Supplementary Materials associated with the manuscript

TRANSPOSABLE ELEMENT EXPRSSION IN TUMORS IS ASSOCIATED WITH IMMUNE INFILTRATION AND INCREASED ANTIGENICITY

Authors: Yu Kong[1], Chris Rose[2†], Ashley A. Cass[3†], Martine Darwish[2], Steve Lianoglou[2], Pete M. Haverty[2], Ann-Jay Tong[2], Craig Blanchette[2], Ira Mellman[2], Richard Bourgon[2], John Greally[1], Suchit Jhunjhunwala[2], Matthew L. Albert[2], Haiyin Chen-Harris[2*]

Affiliations:
[1]Department of Genetics and Center for Epigenomics, Albert Einstein College of Medicine, New York 10461, USA.

[2]Genentech, Inc., 1DNA way, South San Francisco, CA 94080, USA.

[3]Department of Bioinformatics Interdepartmental Program, University of California at Los Angeles, Los Angeles, California, USA.

*Correspondence to: chen.haiyin@gene.com

†These authors made equal contributions.

Materials and Methods

Data manifest: all data used in this paper are from public databases

TCGA
RNAseq fastq files and DNA methylation raw data were downloaded from TCGA GDC. Samples are summarized in **Table S2** (RNAseq) and **Table S5** (DNA methylation) by cancer types. Tumor purity values were downloaded from NIH/NCI GDC PanCanAtlas Publications website: https://gdc.cancer.gov/about-data/publications/pancanatlas

CGP (Genentech Cancer Genome Project)
CGP RNAseq data had been previously deposited into EGA by prior publications (*17*). Samples in each cancer type are summarized in **Table S2.**

Accession No.	Project Name
EGAS00001000334	Genentech Small Cell Lung Cancer (SCLC) Screen
EGAS00001000288	Genentech Colon Cancer Screen
EGAS00001000736	Exome-seq, RNA-Seq, SNP array profiling of gastric tumor samples and cell line
EGAS00001000926	Study of non-clear cell renal cell carcinoma

GBM cell line data
Matched RNAseq, proteome and MHC-I peptidome data previously published by Shraibman et al. 2016 (*39*) on 3 glioblastoma (GBM) cell lines before and after decitabine treatment were analyzed.

REdiscoverTE method

REdiscoverTE uses the light weight-mapping method, *Salmon* (*19*), for repetitive element expression quantification. The method as applied quantifies expression for all REs included in the reference transcriptome. In this study downstream analysis focused on TE expression.

Generating REdiscoverTE reference transcriptome. Salmon version 0.8.2 was used to generate quasi mapping index. The reference transcriptome includes:

1. distinct RNA transcript sequences (n=98,029) from the GENCODE release 26 basic (*20*)
2. RepeatMasker elements (n=5,099,056) from the standard chromosomes, excluding all polyA repetitive elements (**Fig. S1A, Fig. S1B**). (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org. 1996-2010)
3. distinct sequences representing GENCODE RE-containing introns (n=185,403) and excluding any regions overlapping with exons on either strand since we analyzed non-strand-specific RNAseq.

Two transcriptome indices were built, one without (index 1) and the other with (index 2) the inclusion of RE-containing introns. We showed with simulation significant performance improvement by index 2 over index 1 (**Fig. S1E, Fig. S1F, Fig. S1G**). As a result *REdiscoverTE* transcriptome includes the RE-containing introns listed above.

Salmon quantification. *Salmon* version 0.8.2 was used to quantify RNA-seq data with adjustment for GC content bias and sequence specific bias options. *REdiscoverTE* reference transcriptome described above was used. *Salmon* produces quantification results in two ways: transcript-per- kilobase-million (TPM) and number of reads. We have chosen to use read counts for all downstream analysis based on benchmarking analysis detailed in Supplementary information.

Post-*Salmon* quantification, RE and host gene transcripts were aggregated separately. Host isoforms were aggregated to the gene level according to *ensembl* gene ID. All aggregation and downstream analysis of the aggregated expression were performed using R (R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.)

Aggregation of RE expression to the subfamily level. Due to the high degree of sequence homology among copies of REs from the same subfamily, *Salmon* quantification of read counts

at individual REs were aggregated to the level of RE subfamily, family and class according to hierarchies defined by the human Repeatmasker for Hg38 (a.k.a. repName, repFamily, repClass, respectively) by summing the counts of all individual REs from a given subfamily/family/class (**Fig. S1A**).

Defining genomic context of REs in relations to host genes.

To distinguish RE expression by genomic context in relation to genes, we separately tallied RE counts by the genomic locations of individual REs into exonic, intronic and intergenic RE.

To defined whether REs are exonic, intronic or intergenic, we downloaded the annotated human transcriptome from Gencode (*20*) Version 26 Basic GTF/GFF file (https://www.gencodegenes.org). Gencode defines the following basic categories of features: gene, transcript, coding exon (CDS), exon, UTR. We inferred intronic and intergenic regions from these features (**Fig. S1B** pie chart). Then simplified these categories of features into 3 mutually exclusive regions: exons (union of all exons and UTRs), introns (union of intronic regions excluding any overlap with exons) and intergenic regions. The R package *GenomicRanges* (*58*) was used to perform range overlap operations. Repeatmasker RE genomic ranges were overlapped with these simplified host gene features to define whether a given RE is exonic, intronic or intergenic. For REs overlapping multiple contexts (e.g. an RE that resides at an exon-intron boundary), their locations are assigned with the following priority: exon > intron > intergenic. For example, an RE residing at the exon-intron boundary is considered as an exonic RE.

REdiscoverTE performance benchmarking

RSEM simulation

To benchmark the accuracy of *REdiscoverTE*, we carried out extensive RSEM (*21*) simulations to create fastq files where we have ground truth on expression levels (TPM) of all features in the transcriptome. To create realistic gene and RE expression levels we first used RSEM to learn sequence statistics from two TCGA RNA-Seq samples with different proportions of RE reads estimated by REdiscoverTE: one with 5.4% of reads derived from REs as estimated by *REdiscoverTE* (THCA, normal sample, TCGA-EL-A3ZS-11A-11R-A23N-07), and one with 11.5% of reads derived from REs (LAML, tumor sample, TCGA-AB-2955-03A-01T-0734-13), then generated corresponding simulated fastq files based on learned models (**Fig. S1C** step 2, step 3) and two additional modifications described below.

A main goal of the simulation is to evaluate *Salmon*'s ability to quantify gene expression stemming from highly repetitive and similar features. We considered the added complexity where some REs can overlap with host gene features that are also expressing mRNAs, e.g. host gene transcripts or retained introns. The following two modifications were made to the default RSEM simulation process to address these issues (also described in **Fig. S1C**):

1) To evaluate Salmon's ability to distinguish RE expression from overlapping host transcripts (**Fig. 1A** orange reads from RE #2 vs. blue reads from exon), we simulated RE expression at varying levels above the host gene expression. We chose to focus our simulations on those RE subfamilies that have copies residing in all three types of genomic regions with respect to genes: exonic, intronic, and intergenic regions. Out of 15,440 RE subfamilies, there are 3,659 subfamilies contain at least one copy of RE in each of exonic, intronic, and intergenic regions. After exluding simples repeats, there were 1,135 subfamilies that satisfy this criteria (**Fig. S1C** Venn Diagram). We randomly chose 1,000 non-Simple Repeat subfamilies from these 3,659 to evaluate with simulation experiments (**Fig. S1C** workflow). If an RE overlapped with multiple isoforms or genes, for simplicity, we randomly chose one isoform to simulate for every RE residing within the transcript. In total, 1,969,915 REs from 1000 non-Simple Repeat subfamilies were simulated; 63,021 of them overlapped with genes.

2) To evaluate *Salmon*'s ability to distinguish RE expression from retained introns (**Fig. 1A** orange reads from RE #3 vs. green reads from retained intron), two isoforms were simulated for genes with intronic REs: one with the RE-containing intron retained, and one without the intron.

After *RSEM learned* statistical profiles from the two TCGA fastq files, and before generating simulated fastq files, we manually changed the TPM values in the *isoforms.results* output file from rsem-calculate-*expression* in order to generate more variation in intron retention levels and RE to RE-containing gene expression level ratio. **Fig. 1D** provides the final profiles of these simulations.

Comparing *REdiscoverTE* to *RepEnrich*. *RepEnrich* (*14*) is a two-step repetitive elements quantification method: step1 -- alignment of RNA-seq reads to hg38 using Bowtie (*22*), step 2 -- applying RepEnrich script to reads uniquely mapped to repeatmasked regions and multi-mapped reads from step 1 using RepEnrich pre-defined repetitive pseudogenomes as reference. *RepEnrich* pseudogenomes are defined for 1000+ RE subfamilies (excluding simple repeats and low complexity repeats), each is a concatenation of all repetitive elements in the subfamily with additional flanking sequences and spacers.

We benchmarked performance of *REdiscoverTE* against *RepEnrich* on RSEM simulated RNA-seq data. We followed default workflow of RepEnrich. Performance of *REdiscoverTE* and RepEnrich were evaluated using the metric mean absolute relative difference (MARD) at the level of subfamily, where MARD is defined as in *Salmon* publication (*19*):

$$MARD = \frac{1}{N}\sum_{i=1}^{N} \frac{|Salmon\ count_i - Simulated\ count_i|}{Salmon\ count_i + Simulated\ count_i}$$

Here N is the total number of features, where features could be individual RE transcripts or aggregated features such as RE subfamily.

Comparing *REdiscoverTE* quantification of 66 ERVs in TCGA RNAseq data with Rooney et al. Cell 2015

To directly compare with ERV quantification results from Rooney et al. 2015 (*15*), we created a *Salmon* reference transcriptome that included 90k human transcripts and the same 124 sequences for the 66 ERVs analyzed by the authors (from Mayer et al. 2012 (*59*)). RNA-seq from 5,217 TCGA samples (20 cancer types) were quantified by *Salmon* (Version 0.6.0). ERV read counts were normalized by total counts mapped to genes, similar to Rooney et al. where ERV expression was normalized by total counts of reads mapped to genes. Counts per million (CPM) of ERVs from Salmon quantification was then compared with cpm value published in Rooney et al. 2015 Supplementary table S5b (distribution of correlation values in **Fig. S1K**). The expression level for the three ERVs highlighted in Rooney et al. 2015 Fig 4A as 'tumor-specific' plotted for comparison (**Fig. S1J**).

Quantification of TE Expression in RNA-Seq Data with REdiscoverTE

RNA-Seq data processing. TruSeq adapters were trimmed by Cutadapt (http://cutadapt.readthedocs.io/en/stable/) from both TCGA and CGP. *REdiscoverTE* was run as described above for whole transcriptome expression quantification.

Normalization of TE and gene expression. Following expression aggregation: isoform level to gene level, individual REs to RE subfamilies, two expression count matrices were created for each data set, one for gene expression, the other for RE expression. We chose to calibrate both expression matrices using total counts of gene expression, which we considered to be more stable across samples. Expression normalization was performed in R using the Bioconductor packages *edgeR* (*26*) using "RLE" method by function *calcNormFactors*. log2CPM is then calculated with prior count set to 5. After normalization, for the RE expression matrix, we focuses on the 1052 transposable element (TE) subfamilies for downstream analysis in this study.

Differential expression analysis. To control for potential batch effect and patient-to-patient variation, only tumor and matched adjacent normal samples are used for differential expression analysis. Cancer types with fewer than 10 normal samples were excluded from this analysis; 13 TCGA cancer types and 5 CGP cancer types satisfied this threshold. The R packages *limma* and *voom* (*27*, *28*) were used for differential expression analysis using aggregated count matrices as input. Prior to differential expression analysis, two filters were applied to exclude genes or TEs with low expression, requiring (1) at least 10% of samples have counts greater than zero, and (2) a log2(CPM) threshold. The log2(CPM) threshold was determined independently for each indication based on visual inspection of the mean-variance trend (estimated by the *voom* function in limma) to ensure variance was monotonically decreasing for low mean expression. Benjamini-Hochberg (BH) approach was used to control false discovery rate (FDR) within each indication (*60*). Differentially expressed genes/TEs were determined at the threshold of: abs(log2 fold change) > 1 and FDR < 0.05.

Divergence of TEs. *RepeatMasker* divergence score in terms of basepair (bp) difference from consensus sequence in 1000bp was used as proxy of the age of a TE element. For each TE

subfamily, we calculated the average divergence score across all the copies of TEs within the subfamily.

Methylation analysis

TCGA 450k array methylation data processing. Illumina 450k Infinium methylation arrays were processed using the "lumi" (*61*). Raw array data were background corrected (lumiB method) and variance stabilized and normalized (lumiT and lumiN methods). Beta values were calculated per CpG site by flooring intensity values at zero and then calculating

$$Beta = \frac{methylated\ density}{methylated\ density + unmethylated\ density + alpha}$$

where alpha is a regularization parameter set at the default of 100 recommended by Illumina (*62*). M-values were transformed from Beta values by:

$$M = log2(\frac{Beta}{1 - Beta})$$

Liftover of CpG sites in 450k array to hg38. Hg19 annotation of 450k probes was obtained using R package "IlluminaHumanMethylation450kanno.ilmn12.hg19" (Kasper Daniel Hansen (2016). IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. R package version 0.6.0.) Using the liftOver utility provided by UCSC genome browser, physical coordinates of 450K probes in hg19 annotation were lifted to hg38 reference genome. 485,441 out of 485,512 CpGs were successfully converted to Hg38, 71 failed due to position removal in Hg38 assembly.

Identification of differentially methylated cytosine. In order to identify differentially methylated cytosines (DMCs), M values of DNA methylation data were used to fit the linear regression model with tumor/normal status and patient ID as covariate using *lmFit* from the R package *limma* (*27*). To control for potential batch effect and patient-to-patient variation, only tumors (N = 237) with matched normal (N = 236) samples were included for analysis in 10 TCGA cancer types (**Table S5**) that had both DNA methylation and RNA-seq data.

DMCs were defined as CpGs with the average absolute beta value change (Δbeta) >=10% and FDR<0.05. DMCs are called demethylated if the average beta value in tumor is lower than normal samples, and methylated if the average beta value in tumor is higher than normal. Δbeta value for each CpG site is defined for tumor and matched normal sample pairs as:

$$\Delta beta = beta_{tumor} - beta_{normal}$$

Distribution of 450K CpGs in different genomic features: similar to RE expression analysis, CpGs were classified into exonic, intronic and intergenic CpGs according to the genomic features defined above. The genomic distribution of all 450K CpGs and CpGs overlapping with TEs were visualized in **Fig. S3A**. In total, there were 70,004 CpGs located within TEs that overlapped with 59,739 individual elements from 992 TE subfamilies.

Spatial profile of methylation around TEs were analyzed by extracting CpG sites near intergenic TEs. CpGs located outside but nearby individual intergenic TE elements were binned into two categories for further analysis: those within 1kb of TEs and those within 10kb of TEs. Due to the lack of functional annotation for all TE transcripts, the most 5' bp of each individual TE as annotated by *RepeatMasker* was used as proxy for Transcription Start Site (TSS).

For the 1kb methylation spatial analysis, all 450K Array CpG sites within 500bp +/- of TSS for each intergenic TE were extracted using the *findOverlaps* function in the R package *GenomicRanges* with strand information taken into account. This resulted in 90,950 TE-proximal CpG sites for 1,007 TE subfamilies.

For these 1,007 TE subfamilies, CpG sites from a bigger, 10kb window: 5kb up- and down-stream from the start and end coordinates of individual intergenic TEs, were extracted, resulting in 155,360 CpG sites.

For spatial profile analysis (e.g. **Fig. 3F, Fig. 3G**) CpGs within TEs were represented using a proportional distance as follows: TEs from the same subfamily were length normalized to create proportional position within the TE, ranging from 0% to 100% that correspond to the start and end of TE.

Correlation between DNA methylation and TE expression. Two types of correlation analyses were carried out, both using Pearson correlation on aggregated intergenic TE expression at the subfamily level and M values of CpG sites, chosen over beta to obtain higher statistical power. The first one uses the per-sample average M-values at all CpGs within 1kb of TEs (e.g. **Fig. 3E, Fig. S3F** column 3) the second one is performed at each CpG site around 5kb +/- of all intergenic elements in the TE subfamily, using M-values from all samples at the given CpG site (e.g. **Fig. 3F, Fig. S3F** columns 4 and 5). FDR was obtained by adjusting p values for multiple testing (Benjamini & Hochberg) across the 1007 tests within each cancer type. Significant correlation was defined as FDR<0.05 and |cor|>=0.4.

TE demethylation enrichment score. We defined methylation state as the ratio of number of demethylated vs. number of over-methylated DMC sites. Methylation state is 1 when there are equal number of demethylated and over-methylated DMCs, > 1 when there is bias in the direction of demethylation, < 1 when there is bias toward over-methylation.

We then computed a TE demethylation enrichment score (**Table S5**) as the ratio of within-TE methylation state (using DMC CpG sites in intergenic TEs) to global methylation state (using all DMC sites). This enrichment score is 1 when the methylation state in TE is comparable to that of the global methylation state, > 1 when a higher proportion of TE DMCs are demethylated, < 1 when a smaller proportion of TE DMCs are demethylated.

$$TE\ demethylation\ enrichment = \frac{N_{demethylated\ DMC\ in\ TE} \Big/ N_{over-methylated\ DMC\ in\ TE}}{N_{demethylated\ DMC\ anywhere} \Big/ N_{over-methylated\ DMC\ anywhere}}$$

Association between gene signatures and TE expression

Gene signatures and calculation of gene signature scores. Twenty-four gene signatures associated with major cellular pathways related to cancer, DNA damage response (DDR) and immune response were selected from previous publications (**Table S7**). The R package *multiGSEA* (https://github.com/lianos/multiGSEA) was used to score gene signature expression based on singular value decomposition.

Calculation of tumor cellularity scores using *xCell*. In order to estimate the immune content within tumor samples, we applied *xCell* (*33*), a recently developed gene signature-based approach for tissue cellularity de-convolution within RNA-seq data, to TCGA samples and obtained the cellularity enrichment scores for 64 cell types, including lymphoid and myeloid cell types (**Fig. S4C**). We further confirmed the accuracy of *xCell* estimations by examining the concordance between certain cell types (e.g. CD8+ T cells) and related gene signature scores (e.g. CD8+ effector T cells) computed with *multiGSEA* (**Fig. S4C**). In addition, for each sample, we defined *total lymphoid content* as the sum of 21 lymphoid cell scores: CD8+ T-cells, NK cells, CD4+ naive T-cells, B-cells, CD4+ T-cells, CD8+ Tem, Tregs, plasma cells, CD4+ Tcm, CD4+ Tem, memory B-cells, CD8+ Tcm, naive B-cells, CD4+ memory T-cells, pro B-cells, class- switched memory B-cells, Th2 cells, Th1 cells, CD8+ naive T-cells, NKT and Tgd cells. *Total myeloid content* was defined as the sum of 13 cell scores: monocytes, macrophages, DC, neutrophils, eosinophils, macrophages M1, macrophages M2, aDC, basophils, cDC, pDC, iDC, mast cells.

Calculation of correlation between gene signature/xCell scores and TE expression adjusted by tumor purity. For each cancer type, Spearman correlation coefficients between log2CPM expression of 1,052 TE subfamilies and 24 gene signature scores were computed using the R package for partial correlations *ppcor* (https://CRAN.R-project.org/package=ppcor), with tumor purity as a covariate. Benjamini-Hochberg approach was used to control false discovery rate (FDR) within each indication separately.

In addition, Spearman correlation coefficients between 1,052 TEs and 64 *xCell* scores were computed using the same method.

Lasso analysis to identify associations between gene signature and TE expression. To identify top TE subfamilies associated with each of the 24 cellular pathways and gene signatures, we exploited Lasso regularized regression -- generalized linear model via penalized maximum likelihood using the R package *glmnet* (*34*). In order to account for variations of cellular content that existed between tumor samples, we included tumor purity as well as abundance of total lymphoid and myeloid content (*xCell* section above) as parameters in the lasso model. To avoid any possible bias introduced by normal-tumor status, only tumor samples were used in the regression model. The following Full Lasso model was computed within each cancer type separately:

$$\textit{Tumor gene signature score} \sim TE1 + TE2 + \ldots + TE1052 + \textit{tumor purity} + \textit{lymphoid} + \textit{myeloid}$$

where TE expression are in units of log2CPM.

Ten-fold cross-validation was performed for each regression, lasso coefficients at one standard error of the minimum mean cross validation errors (lamda 1SE) were used. Each Lasso fit yielded a sparse set of predictors – variables with non-zero coefficients, corresponding to TE subfamilies with significant contributions to the variability of a given gene signature. We then ranked all 1055 dependent variables (TEs and 3 covariates) by their average absolute coefficient values across cancer types to select the top 20 predictors associated with the gene signature of interest. To produce the Lasso rank coefficient heatmap (**Fig. S4D**), we indicated the rank of these top predictors by their absolute coefficient values within each cancer type. Dots corresponds to a coefficient of zero for a given cancer type (also shown in **Table S8**).
Post Lasso regression, deviance ratio from the models (fraction of deviance explained) were used as R^2 values for these models.

Cellularity model. Relative contribution to gene signature variance from the 3 cellularity scores (purity, lymphoid and myeloid) was estimated from R^2 values of the following linear regression (performed by the *lm* function in R) for each of the 24 gene signature scores:

Tumor gene signature scores ~ tumor purity + lymphoid + myeloid

Top TE linear model. Building on the cellularity model, this linear model includes the 3 cellularity scores as well as up to 6 top TE subfamilies with the highest non-zero lasso coefficient:

Tumor gene signature scores ~ tumor purity + lymphoid + myeloid + TE1 + TE2 + ... + TE6

This model was skipped if all coefficients for TE subfamilies were zero in the corresponding Lasso model.

As both cellularity and top TE linear models are based on linear regression, their R^2 values can be directly compared for a given gene signature.

top TE fractional variance contribution = R^2_topTE_model - R^2_cellularity_model

TE Peptide Identification

Mass spectrometry (MS) raw data files for the global proteome (unenriched peptides) and MHC-bound peptidome (pan-MHCI enriched peptides) were obtained from PRIDE (PXD003790) and SysteMHCAtals (SYSMHC00007), respectively.

To enable identification of TE-derived peptides in the GBM proteome and MHC-bound peptidome data, we collected nucleotide sequences at all individual loci for the 62 TE subfamilies that were significantly over-expressed at either the intergenic or intronic regions upon 5'aza treated condition, performed 6 frame translations (both forward and reverse direction), then fragmented the resulting amino acid sequences at all stop codons. This yielded ~1.1M peptide fragments, ranging 7 to 1,321 amino acids in length. The peptide fragments were combined with the human protein sequences in Uniprot (downloaded Jan 1st of 2017) and common contaminant proteins to create a database used for searching non-MHC enriched mass spectrometry data. TE-derived peptide fragments were further reduced into 4.6M 11mers generated with a moving window of 8 amino acid overlaps, with duplicates removed. This 11mer database was also combined with the human protein sequences in Uniprot and common contaminant to create a database used for searching MHC-enriched mass spectrometry data.

Raw MS data was analyzed using PEAKS Studio (Bioinformatics Solutions Inc., v8.5) (*63*). In brief, raw MS data were refined and sequence tags were identified by a *de novo* search algorithm. Identified sequence tags were used in the assignment of peptide sequences to MS data through a database search. For all database searches the following parameters were used: precursor tolerance = 25 ppm, fragment ion tolerance = 0.02 Da, enzyme = none, variable modifications include deamidation [N or Q, 0.98 Da] and oxidation [M, 15.99 Da], and max number of variable mods = 3. Data were filtered to 1% FDR at the peptide level, but due to TE peptide fragments being represented as multiple "protein" entries within the database protein level FDR was not performed. For MHC-bound peptidome data a median of 3 peptide spectral matches (PSMs) were identified for non-TE peptides, due to this TE peptides were considered high confidence if they had been identified in ≥3 spectra.

Additionally, we performed an identical search against a database that did not include the TE peptide 11-mers in order to determine if TE PSMs mapped to alternative sequences (**Table S12**). A total of 555 PSMs mapped to TE peptides when the 11-mer peptides were included in the database. Of these, 487 failed to match a peptide sequence at 1% FDR when TE peptide 11-mers were excluded from the database. Of the remaining 68 PSMs, 64 matched to Trembl, Uniprot entries which are short RNA transcript reads that likely originate from expressed TE peptides. The remaining 4 spectra matched to alternative proteins in Uniprot (3) or a decoy protein entry (1). If we consider these 4 spectra false observations we can estimate our experimental FDR to be ~1.4%.

For further validation of TE peptide identification, we synthesized 15 of the 83 unique peptides and analyzed them by MS. All MS analyses were performed on an Orbitrap Fusion mass spectrometer (ThermoFisher Scientific, San Jose, CA) with peptides separated over a nano-LC column (100 µm I.D. packed to 25 cm with Waters M-Class BEH 1.7 µm packing material) by a gradient delivered by a Waters NanoAqcuity nano-LC. For each synthetic TE peptide, 250 fmol was injected and analyzed by MS utilizing various collision energies (HCD at 20, 25, 30, and 35 NCE) in order to match fragmentation spectra of Shraibman et. al. Synthetic peptide MS data were analyzed in PEAKS in an identical manner to the MHC-bound peptidome data. Annotated spectra for the synthetic and experimental spectra were manually compared to validate peptide identifications. Through this process we were able to confirm 17 of 18 peptide spectra as visual matches, adding further confidence to TE peptide identifications.

MHC Class I Peptide Exchange

Recombinant HLA-A*03:01 MHCI was refolded in the presence of a conditional peptide ligand that contains a UV sensitive amino acid, as previously described (*64*). The resultant purified, stable complex was incubated in the presence of 100 fold molar excess of synthetic TE derived peptides of interest. HLA-A*03:01 was present at a concentration of 50 ug/ml (1.04 uM) in 25 mM TRIS pH 8.0, 150 mM NaCl, 2mM EDTA, 5% DMSO. The peptide exchange reaction mixture was incubated for 25 min under a UV lamp set to 365 nm to induce cleavage of the UV sensitive amino acid 3-amino-3-(2-nitro)phenyl-propionic acid. Samples were then incubated at room temperature, overnight, to allow for peptide exchange to occur. Upon cleavage of the conditional peptide ligand, synthetic TE derived peptides with suitable properties (affinity, solubility) exchanged into the complex displacing any fragments of the cleaved conditional ligand.

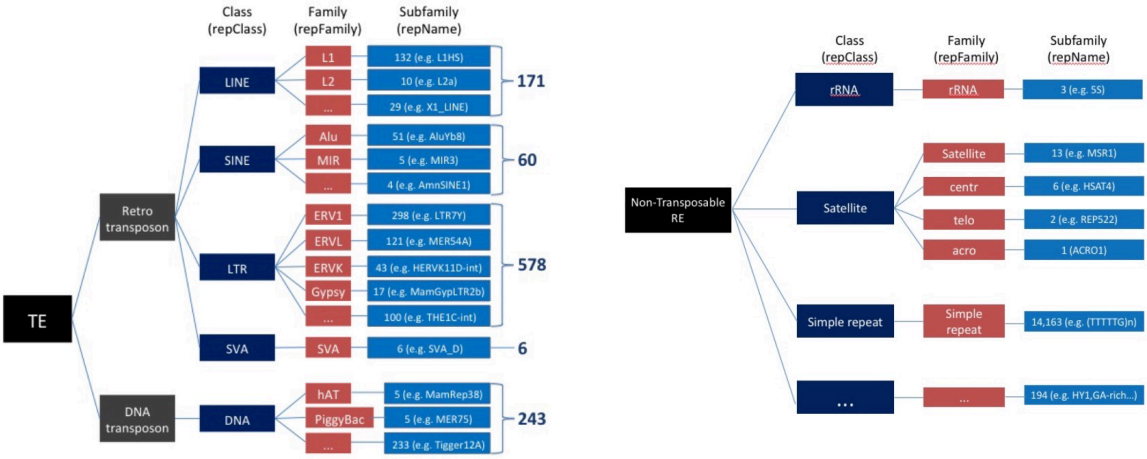
2D-LCMS Characterization of Peptide Exchange

To determine successful exchange of TE derived peptides into HLA-A*03:01 complexes, a 2-dimensional liquid chromatography mass spectrometry method was used. The first dimension LC method employed an analytical SEC column (Agilent AdvanceBio SEC 300Å, 2.7um, 4.6 x 15 mm) to separate intact complex from excess peptide run at an isocratic flow of 0.7 ml/min in 25 mM TRIS pH 8.0, 150 mM NaCl for 10 min. A sampling valve collected the entirety of the complex peak that eluted between 1.90 – 2.13 min in a volume of 160ul and injected it onto the second dimension reversed phase column (Agilent PLRP-S 1000 Å, 8um, 50 x 2.1 mm). The second dimension column was exposed to a gradient of 5-50% mobile phase B in 4.7 min at 0.55 ml/min with the column heated to 80°C. Mobile phase A was 0.05% TFA. Mobile phase B was 0.05% TFA in acetonitrile. The column eluent was sent to an Agilent 6224 TOF LCMS for mass spectrometry data acquisition.

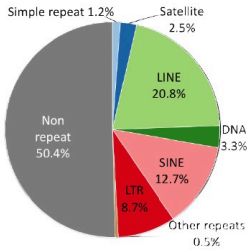
HLA-A*03:01 complex peak area (detected at 280 nm) in the first dimension and mass spec detection of the peptide in the second dimension are used to determine successful exchange. Successful exchange of a peptide into the complex after cleavage of the conditional ligand during the peptide exchange reaction stabilizes the complex and results in nearly complete recovery of the starting complex measured in the first dimension SEC analysis. The peptide that has exchanged into the complex can then be detected in the second dimension, where the complex is run under denaturing conditions with mass spectral analysis allowing for direct detection of the peptide of interest. Unsuccessful peptide exchange reactions result in destabilized complex after the cleavage of the conditional ligand when a peptide fails to bind to and stabilize the complex. This is measured as a reduction in A280 peak area of the complex on SEC and an absence of peptide in the second dimension. In some cases, such as for peptide RLAPRPASR, no reduction in peak area was observed, however the peptide was not detected by mass spectrometry. A small number of peptides, due to their properties, are not captured by the second dimension chromatography column and method. In these cases, the peak area recovery is enough to suggest successful exchange when the proper experimental controls are used.

FigureS1. TE in the human genome and REdiscoverTE

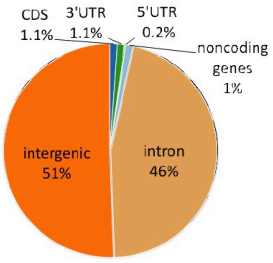
A



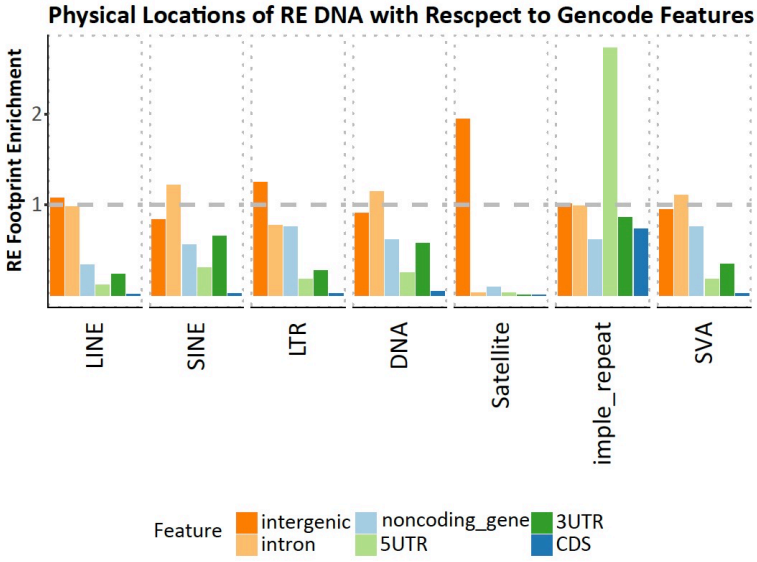
B



The repetitve element content of the human genome (Hg38)



Genomic features in the Gencode human transcriptome (v26 basic, Hg38)



C

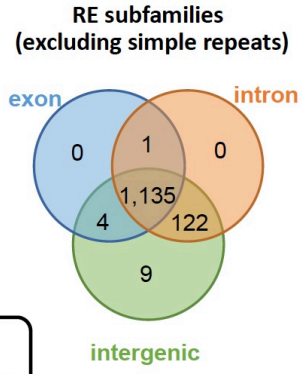
Generate transcriptome fasta files:
- 1,000 subfamilies randomly drawn from 1,135 non-Simple Repeat subfamilies, corresponding to 1,969,915 individual REs in total.
- REs overlapped 63,021 genes

Learn transcriptome patterns from real data:
- rsem-prepare-reference --bowtie2
- rsem-calculate-expression

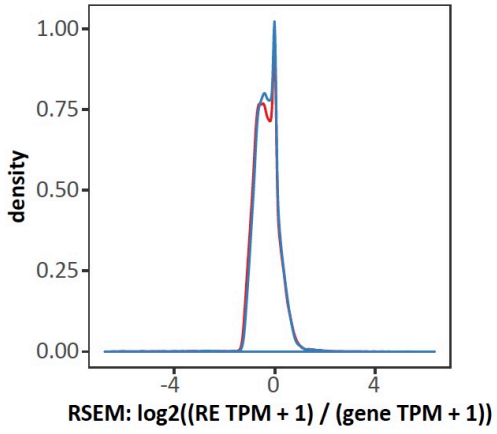
Simulate reads based on learned model:
- Manually update TPM in .isoforms.results output from rsemcalculate-expression
- rsem-simulate-reads

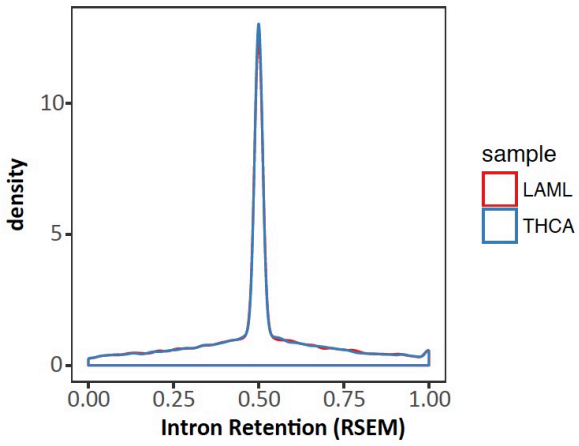
Salmon quantification:
- salmon --quasi index
- salmon quant

Evaluate quantification accuracy:
- simulated read count vs. Salmon's estimated read count

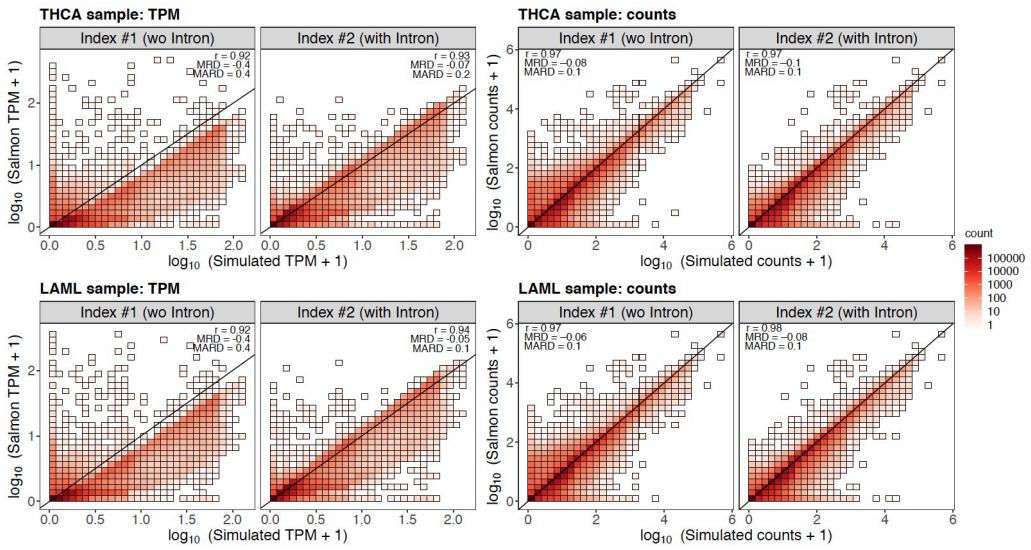


D

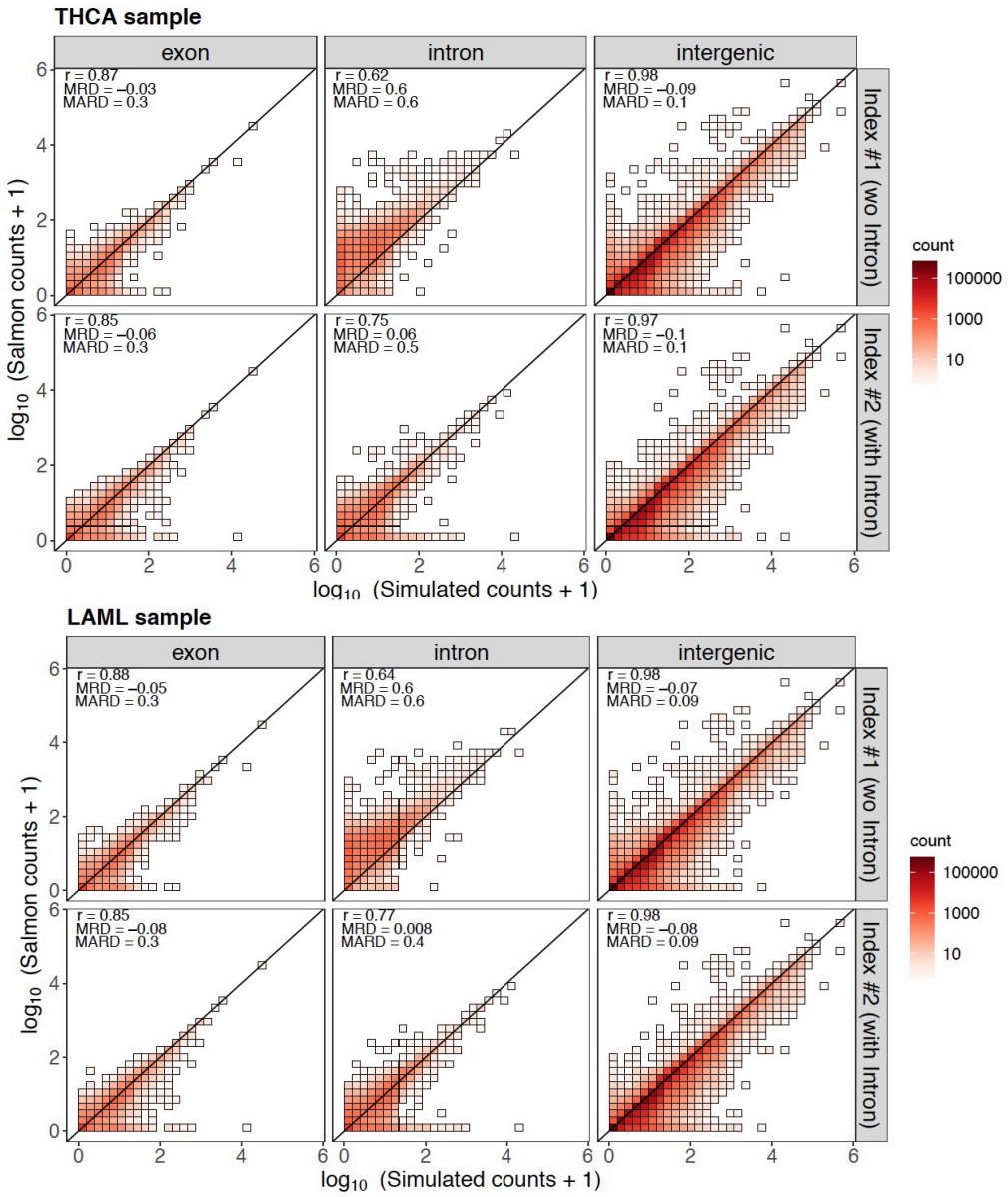




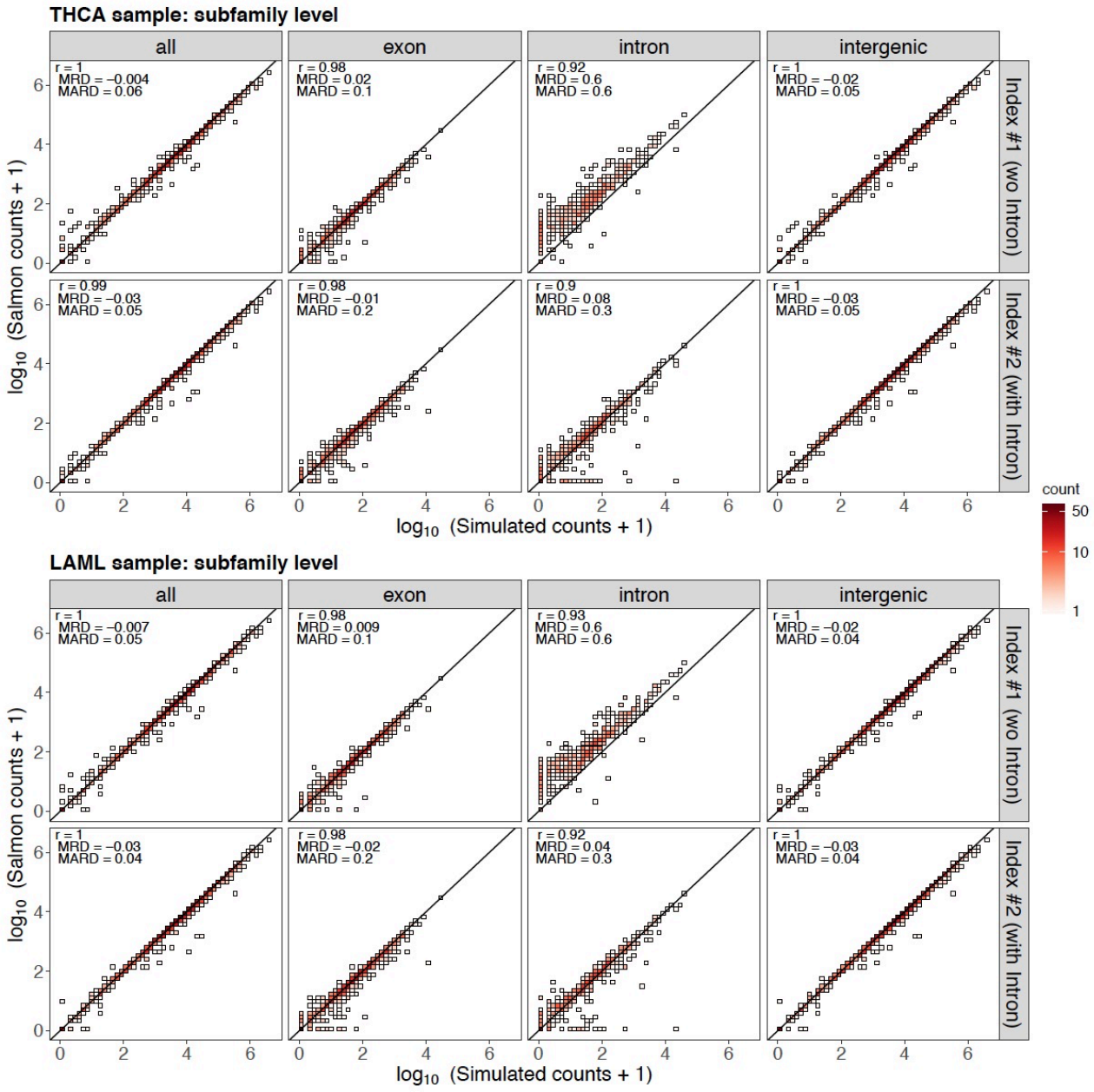
E


THCA sample: TPM
Index #1 (wo Intron)
Index #2 (with Intron)
r = 0.92
MRD = -0.4
MARD = 0.4
r = 0.93
MRD = -0.07
MARD = 0.2
log10 (Salmon TPM + 1)
log10 (Simulated TPM + 1)
THCA sample: counts
r = 0.97
MRD = −0.08
MARD = 0.1
r = 0.97
MRD = −0.1
MARD = 0.1
log10 (Salmon counts + 1)
log10 (Simulated counts + 1)
LAML sample: TPM
r = 0.92
MRD = -0.4
MARD = 0.4
r = 0.94
MRD = -0.05
MARD = 0.1
LAML sample: counts
r = 0.97
MRD = −0.06
MARD = 0.1
r = 0.98
MRD = −0.08
MARD = 0.1
count
100000
10000
1000
100
10
1

F


THCA sample
exon
intron
intergenic
Index #1 (wo Intron)
Index #2 (with Intron)
r = 0.87
MRD = −0.03
MARD = 0.3
r = 0.62
MRD = 0.6
MARD = 0.6
r = 0.98
MRD = −0.09
MARD = 0.1
r = 0.85
MRD = −0.06
MARD = 0.3
r = 0.75
MRD = 0.06
MARD = 0.5
r = 0.97
MRD = −0.1
MARD = 0.1
log10 (Salmon counts + 1)
log10 (Simulated counts + 1)
count
100000
1000
10
LAML sample
exon
intron
intergenic
Index #1 (wo Intron)
Index #2 (with Intron)
r = 0.88
MRD = −0.05
MARD = 0.3
r = 0.64
MRD = 0.6
MARD = 0.6
r = 0.98
MRD = −0.07
MARD = 0.09
r = 0.85
MRD = −0.08
MARD = 0.3
r = 0.77
MRD = 0.008
MARD = 0.4
r = 0.98
MRD = −0.08
MARD = 0.09
log10 (Salmon counts + 1)
log10 (Simulated counts + 1)
count
100000
1000
10

G


THCA sample: subfamily level
all
exon
intron
intergenic
Index #1 (wo Intron)
Index #2 (with Intron)
r = 1
MRD = −0.004
MARD = 0.06
r = 0.98
MRD = 0.02
MARD = 0.1
r = 0.92
MRD = 0.6
MARD = 0.6
r = 1
MRD = −0.02
MARD = 0.05
r = 0.99
MRD = −0.03
MARD = 0.05
r = 0.98
MRD = −0.01
MARD = 0.2
r = 0.9
MRD = 0.08
MARD = 0.3
r = 1
MRD = −0.03
MARD = 0.05
log10 (Salmon counts + 1)
log10 (Simulated counts + 1)
count
50
10
1
LAML sample: subfamily level
all
exon
intron
intergenic
Index #1 (wo Intron)
Index #2 (with Intron)
r = 1
MRD = −0.007
MARD = 0.05
r = 0.98
MRD = 0.009
MARD = 0.1
r = 0.93
MRD = 0.6
MARD = 0.6
r = 1
MRD = −0.02
MARD = 0.04
r = 1
MRD = −0.03
MARD = 0.04
r = 0.98
MRD = −0.02
MARD = 0.2
r = 0.92
MRD = 0.04
MARD = 0.3
r = 1
MRD = −0.03
MARD = 0.04
log10 (Salmon counts + 1)
log10 (Simulated counts + 1)

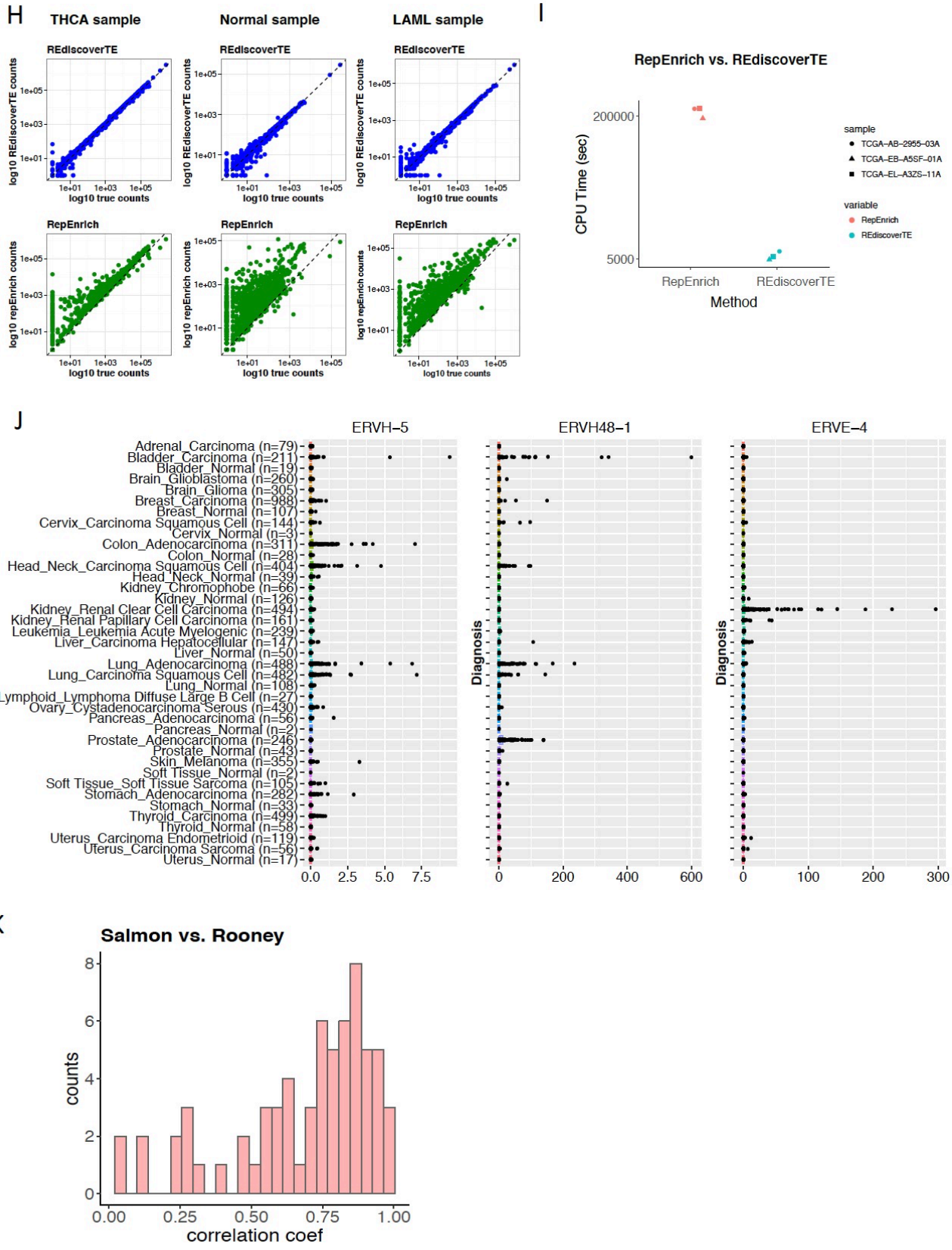

H
THCA sample
Normal sample
LAML sample
REdiscoverTE
RepEnrich
log10 REdiscoverTE counts
log10 repEnrich counts
log10 true counts
I
RepEnrich vs. REdiscoverTE
CPU Time (sec)
200000
5000
RepEnrich
REdiscoverTE
Method
sample
TCGA-AB-2955-03A
TCGA-EB-A5SF-01A
TCGA-EL-A3ZS-11A
variable
RepEnrich
REdiscoverTE
J
ERVH-5
ERVH48-1
ERVE-4
Diagnosis
Adrenal_Carcinoma (n=79)
Bladder_Carcinoma (n=211)
Bladder_Normal (n=19)
Brain_Glioblastoma (n=260)
Brain_Glioma (n=305)
Breast_Carcinoma (n=988)
Breast_Normal (n=107)
Cervix_Carcinoma Squamous Cell (n=144)
Cervix_Normal (n=3)
Colon_Adenocarcinoma (n=311)
Colon_Normal (n=28)
Head_Neck_Carcinoma Squamous Cell (n=404)
Head_Neck_Normal (n=39)
Kidney_Chromophobe (n=66)
Kidney_Normal (n=126)
Kidney_Renal Clear Cell Carcinoma (n=494)
Kidney_Renal Papillary Cell Carcinoma (n=161)
Leukemia_Leukemia Acute Myelogenic (n=239)
Liver_Carcinoma Hepatocellular (n=147)
Liver_Normal (n=50)
Lung_Adenocarcinoma (n=488)
Lung_Carcinoma Squamous Cell (n=482)
Lung_Normal (n=108)
Lymphoid_Lymphoma Diffuse Large B Cell (n=27)
Ovary_Cystadenocarcinoma Serous (n=430)
Pancreas_Adenocarcinoma (n=56)
Pancreas_Normal (n=2)
Prostate_Adenocarcinoma (n=246)
Prostate_Normal (n=43)
Skin_Melanoma (n=355)
Soft Tissue_Normal (n=2)
Soft Tissue_Soft Tissue Sarcoma (n=105)
Stomach_Adenocarcinoma (n=282)
Stomach_Normal (n=33)
Thyroid_Carcinoma (n=499)
Thyroid_Normal (n=58)
Uterus_Carcinoma Endometrioid (n=119)
Uterus_Carcinoma Sarcoma (n=56)
Uterus_Normal (n=17)
K
Salmon vs. Rooney
counts
correlation coef

Fig. S1. Overview of the human repetitive genome and benchmarking *REdiscoverTE*

A. Repeatmasker hierarchical classification (class, family and subfamilies) of human repetitive DNA. Repetitive elements (RE) can be categorized into transposable elements (TE, left) and non-transposable elements (right). Numbers to the right are total counts of subfamilies within each class.

B. Abundance of repetitive elements in the human genome and their physical locations in relations to host genes defined in Gencode (v26 basic). Upper pie-chart: the content of RE in human genome according to the Human Repeatmasker (hg38). Lower pie-chart: the relative footprint of genomic features (CDs, UTRs, introns, noncoding genes) defined in the Gencode transcriptome. Regions outside Gencode features are considered intergenic. Bar-plot: enrichment of RE DNA in different genomic context in terms of Gencode genomic features. Numerical score of 1 corresponds to no enrichment, <1 corresponds to depletion, >1 corresponds to enrichment.

C. *REdiscoverTE* benchmarking workflow: 1) Generate transcriptome for RSEM simulation; 2) RSEM learns of expression pattern from real RNA-seq data; 3) RSEM simulates new RNA-seq fastq based on learned and adjusted statistics; 4) Salmon quantification of RSEM simulated fastq; 5) Evaluate Salmon's performance. This workflow was carried out for two TCGA samples: one LAML sample and one THCA sample. TPM: transcript per kilobase million. Venn diagram on physical locations of RE DNA in relations to genes for all RE subfamilies except those that belong to the class of simple repeats. 1,135 of these RE subfamilies have elements located in all 3 genomic regions (exon, intron, intergenic).

D. Post-hoc profiling of RE-to-transcript abundance in simulated data. Left: distribution of exonic RE to transcript TPM fold change for transcripts containing REs. Right: distribution of intron retention rate. Red: LAML sample. Blue: THCA sample.

E. Accuracy of *REdiscoverTE* RE quantification: TPM vs. counts. Top: simulation based on a TCGA THCA sample. Bottom: simulation based on a TCGA LAML sample. Left two panels: simulated TPM vs. estimated TPM. Right two panels: simulated read counts vs. estimated read counts. Index #1: reference transcriptome without inclusion of introns. Index #2: reference transcriptome that includes all introns containing REs. Performance accuracy is measured in terms of Spearman correlation coefficient (r), mean relative difference (MRD), mean absolute relative difference (MARD).

F. Accuracy of *REdiscoverTE* RE quantification at the individual element level. Elements are categorized according to their genomic context in relations to genes into exonic, intronic and intergenic REs.

G. Accuracy of *REdiscoverTE* RE quantification where counts have been aggregated to the subfamily level.

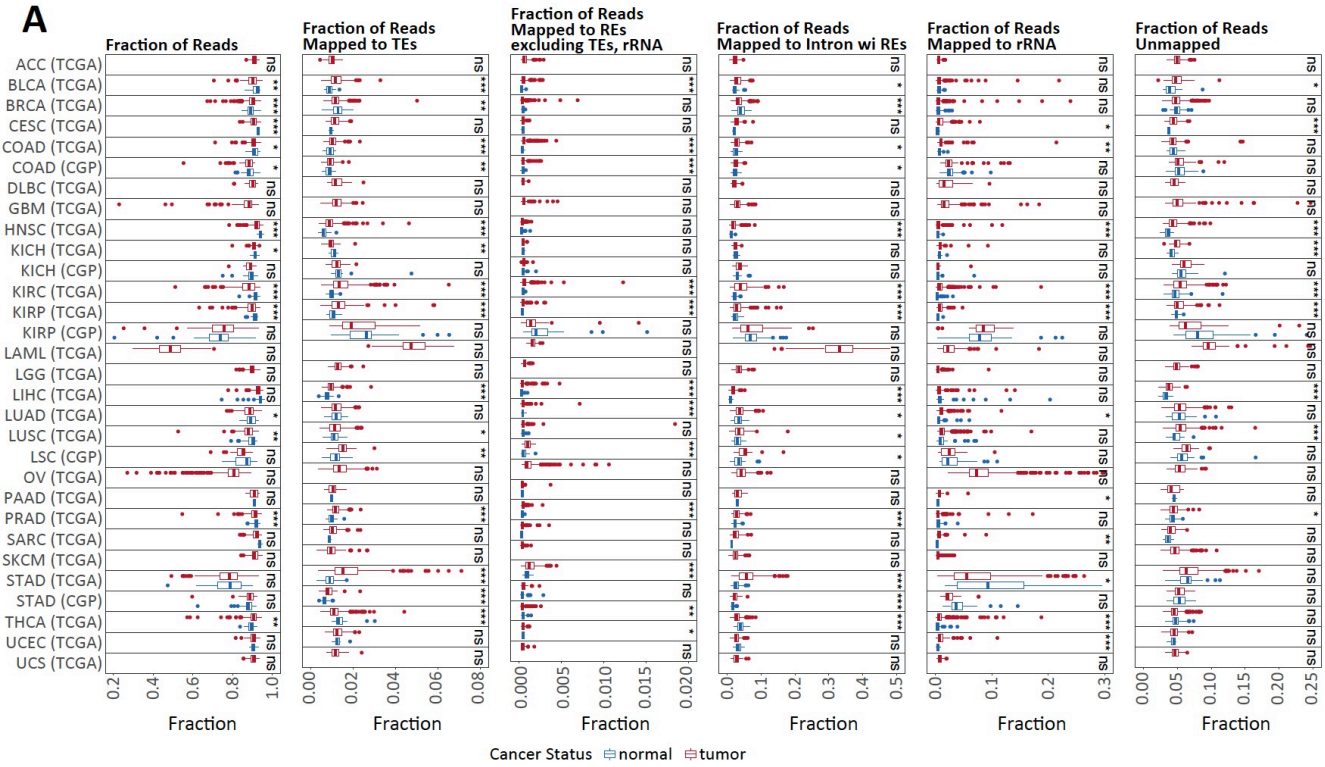
H. Comparison of TE quantification by *REdiscoverTE* to RepEnrich. Each point is one subfamily.

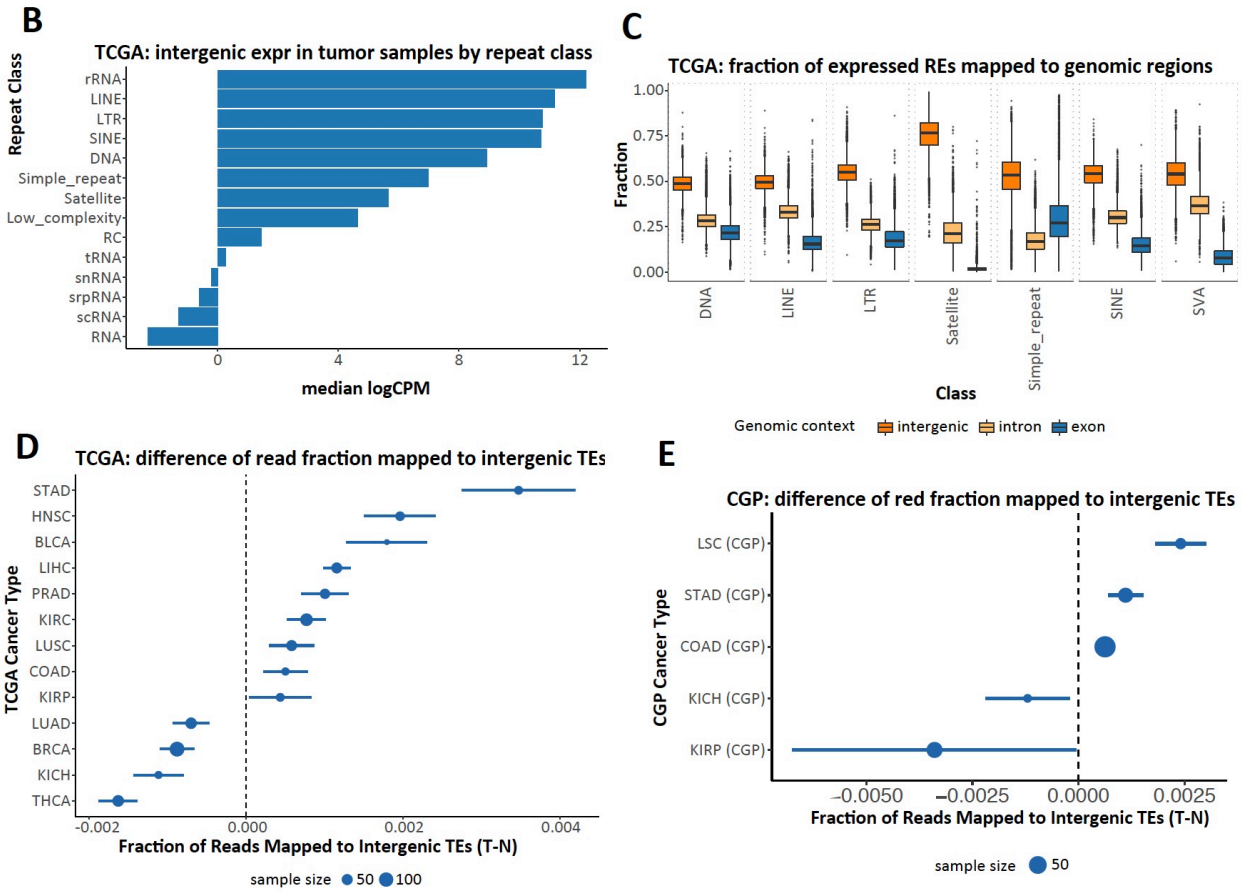
I. Compute time (in second) used by *REdiscoverTE* vs. *RepEnrich* to quantify the same fastq files using the same computer and memory resources.

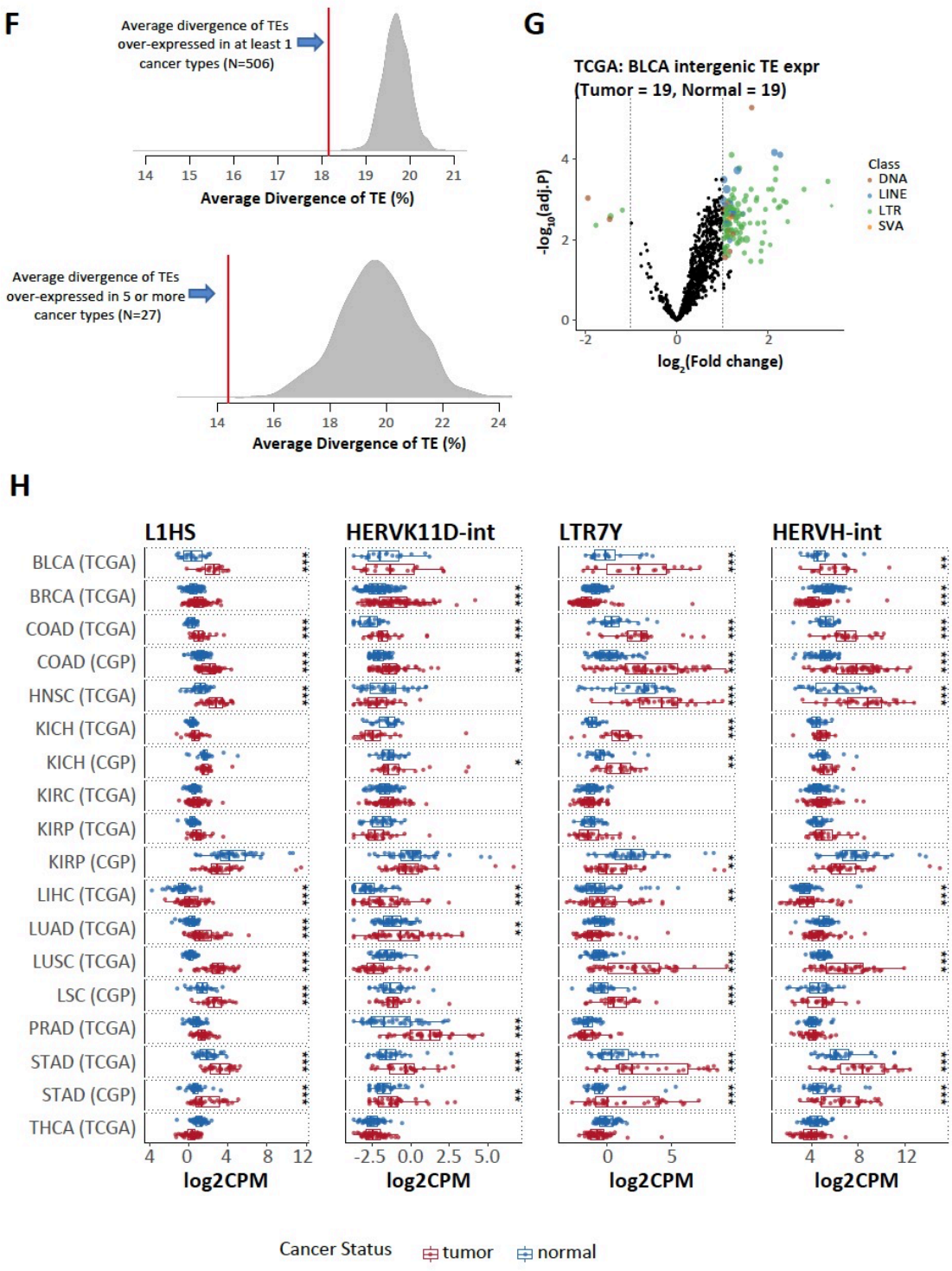
J. *REdiscoverTE* quantification of expression of 3 HERVs in TCGA RNA-seq data (compare to Rooney et al. Cell 2015 Fig4A)

K. Distribution of coefficients from Pearson correlation between *REdiscoverTE* and Rooney et al. Cell 2015 quantifications of 66 HERVs.

FigureS2. Characteristics of TE expression in Cancer




B
TCGA: intergenic expr in tumor samples by repeat class
Repeat Class
rRNA
LINE
LTR
SINE
DNA
Simple_repeat
Satellite
Low_complexity
RC
tRNA
snRNA
srpRNA
scRNA
RNA
0
4
8
12
median logCPM
C
TCGA: fraction of expressed REs mapped to genomic regions
Fraction
1.00
0.75
0.50
0.25
0.00
DNA
LINE
LTR
Satellite
Simple_repeat
SINE
SVA
Class
Genomic context
intergenic
intron
exon
D
TCGA: difference of read fraction mapped to intergenic TEs
TCGA Cancer Type
STAD
HNSC
BLCA
LIHC
PRAD
KIRC
LUSC
COAD
KIRP
LUAD
BRCA
KICH
THCA
-0.002
0.000
0.002
0.004
Fraction of Reads Mapped to Intergenic TEs (T-N)
sample size
50
100
E
CGP: difference of red fraction mapped to intergenic TEs
CGP Cancer Type
LSC (CGP)
STAD (CGP)
COAD (CGP)
KICH (CGP)
KIRP (CGP)
−0.0050
−0.0025
0.0000
0.0025
Fraction of Reads Mapped to Intergenic TEs (T-N)
sample size
50


F
Average divergence of TEs over-expressed in at least 1 cancer types (N=506)
14 15 16 17 18 19 20 21
Average Divergence of TE (%)
Average divergence of TEs over-expressed in 5 or more cancer types (N=27)
14 16 18 20 22 24
Average Divergence of TE (%)
G
TCGA: BLCA intergenic TE expr
(Tumor = 19, Normal = 19)
-log10(adj.P)
log2(Fold change)
Class
DNA
LINE
LTR
SVA
H
L1HS
HERVK11D-int
LTR7Y
HERVH-int
BLCA (TCGA)
BRCA (TCGA)
COAD (TCGA)
COAD (CGP)
HNSC (TCGA)
KICH (TCGA)
KICH (CGP)
KIRC (TCGA)
KIRP (TCGA)
KIRP (CGP)
LIHC (TCGA)
LUAD (TCGA)
LUSC (TCGA)
LSC (CGP)
PRAD (TCGA)
STAD (TCGA)
STAD (CGP)
THCA (TCGA)
log2CPM
Cancer Status
tumor
normal

I

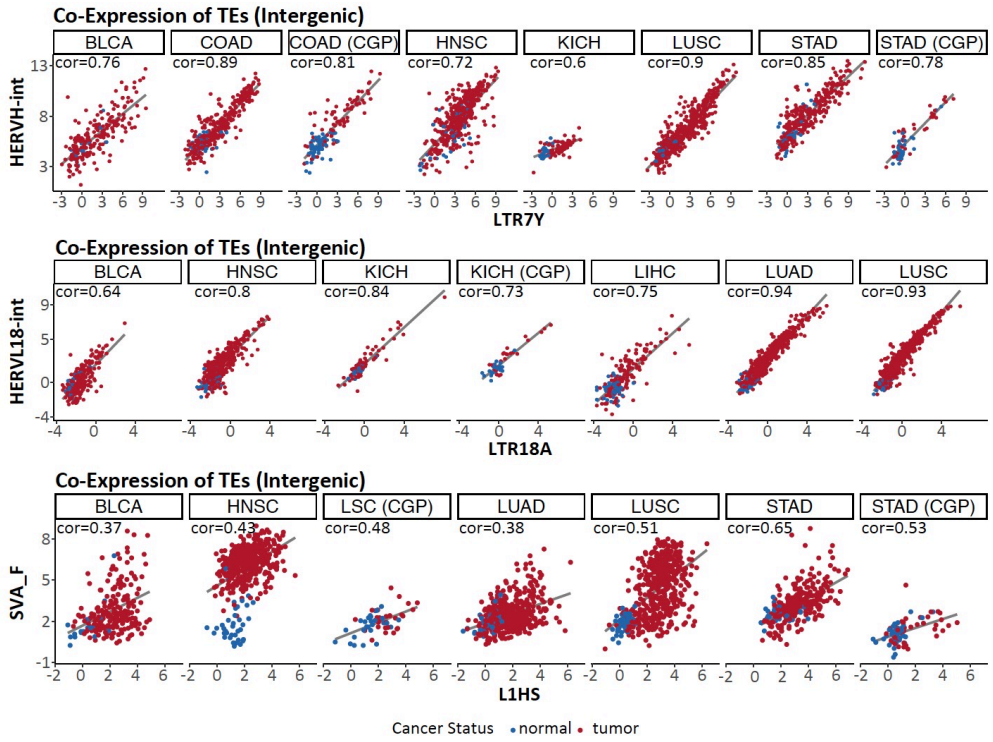


Fig. S2. Characteristics of TE expression in cancer

A. Fractions of TCGA and CGP RNAseq (both are poly-A preps) reads mapping to all features in the *REdiscoverTE* transcriptome (left to right): Gencode v26 basic transcripts, Repeatmasker TEs (across cancer types: median 1.1%, mean 1.3%), Repeatmasker REs (excluding TEs, rRNAs), Gencode RE-containing introns and rRNAs. The last column is the fraction of reads that remained unmapped.

B. RE intergenic expression in TCGA tumor samples from distinct repeat classes.

C. Genomic context of RE expression for top 7 non-rRNA repeat classes in Fig. S2B. All TCGA samples are used for this calculation. For each repeat class, the denominator is total number of reads mapped to that class.

D. Difference in fractions of RNAseq reads mapped to Intergenic TEs in tumor samples compared to matched normal samples across 13 TCGA cancer types (each with at least 10 normal sample). Error bars are standard errors of the mean.

E. Same as Fig. S2D, except for the 5 CGP cancer types. Error bars are standard errors of the mean.

F. Permutation analysis on the divergence of over-expressed TE subfamilies. Divergence of TE subfamilies is known to be inversely associated with age of TE. Top: 506 TEs over-expressed in at least 1 cancer type. Bottom: 27 TEs over-expressed in 5 or more cancer

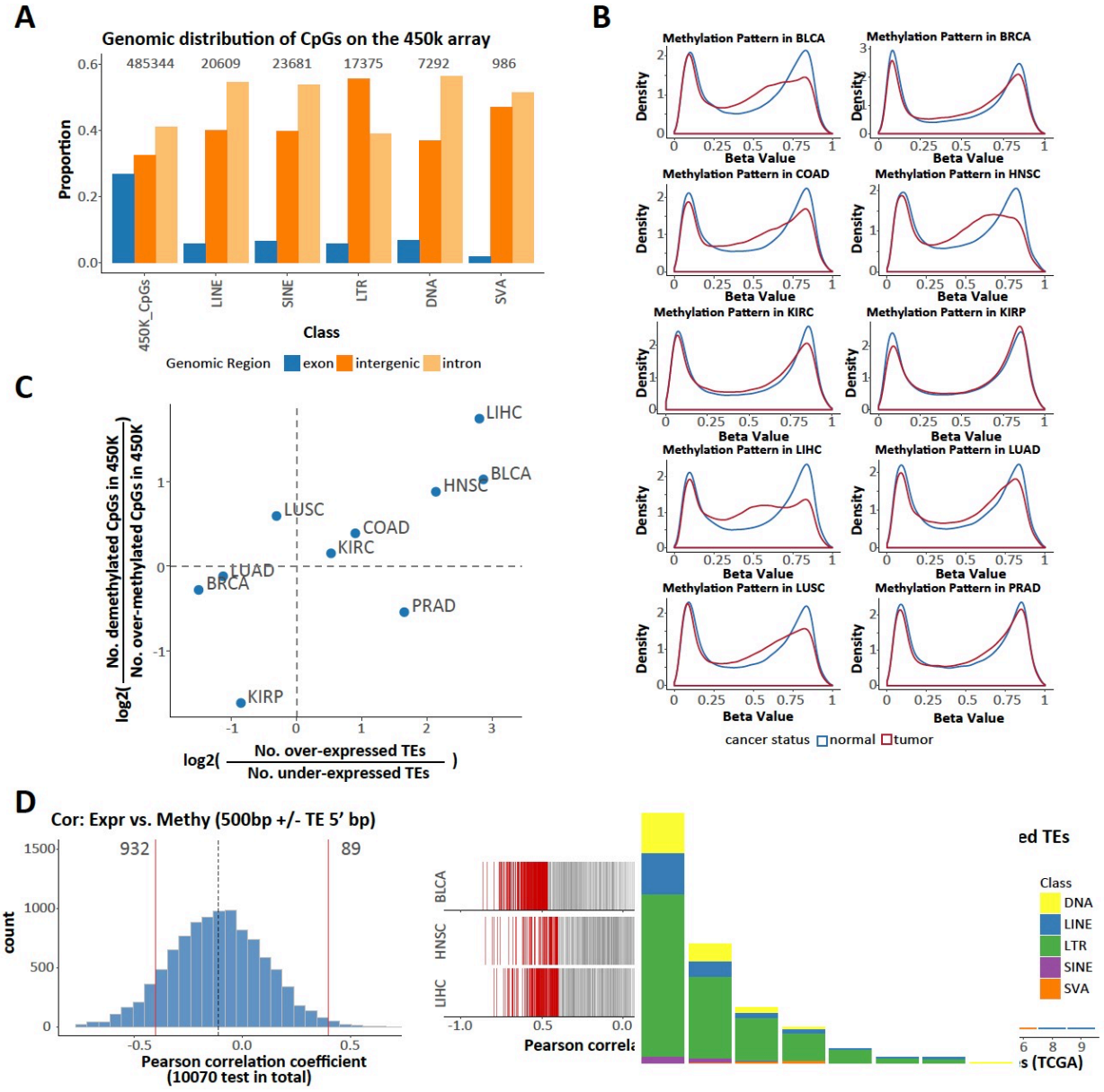
types. Red lines: observed average divergence value of over-expressed TEs. Gray distributions: bootstrapped distributions of mean divergence value of a random sample of TE subfamilies (matching number of subfamilies, 1000x permutations).

G. Example volcano plots of intergenic TE differential expression aggregated to the subfamily level performed on TCGA BLCA, 19 tumor and matched normal samples.

H. Patterns of differential expression for 4 TE subfamilies consistently over-expressed across cancer types in both TCGA and CGP. Only tumor and normal sample pairs are included here. Red: tumor samples. Blue: matched normal samples. Asterisks indicate level of significance in differential expression analysis between tumor and matched normal: * abs(log2 fold change) > 1 & FDR<0.05, ** abs(log2 fold change) > 1 & FDR<0.01, *** abs(log2 fold change) > 1 & FDR<0.001

I. Co-expression of three pairs of over-expressed TEs. Units in log2 CPM. Red: all tumor samples. Blue: available matched normal samples.

FigureS3. Association between TE expression and loss of DNA methylation in cancer



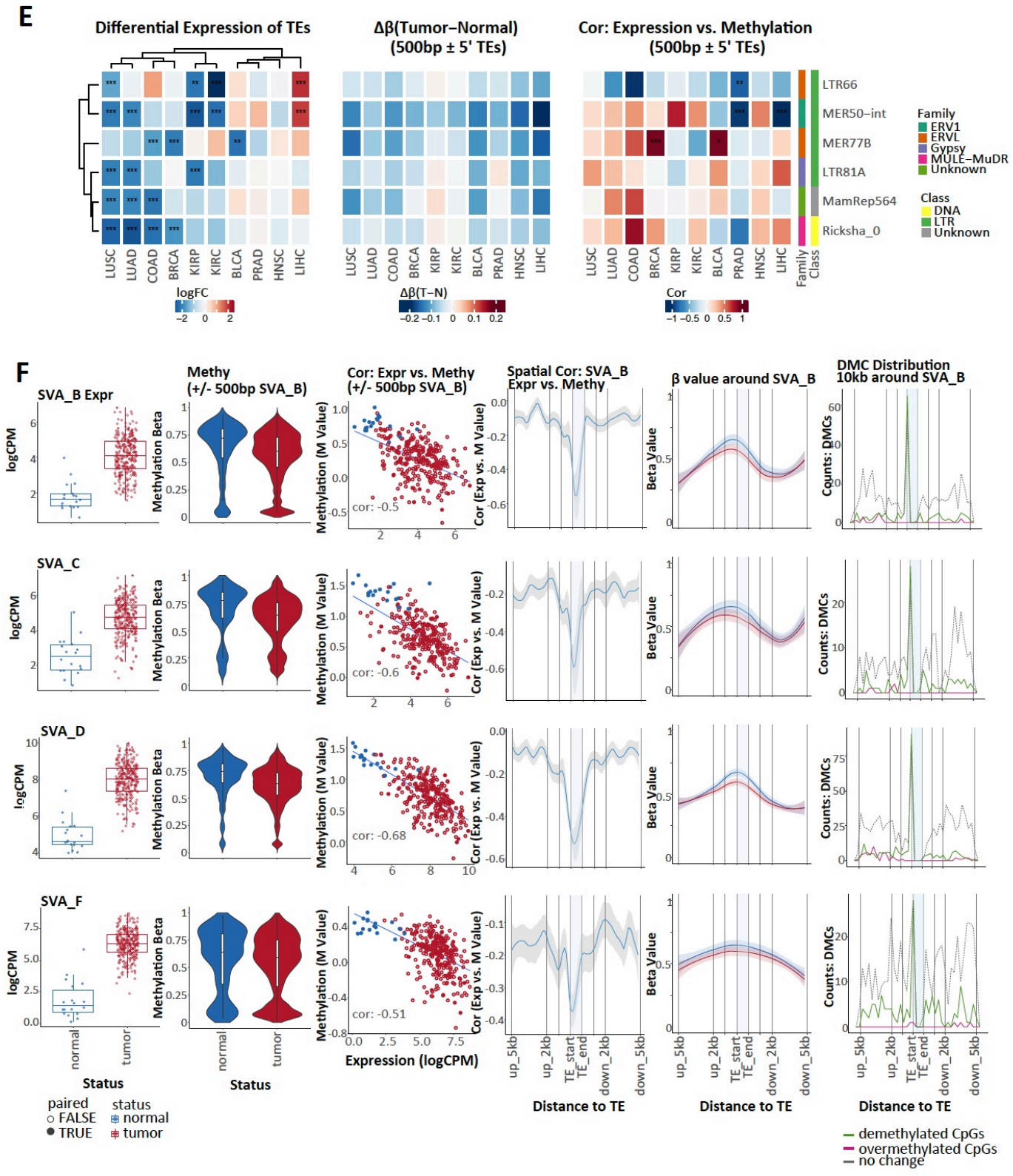


Fig. S3. Association between TE expression and loss of DNA methylation in cancer

A. Illumina 450K array coverage of CpG sites in relations to TE (by class) as well as to genes. Each set of 3 bars represent the fraction of all CpGs, either in its entirety (450K_CpGs) or overlapping with a particularly TE class (e.g. LINE) grouped by their physical location in relations to Gencode host genes (exon, intron and intergenic regions).

B. Global distribution of beta values of all 450k CpGs for 10 TCGA cancer types with methylation data. Only matched tumor-normal samples are used. Red: tumor samples. Blue: matched normal samples.

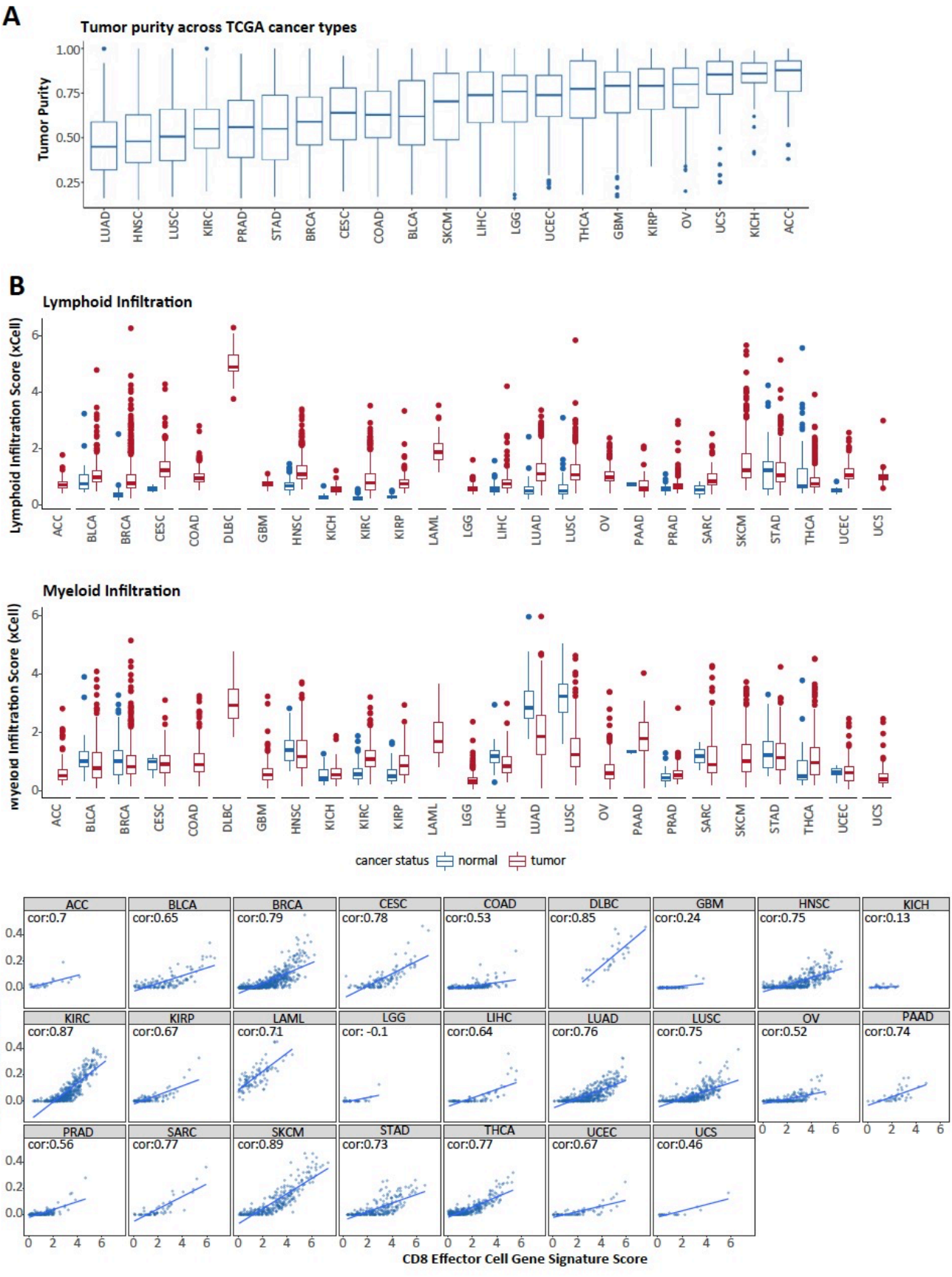
C. The extent of TE mRNA overexpression is strongly correlated with the extent of global CpG demethylation across cancer types.

D. Distribution coefficients of Pearson correlation between intergenic TE expression (N=1007 subfamilies) with average CpG methylation (M-value averaged over 500bp+/- 5' bp of the corresponding TEs) using matched tumor-normal samples across 10 cancer types. Significance threshold: cor < abs(0.4) and BH FDR<0.05. Left: pooled correlation coefficients for 10 TCGA cancer types. Median cor=-0.11. There are 932 significant inverse correlations and 89 positive correlations across the 10 cancer types. Middle: correlation coefficients for 3 most de-methylated cancer types: BLCA, HNSC, LIHC. Red lines indicate significant correlations. Right: Across 10 cancer types there were 431 unique TE subfamilies with significant inverse correlation between expression and methylation. Some TEs show inverse correlation in multiple cancer types. Histogram shows distribution on the recurrence of these inverse correlation.

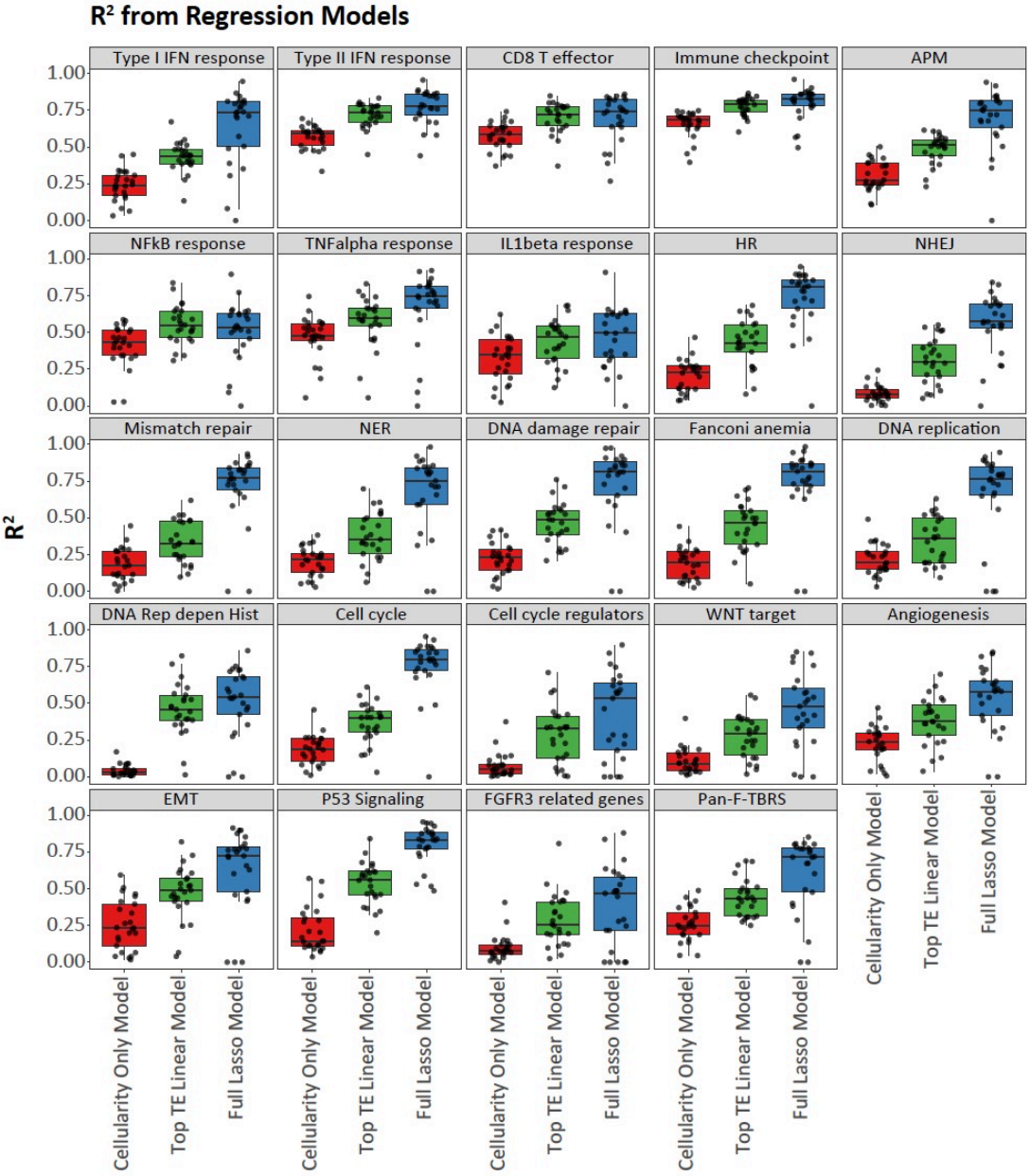
E. Examples of TE subfamilies with reduced-expression in tumor compared to matched normal and their CpG methylation status. Selection criteria: TE subfamilies showed significantly reduced expression in >=3 cancer types. Left: log2 FC values of tumor vs. normal differential expression of TEs (row) across indications (column). Middle: tumor - normal delta beta value at CpG 500bp+/- 5'bp of TE locations. Right: Correlation between intergenic TE expression and methylation M value at CpG 500bp +/- 5'bp of TE locations.

F. Examples from HNSC: expressions of 4 SVA subfamilies are associated with DNA methylation status. Blue: normal sample. Red: tumor samples. Filled circle: tumor samples with matched normal. Open circle tumor samples without matched normal. Grey shading: 95% confidence interval. Column 1: normal and tumor SVA intergenic expression. Column 2: normal and tumor CpG beta values in 500bp+/- around 5'bp of intergenic SVA. Column 3: correlation between SVA intergenic expression and methylation M value (500bp+/- 5'bp intergenic SVA). Column 4: spatial correlation between intergenic SVA expression and CpG methylation M value around 5kb +/- SVAs. SVA gene body is shaded in blue. Column 5: smoothed beta value in tumor and matched normal pairs in 5kb +/- region around SVA. Column 6: spatial distribution of differentially methylated cytosines (DMCs): demethylated CpG sites (green), over-methylated CpG sites(magenta) and CpGs with no change (grey, dashed) around 5kb +/- SVAs.

Figure S4 Characteristics of tumor gene expression profiles in relations to TE

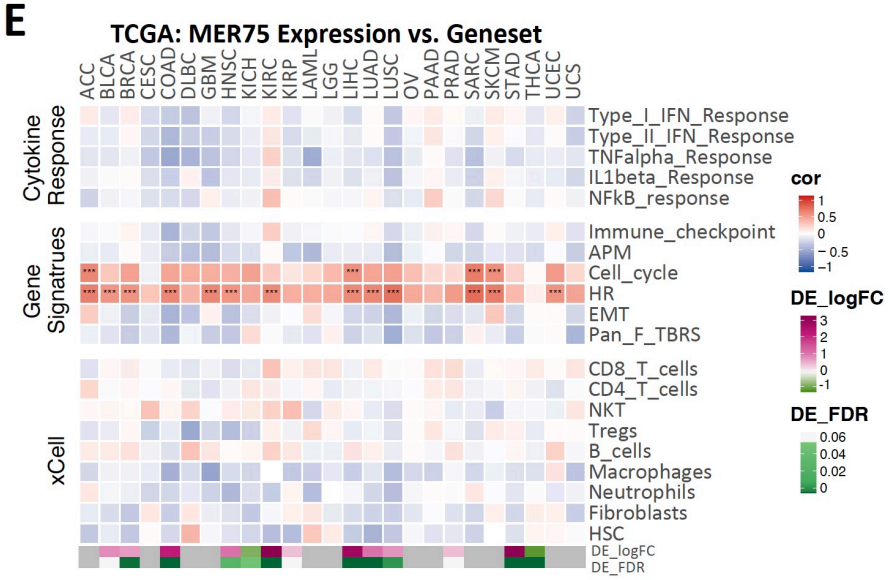


C

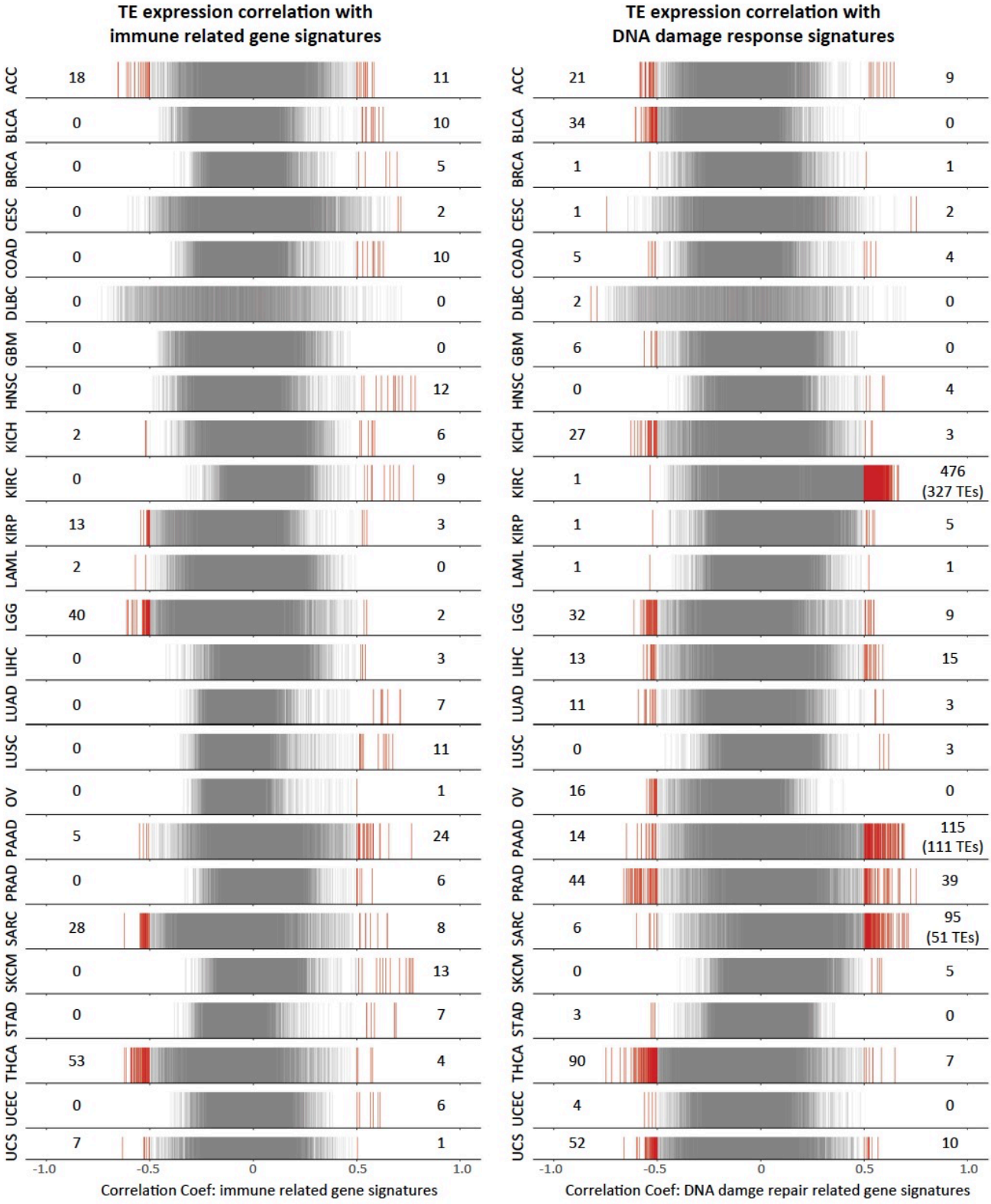

R² from Regression Models
Type I IFN response
Type II IFN response
CD8 T effector
Immune checkpoint
APM
NFkB response
TNFalpha response
IL1beta response
HR
NHEJ
Mismatch repair
NER
DNA damage repair
Fanconi anemia
DNA replication
DNA Rep depen Hist
Cell cycle
Cell cycle regulators
WNT target
Angiogenesis
EMT
P53 Signaling
FGFR3 related genes
Pan-F-TBRS
R²
1.00
0.75
0.50
0.25
0.00
Cellularity Only Model
Top TE Linear Model
Full Lasso Model


D
ACC BLCA BRCA CESC COAD DLBC GBM HNSC KICH KIRC KIRP LAML LGG LIHC LUAD LUSC OV PAAD PRAD SARC SKCM STAD THCA UCEC UCS type
Type I IFN Response
MamGypLTR2b
THE1C–int
myeloid
LTR21B
L1PBb
MER57F
Type II IFN Response
myeloid
lymphoid
purity
MER57F
LTR21B
HERVL74–int
NFkB Response
lymphoid
myeloid
purity
LTR21B
MER57F
HERVL74–int
TNFalpha Response
purity
myeloid
MER57F
MER61F
LTR21B
HERVL74–int
CD8 T Effector
lymphoid
myeloid
purity
MER57F
LTR21B
HERVL74–int
Immune Checkpoint
myeloid
lymphoid
purity
LTR21B
MER57F
HERVL74–int
APM
myeloid
Tigger1a_Mars
MER57F
lymphoid
ERV3–16A3_I–int
Charlie20a
IT1β Response
purity
myeloid
lymphoid
MER57F
MER61F
LTR54
Pan-F-TBRS
purity
Tigger12A
Charlie20a
lymphoid
LTR38B
Tigger2a
EMT
purity
myeloid
Charlie20a
Tigger12A
HERVL74–int
Tigger3
WNT Target
lymphoid
L1M2a
MER70–int
LTR48B
MER66B
LTR84a
FGFR3 Related Genes
lymphoid
HERV16–int
L1MEh
Charlie5
LTR1
LTR9B
HR
MER75
lymphoid
MER4A
MER54A
purity
MER67A
NHEJ
MER67A
HERVIP10B3–int
LTR7B
MER4A
MER75
MLT1F2–int
Mismatch Repair
MER75
lymphoid
purity
MER4A
MER54A
Charlie20a
NER
MER75
lymphoid
purity
MER4A
Charlie20a
MER54A
DDR
MER75
lymphoid
purity
MER4A
MER67A
MER54A
Fanconi Anemia
MER75
lymphoid
MER4A
purity
MER67A
MER54A
DNA Replication
MER75
lymphoid
MER4A
purity
MER67A
MER54A
DNA Rep dep Histones
LTR48B
MER51D
MER34
MER41E
LTR12C
L1ME3D
Cell Cycle
MER75
lymphoid
MER4A
MamGypLTR1d
MER67A
MER54A
Cell Cycle Regulator
LTR17
MER75
LTR27B
MLT1G1–int
HERVL74–int
MER45R
Angiogenesis
purity
Tigger3
myeloid
Tigger12A
lymphoid
LTR22
P53 Signaling
purity
myeloid
MER106A
LTR33
LTR9B
lymphoid
Lasso Coef
0.4
0.2
0
−0.2
−0.4
Variable Category
TE
Cellularity

E


TCGA: MER75 Expression vs. Geneset
ACC
BLCA
BRCA
CESC
COAD
DLBC
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
THCA
UCEC
UCS
Cytokine Response
Type_I_IFN_Response
Type_II_IFN_Response
TNFalpha_Response
IL1beta_Response
NFkB_response
Gene Signatrues
Immune_checkpoint
APM
Cell_cycle
HR
EMT
Pan_F_TBRS
xCell
CD8_T_cells
CD4_T_cells
NKT
Tregs
B_cells
Macrophages
Neutrophils
Fibroblasts
HSC
DE_logFC
DE_FDR
cor
1
0.5
0
−0.5
−1
DE_logFC
3
2
1
0
-1
DE_FDR
0.06
0.04
0.02
0

F


TE expression correlation with immune related gene signatures
TE expression correlation with DNA damage response signatures
ACC 18 11 21 9
BLCA 0 10 34 0
BRCA 0 5 1 1
CESC 0 2 1 2
COAD 0 10 5 4
DLBC 0 0 2 0
GBM 0 0 6 0
HNSC 0 12 0 4
KICH 2 6 27 3
KIRC 0 9 1 476 (327 TEs)
KIRP 13 3 1 5
LAML 2 0 1 1
LGG 40 2 32 9
LIHC 0 3 13 15
LUAD 0 7 11 3
LUSC 0 11 0 3
OV 0 1 16 0
PAAD 5 24 14 115 (111 TEs)
PRAD 0 6 44 39
SARC 28 8 6 95 (51 TEs)
SKCM 0 13 0 5
STAD 0 7 3 0
THCA 53 4 90 7
UCEC 0 6 4 0
UCS 7 1 52 10
-1.0 -0.5 0 0.5 1.0
Correlation Coef: immune related gene signatures
-1.0 -0.5 0 0.5 1.0
Correlation Coef: DNA damge repair related gene signatures

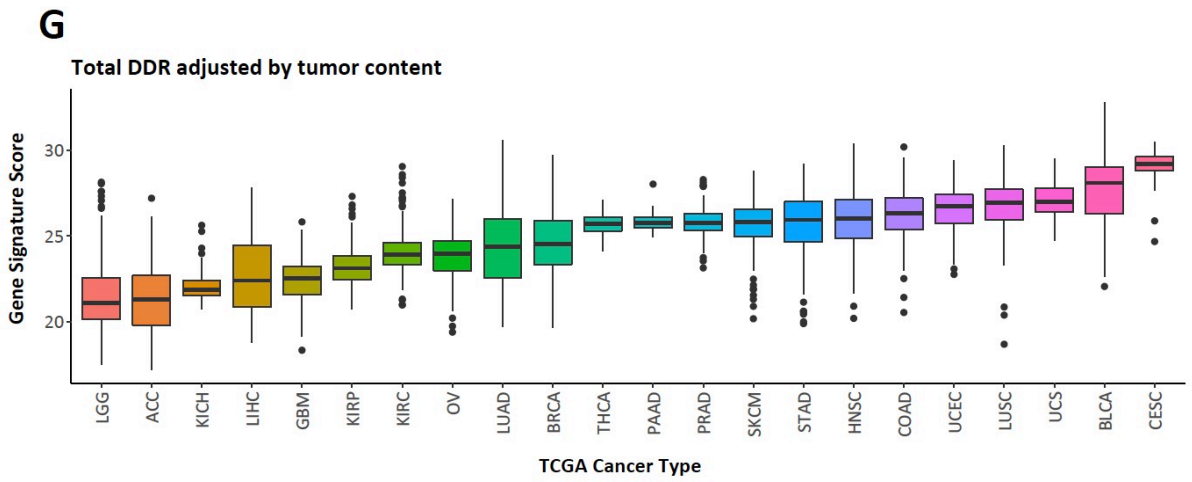


Fig. S4. Characteristics of tumor gene expression profile in relations to TE

A. Distribution of tumor purity score for TCGA samples by cancer type

B. Top panel and middle panels: total lymphoid and myeloid abundance in TCGA tumors, respectively, estimated based on xCell. Total lymphoid score is the sum of xCell scores of CD8+ T-cells, NK cells, CD4+ naive T-cells, B-cells, CD4+ T-cells, CD8+ Tem, Tregs, plasma cells, CD4+ Tcm, CD4+ Tem, memory B-cells, CD8+ Tcm, naive B-cells, CD4+ memory T-cells, pro B-cells, class-switched memory B-cells, Th2 cells, Th1 cells, CD8+ naive T-cells, NKT and Tgd cells. Total myeloid score is the sum of xCell scores of monocytes, macrophages, DC, neutrophils, eosinophils, macrophages M1, macrophages M2, aDC, basophils, cDC, pDC, iDC, mast cells. Bottom panel: correlation between xCell CD8+ T cell score and CD8+ Effector T cell geneset score estimated from multiGSEA. Each panel is one cancer type; each point is one sample. Samples with xCell score of 0 were omitted.

C. R^2 values from 3 regression models on 24 gene signatures in 25 cancer types. Each panel is one gene signature, each point is one cancer type. Red: R^2 from cellularity linear model which includes tumor content, total lymphoid and myeloid scores as predictors. Blue: R^2 of Lasso model which includes 3 aforementioned cellularity parameters and expression level of all 1,052 TEs as predictors. Green: R^2 from linear model taking top 6 TEs predicted by Lasso model and 3 cellularity parameters as predictors.

D. Graphical overview of top hits from Lasso models of gene signature scores across 25 TCGA cancer types. Only the top 6 variables predicting the gene signatures are included in the heatmaps. Top variables were selected based on rank order of the number of cancer types in which a given variable (e.g. a L1HS) had non-zero coefficients. Heatmap colors denote the value of the coefficient from Lasso model -- red and blue correspond to positive and negative coefficient values, respectively. Dots in the heatmap denote zero coefficients assigned by Lasso. Numbers in the heatmap indicate the rank of the absolute value of non-zero coefficients from the Lasso model for a given cancer type. Side bar

colors indicate whether a variable is a TE subfamily (blue) or one of three cellularities (magenta).

E. Heatmap showing association between MER75 expression and gene signatures as well as immune infiltrates estimated by xCell across 25 cancer types. MER75 expression is strongly associated with and DNA damage as well as cell cycle. Heatmap colors denote Spearman correlation coefficient. Differential expression status are denoted at the bottom. *** FDR<0.001, BH corrected.

F. Distribution of coefficients from Spearman correlations between the expression of 1052 TE subfamilies and gene signature scores across 25 TCGA cancer types. Left: pooled correlation coefficients from correlations with 8 immune related gene signatures (Type I IFN Response, Type II IFN Response, NFkB, TNFalpha, CD8 T Effector, Immune checkpoint, Antigen Processing Machinery, IL1b Response)
Right: pooled correlation coefficients from correlations with 6 DNA damage related gene signatures (NHEJ, Homologous Recombination, Mismatch repair, Nucleotide excision repair, DNA damage repair, Fanconi anemia). Correlation was calculated using only tumor samples and controlled for tumor content. Red: significant and strong correlations. Significance threshold: abs(cor)>0.5 & FDR<0.05, BH corrected. Numbers indicate the number of significant correlations except those in parenthesis indicate number of unique TE subfamilies with significant correlations.

G. Distribution of tumor DNA damage response (DDR) scores, adjusted for tumor purity scores. DDR scores are computed as sum of 6 DNA damage related gene signature scores: homologous recombination, NHEJ, DNA damage repair, Fanconi anemia, nucleotide excision repair and mismatch repair (**Table S4**). Adjusted total DDR is the intercept plus residual of linear regression of total DDR score on tumor purity score.

Figure S5. 5-aza-2'-deoxycytidine treatment of GBM cell lines induces TE expression

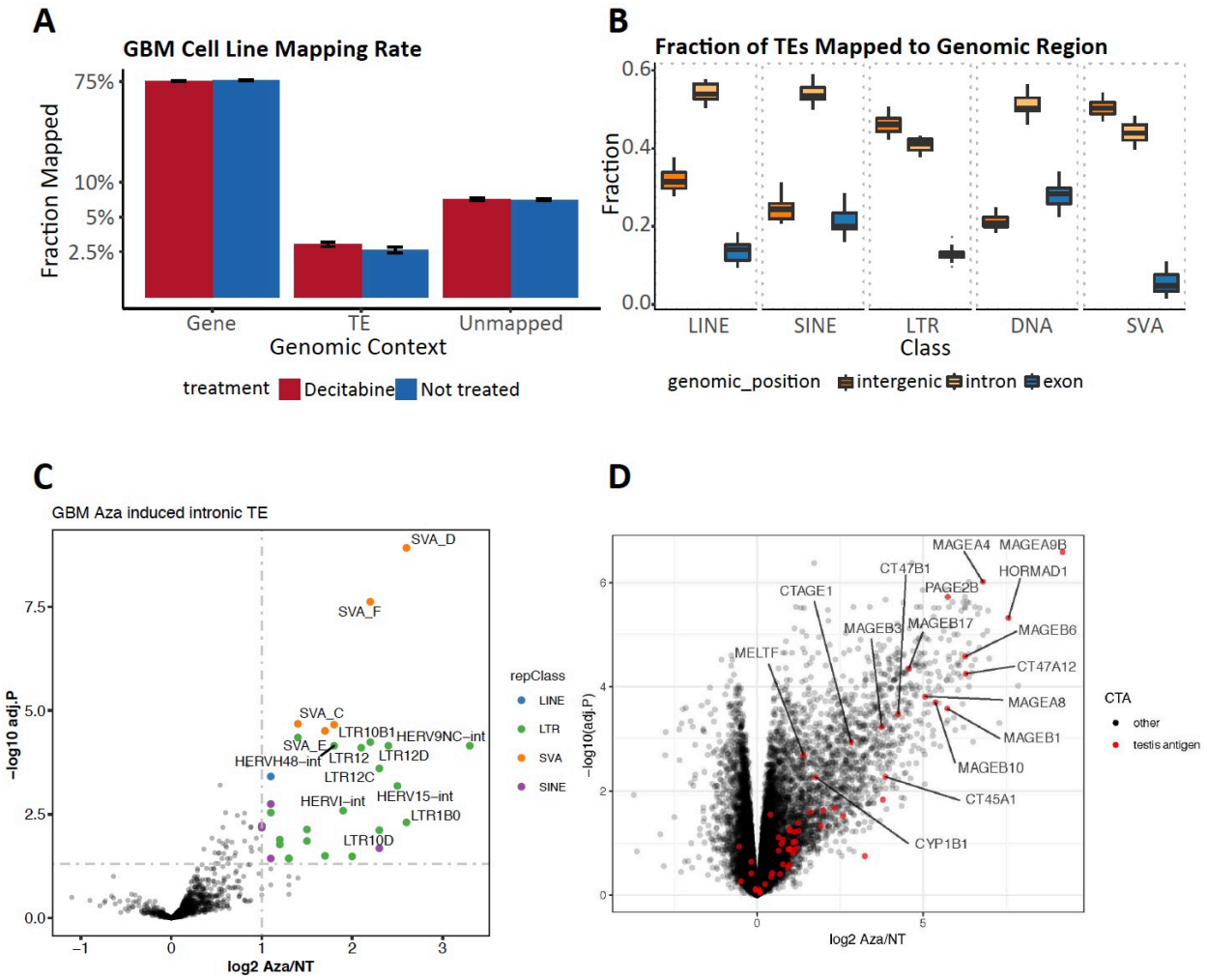


Fig. S5. 5-aza-2'-deoxycytidine treatment of GBM cell lines induces TE expression

A. Fractions of RNA-seq output (rRNA depletion prep) corresponding to reads mapped to Gencode genes, Repeatmasker TEs and unmapped reads. Error bars denote standard error over 12 samples.

B. Fraction of RNA-seq reads mapped to intergenic, intronic and exonic TE elements for each of 5 TE classes.

C. Volcano plot showing differential intronic expression of TE subfamilies, decitabine-treated (Aza) vs. non-treated (NT). TE subfamilies are colored by class at the significance threshold of log2FC > 1 and adjusted p-value < 0.05 and labeled if log2FC>1.5 and adjust p-value < 0.01.

D. Decitabine treatment results many cancer testis antigen. Aza vs. NT volcano plots showing differential expression of Gencode genes. Red: select cancer testis antigens.