits of the mechanotransduction channel of hair cells. Neuron. 2021 Jul 7;109(13):2131-2149.e15. doi:10.1016/J.NEURON.2021.05.007 PubMed PMID: 34089643.

56. Oomen ME, Rodriguez-Terrones D, Kurome M, Zakhartchenko V, Mottes L, Simmet K, et al. An atlas of transcription initiation reveals regulatory principles of gene and transposable element expression in early mammalian development. Cell. 2025 Feb 20;188(4):1156-1174.e20. doi:10.1016/j.cell.2024.12.013 PubMed PMID: 39837330.
57. Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. Development. 2016 Nov 15;143(22):4101. doi:10.1242/DEV.132639 PubMed PMID: 27875251.
58. Payer LM, Steranka JP, Ardeljan D, Walker J, Fitzgerald KC, Calabresi PA, et al. Alu insertion variants alter mRNA splicing. Nucleic Acids Res. 2018 Jan 10;47(1):421. doi:10.1093/NAR/GKY1086 PubMed PMID: 30418605.
59. Morales ME, White TB, Streva VA, DeFreece CB, Hedges DJ, Deininger PL. The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. PLoS Genet. 2015 Mar 11;11(3):e1005016. doi:10.1371/JOURNAL.PGEN.1005016 PubMed PMID: 25761216.
60. Beck CR, Garcia-Perez JL, Badge RM, Moran J V. LINE-1 Elements in Structural Variation and Disease. Annu Rev Genomics Hum Genet. 2011 Jul 13;12:187. doi:10.1146/ANNUREV-GENOM-082509-141802 PubMed PMID: 21801021.
61. Redaelli S, Grati FR, Tritto V, Giannuzzi G, Recalcati MP, Sala E, et al. Olfactory receptor genes and chromosome 11 structural aberrations: Players or spectators? Human Genetics and Genomics Advances. 2024 Apr 11;5(2). doi:10.1016/j.xhgg.2023.100261 PubMed PMID: 38160254.
62. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018 Oct 1;19(10):621. doi:10.1038/s41580-018-0028-8 PubMed PMID: 29946135.
63. Grover D, Majumder PP, Rao CB, Brahmachari SK, Mukerji M. Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. Mol Biol Evol. 2003 Sep 1;20(9):1420–4. doi:10.1093/MOLBEV/MSG153 PubMed PMID: 12832639.
64. Denisko D, Kim J, Ku J, Zhao B, Lee EA. Inverted Alu repeats in loop-out exon skipping across hominoid evolution. bioRxiv. 2025 Mar 11. doi:10.1101/2025.03.07.642063 PubMed PMID: 40161837.
65. Pascarella G, Mikhova M, Parkhi G, Godfrey J, Heyza J, Janovič T, et al. Ku limits aberrant mRNA splicing promoted by intronic antisense Alu elements. bioRxiv. 2025 Nov 20. doi:10.1101/2025.11.20.689478 PubMed PMID: 41332606.
66. Yu T, Yoon J, Zhu Y, Li A, Lee BJ, Moakley DF, et al. Primate-specific adaptation of Ku protects transcriptomic integrity by suppressing Alu-mediated alternative splicing. bioRxiv. 2025 Dec 18. doi:10.64898/2025.12.17.694518 PubMed PMID: 41446183.
67. Humphries ESA, Dart C. Neuronal and Cardiovascular Potassium Channels as Therapeutic Drug Targets: Promise and Pitfalls. J Biomol Screen. 2015 Oct 22;20(9):1055. doi:10.1177/1087057115601677 PubMed PMID: 26303307.
68. Urrutia J, Arrizabalaga-Iriondo A, Sanchez-del-Rey A, Martinez-Ibargüen A, Gallego M, Casis O, et al. Therapeutic role of voltage-gated potassium channels in age-related neurodegenerative diseases. Front Cell Neurosci. 2024 May 17;18:1406709. doi:10.3389/FNCEL.2024.1406709/TEXT

69. Ranjan R, Logette E, Marani M, Herzog M, Tâche V, Scantamburlo E, et al. A Kinetic Map of the Homomeric Voltage-Gated Potassium Channel (Kv) Family. Front Cell Neurosci. 2019 Aug 20;13:450839. doi:10.3389/FNCEL.2019.00358/TEXT
70. Hancks DC, Kazazian HH. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol. 2010 Aug;20(4):234. doi:10.1016/J.SEMCANCER.2010.04.001 PubMed PMID: 20416380.
71. Kirby AE, Loftus M, Golba EC, Lee C, Eichler EE, Marschall T, et al. Structural and transduction patterns of human-specific polymorphic SVA insertions. Mob DNA. 2025 Dec 1;16(1):42. doi:10.1186/S13100-025-00373-W PubMed PMID: 41199327.
72. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 2008 Mar 1;18(3):359–69. doi:10.1101/GR.7146408 PubMed PMID: 18256242.
73. Ferrari R, Grandi N, Tramontano E, Dieci G. Retrotransposons as Drivers of Mammalian Brain Evolution. Life 2021, Vol 11, Page 376. 2021 Apr 22;11(5):376. doi:10.3390/LIFE11050376
74. Homo sapiens genome assembly T2T-CHM13v2.0 - NCBI - NLM [Internet]. [cited 2026 Feb 3]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/
75. Moreau TRJ, Bondet V, Rodero MP, Duffy D. Heterogeneity and functions of the 13 IFN-α subtypes – lucky for some? Eur J Immunol. 2023 Aug 1;53(8):2250307. doi:10.1002/EJI.202250307;JOURNAL:JOURNAL:15214141 PubMed PMID: 37367434.
76. Apostolou E, Thanos D. Virus Infection Induces NF-κB-Dependent Interchromosomal Associations Mediating Monoallelic IFN-β Gene Expression. Cell. 2008 Jul 11;134(1):85–96. doi:10.1016/j.cell.2008.05.052 PubMed PMID: 18614013.
77. Pasquesi GIM, Allen H, Ivancevic A, Barbachano-Guerrero A, Joyner O, Guo K, et al. Regulation of human interferon signaling by transposon exonization. Cell. 2024 Dec 26;187(26):7621-7636.e19. doi:10.1016/J.CELL.2024.11.016 PubMed PMID: 39672162.
78. Tunbak H, Enriquez-Gasca R, Tie CHC, Gould PA, Mlcochova P, Gupta RK, et al. The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. Nature Communications 2020 11:1. 2020 Nov 3;11(1):5387-. doi:10.1038/s41467-020-19170-5 PubMed PMID: 33144593.
79. Hu Z, Xu B, Zhang X, Zhang X ou, Weng Z, Yu T. LOCATE: using Long-read to Characterize All Transposable Elements. bioRxiv. 2025 Mar 2;2025.02.26.640385. doi:10.1101/2025.02.26.640385
80. Eloranta ML, Rönnblom L. Cause and consequences of the activated type I interferon system in SLE. J Mol Med (Berl). 2016 Oct 1;94(10):1103–10. doi:10.1007/S00109-016-1421-4 PubMed PMID: 27094810.
81. Kelly M, Lihua S, Zhe Z, Li S, Yoselin P, Michelle P, et al. Transposable Element Dysregulation in Systemic Lupus Erythematosus and Regulation by Histone Conformation and Hsp90. Clin Immunol. 2018 Dec 1;197:6. doi:10.1016/j.clim.2018.08.011 PubMed PMID: 30149120.
82. Gázquez-Gutiérrez A, Witteveldt J, Heras SR, Macias S. Sensing of transposable elements by the antiviral innate immune system. RNA. 2021 Jul 1;27(7):735. doi:10.1261/rna.078721.121 PubMed PMID: 33888553.
83. Kelly M, Lihua S, Zhe Z, Li S, Yoselin P, Michelle P, et al. Transposable Element Dysregulation in Systemic Lupus Erythematosus and Regulation by Histone

Conformation and Hsp90. Clin Immunol. 2018 Dec 1;197:6. doi:10.1016/J.CLIM.2018.08.011 PubMed PMID: 30149120.
84. Pribus SJ, Osredek I, Otonicar J, Simovic-Lorenz M, Scherer M, Manzano-Sanchez S, et al. Integrative spatial and multi-omic profiling in bladder cancer links L1 retrotransposition to extrachromosomal DNA, genomic instability, and viral mimicry response. bioRxiv. 2025 Aug 2;2025.07.30.667694. doi:10.1101/2025.07.30.667694
85. Gudkov A, Shirokorad V, Kashintsev K, Sokov D, Nikitin D, Anisenko A, et al. Gene Expression-Based Signature Can Predict Sorafenib Response in Kidney Cancer. Front Mol Biosci. 2022 Mar 14;9:753318. doi:10.3389/FMOLB.2022.753318/TEXT
86. Sorokin M, Zolotovskaia M, Nikitin D, Suntsova M, Poddubskaya E, Glusker A, et al. Personalized targeted therapy prescription in colorectal cancer using algorithmic analysis of RNA sequencing data. BMC Cancer 2022 22:1. 2022 Oct 31;22(1):1113-. doi:10.1186/S12885-022-10177-3 PubMed PMID: 36316649.
87. Jovčevska I, Zottel A, Šamec N, Mlakar J, Sorokin M, Nikitin D, et al. High FREM2 Gene and Protein Expression Are Associated with Favorable Prognosis of IDH-WT Glioblastomas. Cancers 2019, Vol 11, Page 1060. 2019 Jul 27;11(8):1060. doi:10.3390/CANCERS11081060
88. Sorokin M, Rabushko E, Efimov V, Poddubskaya E, Sekacheva M, Simonov A, et al. Experimental and Meta-Analytic Validation of RNA Sequencing Signatures for Predicting Status of Microsatellite Instability. Front Mol Biosci. 2021 Nov 23;8:737821. doi:10.3389/FMOLB.2021.737821/TEXT
89. Vladimirova U, Rumiantsev P, Zolotovskaia M, Albert E, Abrosimov A, Slashchuk K, et al. DNA repair pathway activation features in follicular and papillary thyroid tumors, interrogated using 95 experimental RNA sequencing profiles. Heliyon. 2021 Mar 1;7(3):e06408. doi:10.1016/j.heliyon.2021.e06408
90. Fiore D, Cappelli LV, Zhaoqi L, Kotlov N, Sorokina M, Phillip J, et al. A Patient-Derived T-Cell Lymphoma Biorepository Uncovers New Pathogenetic Mechanisms and Host-Related Therapeutic Vulnerabilities [Internet]. 2023. doi:10.2139/SSRN.4529648
91. Shtam T, Naryzhny S, Samsonov R, Karasik D, Mizgirev I, Kopylov A, et al. Plasma exosomes stimulate breast cancer metastasis through surface interactions and activation of FAK signaling. Breast Cancer Research and Treatment 2018 174:1. 2018 Nov 27;174(1):129–41. doi:10.1007/S10549-018-5043-0 PubMed PMID: 30484103.
92. Shtam T, Naryzhny S, Kopylov A, Petrenko E, Samsonov R, Kamyshinsky R, et al. Functional properties of circulating exosomes mediated by surface-attached plasma proteins. Haematologica. 2018;7(4):149–53. doi:10.14740/JH412W
93. Adamyan L, Aznaurova Y, Stepanian A, Nikitin D, Garazha A, Suntsova M, et al. Gene Expression Signature of Endometrial Samples from Women with and without Endometriosis. J Minim Invasive Gynecol. 2021 Oct 1;28(10):1774–85. doi:10.1016/j.jmig.2021.03.011 PubMed PMID: 33839309.
94. Sorokin M, Kholodenko I, Kalinovsky D, Shamanskaya T, Doronin I, Konovalov D, et al. RNA Sequencing-Based Identification of Ganglioside GD2-Positive Cancer Phenotype. Biomedicines 2020, Vol 8, Page 142. 2020 May 30;8(6):142. doi:10.3390/BIOMEDICINES8060142
95. Yudina A, Tazearslan C, Baisangurov A, Nuzhdina E, Lauziere K, Segodin V, et al. Clinical and analytical validation of a combined RNA and DNA exome assay across a

large tumor cohort. Communications medicine. 2025 Dec 1;5(1). doi:10.1038/S43856-025-00934-3 PubMed PMID: 40523952.
96. Liu W, He H, Chicco D. Gene signatures for cancer research: A 25-year retrospective and future avenues. PLoS Comput Biol. 2024 Oct 1;20(10):e1012512. doi:10.1371/JOURNAL.PCBI.1012512 PubMed PMID: 39413055.
97. Ottaiano A, Santorsola M, Sabbatino F, Sirica R, Caraglia F, Ceccarelli A, et al. Interpreting cancer genetics through a two-step "evolutionary cascade hypothesis": bridging neutral and selective perspectives. Journal of Translational Medicine 2026 24:1. 2026 Feb 13;24(1):407-. doi:10.1186/S12967-026-07869-W
98. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;Chapter 4(SUPPL. 25). doi:10.1002/0471250953.BI0410S25 PubMed PMID: 19274634.
99. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science (1979). 2022 Apr 1;376(6588). doi:10.1126/science.abk3112 PubMed PMID: 35357925.
100. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature. 2014 Dec 11;516(7530):242–5. doi:10.1038/NATURE13760 PubMed PMID: 25274305.
101. Barro-Trastoy D, Köhler C. Helitrons: genomic parasites that generate developmental novelties. Trends in Genetics. 2024 May 1;40(5):437–48. doi:10.1016/j.tig.2024.02.002 PubMed PMID: 38429198.
102. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 2025 Jan 6;53(D1):D1243–9. doi:10.1093/NAR/GKAE974 PubMed PMID: 39460617.
103. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2. doi:10.1093/BIOINFORMATICS/BTQ033 PubMed PMID: 20110278.
104. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other on JSTOR [Internet]. [cited 2026 Apr 5]. Available from: https://www.jstor.org/stable/2236101
105. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020 17:3. 2020 Feb 3;17(3):261–72. doi:10.1038/s41592-019-0686-2 PubMed PMID: 32015543.
106. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995 Jan 1;57(1):289–300. doi:10.1111/J.2517-6161.1995.TB02031.X
107. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. SciPy 2010. 2010;92–6. doi:10.25080/MAJORA-92BF1922-011
108. Mays S, Stark S. The use of Fisher's exact test in contingency table analysis in palaeopathology. Int J Paleopathol. 2026 Mar 1;52:135–9. doi:10.1016/J.IJPP.2026.01.005 PubMed PMID: 41653486.

109. Waldmann P. On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction. Front Genet. 2019 Sep 1;10(SEP):899. doi:10.3389/FGENE.2019.00899 PubMed PMID: 31632436.
110. Cardoso DO, Galeno TD. Online evaluation of the Kolmogorov–Smirnov test on arbitrarily large samples. J Comput Sci. 2023 Mar 1;67:101959. doi:10.1016/J.JOCS.2023.101959
111. Klopfenstein D V., Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Scientific Reports 2018 8:1. 2018 Jul 18;8(1):10872-. doi:10.1038/s41598-018-28948-z PubMed PMID: 30022098.
112. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5. doi:10.1109/MCSE.2007.55
113. Plotly Inc. Interactive Data Visualization & Data Apps | Plotly [Internet]. [cited 2026 Apr 5]. Available from: https://plotly.com/
114. GitHub - trevismd/statannotations [Internet]. [cited 2026 Apr 5]. Available from: https://github.com/trevismd/statannotations
115. Waskom ML. seaborn: statistical data visualization. J Open Source Softw. 2021 Apr 6;6(60):3021. doi:10.21105/JOSS.03021
116. GitHub - gecko984/supervenn [Internet]. [cited 2026 Apr 5]. Available from: https://github.com/gecko984/supervenn
117. Hagberg hagberg AA, -Los lanlgov, Schult DA, Swart swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. Python in Science Conference. 2008 Jun 28;11–5. doi:10.25080/TCWV9851
118. Perrone Gary G, Unpingco Gary J, Lu Gary H minn. Network visualizations with Pyvis and VisJS. Proceedings of the 19th Python in Science Conference. 2020 Jun 2;58–62. doi:10.25080/majora-342d178e-008
119. Casper J, Speir ML, Raney BJ, Perez G, Nassar LR, Lee CM, et al. The UCSC Genome Browser database: 2026 update. Nucleic Acids Res. 2026 Jan 6;54(D1):D1331–5. doi:10.1093/NAR/GKAF1250 PubMed PMID: 41251146.
120. team T pandas development. pandas-dev/pandas: Pandas [Internet]. doi:10.5281/ZENODO.19340003
121. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 585:7825. 2020 Sep 16;585(7825):357–62. doi:10.1038/s41586-020-2649-2 PubMed PMID: 32939066.
122. Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities [Internet]. 2025 Jul 7 [cited 2026 Apr 5]. Available from: https://arxiv.org/pdf/2507.06261
123. Biswas SS. ChatGPT for Research and Publication: A Step-by-Step Guide. The Journal of Pediatric Pharmacology and Therapeutics : JPPT. 2023;28(6):576. doi:10.5863/1551-6776-28.6.576 PubMed PMID: 38130350.