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@upenn.edu)

Abstract

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. Here, we leverage Roadmap and GTEx data to investigate the dynamics of TE recruitment 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 co-opted 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 to TE insertions. TE repression similarly follows tissue specific patterns, and LTRs are the most abundant class in repressed regions. Different TE classes are preferentially silenced in different ways: LTRs and LINEs are overrepresented in regions marked by H3K9me3, while other TEs are preferentially repressed by H3K27me3. Young TEs are typically enriched in repressed regions and depleted in active regions. We detect multiple instances of TEs with tissue specific regulatory activity. Such TEs provide binding sites for transcription factors that are master regulators for the given tissue. These tissue specific activated TEs are enriched in intronic enhancers, and significantly affect the expression of the associated genes. Finally we show that SVAs can act as transcriptional activators or repressors in a tissue specific context. We provide an integrated overview of the contribution of TEs to human gene regulation. Expanding previous analyses, we demonstrate that TEs generate regulatory novelty in a tissue specific fashion.

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

Introduction

Transposable elements (TEs) contribute to roughly half of the human genome. Several TE groups maintain transposing activity in humans, including Long Terminal Repeats (LTRs, mostly ERV1s; Tonjes et al. 1996; Medstrand et al. 1998; Fuchs et al. 2013), Long Interspersed Nuclear Elements (LINEs, mostly L1s; Kazazian et al. 1998; Brouha et al. 2003), Short Interspersed Nuclear Elements (SINEs) of the Alu families (Batzer and Deininger 1991; Batzer et al. 1991), and SINE-VNTR-*Alus* (SVAs; Ostertag et al. 2003; Wang et al. 2005).

Multiple elegant studies have demonstrated that TEs play a functional role in eukaryotic gene regulation (McClintock 1950, 1984; Britten and Davidson 1969; Davidson and Britten 1979; Jordan et al. 2003; Bejerano et al. 2006; Wang et al. 2007; Bourque et al. 2008; Sasaki et al. 2008; Markljung et al. 2009; Kunarso et al. 2010; Lynch et al. 2011, 2015; Schmidt et al. 2012; Chuong et al. 2013, 2016; Jacques et al. 2013; Xie et al. 2013; del Rosario et al. 2014; Sundaram et al. 2014; Du et al. 2016; Rayan et al. 2016). 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 (Trizzino et al. 2017). Similarly, other studies have shown that the recruitment of novel regulatory networks in the uterus was likely mediated by ancient mammalian TEs (Lynch et al. 2011, 2015), and that TEs have a role in pluripotency (Macfarlan et al. 2012). Conversely, other researchers have proposed that TE exaptation into regulatory regions is rare (Simonti et al. 2017), and that TE silencing may not be a major driver of regulatory evolution in primates (Ward et al. 2017).

Given these contrasting lines of evidence, we aimed to shed light on the contribution of TEs to the evolution of tissue specific human gene expression regulation. For this purpose, we took advantage of publicly available data (Roadmap Epigenomics Mapping Consortium 2015; GTEx Consortium 2017) to investigate the dynamics of TE recruitment in 24 tissues, and to characterize the contribution of TEs to the regulation of gene expression. A significant fraction of the existing human TEs are enriched in regions of the genome bearing hallmarks of active or repressed chromatin, suggesting they are actively regulated by the cellular machinery. DNA transposons and SINEs represent the most frequently enriched classes across tissues, while LTR-ERV1s are the TEs that more commonly show tissue specific enrichment and active regulatory activity. TE enrichment in active and repressed chromatin exhibits tissue specific patterns. Genes with consistent expression across tissues are less likely to be associated with a local TE insertion, further supporting the role of TEs in mediating tissue specific regulatory programs. We detect multiple instances of TEs with tissue specific activity, and demonstrate that they provide binding sites for transcription factors that are tissue specific master regulators.

Results

Specific TE families are enriched in active and repressed genomic regions

To investigate the extent to which TEs contribute to the regulation of human gene expression, we leveraged publicly available data from the Roadmap Epigenomics Project (2015) and from the GTEx Project (2017). We focused on 24 primary tissues and cell types that were processed by both consortia (Supplementary Table S1). Using five different histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K9me3, and H3K27me3), Roadmap segmented the human genome into 15 regulatory classes, reflecting different degrees and types of regulatory activity. We took advantage of this classification to define active and repressed genomic regions in each of the studied tissues.

To test for TE enrichment in active and repressed regions, we used the TE-Analysis pipeline (Kapusta et al. 2013; https://github.com/4ureliek/TEanalysis; Supplemental File S1). As expected, the majority of human TEs are significantly depleted from regions marked as active from the Roadmap histone modifications (mean 83.9% of TEs; FDR <5%; Supplementary Table S2). Nevertheless, 112 TE families (9.07% of the annotated TE families in the human genome) are significantly enriched in active regions of the genome in at least one tissue (FDR <5%; Fig. 1a; Supplementary Table S2). These data suggest variability across tissues: aorta, brain anterior caudate, and adipose are the most "permissive" tissues, while right atrium and spleen do not show any significant TE enrichment in active regions (Fig. 1a).

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

Conversely, LTRs and LINEs are significantly depleted from active genomic regions of all tissues (Proportion Test $p < 2.2 \times 10^{-16}$ for each tested tissue; Fig. 1b).

Finally, SINE-VNTR-*Alus* (SVAs), which are the least abundant TEs in the human genome (0.12% of the total annotate TEs in the human genome), are significantly overrepresented in active chromatin in 13/24 tissues; Fig. 1b).

We set out to investigate the TEs overlapping active regions. These TEs are depleted in active promoters and intergenic regions, but significantly enriched within active regions inside gene bodies (Fisher's Exact Test *p-values* in Fig. 1c). A possible interpretation for these results could be that genomic regions containing active genes are more frequently accessible, providing a substrate for TEs to insert. Moreover, we speculate that the TEs present in the bodies of active genes are less likely to be silenced than TEs in intergenic regions.

Using the same approach previously described for the active regions, we searched for TEs enriched in repressed genomic regions. Overall, 314 human TE families (25.4%) are significantly enriched in repressed regions of the genome in at least one tissue (FDR <5%; Fig. 2a; Supplementary Table S3). LTRs (predominantly ERV1) represent the large majority of the repressed TEs (Fig. 2b), followed by LINEs (predominantly L1s) and DNA TEs. Notably, ERV LTRs and L1 LINEs are among the most active in the genome, and also have their own regulatory architecture (Klaver et al. 1994; Lavie et al. 2004). We thus surmise that these autonomous active TEs are preferential targets of repressive marks.

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

Different TE repression patterns in the human genome

We examined whether TEs are preferentially repressed via Polycomb Repressive Complex (H3K27me3) or via Heterochromatin (H3K9me3). Overall, 78.6% of the regions classified as repressed in the human genome across all tissues are bound by H3K27me3 (Polycomb Repressive Complex), while 21.4% are marked by H3K9me3 (Heterochromatin conformation). However, when we restrict the analysis to the repressed regions containing a TE, we report an overall higher than expected overlap with H3K27me3 (median across tissues 85.5%; Proportion Test across tissues *p* < 2.2 x 10^{-16}; Supplementary Table S4; Fig. 2C), and a consequent underrepresentation of H3K9me3 (median 15.5%; Supplementary Table S4; Proportion Test *p* < 2.2 x 10^{-16} ; Fig. 2d). In 20/24 of the tested tissues, TEs are marked by H3K27me3 more than expected by chance (Proportion Test *p* < 2.2 x 10^{-16} for each of the 20 significant tissues; Supplementary Table S4). In the remaining four tissues this histone mark is instead underrepresented, while H3K9me3 is overrepresented: breast (H3K27me3 = 76.4%; Supplementary Table S4; Proportion Test *p* < 2.2 x 10^{-16}), aorta (55.1%; Supplementary Table S3; *p* < 2.2 x 10^{-16}), lung (48.9%; Supplementary Table S4; *p* < 2.2 x 10^{-16}), and spleen (26.5%; Supplementary Table S3; *p* < 2.2 x 10^{-16}). Notably, in these four tissues we detect the highest numbers of TE families enriched in repressed regions (Fig. 2a), and the highest proportion of repressed LINEs. This suggests that the heterochromatin state (H3K9me3) may be employed to target specific TE classes and families in a context specific manner (Ward et al. 2017).

We therefore tested whether different TE classes are preferentially repressed by heterochromatic configuration (H3K9me3) or by Polycomb (H3K27me3). LTRs, LINEs, and SVAs are overrepresented in regions marked by H3K9me3 (Fisher's Exact Test *p* < 2.2 x 10^{-16} ; Fig. 2d). Conversely, SINEs and DNA TEs are preferentially repressed by H3K27me3. Notably, SVAs are depleted from the regions marked by H3K27me3 (Fig. 2d).

These findings are consistent with recent reports suggesting that H3K27me3 and H3K9me3 target different transposon types in embryonic stem cells (Walter et al. 2016), and with a study reporting that LINEs, LTRs, and SVAs are the most abundant TEs repressed by H3K9me3 in induced pluripotent stem cells (Ward et al. 2017).

Ancient TEs are enriched in active regions, while young TEs are repressed

We clustered the annotated human TEs in 35 age classes as in ref. (Kapusta et al 2013; e.g. Eutheria, Primates, Hominidae; Supplemental Table S6), and used the TE-Analysis shuffling script to test for enrichment of each age class in a given set of regions (see Methods). Using this approach, we assessed the age of TEs enriched in active and repressed genomic regions. Ancient TE classes (i.e. age classes older than the Eutheria lineage) are enriched in the active regions of all tested tissues (FDR <5%; Supplemental Table S6). These TEs are largely vertebrate or mammalian specific (Supplemental Table S6). Notably, the only tissues with an enrichment of young TEs (specifically primate specific) are blood related (Mononuclear and Lymphoblastoid Cells). These results are in agreement with an elegant study that discovered a key role of primate specific TEs in the regulatory evolution of immune response (Chuong et al. 2016). TE families enriched in active regions across at least 20 of the 24 tissues correspond to DNA TEs and SINEs

(Supplemental Table S2). Despite a lack of general enrichment of young TEs in active regions, 24 *Alu* families are in fact enriched in active regions.

In contrast, young TEs (i.e. TE classes younger than the Eutheria lineage split) are significantly enriched in the repressed regions of most tissues. In particular human specific TEs are enriched in the repressed regions of all brain related tissues (FDR <5%; Supplemental Table S6). These young TEs correspond to ERV LTRs, L1 LINEs and SVAs, but only one family is found enriched in at least 20 tissues (MER52A), which is in line with the large variability across tissues of the TE composition of repressed regions (see above). Collectively, these data suggest that young TEs are predominantly silenced, while older TEs are now more tolerated.

TE insertions are associated with gene expression variance across tissues

We used GTEx data to test if TE insertions affected local gene expression. For this purpose, we first assigned each TE overlapping an active genomic region to its nearest gene transcription start site (TSS). Next, we divided all human genes in four categories (Supplemental Table S7): 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. Based on this classification, genes associated with a TE insertion in regions that are active in at least one tissue are characterized by significantly higher expression variance (normalized by mean expression) than genes either associated to repressed TEs or not associated to a TE (Wilcoxon's Rank Sum Test $p < 2.2 \times 10^{-16}$; Fig. 3). On the other hand, we do not detect significant differences when comparing genes associated to TEs only present in active regions, to genes with

TEs present in both active and repressed regions (Wilcoxon's Rank Sum Test *p* > 0.05; Fig. 3).

These findings suggest that genes with local TEs overlapping active chromatin have higher variability in gene expression across tissues, and that genes consistently expressed across tissues (e.g. housekeeping and other essential genes) may be less tolerant towards TE insertions in their regulatory regions.

Tissue specific activity of TEs

We next investigated whether specific TE families display differential activity across tissues. We compared the relative enrichment in active regions of each TE family across tissues (see methods). TE enrichment varies substantially across tissues (Supplemental Table S5; Fig. 4), and many TEs exhibit tissue specific activity (Fig. 4). For example, HERV15 (LTR) is significantly more enriched in the liver and in the stomach mucosa compared to any other tissue. Motif analysis revealed regions of active histone modification in the liver overlapping HERV15 are enriched in motifs for EOMES (Supplemental File S2). This transcription factor (TF) has a key role in the immune response in the liver, instructing the development of two distinct natural killer cell lineages specific to this tissue (Daussy et al. 2014). Moreover, EOMES is also an established tumor suppressor in Hepatocellular Carcinoma (Gao et al. 2014).

Similarly, X7C (LINE) and Charlie15a (DNA TE), are the most enriched TEs in the breast. We find binding sites for key breast TFs in these TEs such as KLF5 and CPEB1 (Fig. 5a; Supplemental File S2). Notably, KLF5 is an essential regulator of hormonal signaling and breast cancer development (Guo et al. 2010), and is considered a breast cancer suppressor (Chen et al. 2002). Similarly, CPEB1 mediates epithelial-to-mesenchyme transition in breast, and mice depleted of this gene showed increased breast cancer metastatic potential (Nagaoka et al. 2016). Interestingly Charlie15a shows tissues-specific depletion in the mononuclear blood cells (Fig. 4), highlighting its tissue specific regulatory activity.

Analogously, LTR13 is the most active TEs in pancreas and Lymphoblastoid Cells (LCL). These LTR copies are enriched for binding sites for SOX9 and PRDM1/Blimp-1 (Fig. 5d; Supplemental File S2). SOX9 is a master regulator of the pancreatic program (Furuyama et al. 2010), while PRDM1/Blimp-1 has a central role in determining and shaping the secretory arm of mature B Lymphocyte differentiation (Cattoretti et al. 2005).

We next tested whether the co-option of these tissue specifically enriched TEs (Fig. 4, 5a–f) affected the expression of the associated genes. Specifically, we tested the TE families showing the highest degree of tissue specific enrichment (Fig. 4: HERV15/liver, LTR13/LCL, X7C-Charlie15a/breast). With the exception of HERV15/liver (Wilcoxon's Rank Sum Test p > 0.05), in the other tested instances (LTR13/LCL; X7C-Charlie15a/breast) the genes associated with tissue specific TEs are significantly more highly expressed in the tissue in which they are actively enriched compared to all the other tissues (Wilcoxon's Rank Sum Test $p < 2.2 \times 10^{-16}$; Figs. 5b,e). These findings support a key regulatory role for the co-opted TEs.

To better understand how these tissue specific TEs regulate gene expression, we investigated what typology of genomic region they overlap (i.e. promoter, intergenic, gene body). Both X7C/Charlie15a in breast and LTR13 in LCLs are significantly depleted in promoter and intergenic regions, but overrepresented in gene bodies (Figs. 5c, f), 97.8% (X7C/Charlie15a) and 96.4% (LTR13) of them respectively found in introns.

The Roadmap data did not include H3K27ac profiles for all tissues. Therefore, to further characterize these intronic regions, we used publicly available H3K27ac and H3K4me1 Encode data generated from the breast epithelium and from the MCF7 cell line (The Encode Project Consortium 2012). These data reveal that 53.0% of the intronic regions containing X7C or Charlie15a are found within 1 Kb of a H3K27ac or H3K4me1 peak, thus suggesting that most of these regions likely represent breast intronic enhancers. As comparison, only 33.7% of random intronic regions of the same size and number of the ones overlapping X7C/Charlie15a TEs are found within 1 kb of a H3K27ac or H3K4me1 peak (Fisher's Exact Test $p < 2.2 \times 10^{-16}$).

Collectively, these findings point towards a model in which specific TE families, largely belonging to LTR (ERVs) and DNA TE classes, have more regulatory potential than other transposons. Furthermore, our data expand upon previous findings that ERVs that escape repression can have a significant impact on the host gene regulation (Wang et al. 2005; Cohen et al. 2009; Jacques et al. 2013; Chuong et al. 2016; Janoušek et al 2016; Trizzino et al. 2017).

SVAs exhibit tissue specific regulatory activity

In our recent work, we demonstrated that a large fraction of human specific cis-regulatory elements in the liver are SVA transposons, which typically function as transcriptional repressors, at least in this tissue (Trizzino et al. 2017). SVAs are very young transposons, being Hominidae (SVA_A, B, C and D) and human specific (SVA_E and F).

According to Roadmap data, SVAs are enriched in the active regions of 13/25 tissues (Fig. 1b), and mainly corresponded to SVA_A copies (Supplementary Table

S4). We first assessed the contribution of SVAs to gene regulation of two of these tissues: the adipose nuclei and the liver. We chose the liver because of the previous evidence of SVA enrichment in hepatic regulatory regions (Trizzino et al. 2017), and the adipose nuclei as a “control tissue” since it is involved in most of the metabolic pathways that also involve the hepatocytes.

In both tissues, SVAs provide binding sites for key transcription factors (Fig. 5g, j; Supplemental File S2). ZEB1 is the master regulator of adipogenesis (Saykally et al. 2009; Gubelmann et al. 2014), and, based on GTEx data, is ten times more highly expressed in adipose tissue compared to the liver. Similarly, SOX6 contributes to the developmental origin of obesity by promoting adipogenesis, and has a key role in adipocyte differentiation (Leow et al. 2016). Consistent with the data reported for other tissues, active SVAs associated with adipose nuclei and liver are strongly enriched in gene bodies (Figs. 5i, l). Genes associated with SVAs in the adipose nuclei are significantly more highly expressed in this tissue compared to other tissues (Wilcoxon’s Rank Sum Test p = 0.0002; Fig. 5h), suggesting that SVA elements can work as transcriptional activators, at least in the adipose tissue.

In the liver, SVAs in active regions are enriched for hepatic regulators like CPEB1, that mediates insulin signaling in the liver (Fig. 5j; Alexandrov et al. 2012), and STAT3, that regulates liver regeneration and immune response and negatively modulates insulin action (Fig. 5j; He et al. 2011). However, the liver SVAs are also enriched for established transcriptional repressors, like Smad3 (Fig. 5j). Consistently, genes associated with liver active SVAs exhibit lower expression in this tissue compared to all the others (Wilcoxon’s Rank Sum Test $p < 2.2 \times 10^{-16}$; Fig. 5k), supporting the previously proposed repressive role of SVAs in the hepatic system (Trizzino et al. 2017).

Discussion

The contribution of transposable elements (TEs) to gene regulation was proposed over half a century ago (McClintock 1950, 1984; Britten and Davidson 1969, 1979) and considerably expanded over the last two decades, largely due to the advances in next generation sequencing (Jordan et al. 2003; Bejerano et al. 2006; Wang et al. 2007; Bourque et al. 2008; Sasaki et al. 2008; Markljung et al. 2009; Kunarso et al. 2010; Lynch et al. 2011, 2015; Schmidt et al. 2012; Chuong et al. 2013, 2016; Jacques et al. 2013; Xie et al. 2013; del Rosario et al. 2014; Sundaram et al. 2014; Du et al. 2016; Rayan et al. 2016; Simonti et al. 2017; Trizzino et al. 2017; Ward et al. 2017).

In order to gain insights in this topic, we identified TEs enriched in active and repressed genomic regions of 24 human tissues, using Roadmap and GTEx data. Our analyses provide a novel integrated overview of the TE contribution to the human gene regulation across multiple tissues, also correlating the presence of TE copies to tissue specific gene expression. In fact, many of the previous studies have proposed that TEs are frequently enriched in cis-regulatory elements and lncRNAs (Lynch et al. 2011, 2015; Kelley et al. 2012; Kapusta et al. 2013; Trizzino et al. 2017), but the actual effect of the presence of TEs on the associated gene expression was not tested on a large scale. Recent work has evaluated the prevalence of TE-derived DNA in enhancers and promoters across mouse cell lines and primary tissues (Simonti et al. 2017). The present study builds upon this by investigating the effects of TE recruitment on tissue-specific gene expression and characterizing the mechanisms of TE co-option by performing tissue-specific motif analyses.

We demonstrate that ~10% of the TEs identified in the human genome are significantly enriched in active regions (promoters, intergenic enhancers, gene bodies) of 24 different human tissues. In general, we report a high degree of variability in the co-option and repression of different TE families across tissues, and detect multiple instances of TEs displaying tissue specific regulatory function.

We show that DNA TEs are generally enriched in active regions, suggesting a high regulatory potential for these elements (i.e. high potential of providing binding sites for transcription factors), likely a main factor driving TE co-option (Chuong et al. 2016). These copies regulate gene expression in a tissue specific fashion, mainly functioning as substrate for the binding of key, tissue specific, master regulators.

Co-opted TEs are typically distributed along gene bodies, likely functioning as intronic enhancers. We reason that this may be explained by the assumption that TEs located within intra-genic regions are less likely to be repressed or removed. . In agreement with these findings, a recent study has shown that TEs are depleted in human promoters and intergenic enhancers across multiple tissues (Simonti et al. 2017). In this context, we see a correlation between gene expression variance and the insertion of TEs in their loci or regulatory regions. This may suggest that genes consistently expressed across tissues are less prone towards TE co-option, but future analyses in this direction will be needed to further characterize this phenomenon. On the other hand, L1 LINEs and ERV LTRs are the most frequently enriched TE classes in the repressed regions. L1 retrotransposons are among the most active TEs in the human genome (Beck et al. 2010), and several studies have demonstrated that they are also active in brain tissues (e.g. hippocampus), and can contribute to neuronal genetic diversity in mammals (Muotri et al. 2005; Coufal et al. 2009; Sur et al. 2017; Upton et al 2017). Both L1s and LTRs possess their own regulatory architecture, and we speculate that their preferential silencing prevents these TEs from interfering with gene regulatory networks. Despite this, we demonstrate that LTRs that escape repression may be co-opted in a tissue specific manner in the active regulatory regions, putatively as a consequence of their regulatory potential.

In agreement with recent reports (Walter et al. 2016; Ward et al. 2017), different TE classes are preferentially silenced through different modalities: LTRs, LINEs, and SVAs are preferentially repressed via heterochromatin conformation, while SINEs and DNA TEs are predominant among the TEs silenced by the Polycomb Repressive Complex. We show that TEs enriched in repressed regions of most tissues are generally young, while TEs enriched in active regions of most tissues generally predate the split of eutherian mammals. This is consistent with an accumulation of mutations in these ancient copies that would have increased the likelihood to generate binding sites for transcription factors, and thus the probability for the TE to be co-opted in the regulatory networks.

Finally, we demonstrate that SVAs, previously characterized as transcriptional repressors in select cell-types (Savage et al. 2014; Trizzino et al. 2017), can act as both activators or repressors in a tissue specific fashion.

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 TEanalysis pipeline v 4.6 (Kapusta et al. 2013; 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 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 bedfiles, we produced two files: one for the active regions (Roadmap annotations: "TssA", "TssAFlnk", "TxFlnk", "Tx", "TxWk", "EnhG", "Enh", "TssBiv", “EnhBiv”), and one for the repressed regions (Roadmap annotations: "Het", "ReprPC", "ReprPCWk").

For each tissue, we tested for TE enrichment in the “active” and “repressed” bed files using the “TE-analysis_Shuffle_bed.pl” script v 4.3. Specifically, this script assesses which TEs are significantly enriched in a set of features (bed files) by comparing observed overlaps with the average of *N* expected overlaps (here 1000). These expected overlaps were obtained by shuffling the genomic position of TEs. TE annotations were downloaded from the University of California Santa Cruz Genome Browser (RepeatMasker, Hg19 version; Smit et al. 2015–2013).

The "TE-analysis_Shuffle_bed.pl" script was run with Bedtools v2.27.1 (Quinlan and Hall 2010) and the following parameters:

-f Roadmap_BEDFILE (active or repressed)

-q RepeatMasker.out (TE file, hg19)

-n 1000 (number of bootstrap replicates)

-r hg19.chrom.sizes

-g 20141105_hg38_TEage_with-nonTE.txt (distributed with the pipeline)

-s rm (shuffles the TEs within their genomics position)

The script performs a two-tailed permutation test to assess the enrichment (or depletion) of each annotated TE in the given regions (Roadmap regions), thus assigning a *p-value* to each annotated TE. Additionally, we corrected for multiple testing by applying a False Discovery Rate (FDR; Benjamini-Hochberg; Benjamini and Hochberg 1995). Only TEs with FDR < 5% were retained, considered significantly enriched in the given tissue, and used for downstream analyses.

Composition of enriched TEs

To characterize TEs enriched within active and repressed regions of each tissue (e.g. Figs. 1b, 2b), each TE was assigned to one of the major TE classes: DNA transposons, LINEs, LTRs, SINEs, SVAs, according to RepeatMasker annotations. To assess the genomic composition of the enriched TEs (e.g. Figs. 1c, 2c), we considered as: 1) PROMOTERS all of the regions annotated as "TssA" (H3K4me3), "TssAFlnk" (H3K4me3 + H3K4me1), or "TssBiv" (H3K4me3 + H3K27me3); 2) INTERGENIC ENHANCERS: all of the regions annotated as "Enh" (H3K4me1) or "EnhBiv" (H3K4me1 + H3K27me3); 3) GENE BODIES: all of the regions annotated as “EnhG” (H3K4me1 + H3K36me3), "Tx" (strong H3K36me3), or "TxWk" (weak H3K36me3).

Correlation between TE insertion and variance in gene expression

We calculated the variance and mean of the TPM (Transcripts Per Million) for each gene using GTEx data. We assigned each TE overlapping an active or a repressed region to the closest gene, based on the distance to the closest transcription start site. Next, we divided all human genes in 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. Gene expression variance, normalized by mean expression, was compared across the four categories.

Computation of Z-scores for tissue specificity

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

Motif analyses

Motif analyses were performed using the Meme-Suite (Bailey et al. 2009), and specifically with the Meme-ChIP application. Fasta files of the regions of interest were produced using BEDTools v2.27.1. Shuffled input sequences were used as background. *E-values* < 0.001 were used as threshold for significance (Bailey et al. 2009).

Effect of TE co-option on gene expression

For each human gene and for each tissue, GTEx provides the mean of the TPMs (Transcripts Per Million). For each gene associated with a TE of interest (e.g. TEs with tissue specific activity, or TEs associated with an SVA), we used the mean TPMs to compare the expression of genes in the tissue of enrichment Vs the average of the gene expression of the same genes in all the other considered tissues (i.e. mean of TPMs across all the other tissues).

Statistical and genomic analyses

All statistical analyses were performed using R v3.4.1 (R Core Team 2016). Figures were made with the package ggplot2 (Wickham 2009). BEDTools v2.27.1 was used for all the genomic analyses.

Acknowledgements

We thank Roadmap and GTEx Consortia for the generation of invaluable data. MT thanks his current P.I. (Alessandro Gardini, The Wistar Institute) who granted him time and freedom to work on this project.

Authors’ contributions

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

References

1. Alexandrov IM et al. 2012. Cytoplasmic Polyadenylation Element Binding Protein Deficiency Stimulates PTEN and Stat3 mRNA Translation and Induces Hepatic Insulin Resistance. Plos Genet. 8(1):e1002457.
2. Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:W202–08.
3. Batzer MA, Deininger PL. 1991. A human-specific subfamily of Alu sequences. Genomics. 9:481-7.
4. Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL. 1991. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 19:3619–23.
5. Beck, CM et al. 2010. LINE-1 Retrotransposition Activity in Human Genomes. Cell.
6. 2010;141:1159–70.
7. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 441:87–90.
8. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding
9. repertoire via transposable elements. Genome Res. 18:1752–62.
10. Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. Science.
11. 165:349–57.
12. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci USA. 100:5280–5.
13. Cattoretti G, Angelin-Duclos C, Shaknovich R, Zhou H, Wang D, Alobeid B. 2005. PRDM1/Blimp-1 is expressed in human B-lymphocytes committed to the plasma cell lineage. J Pathol. 206:76–86.
14. Chen C, Bhalala HV, Qiao H, Dong JT. 2002. A possible tumor suppressor role of the KLF5 transcription factor in human breast cancer. Oncogene. 21:6567–6572.
15. Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 351:1083–7.
16. Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 45:325–9.
17. Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene. 448:105–14.
18. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. Nature. 460(7259):1127–31.
19. Daussy C, et al. 2014. T-bet and Eomes instruct the development of two distinct natural killer cell lineages in the liver and in the bone marrow. J Exp Med. 3:563–77.
20. Davidson EH, Britten RJ. 1979. Regulation of gene expression: possible role of repetitive sequences. Science. 1979;204:1052–9.

21. del Rosario RCH, Rayan NA, Prabhakar S. 2014. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. Genome Res. 24:1469–84.
22. Du J, Leung A, Trac C, Lee M, Parks BW, Lusis AJ, Natarajan R, Schones DE. 2016. Chromatin variation associated with liver metabolism is mediated by transposable
23. elements. Epigenetics Chromatin. 9:28.
24. Fuchs NV, Loewer S, Daley GQ, Izsvak Z, Lower J, Lower R. 2013. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. Retrovirology;10:115.
25. Furuyama K, et al. 2010. Continuous cell supply from a Sox9-expressing progenitor
26. zone in adult liver, exocrine pancreas and intestine. Nature Genetics. 43(1):35–42.
27. Gao F, et al. 2014. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. Genome Biol. 15:533–46.
28. GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. Nature. 550:204–13.
29. Gubelmann C, et al. 2014. Identification of the transcription factor ZEB1 as a central
30. component of the adipogenic gene regulatory network. eLIFE. 3:e03346.
31. Guo P, Dong X-Y, Zhao KW, Sun X, Li Q, Dong J-T. 2010. Estrogen-induced interaction between KLF5 and estrogen receptor (ER) suppresses the function of ER in ER-positive breast cancer cells. Int J Cancer. 126(1):81–9.
32. He G, Karin M. 2011. NF-κB and STAT3 – key players in liver in ammation and cancer. Cell Res. 21:159-68.
33. Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet 9: e1003504.
34. Janoušek V, Laukaitis CM, Yanchukov A, Karn R. 2016. The role of retrotransposons in gene family expansions in the human and mouse genomes. Genome Biol Evol.
35. 8:2632–50.
36. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19:68–72.
37. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 9:e1003470.
38. Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 332:164–6.
39. Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. 13(11):R107.
40. Klaver B, Berkhout B. 1994. Comparison of 5' and 3' long terminal repeat promoter function in human immunodeficiency virus. J Virol. 68(6):3830–40.

41. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 42:631–4.
42. Lavie L, Esther Maldener E, Brook Brouha B, Meese EU, Mayer J. 2004. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res. 14:2253–60.
43. Leow SC, et al. 2016. The transcription factor SOX6 contributes to the developmental origins of obesity by promoting adipogenesis. Development. 143:950-61.
44. Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat
45. Genet. 43:1154–9.
46. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F, Bauersachs S, et al. 2015. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 10:551–61.
47. Macfarlan TS, et al. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 487:57–63.
48. Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang X, Wang L, Saenz-Vash V, Gnirke A, et al. 2009. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. PLoS Biol. 7:e1000256.
49. McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA. 36:344–55.
50. McClintock B. 1984. The significance of responses of the genome to challenge. Science. 226:792–801.
51. Medstrand P, Mager DL. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol. 72:9782–7.
52. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature.
53. 435(7044):903–10.
54. Nagaoka K, Fujii K, Zhang H, Usuda K, Watanabe G, Ivshina M, Richter JD. 2016. CPEB1 mediates epithelial-to-mesenchyme transition and breast cancer metastasis. Oncogene. 35:2893–901.
55. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. 2003. SVA elements are non autonomous retrotransposons that cause disease in humans. Am J Hum Genet. 73:1444–51.
56. Pavlicev M, Hiratsuka K, Swaggart KA, Dunn C, and Muglia L. 2015. Detecting Endogenous Retrovirus-Driven tissue specific Gene Transcription. Genome Biol Evol. 7(4):1082–97
57. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–2.
58. R Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. https://www.R-project.org/.

59. Rayan NA, Del Rosario RCH, Prabhakar S. 2016. Massive contribution of transposable elements to mammalian regulatory sequences. Semin Cell Dev Biol. 57:51–6.
60. Roadmap Epigenomics Mapping Consortium. 2015. Integrative analysis of 111 reference human epigenomes. Nature. 518:317–30.
61. 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. 2008. Proc Natl Acad Sci USA. 105: 4220–5.
62. Savage AL, et al. 2014. An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS. Plos One. 9(6):e90833.
63. Saykally JN, Dogan S, Cleary MP, Sanders MM. 2009. The ZEB1 Transcription Factor Is a Novel Repressor of Adiposity in Female Mice. PlosONE. 4(12):e8460.
64. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown GD,
65. Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel
66. genome organization and CTCF binding in multiple mammalian lineages. Cell.
67. 148:335–48.
68. Simonti CN, Pavlicev M, Capra JA. 2017. Transposable Element Exaptation into Regulatory Regions is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. Mol Biol Evol. 34(11):2856–69.
69. Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open 4.0. http://
70. www.repeatmasker.org.
71. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 57:289–300.
72. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res.24:1963–76.
73. Sur D, et al. 2017. Detection of the LINE-1 retrotransposon RNA-binding protein ORF1p in different anatomical regions of the human brain. Mobile DNA. 8:17.
74. The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489:57–74.
75. Tonjes RR, et al. 1996. HERV-K: the biologically most active human endogenous retrovirus family. J Acquir Immune Defic Syndr Hum Retrovirol. 13(1):261–7.
76. Trizzino M, Park S, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch V, Brown CD. 2017. Transposable elements are the primary source of novelty in the primate gene regulation. Genome Res. 27:1623–33.
77. Upton KR, et al. 2017. Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 161:228–39.
78. Walter M, Teissandier A, Pérez-Palacios R, Bourchis D. 2016. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. Elife. 5:e11418.

79. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. J Mol Biol. 354:994–1007.
80. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA. 104:18613–8.
81. Ward M, Zhao S, Luo K, Pavlovic B, Karimi MM, Stephens M, Gilad Y. 2017. Silencing of transposable elements may not be a major driver of regulatory evolution in primate induced pluripotent stem cells. BioRxiv. doi: 10.1101/142455.
82. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York.
83. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue specific enhancer landscape. Nat Genet. 45:836–41.

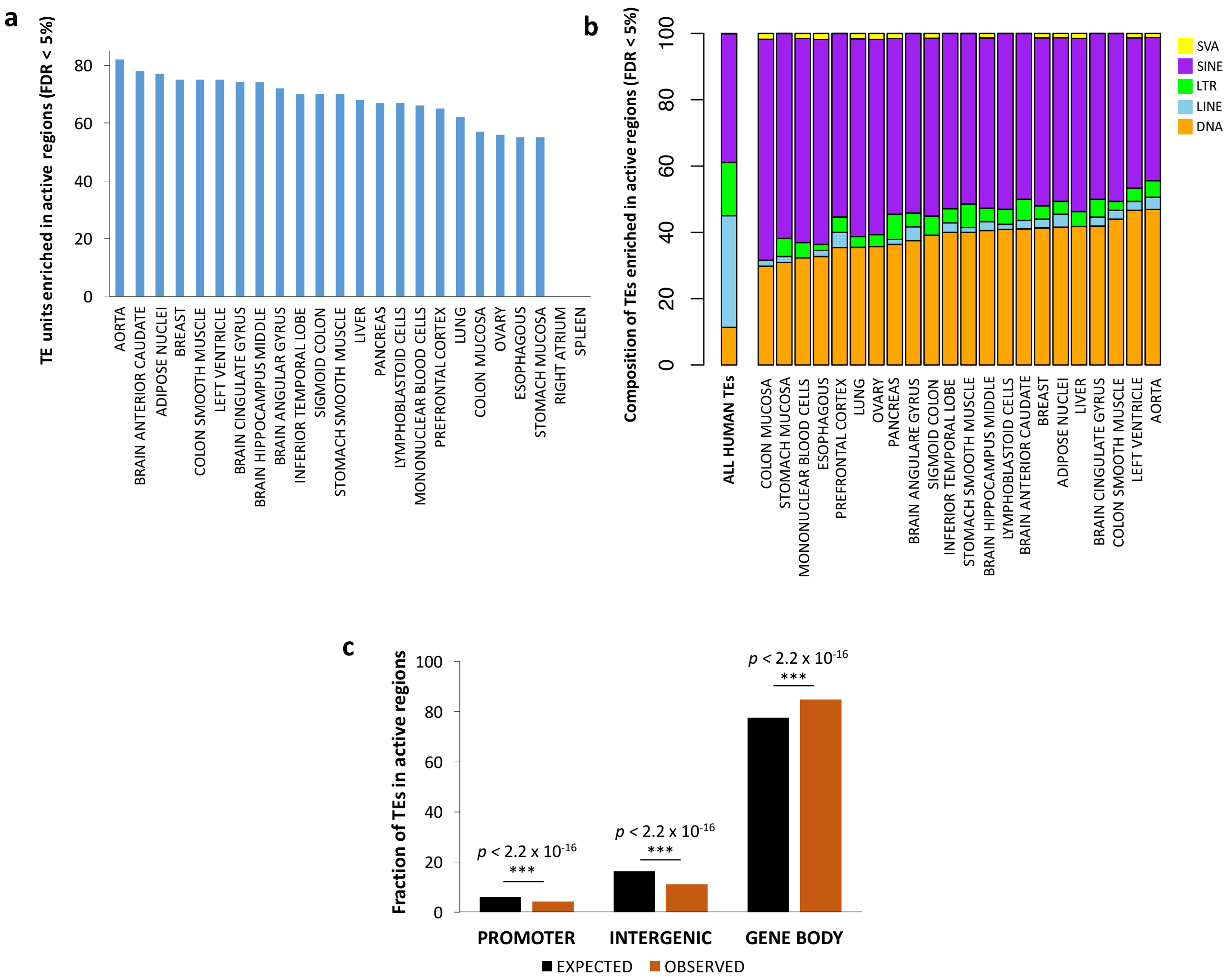

Figure 1 - Transposable elements are enriched in active genomic regions. (A) The plot displays the numbers of enriched TE families in the active genomic regions for each tissue (FDR <5%). (B) Stacked-bar charts show TE class composition for the transposons enriched in active regions (FDR <5%). (C) The TEs found in active regions are depleted from promoters and intergenic regions, while they are enriched in gene bodies.

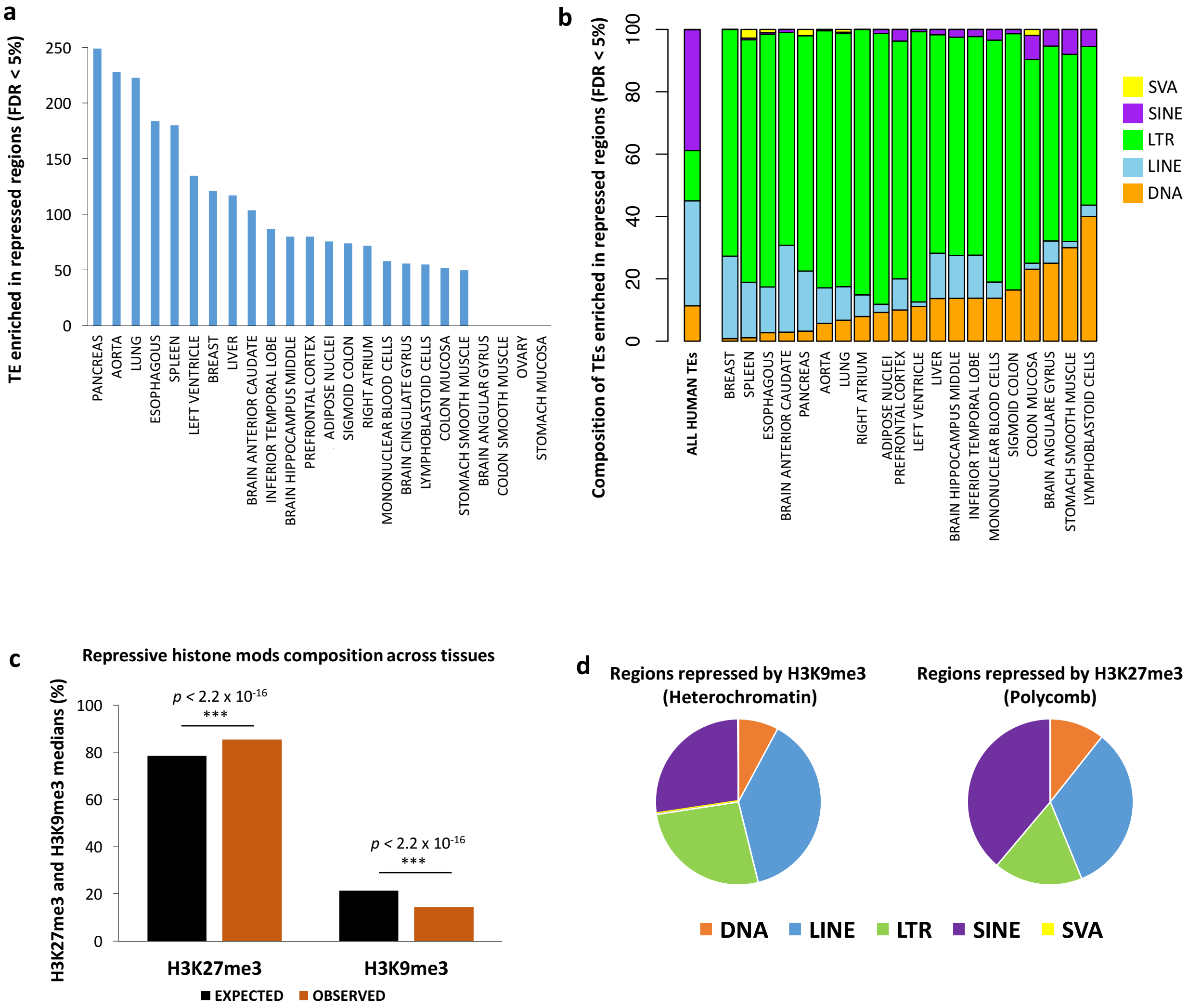

Figure 2 - Transposable elements are enriched in repressed genomic regions. (A) The plot displays the numbers of enriched TE families in the repressed genomic regions for each tissue (FDR <5%). (B) Stacked-chart plot shows TE class composition for the transposons enriched in repressed regions (FDR <5%).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

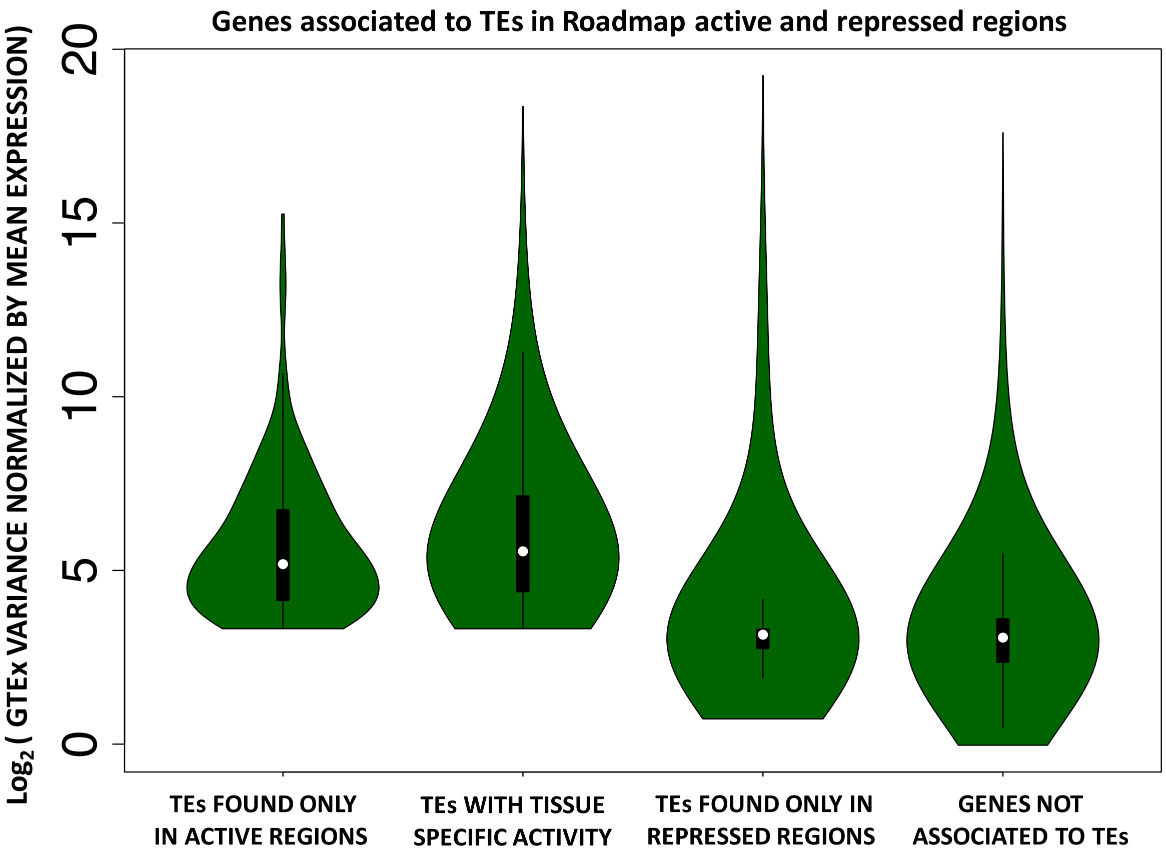

Figure 3 - Genes with higher expression variance are more tolerant towards TE expression. Human genes were split into four categories: 1) Genes associated with TEs that are only found in active regions across tissues; 2) Genes associated with TEs that are found in active or repressed regions in a tissue 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.

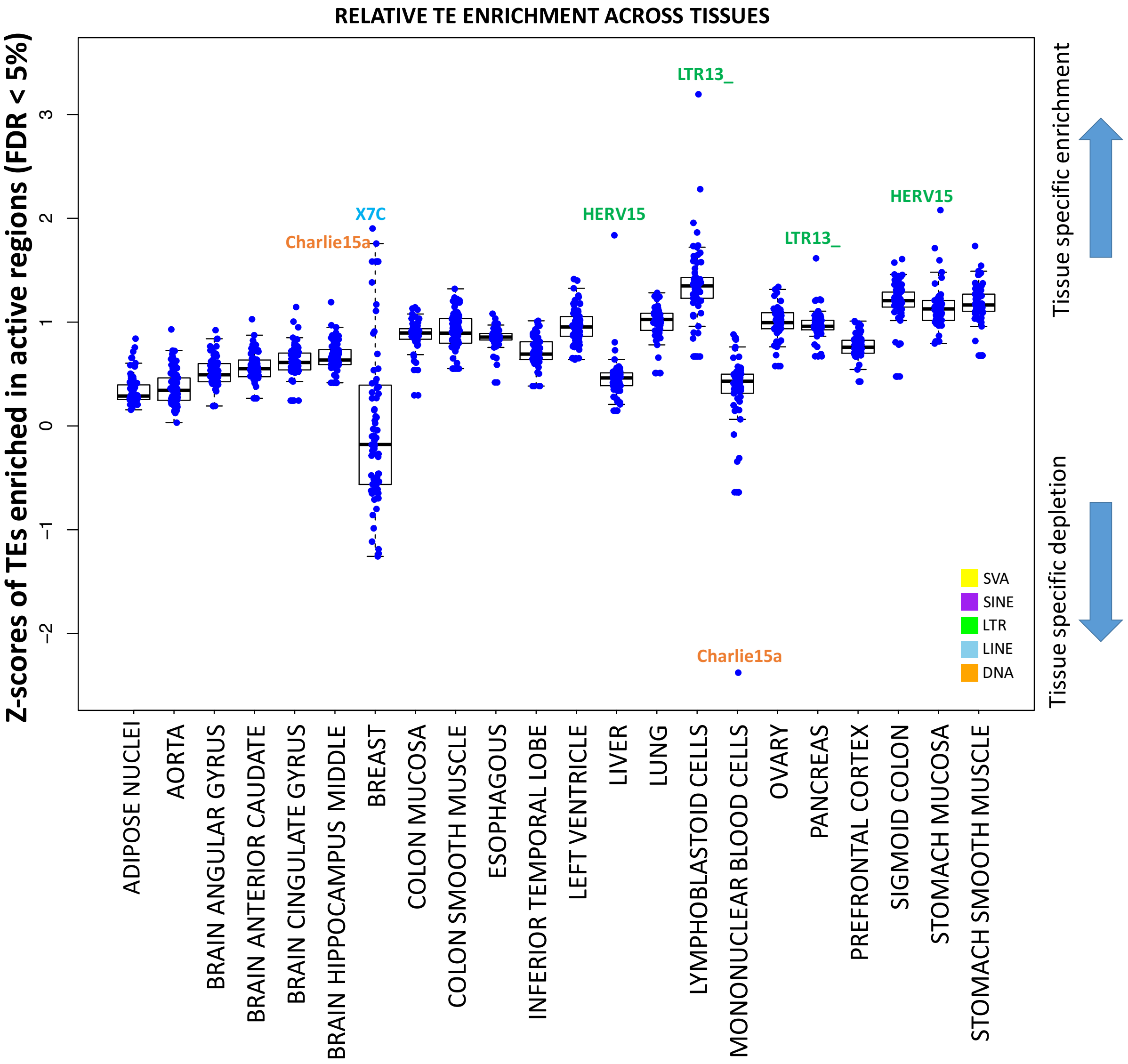

Figure 4 - Transposable elements have tissue specific activity. The plot displays the distribution of the effect sizes (Z-scores from permutation test, see methods) for each TE family enriched in active regions (FDR < 5%), in each tissue. The higher the Z-score, the more tissue specific is the enrichment.

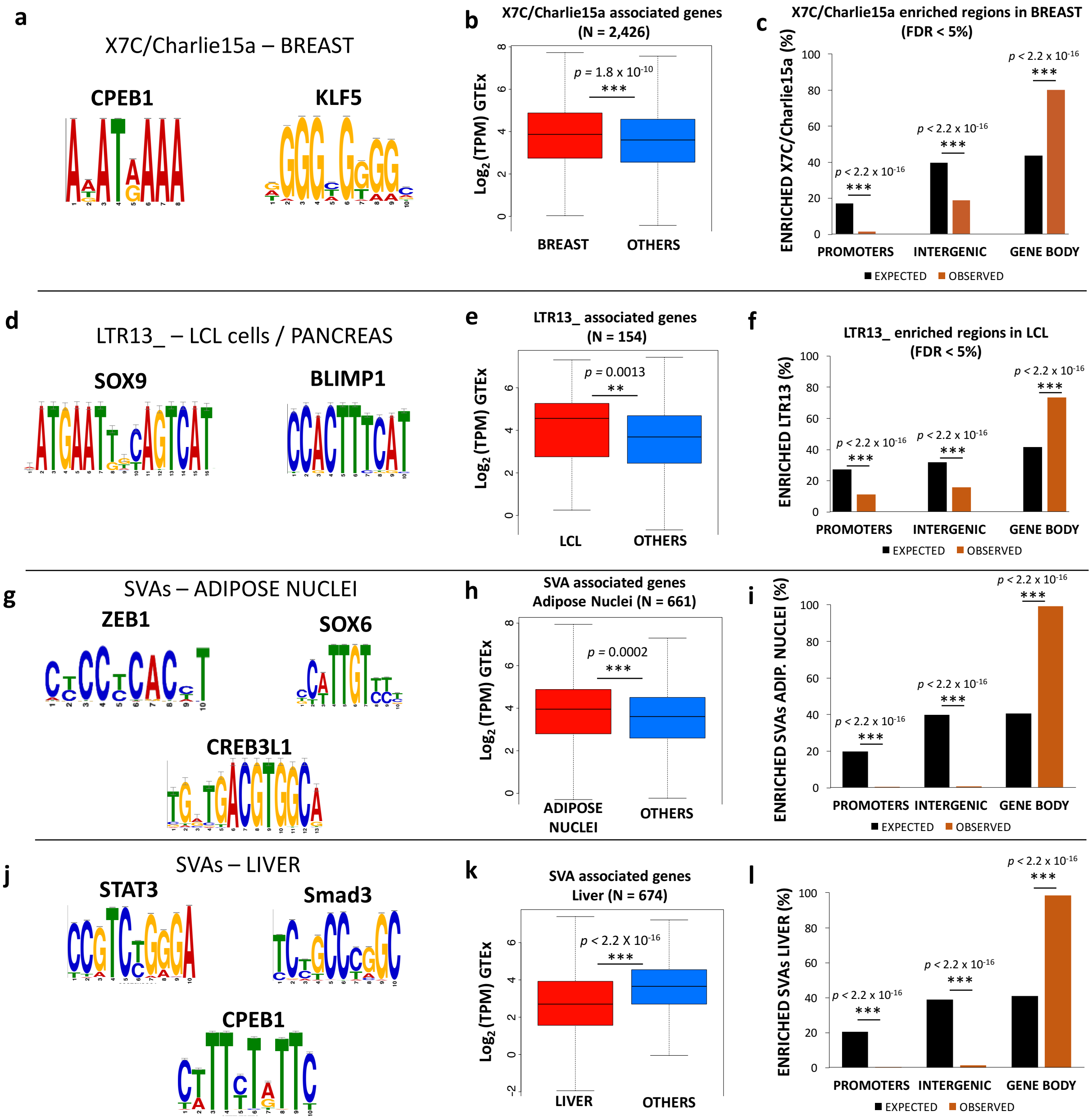

Figure 5 - Tissue specific TEs are enriched for TF binding sites, are mostly intronic, and affect gene expression. (a) Motifs enriched in the regions overlapping X7C and and Charlie15a TEs in the breast. (b) Boxplot comparing mean expression for the genes associated to X7C and and Charlie15a in the breast vs all the other tissues. (c) Genomic composition for 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 composition for 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 composition for 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 composition for the regions overlapping SVAs in the liver.