Transcriptome analysis reveals negative correlation between transcripts of mitochondrial genes and L1HS, and positive correlation between KRAB-ZFPs and older LINE elements in normal somatic tissue

Nicky Chung[1], G.M. Jonaid[1], Sophia Quinton[1], Austin Ross[1], Adrian Alberto[2], Cody Clymer[2], Daphnie Churchill[2], Omar Navarro Leija[2] and Mira V. Han[1,3,]*

[1] School of Life Sciences, University of Nevada, Las Vegas, NV 89154, USA
[2] Department of Computer Science, University of Nevada, Las Vegas, NV 89154, USA
[3] Nevada Institute of Personalized Medicine, Las Vegas, NV 89154, USA

* To whom correspondence should be addressed. Tel: +1-702-774-1503; Fax: 702-895-3956; Email: mira.han@unlv.edu

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors

ABSTRACT

Despite the long-held assumption that transposons are normally only expressed in the germ-line, recent evidence shows that transcripts of LINE sequences are frequently found in the somatic cells. However, the extent of variation in LINE transcript levels across different tissues and different individuals, and the genes and pathways that are co-expressed with LINEs are unknown. Here we report the variation in LINE transcript levels across tissues and between individuals observed in the normal tissues collected for The Cancer Genome Atlas. Mitochondrial genes and ribosomal protein genes were enriched among the genes that showed negative correlation with L1HS in transcript level. We hypothesize that oxidative stress is the factor that leads to both repressed mitochondrial transcription and LINE over-expression. KRAB zinc finger proteins (KZFPs) were enriched among the transcripts positively correlated with older LINE families. The correlation between transcripts of individual LINE loci and individual KZFPs showed highly tissue-specific patterns. There was also a significant enrichment of the corresponding KZFP's binding motif in the sequences of the correlated LINE loci, among KZFP-LINE locus pairs that showed co-expression. These results support the KZFP-LINE interactions previously identified through ChIP-seq, and provide information on the *in vivo* tissue context of the interaction.

INTRODUCTION

LINE transposable elements (TEs) comprise more than 17% of the human genome. Most LINE elements are incapable of retrotransposition, except for the few elements among the youngest families including L1HS. In addition to generating insertional mutations and causing disease through its retrotransposition activities (reviewed in (1)), L1HS also causes damage to the cell through aberrant expression of its sequence (Reviewed in (2)). Aberrant expression of full-length or even partial L1HS sequence that contains the ORF2 is known to cause double-strand DNA breaks (DSBs)

(3, 4). This excess damage in DNA can lead to cell cycle arrest and apoptosis (3, 5), or a senescence-like state (4).

LINE elements have long been thought to be expressed only in the germline cells (6–8), but recently we have learned that both full-length and partial transcripts of LINEs are frequently found in the somatic cells (8–10) with large variation in expression levels across tissue types. A large number of different LINE element sites are expressed in human somatic tissues, and this expression varies among different individuals (11). The level of LINE expression is especially pronounced in cancer cells. Endogenous expression of LINEs have been observed early on in germ cell tumors (GCTs) (12), and GCT-derived embryonal carcinoma cell (ECC) lines (13), and in numerous other types of cancer (14) more recently. There are hypotheses considering LINE activity as a driver of tumorigenesis or LINE expression as one of the hallmarks of cancer (15).

Although there are many reports of LINE expression in the somatic cells, how LINE expression is repressed and de-repressed in human somatic cells is still largely unknown. Based on what we have learned so far, LINE expression is regulated through multiple layers, consisting of transcription factors, DNA methylation, PIWI-interacting RNAs (piRNAs), RNA interference (RNAi), and posttranscriptional host factors. Full-length human LINE elements contain a 5’ Untranslated Region (UTR) that includes an internal RNA polymerase II promoter (16), as well as binding sites for RUNX3 (17), SRY (18) and YY1 (19). LINE expression is also regulated through cell-specific and stage-specific epigenetic mechanisms (2). Expression of LINE full-length transcripts is correlated with differential methylation of the promoter region in various cell lines (45), and in fetal germ cell development in male mice (20). Loss-of-function studies in mice suggest that piRNAs guide de novo methylation to retrotransposons during germline development (21, 22), and an intact pi-RNA pathways is required to prevent L1 mobilization in male meiosis (23). A large number of proteins associated with LINE in humans at the post-transcription stage has been identified through proteomic studies of the LINE ORFs and the ribonucleoprotein (RNP) complex. LINE RNP comprises of its two ORF proteins bound preferentially to their encoding (cis) LINE RNA. Other proteins modulating retrotransposition are also associated with the RNP complex. So far, four studies have utilized immunoprecipitation and proteomics and identified altogether about a hundred proteins (24–27). Recently, CRISPR–Cas9 screen was used to identify proteins that restrict LINE activity (28). The protein MORC2 and the human silencing hub (HUSH) complex was shown to selectively bind evolutionarily young, full-length LINEs located within euchromatic environments, and promote deposition of histone H3 Lys9 trimethylation (H3K9me3) for transcriptional silencing (28).

Among the proteins participating in regulation of TEs, Krüppel associated box-Zinc Finger Proteins (KRAB-ZFPs) are an interesting gene family. The Krüppel associated box (KRAB) domain is a well-known repressor domain and together with the best-known co-factor KAP1 (TRIM28), the KZFP-KAP1 complex has been shown to silence both exogenous retroviruses and EREs during embryonic development (29, 30). Based on this observation, and the pattern of co-evolution of retroviral LTRs and the C2H2-Zinc Finger gene family, it has been hypothesized that the KRAB-ZFPs function in transposable element suppression (31). But except for a few KRAB-ZFPs, most members do not have a characterized function. In an alternative hypothesis, instead of its original role in silencing, it was proposed that KRAB-ZFPs may control the domesticated transposable elements that contribute to the host transcription regulation network (32). Several studies have now reported observations that suggest that older LINE elements are functional members of the host transcriptome, possibly acting as alternative promoters, enhancers or precursors of long non-coding RNAs. Faulkner et al. in 2009, was the first study to provide a global picture of the significant contribution of retrotransposons to human transcriptome in multiple tissue types (33). This report showed that 6-30% of transcripts had transcription start sites located within transposons, and these transposons were tissue-specifically expressed and influenced the transcription of nearby genes. The results were extended by Djebali et al. in 2012 showing again the tissue-specificity of transposon expression, and that most of these transcripts are found in the nuclear part of the cell (34).

Recently, high-throughput RNA-seq data of various types of cancer samples and their normal counterparts have become available in The Cancer Genome Atlas (TCGA) (35–37). By focusing on the normal tissue samples from TCGA, we can access thousands of natural experiments across various types of tissues that show normal variation in LINE transcript levels, and obtain a global picture of LINE expression and regulation in humans. Since the samples are collected from fresh-frozen tissues, LINE transcript levels are observed *in vivo*, complementing the studies that focus on retrotransposition assays or transposon expression in human cell lines. An important strength of this approach is the large number of samples collected for each tissue type. The variation in LINE element transcripts observed in multiple samples within each tissue, allowed us to analyze the co-expression patterns between genes and LINEs. We hypothesized that genes that regulate the transcription level of LINEs would show correlation in expression levels with the LINE transcripts.

We first summarize the survey of LINE expression variation found in the RNA-seq data from 660 samples of normal tissue. We confirm the earlier findings that LINE expression varies across tissue types, with the tissues esophagus, stomach and head and neck showing the highest level of L1HS transcripts. Transcript levels of individual LINE loci are highly tissue specific and within each family only a few individual loci are highly expressed, contributing to the bulk of the transposon transcripts at the family level. We also find large variation in LINE expression across individual samples within each tissue type. The expression of the L1HS family varies by as low as 2-fold in tissues like kidney, to greater than 8-fold in breast, thyroid, and many other tissues.

By analyzing the correlation between individual genes and L1HS transcripts, controlling for the older LINE transcript levels, we found a significant excess of mitochondrial genes and ribosomal protein genes negatively correlated with L1HS in transcript level. Several of the genes that recurrently show strong negative correlation across tissues, have functions related to oxidative stress. We also found an enrichment of KRAB-ZFPs that were positively correlated with older LINE expression, recurrently in multiple tissue types. Binding motifs of the KRAB-ZFP proteins are enriched in the sequences of LINE loci for co-expressed KZFP-LINE pairs, when compared to random KZFP-LINE loci combinations that are not co-expressed.

MATERIAL AND METHODS

RNA-Seq and gene expression quantification in the normal tissues.

We used the gene level quantification provided by The Cancer Genome Atlas (TCGA) for the gene expressions. We collected gene level quantifications for 660 samples from The Cancer Genome Atlas (TCGA). We focused on cancer types that had at least 6 normal samples of RNA-seq data, collected from normal tissue adjacent to the cancer tissue. As a result, 17 different tissue types were included in our analysis: BLCA (Bladder urothelial carcinoma) , BRCA (Breast carcinoma), CHOL (Cholangiocarcinoma), COAD (Colon adenocarcinoma), ESCA (Esophageal adenocarcinoma), HNSC (Head and neck squamous cell carcinoma), KICH (kidney chromophobe), KIRC (kidney renal clear cell carcinoma), KIRP (Kidney renal papillary cell carcinoma), LIHC (Liver hepatocellular carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), PRAD (Prostate adenocarcinoma), READ (Rectum adenocarcinoma), STAD (Stomach adenocarcinoma), THCA (Thyroid carcinoma) and UCEC (Uterine Corpus Endometrial Carcinoma). Number of samples for each tissue is described in Table 1. Although we will use the acronym for the cancer type to describe these tissues, we emphasize again that all our samples come from the normal tissues collected from the same organ of the same patient with the tumour. The tumour tissue samples were not included in our analysis.

Methods for sequencing and data processing of RNA using the RNA-seq protocol have been previously described for The Cancer Genome Atlas (TCGA) (35–37). Briefly, RNA was extracted, prepared into poly(A) enriched Illumina TruSeq mRNA libraries, sequenced by Illumina HiSeq2000 (resulting in paired 48-nt reads), and subjected to quality control. RNA reads were aligned to the hg19 genome assembly using Mapsplice (38). Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1 using RSEM (39). We used the raw_count values in the .rsem.genes.results files, rounded to an integer, as the gene level quantification.

Quantifying TE derived transcripts at the locus and family level

We collected RNA-seq level 1 binary alignment files (.bam files) for 660 samples (Table 1) from The Cancer Genome Atlas. These bam files are results of the RNA-seq and MapSplice protocol described in the previous section. We used a modified version of the software TEtranscripts (40) for quantifying the reads mapping to annotated transposons. TEtranscripts is a software that can quantify both gene and TE transcript levels from RNAseq experiments. It takes into account the ambiguously mapped TE-associated reads by proportionally assigning read counts to the corresponding TE families using an Expectation-Maximization algorithm. We implemented two modification to the original TEtranscripts software. 1) We modified it to report read counts for each individual TE locus in the reference genome in addition to the family level counts. 2) We developed a function to discount the read counts by removing read counts that correspond to transcripts containing TE sequences that originate from pre-mrna or retained introns in the mature RNA (41). If the reads mapping to TEs are part of the pre-mrna or retained introns, we should see continuous mapping of reads that span the introns flanking the TE of interest. We utilized the read depths in the flanking introns to proportionally reduce the number of total reads mapped to the TE using the following formula:

$$R_{IL} = \frac{count_{IL}}{len_{IL} - read\_len}$$

$$count_{TE}' = count_{TE} \ - count_{TE} \times \frac{R_{IL} + R_{IR}}{2R_{L1}}$$

IL : intron left to TE. IR : intron right to TE.

R_{IL} : read depth of the intron left to the TE

$count_{IL}$: read counts mapped to the intron left to the TE (includes multi-mapped reads)

len_{IL} : length of the left intron

$read\_len$: length of the sequencing read

$count_{TE}'$: count of reads mapped to TE after the correction.

In addition to the discounted quantification, we also used the option in the TEtranscripts software to quantify the TE transcripts using only the uniquely mapped reads. Downstream analyses were done using both the discounted quantification based on multi-mapped reads and the uniquely mapped quantification, to assess the impact of uncertainty in multi-mapped reads.

The retrotransposon annotations used were generated from the RepeatMasker tables, obtained from the UCSC genome database and provided by TEtranscripts. For quantifying reads mapping to the TE flanking introns we generated gtf files containing 1) the TE flanking intron positions, 2) the intergenic TE positions, 3) the exonic TE positions (TEs that fall within an exon, including non-coding RNA genes). In case of intronic TEs, we use the algorithm described above to discount the transcripts from pre-mrna or retained introns. In case of intergenic TEs, we count all EM estimated reads mapped to TEs without any discount. In case of exonic TEs, we ignore those counts altogether, and the exonic TEs do not contribute to the locus count nor the family level count. The modified version of the TEtranscripts software and the required gtf files can be found at https://github.com/HanLabUNLV.

Normalization and transformation of read counts

After quantifying the reads mapping to annotated genes and TEs, both the gene level counts, and the TE counts were normalized between samples across all tissue types with DEseq2. We used the default "median ratio method" for normalization in DESeq2 (42). Briefly, the scaling factor for each sample is calculated as median of the ratio, for each gene, of its read count to its geometric mean across all samples. The assumption of the median ratio method is that most genes are not consistently differentially expressed between tissues. If there is systematic difference in ratio between samples, the median ratio will capture the size relationship. But, this assumption may be violated when we are comparing large number of tissues types at the same time, since a large proportion of the genes may be differentially expressed in at least one tissue type, or one of the tissues may be extremely biased in their number of differentially expressed genes. In order to achieve more robust normalization, we used a two-step normalization method called the differentially expressed genes elimination strategy (DEGES) (43). We performed preliminary normalization using the “median ratio method”, filtered out potential differentially expressed genes in the data, found a subset of robust non-differentially expressed genes, and used the subset to perform the second round of “median ratio normalization". The resulting pairwise MA plot between tissues after normalization showed better normalization compared to the regular one-step normalization (Supplementary Figure 1). The size factors for each sample obtained from the two-step normalization on gene counts were then used to normalize the TE quantifications of the same sample. The normalized counts were log2 transformed using the variance stabilizing transformation function in DESeq2 (42, 44) for downstream analysis.

Clustering of samples by expression pattern.

We visualized the expression patterns in the data using the default complete linkage clustering in *hclust*, with the *pheatmap* function in R (45). The top 1000 genes or TEs, with the largest variances were used for the clustering and visualization. For the locus level TE expression, we filtered out all loci that had less than 5 read counts for every sample. To compare the clustering of samples based on gene expression vs. TE expression, we used the normalized Mutual Information measure (46). The hierarchical clusters were cut off at k = 17, the number of different tissue types. Because the resulting clusters were not accurate enough to distinguish between similar tissues, we used a broader tissue grouping to compare with the clusters. The tissues were grouped to 10 broader types based on preliminary clustering: bladder/endometrium (BLCA, UCEC), breast (BRCA), liver/bile duct (CHOL, LIHC), colon/rectum (COAD, READ), esophagus/stomach (ESCA, STAD), head and neck — the squamous epithelium in the mucosal surfaces inside the mouth, nose, and throat (HNSC), kidney (KICH, KIRC, KIRP), lung (LUAD, LUSC), prostate (PRAD), and thyroid (THCA). The broader tissue type of each sample was used as the ground truth. Each resulting cluster was then assigned a group label based on the majority tissue type. Normalized mutual information was calculated by comparing the labels from the clustering to the true class labels.

Correlated expression between genes and LINEs

Before examining genes that are associated with L1HS expression, we first examined the clinical variables to check any confounding variables associated with the L1HS levels in the normal tissue. We tested the variables age, days to death, pathological stage, T staging, N staging, M staging, gender, radiation and race for each tissue type. Radiation therapy was the only clinical variable associated with L1HS expression in the normal tissue of thyroid, and thus was the only clinical variable included as a covariate in our linear models and only for the thyroid tissue.

Correlation between gene and L1HS transcripts were tested in each tissue groups separately, in bladder, breast, liver/bile duct, colon/rectum, stomach/esophagus, head and neck, kidney, lung, prostate and thyroid. We tested 20532 genes for each tissue group using a linear model with log2 L1HS expression as the dependent variable, and log2 gene expression as the independent variable. For a gene to be included in our test, it had to be present in at least eight individual patients. We also required that the gene be expressed with a minimum RPM of 2 in 75% of the samples to be included in the dataset. In addition to the radiation therapy for thyroid tissue, we considered effective library size (sum of all normalized counts) and the batch ID provided by the TCGA project as additional covariates. The linear model we used is described below.

$$\log_2 L1HS \sim \log_2 gene + \log_2 effective\_library\_size + batch + radiation$$

We tested all combination of linear models that can be created by including or excluding these variables. Second-order Akaike Information Criterion (AIC_c) was used to select the best linear model. The coefficient and p-value used are the ones from the best model.

After the preliminary analysis, we found there was significant correlation between the level of L1HS and the level of the older LINE elements. To tease apart the effects, we separated the total LINE quantification into L1HS and older LINEs. *L1HS* was the sum of read counts mapping to all the individual L1HS loci. An additional variable that we called "*oldLINEs*" was quantified as the sum of read counts mapping to all the LINE elements except for L1HS, L1PA2 and L1PA3. We estimated the linear models above for *L1HS*, this time including the covariate *oldLINEs*. To identify genes correlated with older LINE elements, we also estimated the linear models with *oldLINEs* as the dependent variable, using *L1HS* as the covariate.

$\log_2 L1HS \sim \log_2 gene + \log_2 effective\_library\_size + batch + radiation + \log_2 oldLINEs$

$\log_2 oldLINEs \sim \log_2 gene + \log_2 effective\_library\_size + batch + radiation + \log_2 L1HS$

Genes showing recurrent correlation with L1HS and older LINEs in multiple tissues

For the gene-L1HS correlation analysis, we decided to focus on global correlation signals that appear in multiple tissue types. In order to identify these global recurrent correlations, we used two different methods. First, we collected genes that showed significant gene coefficients from the linear model (q-value < 10e-4) in more than one tissue type. But, because this method relies on the p-value of the coefficient, it is biased towards picking up genes in tissues with large sample sizes. So, we also utilized a second measure, called REC score (47), that utilizes the ranks of correlations instead of the absolute p-values to find recurrent correlation across tissues. This rank-based approach ensures that individual tissue types are weighted equally, and limits bias from tissues with large sample sizes or from strong associations measured in only a single tissue type. We modified the REC score slightly to test all gene relationship against a single dependent variable (*L1HS* or *oldLINEs)*.

Let L_k be the ordered vector of negative and positive associations involving mRNA j in cancer type k. The strength of the association between mRNA j and L1HS in cancer type k is defined by the relative rank:

$$rr_{j,k} = \frac{r_{j,k}}{|L_k|} - \frac{1}{2|L_k|},$$

where, $rr_{j,k}$ is the rank of the association between mRNA j in the vector L_k. The rest of the formulas for calculating REC_j, the REC score for mRNA j follow the methods described in (47).

Once genes with recurrent correlations were identified, we used the gene enrichment analysis software DAVID (48) with the positively and negatively correlated gene sets to identify functional clusters that are enriched among the correlated genes. We also derived a ranked list of these genes based on their sign of the gene coefficient and the maximum partial eta-squared values for the gene coefficient from the linear model. We used this pre-ranked list of genes with recurrent correlations to run the Gene Set Enrichment Analysis (GSEA) software (49) against the curated pathways gene set and the GO terms gene set of the MSigDB collections (50).

Correlation between individual LINE loci and KRAB-ZFPs

To understand the positive correlation between LINEs and KRAB-ZFPs, we looked at the correlation between each KZFPs and individual LINE loci in different tissue types. We tested the correlation for 366 KRAB Zinc Finger Proteins that were identified in Imbeault et al. (51) and also found in our gene expression data. Because the search space of pairwise combinations of KZFP and individual LINE loci was too large, we examined the relationship in a step-wise approach. In the first step, we tested the correlation between all pairwise combinations of 366 KZFPs and 146 LINE subfamilies using the TE quantification at the family level in each tissue type. Then, in the second step, once the significantly correlated KZFP and LINE family was identified, we focused on those pairs. We tested the correlation between the expression of the significant KZFP and the expression of each individual locus of the significant LINE family in the tissue where the initial co-expression was found to identify individual LINE loci that are co-expressed with the KZFP.

KRAB-ZFP motif search in the sequences of co-expressed LINE loci

To test the presence of zinc finger motifs in sequences of co-expressed LINE loci, we searched for known KZFP motifs using the software Find Individual Motif Occurrences (FIMO) (52). We used the motifs of the 39 KZFPs identified by Barazandeh, et al. (53) through a comparison of ChIP-Seq results. Taking only the zinc finger genes from our dataset that had a zinc finger motif available from Barazandeh, et al. (53), we had a total of 10,782 pairwise combinations of zinc finger motifs and individual L1 locus that showed significant co-expression. We searched the corresponding motifs in the respective L1 sequence using the default p-value cutoff from the FIMO program, p-value < 0.0001. Afterwards, we summed up the total number of motifs identified in the L1 sequences as well as the total number of unique zinc finger motifs identified in each of the L1 sequences, as some motifs are identified multiple times in an L1 sequence.

To test the significance of our results, we generated two different null models for comparison. For the first null model, we randomly sampled 10,782 pairs of KZFP motifs and L1 loci out of all available KZFPs and L1 loci in our gene expression dataset, regardless of whether they are co-expressed or not. For the second null model, we aimed to construct a null model that reflects the distribution of the L1 families in our significantly co-expressed pairs. We thus started with the 10,782 pairs of zinc finger motifs and L1 sequences that were identified as significantly co-expressed, but we created new pairwise relationships by randomly permuting the LINEs against the KZFPs so that the co-expression relationships are disrupted. We generated 1000 sets of each null model, and tested for the existence of the KZFP motif in the paired L1 sequence. We then summed up the total number of zinc finger motifs identified in the L1 sequences as well as the total number of unique motifs in each of the L1 sequences. We fit a normal distribution to the total number of motifs identified for the 1000 set of null models, and calculated the p-value of our observed count of motifs.

Validation data set from the GTEx project

In addition to the normal samples from TCGA, we also obtained normal lung tissue RNA-seq files from the Genotype-Tissue Expression (GTEx) project (54) as an independent set of samples for validation. GTEx donors are identified through low-post-mortem-interval (PMI) autopsy or organ and tissue transplant settings, in which biospecimens collection can start within 24 hours of death or surgery. It is expected that the majority of the tissue samples in this dataset is free of major disease. The RNA-seq protocol for the GTEx project has been previously described in (54). Briefly, RNA was extracted to a non-strand specific poly(A) selected Illumina TruSeq library, sequenced by HiSeq 2000 into 76bp paired-end reads, with median coverage of 82M total reads. The reads were mapped to the GENCODE 19 annotation using STAR (55). We downloaded 210 aligned bam files from dbGaP for the lung tissue, chosen because of its large sample size compared to other tissues. The analyses described above were repeated identically for these 210 samples.

RESULTS

TE transcripts are quantified with correction to discount TE reads originating from pre-mRNAs or retained introns

In general, we find that there is good correlation between the total raw counts, and the reduced counts of the TE family after our correction (Figure 1). We also visually confirmed several examples of clear L1HS expression that are not part of the surrounding intron (supplementary Figure 2). But there were a few samples with large differences before and after correction, such that the overall distribution of the corrected difference was positively skewed (Figure 1, Supplementary Figure 3). On average, the total read count for L1HS was reduced by around 426 reads for each sample. There were 10 patients with large reduction in L1HS read counts, greater than twice the standard deviation above the mean, and these outliers with large corrections came from two tissue types, the stomach and the esophagus (Table 2). When looked in detail these extremely large corrections came from very few specific L1HS loci. The most extreme case was due to a single L1HS locus, L1HS_dup39 on chromosome 1, 91853148-91853486 (Supplementary Figure 4a). It lies next to an rDNA that is highly expressed in stomach and esophagus. There was transcription read-through of the rDNA that extended beyond the rDNA boundary and continued to transcribe part of the L1HS sequence, beyond the 5-base threshold that we required for being counted as L1HS transcription. Because of this, L1HS_dup39 had a read count of 11473 reads in this sample (patient ID 6852). But after applying our correction method based on the flanking read depths, all the reads mapping to this particular L1HS locus was removed, and the quantification for the particular locus went down to zero as it should. Another example is L1HS_dup898, which was the single locus most frequently identified as the largest correction within each sample. Among the 674 samples that we quantified, L1HS_dup898 showed the largest correction in 310, almost half of the samples (Table 3a). L1HS_dup898 is found on chr 9, 85664455-85670486 in the middle of the first intron of the gene *RASEF*. When we looked into the read alignment in detail in one of the samples with the highest expression (patient ID A4OM), we found that there were large numbers of reads uniquely mapping to L1HS_dup898, but we also found equally large numbers of reads mapping to the surrounding intron on both the left and right side of the TE, indicating intron retention (Supplementary Figure 4b). The estimated read counts for L1HS_dup898 were reduced to zero after the correction.

Although, we focused on L1HS elements as example cases, L1HS is actually not the TE family that showed the largest corrections compared to other TE families. AluSx, AluSx1, AluJb, AluY, MIR3, MIRb, etc. showed the largest corrections, several orders of magnitude larger reduction in read counts compared to L1HS (Table 3b). As we have seen in the case of L1HS, a few specific loci were responsible for a major proportion of the correction for a TE family, depending on the tissue type. This was due to specific transcription read-throughs or intron retentions that happen more frequently in certain tissue types.

We also found cases where the method corrected for erroneous TE quantifications due to TEs embedded within long non-coding RNAs (lncRNAs). One TE locus that was most frequently identified with the largest correction in the sample was AluSx1_dup59209. AluSx1_dup59209 had very high transcript levels with an average read count of 18863 in thyroid and head and neck tissue. But when we looked into the alignment in detail, we found that the Alu element was embedded in a lncRNA gene called *TTTY14*. The reads mapping here were counted as AluSx1 transcripts in the original quantification. In the alignment, we see that AluSx1_dup59209 has accumulated enough mutations, such that almost all the reads mapped to it are uniquely mapped reads. It looks to be a novel case of an *Alu* domestication, where an *Alu* insertion or a secondary duplication of an original *Alu* insertion became part of a testis specific RNA gene (56). Since we added a step to remove all read counts mapping to TEs if the TE falls within exons or non-coding RNAs, the read counts for AluSx1_dup59209 were reduced to zero in all samples, removing the errors from the original quantification (Supplementary Figure 4c).

LINE expression shows tissue-specific expression patterns among the normal somatic tissues

There have been multiple reports of tissue specific expression of TEs in the human genome, starting from Faulkner et al. in 2009 (33) to Philippe et al. 2017 (57) more recently. We were still surprised at the extent of tissue specific information contained in the expression pattern of individual LINE loci across the genome. During our preliminary analysis examining the data quality and exploring the data, we unexpectedly found that we could cluster each sample into their broader tissue groupings, based on LINE expression patterns alone without relying on any genes at all. We checked the quality of the data by clustering the samples based on the 1000 genes with the largest variance in expression level, and confirmed that the samples were correctly grouped into the tissues based on gene expressions (Supplementary Figure 5a). Then we explored the expression of each LINE family and found that the family level quantification of LINE expression did not contain enough information to correctly cluster the samples into their corresponding tissue (Supplementary Figure 5b). But when we used the expression levels of 1000 individual TE loci with the largest variance, the samples were again correctly classified into their corresponding tissue groups, even without any information from the genes (Figure 2). We used normalized mutual information between the different clustering results and the ground truth (the true tissue group) to evaluate the quality of clustering. Normalized mutual information was compared for clustering results based on gene expression, family level LINE expression, locus level LINE expression and random assignments. We found that the locus level LINE expression was as predictive of tissue groups as the genes (Table 4).

Compared to other older LINE elements, the tissue specificity was less pronounced for L1HS expression. Still we found large variation in the amount of L1HS expression in different somatic tissues across normal tissue samples (Figure 3), consistent with the previous observation in adult human tissues (10) and in human cell lines (57). There were at least a moderate level of L1HS expression across all the tissue types we examined. Tissue with the lowest L1HS level were the liver and the bile ducts, with the median value at about 300 reads per million. Head and neck tissue, esophageal tissue and stomach tissue showed the highest level of L1HS expression in the normal tissue samples, with a median of around 1500 reads per million. This was even after large amount of corrections in stomach and esophagus samples seen in Figure 1. This is consistent with the observation by Belancio et al. showing high levels of full length LINE expressed in the adult esophagus and stomach tissue, at about 80% and 150% relative to the levels in HeLa cells (10).

When we looked at expression of each individual L1HS loci, we found that certain L1HS loci and certain individual samples were over-represented in the list of the top 100 L1HS loci with the largest read counts (Table 5a, Supplementary Table 1). For example, the locus with the highest expression level across all L1HS loci, L1HS_dup241 shows up 23 times in the top 100 list in 23 different individual samples. L1HS_dup241 is a full length L1HS element on chromosome 3 (chr3:26439509-26445536), and it is most frequently the most highly expressed L1HS locus in the genome. It is possible that the few L1HS loci that we identified to be highly expressed are also the handful of source L1s (hot L1's) that are responsible for the majority of retrotransposition events (58–60). We tested the overlap between the 23 L1HS loci in our top 100 highly expressed list, and the 28 L1HS elements in Table S3 of Tubio et al. (60) We found 6 elements overlapping between the two lists. Considering there are a total of 863 annotated L1HS loci in the human genome (hg19), 6 overlaps are more than what we would expect if the lists were independent. Similar to the over-represented loci, certain tissue samples were also over-represented in the top 100 list. For example, a head and neck tissue sample from patient ID 8663 shows up 9 times in the top 100 list with 9 different L1HS loci (Table 5b, Supplementary Table 1). This suggests that multiple L1HS loci are upregulated by a common regulatory mechanism across different chromosomes in certain individuals. There were six stomach tissue samples, five head and neck, and five esophagus tissue represented in the top 100 list, again showing that head and neck, esophagus and stomach showed higher L1HS expression not only at the family level but also at the individual locus level (Table 5b).

Specific sub-families of SINEs, ERVs and MERs show recurrent positive correlation with LINE expression in multiple normal tissues

Since TEtranscripts quantifies the reads mapped to different transposon families (40), we were able to study the co-expression of different transposons with L1HS. Based on the estimates of TEtranscripts, we found multiple LINE subfamilies closely related to L1HS showing up with the strongest correlation in expression levels with L1HS (Supplementary Table 2a). In fact, the correlations between L1HS and many repeat families were much stronger than the relationship between L1HS and any of the genes. The correlation between closely related LINE subfamilies is expected because reads from transposons that map to sequences that are indistinguishable between subfamilies are assigned to multiple subfamilies with proportional weight by TEtranscripts using an Expectation-Maximization algorithm. What was not expected, was that we also found several DNA transposons and Endogenous retroviruses (ERVs) that are highly and recurrently correlated with L1HS expression in multiple tissues (Supplementary Table 2b). Since there was wide-spread correlation between different LINE subfamilies and multiple transposons outside LINEs, we decided to tease apart the correlation using a multivariate model including separate quantification for the L1HS element and the older LINEs. When we controlled for the RNA level of older LINE elements as a covariate, most of the correlation seen between non-LINE transposons and L1HS disappeared, and the elements correlated with L1HS was limited to the most closely related L1 elements (L1PAs), but no other transposon families (Supplementary Table 2c). On the other hand, when we looked at elements that are correlated to older LINEs, controlling for correlation with L1HS, we found several *Alu* elements recurrently correlated with older LINEs, (Supplementary Table 2d). Considering there is no sequence similarity between the SINEs, MERs, ERVs and the LINEs, the correlation among these diverse transposons is probably due to a common regulatory mechanism that is de-repressing these transposons and LINEs at the same time. There have been reports of such co-expression of ERVs and LINEs in cancerous tissues (61, 62), possibly through concordant hypomethylation (63).

Expression of mitochondrial proteins are negatively correlated with L1HS expression

Based on our results showing widespread correlation between L1HS and older LINEs, we decided to examine genes co-expressed with L1HS and older LINEs separately, controlling for the other variable as a covariate in the linear model. We found 3650 genes that showed negative correlation in expression with L1HS (q-value < 10e-4) in at least one tissue (Supplementary Table 3). There were 1056 genes that were negatively correlated with L1HS in more than one tissue, *i.e.* the correlation was replicated in at least two tissues. Figure 4 shows the top 12 genes with the largest partial eta-squared estimated across all tissues.

We did a gene set enrichment analysis (GSEA) using DAVID for the 1056 genes. The top three most enriched clusters were mitochondrial transit peptide / mitochondrial inner membrane, oxidative phosphorylation and ribosome and ribonucleoprotein (including mitochondrial ribosome proteins) (Table 6, Supplementary Table 4). This result was robust to the threshold we used for significance. When we used a more stringent cutoff of q-value < 10e-5, we found 326 genes negatively correlated with L1HS in more than one tissue, but the top three most enriched annotation clusters were still the same for the 326 genes. The result was also robust to the software we used for the gene set enrichment analysis (DAVID or GSEA)(48, 49), and to the annotation database we used for the analysis (curated pathways or GO terms). With GSEA on curated pathways, the pathways KEGG: OXIDATIVE_PHOSPHORYLATION, REACTOME: RESPIRATORY_ELECTRON_TRANSPORT were on the top of the list. With GSEA on GO terms, the terms MITOCHONDRIAL_PART, MITOCHONDRION, MITOCHONDRIAL_ENVELOPE were on the top of the list (Supplementary Table 5, Supplementary Figure 6). When we look at the list of genes significant in these categories, we find that there is almost an across-the-board reduction in transcripts for mitochondria and the ribosome (Figure 5) in the cells that have high level of L1HS transcript levels. Nuclear encoded mitochondrial genes are known to be co-regulated with mtDNA encoded genes, and ribosomal protein genes are co-regulated with rDNAs. We did not include mtDNA encoded genes, nor rDNAs in our quantification, so we could not test whether they are also negatively correlated with L1HS expression. The enrichment of mitochondrial genes among the genes negatively correlated with L1HS was significant in models both including and excluding expression of older LINE elements as a covariate. In contrast, there was no enrichment of mitochondrial genes when we looked at the genes correlated with the expression of older LINE elements adjusting for L1HS level as a covariate (Figure 6). This shows that the negative correlation between the expression of mitochondrial genes and LINE expression is either a common relationship among all LINE families or stronger for L1HS compared to older LINE elements.

To understand how the uncertainty in reads mapping to TEs affect our results, we decided to examine the correlation only using the uniquely and unambiguously mapped reads. We used the option in the TEtranscripts software to quantify TEs only based on the uniquely mapped reads, and replicated the gene-L1HS co-expression analysis. Overall, the significance of the correlations dropped, due to the lower power coming from lower number of reads. We found 6 genes negatively correlated with L1HS with q-value < 10-e4 and 96 genes using a less stringent cutoff (q-value < 10-e3). The six genes identified are *CA4, CCDC151, PARD6A, RASSF4, SLC29A2, UBTD1*, and there are no common functions across these genes, and no obvious link to mitochondria was found. But with the 96 genes, we found that mitochondrial inner membrane and mitochondrial transit peptide were still the most enriched category among the genes negatively correlated with L1HS measured by uniquely mapped reads (Supplementary Table 6).

We also looked at the recurrent correlations using REC score (Jacobsen et al. NSMB 2013). A strong negative REC score reflects that the mRNA-L1HS relationship generally show correlation in the same direction across the studied tissue types. Compared to the list of correlated genes we examined above, that were identified by a significant coefficient in the linear model in at least two or more tissue groups, this list is based on the rank of correlations and doesn't rely on the significance of the *p-value* in the linear model. Table 7 shows the top 20 genes that have the best REC scores. None of the genes had a REC score that is less that 6.2, which is considered as the threshold for significance, showing that no gene was consistently co-expressed with L1HS across all tissues. But, when we used the top 100 genes with the most negative REC scores and examined the enriched functions, we again found that the most enriched cluster was mitochondrial respiratory chain and oxidative phosphorylation.

When examining the top 20 genes that show recurrent negative correlation with L1HS across tissues (Table 7), we found interesting interaction and overlap in function among the genes. *DHPS* encodes the Deoxyhypusine Synthase, the enzyme that cleaves and transfers spermidine that is necessary for the hypusination of eIF5A. Depletion of eIF5A leads to endoplasmic reticulum stress and unfolded protein response (64). The function of *AIP* is unknown but it is reported to interact with the aryl hydrocarbon receptor (AhR). A tight coupling of AhR and Nrf2-dependent regulation in the prevention of quinone-induced oxidative stress and ER stress has been reviewed in (65). *CREB3* is a transcription factor that activates unfolded protein response (UPR) target genes during endoplasmic reticulum (ER) stress response (66). *PARK7* and *ECSIT* are mitochondrial genes. *PARK7* has a role in protection against oxidative damage, and *PARK7* deficient mice under oxidative stress show impaired mitochondrial biogenesis and respiratory chain deficiency (Billia et al. PNAS 2013). *ECSIT* is a mitochondrial complex I associated gene that has been shown to regulate the production of mitochondrial reactive oxygen species (mROS) following engagement of Toll-like receptors (TLRs) (67). *ECSIT*-deleted cells show complete loss of mitochondrial complex I function, increase in mROS production, and accumulate damaged mitochondria because of defective mitophagy (68). *PARK7* and *ECSIT* both interact with the mitophagy regulator *PINK1* (69). *PARK7* also regulates lipid rafts mediated endocytosis, and *FLOT1* is one of the main components of lipid rafts (70). Knockout of *PARK7* (*DJ*-1) have been reported to decrease the expression of *FLOT1* (71). Knock down of *FLOT1* results in increased PERK and eIF2α phosphorylation indicating endoplasmic reticulum (ER) stress (72). *PARD6A* is a cell polarity gene that is required for endocytic trafficking (73).

In addition to the genes *DHPS, FLOT1, AIP* and *CREB3*, the overall transcriptional repression of ribosomal proteins seen in Table 6 and Figure 6 also led us to consider the possibility that the cells that have high level of L1HS transcripts may be exhibiting activation of the unfolded protein response (UPR) due to ER stress. Although the UPR pathway is mainly regulated at the protein level through phosphorylation and cleavage, we examined whether any genes in the UPR pathway showed correlation with L1HS at the transcript level that was replicated in at least two tissue types. We found that L1HS transcript level was positively correlated with *EIF2AK2*(*PKR*), *ERN1*(*IRE1α*), and *NFE2L2* (*NRF2*), but negatively correlated with *ATF4*(*CREB2*) and *CHOP*(*DDIT3*) (Figure 7). We did not observe correlation between L1HS and the other transducers of UPR, *PERK* or *ATF6*, nor did we see any correlation between *KEAP1* and L1HS. We did see that the gene *SQSTM1*, encoding p62 that binds to KEAP1, was negatively correlated with L1HS. We also saw several chaperone proteins, ubiquitination factors and other regulators of ER stress correlated with L1HS. *DNAJC5, HSPA4L, UBE4B, MBTPS2, ITPR1, ITPR2,* and *ITPR3* was positively correlated with L1HS and *DNAJC4, DNAJC17, DNAJC19, UBE2G2, SSR2 (TRAPB), BNIP1* was negatively correlated with L1HS in at least two tissues. Excluding the annotations related to mitochondria, the functional annotations “proteasome”, “autophagy”, “endosome”, and “antioxidant” were the top annotations enriched in the negatively correlated gene set (Table 6).

Since *XBP1* is one of the genes in the UPR pathway regulated through splicing, we examined if the XBP1s isoform correlates with the L1HS level in our samples. We tested the log2 transformed raw counts of the XBP1s isoform against the log2 transformed L1HS level, and we also tested the ratio of the spliced isoform over all transcripts XBP1s/(XBP1u+XBP1s) against the log2 transformed L1HS level. But neither quantification of the spliced version of *XBP1* correlated with L1HS consistently across tissues. We found that the ratio of the spliced form positively correlated with L1HS levels in the colorectal samples, but it was not replicated in any other tissue.

To find whether there are common transcription factors upstream of the genes correlated with L1HS, we took our list of 1056 genes negatively correlated with L1HS and looked for common regulatory elements using the web server DiRE (https://dire.dcode.org) (74). We found that the transcription factor binding sites (TFBS) of *NRF2* and *ELK1* are the most frequently identified motifs for our list of genes. 476 genes out of 1056 had at least one TFBS identified, and of those 99 (20.69%) included at least one *NRF2* binding site and 105 (22.52%) included at least one *ELK1* binding site. Unlike *NRF2*, we did not find any significant correlation between *ELK1* and L1HS. But, as we describe in the next section, several genes in the MAPK signaling pathway were positively correlated with L1HS in multiple tissues.

Expression of chromatin modifiers are positively correlated with L1HS expression

Compared to the genes in negative correlation with L1HS, there was a larger number of genes in positive correlation with L1HS in expression. We found 4338 genes that showed positive correlation in expression with L1HS (q-value < 10e-4) in at least one tissue out of the ten tissue groups (Supplementary Table 2). There were 1318 genes that were positively correlated with L1HS in more than one tissue. We did a gene set enrichment analysis using DAVID for the 1318 genes that are positively correlated with L1HS in more than one tissue. After excluding the very general categories of "kinase" and "transcription" that consisted of 293 and 768 genes respectively, we saw SH3 domains, DNA damage and repair, bromodomains, PHD fingers, and chromodomain helicases as most enriched functional categories for genes positively correlated with L1HS expression (Table 8, Supplementary Table 4, Figure 8).

Bromodomains are acetyllysine binding domains and PHD domains are methyllysine binding domains. They are both chromatin readers that recognize modified chromatin and function in epigenetic control of gene transcription. PHD and bromodomains are found in tandem in the C-terminal of several chromatin-associated proteins, including the transcriptional intermediary factor 1 (TIF1) family. The best-known member of the TIF1 family with PHD-bromo domains is the co-repressor *TRIM28* (*KAP1*) that binds to multiple KRAB C2H2 zinc-finger family of proteins and recruits associated mediators of histone and DNA methylation to silence endogenous retroviruses (75). *KAP1* did not show significant correlation with L1HS in our data in any tissue. One other member of the TIF1 family, *TRIM33*(*TIF1γ*) showed positive correlation with L1HS in four tissues, esophagus/stomach, head and neck, kidney and thyroid gland. There was no direct relationship between whether the protein had an activating or repressing role, and whether the protein was positively or negatively correlated in expression with L1HS. For example, *KDM5A*, *KDM5B* (Figure 4), and *KDM5C*, KDM5 family of H3K4 lysine demethylases that function as transcriptional corepressors were all positively correlated with L1HS in expression. On the other hand, *TET2* and *TET3* zinc finger proteins that can actively de-methylate DNA through oxidization of 5-methylcytosine were also positively correlated in expression with L1HS.

Because the list of genes was too long, we didn't list all the genes for the kinase category or the transcription category. But some notable genes in those categories include genes in the MAPK signaling pathway. *IGF1R, SOS1, NRAS, BRAF, MAPK1*(*ERK1*), and *RPS6KA6*(*RSK4)* were all positively correlated with L1HS in multiple tissues. *GSK3B* and *PI3K(PIK3CA*) also showed positive correlation with L1HS levels. *EP300* (*p300*) and *CREBBP*(*CBP*), histone acetyltransferases that act as coactivators were positively correlated with L1HS, and *CARD14*(*CARMA2*), *EIF2AK2*(*PKR*), *TRAF6* and *NF-κB* were positively correlated with L1HS in transcript levels.

Table 9 shows the genes with the best REC scores that have recurrent positive correlation across tissues. We removed genes *RASEF*, *MAST4*, and *LPP* from the list, because they have L1HS sequences in their introns, that can lead to spurious positive correlation. Although we correct for the TEs embedded in introns, the correction is not complete, as we discuss in the Discussion section below.

Expression of KRAB-ZFPs are positively correlated with expression of older LINE elements in the normal tissues

As can be seen in Figure 9, various zinc finger proteins are both positively and negatively correlated with L1HS and older LINEs. But, overwhelmingly more zinc fingers are positively correlated with LINE expression, rather than negatively. Also, if we focus on the Krüppel associated box (KRAB) domain, zinc fingers with KRAB domains were significantly enriched among genes positively correlated with older LINE elements but less among genes correlated with the young and active L1HS (Supplementary Table 4). There are three ZFPs that show recurrent negative correlation with L1HS in multiple tissues, *ZNF511*, *ZNF32*, *ZNF174* (Table 7), but none of these were KRAB-ZFPs, since they were missing the KRAB domain. We initially predicted a negative relationship given the known repressive function of the KRAB domain, and the well characterized role of *KAP1* in TE silencing. But again, as seen with the chromatin modifiers, the functional role of the protein and the co-expression relationship is not simplistic, and one cannot predict whether the protein is a repressor or an activator from the direction of co-expression. To understand this relationship between KZFPs and LINE elements, we decided to examine the co-expression in more detail by looking at all the subfamilies within LINEs and each LINE locus within the subfamilies.

Among the ~400 KZFPs that were identified by Imbeault et al. (51), we found 366 KZFPs in our expression data. We first tested the correlation between those 366 KZFPs and 294 LINE subfamilies using the family level expression quantification. We found 31,187 pairs with significant correlations (qvalue < 0.0001), amounting to about 2.90% out of a total of 1,076,040 pairwise combinations (366 KZFPs x 294 LINE subfamilies x 10 tissue groups). There were 29,484 pairs with positive correlation and 1,703 with negative correlation at the LINE family level, Next, focusing on those 31,187 KZFP, LINE family pairs with significant correlation, we then tested the focal KZFP against the expression level of all individual LINE loci belonging to that family of LINE in the tissue where the significant relationship was observed. There were 49,313 KZFP and LINE locus pairs that were significantly correlated (qvalue < 0.0001 and partial-eta^2 > 0.4), 41,067 pairs with positive correlation and 8,246 with negative correlation at the individual LINE locus level. This is about 0.05 % of all 90,696,187 pairwise combinations that we tested. Bearing in mind that we were only testing ZFP-LINE locus pairs that we already knew had a correlation at the family level, this tells us that most of the correlation we observed at the LINE family level were due to only a small number of individual LINE loci contributing to the correlation.

The correlation between KZFP and each individual LINE locus was very tissue-specific, as 37,994 out of 41,067 KZFP-LINE locus pairs that were positively co-expressed were significant in only one tissue (92.5%). Still, there were some ZFPs that correlated with multiple loci within a family, or even multiple families. For example, *ZNF780B, ZNF587, ZNF793* had very promiscuous correlations across the board with multiple LINE families, multiple loci within a LINE family, and in multiple tissues. These zinc finger proteins and the LINE families were all correlated among each other, creating a network cluster of co-expression. But, for the most part, most correlations appeared between specific zinc finger proteins and specific LINE families, in a limited number of LINE loci, and in specific tissues. For example, *ZIM3* showed positive correlation with L1 family L1PBb. When correlation was tested between *ZIM3* and all L1PBb loci, only 221 out of 3415 (6.4%) of the L1PBb loci showed positive correlation with *ZIM3*, and of those 209 (94.5%) showed positive correlation in the tissue group esophagus and stomach. Figure 10 shows an example of correlation between KZFPs and individual L1M2b loci in the breast tissue.

Again, confirming what we found in our initial analysis, the families that were found as significantly co-expressed with KZFPs were mostly older LINE families. The total count of significant loci were largest for L2 families, L2c, L2a, L2b and L2 in that order, but, that is mainly because L2 elements have the largest number of loci in the human genome. If we adjust for the total number of loci available per family, then the families that are most frequently co-expressed with the ZFPs were elements that are shared in the primate lineages, such as L1PA12, L1PBb, L1P2, L1PA6, L1PA7. The list did not include the youngest element L1HS or L1PA2, L1PA3. There are five KRAB-ZFPs that have been shown to bind to the youngest L1HS sequence in earlier Chip-Seq studies, *ZNF425, ZNF382, ZNF84, ZNF141* and *ZNF649* (51, 76). We looked at whether any of these proteins showed correlation with L1HS in the expression level, but we did not find any KRAB-ZFPs that showed significant correlation with L1HS.

Binding motifs of KZFPs are enriched in the sequence of the co-expressed LINE locus among KZFP-LINE locus pairs that show significant co-expression

We searched for the known motifs of KZFPs identified in Imbeault et al. (51) in the sequences of the L1 loci found to be significantly correlated with said KZFP. We found a total of 4222 identified motifs in our total set of co-expressed KZFP-L1 pairs, 2256 of them being unique. From the 1000 simulations of our first null model, where we sampled randomly from all zinc fingers and L1 instances available in our dataset for each run, the number of motifs we found ranged from 3127 to 3627, with a mean of 3351.466, median of 3355, and a standard deviation of 78.999. In the second null model, we utilized the same dataset of co-expressed zinc fingers and L1 sequences as our experimental model but randomly permuted the L1 sequences, so the pairwise relationship do not reflect the co-expression any more. With 1000 simulations we observed the number of motifs ranging from 2935 to 3519, with a mean of 3255.095, a median of 3254, and a standard deviation of 81.481. Our observation of 4222 identified motifs in the co-expressed pairs showed a p-value of zero under both null model (Figure 11), indicating a significant excess of the corresponding motifs in the KZFP-L1 locus pairs that show significant co-expression.

Results are replicated in an independent dataset of normal tissue samples of the lung.

All our samples come from supposedly normal tissues that were selected by pathologists based on histology, although they are adjacent to the cancer cells. But, since suppression of mitochondria and oxidative phosphorylation, replaced by enhanced aerobic glycolysis is one of the hallmarks of cancer, referred to as the Warburg effect, there is a possibility that our results reflect the cancerous environment that these "normal tissue" are experiencing, including hypoxic conditions. To find out if we could replicate the results in a true normal tissue, we used an independent dataset from the GTEX project to test the correlation with the same analysis.

First, we confirmed that the lung tissues clustered well with the lung tissues from the TCGA dataset based on the gene expression patterns (Supplementary Figure 7a). When examining the LINE expression, there were a few outliers in the GTEX lung samples with unusually high LINE expression that formed a cluster of their own separate from the other lungs, and instead grouping with two other stomach samples. These samples were interesting, because they showed LINE expressions that were significantly higher than any of the samples seen in the TCGA dataset, yet the gene expressions were not especially different from other lungs in GTEX nor in TCGA (Supplementary Figure 7 a vs. b).

The negative relationship between mitochondrial genes, ribosomal protein genes and L1HS expression was replicated in the lung transcriptome data from GTEX in the linear model without controlling for older LINEs as the covariate, but was not replicated when we controlled for the older LINEs. We are not sure why there is this discrepancy. But, the TCGA data is a more robust result because we only included genes that showed correlation in multiple tissues. With the GTEX data, we only had one tissue type, the lung, so we did not require the additional filtering of recurrent correlation. When examining the linear model without controlling for older LINEs as a covariate, we confirmed that the main results were replicated. There was enrichment of terms mitochondrial genes, ribosomal protein genes, response to reactive oxygen species, autophagy, proteasome, among the genes negatively correlated with L1HS. There was enrichment of KRAB-ZFPs, among genes positively correlated with both L1HS and older LINE elements.

DISCUSSION

Limitations to the quantification and correction

Quantifying transposon transcripts is a difficult problem, due to their ambiguity in short read mapping because of repeated content in the reference genome. Current state of the art methods rely on Expectation-Maximization to account for the uncertainty in multi-mapped reads (40). Because the estimation of multi-mapped reads start from values initialized based on the uniquely mapped reads, the final estimates tend to correlate well with the uniquely mapped read counts. This works well with gene transcripts where more than 90% of the reads are estimated to map uniquely across all genes (77), and the variance in the mappability across genes is not large. But with transposons, not only are the mappability lower than the genes, the variance of the mappability across different loci are huge due to the large number of similar sequences. This can lead to biases in the quantification if one locus has high mappability due to more unique mutations, while another locus has lower mappability due to smaller number of unique mutations by chance. Focusing only on uniquely mapped reads doesn't really solve this problem, and will lead to equally biased quantification, as evidenced by the high correlation observed between uniquely mapped reads and the total reads including multi-mapped reads.

In addition, the reference genome that are used for mapping the transposon transcripts do not include all the polymorphic TE insertions in the human population. If the transcript originated from a polymorphic TE that is not present in the human reference genome, it will be redirected to the most similar locus that is present in the reference genome. This will not be a serious problem when quantification is done at the family level, but if one is interested in locus level quantification it becomes a serious problem, especially because the polymorphic loci could potentially be the more active loci that are expressed at higher levels compared to other loci of the same family.

Another complication in TE transcript quantification is that TEs are frequently embedded within introns that are transcribed before they are processed, or sometimes fail to be spliced out, or embedded within exons or non-coding RNAs that are expressed in different conditions (41). To account for this source of error, we introduced a method to correct for TE reads coming from retained introns or pre-mRNA. Although we observed large corrections for specific TE elements embedded within introns, the correction is not complete. The main limitation to our approach stems from the fact that the correction is done after the EM algorithm, after the multi-mapped counts are probabilistically assigned to multiple TE loci in the genome. The correction only removes the equivalent of read counts assigned to the problematic locus, and the uncertainty in the assignment through EM means there may be remaining reads assigned to alternative locations, that actually originate from retained introns as well. A more accurate approach would be to correct for the read counts from retained introns before the EM algorithm based on the uniquely mapped reads, and then run EM based on the corrected counts. But, estimating the depth of the repeat region using uniquely mapped reads is a difficult problem. The effective length of the uniquely mapped region is difficult to estimate, because again mappability varies from locus to locus for any TE, depending on the unique mutations it has accumulated. So, for this study, we decided to use the easier approach to run EM first, and probabilistically assign the TE reads, and then correct based on the expected read depth across the length of the TE locus.

Although we partially dealt with the third problem described above, we have ignored the first and second problems, and we recognize the limitations of the current methods. An important future study would be to study the mappability of individual TE loci carefully, including the known polymorphic sites, and to design a software for TE quantification that can take into account the mappability of each locus in its EM algorithm, as well as correct for the retained introns while considering the effective length of the uniquely mappable region within the TE. Despite these limits, we believe the main results of gene correlations found with L1HS and older LINEs are not affected by the quantification. The uncertainty lies in the assignment of reads to each LINE families, but the total reads coming from LINEs that are found in the RNAseq are valid. Our correlation analysis was done at the family level for each tissue separately. We may have under- or over-estimated the L1HS read counts, but as long as the errors are similar in distribution within each tissue type, and the errors are not systematically biased to be associated with a certain set of genes, it should not be a serious problem. We tried to make the results robust by dropping the reads mapping to LINEs most closely related to L1HS, *i.e.* L1PA2 and L1PA3, and requiring the correlation be replicated in at least two tissue types. We also confirmed that the results were replicated using uniquely mapped reads only.

Stress and TE expression

Initially, when we started the project, our goal was to identify candidate genes involved in transposon control, based on the co-expression analysis. But, once the analysis was done, the results were pointing to what induces LINE expression, rather than what suppresses LINE expression. Among the genes known to function in transposon control, *RNaseH2C* (Figure 4), and *RNaseH2A* showed negative correlation with L1HS, *MPHOSPH8* encoding the protein MPP8 and *C3orf63* encoding TASOR, part of the HUSH complex identified in Liu et al. (28), showed positive correlation with L1HS in multiple tissues in our results. But several well-known genes with functions in transposon control, *e.g. MORC2, SIRT6, KAP1, SAMHD1, MOV10, ZAP, C12orf35* (human ortholog of *RESF1*), etc. are missing in our list of significantly correlated transcripts. Occasionally they would show up as positively correlated in one of the tissues, but not replicated in more than one tissue, which was what we required to be called as correlated genes. Instead, the major theme that emerged from our results is stress response. Genes involved in oxidative stress such as *ISCU*, *PARK7*, or *ECSIT* and genes such as *DHPS*, *FLOT1* and *CREB3* in the UPR pathway, showed negative correlation across multiple tissues. That leads to the hypothesis that oxidative stress or ER stress may be inducing the high expression of LINEs. Mitochondrial damage, oxidative stress and proteotoxic stress are tightly linked, and causal relationship in all directions have been observed. Misfolded protein accumulation can induce reactive oxygen species (ROS), and constitutively activated UPR has been shown to induce expression of Ty2 transposon in yeast (78, 79). Defective mitochondria result in elevated mROS production, and mROS can lead to more damage in the mitochondrial DNA. It is also possible that L1HS expression is the cause, not the result, of the stress response. Observation of co-localization between LINE elements and stress granules (80) supports the co-occurrence of LINE expression and UPR, but how expression of ORF proteins and the damage to DNA are linked with UPR is not well understood (81, 82). We found that transcripts of *PKR* and *NRF2* are positively correlated with L1HS, and the binding motif of *NRF2* are enriched among the genes negatively correlated with L1HS. But, we did not find higher level or higher proportion of the active isoform XBP1(S) in samples with higher L1HS levels. It is possible that other transducers are involved, but further experiments would be needed to evaluate the relationship between oxidative stress, ER stress and the expression of L1HS.

Ever since the initial discovery of transposons, stress has been associated with transposon activity (83). Numerous reports have documented transposon expression and transposon activity induced by various biotic and abiotic stresses (84–87). In humans, various stresses have been shown to induce LINE1 transcription or activation including chemical compounds (88–90), radiation (91, 92), oxidative stress (93) and aging (94). Most of these studies have observed L1 activity *in vitro*, by exposing cultured cells to stress factors and assaying the retrotransposition activity (but see (95) in mouse). Our study is the first that reports observation on L1 expression *in vivo* in normal human tissue samples, and shows widespread association between L1HS expression and genes modulating stress response. In this light, it is interesting that we were not able to observe correlation between age and L1 transcript levels in our data. It may be because, although our samples are all normal tissue, they come from cancer patients and thus the age distribution is skewed towards older age. The range is from 15 to 90, but the median age across all samples is 62. We did find association between past radiation therapy and L1HS expression level but only in the thyroid tissue. The exposure and damage to normal tissues resulting from radiation therapy has been long recognized (96), and the sensitivity to radiation is known to vary depending on the tissue or organ, as well as the genetics of the patient (97). Our results show that one of the responses to radiation damage in the adjacent thyroid tissue may be a higher level of L1HS expression (Supplementary Figure 8) and it may have implications for future rate of recurrence.

TE expression in disease

The genes and pathways that we find to be correlated with L1HS expression are interesting when we think about the accumulated evidence of transposon activity in cancer, and more recent reports of transposon dysregulation in neurodegenerative diseases. Suppressed mitochondrial function and altered metabolism, are one of the hallmarks of cancer, and abnormal MAPK signaling is implicated in a wide range of cancers. Oxidative stress has also been linked to cancer through increased DNA damage, and genome instability. In our preliminary analysis, we observed that the same correlations are found in the cancer tissue samples as well, but we focused on the normal tissues for the scope of this study, to avoid the confounding factors of malignant transformation. The fact that we observe these correlations in normal tissue, albeit obtained adjacent to the cancer tissue, could mean that LINE dysregulation is one of the earlier steps in the initiation of cancer and could potentially contribute to the progression of cancer through the damage that they cause to the DNA. Similarly, mitochondrial dysfunction, oxidative stress and proteotoxic stress, and more recently transposon expression are some of the common features that are observed in the tissue samples or animal models of neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS) (98–100) and Alzheimer (101). Our observation that the link between mitochondria, ER stress, and LINE expression is observed across multiple tissues other than brain that are considered normal and healthy, indicate that cellular stress, transcriptional response and the induction of L1 expression may be a fundamental process that happens quite frequently across different organs. Whether such expression becomes a burden to the cell and contributes to disease, and if so, why the process leads to more serious consequences in certain tissues such as brain are questions that will need to be explored.

TEs and KRAB-ZFPs

Although LINE elements have been studied for a long time, their ubiquitous and highly tissue-specific expression patterns are starting to be appreciated only recently. The fact that LINEs compose close to 19% of the human genome is frequently emphasized, but the fact that there is observable amount of LINE transcripts in human RNA-seq data has mostly been ignored or regarded as a nuisance without any functional relevance. But, recently, important regulatory roles for LINEs are emerging with observations of contribution to transcription start sites (33), active transcription during early development (102), and even critical function similar to long non-coding RNAs that guide chromatin-remodeling complexes to specific loci in the genome (103). ChIP-Seq studies on KRAB-ZFPs have identified extensive binding between this family of proteins and transposable elements including LINEs (51, 76), implying a role for regulating TE expression. Our results add complementary evidence at the RNA level, by revealing widespread correlation in transcripts across multiple KZFPs and LINE elements that are generally tissue specific and LINE locus specific. We also found that despite the overall sequence similarity among the LINE elements, the transcriptionally correlated pairs showed enrichment of the specific ZFP motifs in the LINE sequence, compared to randomly generated pairs. Given the correspondence between the presence of the binding motif and the co-expression relationship between KZFP and LINEs, it is possible that the co-expression we are observing is leaky transcription accompanying KZFP binding at the LINE locus, similar to the observations of transcription in enhancer RNAs. The tissue-specificity of the correlated expression hints that the *in vivo* binding patterns or the effect of the binding may be different depending on the tissues. Our results may be informative in designing future experiments on the types of cell lines or tissues and the ZFP proteins to explore with ChIP-seq assays. We could also potentially utilize the transcriptional correlation to predict motifs for about one third of the KZFPs, of which we have no information on their binding patterns yet.

AVAILABILITY

The pipeline and data can be found at https://github.com/HanLabUNLV/LINEexpression. The modified version of the TEtranscripts (40) software can be found at https://github.com/HanLabUNLV/tetoolkit.

FUNDING

This work was supported by the National Institutes of Health [GM116108 to M.V.H.]. Funding for open access charge: National Institutes of Health.

ACKNOWLEDGEMENT

REFERENCES

1. Chen,J.-M., Stenson,P.D., Cooper,D.N. and Férec,C. (2005) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum. Genet.*, **117**, 411–427.

2. Rosser,J.M. and An,W. (2012) L1 expression and regulation in humans and rodents. *Front. Biosci. Elite Ed.*, **4**, 2203–2225.

3. Gasior,S.L., Wakeman,T.P., Xu,B. and Deininger,P.L. (2006) The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *J. Mol. Biol.*, **357**, 1383–1393.

4. Wallace,N.A., Belancio,V.P. and Deininger,P.L. (2008) L1 mobile element expression causes multiple types of toxicity. *Gene*, **419**, 75–81.

5. Belgnaoui,S.M., Gosden,R.G., Semmes,O.J. and Haoudi,A. (2006) Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer Cell Int.*, **6**, 13.

6. Branciforte,D. and Martin,S.L. (1994) Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol. Cell. Biol.*, **14**, 2584–2592.

7. Trelogan,S.A. and Martin,S.L. (1995) Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc. Natl. Acad. Sci.*, **92**, 1520–1524.

8. Ergün,S., Buschmann,C., Heukeshoven,J., Dammann,K., Schnieders,F., Lauke,H., Chalajour,F., Kilic,N., Strätling,W.H. and Schumann,G.G. (2004) Cell Type-specific Expression of LINE-1 Open Reading Frames 1 and 2 in Fetal and Adult Human Tissues. *J. Biol. Chem.*, **279**, 27753–27763.

9. Kubo,S., Seleme,M. del C., Soifer,H.S., Perez,J.L.G., Moran,J.V., Kazazian,H.H. and Kasahara,N. (2006) L1 retrotransposition in nondividing and primary human somatic cells. *Proc. Natl. Acad. Sci.*, **103**, 8036–8041.

10. Belancio,V.P., Roy-Engel,A.M., Pochampally,R.R. and Deininger,P. (2010) Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res.*, **38**, 3909–3922.

11. Rangwala,S.H., Zhang,L. and Kazazian,H.H. (2009) Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol.*, **10**, R100.

12. Skowronski,J., Fanning,T.G. and Singer,M.F. (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.*, **8**, 1385–1397.

13. Skowronski,J. and Singer,M.F. (1985) Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc. Natl. Acad. Sci. U. S. A.*, **82**, 6050–6054.

14. Bratthauer,G.L., Cardiff,R.D. and Fanning,T.G. (1994) Expression of LINE-1 retrotransposons in human breast cancer. *Cancer*, **73**, 2333–2336.

15. Rodić,N., Sharma,R., Sharma,R., Zampella,J., Dai,L., Taylor,M.S., Hruban,R.H., Iacobuzio-Donahue,C.A., Maitra,A., Torbenson,M.S., *et al.* (2014) Long Interspersed Element-1 Protein Expression Is a Hallmark of Many Human Cancers. *Am. J. Pathol.*, **184**, 1280–1286.

16. Swergold,G.D. (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.*, **10**, 6718–6729.

17. Yang,N., Zhang,L., Zhang,Y. and Kazazian Jr,H.H. (2003) An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res.*, **31**, 4929–4940.

18. Tchénio,T., Casella,J.-F. and Heidmann,T. (2000) Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res.*, **28**, 411–415.

19. Athanikar,J.N., Badge,R.M. and Moran,J.V. (2004) A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.*, **32**, 3846–3855.

20. Thayer,R.E., Singer,M.F. and Fanning,T.G. (1993) Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1 -encoded protein. *Gene*, **133**, 273–277.

21. Kuramochi-Miyagawa,S., Watanabe,T., Gotoh,K., Totoki,Y., Toyoda,A., Ikawa,M., Asada,N., Kojima,K., Yamaguchi,Y., Ijiri,T.W., *et al.* (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev.*, **22**, 908–917.

22. Aravin,A.A., Sachidanandam,R., Bourc'his,D., Schaefer,C., Pezic,D., Fejes Toth,K., Bestor,T. and Hannon,G.J. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol. Cell*, **31**, 785–799.

23. Newkirk,S.J., Lee,S., Grandi,F.C., Gaysinskaya,V., Rosser,J.M., Vanden Berg,N., Hogarth,C.A., Marchetto,M.C.N., Muotri,A.R., Griswold,M.D., *et al.* (2017) Intact piRNA pathway prevents L1 mobilization in male meiosis. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, E5635–E5644.

24. Goodier,J.L., Cheung,L.E. and Kazazian,H.H. (2013) Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res.*, **41**, 7401–7419.

25. Taylor,M.S., LaCava,J., Dai,L., Mita,P., Burns,K.H., Rout,M.P. and Boeke,J.D. (2016) Characterization of L1-Ribonucleoprotein Particles. *Methods Mol. Biol. Clifton NJ*, **1400**, 311–338.

26. Peddigari,S., Li,P.W.-L., Rabe,J.L. and Martin,S.L. (2013) hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. *Nucleic Acids Res.*, **41**, 575–585.

27. Moldovan,J.B. and Moran,J.V. (2015) The Zinc-Finger Antiviral Protein ZAP Inhibits LINE and Alu Retrotransposition. *PLOS Genet.*, **11**, e1005121.

28. Liu,N., Lee,C.H., Swigut,T., Grow,E., Gu,B., Bassik,M.C. and Wysocka,J. (2017) Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature*, **553**, 228.

29. Wolf,D. and Goff,S.P. (2009) Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature*, **458**, 1201–1204.

30. Jacobs,F.M.J., Greenberg,D., Nguyen,N., Haeussler,M., Ewing,A.D., Katzman,S., Paten,B., Salama,S.R. and Haussler,D. (2014) An evolutionary arms race between KRAB zinc-finger genes *ZNF91/93* and SVA/L1 retrotransposons. *Nature*, **516**, 242–245.

31. Rowe,H.M. and Trono,D. (2011) Dynamic control of endogenous retroviruses during development. *Virology*, **411**, 273–287.

32. Trono,D. (2015) Transposable Elements, Polydactyl Proteins, and the Genesis of Human-Specific Transcription Networks. *Cold Spring Harb. Symp. Quant. Biol.*, **80**, 281–288.

33. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T., *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.

34. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

35. The Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330.

36. The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

37. The Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519.

38. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M., *et al.* (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178–e178.

39. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

40. Jin,Y., Tam,O.H., Paniagua,E. and Hammell,M. (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinforma. Oxf. Engl.*, **31**, 3593–3599.

41. Deininger,P., Morales,M.E., White,T.B., Baddoo,M., Hedges,D.J., Servant,G., Srivastav,S., Smither,M.E., Concha,M., DeHaro,D.L., *et al.* (2017) A comprehensive approach to expression of L1 loci. *Nucleic Acids Res.*, **45**, e31–e31.

42. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

43. Kadota,K., Nishiyama,T. and Shimizu,K. (2012) A normalization strategy for comparing tag count data. *Algorithms Mol. Biol.*, **7**, 5.

44. Huber,W., von,H.A., Sueltmann,H., Poustka,A. and Vingron,M. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.*, **2**.

45. Kolde,R. (2018) pheatmap: Pretty Heatmaps.

46. Kieffer,J. (1994) *SIAM Rev.*, **36**, 509–511.

47. Jacobsen,A., Silber,J., Harinath,G., Huse,J.T., Schultz,N. and Sander,C. (2013) Analysis of microRNA-target interactions across diverse cancer types. *Nat. Struct. Mol. Biol.*, **20**, 1325–1332.

48. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

49. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**, 15545–15550.

50. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

51. Imbeault,M., Helleboid,P.-Y. and Trono,D. (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, **543**, 550–554.

52. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

53. Barazandeh,M., Lambert,S.A., Albu,M. and Hughes,T.R. (2018) Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3 Bethesda Md*, **8**, 219–229.

54. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

55. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.*, **29**, 15–21.

56. Britten,R.J. (1997) Mobile elements inserted in the distant past have taken on important functions. *Junk DNA Role Evol. Non-Coding Seq.*, **205**, 177–182.

57. Philippe,C., Vargas-Landin,D.B., Doucet,A.J., Essen,D. van, Vera-Otarola,J., Kuciak,M., Corbin,A., Nigumann,P. and Cristofari,G. (2016) Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife*, **5**, e13926.

58. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci.*, **100**, 5280–5285.

59. Beck,C.R., Collier,P., Macfarlane,C., Malig,M., Kidd,J.M., Eichler,E.E., Badge,R.M. and Moran,J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, **141**, 1159–1170.

60. Tubio,J.M.C., Li,Y., Ju,Y.S., Martincorena,I., Cooke,S.L., Tojo,M., Gundem,G., Pipinikas,C.P., Zamora,J., Raine,K., *et al.* (2014) Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, **345**, 1251343–1251343.

61. Desai,N., Sajed,D., Arora,K.S., Solovyov,A., Rajurkar,M., Bledsoe,J.R., Sil,S., Amri,R., Tai,E., MacKenzie,O.C., *et al.* (2017) Diverse repetitive element RNA expression defines epigenetic and immunologic features of colon cancer. *JCI Insight*, **2**.

62. Solovyov,A., Vabret,N., Arora,K.S., Snyder,A., Funt,S.A., Bajorin,D.F., Rosenberg,J.E., Bhardwaj,N., Ting,D.T. and Greenbaum,B.D. (2018) Global Cancer Transcriptome Quantifies Repeat Element Polarization between Immunotherapy Responsive and T Cell Suppressive Classes. *Cell Rep.*, **23**, 512–521.

63. Menendez,L., Benigno,B.B. and McDonald,J.F. (2004) L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. *Mol. Cancer*, **3**, 12.

64. LaRocca,T.J., Gioscia-Ryan,R.A., Hearon,C.M. and Seals,D.R. (2013) The autophagy enhancer spermidine reverses arterial aging. *Mech. Ageing Dev.*, **134**, 314–320.

65. Bock,K.W. (2012) Ah receptor- and Nrf2-gene battery members: Modulators of quinone-mediated oxidative and endoplasmic reticulum stress. *Biochem. Pharmacol.*, **83**, 833–838.

66. Liang,G., Audas,T.E., Li,Y., Cockram,G.P., Dean,J.D., Martyn,A.C., Kokame,K. and Lu,R. (2006) Luman/CREB3 Induces Transcription of the Endoplasmic Reticulum (ER) Stress Response Protein Herp through an ER Stress Response Element. *Mol. Cell. Biol.*, **26**, 7999–8010.

67. West,A.P., Brodsky,I.E., Rahner,C., Woo,D.K., Erdjument-Bromage,H., Tempst,P., Walsh,M.C., Choi,Y., Shadel,G.S. and Ghosh,S. (2011) TLR signalling augments macrophage bactericidal activity through mitochondrial ROS. *Nature*, **472**, 476–480.

68. Carneiro,F.R., Lepelley,A., Seeley,J.J., Hayden,M.S. and Ghosh,S. (2018) An Essential Role for ECSIT in Mitochondrial Complex I Assembly and Mitophagy in Macrophages. *Cell Rep.*, **22**, 2654–2666.

69. Hao,L.-Y., Giasson,B.I. and Bonini,N.M. (2010) DJ-1 is critical for mitochondrial function and rescues PINK1 loss of function. *Proc. Natl. Acad. Sci.*, **107**, 9747.

70. Kim,K.S., Kim,J.S., Park,J.-Y., Suh,Y.H., Jou,I., Joe,E.-H. and Park,S.M. (2013) DJ-1 Associates with lipid rafts by palmitoylation and regulates lipid rafts-dependent endocytosis in astrocytes. *Hum. Mol. Genet.*, **22**, 4805–4817.

71. Kim,J.-M., Cha,S.-H., Choi,Y.R., Jou,I., Joe,E.-H. and Park,S.M. (2016) DJ-1 deficiency impairs glutamate uptake into astrocytes via the regulation of flotillin-1 and caveolin-1 expression. *Sci. Rep.*, **6**, 28823.

72. Fork,C., Hitzel,J., Nichols,B.J., Tikkanen,R. and Brandes,R.P. (2014) Flotillin-1 facilitates toll-like receptor 3 signaling in human endothelial cells. *Basic Res. Cardiol.*, **109**, 439.

73. Silva-Ayala,D., López,T., Gutiérrez,M., Perrimon,N., López,S. and Arias,C. (2013) Genome-wide RNAi screen reveals a role for the ESCRT complex in rotavirus cell entry.

74. Gotea,V. and Ovcharenko,I. (2008) DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res.*, **36**, W133–W139.

75. Rowe,H.M., Jakobsson,J., Mesnard,D., Rougemont,J., Reynard,S., Aktas,T., Maillard,P.V., Layard-Liesching,H., Verp,S., Marquis,J., *et al.* (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, **463**, 237–240.

76. Najafabadi,H.S., Mnaimneh,S., Schmitges,F.W., Garton,M., Lam,K.N., Yang,A., Albu,M., Weirauch,M.T., Radovani,E., Kim,P.M., *et al.* (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.*, **33**, 555–562.

77. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

78. Hashimoto,M., Rockenstein,E., Crews,L. and Masliah,E. (2003) Role of protein aggregation in mitochondrial dysfunction and neurodegeneration in Alzheimer's and Parkinson's diseases. *Neuromolecular Med.*, **4**, 21–36.

79. Kimata,Y., Ishiwata-Kimata,Y., Yamada,S. and Kohno,K. (2006) Yeast unfolded protein response pathway regulates expression of genes for anti-oxidative stress and for cell surface proteins. *Genes Cells Devoted Mol. Cell. Mech.*, **11**, 59–69.

80. Goodier,J.L., Zhang,L., Vetter,M.R. and Kazazian,H.H.J. (2007) LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. *Mol. Cell. Biol.*, **27**, 6469–6483.

81. Dicks,N., Gutierrez,K., Michalak,M., Bordignon,V. and Agellon,L.B. (2015) Endoplasmic Reticulum Stress, Genome Damage, and Cancer. *Front. Oncol.*, **5**, 11.

82. Pasquarella,A., Ebert,A., Pereira de Almeida,G., Hinterberger,M., Kazerani,M., Nuber,A., Ellwart,J., Klein,L., Busslinger,M. and Schotta,G. (2016) Retrotransposon derepression leads to activation of the unfolded protein response and apoptosis in pro-B cells. *Development*, **143**, 1788.

83. McClintock,B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792.

84. Pouteau Sylvie, Grandbastien Marie-Angèle and Boccara Martine (1994) Microbial elicitors of plant defence responses activate transcription of a retrotransposon. *Plant J.*, **5**, 535–542.

85. Mhiri,C., Morel,J.-B., Vernhettes,S., Casacuberta,J.M., Lucas,H. and Grandbastien,M.-A. (1997) The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Mol. Biol.*, **33**, 257–266.

86. Hirochika,H., Sugimoto,K., Otsuki,Y., Tsugawa,H. and Kanda,M. (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci.*, **93**, 7783.

87. Ito,H., Gaubert,H., Bucher,E., Mirouze,M., Vaillant,I. and Paszkowski,J. (2011) An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*, **472**, 115.

88. Okudaira,N., Iijima,K., Koyama,T., Minemoto,Y., Kano,S., Mimori,A. and Ishizaka,Y. (2010) Induction of long interspersed nucleotide element-1 (L1) retrotransposition by 6-formylindolo[3,2-b]carbazole (FICZ), a tryptophan photoproduct. *Proc. Natl. Acad. Sci.*, **107**, 18487–18492.

89. Stribinskis,V. and Ramos,K.S. (2006) Activation of Human Long Interspersed Nuclear Element 1 Retrotransposition by Benzo(a)pyrene, an Ubiquitous Environmental Carcinogen. *Cancer Res.*, **66**, 2616–2620.

90. Terasaki,N., Goodier,J.L., Cheung,L.E., Wang,Y.J., Kajikawa,M., Kazazian,H.H.,Jr and Okada,N. (2013) In Vitro Screening for Compounds That Enhance Human L1 Mobilization. *PLOS ONE*, **8**, e74629.

91. Banaz-Yaşar,F., Gedik,N., Karahan,S., Diaz-Carballo,D., Bongartz,B.M. and Ergün,S. (2012) LINE-1 Retrotransposition Events Regulate Gene Expression After X-Ray Irradiation. *DNA Cell Biol.*, **31**, 1458–1467.

92. Farkash,E.A., Kao,G.D., Horman,S.R. and Prak,E.T.L. (2006) Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res.*, **34**, 1196–1204.

93. Giorgi,G., Marcantonio,P. and Del Re,B. (2011) LINE-1 retrotransposition in human neuroblastoma cells is affected by oxidative stress. *Cell Tissue Res.*, **346**, 383–391.

94. Van Meter,M., Kashyap,M., Rezazadeh,S., Geneva,A.J., Morello,T.D., Seluanov,A. and Gorbunova,V. (2014) SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat. Commun.*, **5**, 5011.

95. Muotri Alysson R., Zhao Chunmei, Marchetto Maria C.N. and Gage Fred H. (2009) Environmental influence on L1 retrotransposons in the adult hippocampus. *Hippocampus*, **19**, 1002–1007.

96. Stone,H.B., Coleman,C.N., Anscher,M.S. and McBride,W.H. (2003) Effects of radiation on normal tissue: consequences and mechanisms. *Lancet Oncol.*, **4**, 529–536.

97. Barnett,G.C., West,C.M.L., Dunning,A.M., Elliott,R.M., Coles,C.E., Pharoah,P.D.P. and Burnet,N.G. (2009) Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype. *Nat. Rev. Cancer*, **9**, 134–142.

98. Li,W., Jin,Y., Prazak,L., Hammell,M. and Dubnau,J. (2012) Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLOS ONE*, **7**, e44099.

99. Li,W., Lee,M.-H., Henderson,L., Tyagi,R., Bachani,M., Steiner,J., Campanac,E., Hoffman,D.A., von Geldern,G., Johnson,K., *et al.* (2015) Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.*, **7**, 307ra153.

100. Krug,L., Chatterjee,N., Borges-Monroy,R., Hearn,S., Liao,W.-W., Morrill,K., Prazak,L., Rozhkov,N., Theodorou,D., Hammell,M., *et al.* (2017) Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLOS Genet.*, **13**, e1006635.

101. Sun,W., Samimi,H., Gamez,M., Zare,H. and Frost,B. (2018) Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nat. Neurosci.*, **21**, 1038–1048.

102. Jachowicz,J.W., Bing,X., Pontabry,J., Bošković,A., Rando,O.J. and Torres-Padilla,M.-E. (2017) LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.*, **49**, 1502.

103. Percharde,M., Lin,C.-J., Yin,Y., Guan,J., Peixoto,G.A., Bulut-Karslioglu,A., Biechele,S., Huang,B., Shen,X. and Ramalho-Santos,M. (2018) A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*, **174**, 391–405.e19.

TABLE AND FIGURES LEGENDS

Table 1. Tissue types and the number of normal tissue samples

RNA-seq data of normal samples were collected from the Cancer Genome Atlas. BLCA (Bladder urothelial carcinoma) , BRCA (Breast carcinoma), CHOL (Cholangiocarcinoma), COAD (Colon adenocarcinoma), ESCA (Esophageal adenocarcinoma), HNSC (Head and neck squamous cell carcinoma), KICH (kidney chromophobe), KIRC (kidney renal clear cell carcinoma), KIRP (Kidney renal papillary cell carcinoma), LIHC (Liver hepatocellular carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), PRAD (Prostate adenocarcinoma), READ (Rectum adenocarcinoma), STAD (Stomach adenocarcinoma), THCA (Thyroid carcinoma) and UCEC (Uterine Corpus Endometrial Carcinoma)

Table 2. Patients with largest difference in L1HS level after correcting for pre-mRNA/retained introns. TCGA patient id, read counts before and after correction, read counts removed, and their tissue type.

Table 3. Transposon loci that show large difference after correcting for pre-mRNA/retained introns. a. L1HS or b. other transposon loci embedded within introns or exons of genes that frequently result in the largest correction in each sample. Locus id, genomic location, surrounding gene and structure the TE is embedded in, and the maximum number of reads removed in a sample.

Table 4. Comparison of clustering outcomes

Normalized Mutual Information was used to measure the agreement between the clustering results and the reference clustering. Clusters were generated by random permutations and hierarchical clustering based on different variables: gene expression, LINE family expression and LINE locus expression.

Table 5. Top 100 L1HS loci with the largest read counts and the samples they are expressed in. a. L1HS loci with the largest read counts. Locus ID, frequency of the locus in the top 100, chromosomal location, length, whether the same locus in found as a source in Tubio et al. (60) b. samples with the largest L1HS read counts. Sample ID, frequency of the sample in the top 100, tissue type.

Table 6. Enriched annotations among genes negatively correlated with L1HS transcripts in at least two or more tissues.

Enriched annotation clusters identified with DAVID. Rank, annotation, annotation terms, enrichment score, number of genes in the annotation cluster, gene names. Enrichment score is the geometric mean of the enrichment scores (modified fisher's exact test) for all terms in the cluster. Rank 4 to rank 8 and rank 10 are additional annotation clusters related with mitochondria that we omitted for brevity. The full table is included in Supplementary Table 4.

Table 7. genes showing highly recurrent negative correlation with L1HS based on REC score.
20 genes with the highest REC score reflecting recurrent negative correlation across multiple tissues.

Table 8. Enriched annotations among genes positively correlated with L1HS transcripts in at least two or more tissues.
Enriched annotation clusters identified by DAVID. Rank, annotation, annotation terms, enrichment score, number of genes in the annotation cluster, gene names. Enrichment score is the geometric mean of the enrichment scores (modified fisher's exact test) for all terms in the cluster. Gene names for the kinase cluster and the transcription cluster are omitted for brevity. The full table is included in Supplementary Table 4.

Table 9. genes showing highly recurrent positive correlation with L1HS based on REC score.
20 genes with the highest REC score reflecting recurrent positive correlation across multiple tissues.

Figure 1. Comparison of TE read counts before and after correcting for pre-mRNAs/retained introns.
a. read counts mapped to L1HS vs. corrected read counts mapped to L1HS after removing transcripts from pre-mRNAs/retained introns. b. read counts uniquely mapped to L1HS vs. corrected read counts. Read counts are log2 transformed.

Figure 2. Heatmap based on 1000 LINE loci with largest variance in expression level.
Heatmap is generated using hierarchical clustering on individual transcript levels of 1000 LINE loci that show largest variation across the samples. Rows represent individual LINE loci, and columns represent individual samples. Tissues are labeled by 17 types defined by TCGA and by 10 broader types defined by grouping the similar tissues together.

Figure 3. Variation in L1HS expression across tissues and individuals
L1HS read counts by tissue, log2 transformed using variance stabilizing transformation in DEseq2, after two rounds of normalization as described in the methods.

Figure 4. Genes with the largest negative and positive correlation with L1HS transcript level, based on the partial eta-squared estimate across multiple tissues.
Left side shows the genes negatively correlated with L1HS transcripts and the right side shows the genes positively correlated with L1HS transcripts. The genes are selected by their overall estimate of partial eta-squared for the gene transcript level in the linear model summed across all tissues.

Figure 5. genes correlated with L1HS in the mitochondria and ribosome.
a. nuclear encoded mitochondrial genes correlated with L1HS in at least two or more tissues. b. ribosomal protein genes correlated with L1HS in at least two or more tissues. Only the genes significant in more than two tissues are colored. The intensity of the colors represents the magnitude

of the coefficient (maximum among all tissues), while the red and green represent the direction of the relationship.

Figure 6. mitochondrial genes and ribosomal protein genes are enriched among genes negatively correlated with L1HS.
a. mitochondrial genes are negatively correlated with L1HS but not with older LINEs. b. ribosomal protein genes and ribonucleoproteins are negatively correlated with L1HS but not with older LINEs. The four columns represent genes either positively or negatively correlated, and with either L1HS as the dependent variable controlling for older LINEs as the covariate, or older LINEs as the dependent variable controlling for L1HS as the covariate. The exact linear models are described in the methods. If the annotation term is enriched among the genes in the category, it is visualized with the color representing the p-value and the area of the circle representing the proportion. The p-value is based on the EASE score from DAVID, and the percentage is based on the proportion of genes observed in our set compared to the total number of genes annotated with the term.

Figure 7. correlation with genes in the Unfolded Protein Response pathway.
Genes in the UPR pathway showing correlation with L1HS in at least two or more tissues are plotted with the regression line.

Figure 8. bromodomains, PHD domains and SET domains are enriched among genes positively correlated with L1HS.
a. bromodomains are positively correlated with L1HS but not with older LINEs. b. PHD domains and SET domains are negatively correlated with L1HS but not with older LINEs. The four columns represent genes either positively or negatively correlated, and with either L1HS as the dependent variable controlling for older LINEs as the covariate, or older LINEs as the dependent variable controlling for L1HS as the covariate. The exact linear models are described in the methods. If the annotation term is enriched among the genes in the category, it is visualized with the color representing the p-value and the area of the circle representing the proportion. The p-value is based on the EASE score from DAVID, and the percentage is based on the proportion of genes observed in our set compared to the total number of genes annotated with the term.

Figure 9. KRAB-Zinc Finger Proteins are enriched among genes positively correlated with older LINE elements.
Zinc Finger Proteins are positively correlated with both L1HS and with older LINEs. KRAB domains show stronger positive correlation with older LINEs than with L1HS. The four columns represent genes either positively or negatively correlated, and with either L1HS as the dependent variable controlling for older LINEs as the covariate, or older LINEs as the dependent variable controlling for L1HS as the covariate. The exact linear models are described in the methods. If the annotation term is enriched among the genes in the category, it is visualized with the color representing the p-value and the area of the circle representing the proportion. The p-value is based on the EASE score from

DAVID, and the percentage is based on the proportion of genes observed in our set compared to the total number of genes annotated with the term.

Figure 10. Correlation between the L1M2b LINE family and KRAB Zinc Finger Proteins in breast tissue.

Correlation between individual LINE loci in the L1M2b family are plotted against several KZFPs. KZFPs showed here are the set of KZFPs that initially showed correlation with L1M2b transcripts at the family level in breast. Colors indicate the direction and the magnitude of the coefficient, and the area of the squares indicate the p-value, log transformed.

Figure 11. Number of ZFP binding motifs identified in the LINE sequence of the co-expressed ZFP-LINE locus pairs.

The number of ZFP binding motifs identified in the LINE sequence of the co-expressed ZFP-LINE locus pairs is compared to the distribution of the number of ZFP binding motifs found in the set of LINE sequences representing the null model. a. null model 1: the set of LINE sequences were generated by randomly sampling LINE loci among all LINEs. b. null model 2: the set of LINE sequences were generated by permutation of the LINE sequence in the co-expressed set, such that the LINE family distribution remains the same, but the ZFP-LINE relationships are disrupted.

TABLES

Table 1. Tissue types and the number of normal tissue samples

tissue	samples
BLCA	19
BRCA	113
CHOL	9
COAD	25
ESCA	11
HNSC	42
KICH	24
KIRC	72
KIRP	30
LIHC	50
LUAD	58
LUSC	51
PRAD	52
READ	6
STAD	32
THCA	59
UCEC	7
Total	660

Table 2. Patients with largest difference in L1HS level after correcting for pre-mRNA/retained introns.

Patient ID	family	read counts before	read counts after	correction	tissue
6852	L1HS	11967	353	11614	STAD
8462	L1HS	9337	6614	2723	STAD
A4GH	L1HS	5998	3908	2090	STAD
A4GY	L1HS	9167	5915	3252	STAD
A4OF	L1HS	4882	2872	2010	ESCA
A4OG	L1HS	6177	4210	1967	ESCA
A4OJ	L1HS	5701	3880	1821	ESCA
A4OM	L1HS	11490	6712	4778	ESCA
AB1V	L1HS	7628	5615	2013	STAD
AB1X	L1HS	9720	7330	2390	STAD

Table 3. Transposon loci that show large difference after correcting for pre-mRNA/retained introns.

a.

locus	chr	start	end	surrounding gene	L1HS embedded in	# of samples	Max correction
L1HS_dup898	9	85664455	85670486	*RASEF*	Intron 1	310	2331
L1HS_dup961	X	67262397	67263226	*OPHN1*	Exon 25	81	285
L1HS_dup144	2	71638605	71644631	*ZNF638*	Intron 20	69	100
L1HS_dup877	9	20655632	20658801	*FOCAD*	Exon 1, Intron 1, exon 2	61	137
L1HS_dup1524	22	29059272	29065303	*TTC28*	Intron 1	34	116

b.

locus	chr	start	end	surrounding gene	TE embedded in	# of samples	Max correction
AluSx1_dup59209	Y	21153222	21153521	*TTTY14*	Exon 1	116	43514
MIRc_dup74805	12	50351953	50352157	*AQP2*	Exon 4	69	61491
MIRc_dup47590	8	22021288	22021431	*SFTPC*	Intron 4, Exon 5	65	456684
AluJb_dup119100	17	16344881	16345132	*C17orf76-AS1*	Intron 4, Exon 5	58	28279
MIRb_dup137684	10	81315669	81315913	*SFTPA2*	Exon 5	40	317785
AluSz6_dup3320	1	207102295	207102608	*PIGR*	Exon 11	36	55137

Table 4. Comparison of clustering outcomes

clusters	Normalized Mutual Information
Tissue types randomly permuted without replacement	0.000229
Tissue types randomly permuted with replacement	0.000264
Hierarchical clustering based on LINE family level expression	0.306346
Hierarchical clustering based on LINE locus level expression	0.914247
Hierarchical clustering based on gene expression	0.929098

Table 5. Top 100 L1HS loci with the largest read counts and the samples they are expressed in.

a.

locus	Freq in top 100	chr	strand	start	end	length	Tubio
L1HS_dup241	23	3	-	26439509	26445536	6028	
L1HS_dup744	18	7	-	65751841	65757871	6031	Y
L1HS_dup605	8	5	-	172829800	172835831	6032	
L1HS_dup425	6	4	-	88268276	88274298	6023	
L1HS_dup170	5	2	-	130173261	130174973	1713	
L1HS_dup795	5	8	-	73787793	73793823	6031	Y
L1HS_dup1353	4	15	-	55218235	55224297	6063	
L1HS_dup413	4	4	+	80888062	80894087	6026	Y
L1HS_dup469	4	4	-	137214650	137220701	6052	
L1HS_dup1256	3	12	-	126783557	126789584	6028	
L1HS_dup735	3	7	-	49719865	49725896	6032	
L1HS_dup1222	2	12	-	51956416	51962441	6026	
L1HS_dup195	2	2	-	176556194	176559365	3172	
L1HS_dup274	2	3	-	89509976	89516006	6031	Y
L1HS_dup412	2	4	+	80858870	80864900	6031	
L1HS_dup607	2	6	-	2418009	2424037	6029	
L1HS_dup1028	1	X	+	134161589	134162039	451	
L1HS_dup1137	1	11	+	5735918	5738813	2896	
L1HS_dup1252	1	12	+	101539822	101545842	6021	
L1HS_dup1380	1	16	-	33755045	33761079	6035	
L1HS_dup1499	1	20	+	23406746	23412777	6032	Y
L1HS_dup690	1	6	-	133341856	133347885	6030	
L1HS_dup924	1	X	-	11953208	11959433	6226	Y

b.

sample ID	Freq in top 100	tissue
8663	9	STAD
7177	7	HNSC
7261	5	HNSC
AB1X	5	STAD
6962	4	HNSC
A4OF	4	ESCA
A4OG	4	ESCA
6935	3	HNSC
7968	3	STAD
A4OJ	3	ESCA
A6RE	3	ESCA
5721	2	STAD
6943	2	HNSC
8462	2	STAD
A20U	2	BLCA
A43C	2	ESCA
A4G3	2	STAD

Table 6. Enriched annotations among genes negatively correlated with L1HS transcripts in at least two or more tissues.

rank	annotation	terms	Enrichment score	# of genes	Gene names
1	mitochondrial inner membrane, mitochondrial transit peptide	GO:0005743, GO:0005739, transit peptide	38.79	250	*ACAA2, ACAT1, ACOT8, ANXA6, APOPT1, ARAF, ARGLU1, ARL2, ATP5E, ATP5G2, ATP5G3, ATP5H, ATP5J, ATP5L, ATP5O, ATP6V1E1, BAD, BCAT2, BCKDK, BCS1L, BLOC1S1, BNIP1, BOLA1, C12orf10, C14orf119, C14orf2, C19orf12, C19orf70, C21orf33, C6orf136, CARS2, CCT7, CDK5RAP1, CHCHD1, CHCHD2, CHCHD5, CISD3, CLPP, CMC1, CMC4, COA3, COA5, COA6, COMT, COQ10A, COQ8B, COX14, COX17, COX4I1, COX4I2, COX5B, COX6A1, COX6B1, COX6C, COX7A1, COX7A2, COX7B, CYC1, DGUOK, DMAC2, DNAJC19, DNAJC4, DTYMK, ECH1, ECI1, ECSIT, ETFRF1, FAM110B, FAM162A, FASTK, FDX1L, FIS1, FKBP8, FMC1, FUNDC2, G0S2, GADD45GIP1, GCDH, GCK, GFER, GLRX, GLRX5, GNG5, GNPAT, GPX4, HAGH, HARS, HAX1, HDDC2, HEBP2, HIBADH, HIGD2A, HINT2, HMGCL, HTRA2, IMMP2L, ISCA1, ISCA2, ISCU, JTB, LYRM4, MAP1LC3B, MDH1, METTL17, MFF, MIGA2, MINOS1, MOAP1, MPC2, MPDU1, MPG, MPV17L2, MRPL11, MRPL16, MRPL18, MRPL2, MRPL20, MRPL21, MRPL22, MRPL23, MRPL24, MRPL27, MRPL28, MRPL34, MRPL38, MRPL40, MRPL41, MRPL43, MRPL48, MRPL51, MRPL52, MRPL53, MRPL54, MRPL55, MRPL57, MRPS15, MRPS18A, MRPS21, MRPS24, MRPS25, MRPS36, MRPS5, MRPS6, MRPS7, MSRB2, MTCH1, MTERF2, MTFR1L, MTG1, MTIF3, MUTYH, NDUFA1, NDUFA11, NDUFA12, NDUFA13, NDUFA2, NDUFA4, NDUFA5, NDUFA7, NDUFA8, NDUFAF4, NDUFAF5, NDUFB1, NDUFB10, NDUFB11, NDUFB2, NDUFB4, NDUFB5, NDUFB7, NDUFB8, NDUFB9, NDUFC2, NDUFS3, NDUFS4, NDUFS6, NDUFV1, NDUFV2, NFU1, NIPSNAP2, NIPSNAP3A, NTHL1, OGG1, PACS2, PAM16, PARK7, PDK4, PHYH, PLGRKT, PLPBP, PNKD, PNKP, PPOX, PPP3CC, PRDX3, PRDX5, PSMB3, PTPMT1, PTS, RILP, RNF5, ROMO1, RPS14, RSAD1, SCCPDH, SDHAF1, SDHAF2, SDHB, SIVA1, SLC25A11, SLC25A14, SLC25A19, SLC25A26, SLC25A27, SLC25A28, SLC25A38, SLC25A4, SLC25A6, SMDT1, SMIM20, SMIM26, SPG7, SRI, SUCLG1, TAZ, TFB1M, TIMM17B, TIMM44, TIMMDC1, TMEM11, TMEM126A, TMEM126B, TMEM14C, TOMM22, TOMM40L, TOMM5, TOMM6, TOMM7, TRMT1, TSTD1, TUSC2, TXN2, UQCC2, UQCC3, UQCR10, UQCR11, UQCRB, UQCRH, UQCRQ, UROS, USMG5*
2	Oxidative phosphorylation	GO:0005743, hsa00190, hsa05012, hsa05016, hsa05010, GO:0032981, hsa04932, GO:0005747, GO:0006120, GO:0008137, hsa01100	20.35	274	*ACAA2, ACAT1, ACOT8, ADI1, AK1, AKR1B1, AMY2B, AP3S1, AP4B1, APIP, ARF5, ARL3, ASNA1, ASNS, ATF4, ATG3, ATG4B, ATP5E, ATP5G2, ATP5G3, ATP5H, ATP5J, ATP5L, ATP5O, ATP6V0B, ATP6V0E1, ATP6V1E1, ATP6V1F, ATP6V1G1, B3GALT4, BAD, BBS4, BCAT2, BCS1L, BNIP1, BTF3, C19orf70, CD63, CDIPT, CERS1, CHCHD1, CHKB, CHMP2A, CHMP4A, CHMP4B, CHMP5, CHMP6, COA3, COMMD3, COMMD9, COMT, COPE, COQ10A, COX17, COX4I1, COX4I2, COX5B, COX6A1, COX6B1, COX6C, COX7A1, COX7A2, COX7B, CREB3, CYC1, CYGB, CYP2R1, DAD1, DBNL, DCTN2, DDIT3, DGAT1, DGUOK, DNAJC19, DPM3, DTYMK, DYNLRB1, ECI1, ECSIT, FADS3, FDX1L, FXYD1, GABARAP, GABARAPL2, GALM, GAMT, GCDH, GCK, GLRX, GUK1, HIBADH, HIGD2A, HIKESHI, HMGCL, HTRA2, HYI, IFT27, IMMP2L, ISYNA1, LCN12, LPL, MDH1, MINOS1, MITD1, MLX, MPC2, MPDU1, MPV17L2, MRPL11, MRPL16, MRPL18, MRPL2, MRPL20, MRPL21, MRPL22, MRPL23, MRPL24, MRPL27, MRPL28, MRPL34, MRPL38, MRPL40, MRPL41, MRPL43, MRPL48, MRPL51, MRPL52, MRPL53, MRPL54, MRPL55, MRPS15, MRPS18A, MRPS21, MRPS24, MRPS25, MRPS36, MRPS5, MRPS6, MRPS7, MTCH1, MTG1, MVB12A, NDUFA1, NDUFA11, NDUFA12, NDUFA13, NDUFA2, NDUFA4, NDUFA5, NDUFA7, NDUFA8, NDUFAF4, NDUFAF5, NDUFB1, NDUFB10, NDUFB11, NDUFB2, NDUFB4, NDUFB5, NDUFB7, NDUFB8, NDUFB9, NDUFC2, NDUFS3, NDUFS4, NDUFS6, NDUFV1, NDUFV2, NME2, NTPCR, NUP85, OAZ1, ORAI1, P2RX4, PAFAH1B3, PAM16, PARK7, PDK4, PGP, PIGP, PIP5KL1, PLA2G16, PLLP, PNMT, POLE4, POLR2F, POLR2G, POLR2H, POLR2I, POLR2J, POLR3GL, PPCDC, PPOX, PPP3CC, PRDX6, PRKAG2, PSENEN, PTPMT1, PTS, RABGEF1, RAMP1, RAMP2, RANGRF, RILP, RINL, ROMO1, S100A13, SAT2, SDHB, SELENOK, SERP2, SERPINA5, SFT2D1, SLC14A1, SLC25A11, SLC25A14, SLC25A19, SLC25A26, SLC25A27, SLC25A28, SLC25A38, SLC25A4, SLC25A6, SLC29A1, SLC2A8, SLC30A2, SLC39A13, SLC41A2, SLC50A1, SMDT1, SMIM20, SNF8, SNX17, SNX21, SNX3, ST3GAL3, STARD10, STX10, SUCLG1, TAZ, THTPA, TIMM17B, TIMM44, TIMMDC1, TMED1, TMEM11, TMEM126A, TMEM126B, TMEM38B, TMEM9, TOMM22, TOMM40L, TOMM5, TOMM6, TOMM7, TRAPPC1, TRAPPC6A, TXN2, UBE2G2, UNC50, UQCC2, UQCC3, UQCR10, UQCR11, UQCRB, UQCRH, UQCRQ, UROD, UROS, USE1, VAMP8, VPS28, VPS4A, VTI1B*
3	Ribonucleoprotein, Ribosomal protein Mitochondrial ribosome, translation	GO:0003735, GO:0006412, hsa03010, GO:0005840, GO:0070125, GO:0070126, GO:0006614, GO:0006413, GO:0005762, GO:0019083, GO:0006364, GO:0000184, GO:0005761, GO:0022625, GO:0022627, GO:0032543, GO:0003723, GO:0044822, GO:0005763, GO:0015935, GO:0005925	15.86	180	*ACAA2, ANXA6, ARL6IP4, BTF3, BUD23, C11orf68, C19orf66, C1D, CD81, CELF6, CHCHD1, COA6, CWC15, DDX49, DHPS, DNAJC17, EIF1B, EIF2B1, EIF2B5, EIF2D, EIF3G, EIF3I, EIF3K, EMG1, ERH, EXOSC1, EXOSC8, FAM103A1, FAM50A, FASTK, FAU, FKBP3, FLOT1, FRG1, GADD45GIP1, GTF3A, HARS, HOXB6, ILK, ITGA2B, ITGB1BP1, KIF22, LGALS1, LSM1, LSM10, LSM2, LSM4, MCTS1, METTL17, MPV17L2, MRPL11, MRPL16, MRPL18, MRPL2, MRPL20, MRPL21, MRPL22, MRPL23, MRPL24, MRPL27, MRPL28, MRPL34, MRPL38, MRPL40, MRPL41, MRPL43, MRPL48, MRPL51, MRPL52, MRPL53, MRPL54, MRPL55, MRPL57, MRPS15, MRPS18A, MRPS21, MRPS24, MRPS25, MRPS36, MRPS5, MRPS6, MRPS7, MT3, MTG1, NDUFA7, NELFE, NHP2, NME2, NOL12, NOP10, NOSIP, NSUN5, NTPCR, NUP85, PABPC1L, PARVB, PCBP4, PCID2, PIH1D1, PNISR, POP4, PPIA, PPIE, PRPF40B, PSMA1, PSMD4, RBBP7, RNASEH2A, RPL13AP3, RPL14, RPL15,*

					RPL18, RPL18A, RPL19, RPL24, RPL27A, RPL29, RPL3, RPL32, RPL35, RPL36, RPL37A, RPL38, RPL41, RPL5, RPL7A, RPL8, RPP21, RPP25L, RPP30, RPS10, RPS11, RPS13, RPS14, RPS16, RPS19BP1, RPS21, RPS27L, RPS5, RPS6, RPS9, RPSA, RRAS, RSL24D1, RTN4, RTRAF, SF3B6, SLC25A11, SLC25A14, SLC25A19, SLC25A26, SLC25A27, SLC25A28, SLC25A38, SLC25A4, SLC25A6, SNRPB, SNRPC, SNRPD2, SNRPGP15, SNRPN, SNU13, SRP14, SUCLG1, SUGP1, SUMO1, SUMO2, TFB1M, TIA1, TPD52L2, TPT1, TRMT1, TRNAU1AP, TSPAN4, TWF2, U2AF1, U2AF1L4, UNC50, YBX1, ZCRB1
4~8 more annotations associated with mitochondria.					
9	proteasome complex	GO:0006521, GO:0000502, GO:0051436, hsa03050, GO:0051437, GO:0051603, GO:0031145, IPR001353, GO:0043161, GO:0038061, GO:0005839, GO:0004298, GO:0043488, GO:0002479, GO:0060071, GO:0090090, IPR023333, GO:0033209, IPR016050, GO:0002223, GO:0090263, GO:0000209, GO:0016032, GO:0050852, GO:0000165, GO:0038095, GO:0006511	2.533	81	*ABTB1, ACOT8, AES, AMZ2, ANAPC11, ANAPC15, ANAPC16, ARAF, ATG4B, BTBD6, CAPN12, CDC34, CLPP, COMMD7, COPS6, CREB3, CTDNEP1, CTSF, CTSL, CTSZ, DDIT3, DOK5, DPP7, EXOSC1, EXOSC8, FBXW5, FKBP8, GADD45GIP1, GFRA2, GPC4, GTF2A2, HTRA2, ICAM2, ILK, IMMP2L, KAT5, LAMTOR2, NEDD8, NPEPL1, NUP85, OAZ1, OTUB1, PACS2, PARD6A, PARK7, PCBP4, PCID2, PFDN5, POMP, PPP4C, PSMA1, PSMA2, PSMA7, PSMB1, PSMB3, PSMB4, PSMB6, PSMC3, PSMD13, PSMD4, PSMD9, PSMG3, RAD23A, RBX1, RNF166, R NF167, RTRAF, SDHAF2, SEC11A, SERPINA5, SLC25A4, SNAPIN, SPG7, SQSTM1, STUB1, SUMO1, TMEM88, TNFSF12, UBE2G2, UBXN1, WDR83*
11	Autophagy	GO:0016236, GO:0000422, GO:0000045	2.21	23	*ATG3, ATG4B, BNIP1, FIS1, FUNDC2, GABARAP, GABARAPL2, LAMTOR1, LAMTOR2, LAMTOR4, LAMTOR5, MAP1LC3B, PACS2, PARK7, PRKAG2, SQSTM1, TMEM208, TOMM22, TOMM5, TOMM6, TOMM7, VAMP8, WDR45*
12	ESCRT complex, late endosome membrane	GO:0039702, GO:0036258, GO:0000920, GO:0031902, GO:0006914, GO:0016197, GO:0019058, GO:0000815, IPR005024, GO:0006997, GO:0007080, GO:0007034, GO:0030496, GO:0010008, hsa04144, GO:0005768	2.21	57	*ACKR4, ANXA6, ARF5, ARL3, ATG4B, ATP6V0B, ATP6V0E1, ATP6V1E1, BLOC1S1, BNIP1, C19orf12, CD320, CD63, CHMP2A, CHMP4A, CHMP4B, CHMP5, CHMP6, DBNL, DCTN3, DPY30, ERH, FLOT1, GABARAP, GABARAPL2, JTB, KIF22, LAMTOR1, LAMTOR2, MAP1LC3B, MITD1, MVB12A, PARD6A, PARK7, PDCD6, PIP5KL1, PKN1, PPIA, PTP4A2, RABGEF1, RILP, SCCPDH, SH3GLB2, SNF8, SNX17, SNX21, SNX3, SQSTM1, TMEM208, TMEM230, TMEM9, VAMP8, VPS28, VPS4A, VTI1B, WDR45, WDR83*
13	Thioredoxin, redox-active	IPR012336, GO:0098869, IPR013766, GO:0045454, GO:0042744, GO:0000302, GO:0005623	2.15	27	*CYGB, DDIT3, GCK, GLRX, GLRX5, GPX4, GSTM2, GSTO1, MGST3, MIEN1, MRPL43, MRPS25, NDUFA2, NDUFV2, PARK7, PRDX2, PRDX3, PRDX5, PRDX6, SELENOM, SELENOT, SELENOW, TXN2, TXNDC11, TXNDC15, TXNL4A, VKORC1*

Table 7. genes showing highly recurrent negative correlation with L1HS based on REC score.

rank	gene	REC	rank	gene	REC
1	*ZNF511*	-5.428	11	*PARK7*	-4.242
2	*DHPS*	-4.855	12	*PKIG*	-4.207
3	*FLOT1*	-4.802	13	*ZNF32*	-4.052
4	*CCDC107*	-4.752	14	*COPS6*	-3.942
5	*AIP*	-4.731	15	*R3HCC1*	-3.884
6	*FEZ2*	-4.674	16	*PPP2R3C*	-3.841
7	*PARD6A*	-4.607	17	*RNASEH2C*	-3.735
8	*CREB3*	-4.508	18	*ECSIT*	-3.700
9	*C19orf42*	-4.361	19	*CUEDC2*	-3.650
10	*ANAPC16*	-4.321	20	*ZNF174*	-3.647

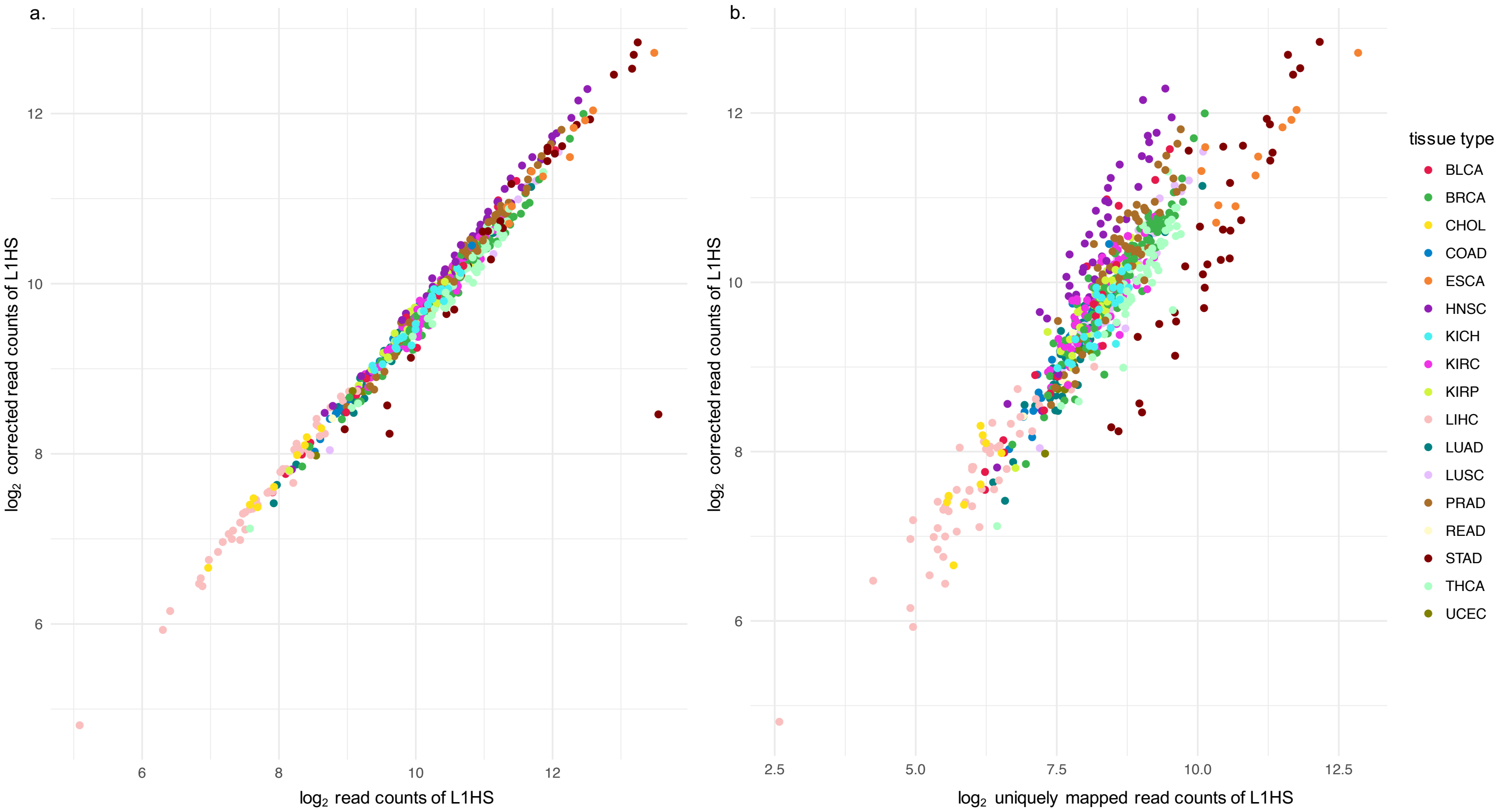
Table 8. Enriched annotations among genes positively correlated with L1HS transcripts in at least two or more tissues.

rank	annotation	terms	Enrichment score	# of genes	Gene names
1	kinase	nucleotide phosphate-binding region, GO:0005524, GO:0004674, IPR011009, IPR000719, IPR008271, GO:0004672, GO:0006468, IPR017441, SM00220, GO:0046777	9.15	293	
2	SH3 domain	IPR001452, SM00326	8.05	44	*ABI1, ABI2, ABL1, ARHGAP32, ARHGAP42, ARHGEF38, ARHGEF5, BAIAP2, CASK, CD2AP, CRKL, CSK, CTTN, DBNL, DLG1, DLG5, DOCK5, DST, EPS8L1, FNBP1L, ITSN2, MACC1, MACF1, MAP3K9, MPP6, MYO1E, NEBL, PPP1R13B, PPP1R13L, PTK6, RASA1, SASH1, SH3GL1, SH3PXD2A, SH3RF1, SH3RF2, SNX33, SPATA13, SRGAP1, SRGAP2, TJP1, TRIO, VAV3, YES1*
3	DNA damage, DNA repair	GO:0006974, GO:0006281	7.07	80	*ABL1, APC, ASCC3, ATAD5, ATF2, ATMIN, ATR, ATRX, BACH1, BAZ1B, BOD1L1, BRCA2, BRIP1, CHD2, CLOCK, DTX3L, E2F7, EMSY, EPC2, ERCC4, ERCC6, ERCC6L2, EYA3, FANCD2, FANCE, FANCI, FBXO45, FMR1, FNIP2, FOXO1, FTO, GTF2H3, HIPK2, HUWE1, INO80D, INTS7, MAPK1, MMS22L, NCOA6, NIPBL, PALB2, PAXIP1, PDS5A, POLQ, PRKDC, RAD50, RBBP5, RECQL4, REV3L, RFWD3, RIF1, RNF168, RNF169, SETD7, SETX, SMARCAD1, SMC3, SMC6, STXBP4, SUPT16H, SUSD6, TAF1, TAOK1, TICRR, TOPBP1, TP53BP1, TP63, TP73, TRAF6, TRIP12, TRRAP, UBR5, UHRF1, USP7, VAV3, WRN, XRCC5, YAP1, ZBTB38, ZRANB3*
4	transcription	GO:0005634, GO:0006351, IPR013083, GO:0003677, GO:0006355, GO:0045893, GO:0003700, GO:0008270, GO:0045944, GO:0003713, zinc finger region, IPR007087, IPR015880, IPR013087, GO:0046872, GO:0003676, SM00355, IPR001909, SM00349	6.12	768	
5	bromodomain	IPR019787, IPR001965, IPR001487, SM00249, IPR018359, SM00297, zinc finger region, IPR018501, SM00571, GO:0070577	5.29	37	*ASH1L, ATAD2B, ATRX, BAZ1B, BAZ2A, BAZ2B, BPTF, BRPF3, BRWD3, CECR2, CHD4, CREBBP, EP300, JADE3, KAT6A, KAT6B, KDM5A, KDM5B, KDM5C, KIAA2026, KMT2A, KMT2C, KMT2D, KMT2E, NSD1, NSD3, PBRM1, PHF12, PHF14, PHF3, PYGO1, RSF1, TAF1, TAF1L, TCF20, TRIM33, UHRF1*
6	PHD finger and SET domain	IPR011011, IPR019787, IPR001965, IPR013083, SM00249, zinc finger region, IPR019786, IPR001214, SM00317, GO:0018024, IPR003616, GO:0035097, hsa00310, GO:0051568, IPR006560, SM00508, SM00570, GO:0042800, GO:0008168	4.71	117	*AEBP2, ALDH3A2, ANKIB1, ARID1A, ARID1B, ARID2, ASH1L, ATRX, ATXN7L3, BAZ1B, BAZ2A, BAZ2B, BPTF, BRPF3, BSN, CARNMT1, CBLC, CBX6, CECR2, CHD2, CHD4, CHD6, CHD7, CHD8, CHD9, CREBBP, DCAF1, DTX3L, EEA1, EMSY, EP400, EPC2, EXPH5, EYA3, EZH2, FGD6, FOXA1, FYCO1, JADE3, JARID2, KANSL1, KAT6A, KAT6B, KDM5A, KDM5B, KDM5C, KMT2A, KMT2C, KMT2D, KMT2E, KMT5B, LTN1, MAP3K1, MARCH6, MARCH8, MID2, MTMR3, MTR, MYCBP2, NCOA6, NCOR1, NSD1, NSD3, PAXIP1, PBRM1, PCLO, PHF12, PHF14, PHF3, PIAS2, PIKFYVE, PRDM10, PRDM4, PYGO1, RBBP5, RC3H1, RC3H2, RCOR1, RFFL, RFWD3, RMND5A, RNF111, RNF145, RNF168, RNF169, RNF39, RNF6, RNMT, RSF1, RSPRY1, SCAF11, SETD2, SETD5, SETD7, SFMBT1, SH3RF1, SH3RF2, SMARCA5, SMARCAD1, SMARCC1, SUDS3, SYTL1, TCF20, TET2, TET3, TRAF6, TRIM33, TRIM56, TRRAP, UBE4B, UBN1, UBR1, UHRF1, USP22, WDFY2, WDFY3, ZMIZ1*
7	chromodomain, helicase	IPR000330, IPR014001, IPR001650, SM00487, SM00490, GO:0016569, IPR027417, IPR016197, IPR023780, GO:0004386, IPR000953, GO:0008026, GO:0032508, SM00298, IPR006576, SM00592, IPR002464, IPR011545, IPR023779	4.55	109	*AEBP2, AGAP1, AQR, ARHGAP5, ARID1A, ARID1B, ARID2, ARID4B, ARL4C, ASCC3, ASPM, ATAD2B, ATAD5, ATL3, ATRX, BMS1, BPTF, BRIP1, BTAF1, CARD14, CASK, CBX6, CDYL2, CECR2, CHD2, CHD4, CHD6, CHD7, CHD8, CHD9, DDX21, DHX32, DHX33, DHX8, DICER1, DLG1, DLG5, DYNC1H1, EMSY, EP400, ERCC6, ERCC6L2, FOXA1, GNA11, GNA15, GSPT1, HELZ, IQGAP1, JARID2, KIF11, KIF13A, KIF18B, KIF1B, KIF20A, KIF21A, KIF27, KIF3A, KIF3B, KIF5C, LONP2, MAGI3, MCM4, MDN1, MFHAS1, MPHOSPH8, MPP6, MYO1B, MYO1D, MYO1E, MYO5A, MYO5B, MYO6, MYO9A, N4BP2, NCOR1, NRAS, PBRM1, POLQ, RAB10, RAB14, RAB27B, RAB3B, RAD50, RAD54L2, RASEF, RECQL4, RSF1, SBNO1, SEPT3, SETX, SFMBT1, SMARCA5, SMARCAD1, SMARCC1, SMC3, SMC4, SMC6, SNRNP200, TANC1, TANC2, TJP1, TTF2, UBN1, VPS4B, WRN, XRCC5, YLPM1, YTHDC2, ZRANB3*

Table 9. genes showing highly recurrent positive correlation with L1HS based on REC score.

rank	gene	REC	rank	gene	REC
1	*CSNK1G1*	-6.136	11	*RSC1A1*	-3.971
2	*DIP2B*	-5.235	12	*SPTBN2*	-3.913
3	*DOCK5*	-4.968	13	*MAP3K9*	-3.908
4	*ARHGAP32*	-4.966	14	*TRIP12*	-3.778
5	*ZNF462*	-4.624	15	*RAB27B*	-3.774
6	*PAK6*	-4.530	16	*PPM1L*	-3.748
7	*EYA3*	-4.530	17	*KLF16*	-3.641
8	*ZNF827*	-4.466	18	*FAM115A*	-3.618
9	*KIAA1244*	-4.050	19	*CARD14*	-3.568
10	*GTF3C4*	-4.011	20	*SON*	-3.550

Figure 1


a.
b.
log2 corrected read counts of L1HS
log2 read counts of L1HS
log2 uniquely mapped read counts of L1HS
6
8
10
12
2.5
5.0
7.5
10.0
12.5
tissue type
BLCA
BRCA
CHOL
COAD
ESCA
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
READ
STAD
THCA
UCEC

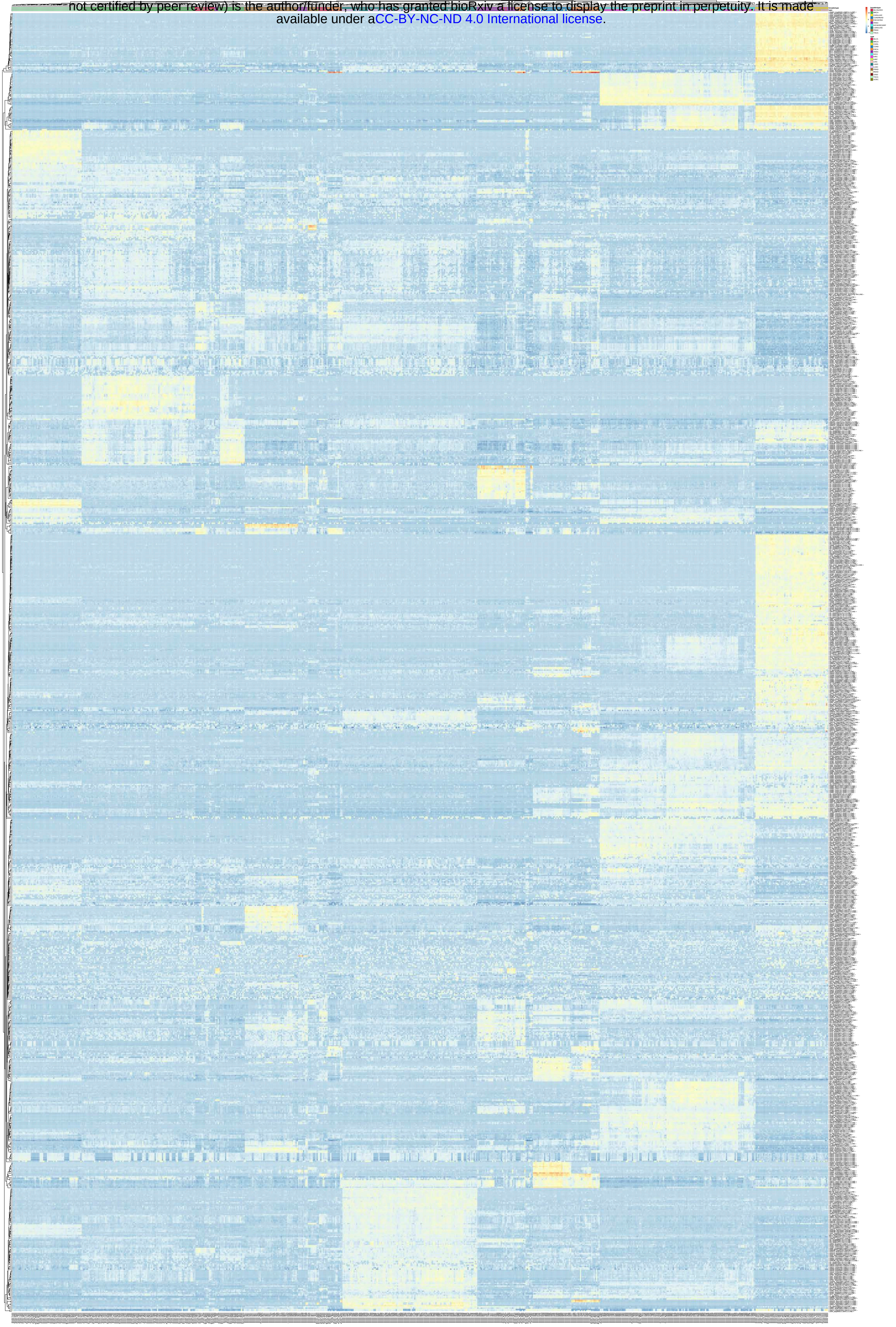
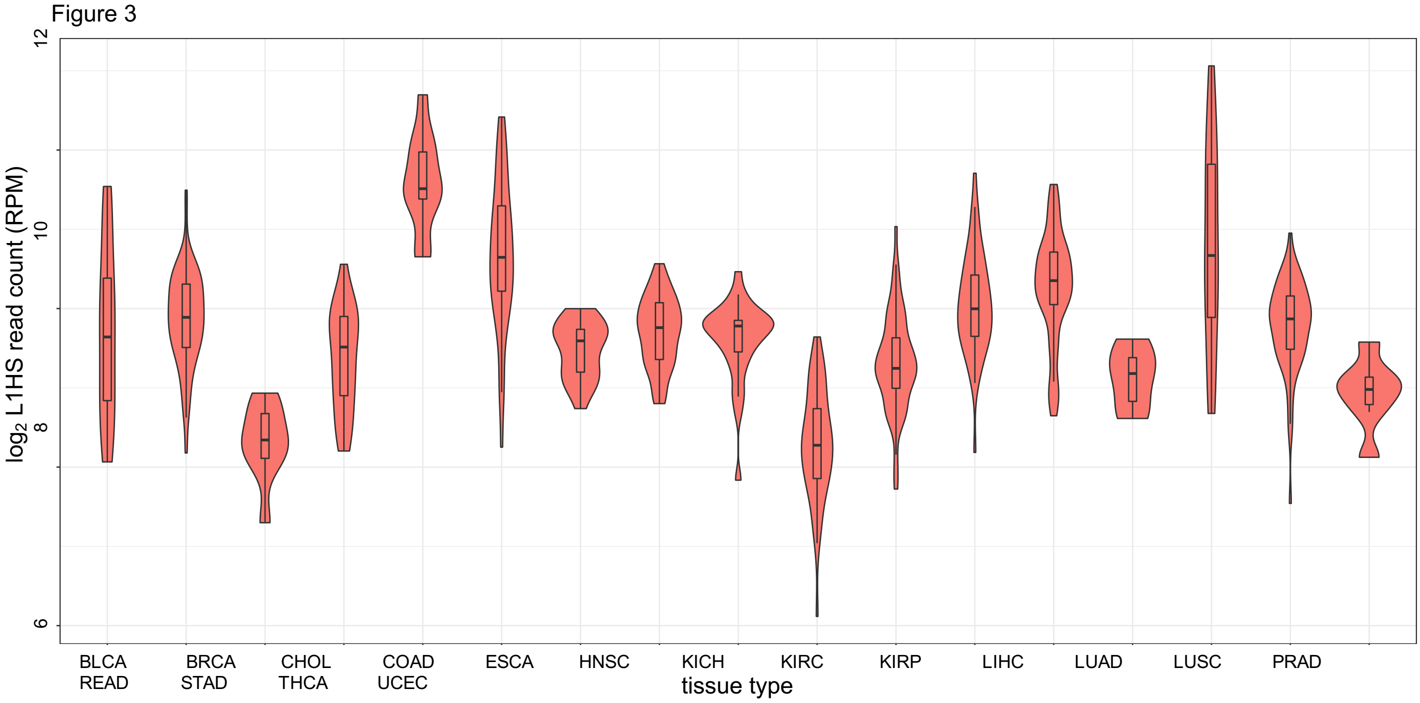

10
5
0
−5
broadertype
BLCAUCEC
BRCA
CHOLLIHC
COADREAD
ESCASTAD
HNSC
KICHKIRCKIRP
LUADLUSC
PRAD
THCA
type
BLCA
BRCA
CHOL
COAD
ESCA
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
READ
STAD
THCA
UCEC

Figure 3



log$_2$ L1HS read count (RPM)
12
10
8
6
BLCA
READ
BRCA
STAD
CHOL
THCA
COAD
UCEC
ESCA
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
tissue type

Figure 4

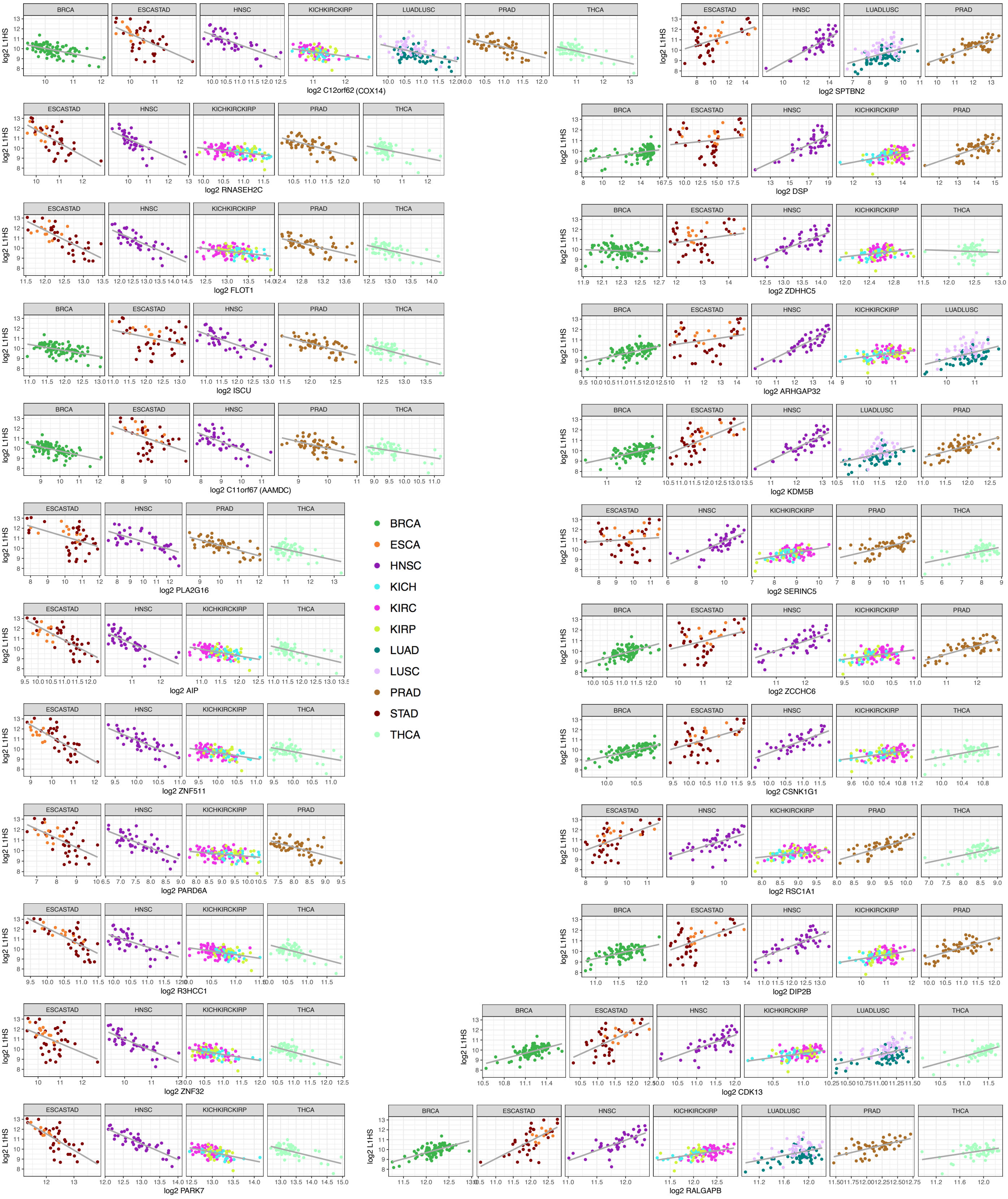
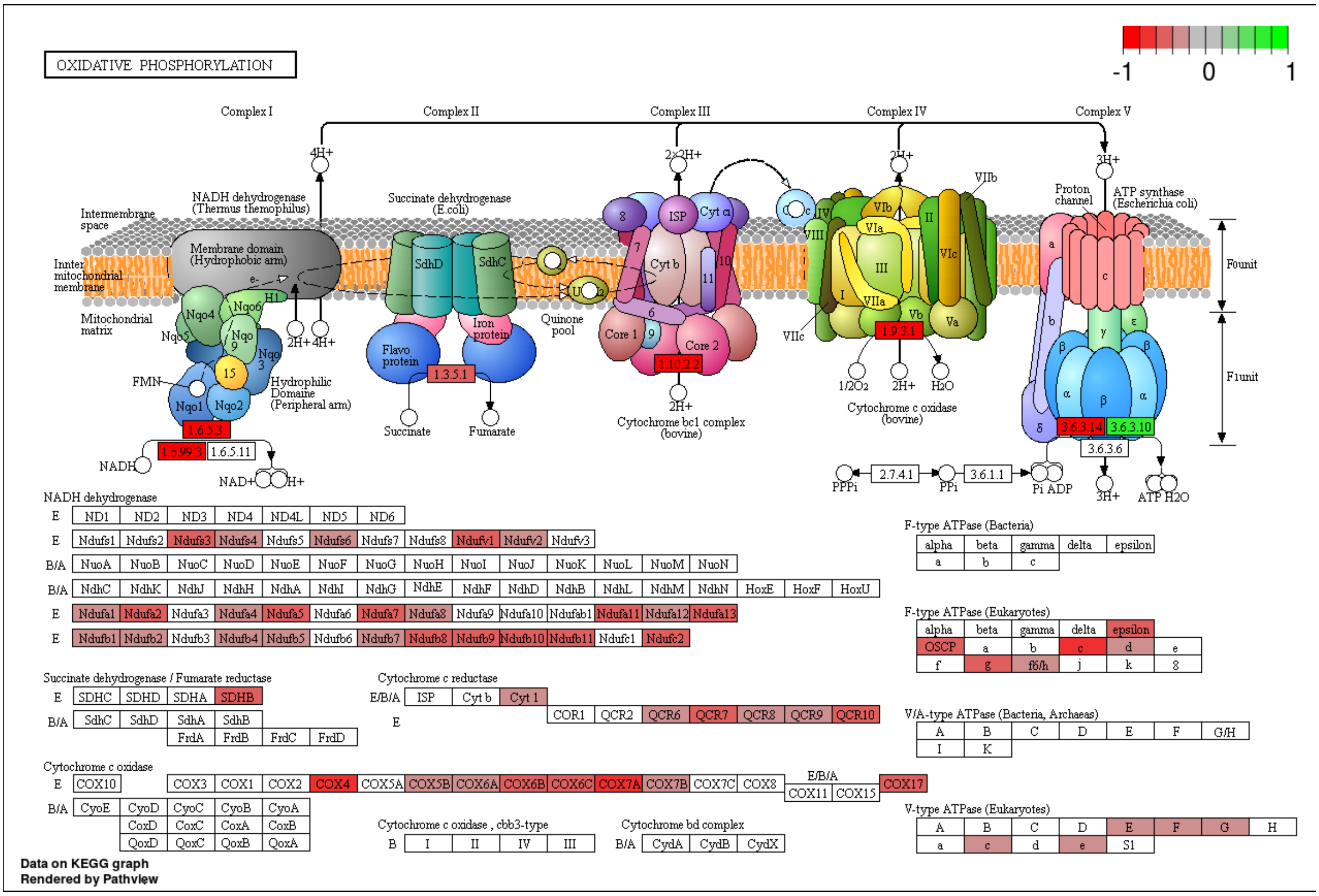

BRCA
ESCASTAD
HNSC
KICHKIRCKIRP
LUADLUSC
PRAD
THCA
log2 L1HS
log2 C12orf62 (COX14)
log2 SPTBN2
log2 RNASEH2C
log2 DSP
log2 FLOT1
log2 ZDHHC5
log2 ISCU
log2 ARHGAP32
log2 C11orf67 (AAMDC)
log2 KDM5B
log2 PLA2G16
log2 SERINC5
log2 AIP
log2 ZCCHC6
log2 ZNF511
log2 CSNK1G1
log2 PARD6A
log2 RSC1A1
log2 R3HCC1
log2 DIP2B
log2 ZNF32
log2 CDK13
log2 PARK7
log2 RALGAPB
● BRCA
● ESCA
● HNSC
● KICH
● KIRC
● KIRP
● LUAD
● LUSC
● PRAD
● STAD
● THCA

Figure 5

a.


OXIDATIVE PHOSPHORYLATION
-1
0
1
Complex I
Complex II
Complex III
Complex IV
Complex V
4H+
2×2H+
2H+
3H+
NADH dehydrogenase (Thermus themophilus)
Succinate dehydrogenase (E.coli)
Intermembrane space
Innter mitochondrial membrane
Mitochondrial matrix
Membrane domain (Hydrophobic arm)
Hydrophilic Domaine (Peripheral arm)
FMN
Nqo1
Nqo2
Nqo4
Nqo5
Nqo6
Nqo9
Nqo3
15
H1
e-
2H+ 4H+
1.6.5.3
1.6.99.3
1.6.5.11
NADH
NAD+
H+
SdhD
SdhC
Flavo protein
Iron protein
1.3.5.1
Succinate
Fumarate
Quinone pool
Q
UQ
ISP
Cyt c1
Cyt b
Core 1
Core 2
1.10.2.2
Cytochrome bc1 complex (bovine)
Cyt c
VIIb
VIa
III
VIc
VIIa
Vb
Va
VIIc
VIII
IV
II
I
1.9.3.1
1/2O2
2H+
H2O
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
F0unit
F1unit
3.6.3.14
3.6.3.10
3.6.3.6
2.7.4.1
3.6.1.1
PPPi
PPi
Pi ADP
3H+
ATP H2O
NADH dehydrogenase
E ND1 ND2 ND3 ND4 ND4L ND5 ND6
E Ndufs1 Ndufs2 Ndufs3 Ndufs4 Ndufs5 Ndufs6 Ndufs7 Ndufs8 Ndufv1 Ndufv2 Ndufv3
B/A NuoA NuoB NuoC NuoD NuoE NuoF NuoG NuoH NuoI NuoJ NuoK NuoL NuoM NuoN
B/A NdhC NdhK NdhJ NdhH NdhA NdhI NdhG NdhE NdhF NdhD NdhB NdhL NdhM NdhN HoxE HoxF HoxU
E Ndufa1 Ndufa2 Ndufa3 Ndufa4 Ndufa5 Ndufa6 Ndufa7 Ndufa8 Ndufa9 Ndufa10 Ndufab1 Ndufa11 Ndufa12 Ndufa13
E Ndufb1 Ndufb2 Ndufb3 Ndufb4 Ndufb5 Ndufb6 Ndufb7 Ndufb8 Ndufb9 Ndufb10 Ndufb11 Ndufc1 Ndufc2
Succinate dehydrogenase / Fumarate reductase
E SDHC SDHD SDHA SDHB
B/A SdhC SdhD SdhA SdhB
FrdA FrdB FrdC FrdD
Cytochrome c reductase
E/B/A ISP Cyt b Cyt 1
E COR1 QCR2 QCR6 QCR7 QCR8 QCR9 QCR10
Cytochrome c oxidase
E COX10 COX3 COX1 COX2 COX4 COX5A COX5B COX6A COX6B COX6C COX7A COX7B COX7C COX8 COX11 COX15 COX17
E/B/A
B/A CyoE CyoD CyoC CyoB CyoA
CoxD CoxC CoxA CoxB
QoxD QoxC QoxB QoxA
Cytochrome c oxidase , cbb3-type
B I II IV III
Cytochrome bd complex
B/A CydA CydB CydX
F-type ATPase (Bacteria)
alpha beta gamma delta epsilon
a b c
F-type ATPase (Eukaryotes)
alpha beta gamma delta epsilon
OSCP a b c d e
f g f6/h j k 8
V/A-type ATPase (Bacteria, Archaeas)
A B C D E F G/H
I K
V-type ATPase (Eukaryotes)
A B C D E F G H
a c d e S1
Data on KEGG graph
Rendered by Pathview

b.

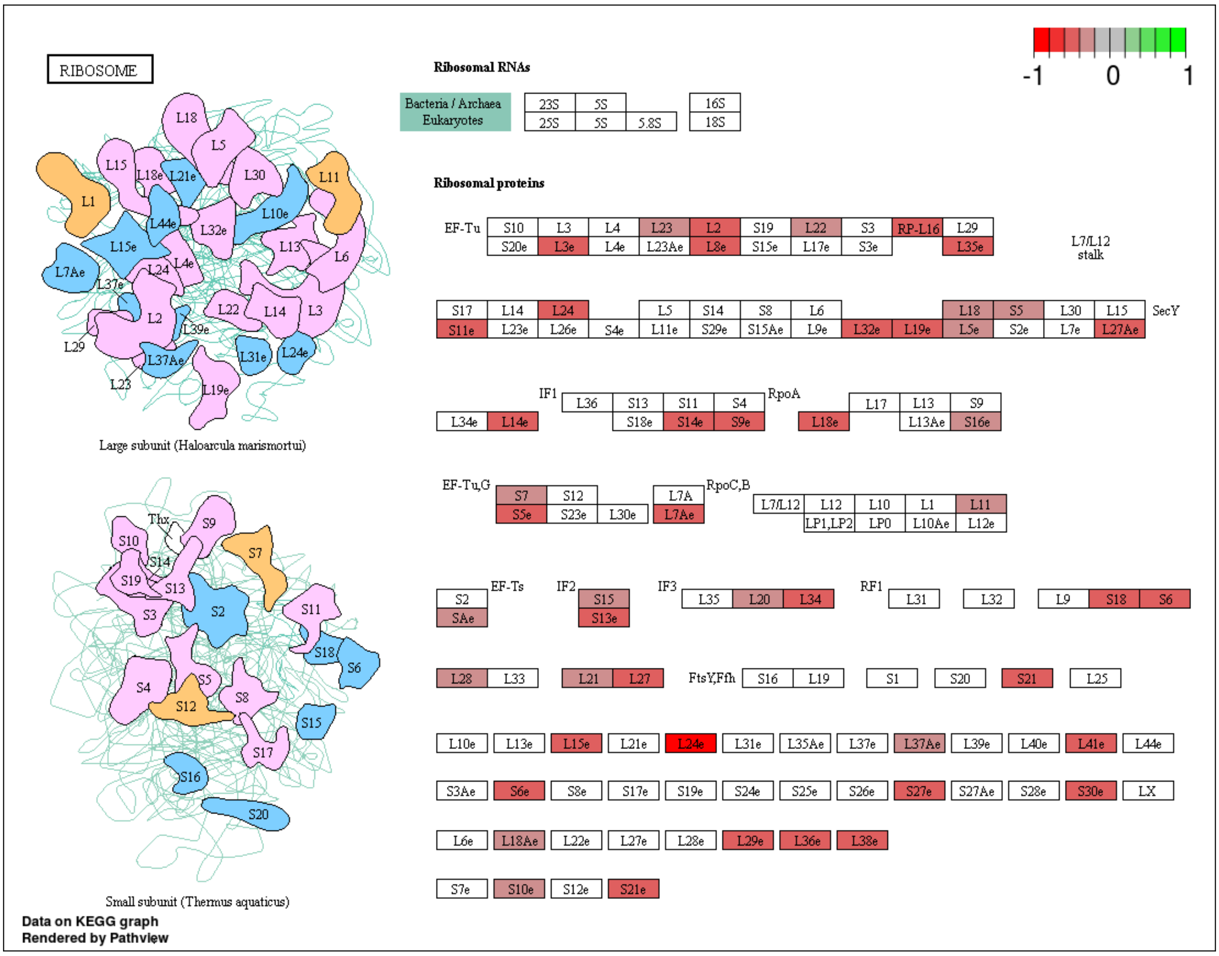

RIBOSOME
-1
0
1
Ribosomal RNAs
Bacteria / Archaea
Eukaryotes
23S 5S 16S
25S 5S 5.8S 18S
Ribosomal proteins
EF-Tu
S10 L3 L4 L23 L2 S19 L22 S3 RP-L16 L29
S20e L3e L4e L23Ae L8e S15e L17e S3e L35e
L7/L12 stalk
S17 L14 L24 L5 S14 S8 L6 L18 S5 L30 L15 SecY
S11e L23e L26e S4e L11e S29e S15Ae L9e L32e L19e L5e S2e L7e L27Ae
IF1
RpoA
L36 S13 S11 S4 L17 L13 S9
L34e L14e S18e S14e S9e L18e L13Ae S16e
EF-Tu,G
RpoC,B
S7 S12 L7A
S5e S23e L30e L7Ae
L7/L12 L12 L10 L1 L11
LP1,LP2 LP0 L10Ae L12e
S2 EF-Ts IF2 S15 IF3 L35 L20 L34 RF1 L31 L32 L9 S18 S6
SAe S13e
L28 L33 L21 L27 FtsY,Ffh S16 L19 S1 S20 S21 L25
L10e L13e L15e L21e L24e L31e L35Ae L37e L37Ae L39e L40e L41e L44e
S3Ae S6e S8e S17e S19e S24e S25e S26e S27e S27Ae S28e S30e LX
L6e L18Ae L22e L27e L28e L29e L36e L38e
S7e S10e S12e S21e
L18 L5 L15 L18e L21e L30 L11 L1 L10e L44e L32e L15e L13 L6 L7Ae L4e L24 L37e L2 L22 L14 L3 L39e L29 L37Ae L31e L24e L23 L19e
Large subunit (Haloarcula marismortui)
Thx S9 S10 S14 S7 S19 S13 S2 S3 S11 S18 S6 S4 S5 S12 S8 S15 S17 S16 S20
Small subunit (Thermus aquaticus)
Data on KEGG graph
Rendered by Pathview

Figure 6

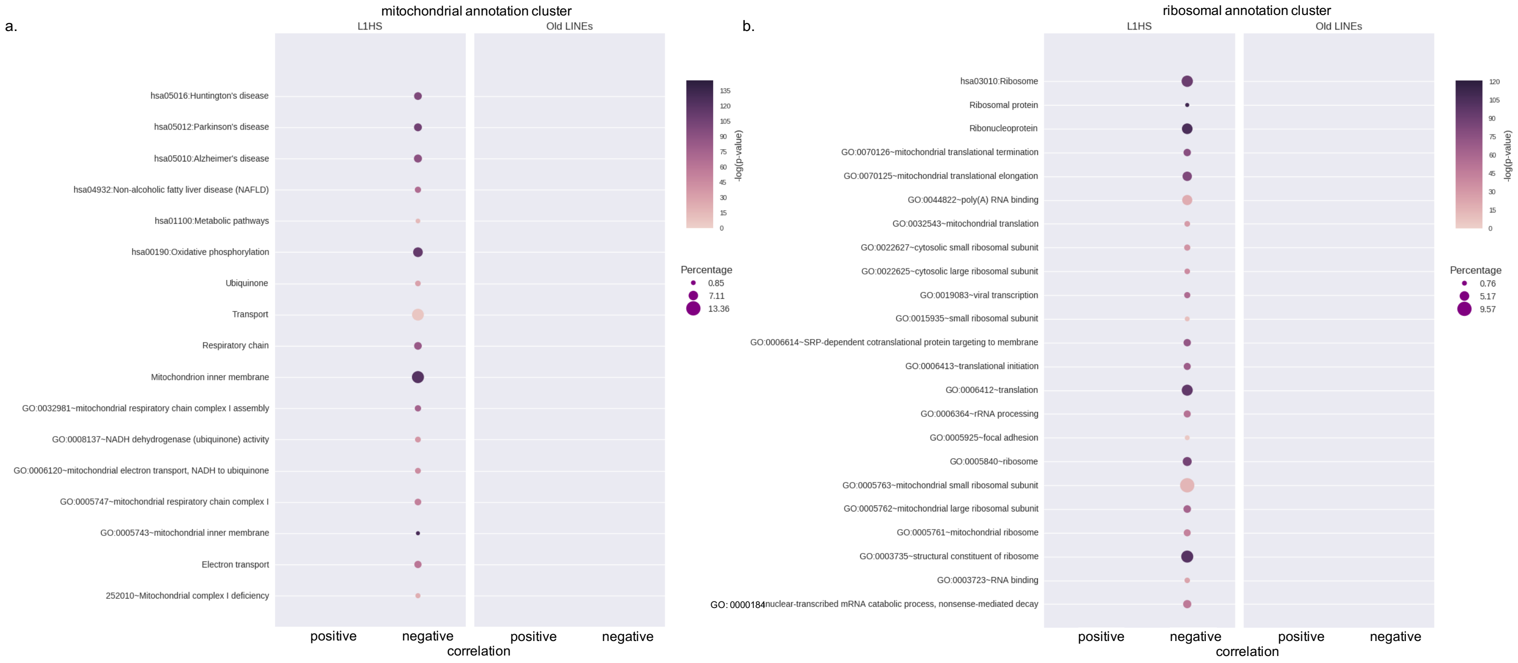

a.
mitochondrial annotation cluster
L1HS
Old LINEs
hsa05016:Huntington's disease
hsa05012:Parkinson's disease
hsa05010:Alzheimer's disease
hsa04932:Non-alcoholic fatty liver disease (NAFLD)
hsa01100:Metabolic pathways
hsa00190:Oxidative phosphorylation
Ubiquinone
Transport
Respiratory chain
Mitochondrion inner membrane
GO:0032981~mitochondrial respiratory chain complex I assembly
GO:0008137~NADH dehydrogenase (ubiquinone) activity
GO:0006120~mitochondrial electron transport, NADH to ubiquinone
GO:0005747~mitochondrial respiratory chain complex I
GO:0005743~mitochondrial inner membrane
Electron transport
252010~Mitochondrial complex I deficiency
positive
negative
positive
negative
correlation
-log(p-value)
135
120
105
90
75
60
45
30
15
0
Percentage
0.85
7.11
13.36
b.
ribosomal annotation cluster
L1HS
Old LINEs
hsa03010:Ribosome
Ribosomal protein
Ribonucleoprotein
GO:0070126~mitochondrial translational termination
GO:0070125~mitochondrial translational elongation
GO:0044822~poly(A) RNA binding
GO:0032543~mitochondrial translation
GO:0022627~cytosolic small ribosomal subunit
GO:0022625~cytosolic large ribosomal subunit
GO:0019083~viral transcription
GO:0015935~small ribosomal subunit
GO:0006614~SRP-dependent cotranslational protein targeting to membrane
GO:0006413~translational initiation
GO:0006412~translation
GO:0006364~rRNA processing
GO:0005925~focal adhesion
GO:0005840~ribosome
GO:0005763~mitochondrial small ribosomal subunit
GO:0005762~mitochondrial large ribosomal subunit
GO:0005761~mitochondrial ribosome
GO:0003735~structural constituent of ribosome
GO:0003723~RNA binding
GO: 0000184nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
positive
negative
positive
negative
correlation
-log(p-value)
120
105
90
75
60
45
30
15
0
Percentage
0.76
5.17
9.57

Figure 7.

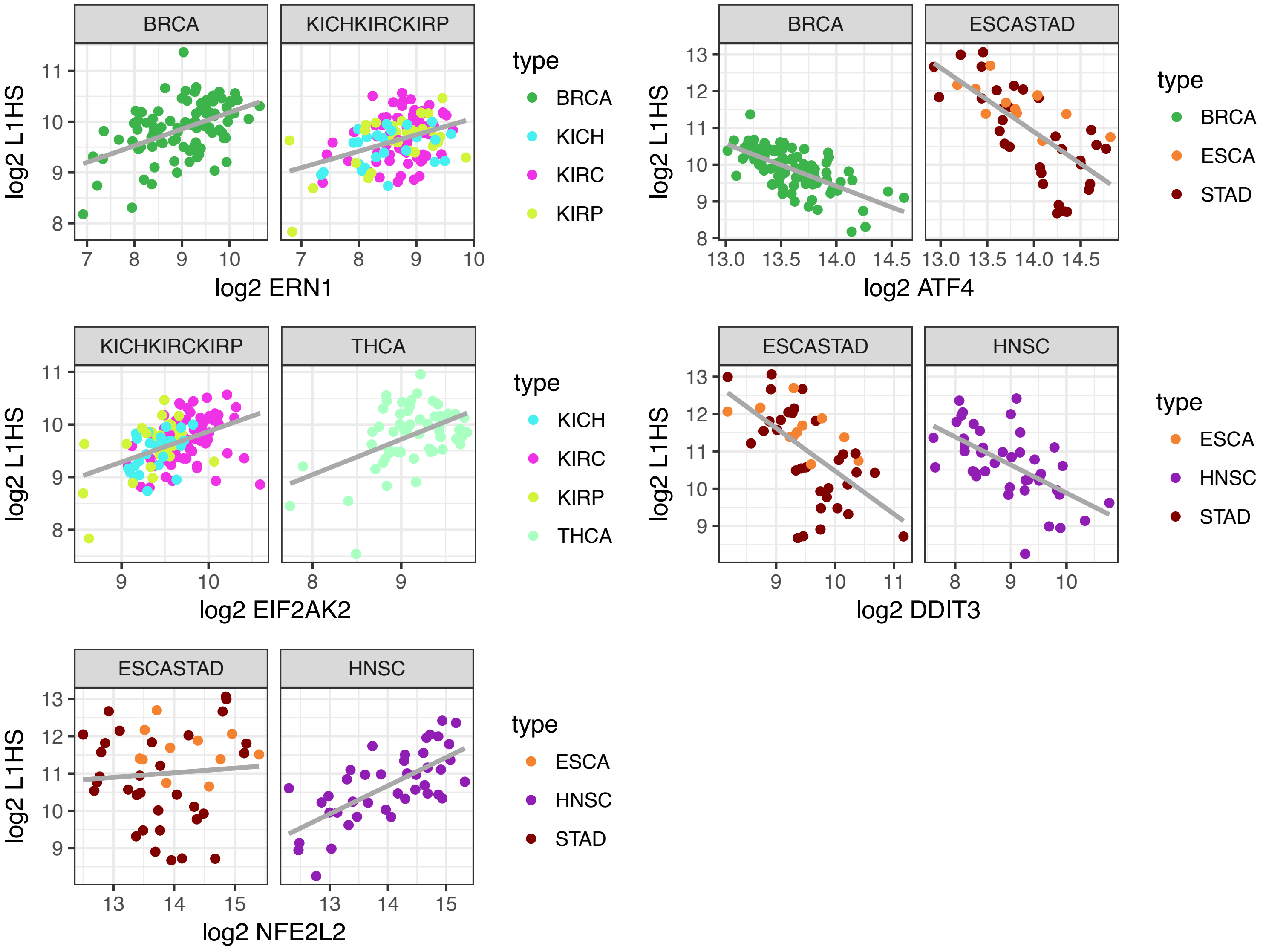

BRCA
KICHKIRCKIRP
log2 L1HS
11
10
9
8
7
8
9
10
log2 ERN1
type
BRCA
KICH
KIRC
KIRP
BRCA
ESCASTAD
13
12
11
10
9
8
13.0
13.5
14.0
14.5
log2 ATF4
type
BRCA
ESCA
STAD
KICHKIRCKIRP
THCA
log2 EIF2AK2
type
KICH
KIRC
KIRP
THCA
ESCASTAD
HNSC
log2 DDIT3
type
ESCA
HNSC
STAD
ESCASTAD
HNSC
log2 NFE2L2
type
ESCA
HNSC
STAD

Figure 8

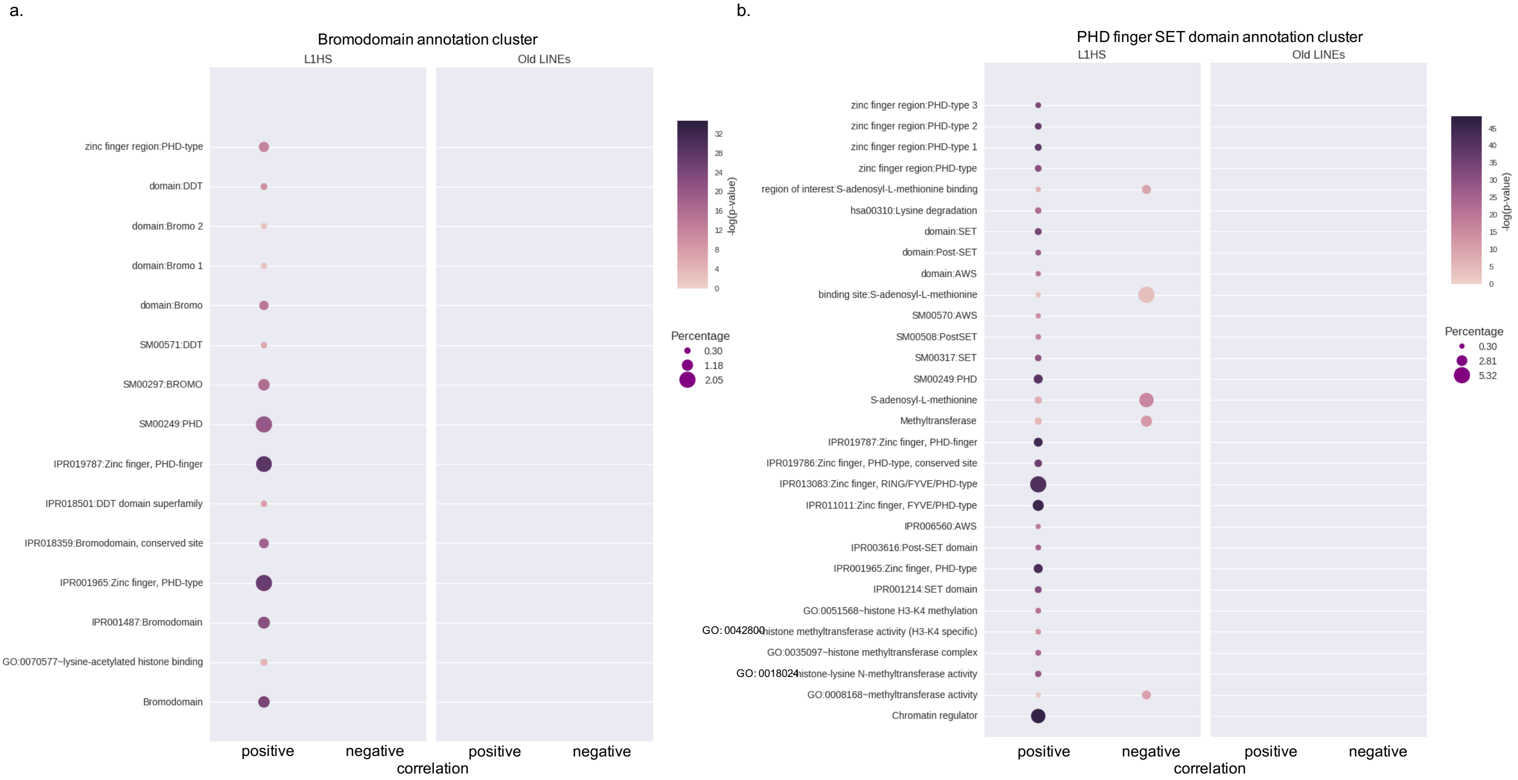

a.
Bromodomain annotation cluster
L1HS
Old LINEs
zinc finger region:PHD-type
domain:DDT
domain:Bromo 2
domain:Bromo 1
domain:Bromo
SM00571:DDT
SM00297:BROMO
SM00249:PHD
IPR019787:Zinc finger, PHD-finger
IPR018501:DDT domain superfamily
IPR018359:Bromodomain, conserved site
IPR001965:Zinc finger, PHD-type
IPR001487:Bromodomain
GO:0070577~lysine-acetylated histone binding
Bromodomain
positive
negative
positive
negative
correlation
32
28
24
20
16
12
8
4
0
-log(p-value)
Percentage
0.30
1.18
2.05
b.
PHD finger SET domain annotation cluster
L1HS
Old LINEs
zinc finger region:PHD-type 3
zinc finger region:PHD-type 2
zinc finger region:PHD-type 1
zinc finger region:PHD-type
region of interest:S-adenosyl-L-methionine binding
hsa00310:Lysine degradation
domain:SET
domain:Post-SET
domain:AWS
binding site:S-adenosyl-L-methionine
SM00570:AWS
SM00508:PostSET
SM00317:SET
SM00249:PHD
S-adenosyl-L-methionine
Methyltransferase
IPR019787:Zinc finger, PHD-finger
IPR019786:Zinc finger, PHD-type, conserved site
IPR013083:Zinc finger, RING/FYVE/PHD-type
IPR011011:Zinc finger, FYVE/PHD-type
IPR006560:AWS
IPR003616:Post-SET domain
IPR001965:Zinc finger, PHD-type
IPR001214:SET domain
GO:0051568~histone H3-K4 methylation
GO: 0042800histone methyltransferase activity (H3-K4 specific)
GO:0035097~histone methyltransferase complex
GO: 0018024histone-lysine N-methyltransferase activity
GO:0008168~methyltransferase activity
Chromatin regulator
positive
negative
positive
negative
correlation
45
40
35
30
25
20
15
10
5
0
-log(p-value)
Percentage
0.30
2.81
5.32

Figure 9

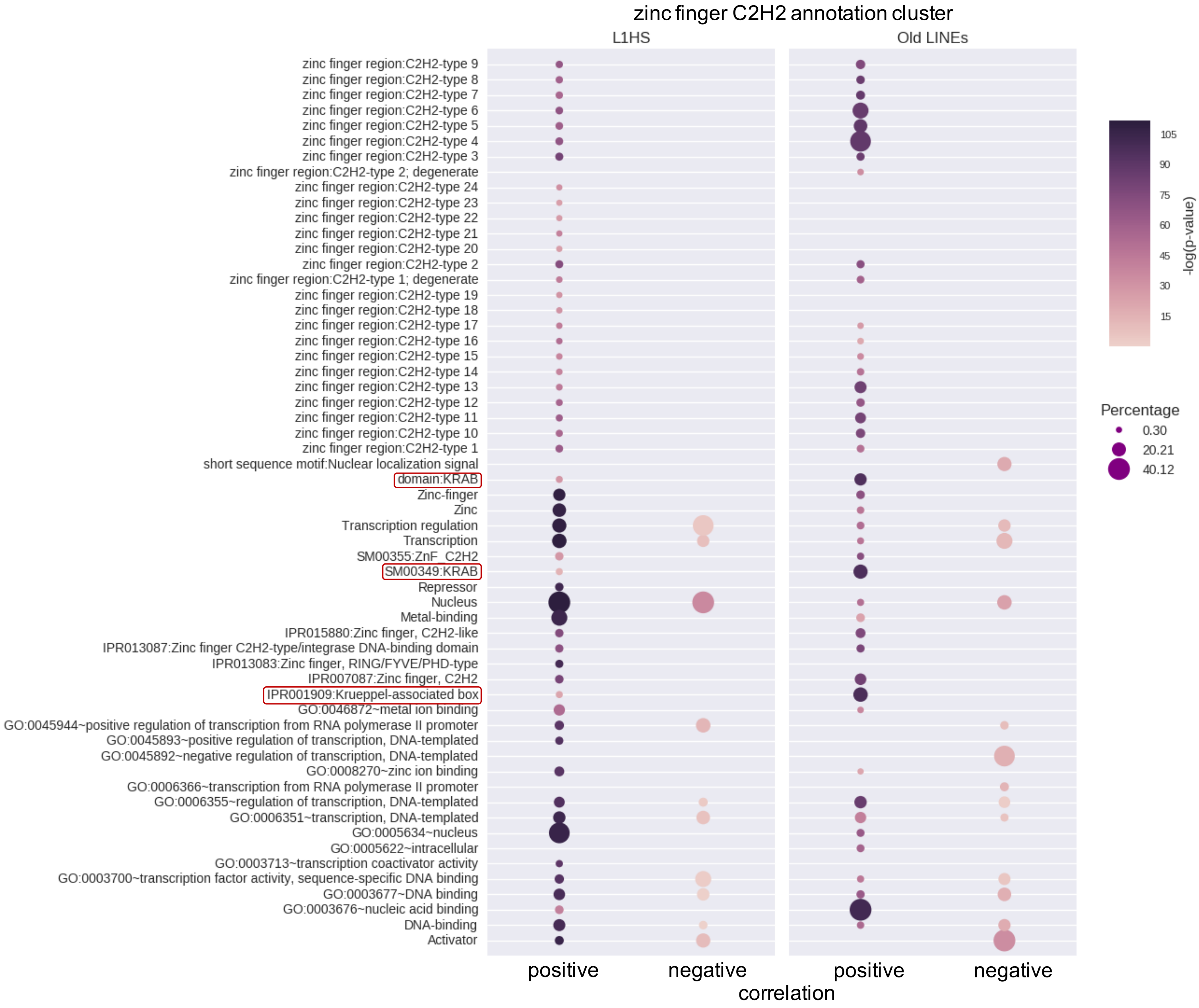

zinc finger C2H2 annotation cluster
L1HS
Old LINEs
zinc finger region:C2H2-type 9
zinc finger region:C2H2-type 8
zinc finger region:C2H2-type 7
zinc finger region:C2H2-type 6
zinc finger region:C2H2-type 5
zinc finger region:C2H2-type 4
zinc finger region:C2H2-type 3
zinc finger region:C2H2-type 2; degenerate
zinc finger region:C2H2-type 24
zinc finger region:C2H2-type 23
zinc finger region:C2H2-type 22
zinc finger region:C2H2-type 21
zinc finger region:C2H2-type 20
zinc finger region:C2H2-type 2
zinc finger region:C2H2-type 1; degenerate
zinc finger region:C2H2-type 19
zinc finger region:C2H2-type 18
zinc finger region:C2H2-type 17
zinc finger region:C2H2-type 16
zinc finger region:C2H2-type 15
zinc finger region:C2H2-type 14
zinc finger region:C2H2-type 13
zinc finger region:C2H2-type 12
zinc finger region:C2H2-type 11
zinc finger region:C2H2-type 10
zinc finger region:C2H2-type 1
short sequence motif:Nuclear localization signal
domain:KRAB
Zinc-finger
Zinc
Transcription regulation
Transcription
SM00355:ZnF_C2H2
SM00349:KRAB
Repressor
Nucleus
Metal-binding
IPR015880:Zinc finger, C2H2-like
IPR013087:Zinc finger C2H2-type/integrase DNA-binding domain
IPR013083:Zinc finger, RING/FYVE/PHD-type
IPR007087:Zinc finger, C2H2
IPR001909:Krueppel-associated box
GO:0046872~metal ion binding
GO:0045944~positive regulation of transcription from RNA polymerase II promoter
GO:0045893~positive regulation of transcription, DNA-templated
GO:0045892~negative regulation of transcription, DNA-templated
GO:0008270~zinc ion binding
GO:0006366~transcription from RNA polymerase II promoter
GO:0006355~regulation of transcription, DNA-templated
GO:0006351~transcription, DNA-templated
GO:0005634~nucleus
GO:0005622~intracellular
GO:0003713~transcription coactivator activity
GO:0003700~transcription factor activity, sequence-specific DNA binding
GO:0003677~DNA binding
GO:0003676~nucleic acid binding
DNA-binding
Activator
positive
negative
positive
negative
correlation
105
90
75
60
45
30
15
-log(p-value)
Percentage
0.30
20.21
40.12

Figure 10.

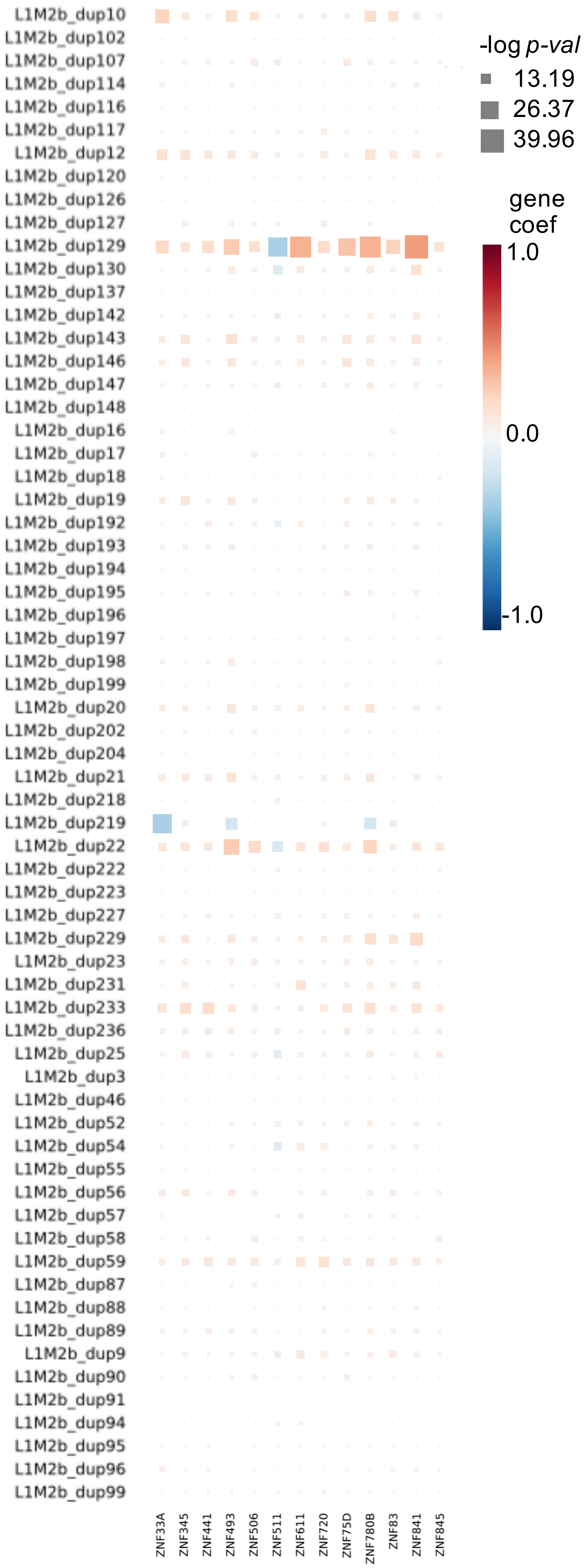
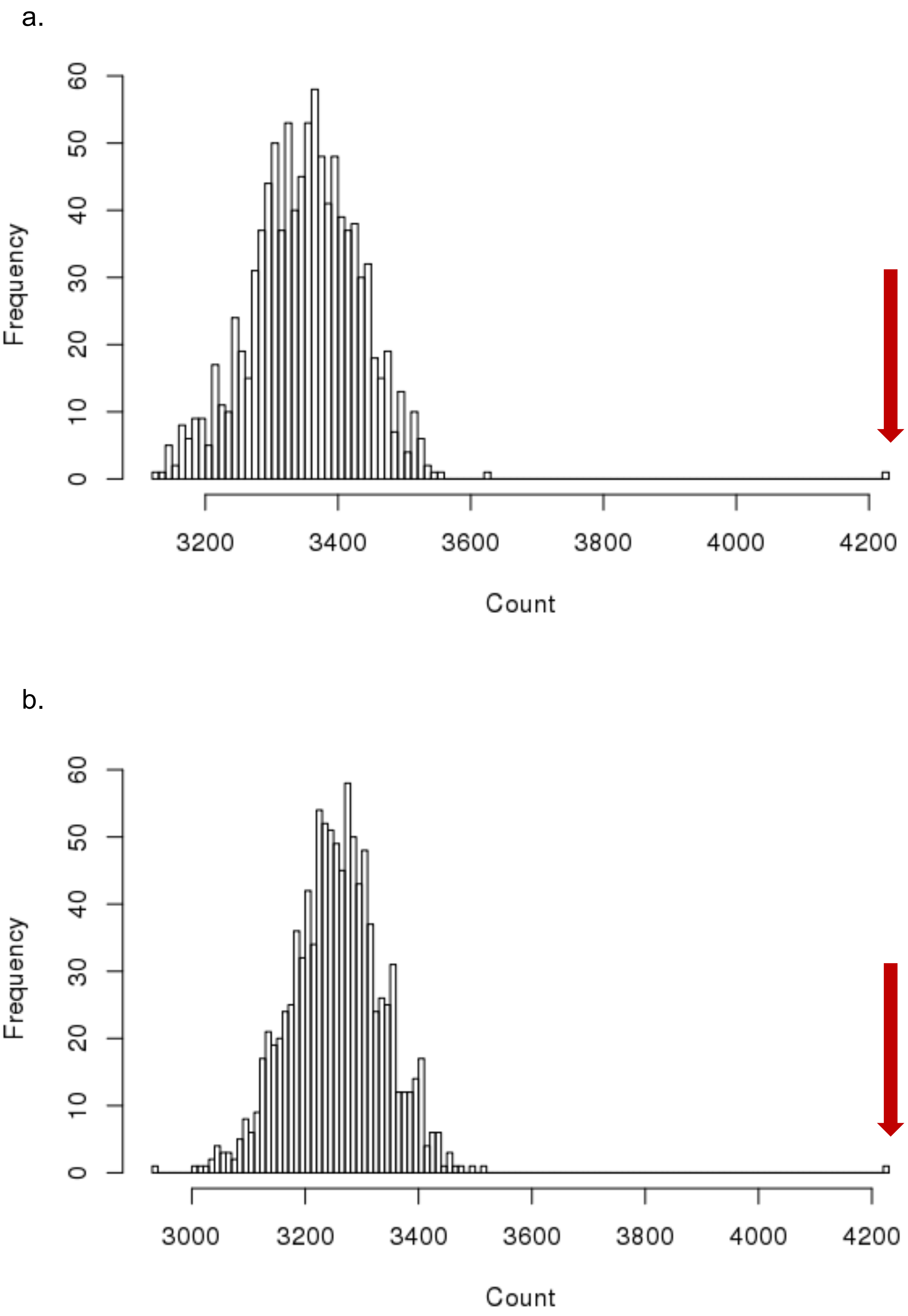

L1M2b_dup10
L1M2b_dup102
L1M2b_dup107
L1M2b_dup114
L1M2b_dup116
L1M2b_dup117
L1M2b_dup12
L1M2b_dup120
L1M2b_dup126
L1M2b_dup127
L1M2b_dup129
L1M2b_dup130
L1M2b_dup137
L1M2b_dup142
L1M2b_dup143
L1M2b_dup146
L1M2b_dup147
L1M2b_dup148
L1M2b_dup16
L1M2b_dup17
L1M2b_dup18
L1M2b_dup19
L1M2b_dup192
L1M2b_dup193
L1M2b_dup194
L1M2b_dup195
L1M2b_dup196
L1M2b_dup197
L1M2b_dup198
L1M2b_dup199
L1M2b_dup20
L1M2b_dup202
L1M2b_dup204
L1M2b_dup21
L1M2b_dup218
L1M2b_dup219
L1M2b_dup22
L1M2b_dup222
L1M2b_dup223
L1M2b_dup227
L1M2b_dup229
L1M2b_dup23
L1M2b_dup231
L1M2b_dup233
L1M2b_dup236
L1M2b_dup25
L1M2b_dup3
L1M2b_dup46
L1M2b_dup52
L1M2b_dup54
L1M2b_dup55
L1M2b_dup56
L1M2b_dup57
L1M2b_dup58
L1M2b_dup59
L1M2b_dup87
L1M2b_dup88
L1M2b_dup89
L1M2b_dup9
L1M2b_dup90
L1M2b_dup91
L1M2b_dup94
L1M2b_dup95
L1M2b_dup96
L1M2b_dup99
ZNF33A
ZNF345
ZNF441
ZNF493
ZNF506
ZNF511
ZNF611
ZNF720
ZNF75D
ZNF780B
ZNF83
ZNF841
ZNF845
-log p-val
13.19
26.37
39.96
gene coef
1.0
0.0
-1.0

Figure 11


a.
Frequency
0
10
20
30
40
50
60
3200
3400
3600
3800
4000
4200
Count
b.
Frequency
0
10
20
30
40
50
60
3000
3200
3400
3600
3800
4000
4200
Count