Transposable elements generate regulatory novelty in a tissue-specific fashion

Marco Trizzino[1,2,*], Aurélie Kapusta[3,4] and Christopher D. Brown[2,5,*]

[1]Gene Expression and Regulation Program, The Wistar Institute, Philadelphia, PA, USA

[2]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA

[3]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

[4]USTAR, Center for Genetic Discovery, Salt Lake City, UT, USA

[5]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA

*Correspondence: M.T. (marco.trizzino83@gmail.com), and C.D.B (chrbro@pennmedicine.upenn.edu)

Abstract

Background

Transposable elements (TE) are an important source of evolutionary novelty in gene regulation. However, the mechanisms by which TEs contribute to gene expression are largely uncharacterized.

Results

Here, we leverage Roadmap and GTEx data to investigate the association of TEs with active and repressed chromatin in 24 tissues. We find 112 human TE types enriched in active regions of the genome across tissues. SINEs and DNA transposons are the most frequently enriched classes, while LTRs are often enriched in a tissue-specific manner. We report across-tissue variability in TE enrichment in active regions. Genes with consistent expression across tissues are less likely to be associated with TE insertions. TE presence in repressed regions similarly follows tissue-specific patterns. Moreover, different TE classes correlate with different repressive marks: Long Terminal Repeat Retrotransposons (LTRs) and Long Interspersed Nuclear Elements (LINEs) are overrepresented in regions marked by H3K9me3, while the other TEs are more likely to overlap regions with H3K27me3. Young TEs are typically enriched in repressed regions and depleted in active regions. We detect multiple instances of TEs that are enriched in tissue-specific active regulatory regions. Such TEs contain binding sites for transcription factors that are master regulators for the given tissue. These TEs are enriched in intronic enhancers, and their tissue-specific enrichment correlates with tissue-specific variations in the expression of the nearest genes.

Conclusions

We provide an integrated overview of the contribution of TEs to human gene regulation. Expanding previous analyses, we demonstrate that TEs can potentially contribute to the turnover of regulatory sequences in a tissue-specific fashion.

Keywords: Transposons, gene regulation, tissue-specific, transcription factors

Background

Sequences derived from transposable element (TE) insertions make up roughly half of the length of the human genome. Several TE groups still show transposing activity in humans, including Long Terminal Repeat Retrotransposons (mostly ERV1-LTRs; [1–3]), Long Interspersed Nuclear Elements (LINEs, mostly L1s; [4–5]), Short Interspersed Nuclear Elements (SINEs) of the Alu families [6,7], and SINE-VNTR-*Alus* (SVAs; [8,9]).

Multiple elegant studies have demonstrated that TE sequences play a functional role in eukaryotic gene regulation [10–32]. Consistently, we recently demonstrated that TEs are the primary source of evolutionary novelty in primate gene regulation, and reported that the large majority of newly evolved human and ape specific liver cis-regulatory elements are derived from TE insertions [33]. Similarly, other studies have shown that the recruitment of novel regulatory networks in the uterus was likely mediated by ancient mammalian TEs [21,22], and that TEs have a role in pluripotency [34]. Conversely, other researchers have proposed that TE exaptation into regulatory regions is rare [35], and that TE silencing may not be a major driver of regulatory evolution in primates [36].

Given these contrasting lines of evidence, we aimed to shed light on the contribution of TEs to the evolution of the tissue-specific regulation of human gene expression. For this purpose, we took advantage of publicly available data [37,38] to investigate patterns of TE overlap with tissue-specific histone modification states and to characterize the contribution of TEs to tissue-specific gene expression. We find that a significant fraction of the existing human TEs are enriched in regions of the genome bearing epigenetic hallmarks of active or repressed chromatin, suggesting they could potentially be actively regulated by the cellular machinery. DNA

106 transposons and SINEs represent the most frequently enriched classes across
107 tissues, while LTR-ERV1s are the TEs that more commonly show tissue-specific
108 enrichment and active regulatory activity. TE enrichment in active and repressed
109 chromatin exhibits tissue-specific patterns. Genes with consistent expression across
110 tissues are less likely to be associated with a local TE insertion. We detect multiple
111 instances of TEs showing tissue-specific enrichment in active and repressed regions,
112 and demonstrate that they contain binding sites for transcription factors that are
113 tissue-specific master regulators.
114

115 ## Results

116 **Specific TE families are enriched in active and repressed genomic regions**

117 To investigate the extent to which TEs contribute to the regulation of human
118 gene expression, we leveraged publicly available data from the Roadmap
119 Epigenomics Project [37] and from the GTEx Project [38]. We focused on 24
120 primary tissues and cell types that were processed by both consortia
121 (Supplementary Table S1). Using five different histone modifications (H3K4me1,
122 H3K4me3, H3K36me3, H3K9me3, and H3K27me3), Roadmap segmented the
123 human genome into 15 regulatory classes, reflecting different degrees and types of
124 regulatory activity. We took advantage of this classification to define active
125 (H3K4me1, H3K4me3, H3K36me3) and repressed (H3K9me3, and H3K27me3)
126 chromatin regions in each of the studied tissues.

127 To test for TE enrichment in active and repressed chromatin, we used the TE-
128 Analysis pipeline ([39]; https://github.com/4ureliek/TEanalysis; Supplemental File
129 S1). This pipeline is designed to output the TE composition of given features, such
130 as TE counts and TE amounts, aiming to detect potential TE enrichments in the

131 select features. As expected, we find that the majority of human TEs are significantly
132 depleted from regions marked as active by Roadmap histone modifications (mean
133 83.9% of TEs; FDR <5%; Supplementary Table S2). Nevertheless, 112 TE families
134 (9.07% of the annotated TE families in the human genome) are significantly enriched
135 in active chromatin in at least one tissue (FDR <5%; Fig. 1a; Supplementary Table
136 S2). These data suggest variability across tissues: aorta, brain anterior caudate, and
137 adipose are the most "permissive" tissues, while right atrium and spleen do not show
138 any significant TE enrichment in active regions (Fig. 1a).

139 SINEs and "cut and paste" DNA transposons are the classes most frequently
140 enriched in active chromatin (Fig. 1b). SINE families, the most abundant human TEs
141 (38.8% of the total), correspond to 43–66% of the TEs enriched in active regions
142 (FDR < 5%), these fractions being more than expected by chance in all tissues
143 (Proportion Test $p < 2.2 \times 10^{-16}$ for each tested tissue). Similarly, DNA TEs, that
144 account for 11.3% of the annotated TEs, represent 29–47% of the transposons
145 enriched in active regions (Proportion Test $p < 2.2 \times 10^{-16}$ for each tested tissue). In
146 general, SINE-Alu elements are the most commonly enriched TEs (Supplementary
147 Table S2).

148 Conversely, LTRs and LINEs are significantly depleted from active genomic
149 regions of all tissues (Proportion Test $p < 2.2 \times 10^{-16}$ for each tested tissue; Fig. 1b).
150 Finally, SINE-VNTR-*Alus* (SVAs), which are the least abundant TEs in the human
151 genome (0.12% of the total annotate TEs in the human genome), are significantly
152 overrepresented in active chromatin in 13/24 tissues; Fig. 1b).

153 We set out to investigate the TEs overlapping active regions. These TEs are
154 depleted in active promoters and intergenic regions, but significantly enriched within
155 active regions inside gene bodies, and in particular in introns (Fisher's Exact Test *p*-

156 *values* in Fig. 1c). More specifically, 96.3% of TEs enriched in gene bodies overlap
157 introns, in line with the normally observed distribution of introns and exons in the
158 human genome (Fig. 1c, Fisher's Exact Test $p > 0.05$). We speculate that genomic
159 regions containing active genes are more frequently accessible, thus providing a
160 substrate for TEs to insert. Moreover, TEs present in the bodies of active genes may
161 be less likely to be silenced than TEs in intergenic regions.

162 Using the same approach previously described for the active regions, we
163 searched for TEs enriched in repressed genomic regions. Overall, 314 human TE
164 families (25.4%) are significantly enriched in repressed regions of the genome in at
165 least one tissue (FDR <5%; Fig. 2a; Supplementary Table S3). LTRs (predominantly
166 ERV1) represent the large majority of the repressed TEs (Fig. 2b), followed by LINEs
167 (predominantly L1s) and DNA TEs. Notably, ERV LTRs and L1 LINEs are among
168 the most active TEs in the genome, and also have their own regulatory architecture
169 [40, 41]. We thus surmise that these autonomous active TEs may be preferential
170 targets of repressive marks.

171 We note a very high variability in the number TE families enriched in
172 repressed regions across tissues (Fig. 2a), as well as large differences in the
173 composition of enriched TE classes in the repressed regions. Notably, the tissues
174 that harbor the highest number of TE families enriched in repressed regions
175 (pancreas, aorta, lung, spleen, esophagus, breast, and liver; Fig. 2a) are also those
176 displaying the highest numbers of enriched LINEs in the same repressed regions
177 (Fig. 2b).

178

179 **Different TE repression patterns in the human genome**

180 We examined whether TEs preferentially overlap regions repressed via
181 Polycomb Repressive Complex (H3K27me3) or via Heterochromatin (H3K9me3).
182 Overall, 78.6% of the regions classified as repressed in the human genome across
183 all tissues are bound by H3K27me3 (Polycomb Repressive Complex), while 21.4%
184 are marked by H3K9me3 (Heterochromatin conformation). However, when we
185 restrict the analysis to the repressed regions containing a TE, we report an overall
186 higher than expected overlap with H3K27me3 (median across tissues 85.5%;
187 Proportion Test across tissues $p < 2.2 \times 10^{-16}$; Supplementary Table S4; Fig. 2C),
188 and a consequent underrepresentation of H3K9me3 (median 15.5%; Supplementary
189 Table S4; Proportion Test $p < 2.2 \times 10^{-16}$; Fig. 2d). In 20/24 of the tested tissues,
190 TEs are marked by H3K27me3 more than expected by chance (Proportion Test $p <$
191 2.2×10^{-16} for each of the 20 significant tissues; Supplementary Table S4). In the
192 remaining four tissues this histone mark is instead underrepresented, while
193 H3K9me3 is overrepresented: breast (H3K27me3 = 76.4%; Supplementary Table
194 S4; Proportion Test $p < 2.2 \times 10^{-16}$), aorta (55.1%; Supplementary Table S3; $p < 2.2$
195 $\times 10^{-16}$), lung (48.9%; Supplementary Table S4; $p < 2.2 \times 10^{-16}$), and spleen (26.5%;
196 Supplementary Table S3; $p < 2.2 \times 10^{-16}$). Notably, in these four tissues we detect
197 the highest numbers of TE families enriched in repressed regions (Fig. 2a), and the
198 highest proportion of repressed LINEs. We speculate that the heterochromatin state
199 (H3K9me3) may be employed to target specific TE classes and families in a context
200 specific manner [36].

201 We therefore tested whether different TE classes correlate with either
202 heterochromatin (H3K9me3) or with Polycomb repressed chromatin (H3K27me3).
203 LTRs, LINEs, and SVAs are overrepresented in regions marked by H3K9me3
204 (Fisher's Exact Test $p < 2.2 \times 10^{-16}$; Fig. 2d). Conversely, SINEs and DNA TEs are

205 significantly more likely to overlap H3K27me3 than expected by chance (Fisher's
206 Exact Test $p < 2.2 \times 10^{-16}$; Fig. 2d). Notably, SVAs are depleted from the regions
207 marked by H3K27me3 (Fig. 2d).
208 These findings are consistent with recent reports suggesting that H3K27me3 and
209 H3K9me3 target different transposon types in embryonic stem cells [42], and with a
210 study reporting that LINEs, LTRs, and SVAs are the most abundant TEs repressed
211 by H3K9me3 in induced pluripotent stem cells [42].
212
213 **Ancient TEs are enriched in active regions, while young TEs are repressed**
214 We clustered the annotated human TEs in 35 age classes as in ref. 39 (e.g.
215 Eutheria, Primates, Hominidae; Supplemental Table S6), and used the TE-Analysis
216 shuffling script to test for enrichment of each age class in a given set of regions (see
217 Methods). Using this approach, we assessed the age of TEs enriched in active and
218 repressed genomic regions. Ancient TE classes (i.e. age classes older than the
219 Eutheria lineage) are enriched in the active regions of all tested tissues (FDR <5%;
220 Supplemental Table S6). These TEs are largely vertebrate or mammalian specific
221 (Supplemental Table S6). Notably, the only tissues with an enrichment of young TEs
222 (specifically primate specific) are blood related (Mononuclear and Lymphoblastoid
223 Cells). These results are in agreement with an elegant study that discovered a key
224 role of primate specific TEs in the regulatory evolution of immune response [25]. TE
225 families enriched in active regions across at least 20 of the 24 tissues correspond to
226 DNA TEs and SINEs (Supplemental Table S2). Despite a lack of enrichment of all
227 young TEs taken together in active regions, 24 *Alu* families are in fact enriched in
228 active regions.

229 In contrast, young TEs (i.e. TE classes younger than the Eutheria lineage
230 split) are significantly enriched in the repressed regions of most tissues. In particular
231 human specific TEs are enriched in the repressed regions of all brain related tissues
232 (FDR <5%; Supplemental Table S6). These young TEs correspond to ERV LTRs, L1
233 LINEs, and SVAs, but only one family is found enriched in at least 20 tissues
234 (MER52A), which is in line with the broad cross-tissue variability of the TEs enriched
235 in repressed chromatin regions (see above).
236 Collectively, these data suggest that young TEs are predominantly silenced, while
237 the older TE fragments still detectable in the human genome are now more tolerated.
238

239 **TE insertions are associated with gene expression variance across tissues**

240 We employed GTEx data to test if TE insertions affect local gene expression.
241 For this purpose, we first assigned each TE overlapping an active genomic region to
242 its nearest gene transcription start site (TSS). Next, we divided all human genes in
243 four categories (Supplemental Table S7): 1) Genes associated with TEs that are only
244 found in active regions across tissues; 2) Genes associated with TEs that are found
245 in active or repressed regions in a tissue-specific fashion; 3) Genes associated with
246 TEs that are only found in repressed regions; 4) Genes never associated with TE
247 insertions. Based on this classification, genes associated with a TE insertion in
248 regions that are active in at least one tissue are characterized by significantly higher
249 expression variance (normalized by mean expression) than genes either associated
250 to repressed TEs or not associated to a TE (Wilcoxon's Rank Sum Test $p < 2.2 \times 10^{-}$
251 ^{16}; Fig. 3). Similarly, the genes associated with TEs exclusively found in active
252 regions have significantly higher expression variance than the genes associated with
253 TEs present in both active and repressed regions (Wilcoxon's Rank Sum Test p =

254 9.91 x 10^{-8}; Fig. 3). We reasoned that TE insertions may happen more likely at
255 longer genes located in gene deserts. However, even after correcting our model for
256 gene density and gene length, the gene expression variance is still positively
257 correlated with TE insertion in active regions (linear regression $p < 2.2 \times 10^{-16}$).
258 Together, these findings suggest that genes with local TEs overlapping active
259 chromatin have higher variability in gene expression across tissues, and that genes
260 consistently expressed across tissues (e.g. housekeeping and other essential genes)
261 may be less tolerant towards TE insertions in their regulatory regions.
262
263 **Tissue-specific TE enrichment in active regions correlates with tissue-specific**
264 **gene expression**
265 We compared the relative enrichment in active regions of each TE family
266 across tissues. Specifically, for each TE enriched in active regions (FDR < 5%), we
267 leveraged the Odd Ratios from the permutation test of the TE-Analysis pipeline to
268 compute Z-scores (i.e. effect sizes; see methods), and compare them across
269 tissues. We find that TE enrichment varies substantially across tissues
270 (Supplemental Table S5; Fig. 4), and many TEs exhibit tissue-specific enrichment in
271 active chromatin (Fig. 4). For example, HERV15 (LTR) is significantly more enriched
272 in the liver and in the stomach mucosa compared to any other tissue (Fig. 4). Motif
273 analysis revealed that the liver regions of active histone modification overlapping
274 HERV15 are enriched in motifs for EOMES (Supplemental File S2). This
275 transcription factor (TF) has a key role in the hepatic immune response, instructing
276 the development of two distinct natural killer cell lineages specific to this tissue [43].
277 Moreover, EOMES is also an established tumor suppressor in Hepatocellular
278 Carcinoma [44]. Notably, HERV15 was recovered as significantly enriched in the

279 human liver enhancers also in our previous study [33], suggesting that the findings of
280 the present analysis are not likely to represent batch-specific effects of the Roadmap
281 data.

282 Similarly, X7C (LINE) and Charlie15a (DNA TE), are the most enriched TEs
283 within regions bearing active chromatin state in the breast. In the sequence of these
284 we find enrichment for binding sites for key breast TFs as KLF5 and CPEB1 (Fig. 5a;
285 Supplemental File S2). Notably, KLF5 is an essential regulator of hormonal
286 signaling and breast cancer development [45], and is considered a breast cancer
287 suppressor [46]. Similarly, CPEB1 mediates epithelial-to-mesenchyme transition in
288 breast, and mice depleted of this gene showed increased breast cancer metastatic
289 potential [47]. Interestingly Charlie15a shows tissues-specific depletion in the
290 mononuclear blood cells (Fig. 4), highlighting a potential tissue-specific regulatory
291 activity.

292 To assess the robustness of the enrichment of X7C and Charlie15a in the breast, we
293 ran the TE-Analysis pipeline on publicly available H3K27ac and H3K4me1 data
294 generated by Encode from the breast epithelium and from the MCF7 cell line [48].
295 Notably, these two TEs were also significantly enriched in the Encode data (FDR <
296 5%), suggesting that batch effects are unlikely strong drivers of this trend.

297 Analogously, LTR13_ is the most enriched TE in the active chromatin of
298 pancreas and Lymphoblastoid Cell Line (LCL). These LTR copies are enriched for
299 binding sites for SOX9 and PRDM1/Blimp-1 (Fig. 5d; Supplemental File S2). SOX9
300 is a master regulator of the pancreatic program [49], while PRDM1/Blimp-1 has a
301 central role in determining and shaping the secretory arm of mature B Lymphocyte
302 differentiation [50].

303 We next tested whether tissue-specific TE enrichment in active chromatin
304 (Fig. 4, 5a–f) correlates with tissue-specific-changes in gene expression.
305 Specifically, we tested the TE families showing the highest degree of tissue-specific
306 enrichment (Fig. 4: HERV15/liver, LTR13_/LCL, X7C-Charlie15a/breast). With the
307 exception of HERV15/liver (Wilcoxon's Rank Sum Test $p > 0.05$), in the other tested
308 instances (LTR13_/LCL; X7C-Charlie15a/breast) the tissue-specific enrichment of
309 the TEs in active chromatin regions is associated with a significant change in the
310 associated gene expression (Wilcoxon's Rank Sum Test *p-values* in Figs. 5b,e).
311 These findings support a possible regulatory role for the co-opted TEs.

312 To better understand how these tissue-specific TEs may be involved in the
313 regulation of gene expression, we investigated what typology of genomic region they
314 overlap (i.e. promoter, intergenic, introns, exons). Both X7C/Charlie15a in breast
315 and LTR13_ in LCLs are significantly depleted in promoter and intergenic regions,
316 but overrepresented in gene bodies (Figs. 5c, f), 97.8% (X7C/Charlie15a) and 96.4%
317 (LTR13_) of them respectively found in introns.

318 The Roadmap data did not include H3K27ac profiles for all tissues. Therefore,
319 to further characterize these intronic regions, we leveraged again the publicly
320 available H3K27ac and H3K4me1 Encode data for the breast (Breast epithelium and
321 MCF7 cell line; [48]). These data reveal that 57.0% of the intronic regions containing
322 X7C or Charlie15a overlap a H3K27ac or H3K4me1 peak, thus suggesting that most
323 of these regions likely represent breast intronic enhancers. As comparison, only
324 33.7% of random intronic regions of the same size and number of the ones
325 overlapping X7C/Charlie15a TEs are overlap a H3K27ac or H3K4me1 peak (Fisher's
326 Exact Test $p < 2.2 \times 10^{-16}$).

327 Collectively, these findings point towards a model in which specific TE
328 families, largely belonging to LTR (ERVs) and DNA TE classes, have more
329 regulatory potential than other transposons. Furthermore, our data expand upon
330 previous findings suggesting that ERVs that escape repression can have a
331 significant impact on the host gene regulation [9, 25, 26, 33, 51, 52].
332
333 **SVAs exhibit tissue-specific regulatory activity**
334 In our recent work, we demonstrated that a large fraction of human specific
335 cis-regulatory elements in the liver are SVA transposons, which typically function as
336 transcriptional repressors, at least in this tissue [33]. SVAs are very young
337 transposons, being Hominidae (SVA_A, B, C and D) and human specific (SVA_E
338 and F). According to Roadmap data, SVAs are enriched in the active regions of
339 13/25 tissues (Fig. 1b), and mainly corresponded to SVA_A copies (Supplementary
340 Table S4). We first assessed the potential contribution of SVAs to gene regulation of
341 two of these tissues: the adipose nuclei and the liver.
342 In both tissues, SVAs provide binding sites for key transcription factors (Fig.
343 5g, j; Supplemental File S2). ZEB1 is the master regulator of adipogenesis [53, 54],
344 and, based on GTEx data, is ten times more highly expressed in adipose tissue
345 compared to the liver. Similarly, SOX6 contributes to the developmental origin of
346 obesity by promoting adipogenesis, and has a key role in adipocyte differentiation
347 [55]. Consistent with the data reported for other tissues, SVAs associated with
348 active chromatin in adipose nuclei and liver are strongly enriched in gene bodies
349 (Figs. 5i, l). Genes associated with SVAs in the adipose nuclei are significantly more
350 highly expressed in this tissue compared to other tissues (Wilcoxon's Rank Sum

351 Test p = 0.0002; Fig. 5h), suggesting that SVA elements can work as transcriptional
352 activators, at least in the adipose tissue.

353 In the liver, SVAs in active regions are enriched for hepatic regulators like
354 CPEB1, that mediates insulin signaling in the liver (Fig. 5j; [56]), and STAT3, that
355 regulates liver regeneration and immune response and negatively modulates insulin
356 action (Fig. 5j; [57]). However, the liver SVAs are also enriched for established
357 transcriptional repressors, like Smad3 (Fig. 5j). Consistently, genes associated with
358 liver active SVAs exhibit lower expression in this tissue compared to all the others
359 (Wilcoxon's Rank Sum Test $p < 2.2 \times 10^{-16}$; Fig. 5k), supporting the previously
360 proposed repressive role of SVAs in the hepatic system [33].

361

362 ## Discussion

363 The contribution of transposable elements (TEs) to gene regulation was proposed
364 over half a century ago [10–13] and considerably expanded over the last two
365 decades, largely due to the advances in next generation sequencing [14–36].
366 In order to gain insights in this topic, we identified TEs enriched in active and
367 repressed genomic regions of 24 human tissues, using Roadmap and GTEx data.
368 Our analyses provide a novel integrated overview of the potential impact of TEs to
369 the human gene regulation across multiple tissues, correlating the enrichment of TE
370 copies in active chromatin to tissue-specific gene expression. In fact, many of the
371 previous studies have proposed that TEs are frequently enriched in cis-regulatory
372 elements and lncRNAs [21, 22, 33, 39, 58], but the actual effect of the presence of
373 TEs on the associated gene expression was not tested on a large scale.
374 Recent work has evaluated the prevalence of TE-derived DNA in enhancers and
375 promoters across mouse cell lines and primary tissues [35]. The present study builds

376 upon this by investigating the dynamics of TE recruitment and the potential effects
377 on tissue-specific gene expression.
378 We demonstrate that ~10% of the TEs identified in the human genome are
379 significantly enriched in active regions (promoters, intergenic enhancers, intronic
380 enhancers) of 24 different human tissues. In general, we report a high degree of
381 variability of TE enrichment in the active and repressed genome across tissues, and
382 detect multiple instances of TEs displaying potential tissue-specific regulatory
383 function. We acknowledge that the correlation between tissue-specific TE
384 enrichment in active regions and the tissue-specific changes in gene expression
385 does not necessarily underly a causal role for the TEs. On the other hand, while it is
386 possible that the changes in gene expression are simply due to the presence of a
387 tissue-specific active histone mark, we also find that in all of the tested cases the
388 enriched TE sequence provides binding sites for transcription factors that are master
389 regulators for that specific tissue. This is consistent with the changes in the gene
390 expression of associated (i.e. adjacent) genes and could explain why these TE
391 insertions are retained by selection.
392 Enriched TEs are typically distributed along gene bodies, likely functioning as
393 intronic enhancers. We reason that this may be explained by the assumption that
394 TEs located within intra-genic regions are less likely to be repressed or removed. In
395 agreement with these findings, a recent study has shown that TEs are depleted in
396 human promoters and intergenic enhancers across multiple tissues [35]. In this
397 context, we see a correlation between gene expression variance and the insertion of
398 TEs in their loci or regulatory regions. This may suggest that genes consistently
399 expressed across tissues are less prone towards TE co-option in their regulatory

400 networks, but future analyses in this direction will be needed to further characterize
401 this phenomenon.
402 On the other hand, L1 LINEs and ERV LTRs are the most frequently enriched TE
403 classes in the repressed regions. L1 retrotransposons are among the most active
404 TEs in the human genome [59], and several studies have demonstrated that they are
405 also active in brain tissues (e.g. hippocampus), and can contribute to neuronal
406 genetic diversity in mammals [60–63]. Both L1s and LTRs possess their own
407 regulatory architecture, and we speculate that their preferential silencing prevents
408 these TEs from interfering with gene regulatory networks. Despite this, we
409 demonstrate that LTRs that escape repression may be co-opted in a tissue-specific
410 manner in the active regulatory regions, putatively as a consequence of their
411 regulatory potential.
412 We show that TEs enriched in repressed regions of most tissues are generally
413 young, while TEs enriched in active regions of most tissues generally predate the
414 split of eutherian mammals. This is consistent with an accumulation of mutations in
415 these ancient copies that would have increased the likelihood to generate binding
416 sites for transcription factors, and thus the probability for the TE to be co-opted in the
417 regulatory networks. An alternative explanation could be that young TE insertions in
418 active chromatin regions are more likely to be removed by purifying selection than
419 the new insertions in repressed regions, since the latter are more likely to have a
420 neutral impact.
421 Finally, we demonstrate that SVAs, previously characterized as transcriptional
422 repressors in select cell-types [33, 64], can act as both activators or repressors in a
423 tissue-specific fashion.
424

Conclusions

In summary, we present a comprehensive overview of the contribution of TE copies to human gene regulation: not only do they provide an important source of evolutionary novelty for the genome, but they can also function with tissue-specific patterns, modulating the expression of key genes and pathways.

Methods

TE-Analysis pipeline

To test for TE enrichment in active and repressed regions, we used the TE-Analysis pipeline v 4.6 ([39]; https://github.com/4ureliek/TEanalysis). This pipeline is designed to output the TE composition of given features, such as TE counts and TE amounts, aiming to detect potential TE enrichments in the select features. Roadmap annotated BED files (i.e. files listing the coordinates of annotated genomic regions) for each of the 24 tissues were downloaded (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/; last access: 10/4/2017). One file per tissue was downloaded (TISSUE_ID_coreMarks_dense.bed.gz"; Supplementary Table S1). From each of the 24 BED files, we produced two different files: one for the regions enriched with epigenomics hallmarks of active chromatin (hereafter "active regions". Histone marks: H3K4me1, H3K36me3, H3K4me3. Roadmap annotations: "TssA", "TssAFlnk", "TxFlnk", "Tx", "TxWk", "EnhG", "Enh", "TssBiv", "EnhBiv"), and one for the regions with signature of repressed chromatin (hereafter: "repressed regions". Histone marks: H3K27me3, H3K9me3. Roadmap annotations: "Het", "ReprPC", "ReprPCWk").

449 For each tissue, we tested for TE enrichment in the “active” and “repressed”
450 BED files using the “TE-analysis_Shuffle_bed.pl” script v 4.3. Specifically, this script
451 assesses which TEs are significantly enriched in a set of features (BED files) by
452 comparing observed overlaps with the average of *N* expected overlaps (here 1000).
453 These expected overlaps were obtained by shuffling the genomic position of TEs. TE
454 annotations were downloaded from the University of California Santa Cruz Genome
455 Browser (RepeatMasker, Hg19 version; [65]).
456 The “TE-analysis_Shuffle_bed.pl” script was run with Bedtools v2.27.1 [66] and the
457 following parameters:
458 -f Roadmap_BEDFILE (active or repressed)
459 -q RepeatMasker.out (TE file, hg19)
460 -n 1000 (number of bootstrap replicates)
461 -r hg19.chrom.sizes
462 -g 20141105_hg38_TEage_with-nonTE.txt (distributed with the pipeline)
463 -s rm (shuffles the TEs within their genomics position)
464
465 The script performs a two-tailed permutation test to assess the enrichment (or
466 depletion) of each annotated TE in the given regions (Roadmap regions), thus
467 assigning a *p-value* to each annotated TE. Additionally, we corrected for multiple
468 testing by applying a False Discovery Rate (FDR; [67]). Only TEs with FDR < 5%
469 were retained, considered significantly enriched in the given tissue, and used for
470 downstream analyses.
471
472 **Composition of enriched TEs**

473 To characterize TEs enriched within active and repressed regions of each tissue
474 (e.g. Figs. 1b, 2b), each TE was assigned to one of the major TE classes: DNA
475 transposons, LINEs, LTRs, SINEs, SVAs, according to RepeatMasker annotations.
476 To assess the genomic distribution of the enriched TEs (e.g. Figs. 1c, 2c), we
477 considered as 1) PROMOTERS: all of the regions found within +/- 1 Kb from an
478 annotated TSS (Gencode_v19 comprehensive annotations). 2) GENE BODIES: all
479 of the regions overlapping an annotated gene but not overlapping the promoter
480 region. 3) INTERGENIC - all of the regions not overlapping an annotated gene and
481 distant > 1 Kb from a TSS.
482
483 **Correlation between TE insertion and variance in gene expression**
484 We calculated the variance and mean of the TPM (Transcripts Per Million) for each
485 gene using GTEx data. We assigned each TE overlapping an active or a repressed
486 region to the closest gene, based on the distance to the nearest transcription start
487 site. Next, we divided all human genes in four categories: 1) Genes associated with
488 TEs that are only found in active regions across tissues; 2) Genes associated with
489 TEs that are found in active or repressed regions in a tissue-specific fashion; 3)
490 Genes associated with TEs that are only found in repressed regions; 4) Genes never
491 associated with TE insertions. Gene expression variance, normalized by mean
492 expression, was compared across the four categories. Gene density and gene length
493 were used as covariates for the model. Specifically, gene density was calculated as
494 the amount of exonic sequence present within +/- 100 Kb from each gene. In
495 summary, the following model was used:
496
497 lm(normalized_variance~CATEGORY+gene_length+gene_density)

498

499 Variance was normalized by average expression across tissues.

500

501 **Computation of Z-scores for tissue-specificity**

502 For each TE enriched in active regions (FDR < 5%), we used the Odd Ratios (OR)
503 from the permutation test of the TE-Analysis pipeline to compute Z-scores with the
504 following equation: (OR – mean(OR)) / sd(OR). Z-scores can be found in
505 Supplemental Table S5.

506

507 **Motif analyses**

508 Motif analyses were performed using the Meme-Suite [68], and specifically with the
509 Meme-ChIP application. Fasta files of the regions of interest were produced using
510 BEDTools v2.27.1. Shuffled input sequences were used as background. *E-values* <
511 0.001 were used as threshold for significance [68].

512

513 **Testing for TE co-option on gene expression**

514 For each human gene and for each tissue, GTEx provides the mean of the TPMs
515 (Transcripts Per Million). To test whether tissue-specific TE enrichment correlates
516 with tissue-specific changes in gene expression, for each gene associated with a TE
517 of interest, we used the mean TPMs to compare the expression of genes in the
518 tissue of enrichment Vs the average of the gene expression of the same genes in all
519 the other considered tissues (i.e. mean of TPMs across all the other tissues).

520

521 **Statistical and genomic analyses**

522 All statistical analyses were performed using R v3.4.1 [69]. Figures were made with
523 the package ggplot2 [70]. BEDTools v2.27.1 was used for all the genomic analyses.
524

525 **Acknowledgements**

526 We thank Roadmap and GTEx Consortia for the generation of invaluable data. MT
527 thanks his current P.I. (Alessandro Gardini, The Wistar Institute) who granted him
528 time and freedom to work on this project. We also thank the two anonymous
529 reviewers for their valuable suggestions and insights.
530

531 **Authors' contributions**

532 MT and CDB designed the project. MT, AK, and CDB analyzed the data. MT, AK,
533 and CDB wrote and approved the manuscript.
534

535 **Ethics approval and consent to participate**: N/A

536 **Consent for publication:** N/A

537 **Availability of data and material:** N/A

538 **Competing interests:** The authors declare no competing interests.

539 **Funding:** N/A

540

541 **References**

542 1. Tonjes RR, et al. HERV-K: the biologically most active human endogenous retrovirus
543 family. J Acquir Immune Defic Syndr Hum Retrovirol. 1996;13(1):261–7.
544 2. Medstrand P, Mager DL. Human-specific integrations of the HERV-K endogenous
545 retrovirus family. J Virol. 1998;72:9782–7.

3. Fuchs NV, Loewer S, Daley GQ, Izsvak Z, Lower J, Lower R. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. 2013. Retrovirology;10:115.
4. Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 1988;332:164–6.
5. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci USA. 2003;100:5280–5.
6. Batzer MA, Deininger PL. A human-specific subfamily of Alu sequences. Genomics. 1991;9:481-7.
7. Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 1991;19:3619–23.
8. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. SVA elements are non autonomous retrotransposons that cause disease in humans. Am J Hum Genet. 2003;73:1444–51.
9. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. SVA elements: a hominid-specific retroposon family. J Mol Biol. 2005;354:994–1007.
10. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA. 1950;36:344–55.
11. McClintock B. The significance of responses of the genome to challenge. Science. 1984;226:792–801.
12. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969;165:349–57.
13. Davidson EH, Britten RJ. Regulation of gene expression: possible role of repetitive sequences. Science. 1979;204:1052–9.
14. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 2003;19:68–72.
15. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441:87–90.
16. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA. 2007;104:18613–8.
17. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008;18:1752–62.
18. Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci USA. 2008;105: 4220–5.

589 19. Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang
590 X, Wang L, Saenz-Vash V, Gnirke A, et al. ZBED6, a novel transcription factor
591 derived from a domesticated DNA transposon regulates IGF2 expression and muscle
592 growth. PLoS Biol. 2009;7:e1000256.
593 20. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G.
594 Transposable elements have rewired the core regulatory network of human embryonic
595 stem cells. Nat Genet. 2010;42:631–4.
596 21. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene
597 regulatory networks contributed to the evolution of pregnancy in mammals. Nat
598 Genet. 2011;43:1154–9.
599 22. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D,
600 Sheikh SZ, Grützner F, Bauersachs S, et al. Ancient transposable elements
601 transformed the uterine regulatory landscape and transcriptome during the evolution
602 of mammalian pregnancy. Cell Rep. 2015;10:551–61.
603 23. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown
604 GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel
605 genome organization and CTCF binding in multiple mammalian lineages. Cell.
606 2012;148:335–48.
607 24. Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as
608 species-specific enhancer elements in the placenta. Nat Genet. 2013;45:325–9.
609 25. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through
610 co-option of endogenous retroviruses. Science. 2016;351:1083–7.
611 26. Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory
612 sequences are derived from transposable elements. PLoS Genet 9: e1003504.
613 27. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL,
614 Ligon KL, et al. DNA hypomethylation within specific transposable element families
615 associates with tissue-specific enhancer landscape. Nat Genet. 2013;45:836–41.
616 28. del Rosario RCH, Rayan NA, Prabhakar S. Noncoding origins of anthropoid traits and
617 a new null model of transposon functionalization. Genome Res. 2014;24:1469–84.
618 29. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T.
619 Widespread contribution of transposable elements to the innovation of gene
620 regulatory networks. Genome Res. 2014;24:1963–76.
621 30. Pavlicev M, Hiratsuka K, Swaggart KA, Dunn C, and Muglia L Detecting
622 Endogenous Retrovirus-Driven Tissue-Specific Gene Transcription. Genome Biol
623 Evol. 2015;7(4):1082–97
624 31. Du J, Leung A, Trac C, Lee M, Parks BW, Lusis AJ, Natarajan R, Schones DE.
625 Chromatin variation associated with liver metabolism is mediated by transposable
626 elements. Epigenetics Chromatin. 2016;9:28.
627 32. Rayan NA, Del Rosario RCH, Prabhakar S. Massive contribution of transposable
628 elements to mammalian regulatory sequences. Semin Cell Dev Biol. 2016;57:51–6.
629 33. Trizzino M, Park S, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry
630 GH, Lynch V, Brown CD. Transposable elements are the primary source in the
631 primate gene regulation. Genome Res. 2017;27:1623–33.

632 34. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous
633 retrovirus activity. Nature. 2012;487:57–63.
634 35. Simonti CN, Pavlicev M, Capra JA. Transposable Element Exaptation into
635 Regulatory Regions is Rare, Influenced by Evolutionary Age, and Subject to
636 Pleiotropic Constraints. Mol Biol Evol. 2017;34(11):2856–69.
637 36. Ward M, Zhao S, Luo K, Pavlovic B, Karimi MM, Stephens M, Gilad Y. Silencing of
638 transposable elements may not be a major driver of regulatory evolution in primate
639 induced pluripotent stem cells. eLife. 2018
640 37. Roadmap Epigenomics Mapping Consortium. Integrative analysis of 111 reference
641 human epigenomes. Nature. 2015;518:317–30.
642 38. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature.
643 2017;550:204–13.
644 39. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M,
645 Feschotte C. Transposable elements are major contributors to the origin,
646 diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet.
647 2013;9:e1003470.
648 40. Klaver B, Berkhout B. Comparison of 5' and 3' long terminal repeat promoter function
649 in human immunodeficiency virus. J Virol. 1994;68(6):3830–40.
650 41. Lavie L, Esther Maldener E, Brook Brouha B, Meese EU, Mayer J. The human L1
651 promoter: Variable transcription initiation sites and a major impact of upstream
652 flanking sequence on promoter activity. Genome Res. 2004;14:2253–60.
653 42. Walter M, Teissandier A, Pérez-Palacios R, Bourchis D. 2016. An epigenetic switch
654 ensures transposon repression upon dynamic loss of DNA methylation in embryonic
655 stem cells. Elife. 2016;5:e11418.
656 43. Daussy C, et al. T-bet and Eomes instruct the development of two distinct natural
657 killer cell lineages in the liver and in the bone marrow. J Exp Med. 2014;3:563–77.
658 44. Gao F, et al. Integrated analyses of DNA methylation and hydroxymethylation reveal
659 tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular
660 carcinoma. Genome Biol. 2014;15:533–46.
661 45. Guo P, Dong X-Y, Zhao KW, Sun X, Li Q, Dong J-T. Estrogen-induced interaction
662 between KLF5 and estrogen receptor (ER) suppresses the function of ER in ER-
663 positive breast cancer cells. Int J Cancer. 2010;126(1):81–9.
664 46. Chen C, Bhalala HV, Qiao H, Dong JT. A possible tumor suppressor role of the KLF5
665 transcription factor in human breast cancer. Oncogene. 2002;21:6567–6572.
666 47. Nagaoka K, Fujii K, Zhang H, Usuda K, Watanabe G, Ivshina M, Richter JD. CPEB1
667 mediates epithelial-to-mesenchyme transition and breast cancer metastasis. Oncogene.
668 2016;35:2893–901.
669 48. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in
670 the human genome. Nature. 2012;489:57–74.
671 49. Furuyama K, et al. 2010. Continuous cell supply from a Sox9-expressing progenitor
672 zone in adult liver, exocrine pancreas and intestine. Nature Genetics. 2010;43(1):35–
673 42.

674 50. Cattoretti G, Angelin-Duclos C, Shaknovich R, Zhou H, Wang D, Alobeid B.
675 PRDM1/Blimp-1 is expressed in human B-lymphocytes committed to the plasma cell
676 lineage. J Pathol. 2005;206:76–86.
677 51. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for
678 human genes: a critical assessment. Gene. 2009;448:105–14.
679 52. Janoušek V, Laukaitis CM, Yanchukov A, Karn R. The role of retrotransposons in
680 gene family expansions in the human and mouse genomes. Genome Biol Evol.
681 2016;8:2632–50.
682 53. Saykally JN, Dogan S, Cleary MP, Sanders MM. The ZEB1 Transcription Factor Is a
683 Novel Repressor of Adiposity in Female Mice. PlosONE. 2009;4(12):e8460.
684 54. Gubelmann C, et al. Identification of the transcription factor ZEB1 as a central
685 component of the adipogenic gene regulatory network. eLIFE. 2014;3:e03346.
686 55. Leow SC, et al. The transcription factor SOX6 contributes to the developmental
687 origins of obesity by promoting adipogenesis. Development. 2016;143:950-61.
688 56. Alexandrov IM et al. 2012. Cytoplasmic Polyadenylation Element Binding Protein
689 Deficiency Stimulates PTEN and Stat3 mRNA Translation and Induces Hepatic
690 Insulin Resistance. Plos Genet. 2012;8(1):e1002457.
691 57. He G, Karin M. NF-κB and STAT3 – key players in liver in ammation and cancer.
692 Cell Res. 2011;21:159-68.
693 58. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long
694 noncoding RNAs. 2012;13(11):R107.
695 59. Beck, CM et al. LINE-1 Retrotransposition Activity in Human Genomes. Cell.
696 2010;141:1159–70.
697 60. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic
698 mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature.
699 2005;435(7044):903–10.
700 61. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M,
701 O'Shea KS, Moran JV, Gage FH. L1 retrotransposition in human neural progenitor
702 cells. Nature. 2009;460(7259):1127–31.
703 62. Upton KR, et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell.
704 2017;161:228–39.
705 63. Sur D, et al. Detection of the LINE-1 retrotransposon RNA-binding protein ORF1p in
706 different anatomical regions of the human brain. Mobile DNA. 2017;8:17.
707 64. Savage AL, et al. An evaluation of a SVA retrotransposon in the FUS promoter as a
708 transcriptional regulator and its association to ALS. Plos One. 2014;9(6):e90833.
709 65. Smit A, Hubley R, Green P. RepeatMasker Open 4.0. 2013–2015. http://
710 www.repeatmasker.org.Benjamini Y, Hochberg Y. Controlling the false discovery
711 rate: a practical and powerful approach to multiple testing. J R Stat Soc B.
712 1995;57:289–300.
713 66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
714 features. Bioinformatics. 2010;26:841–2.
715 67. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a prac- tical and
716 powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300.

68. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–08
69. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. https://www.R-project.org/.
70. Wickham H. ggplot2: elegant graphics for data analysis. 2009. Springer-Verlag, New York.

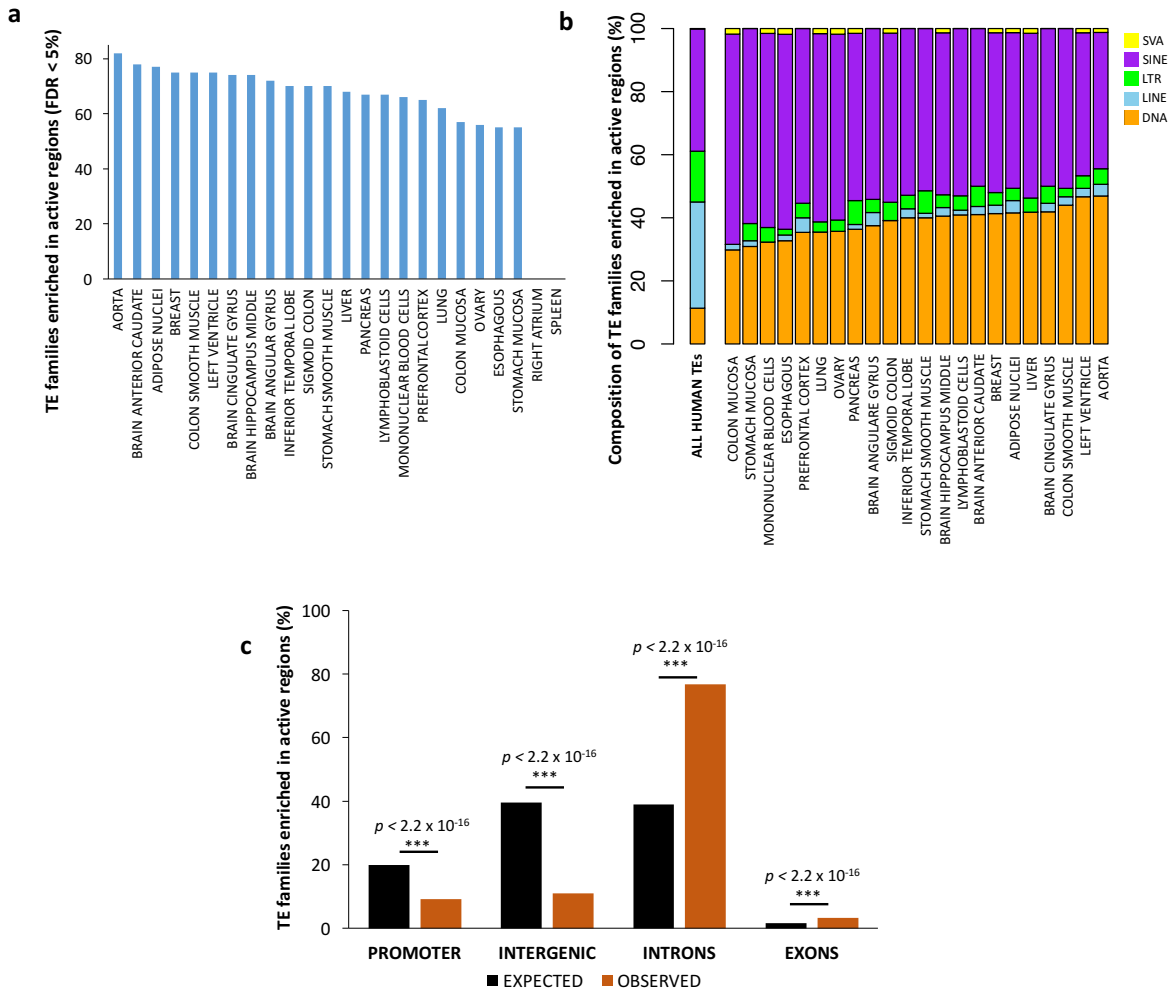

Figure 1 - Transposable elements are enriched in active genomic regions. (A) The plot displays the numbers of enriched TE families in the active genomic regions for each tissue (FDR <5%). The distribution suggests a tissue-specific pattern. (B) Stacked-bar charts show TE class composition for the TE families enriched in active regions (FDR <5%). SINE and DNA transposons are the dominant TEs enriched in active regions. (C) The TEs enriched in active regions are depleted from promoters and intergenic regions, while they are significantly enriched in intronic regions.

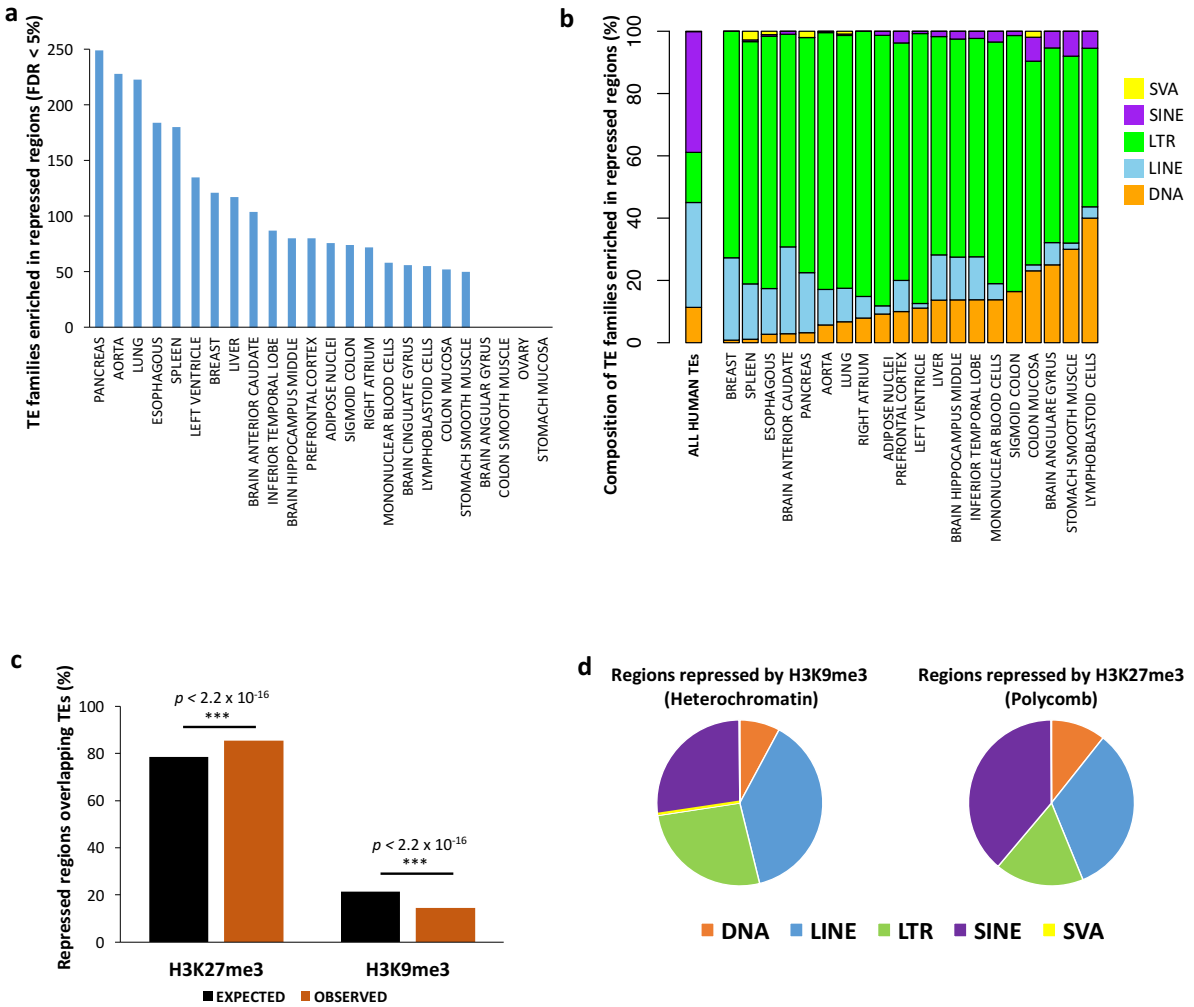

Figure 2 - Transposable elements are enriched in repressed genomic regions. (A) The plot displays the numbers of enriched TE families in the repressed genomic regions for each tissue (FDR <5%). The distribution suggests a tissue-specific pattern. (B) Stacked-chart plot shows class composition for the TE families enriched in repressed regions (FDR <5%). (C) Across tissues, the repressed TEs overlap H3K27me3 more than expected by chance, while H3K9me3 is underrepresented. (D) Pie-charts show class composition for the TEs overlapping H3K27me3 and H3K9me3.

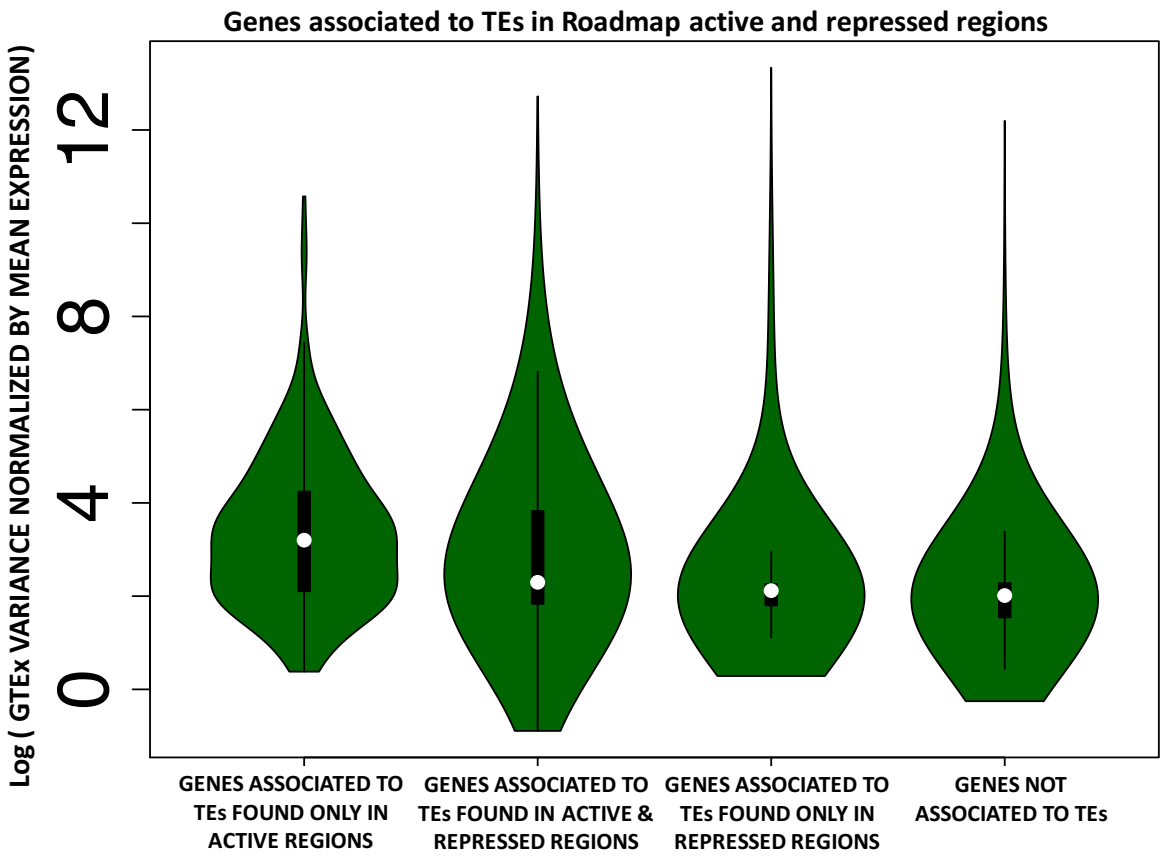

Figure 3 - Genes with higher expression variance are more tolerant towards TE expression. Human genes were split into four categories: 1) Genes associated with TEs that are only found in active regions across tissues; 2) Genes associated with TEs that are found in active or repressed regions in a tissue-specific fashion; 3) Genes associated with TEs that are only found in repressed regions; 4) Genes never associated with TE insertions. The violin plots display the distribution of the GTEx gene expression variance, normalized by mean expression, for each of the four categories.

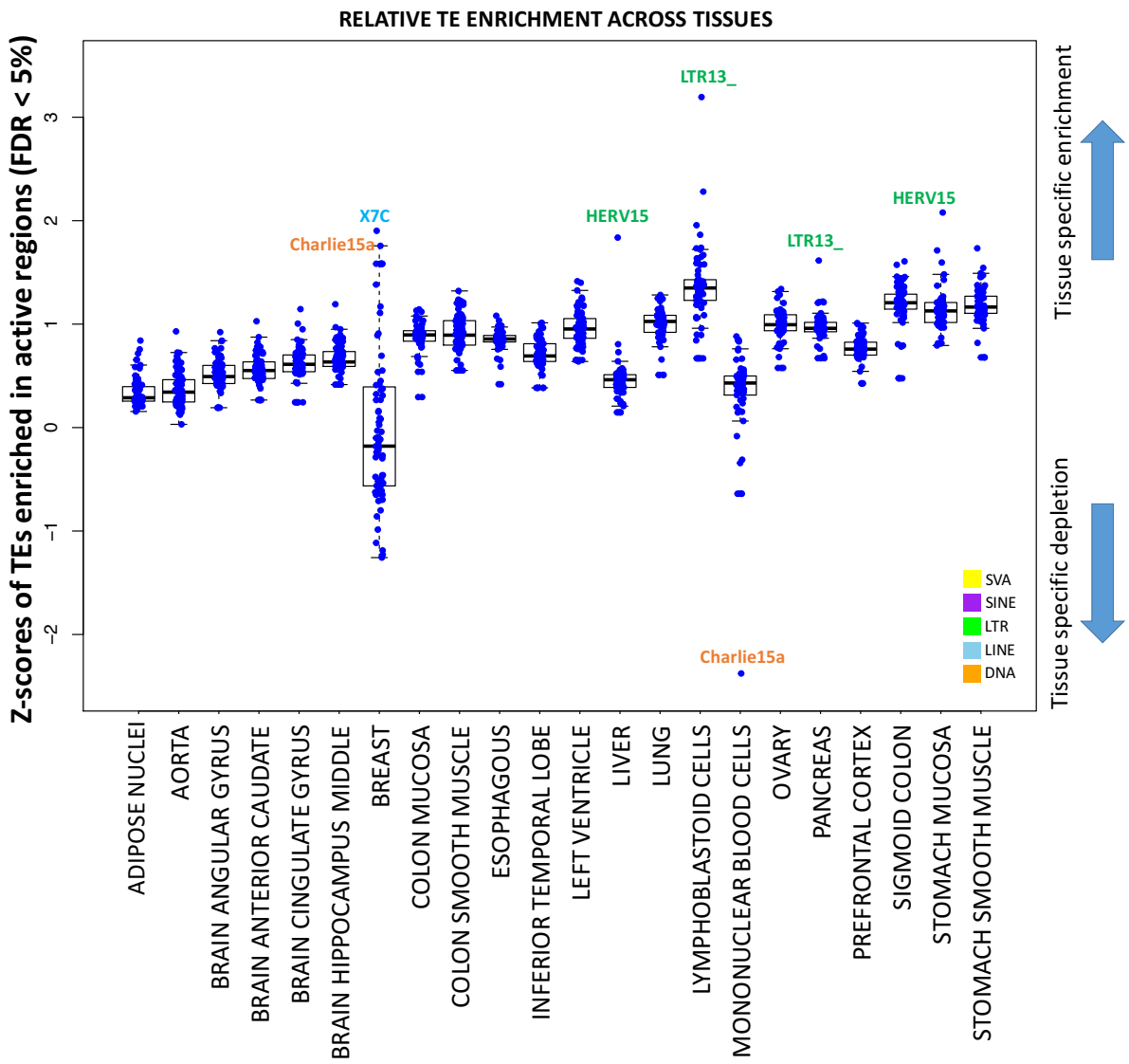

Figure 4 - Transposable elements have tissue-specific enrichment in active regions. The plot displays the distribution of the effect sizes (Z-scores from permutation test, see methods) for each TE enriched in active regions (FDR < 5%), in each tissue. The higher the Z-score, the more tissue-specific is the enrichment.

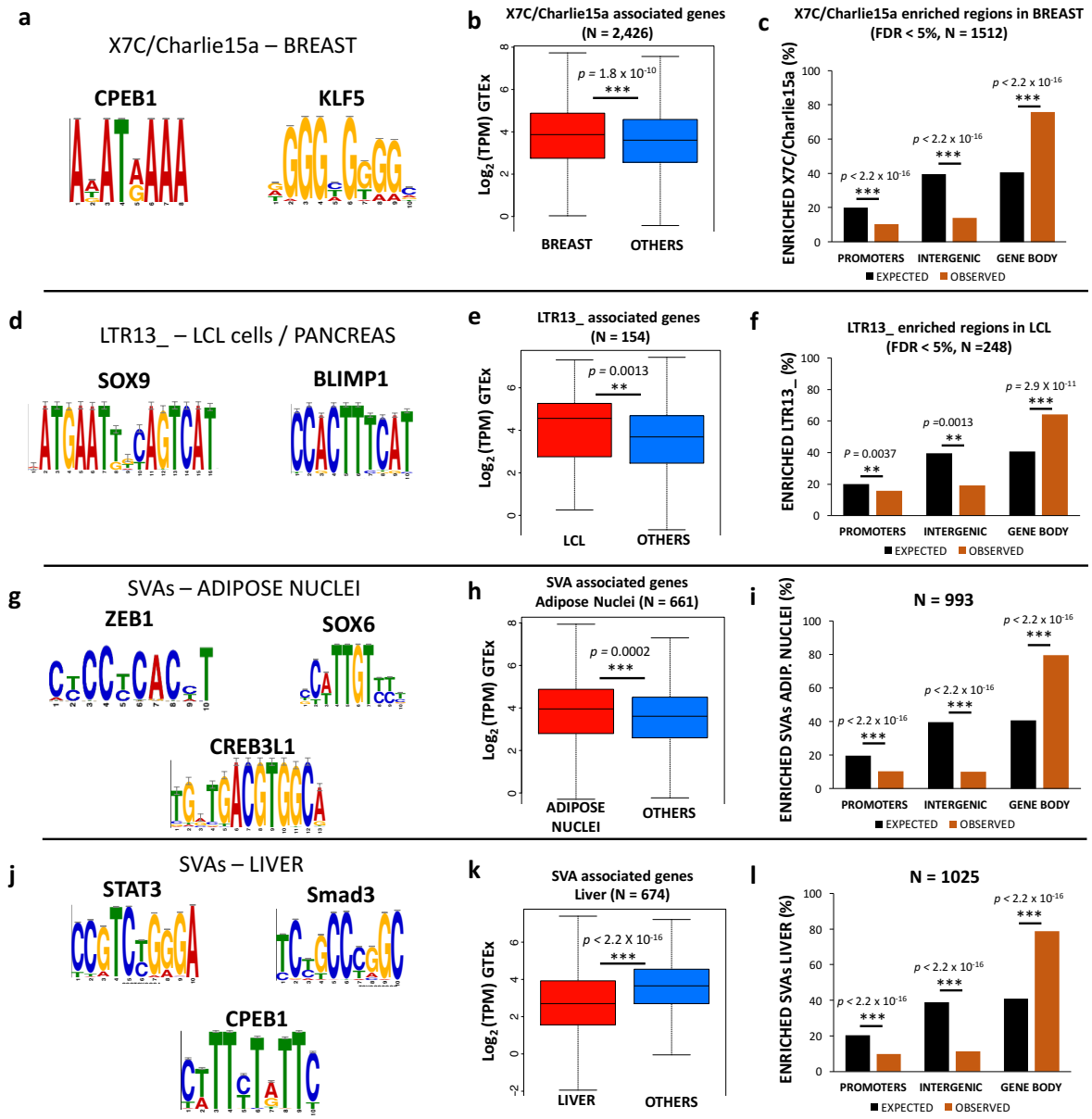

Figure 5 - Tissue-specific TEs are enriched for TF binding sites, are mostly intronic, and affect gene expression. (a) Motifs enriched in the regions overlapping X7C and and Charlie15a TEs in the breast. (b) Boxplot comparing mean expression for the genes associated to X7C and and Charlie15a in the breast vs all the other tissues. (c) Genomic distribution of the regions overlapping X7C and and Charlie15a TEs in the breast. (d) Motifs enriched in the regions overlapping LTR13_ TEs in pancreas and LCL cells. (e) Boxplot comparing mean expression for the genes associated to LTR13_ in the LCLs vs all the other tissues. (f) Genomic distribution of the regions overlapping LTR13_ in the LCLs. (g) Motifs enriched in the regions overlapping SVAs in the adipose nuclei. (h) Boxplot comparing mean expression for the genes associated to SVAs in the adipose nuclei vs all the other tissues. (i) Genomic distribution of the regions overlapping SVAs in the adipose nuclei. (j) Motifs enriched in the regions overlapping SVAs in the liver. (k) Boxplot comparing mean expression for the genes associated to SVAs in the liver vs all the other tissues. (l) Genomic distribution of the regions overlapping SVAs in the liver.