

8-1-2018

## Expression Of Line-1 In Human Somatic Tissues And The Factors Correlated With Line- 1 Expression

G. M. Jonaid  
gmdb97@gmail.com

Follow this and additional works at: <https://digitalscholarship.unlv.edu/thesesdissertations>



Part of the [Bioinformatics Commons](#), and the [Biology Commons](#)

---

### Repository Citation

Jonaid, G. M., "Expression Of Line-1 In Human Somatic Tissues And The Factors Correlated With Line- 1 Expression" (2018). *UNLV Theses, Dissertations, Professional Papers, and Capstones*. 3363.  
<https://digitalscholarship.unlv.edu/thesesdissertations/3363>

This Thesis is protected by copyright and/or related rights. It has been brought to you by Digital Scholarship@UNLV with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Thesis has been accepted for inclusion in UNLV Theses, Dissertations, Professional Papers, and Capstones by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact [digitalscholarship@unlv.edu](mailto:digitalscholarship@unlv.edu).

EXPRESSION OF LINE-1 IN HUMAN SOMATIC TISSUES AND THE FACTORS CORRELATED WITH  
LINE-1 EXPRESSION

By

GM Jonaid  
Bachelor of Science – Microbiology  
University of Dhaka  
2012

A thesis submitted in partial fulfillment  
of the requirements for the

Master of Science – Biological Sciences

School of Life Sciences  
College of Sciences  
The Graduate College

University of Nevada, Las Vegas  
August 2018



## **Thesis Approval**

The Graduate College  
The University of Nevada, Las Vegas

August 17, 2018

This thesis prepared by

G.M. Jonaid

entitled

Expression Of Line-1 In Human Somatic Tissues And The Factors Correlated With Line-1 Expression

is approved in partial fulfillment of the requirements for the degree of

Master of Science – Biological Sciences  
School of Life Sciences

Mira Han, PhD  
*Examination Committee Chair*

Kathryn Hausbeck Korgan, Ph.D.  
*Graduate College Interim Dean*

Martin Schiller, PhD  
*Examination Committee Member*

Philippos Tsourkas, PhD  
*Examination Committee Member*

Qing Wu, PhD  
*Graduate College Faculty Representative*

## ***Abstract***

Despite the long-held assumption that transposons are normally only expressed in the germ-line, recently we discovered that full length or partial transcripts of LINE1 are frequently found in the somatic cells. However, the extent of variation in LINE1 levels across different tissues and different individuals, and the genes and pathways that are co-expressed with LINE1 are unknown. Co-expressed genes may be candidate genes that are functioning in transposon silencing. Here, we report the extent of variation in L1HS expression levels across cancer tissues and healthy tissues collected for The Cancer Genome Atlas (TCGA). L1HS is overexpressed in most of the cancer types we have studied. Our results confirm earlier reports of higher L1HS expression in the esophagus and stomach tissues. We also show that mitochondrial genes are enriched among the genes whose expression is negatively correlated with L1HS expression and that PHD fingers, bromodomains and KRAB-zinc fingers (KRAB-ZFPs) are enriched among the genes positively co-expressed with L1HS. Additionally, we studied the association of L1HS transcript level with miRNA expression, and we found several candidate miRNAs that are significantly correlated with L1HS expression.

# ***Table of Contents***

<b><i>Abstract</i></b>	iii
<b><i>List of Figures</i></b>	vi
<b><i>List of Tables</i></b>	vii
<b><i>Chapter 1: Introduction</i></b>	1
<b><i>Chapter 2: Differential expression of TEs</i></b>	9
2.1 Methods	9
2.1.1 Data	9
2.1.2. RNA Seq:	12
2.1.3 TEtranscripts	13
2.1.4 Normalization	14
2.1.5 DESeq2	16
2.1.6 edgeR	17
2.2 Results	18
2.2.1. Retrotransposon expression varies across different tissue types and individuals	18
2.2.2 L1HS consistently over-expressed in multiple cancer types	20
2.2.3. HTSeq comparison	35
2.2.4. TCGA vs STAR	38
2.2.5. DESeq2 vs edgeR	43
2.2.6. Conclusion	44
<b><i>Chapter 3: Gene and L1HS coexpression</i></b>	45
3.1 Introduction	45
3.2 Methods	47
3.2.1 Linear Regression	47
3.2.2. REC score	48
3.3 Results	49
3.3.1 Positively correlated expression of Genes with the L1HS expression	49
3.3.2 Correlation with other TE families	56
3.3.3 Mitochondrial genes are enriched in negatively correlated genes with L1HS	58

3.3.4 Conclusion	60
<b>Chapter 4: miRNA and L1HS coexpression</b>	61
4.1 Introduction	61
4.2 Methods	61
4.2.1 Data	61
4.2.2. Linear Regression	63
4.2.3. Linear Mixed Model	64
4.2.4. REC Scores	65
4.2.5. TargetScan	65
4.2.6. MiRanda	67
4.3 Results	68
4.3.1. Several miRNAs are correlated with L1HS	68
4.3.2. Mixed Linear Model predicts 38 potential miRNAs	71
4.3.3. REC Scores did not predict significant miRNAs	74
4.3.4. TargetScan predicts binding site of L1HS in several miRNAs	75
4.3.5. Miranda predicts binding site of L1HS in several miRNAs	79
4.4. Conclusion	82
<b>Chapter 5: Reference</b>	83
<b>Chapter 6: Curriculum Vitae</b>	94

## ***List of Figures***

Figure 1: RNA seq workflow implemented by tcga	8
Figure 2: RNA seq method	12
Figure 3: L1hs expression in different normal tissues..	18
Figure 4: Frequency of l1hs expression over different individuals.	19
Figure 5: Expression of l1hs in different type of cancer tissues	20
Figure 6: L1hs expression in different types of cancer..	22
Figure 7: Expression of retrotransposons in different cancer tissues..	34
Figure 8: Comparison between htseq and tetra transcripts for raw counts of tes. .	36
Figure 9: Comparison of htseq and tetra transcripts for raw count reads of genes..	37
Figure 10: Comparison of gene counts between star pipeline and tcga pipeline for a patie.....	39
Figure 11: Comparison of gene counts between star pipeline and tcga pipeline for a.....	40
Figure 12: Comparison of gene counts between star pipeline and tcga pipeline for a.....	41
Figure 13: Comparison of gene counts between star pipeline and tcga pipeline for a.....	42
Figure 14: L1hs expression in different types of tissues ( edgeR results)..	43
Figure 15: Genes positively correlated with l1hs expression in multiple cancer types	50
Figure 16: Genes negatively correlated with l1hs expression in multiple cancer types	51
Figure 17: Linear mixed model fit mirna hsa-mir-29b-2.	71
Figure 18: Target Site for hsa-mir-138-2-5p.....	79

## ***List of Tables***

Table 1: Number of patients in different type of cancers with corresponding healthy tissue	11
Table 2: DAVID enrichment score	49
Table 3: Top genes predicted by REC score	55
Table 4: REC score of L1HS with another repeat family	57
Table 5: Mitochondrial genes negatively correlated with L1HS	59
Table 6: Mixed model	64
Table 7: Linear Regression Results. Per miRNA, we performed a linear regression....	68
Table 8: Linear Regression Results. Per miRNA we performed a linear regression.....	69
Table 9: Linear Regression Results. Per miRNA we performed a linear regression.....	70
Table 10: List of Significant MiRNAs. MiRNAs with significant correlation based.....	73
Table 11: REC Scores Results in Python. This list contains miRNA with the highest REC scores.	74
Table 12: List of miRNA transcripts from the short CDSs.....	76
Table 13: List of miRNA transcripts from the long CDS.....	78
Table 14: List of miRNA transcript 8-mer for short CDS. This list all the miRNAs.....	79
Table 15: List of miRNA transcript 8-mer for long CDS. This list all the miRNAs.....	79
Table 16: microRNA predicted by the mixed model, Targetscan, Rec score and MiRanda	80



## ***Chapter 1: Introduction***

Transposable elements, discovered by Barbara McClintock in the 1940s, are genetic elements that can multiply themselves in a genome (Cordaux & Batzer, 2009; Fedoroff, 2012). Transposable elements can be classified into two categories: DNA transposon and retrotransposon. DNA transposons can cut themselves from the genome and paste themselves in other regions of the genome. On the other hand, retrotransposons multiply themselves by copy-paste mechanism through RNA-intermediates that are reverse-transcribed and inserted at new genomic locations. Retrotransposons can be divided into two main categories: LTR (long terminal repeats) and non-LTR.

LTR retrotransposons have a direct repeat of a few hundred base pairs long at each end. LTR retrotransposons are 5-7 kb long. LTR retrotransposons encode two open reading frames (ORFs): one ORF is equivalent to viral structural proteins named *gag* and the other, named as *pol*, is a protein consist of an aspartic protease (AP), a reverse transcriptase (RT), an RNase H and an integrase (INT) (Cordaux & Batzer, 2009).

Non-long terminal repeat retrotransposons (non-LTR) exist in most eukaryotic genomes and they are the most abundant genome sequences in the human genome (Cordaux & Batzer, 2009). Non-LTRs do not carry LTR at the ends of their genome; instead, they act like integrated mRNA. Non-LTRs can be divided into long interspersed elements (LINEs) and short interspersed elements (SINEs). Among the non-LTRs, L1, Alu, and SVA are the only transposable elements are

shown to be currently active in replicating in the human genome (Hancks & Kazazian, 2012; Mills, Bennett, Iskow, & Devine, 2007). Non-LTRs have two open reading frames, one encoding an RNA binding protein and the other encoding a nuclease, a reverse transcriptase and in some cases an RNase H domain (Mathias, Scott, Kazazian, Boeke, & Gabriel, 1991).

Retrotransposons affect the human genome by inserting mutations, causing genomic instability and variations in gene expression (Cordaux & Batzer, 2009). Although retrotransposons exist in a vast region of the human genome, they are inactive, mostly controlled by epigenetic regulations such as DNA methylation and RNAi mechanisms that restrict expression (Slotkin & Martienssen, 2007). But, when human DNA is hampered by some types of genomic instability such as cancer, retrotransposons become active in the human genome, partly due to ineffective regulation by host genome. LINEs, particularly the L1 family, have been reported to be activated in numerous cancer types such as colorectal carcinoma, breast carcinoma, liver hepatocellular carcinoma, etc (Beck, Garcia-Perez, Badge, & Moran, 2011; Callinan & Batzer, 2006; Hancks & Kazazian, 2012). SINEs, especially the *alu* family, have been reported to cause genomic instability in various types of cancers as many studies investigated genome-wide tracking of *alu* repeats in cancer and normal cell lines.

LINE-1 (L1) transposable elements (TEs) comprise more than 17% of the human genome (Cordaux & Batzer, 2009). In addition to generating insertion mutagenesis, it also causes damage to the cell through the aberrant expression of its sequence (Belgnaoui, Gosden, Semmes, & Haoudi, 2006). The L1 expression can cause insertion mutations, genomic

instability, alterations in gene expression and genetic innovation (Belgnaoui et al., 2006; Cordaux & Batzer, 2009). Insertional mutagenesis by L1 and other retroelements can cause numerous types of human disease; for instances, L1, *Alu*, and SVA insertions have been detected in haemophilia, cystic fibrosis, Apert syndrome, neurofibromatosis,  $\beta$ -thalassemia, hypercholesterolemia, breast and colon cancer (Beck, Garcia-Perez, Badge, & Moran, 2011; Callinan & Batzer, 2006). Expression of one of the L1 proteins, ORF2, in human cancer cells can cause DNA double-strand breaks (Gasior, Wakeman, Xu, & Deininger, 2006). Furthermore, the L1 expression has been reported to induce cell cycle arrest and apoptosis (Belgnaoui et al., 2006).

LINE-1 has long been thought to be expressed only in the germline cells (Kano et al., 2009), but both full-length and partial transcripts of LINE-1 are frequently found in the somatic cells with large variation in expression levels across tissue types (Belancio, Roy-Engel, Pochampally, & Deininger, 2010). The level of L1 expression is especially pronounced in cancer cells (Belgnaoui et al., 2006; Bratthauer, Cardiff, & Fanning, 1994; Rodić et al., 2014; Xiao-Jie, Hui-Ying, Qi, Jiang, & Shi-Jie, 2016). One of the earliest reports of detection of L1 expression outside the germline cells was in teratocarcinoma (Skowronski, Fanning, & Singer, 1988). Since then, overexpression of L1s have been identified in numerous types of cancer (Bratthauer et al., 1994; Rodić et al., 2014), and some even consider L1 expression as one of the hallmarks of cancer (Rodić et al., 2014) or L1 activity as a driver of tumorigenesis (Rodić et al., 2014).

Although there are many reports of L1 expression in cancer, how the L1 expression is repressed in human somatic cells and de-repressed in cancer cells is still largely unknown. Based on what we know from model organisms, *Drosophila melanogaster*, and *Mus musculus*, the L1 expression is regulated through several silencing mechanisms including DNA methylation, Eset-mediated histone modification, endo-siRNAs, miRNAs, and posttranslational modification (Ghildiyal et al., 2008; Martin & Branciforte, 1993; Reichmann et al., 2012).

To study L1 activity, *M. musculus* is used as a model because it contains L1s that are found in mammals. Although L1s are relatively inactive, a synthetic L1 element can be used to track retrotransposon activity. Human and *M. musculus* L1s can transpose between each organism, which makes for a comparative model (An et al., 2006).

*M. musculus* embryonic stem cells are standard model to analyze L1 expression because they can be monitored through different methods including microarray technology through analyzing repeat elements. A histone deacetylase, Hdac1, is a regulator for L1 expression in embryonic stem cells (Reichmann et al., 2012). Computational approaches identified other mechanisms that silence retrotransposon expression including DNA methylation, ESET-mediated histone modification, and Ring1B/Eed-containing polycomb repressive complexes (Reichmann et al., 2012). Analysis of these varying mechanisms shows that different transposons are regulated by different mechanisms and they are utilized simultaneously to regulate the transposable element expression within *M. musculus*. SiRNA found in *D. melanogaster* can be generated exogenously or endogenously, but endo-

siRNAs are believed to regulate L1 expressions, not only in *D. melanogaster* but also humans (Ghildiyal et al., 2008). SiRNA interacts with RNAi to regulate gene expression. SiRNA is comparable to miRNA because both silence gene expression (Ghildiyal et al., 2008). Another mechanism by which L1 activity is repressed is by RNA-induced Silencing Complex (RISC) and Piwi-interacting RNA (piRNA) pathway. Small interfering RNA (siRNA) (Ghildiyal et al., 2008) and miRNA are known to repress the movement of retrotransposons (“Transposable elements and miRNA: Regulation of genomic stability and plasticity: Mobile Genetic Elements: Vol 6, No 3,” n.d.). One of the miRNAs that are cited by several literatures to repress L1 retrotransposition is mir-128 (Hamdorf et al., 2015; M. Li et al., 2013).

Recent studies have found that L1 activity is restricted by several mechanisms including DNA methylation (Nagamori et al., 2015; Zamudio et al., 2015) and post-transcriptional regulation. APOBEC proteins participate in post-transcriptional regulation by inhibiting L1 retrotransposition (Chen et al., 2006; Wissing, Montano, Garcia-Perez, Moran, & Greene, 2011). Another protein named MOV10, an RNA helicase, also inhibits L1 retro-activity by conjugating with RNPs (Arjan-Odedra, Swanson, Sherer, Wolinsky, & Malim, 2012; Goodier, Cheung, & Kazazian, 2012, p. 10; X. Li et al., 2013). In many cases, some transcription factors can promote L1 retrotransposition. Some of these transcription factors are RUNX3, SOX2, SP1, and YY1, etc. Methylation of retrotransposon promoter transcriptionally represses retrotransposition to control expression in the human somatic cell.

In order to understand how L1 expression is regulated in human somatic cells, we utilized the large variation found in L1 expression levels in cancer cells across cancer types and across individual patients. We tested for co-expression between individual genes/miRNAs and L1s.

### **TCGA data**

The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. TCGA has collected genomic data for more than 11,000 patients for 33 types of cancer.

TCGA has following data available in TCGA data portal:

- Clinical information about participants in the program
- Metadata about the samples (e.g. the weight of a sample portion, etc.)
- Histopathology slide images from sample portions
- Molecular information derived from the samples (e.g. mRNA/miRNA expression, protein expression, copy number, etc.)

In addition to collecting high-quality tumor samples, TCGA also collects high-quality non-tumor samples. The purpose is to compare the abnormalities associated with cancer compared to healthy tissues.

- For most disease studies, TCGA collect and analyze normal blood samples for the majority of participants with that specific diseases
- In case of missing normal blood sample, TCGA collects normal tissue sample from that specific patient to use for germline control in DNA assays

- But, a normal blood sample cannot be used for RNA assay as RNA profile differs in different tissues. To address this issue, TCGA collects normal tissue sample for that specific organ. Ultimately, this procedure is useful for gene expression analysis in cancer versus normal tissues.

TCGA has the data available for de-identified clinical and demographic data, gene expression data, copy number alterations in regions of the genome, epigenetic data, anonymized single amplicon DNA sequence data, primary sequence data (BAM and FASTQ files), SNP6 array level 1 and level 2 data, exon array level 1 and level 2 data and VCFs.

TCGA pipeline:

First, TCGA samples are collected by Tissues Source Sites (TSS) and then sent to Biospecimen Core Resources (BCRs). Second, The BCRs submit clinical data and metadata to the Data Coordinating Center (DCC) and analytes to the Genome Characterization Centers (GCCs) and Sequencing Centers (GSCs), where mutation calls are generated and then submitted to the DCC. Third, GSCs submit trace files, sequences and alignment mappings to the Cancer Genomics Hub (CGHub) as well. Data submitted to the DCC and CGHub are made available to the research community and Genome Data Analysis Centers (GDACs). Finally, Analysis pipelines and data results produced by GDACs are served to the research community via the DCC.

## Genomic Data Commons (GDC):

GDC is a research program for the National Cancer Institute. It stores the genomic datasets from TCGA. It has access to standardized bio specimen, clinical and molecular data. GDC also generates high-level data based on Human reference genome build GRCh38 which includes germline mutations, somatic mutations, RNA sequence and quantifications, miRNA sequences and qualifications, and SNP based copy number variations.

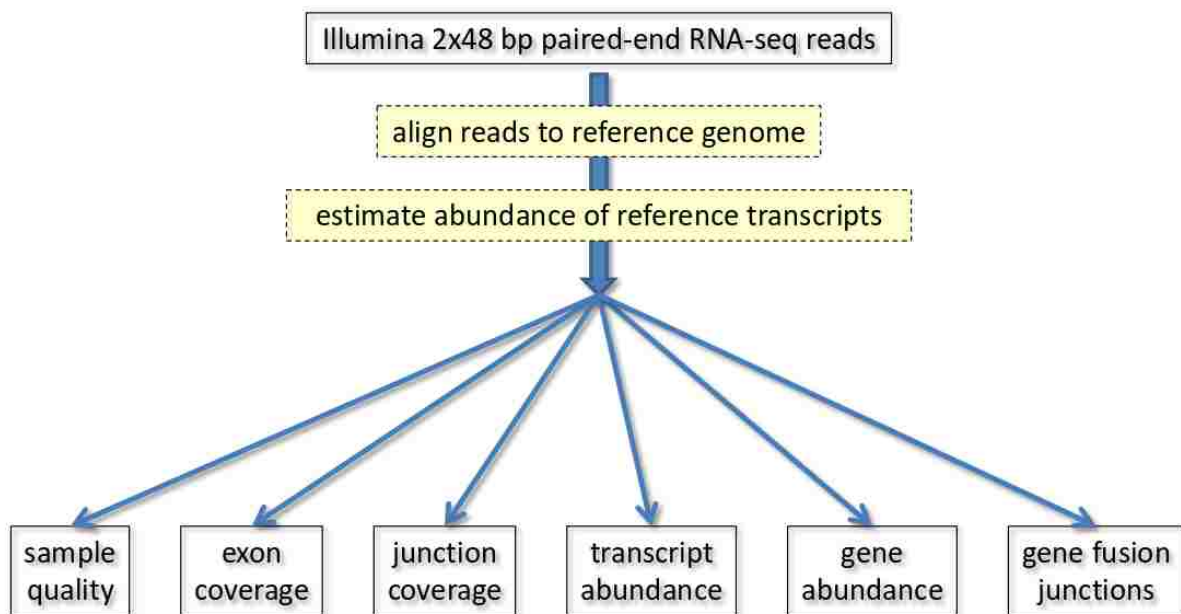


Figure 1: RNA seq workflow implemented by TCGA



## ***Chapter 2: Differential expression of TEs***

### **2.1 Methods**

#### **2.1.1 Data**

We downloaded RNA-seq data for 634 cancer patients from TCGA (The Cancer Genome Atlas) consisting of both cancer tissues and corresponding normal tissues. TCGA has both raw sequencing data (fastq format) and data mapped to the human reference genome (bam format). Though TCGA has more RNA-Seq data for more than 33 different types of cancer, We analyzed only 17 different cancer types for gene expression study because TCGA did not have enough RNA-Seq data for corresponding normal tissues for other cancer types (Table 1). The 17 cancer types included were Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Cholangiocarcinoma (CHOL), Esophageal Carcinoma (ESCA), Head and Neck Squamous Cell carcinoma (HNSC), Kidney chromophore (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell carcinoma (LUSC), Pancreatic adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate Adenocarcinoma (PRAD), Stomach Adenocarcinoma (STAD), Thyroid Carcinoma (THCA), and Uveal Melanoma (UCEC). We downloaded bam files and ran TETranscripts software (version 1.5.0) for counting raw reads mapped to annotated human genome file (gtf format). We downloaded the separate annotated file for genes and retrotransposons, both of which were downloaded from TETranscripts repository hub. Once raw count reads mapped to reference genome (*hg19*) generated, we ran Deseq2 for both normalization and expression of genes and retrotransposons. We analyzed

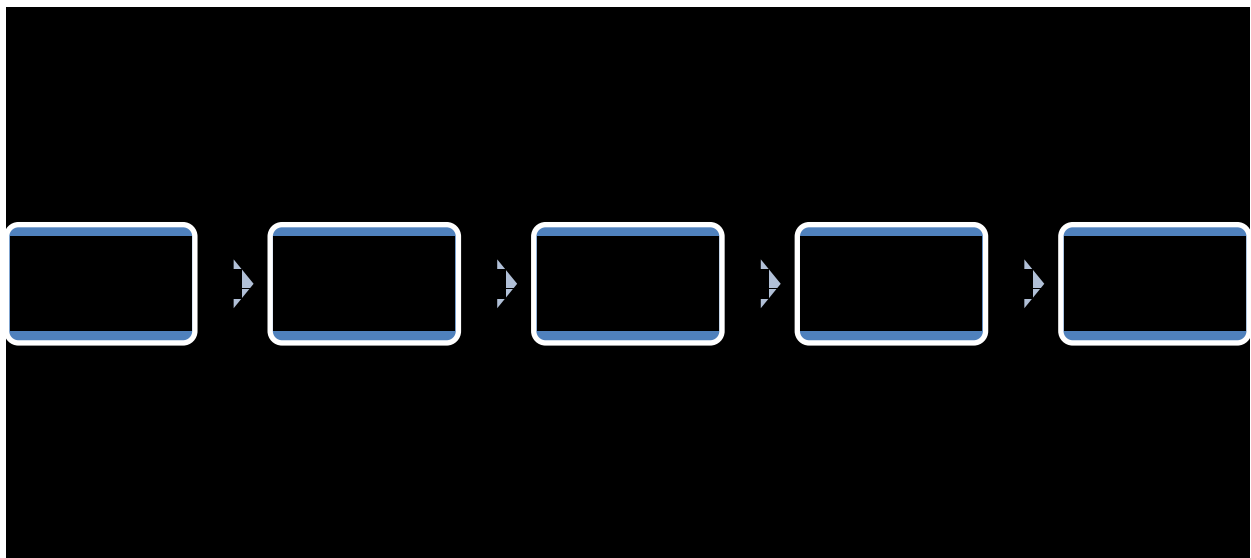
each cancer type for differential expression of genes and retrotransposons where we took cancer tissue as experimental variable and normal tissue as control variable.

*Table 1: Number of patients in different type of cancers with corresponding healthy tissue*

<b>Cancer type</b>	<b>Number of Patients</b>
BLCA	19
BRCA	98
CHOL	9
COAD	48
ESCA	13
HNSC	43
KICH	25
KIRC	63
KIRP	32
LIHC	50
LUSC	49
LUAD	53
PRAD	36
READ	9
UCEC	23
STAD	8

### 2.1.2. RNA Seq:

RNA sequencing is rapidly becoming popular for gene expression study. In RNA seq, mRNA (poly-A) is converted to cDNA and then sequenced. Once it is sequenced, then it is mapped to reference genome. After mapping to reference genome, it is analyzed for differential expression study. RNA Seq has several advantages over other traditional technology (Wang, Gerstein, & Snyder, 2009). First, it is not limited to detecting transcript levels just based on the reference genome. This offers an advantage for non-model organisms whose reference genome is yet to be discovered. Second, RNA Seq is highly accurate in quantifying gene expression level in contrast to DNA microarray study which is not sensitive in terms of very high or low gene expressions. Third, it is a very cost effective method.



*Figure 2: RNA Seq method. In RNA seq, mRNA (poly-A) is converted to cDNA and then sequenced. Once it is sequenced, then it is mapped to reference genome. After mapping to reference genome, it is analyzed for differential expression study.*

### 2.1.3 TEtranscripts

We ran TEtranscripts to quantify raw reads count mapped to reference genome hg19 (Jin, Tam, Paniagua, & Hammell, 2015). It quantifies both gene and transposon. The input files for TEtranscripts are alignment files such as BAM or SAM files, annotated gene and TE files which are in Gene Transfer Format (GTF) files. To enrich multi-mapped reads we ran two additional commands during mapping to the human reference genome using STAR. Those additional commands are : `--outFilterMultimapNmax 100` and `--winAnchorMultimapNmax 200`. One of the key steps in TEtranscripts involves distributing the mapped reads among annotated genes and TEs that overlap those genomic regions. Unique reads which are applicable to most of the genes are relatively easier to distribute. However, multiple reads which are mapped to multiple genomic locations are not easy to handle. Because many TEs has similar sequences, it becomes more complicated to assign the multiple reads those arises from TEs. However, to handle this complexity, TEtranscripts takes account of the sequence similarities at the different level of hierarchy in TEs to distribute reads amongst closest related TE families. Once the reads are assigned, TEtranscripts performs EM (expectation maximization) algorithms to determine the maximum likelihood probability of multi-reads assignment of all of TE transcript levels. First, EM algorithm computes the fractional distribution of each mapped reads of each TEs, which is E-step. Next, it estimates the relative abundances of every TE transcripts (M-step) till the estimated relative abundances converge.

#### 2.1.4 Normalization

Normalization is pivotal for quantifying accurate gene expression levels in the sample. Currently, there are numerous methods for quantifying normalization. Most of the normalization methods take account of a couple of main factors: sequencing depth, gene length, and RNA composition. When it is necessary to compare gene expression between different samples, sequencing depth is important. Because if the library size for one of the samples is higher than other, gene expression level for it might be higher than other. But, it might be false positive results because of the larger library size. That is why it is important to normalize for library size.

Second, gene length is important when gene expression levels within the same sample are compared. If one of the genes is longer than others, it might have more reads than others. Ultimately it might have more expression. Another important factor is RNA composition. While comparing gene expression between the sample, RNA composition might be a factor that might skew the result.

Normalization methods that exist: CPM (counts per million), TPM (Transcripts per kilobase million), FPKM (Fragments per kilo base per million). CPM only takes account of sequencing depth for normalization. Both TPM and FPKM take account of sequencing depth and gene length for normalization. However, FPKM is not recommended for differential expression between samples. DESEQ2 utilizes a median ratio method which accounts for both sequencing depth and RNA composition. As the purpose of differential expression is to compare gene expression level between samples, not within the sample, DESEQ2 is optimal (Love, Huber, &

Anders, 2014). One of the key assumptions made by a median of ratios method applied by Deseq2 is that not all of the genes are differentially expressed.

### 2.1.5 DESeq2

DESeq2 is an R/Bioconductor package for normalization and detection of differentially expressed genes in different conditions (Love, Huber, & Anders, 2014). The input for DESeq2 is raw count data mapped to reference genome generated from different counting tools such as HTSeq or TETranscripts. For each gene, DESeq2 fits a generalized linear model (GLM). It follows negative binomial distribution to model the count reads. The mean of the distribution is considered to be proportional to the cDNA concentration from the sample. To account for sequencing depth between samples, it calculates size factors applying the median of ratios method.



### 2.1.6 edgeR

For gene expression analysis we also ran edgeR: a Bioconductor package for differential expression analysis of digital gene expression data (Robinson, McCarthy, & Smyth, 2010). edgeR models count data--generated from different tools --applying an overdispersed Poisson model. Additionally, it utilizes empirical Bayes model to moderate the overdispersion across genes. edgeR expects data to be summarized into count table where rows are genes and columns are different samples in different conditions. It models the data as negative binomial distribution.

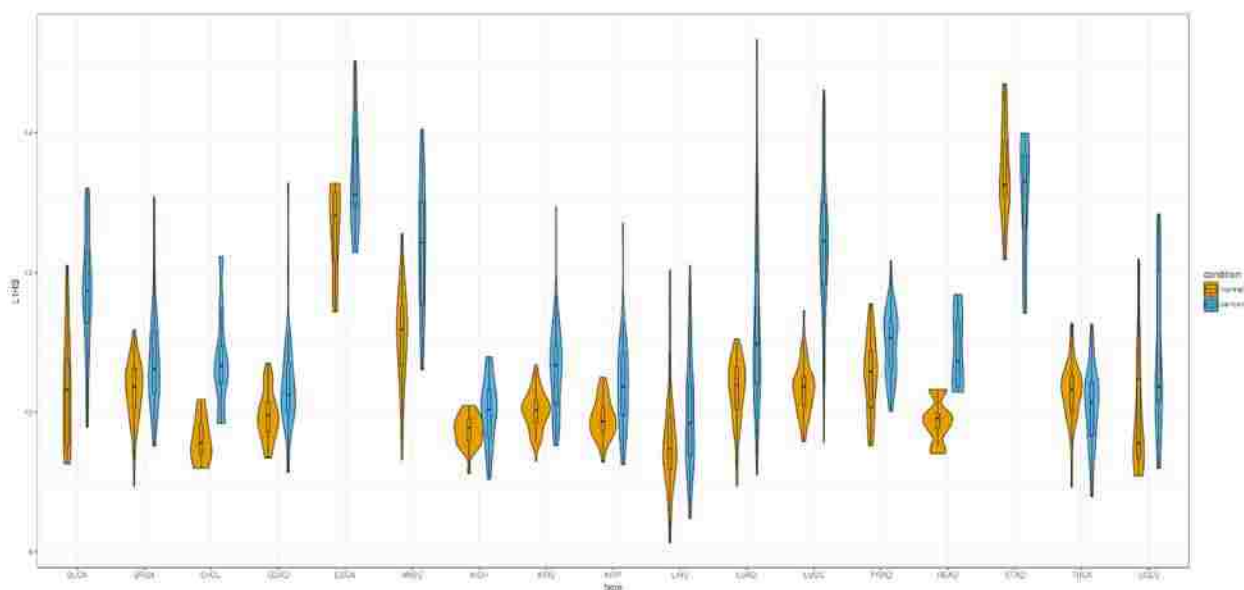
EdgeR separates biological variations from technical variations.

edgeR calculates the genewise dispersions by conditional maximum likelihood. It applies an empirical Bayes procedure to shrink the dispersions towards a consensus value taking information between genes. Then, it calculates differential expression of genes using an exact test similar to Fisher exact test adapted for over-dispersed data.

## 2.2 Results

### 2.2.1. Retrotransposon expression varies across different tissue types and individuals

We found that retrotransposon expression is not uniform across different tissue types. We measured the normalized (reads per million) reads that are mapped to L1HS gene in human reference genome with the help of DESeq2 for different types of tissues and compared them to each other. There are large variations of L1HS across different normal tissues -- with stomach and esophagus have the highest counts (**FIGURE 3** ). Overall, esophagus tissue had the most counts for L1HS.



*Figure 3: L1HS expression in different normal tissues. Raw count reads mapped to hg19 are generated by TETranscripts and normalized by DESeq2. In Y axis, L1HS expression is shown in log2RPM of normalized counts for both normal and cancer tissue. X axis represents different type of cancer and their corresponding healthy tissues.*

Additionally, we measured the L1HS expression (log2 fold change of cancer vs normal tissue) across 634 individual patients. We found that there is also a large variation of L1HS expression

across individual levels (**FIGURE 4**). Despite many of the patients have a less L1HS expression ( $\log_2$  fold change  $< 0$ ) in cancer compared to the corresponding normal tissues, most of the patients have a more L1HS expression ( $\log_2$  fold change  $> 0$ ).

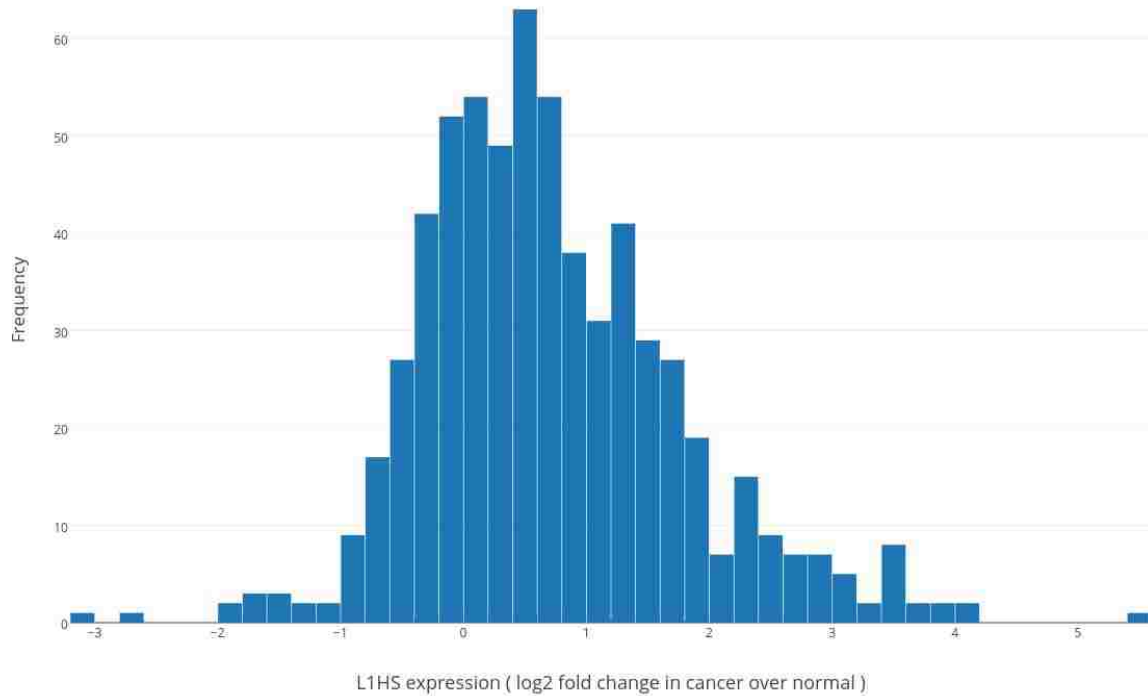


Figure 4: Frequency of L1HS expression over different individuals. L1HS expression is shown as  $\log_2$  fold change of cancer over normal tissue. The number of individuals is 634.

### 2.2.2 L1HS consistently over-expressed in multiple cancer types

We analyzed TCGA (The Cancer Genome Atlas) cancer datasets for differential expression analysis of retrotransposons. We analyzed RNA-seq data from 634 patients for different types of cancer. We used mapped RNA-seq data that are available in TCGA repository and ran Tetranscripts software to produce raw read counts mapping to the genes and retrotransposons sequences. After generating raw read counts, we applied default DESeq2 normalization process to normalize the read counts.

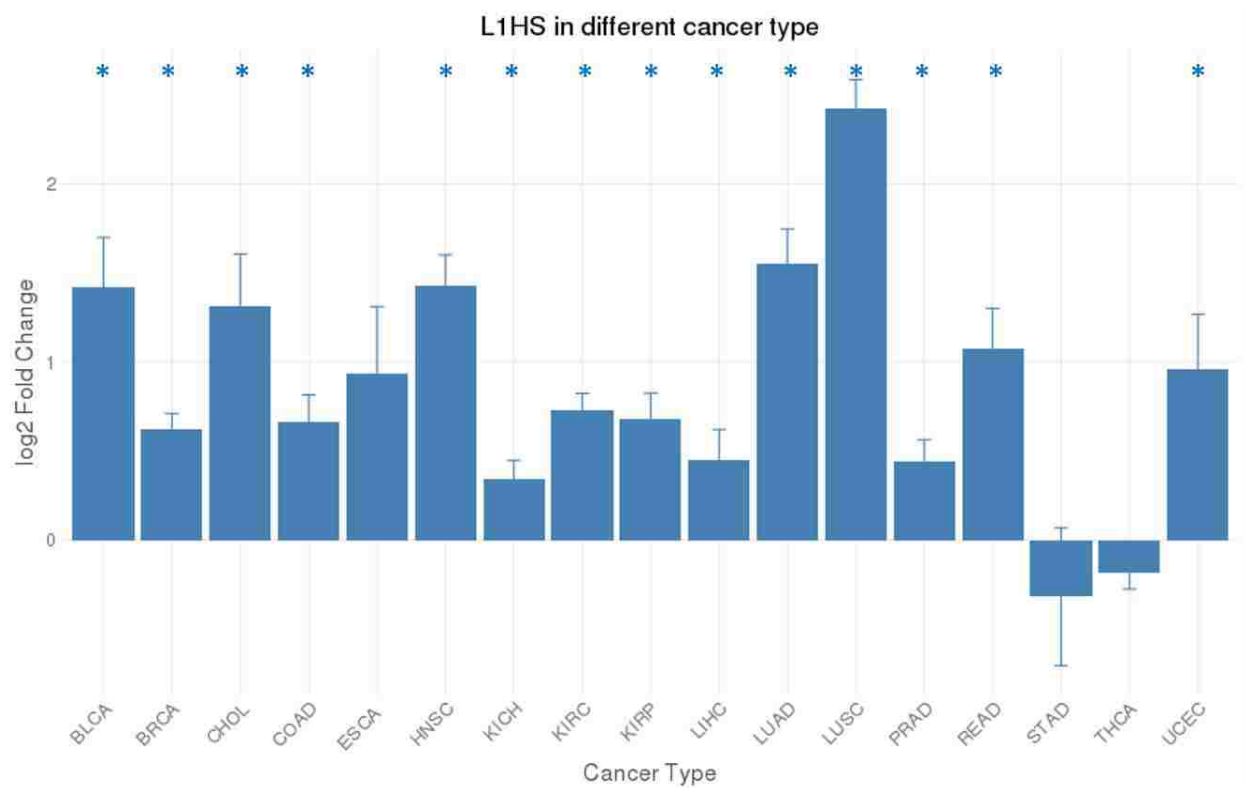
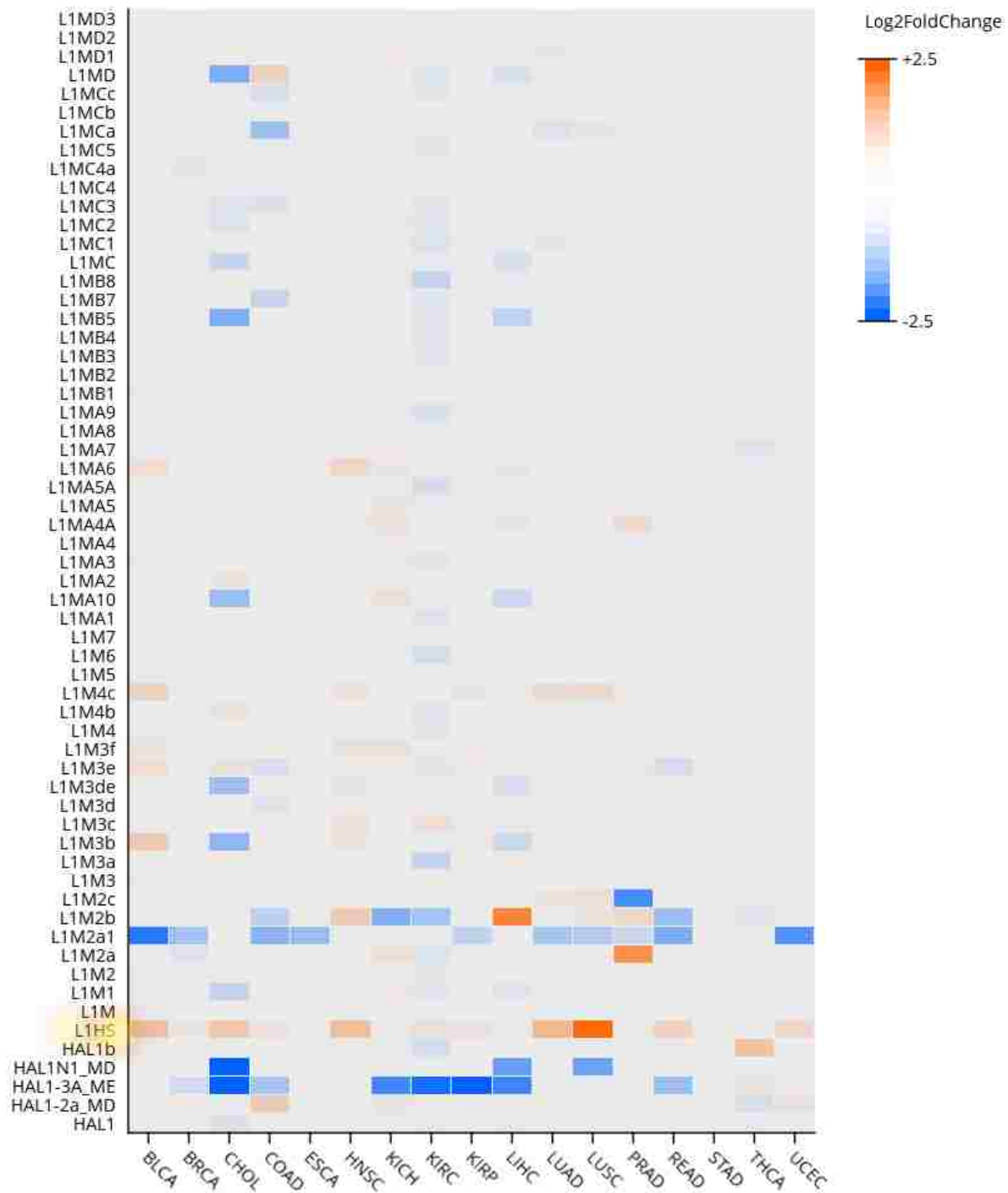


Figure 5: L1HS expression in different types of cancer. Data are shown as log2 fold change for cancer over normal tissues. For most of them FDR < 0.05 except ESCA, STAD, and THCA. Raw count reads mapped to hg19 are generated by Tetranscripts and gene expression level is measured by DESeq2. Normal tissues for corresponding cancers have been used for control. '\*' indicates FDR < 0.05.

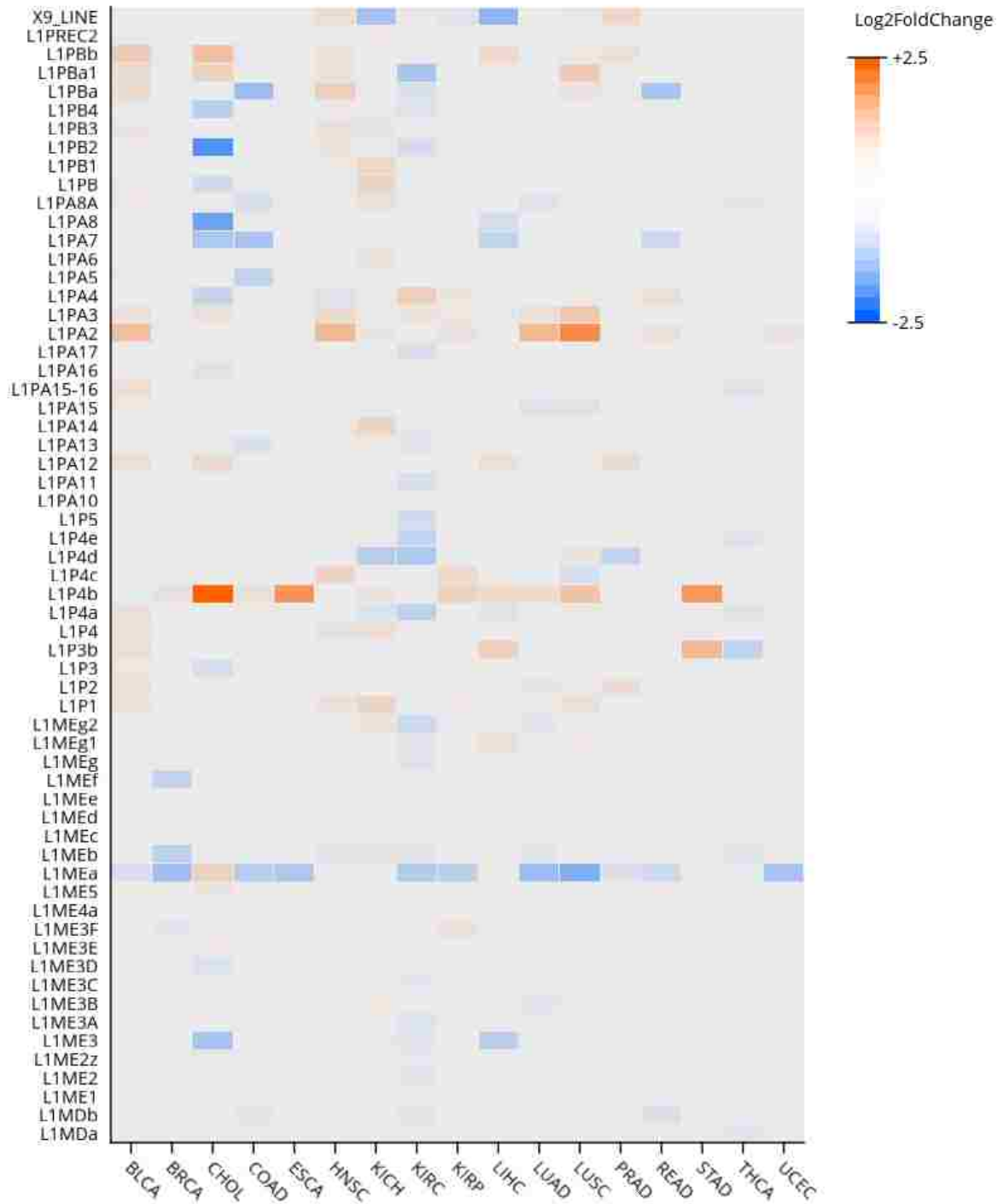
We have observed that numerous types of LINE-1 families and alu families are significantly ( $\text{FDR} < 0.05$ ) overexpressed in different cancer tissue such as colorectal carcinoma, liver hepatocellular carcinoma, kidney renal adenocarcinoma, kidney papillary carcinoma, breast carcinoma, esophageal carcinoma, head and neck carcinoma, prostate carcinoma, etc., compared to corresponding normal tissues (**FIGURE 5, 6.1, 6.2**). The most significant finding in our study is that L1HS, one LINE-1 sub-family, is overexpressed across most of the cancer types of our study (**FIGURE 5, 6.1, 6.2**). Log2 fold changes of L1HS in different types of cancer are shown in figure 3. We also found that log2 fold change was significantly higher for both lung squamous cell cancer and lung adenocarcinoma compared to other cancer types. FDR (false discovery rate)  $> 0.05$  is applied for significance.

## Expression of Retrotransposons in Cancer



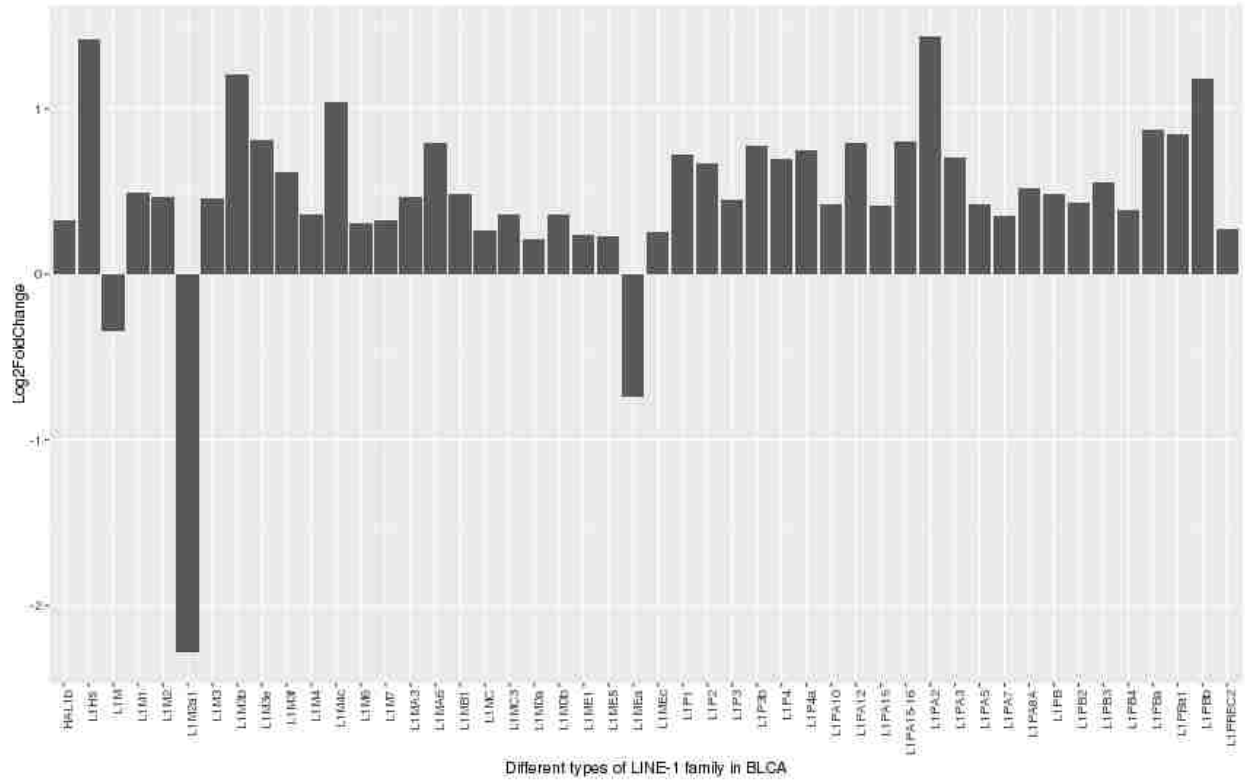
*Figure 6.1: Expression of retrotransposons in different cancer tissues. Expression measured as log2 ratio of normalized count of cancer over normal for each gene. X axis represents different cancer type and Y axis represents different repeat families. L1HS is highlighted yellow.*

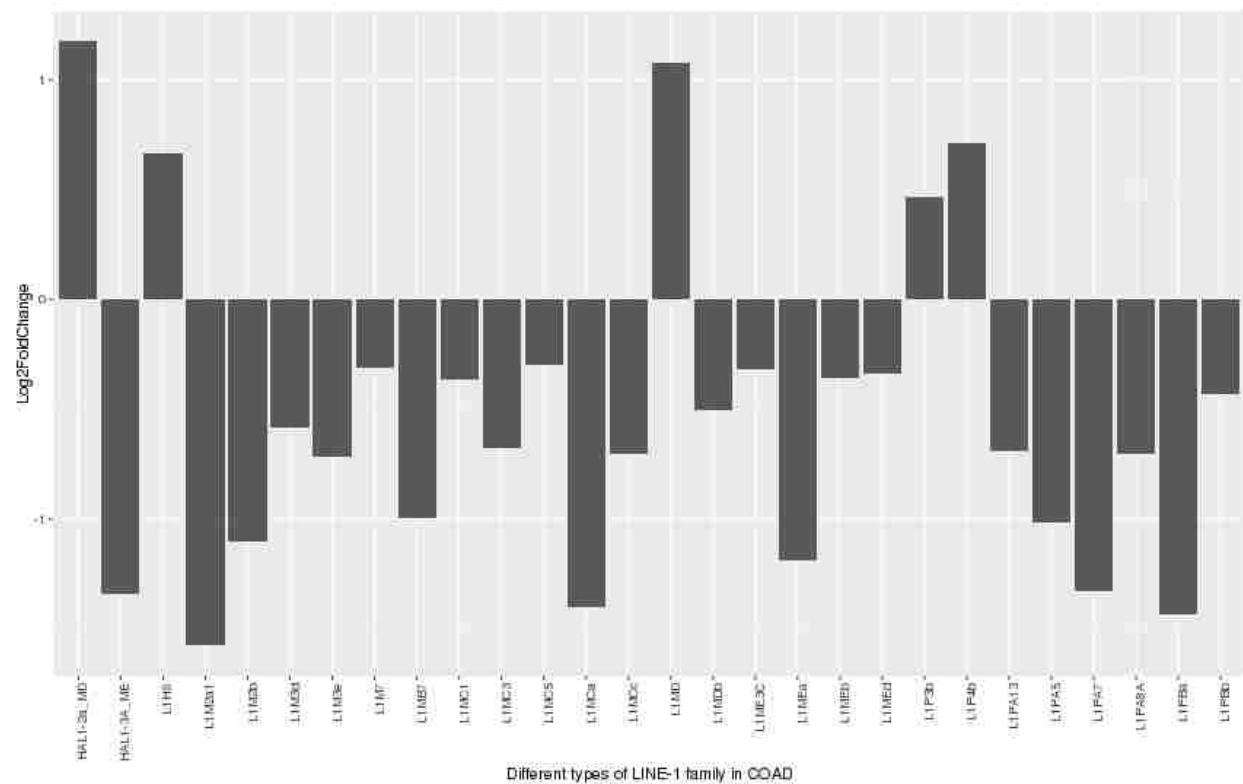
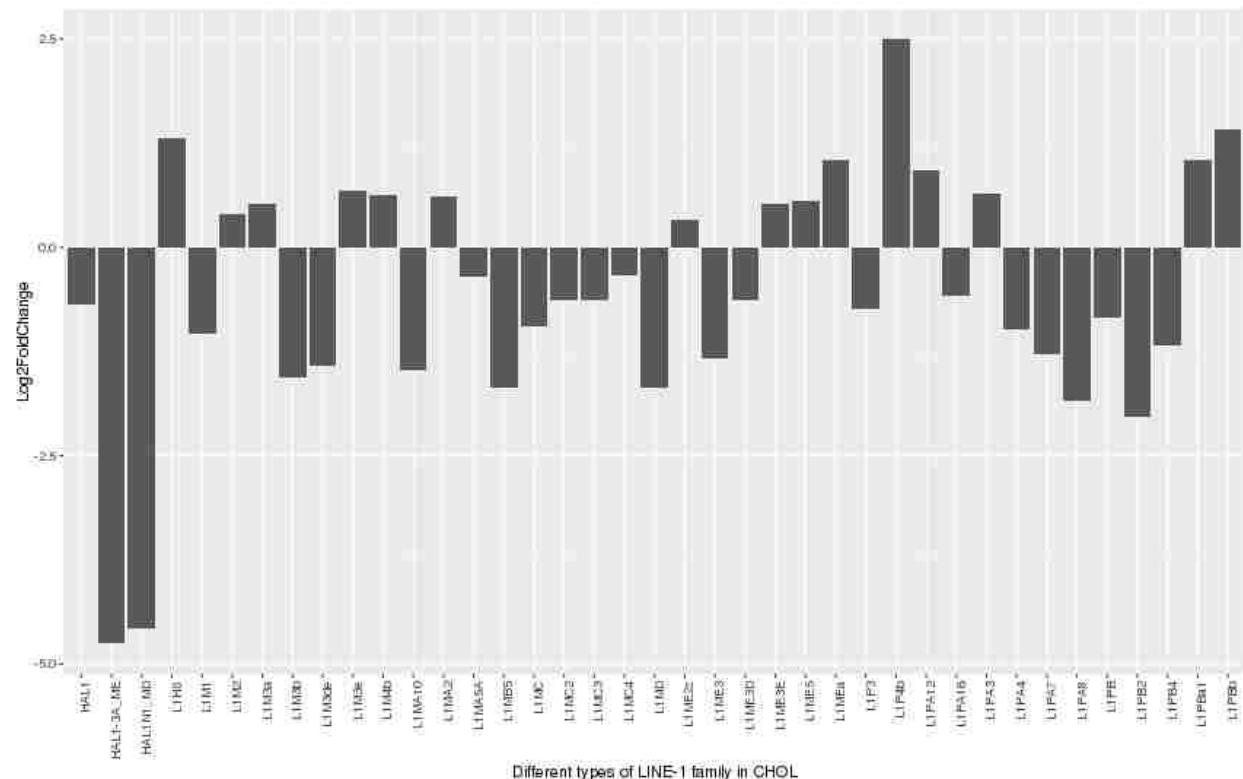
## Expression of Retrotransposons in Cancer





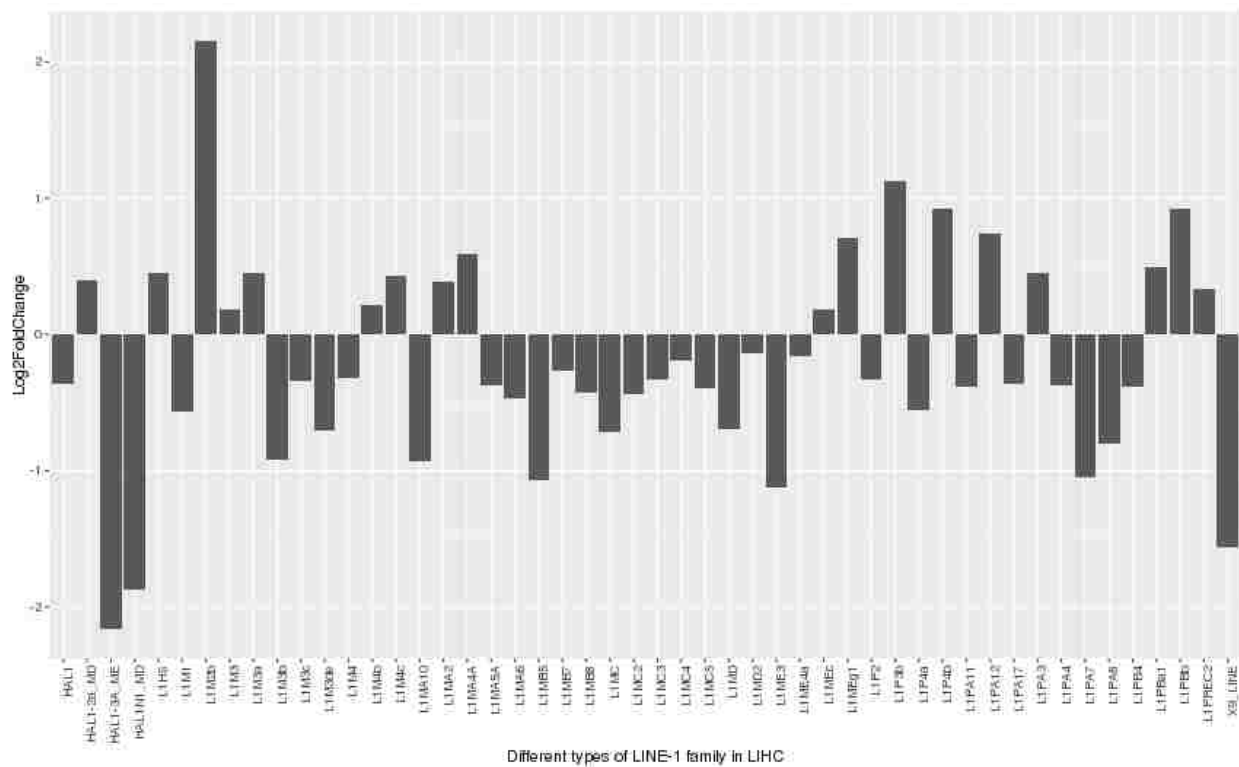
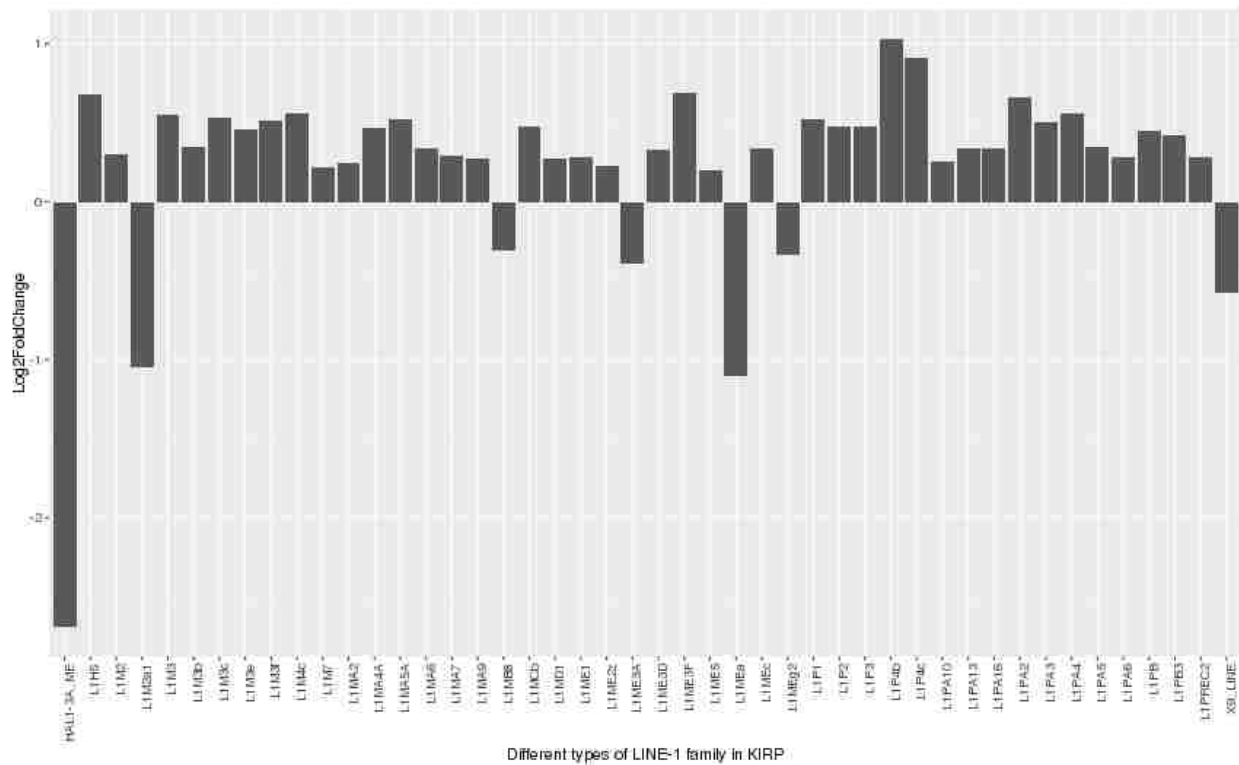
*Figure 6.2: Expression of retrotransposons in different cancer tissues. Expression measured as log2 ratio of normalized count of cancer over normal for each gene. X axis represents different cancer type and Y axis represents different repeat families. L1HS is highlighted yellow.*

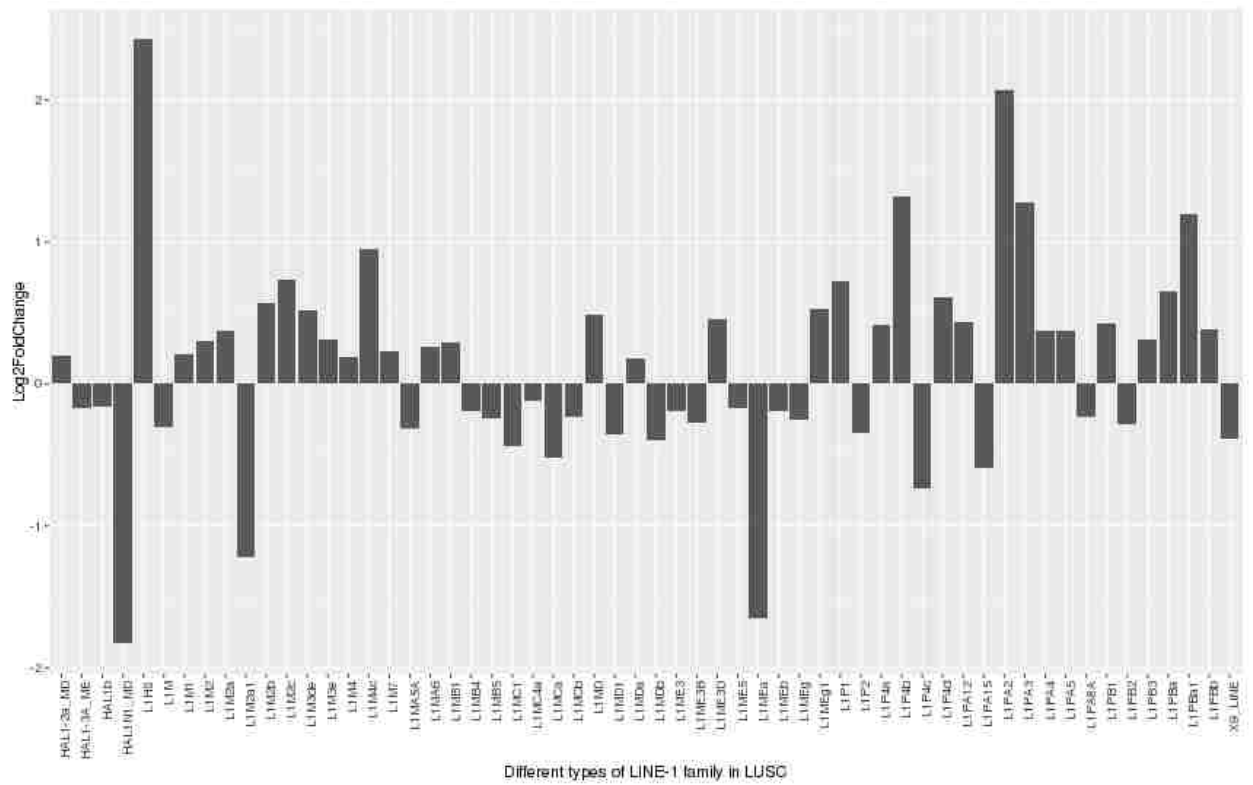
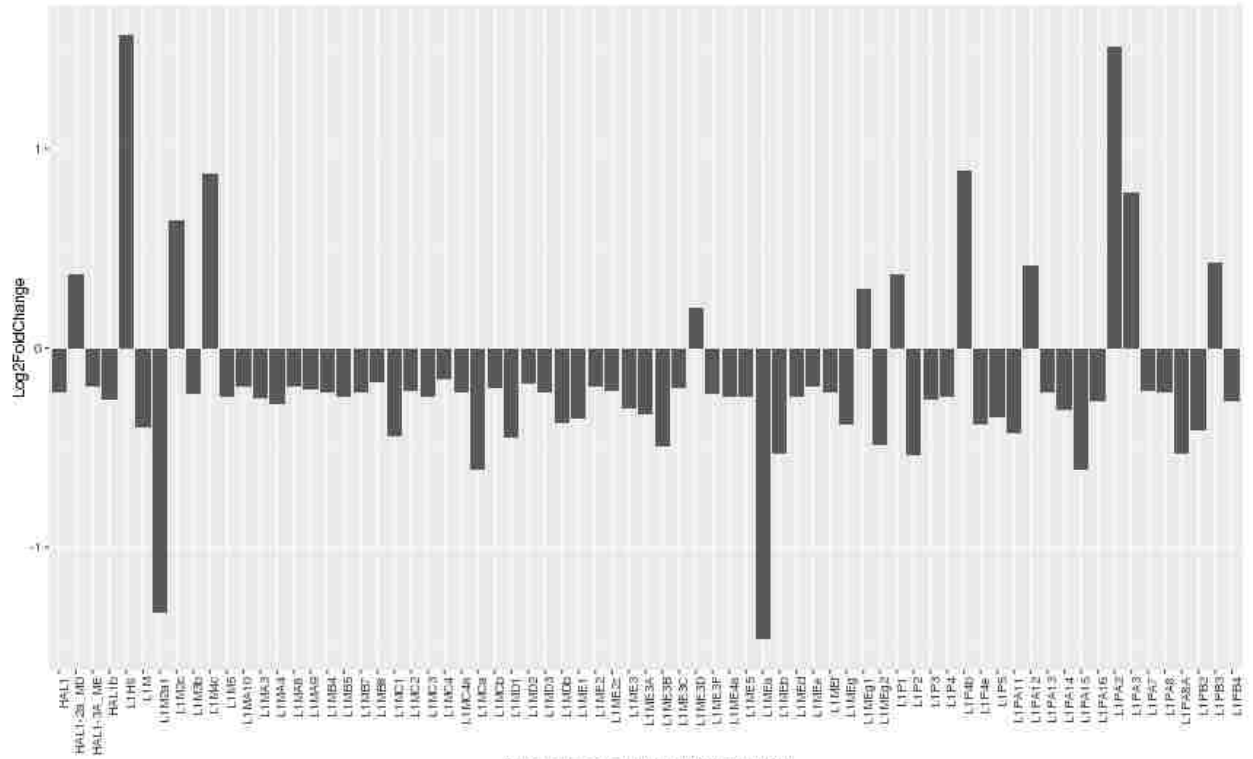


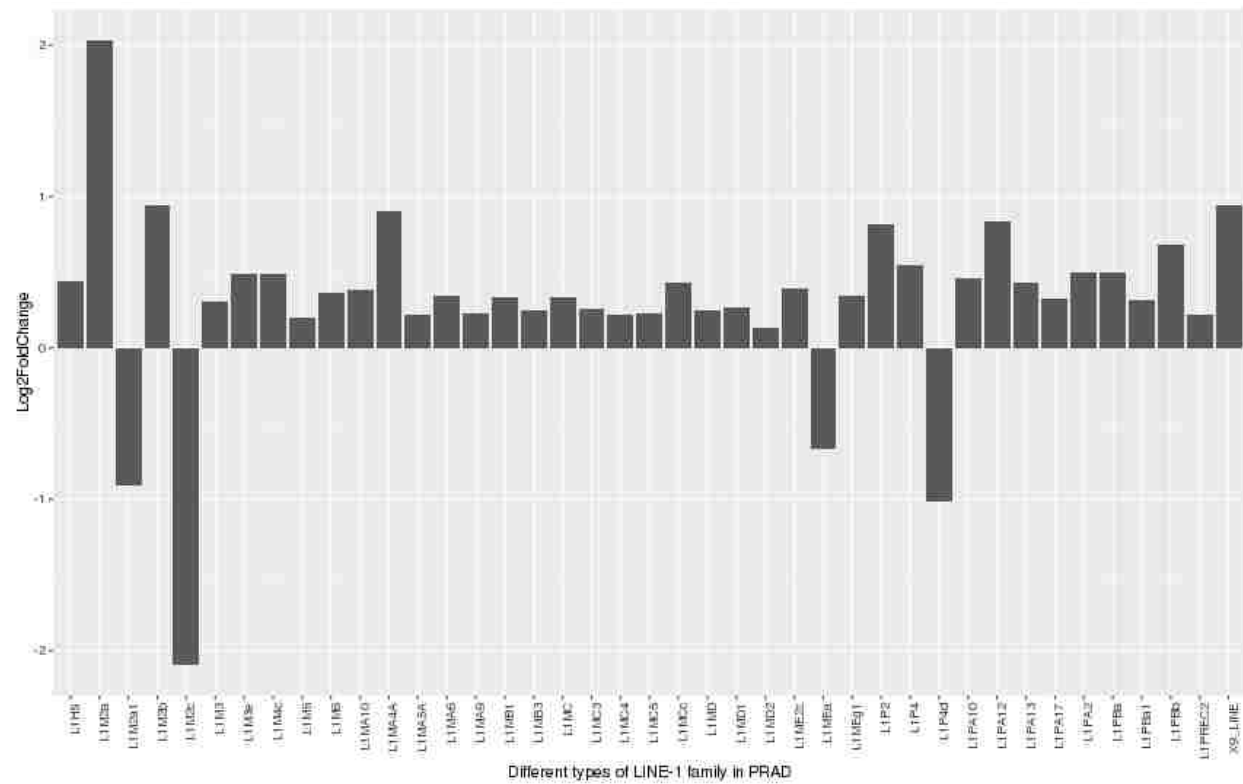




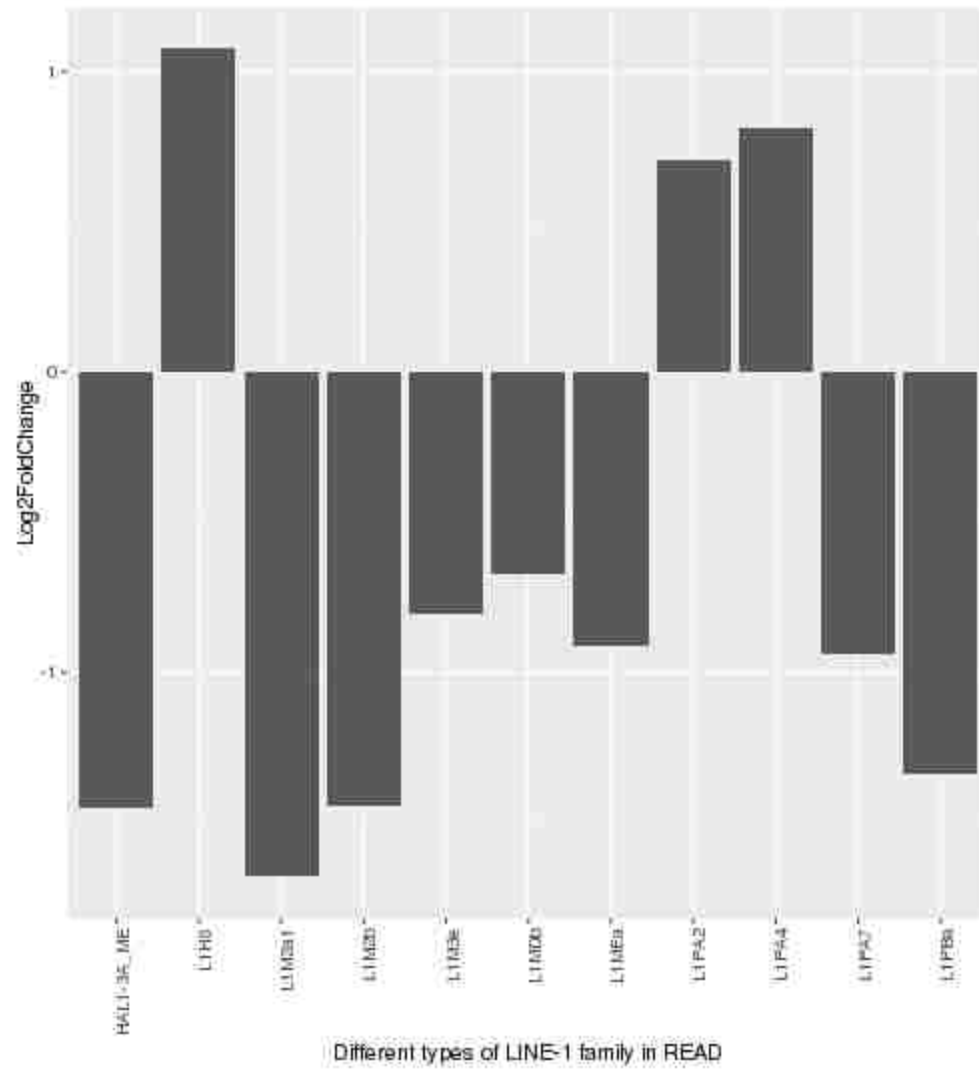












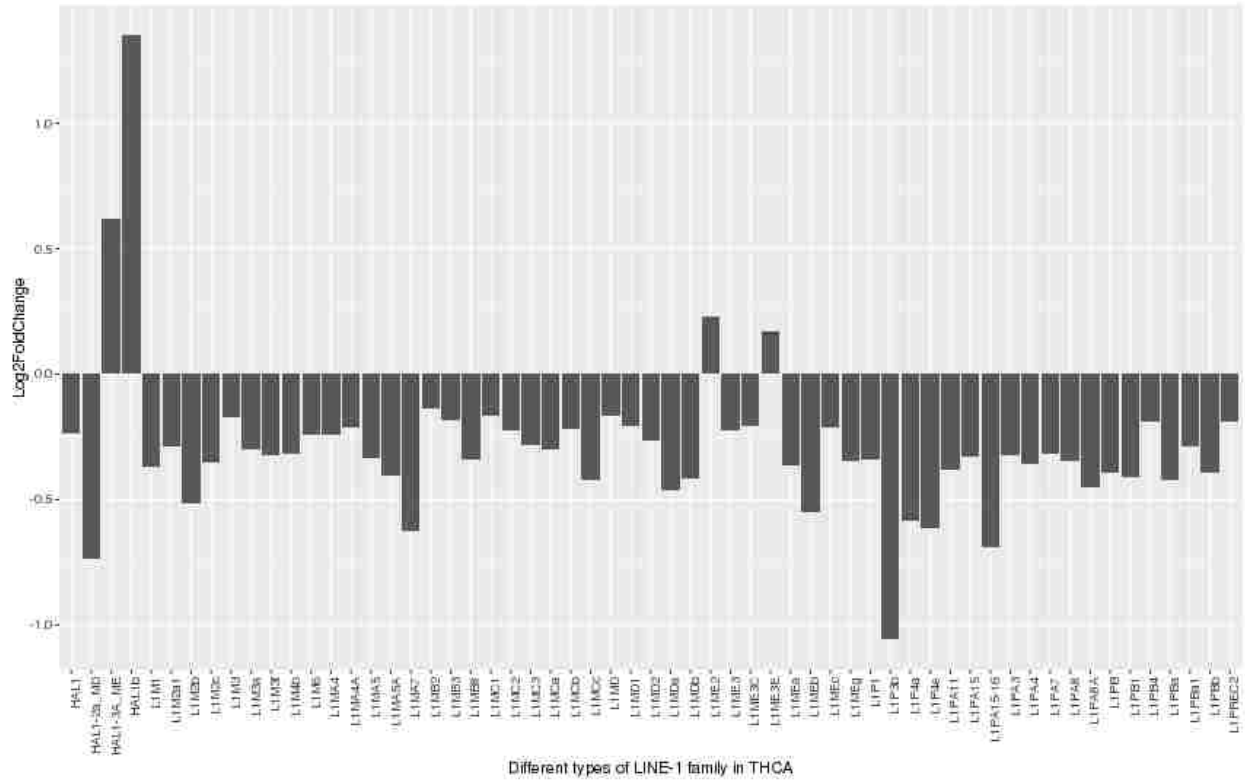


Figure 7: Expression of retrotransposons in different type of cancer tissues. Expression is shown as log2 fold change of normalized counts of cancer over normal for each gene. Only significantly (FDR < 0.05) upregulated or downregulated transposons are shown in these figures.

### 2.2.3. HTSeq comparison

Similar to TEtranscripts, HTSeq counts the raw reads mapped to reference genome (Anders, Pyl, & Huber, 2015). Unlike TEtranscripts, it is not suitable for multi-mapped reads as it ignores the reads that map to multiple locations. However, as most of the genes have only unique reads, we compared the counts for genes between HTSeq and TEtranscripts. Hypothetically, we do not expect to see any difference in terms of gene counts between both tools. Expectedly, we did not notice any difference in terms of gene counts between the two tools. But, we observed the difference in terms of TE counts. **(FIGURES 8 AND 9).**

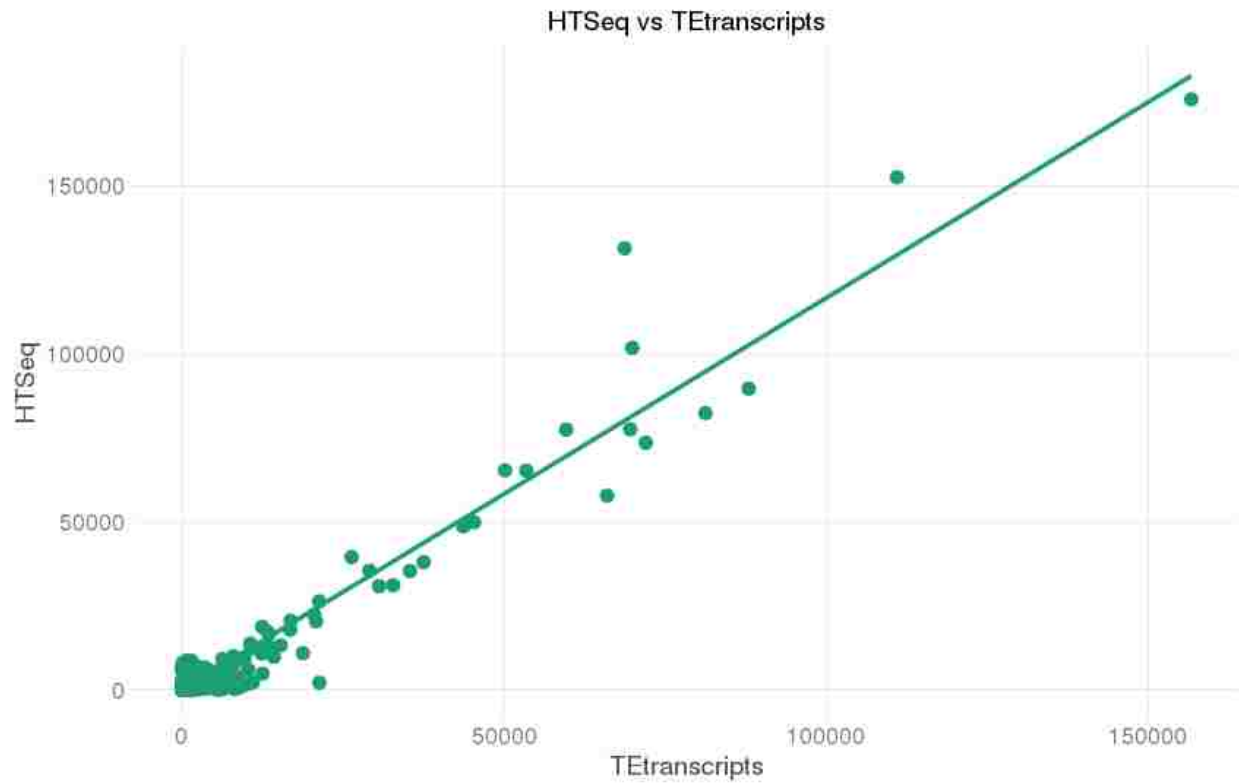


Figure 8: Comparison between HTSeq and Tetranscripts for raw counts of TEs. The Pearson correlation coefficient between HTSeq and Tetranscripts for TEs is 0.9727572.

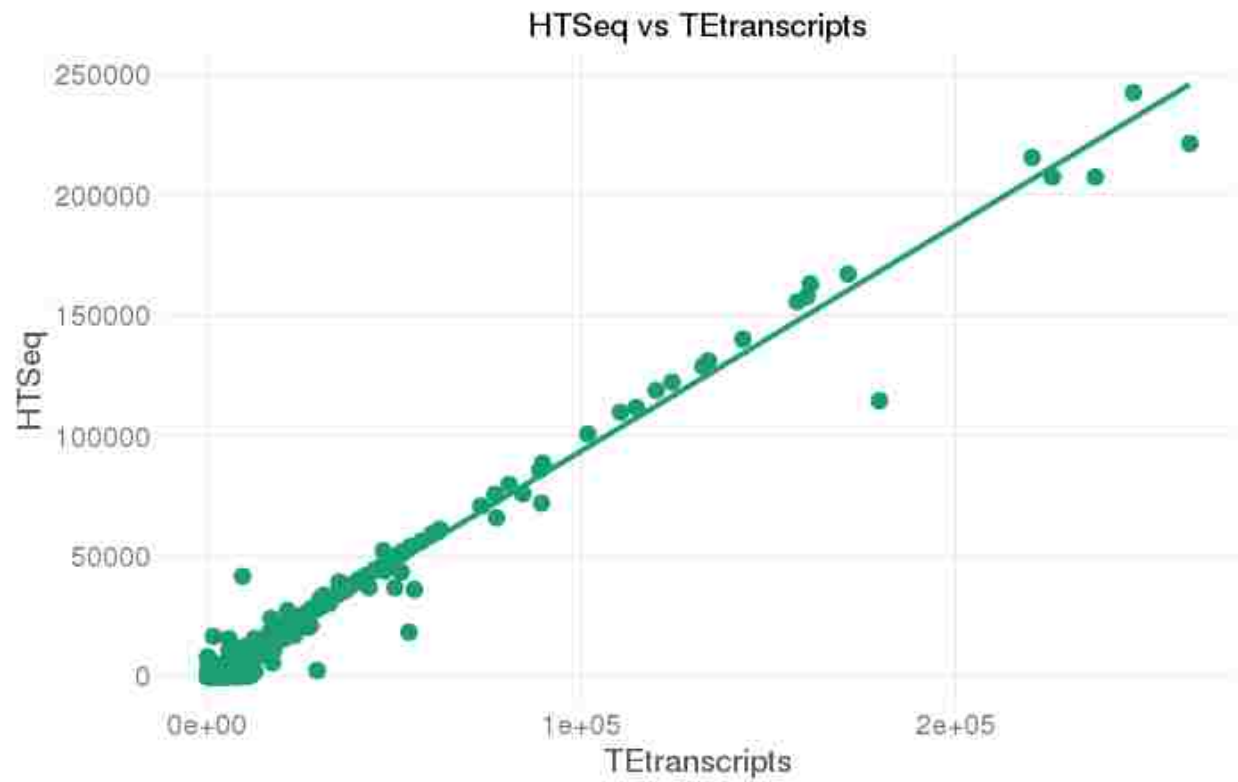


Figure 9: Comparison of HTSeq and Tetranscripts for raw count reads of Genes. The Pearson correlation coefficient between HTSeq and Tetranscripts for the gene is 0.9921285.

#### 2.2.4. TCGA vs STAR

The RNA-Seq files, which were in bam format, analyzed for our study were mapped by TCGA. Additionally, we downloaded fastq files for those RNA Seq files and mapped to *hg19* genome with STAR tools. To enrich multi-mapped reads we ran two additional commands during mapping to the human reference genome using STAR. Those additional commands are : `--outFilterMultimapNmax 100` and `--winAnchorMultimapNmax 200`. Next, we followed the conventional pipeline to generate raw count reads data in TEtranscripts. For this comparison, we analyzed two patients. Once raw data are generated, we calculated Pearson correlation. The correlations for both of the patients were 0.9738137 and 0.9743876 respectively (**FIGURE 10-13**). As the raw count reads from mapped RNA seq files from TCGA were identical to STAR, we opted for the bam files from TCGA to be run directly without downloading fastq files to save time.

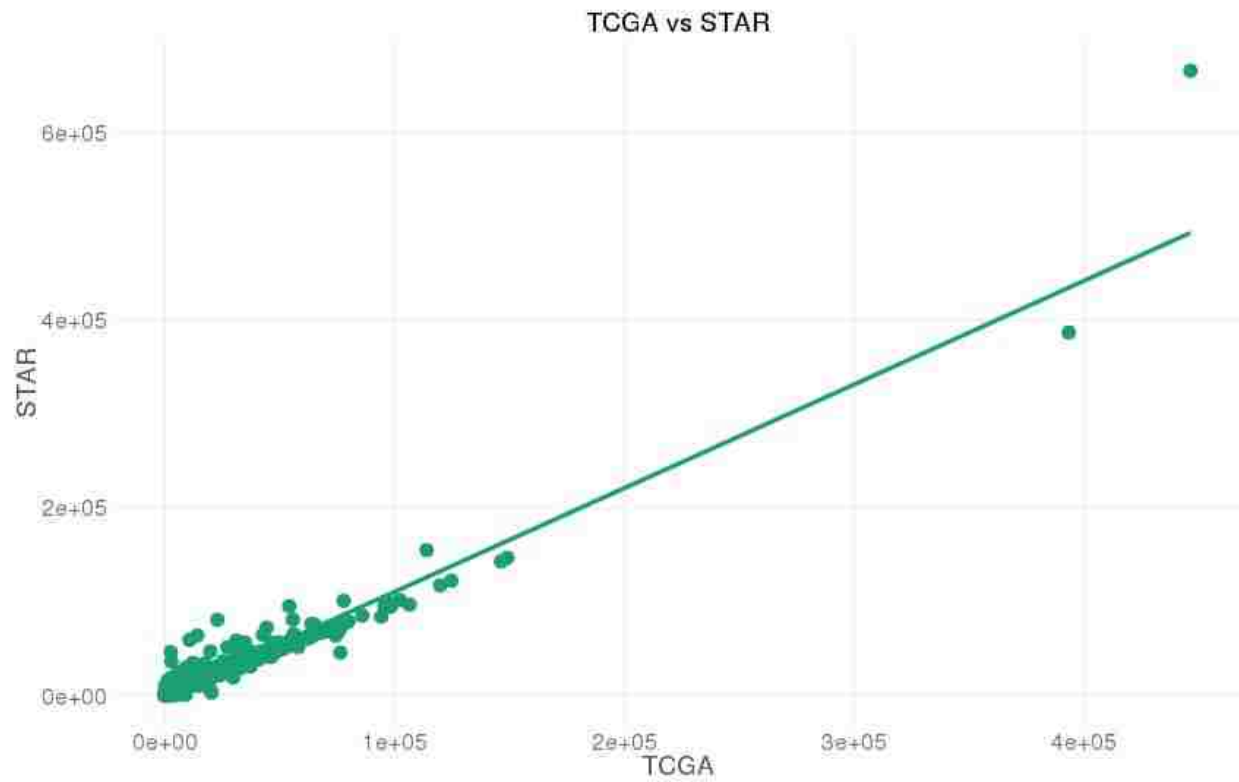


Figure 10: Comparison of gene counts between STAR pipeline and TCGA pipeline for a patient with id AOCE. The correlation is 0.997518.

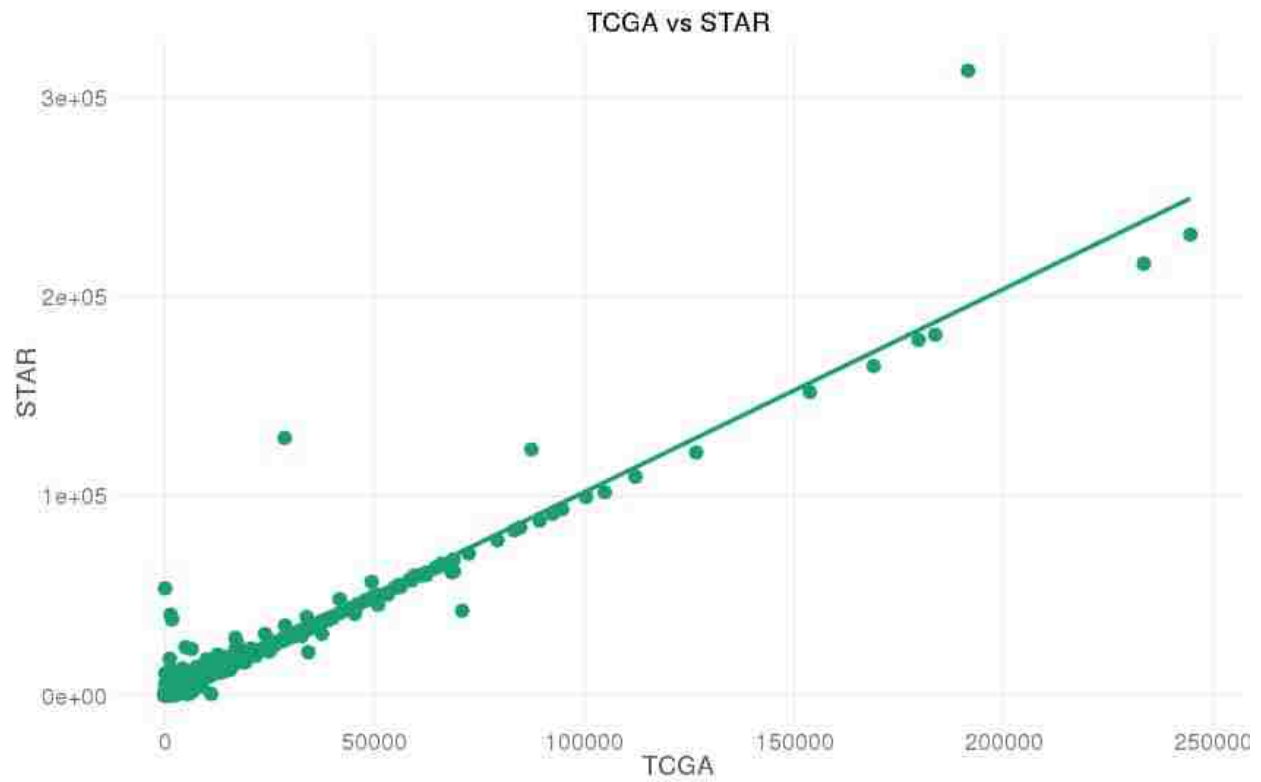


Figure 11: Comparison of gene counts between STAR pipeline and TCGA pipeline for a patient with id AOD9. The correlation is 0.9982605.



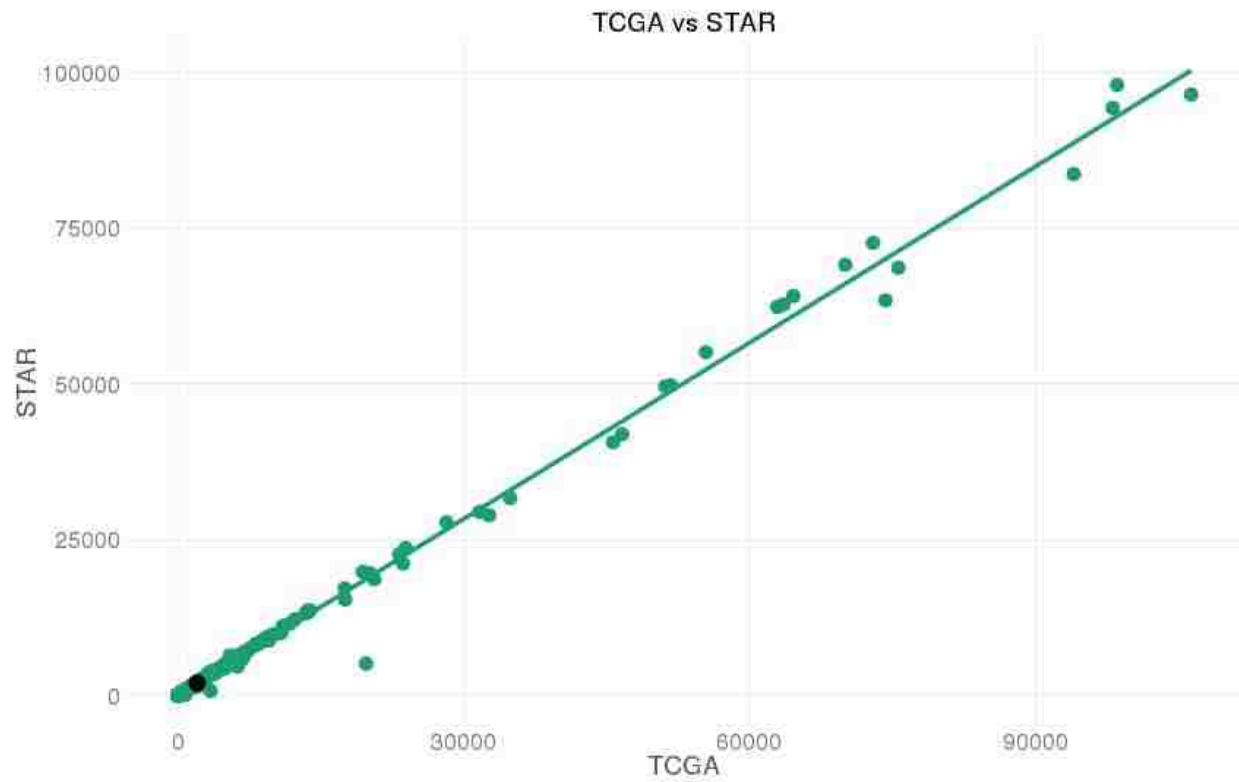


Figure 12: Comparison of TE counts between STAR pipeline and TCGA pipeline for a patient with id AOCE. The correlation is 0.9738137.

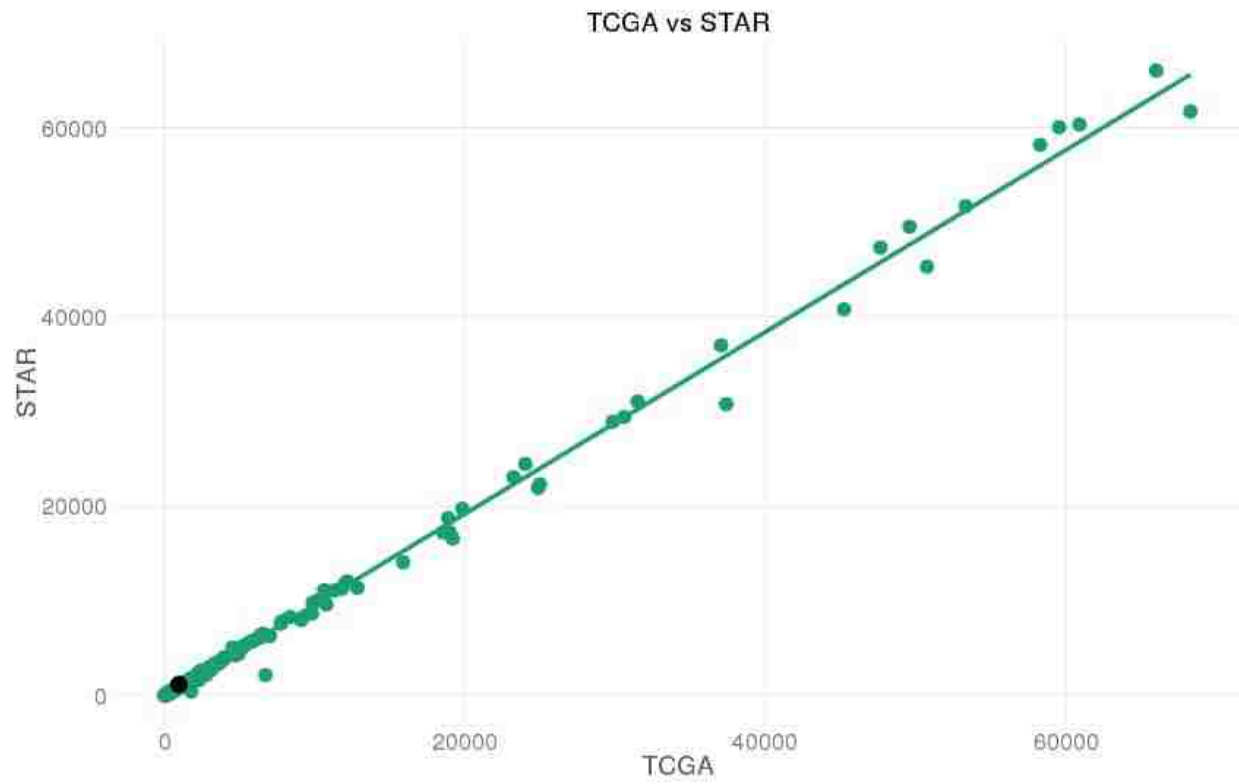


Figure 13: Comparison of TE counts between STAR pipeline and TCGA pipeline for a patient with id AOD9. The correlation is 0.9743876.

### 2.2.5. DESeq2 vs edgeR

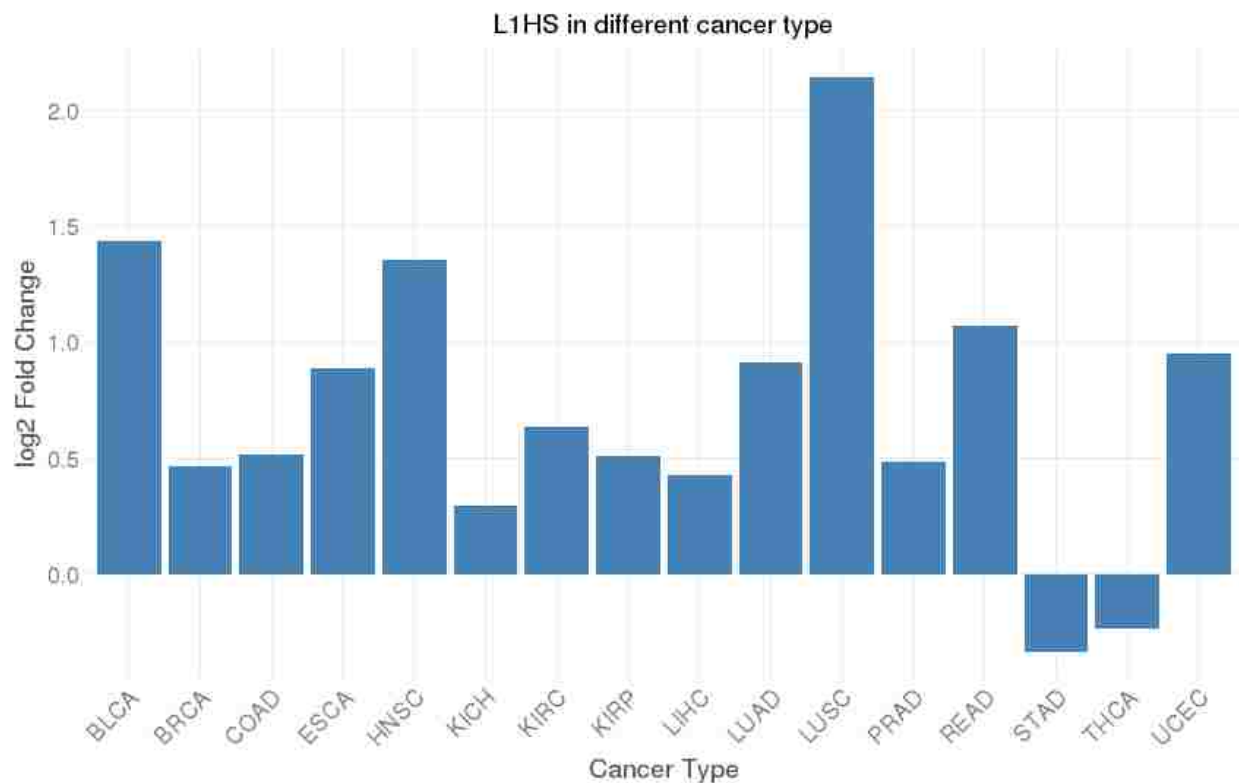


Figure 14: L1HS expression in different types of tissues ( edgeR Results). RNA Seq is downloaded from TCGA. Then, TETranscripts has been utilized for raw count data reads mapped to reference genome hg19. edgeR has been used for normalization and differential expression.

After running edgeR for the same number of samples run by DESeq2, we looked for transposable elements those were differentially expressed in cancer tissues compared to respective normal tissues. We found that L1HS, which was overexpressed in DESeq2 results, also significantly are overexpressed in edgeR results (**FIGURE 14**). Additionally, L1HS was not overexpressed in STAD (stomach adenocarcinoma) and THCA (Thymus Carcinoma) as well as DESeq2.

### 2.2.6. Conclusion

We observed large variations of L1HS across different normal tissues -- with stomach and esophagus have the highest counts (**FIGURE 3**). Overall, esophagus tissue had the most counts for L1HS. We observed that numerous types of LINE-1 families and alu families are significantly ( $FDR < 0.05$ ) overexpressed in different cancer tissue such as colorectal carcinoma, liver hepatocellular carcinoma, kidney renal adenocarcinoma, kidney papillary carcinoma, breast carcinoma, esophageal carcinoma, head and neck carcinoma, prostate carcinoma, etc., compared to corresponding normal tissues (**FIGURE 5, 6.1, 6.2**). L1HS, one LINE-1 sub-family, is overexpressed across most of the cancer types of our study (**FIGURE 5, 6.1, 6.2**).

## ***Chapter 3: Gene and L1HS coexpression***

### **3.1 Introduction**

We sought to find the relative expression level of genes that are known to control L1HS expression. There are numerous genes that are known to control L1HS expression. Some genes are known to restrict L1HS expression; on the other hand, some genes are known to promote L1HS expression. Transcription factors that are known in L1HS regulation include RUNX3, SOX2, YY1, and SP1. Mutations to the binding site of these transcription factors regulation L1HS expression ("A YY1-binding site is required for accurate human LINE-1 transcription initiation," n.d.; Becker, Swergold, Ozato, & Thayer, 1993, p. 1; Minakami et al., 1992; Yang, Zhang, Zhang, & Kazazian Jr, 2003). HNRNPC prevents alu mediated excessive exonization in the cell (Zarnack et al., 2013). One notable retrotransposon restriction factor is APOBEC3 (Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) (Chen et al., 2006; Chiu & Greene, 2008; Holmes, Malim, & Bishop, 2007; Kinomoto et al., 2007; Niewiadomska et al., 2007; Willems & Gillet, 2015; Wissing et al., 2011). Humans possess seven APOBEC3s: A3A, A3B, A3C, A3D, A3F, A3G and A3H. All of the APOBEC3 family are reported to have anti-retrotransposon activity; some of them are also involved in inhibiting retroviral activity. Despite all of the APOBEC3s have anti-retrotransposon activity, A3A and A3B are reported to be most efficient to prevent retrotransposition. Another unit-retrotransposition factor is SAMHD1 (SAM domain and HD domain 1). SAMHD1 is reported to decrease retrotransposition as well as retroviral activity by decreasing available dNTPs for the cell (Clifford et al., 2014; Hu et al., 2015; Kretschmer et al.,

2015). In retroviruses, SAMHD1 also inhibits viral replication by having ribonuclease activity, a function not yet reported for retrotransposon.

Another way retrotransposons are restricted post-transcriptional silencing by small interfering RNA (siRNA). SiRNAs are known to mediate retrotransposon regulation by RNA degradation or epigenetic regulation (Ghildiyal et al., 2008; Obbard, Gordon, Buck, & Jiggins, 2009; Zhai et al., 2015). Numerous genes are involved in this process. The microprocessor, which is formed when DGCR8 binds to DROSHA, splits primary miRNAs, which are then processed in the cytoplasm to mature miRNAs by DICER and further processed by AGO containing RISCs. MOV10 is a member of ATP dependant helicases which precludes all kind of non-LTR retrotransposition in human cell culture. Additionally, MOV10 reduces L1 RNA levels in the post-transcriptional stage. MOV10L1, a MOV10 paralog, is also reported to have anti-retrotransposition activity (Goodier et al., 2012; X. Li et al., 2013, p. 10). RNase L (Ribonuclease L) is an endonuclease single-stranded region of RNAs, which ultimately restricts L1 retrotransposition in human cell culture (Silverman, 2007). TREX1, another factor, reduces the accumulation of reverse transcribed cDNA and has been shown to reduce L1 activity in the cell line (Hasan & Yan, 2014; Stetson, Ko, Heidmann, & Medzhitov, 2008; Yan, Regalado-Magdos, Stiggelbout, Lee-Kirsch, & Lieberman, 2010).

## 3.2 Methods

### 3.2.1 Linear Regression

We are interested in finding genes that are correlated with L1HS expression levels. Linear regressions were implemented on L1HS expression levels (RPM) and gene counts per million data for individual cancers. We took the log<sub>2</sub> value of all our data to create a normalized distribution where all zeros were replaced by the minimum non-zero value in our data set. The relationships were derived from a correlation between individual gene counts per million data and the L1HS expression levels. We analyzed patients per cancer type to make our correlations because of varying normal expression in tissues.

We applied linear correlations for each cancer type. Furthermore, we calculated fold change, which is the proportion of dysregulation in expression of L1HS and genes compared to their normal tissue expression. The independent variable was the log<sub>2</sub> fold change of gene, and the dependent variable was the log<sub>2</sub> fold change of L1HS. We only considered patient gene that had all non-zero values in our regressions.

### 3.2.2. REC score

REC scores are cross-cancer association recurrence scores. They are a measure of a regulatory relationship in different types of cancer (Jacobsen et al., 2013). We applied p-values generated from the linear regression fold change of L1HS and gene counts per cancer. Then a rank was assigned to the smallest p-values. Each rank was divided by the total number of ranks in each cancer. The same procedure was done for the inverted. The H0 score, where H0 is the null hypothesis, was conducted on both tails. The two-tailed H0 scores were placed in chi-square distribution with two times the number of cancers, creating the degree of freedom for that gene. The REC score combines both tails measuring a log10 calculation on the lowest valued chi-square result.



## 3.3 Results

### 3.3.1 Positively correlated expression of Genes with the L1HS expression

Table 2: DAVID enrichment score

annotation	score	count	genes
Bromodomain (IPR001487,SM00297,IPRO18359)	11.22	16	ASH1L, ATAD2B, BAZ2A, BAZ2B, BPTF, BRWD1, BRWD3, CREBBP, EP300, KIAA2026, PBRM1, PHIP, SMARCA2, TAF1, TRIM33, TRIM66
Zinc finger C2H2 type (GO:0006355,IPR007087,IPR015880,GO:0003676,IPRO13087,SM00355,GO:0005622,IPR001909,GO:0046872,SM00349,GO:0003700)	8.14	229	ADNP, ARGLU1, ARHGAP5, ARID1A, ARID1B, ARID2, ARID4B, ASCC3, ASH1L, ATP11B, ATP13A4, ATP7A, ATR, ATRX, BACH1, BAZ2A, BAZ2B, BCLAF1, BPTF, BRAF, BTAF1, CBFA2T2, CCAR1, CDC42BPA, CDH1, CELF1, CHD2, CHD6, CHD7, CHD8, CHD9, CNOT6, CPSF6, CREBBP, CTNND1, CTR9, DGKH, DMTF1, DSC2, DST, EP300, EP400, ERBB3, ERCC6, ERI2, EXPH5, EYA3, FAM120B, FAM126B, FGD6, GATAD2B, GCC2, GOLGA1, GOLGB1, GPATCH8, HECTD1, HELZ, HERC2, HIPK1, HIST4H4, HMBOX1, HUWE1, IKBKAP, IKZF2, INO80D, INTS2, JMJD1C, KDM2A, KDM3B, KDM4C, KDM5A, KDM5B, KDM6A, KLF8, LAT51, LCOR, LIN54, LMLN, LNX2, MACF1, MAP3K4, MAP4K5, MAPK1, MARCH6, MASP2, MAST4, MBNL3, MBTD1, MDM4, MED1, MED23, MGA, MLLT10, MTF1, MTMR3, MTR, MYCBP2, MYSM1, NCOA3, NCOR1, NFAT5, NOTCH2, NSD1, PAN3, PBRM1, PCLO, PHC3, PHF3, PHF8, PIAS1, PIKFYVE, PLAGL2, PLEKHM3, PLXNB1, PNN, POGK, POGZ, PPP2R3A, PRDM10, PRDM2, PRKDC, PTPN14, RAB10, RAB27B, RAB5B, RAD54L2, RALGAP1, RALGPS1, RALGPS2, RANBP2, RAPGEF2, RBAK, RBBP6, RBM25, RBM26, RBM33, RCAN3, RFW3, RFPD3, RFPD4, RNF169, RNPC3, RREB1, RSC1A1, RUFY2, SETD2, SIK3, SMARCA2, SMG1, SON, SP1, SPEN, TAB3, TAF1, TCF20, TET2, TET3, TNKS, TOP2B, TOPBP1, TP53BP1, TRIM33, TRIM66, TRIP11, TRRAP, TSSK4, UBR1, UBR4, UBR5, UNC13B, USP34, VPS13A, VPS13D, WASL, WDFY2, WDFY3, WRN, WSB1, YLPM1, YTHDC2, ZBTB40, ZBTB43, ZCCHC11, ZCCHC14, ZCCHC6, ZFC3H1, ZFH3, ZMAT1, ZMYM1, ZMYM2, ZMYM6, ZNF10, ZNF121, ZNF154, ZNF221, ZNF234, ZNF236, ZNF264, ZNF326, ZNF33A, ZNF397, ZNF417, ZNF426, ZNF462, ZNF493, ZNF518A, ZNF562, ZNF587, ZNF594, ZNF608, ZNF611, ZNF621, ZNF638, ZNF654, ZNF655, ZNF662, ZNF664, ZNF713, ZNF770, ZNF780B, ZNF793, ZNF800, ZNF808, ZNF83, ZNF84, ZNF841, ZNF860, ZNF91, ZSCAN29
Zinc finger, RING/FYVE/PHD-type (IPR011011,IPR001965,IPR013083,IPR019787,SM00249,IPR019786)	5.89	34	ASH1L, ATRX, BAZ2A, BAZ2B, BPTF, CREBBP, EXPH5, FGD6, KDM2A, KDM4C, KDM5A, KDM5B, LNX2, MARCH6, MDM4, MLLT10, MTMR3, MYCBP2, NSD1, PCLO, PHF3, PHF8, PIKFYVE, RBBP6, RFW3, RNF169, RUFY2, TCF20, TRIM33, TRIM66, UBE4A, UBR1, WDFY2, WDFY3
Helicase (IPR000330,IPR001650,IPR014001,IPR006576,SM00592,GO:0016569,SM00490,SM00487,GO:0004386,IPR027417,IPR016197,IPR023780,IPR000953,GO:0008094,GO:0016817,SM00298)	4.18	68	ABCA5, AQR, ARHGAP5, ARID1A, ARID1B, ARID2, ARID4B, ASCC3, ATAD2B, ATAD5, ATP11B, ATP13A4, ATP7A, ATRX, BPTF, BTAF1, CHD2, CHD6, CHD7, CHD8, CHD9, CNOT6, DLG1, DLG5, DNAH1, EP400, ERCC6, ERI2, EYA3, HELZ, KIF20B, KIF21A, LMLN, MAGI3, MASP2, MBTD1, MDN1, MGEA5, MPP7, MTMR3, MYO6, MYSM1, N4BP2, NCOR1, OTUD4, PBRM1, PEX1, PTPN13, PTPN14, RAB10, RAB27B, RAB5B, RAD54L2, SENP1, SENP6, SETX, SMARCA2, SMC2, SMC5, SPEF2, TANC2, USP24, USP34, USP46, USP48, WRN, YLPM1, YTHDC2
Armadillo-like helical, HEAT repeat (IPR011989)	4.08	26	AP1G1, ARFGEF1, ARID1A, ARID1B, ATR, BTAF1, CAND1, CLASP1, CTNND1, HEATR5B, HECTD1, KIAA0368, KIAA1468, MON2, MTOR, NIPBL, PDS5A, PRKDC, PUM2, RICTOR, RIF1, SCYL2, SF3B1, SMG1, TRRAP, WDFY3
Bromodomain	3.14	7	BRWD1, BRWD3, HERC1, PBRM1, PHIP, STXBPS, TAF1

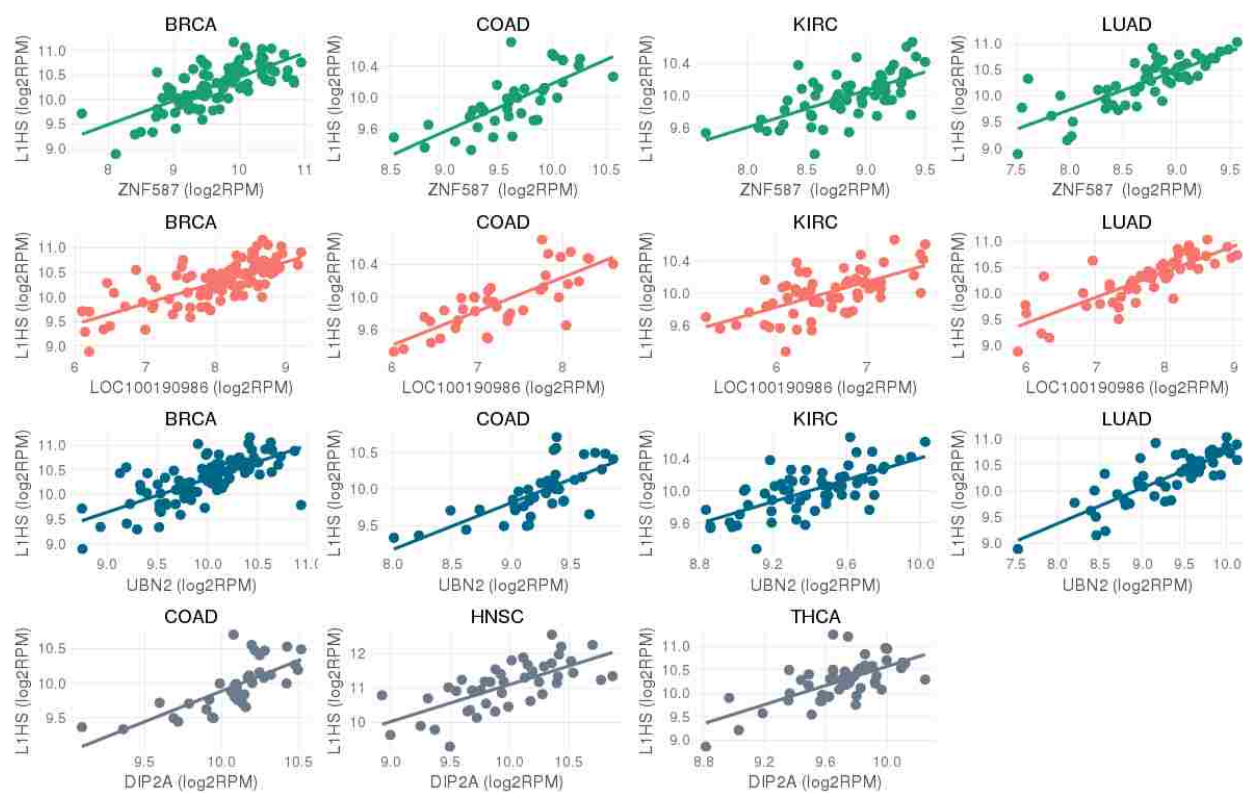


Figure 15: Genes positively correlated with L1HS expression in multiple cancer types

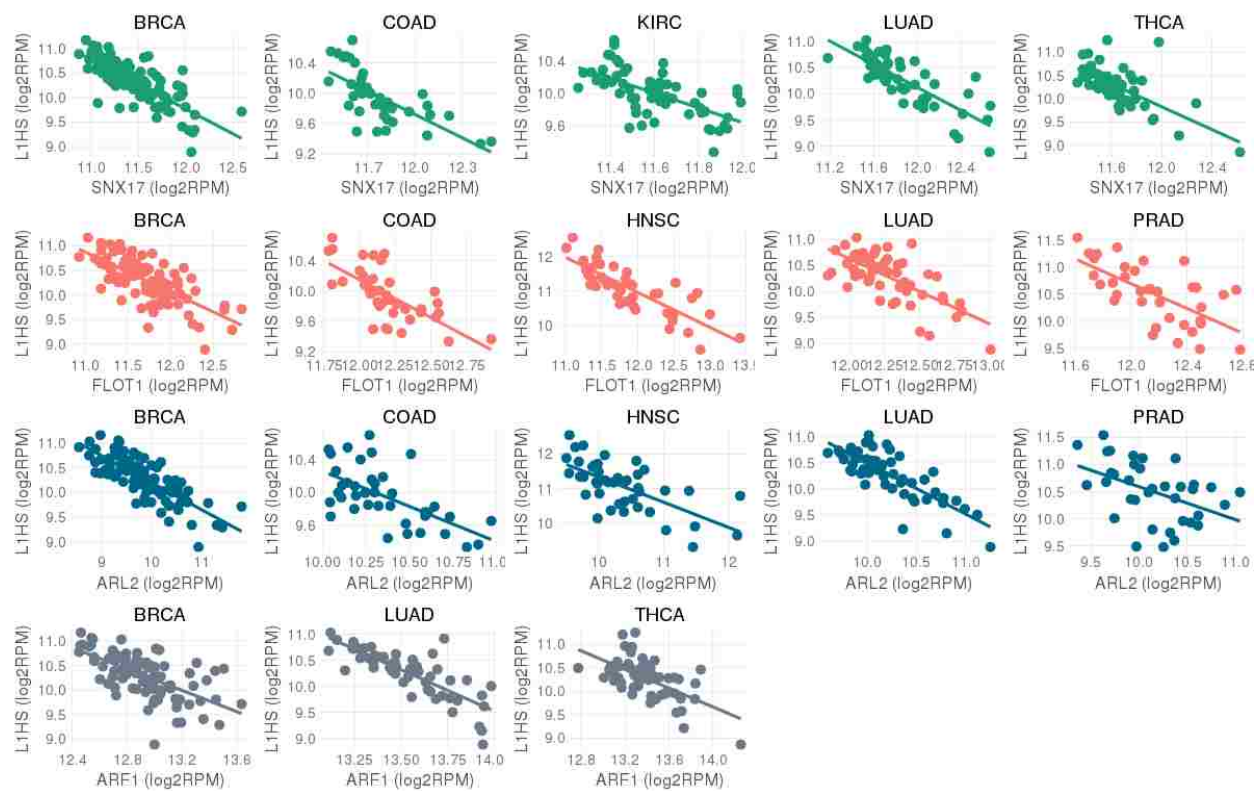


Figure 16: Genes negatively correlated with L1HS expression in multiple cancer types

To study the genes that are co-expressed with L1HS, we relied on the RNA-seq gene level quantification provided by TCGA estimated with the RSEM algorithm. For each gene within each tissue, we tested several linear models with L1HS level as the dependent variable and the individual gene expression level as the independent variable, taking into account the batch effect and the total transcript count as covariates. The coefficients, p-value, and q-value for each gene were estimated based on the best model with the smallest AICc.

We found 2618 genes that showed positive correlation in expression with L1HS (q-value < 10e-4) in at least one tissue. There were 515 genes that were positively correlated with L1HS in more than one tissue. We did a gene set enrichment analysis in DAVID for the 515 genes that are positively correlated with L1HS in more than one tissue. After excluding the very general category of “transcription regulation” that consisted of 266 genes, the top three most enriched clusters were the bromodomains, the PHD fingers, and the C2H2 zinc fingers including the Kruppel-associated box zinc-finger proteins (KRAB-ZFPs) (**TABLE 2** ).

DAVID enrichment score: minus log transformation on the geometric mean of all the enrichment P-values (from Fisher's exact test) for each annotation term associated with the gene members in the group. Enrichment score of 1.3 is equivalent to a non-log scale average p-value of 0.05.

Bromodomains are acetyllysine binding domains and PHD domains are methyl-lysine binding domains. They are both chromatin readers that recognize modified chromatin and function in the epigenetic control of gene transcription. PHD and bromodomains are found in tandem in

the C-terminal of several chromatin-associated proteins (Meroni & Diez-Roux, 2005; Tallant et al., 2015), including the transcriptional intermediary factor 1 (TIF1) family. Two members of the TIF1 family, TRIM33 (TIF1 $\gamma$ ) and TRIM66 (TIF1 $\delta$ ) showed a positive correlation with L1HS in multiple tissues in our analysis. TRIM33, which is known to bind to histone (H3) and repress TFG- $\beta$  signaling pathways, was significantly associated with L1HS in four different tissues, breast (BRCA), squamous cells in the head and neck (HNSC), lung (LUAD) and the thyroid gland (THCA). TRIM66, which is known to be testis-specific and to interact with HP1 $\gamma$  (Khetchoumian et al., 2004) was strongly associated with L1HS levels in breast and kidney (KIRC).

The second enriched category of genes are the C2H2 zinc fingers including the KRAB-ZFPs, many of which directly interact with PHD-bromodomain proteins. The function of most KZFPs are unknown, but a few have been demonstrated to repress transposable elements in embryonic stem (ES) cells (Castro-Diaz et al., 2014; Wolf & Goff, 2009). The best-known example of the interaction between KRAB-ZFPs and PHD-bromodomain proteins is the co-repressor KAP1 (TRIM28) that binds to multiple KRAB C2H2 zinc-finger family of proteins and recruits associated mediators of histone and DNA methylation (Rowe 2010). KAP1 did not show a positive correlation with L1HS in our data, as it actually showed marginal negative correlation with a q-value < 0.001 in breast and q-value < 0.05 in the thyroid gland. There are five KRAB-ZFPs that have been shown to bind to L1HS sequence in earlier Chip-Seq studies, ZNF425, ZNF382, ZNF84, ZNF141 and ZNF649 (Imbeault, Helleboid, & Trono, 2017; Najafabadi et al., 2015). Although binding doesn't necessarily correspond with expression, we looked at whether any of these showed association with L1HS in the expression level. Of these five, ZNF84 showed positively correlated expression with L1HS in the breast and lung tissues, and ZNF141 showed a

positive correlation with L1HS in the breast tissue. Among the KRAB-ZFPs, ZNF37B, which is known as a pseudogene, and ZNF587, that hasn't been studied very much, were the two genes that showed positive correlation in the broadest tissue types. ZNF37B showed significant correlation with L1HS in breast, kidney, lung, and thyroid gland and ZNF587 showed significant correlation with L1HS in breast, kidney, and lung. ZNF587 have been tested for binding to L1HS in the 293T cell line and showed no evidence of binding (Imbeault et al., 2017). Considering the repressing function of KRAB domains, we also looked at whether expression of any of the KRAB-ZFPs was negatively correlated with L1HS expression, but we only found four KZFPs that showed significant negative correlation with L1HS in more than one tissue: ZNF32, ZNF358, ZNF511, and ZNF576. Thus, KRAB-ZFPs were not over-represented among the genes negatively correlated with L1HS in expression.

Since Tetrascripts software provided us with quantification for genes as well as repeat elements, we could duplicate the gene-L1HS co-expression analysis taking the quantification obtained through Tetrascripts instead of the quantification provided by TCGA. Although there were minor differences in the correlation ranks of individual genes between the two quantifications, the highest enrichment of PhD fingers, bromodomains and KRAB-ZFPs were found in both results.

Since the linear regressions were done for each tissue separately and the significance depends on the number of samples for each tissue type, we found that the correlation in the tissues, breast, kidney, thyroid, and lung, which have the largest number of samples, were over-represented in our significant results. So, we decided to use a complementary approach called

REC score 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 cancer types are weighted equally, and limits bias from cancer data sets with large sample sizes or from strong associations measured in only a single cancer type. A strong negative REC score reflects that the miRNA-mRNA pair generally show anti-correlation across the studied cancer types, and this is evidence of a functional miRNA target relationship (Jacobsen et al., 2013). Based on the REC score, there were 116 genes that showed significant positive correlation with L1HS (REC score < -6.2), including long noncoding RNA genes, and a gene involved in chromatin remodeling. Table 3 lists the top 10 genes with the most recurrent positive correlation with L1HS across multiple tissues. The figure shows the correlation with L1HS in the tissues that were found significant.

*Table 3: Top genes predicted by REC score*

rank	gene	REC
1	LOC100190986	-13.26616995
2	NKTR	-12.472002
3	UBN2	-11.45355791
4	ARHGAP32	-11.18303168
5	ZNF587	-11.05819585
6	ZNF37B	-10.43415864
7	FLJ45340	-10.43328428
8	WDR52	-10.40008123
9	BAT2L2	-10.35514121
10	THOC2	-10.0873781

### 3.3.2 Correlation with other TE families

Since Tetrascripts quantifies the reads mapped to different transposon families, we were also able to study the coexpression of different transposons with L1HS. Based on the estimates of Tetrascripts, we found multiple LINE1 subfamilies showing up with the strongest correlation in expression levels with L1HS at the top of the list (**TABLE 4**). In fact, the correlations between L1HS and many repeat families were much stronger than the relationship between L1HS and any of the genes. Based on the REC score that summarizes the ranks of recurrent correlations in multiple tissues, we found that the top 129 transcripts that are in recurrent correlations with L1HS are all repeat elements with various L1PAs occupying the top of the ranks. It is only by the 130th rank in REC score that we observed the first non-transposon gene in correlation with L1HS, which is NKTR, the Natural Killer Cell Triggering Receptor protein. (Although NKTR is the second-ranking gene in the analysis based on TCGA-quantification, it is the first ranking gene in the analysis based on the Tetrascripts-quantification). The correlation between L1 subfamilies is expected because reads from transposons that are indistinguishable between subfamilies are assigned to multiple subfamilies with proportional weight by Tetrascripts applying an Expectation-Maximization algorithm. What was not expected, was that we also found several Endogenous retroviruses (ERVs) that are highly and recurrently correlated with L1HS expression in multiple tissues (Table 5). Since there is no sequence similarity between the ERVs and the LINEs, we conclude that it is due to a common regulatory mechanism that is de-repressing or up-regulating these ERVs and L1HS at the same time. There have been reports of co-expression of ERVs and LINE-1s in cancerous tissues (Desai et al. JCI Insight. 2017), possibly through concordant hypomethylation (Menendez et al., Molecular Cancer 2004). This is the first report



of co-expression of L1HS and specific families of ERVs in multiple normal tissues. In colon cancer cells studied by Desai et al, HERV-K elements were coexpressed with LINE-1, but HERV-H elements were not. HERV-K elements are human-specific and are the most active LTRs in the human genome (Konstantin Khodosevich Comp Funct Genomics 2002). But, in the normal cells we examined, HERV-K elements were not as highly correlated with an L1HS expression as some ERV-L or ERV1 elements. The ERV-K element with the highest REC score was HERVK11 ranked at 242th with a REC score of -7.214.

*Table 4: REC score of L1HS with another repeat family*

REC score rank	LINE1 family	REC	REC score rank	ERV family	REC
1	L1HS:L1:LINE	-Inf	17	ERVL-B4-int:ERVL:LTR	-12.571
2	L1M2:L1:LINE	-Inf	19	MLT1E1A:ERVL-MaLR:LTR	-12.414
3	L1MA4A:L1:LINE	-Inf	25	HERVH48-int:ERV1:LTR	-11.875
4	L1P1:L1:LINE	-Inf	26	Tigger1:TcMar-Tigger:DNA	-11.767
5	L1PA10:L1:LINE	-Inf	29	MLT1A:ERVL-MaLR:LTR	-11.707
6	L1PA13:L1:LINE	-Inf	30	ERVL-E-int:ERVL:LTR	-11.564
7	L1PA2:L1:LINE	-Inf	31	LTR9B:ERV1:LTR	-11.538
8	L1PA3:L1:LINE	-Inf	35	MER21C:ERVL:LTR	-11.373
9	L1PA4:L1:LINE	-Inf	36	MSTB1:ERVL-MaLR:LTR	-11.305
10	L1PA5:L1:LINE	-Inf	38	LTR40a:ERVL:LTR	-11.116
11	L1PA7:L1:LINE	-Inf	41	MER5A:hAT-Charlie:DNA	-10.987

### 3.3.3 Mitochondrial genes are enriched in negatively correlated genes with L1HS

Compared to the genes in positive correlation with L1HS, there was overall less number of genes in negative correlation with L1HS in expression. We found 1891 genes that showed a negative correlation in expression with L1HS ( $q\text{-value} < 10e-4$ ) in at least one tissue. There were 454 genes that were negatively correlated with L1HS in more than one tissue. We did a gene set enrichment analysis in DAVID for the 454 genes that were negatively correlated with L1HS in more than one tissue. The top three most enriched clusters were mitochondrial transit peptide, mitochondrial inner membrane, ribosome and ribonucleoprotein, and oxidative phosphorylation (**TABLE 5**).

Table 5: Mitochondrial genes negatively correlated with L1HS

annotation	Score	Count	Genes
mitochondrial transit peptide, mitochondrial inner membrane (GO:0005743)	22.32	82	ACADS, ARL2, ATP5D, ATP5G2, ATP5G3, ATP5J, AURKAIP1, BAD, BLOC1S1, C19ORF12, C19ORF70, C21ORF33, CHCHD1, CHCHD2, CISD3, CLPP, COX6A1, ECSIT, FIS1, GADD45GIP1, GPX4, HAGH, HAX1, HIGD2A, HMGCL, HSD17B10, HTRA2, MPG, MRPL11, MRPL12, MRPL16, MRPL17, MRPL18, MRPL20, MRPL24, MRPL27, MRPL34, MRPL38, MRPL41, MRPL43, MRPL52, MRPL53, MRPL54, MRPS11, MRPS12, MRPS15, MRPS23, MRPS24, MRPS34, MSRB2, NDUFA11, NDUFA13, NDUFA2, NDUFA8, NDUFAF3, NDUFB10, NDUFB2, NDUFB6, NDUFB7, NDUFB8, NDUFC2, NDUF56, NDUF58, NFU1, NUDT1, PACS2, PARK7, PEMT, PHB, PRDX3, PRELID1, ROMO1, RSAD1, SLC25A6, TIMM17B, TIMM50, TIMM8B, TMEM126A, TRIAP1, TSPO, UQCR11, UQCRCQ
mitochondrial ribosomal protein, ribonucleoprotein (GO:0070125,GO:0070126,GO:0003735,GO:0005762,GO:0005840,GO:0006412,hsa03010,GO:0005761)	14.83	45	AURKAIP1, CHCHD1, FAU, GADD45GIP1, IMP3, LSM1, LSM10, LSM2, LSM4, MRPL11, MRPL12, MRPL16, MRPL17, MRPL18, MRPL20, MRPL24, MRPL27, MRPL34, MRPL38, MRPL41, MRPL43, MRPL52, MRPL53, MRPL54, MRPS11, MRPS12, MRPS15, MRPS23, MRPS24, MRPS34, NHP2, RPL13, RPL3, RPL35, RPS12, RPS19, RPS19BP1, RPS9, SLBP, SLC25A6, SNRNPB, SNRPC, SNRPD2, SNRPN, SRP14
oxidative phosphorylation (hsa05016,hsa05012,GO:0005747,GO:0032981,hsa00190,hsa05010,GO:0006120,GO:0008137,hsa04932,GO:0070469,hsa01100)	8.58	65	ACADS, ADSL, AK1, AP2M1, APH1A, APRT, ATP5D, ATP5G2, ATP5G3, ATP5J, ATP6V1E1, BAD, C19ORF70, CLTA, COX6A1, CRYL1, DCTN2, DCTPP1, DPM3, DTYMK, ECSIT, GUK1, HIGD2A, HMGCL, HSD17B10, HTRA2, NDUFA11, NDUFA13, NDUFA2, NDUFA8, NDUFAF3, NDUFB10, NDUFB2, NDUFB6, NDUFB7, NDUFB8, NDUFC2, NDUF56, NDUF58, NME2, PARK7, PEMT, PGAM1, PGLS, PHB, POLE4, POLR2C, POLR2F, POLR2G, POLR2L, ROMO1, RPN1, SAT2, SLC25A6, ST6GALNAC4, THTPA, TIMM17B, TIMM50, TIMM8B, TMEM126A, TPI1, UQCR11, UQCRCQ, UROD, UROS

### 3.3.4 Conclusion

Overall, we discovered numerous genes that are both positively and negatively correlated with L1HS expression. Through REC score study, we predicted top genes that might be associated with L1HS expression. Additionally, we detected top LINE1 families and other repeat families that are highly correlated with L1HS expression. We predicted functions enriched in positively and negatively correlated genes by DAVID enrichment study. Mitochondrial functions are enriched in negatively correlated genes. On the other hand, bromodomains and zinc finger families are enriched in positively correlated genes.

## ***Chapter 4: miRNA and L1HS coexpression***

### **4.1 Introduction**

MicroRNAs (miRNAs) are short RNA sequences known to play an integral role in gene expression and cell differentiation (MacFarlane & Murphy, 2010). L1HS is a LINE-1 retrotransposon still active in the human genome. L1HS is known to be responsible for several diseases in humans. Our goal was to further understand miRNA dysregulation of the L1HS transposon. We compared L1HS to miRNA expression levels to find significant correlations. From the correlations, computational methods predicted miRNA targets in L1HS's coding regions and 3' UTR.

### **4.2 Methods**

#### **4.2.1 Data**

We collected 1046 miRNA patient files from The Cancer Genome Atlas (TCGA) for 530 patients. The 18 cancer types included were Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Esophageal Carcinoma (ESCA), Head and Neck Squamous Cell carcinoma (HNSC), Kidney chromophore (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell carcinoma (LUSC), Pancreatic adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate Adenocarcinoma (PRAD), Stomach Adenocarcinoma (STAD), Thyroid Carcinoma (THCA), and Uveal Melanoma (UCEC).

We computed gene expression levels in DESeq1 & DESeq2 (Love et al., 2014) for the 630 patients for all genes and TEs. We sometimes took just L1HS cancer read counts per million (RPM) values, but for fold change, we took L1HS normal RPM for the experiments also which being a proper subset of the former reduces the data patients to 530. We omitted 160 miRNAs with zero values across all patients. As with L1HS, we considered just cancer RPM except with experiments with fold-change.

For a set of fold change data including linear regressions and REC scores, we analyzed with matched patient cancer tissue samples of miRNA, L1HS's expression levels, normal tissue sample miRNA, and L1HS expression level counts.

To identify possible target sites, we collected miRNA transcript sequences from MirBase. We downloaded a consensus L1HS cDNA sequence from RepBase. Furthermore, to consider conservation of LINE1 elements, we gathered other mammalian L1 sequences for Homo sapiens, Callithrix jacchus, Sus scrofa, Bos taurus, Microcebus murinus, Rattus norvegicus, Felis catus, Canis lupus familiaris, Mus musculus, Pan troglodytes, and Tarsius syrichta from RepBase.

#### 4.2.2. Linear Regression

We are interested in finding miRNAs that are correlated with L1HS expression levels. Linear regressions were used on L1HS expression levels (RPM) and miRNA counts per million data for individual cancers. We took the log2 value of all our data to create a normalized distribution where all zeros were replaced by the minimum non-zero value in our data set. The relationships were derived from a correlation between individual miRNA counts per million data and the L1HS expression levels. We used patients per cancer type to make our correlations because of varying normal expression in tissues.

Several studies show that an miRNA could be responsible for dysregulation of expression in one tissue type but not necessarily others (Ahluwalia et al., 2008; Filipowicz, Bhattacharyya, & Sonenberg, 2008; MacFarlane & Murphy, 2010). Therefore, we used linear correlations for each cancer type. Furthermore, we used fold change, which is the proportion of dysregulation in expression of L1HS and miRNAs compared to their normal tissue expression. The independent variable was the log2 fold change of miRNA, and the dependent variable was the log2 fold change of L1HS. We only considered patient miRNA that had all non-zero values in our regressions.

### 4.2.3. Linear Mixed Model

We considered the possibility that miRNA and L1HS expression levels could vastly differ across different types of cancer tissues. Therefore we fitted a linear mixed model (Bates et al., 2014; Bolker et al., 2009) where we consider every cancer type to be a category. Using Python's StatsModel library we fitted a mixed linear model per miRNA. Our dependent variable is the natural log of miRNA cancer counts per million. Our independent variable is the natural log of L1HS cancer RPM level.

We considered four possible linear models. We performed Likelihood Ratio Test (LRT) across nested models and compared Akaike Information Criterion (AIC) otherwise. Across our miRNA data Model, 0 best predicted our data. We took separate intercepts per cancer type as our random effect. For every miRNA, we fit a mixed linear model computing a p-value, log-likelihood score, slope coefficient, global intercept, and individual intercepts per cancer type, and graph.

Table 6: Mixed model

Model Name	Model Description	Degrees of Freedom	Nesting
Model 0	Random intercepts model	4	Nested in Model 3 and Model 2
Model 1	Random slopes model	4	Nested in Model 3 and Model 2
Model 2	Random intercepts and slopes assuming correlation = 0	5	None
Model 3	Random intercepts and slopes	6	Nested in Model 2



#### 4.2.4. REC Scores

REC scores are cross-cancer association recurrence scores. They are a measure of a regulatory relationship in different types of cancer (Jacobsen et al., 2013). We used p-values generated from the linear regression fold change of L1HS and miRNA counts per cancer. Then a rank was assigned to the smallest p-values. Each rank was divided by the total number of ranks in each cancer. The same procedure was done for the inverted. The H0 score, where H0 is the null hypothesis, was conducted on both tails. The two-tailed H0 scores were placed in chi-square distribution with two times the number of cancers, creating the degree of freedom for that miRNA. The REC score combines both tails taking a log10 calculation on the lowest valued chi-square result.

#### 4.2.5. TargetScan

TargetScan finds miRNA targets sites of miRNA at the 3' UTR of mature mRNA sequences using context++ scores (D. Betel, Wilson, Gabow, Marks, & Sander, 2007; Doron Betel, Koppal, Agius, Sander, & Leslie, 2010; Didiano & Hobert, 2006). Context scores take into account 17 features including sequence conservation, seed pairing stability, site type e.g. 8-mer, 7-mer, 6-mer. We ran TargetScan 7.0 to calculate context scores and find 8-mer target sites. We downloaded sequences from Repbase which we aligned with Seaview Muscle. The miRNA mature taxon sequence data was from MirBase and miRNA family input files were made from TargetScan's miRNA family file. Open reading frames (ORFs) were found with an NCBI gene annotation file for L1HS. We found L1PT's ORF sequence using OrfFinder, the sequence was analogous to the ORFs found in L1HS. Each part of the sequence had an affected isoform ratio (AIR) of 100. AIRs

are important because the scores current scores can be increased five-fold considering this variable.

TargetScan relies on an input of an ORF sequence per gene. L1HS has two ORFs (McMillan & Singer, 1993) which were tested separately. TargetScan takes into account the length of the ORF sequence, which explains difference in context scores between our short-ORF and long-ORF.

#### 4.2.6. MiRanda

We are interested in finding MiRNA target sites for L1HS regulation. The software MiRanda takes a MiRNA transcript and the cDNA sequence of L1HS and predicts possible target sites. Based on the results of the mixed model we chose MiRNAs that had a significant correlation. For each MiRNA, we tested all known transcripts against the L1HS cDNA sequence. MiRanda predicted multiple possible MiRNA target sites in the 3' UTR and CDS of L1HS. These results match the predictions of TargetScan.

## 4.3 Results

### 4.3.1. Several miRNAs are correlated with L1HS

The linear regression by miRNA including all cancers did not yield significant results. We found the correlation scores were below  $|.4|$ , this did not meet our threshold. This is attributed to each cancer has its own expression relation a miRNA and differential gene expression based on each tissue. The data from the linear regression by miRNA shows that various miRNAs are active in different tissue samples. We evaluated data points in one linear regression, which did not look at these underlying patterns and skewed results.

For our linear regressions per cancer type results, we found several high correlation coefficients, however, most cancer types have few data points, and therefore we cannot confidently conclude much from these results.

*Table 7: Linear Regression Results. Per miRNA, we performed a linear regression on all cancers without fold change. This figure shows our top results based on r-value.*

<b>MiRNA</b>	<b>P-value</b>	<b>R-value</b>	<b>Q-value</b>
hsa-mir-203	4.26E-22	0.402	2.69E-19
hsa-mir-944	7.28E-22	0	1.70E-13
hsa-mir-194-1	1.79E-16	-0.347	1.49E-14
hsa-mir-200c	1.29E-21	0.398	2.69E-19

*Table 8: Linear Regression Results. Per miRNA we performed a linear regression on fold change of all cancers. This figure shows our top results based on r-value.*

<b>MiRNA</b>	<b>P-value</b>	<b>R-value</b>	<b>Q-value</b>
hsa-mir-30a	4.15E-08	-0.235	6.05E-07
hsa-mir-145	0.702	0.016	0.788
hsa-mir-210	2.58E-19	0.377	1.81E-16
hsa-let-7f-1	0.043	0	0.0962

Table 9: Linear Regression Results. Per miRNA we performed a linear regression on each cancer type. This table shows our top results based on r-value.

MiRNA	Cancer type	R-value	P-value	Q-Val	Number of Patients
hsa-mir-204	PRAD	-0.654	$x < 1.0e-4$	0.0389148615	36
hsa-mir-375	PRAD	0.647	$x < 1.0e-4$	0.0389148615	36
hsa-mir-374a	KICH	-0.615	0.001	0.1172555	25
hsa-mir-503	KICH	-0.614	0.001	0.1172555	25
hsa-mir-505	KICH	-0.610	0.001	0.1172555	25
hsa-mir-1468	UCEC	0.604	0.004	0.23811063639393937	21
hsa-mir-382	UCEC	0.600	0.004	0.23811063639393937	21
hsa-mir-9-1	KIRP	0.516	0.003	0.7401719	32
hsa-mir-9-2	KIRP	0.511	0.003	0.7401719	32
hsa-mir-200c	HNSC	0.506	6.27E-4	0.1725512	42

### 4.3.2. Mixed Linear Model predicts 38 potential miRNAs

We compute 731 mixed linear model, one for each miRNA and a graph of the fitted model. We compiled a list of 38 MiRNAs with significant q-values based on the fitting of the linear mixed models.

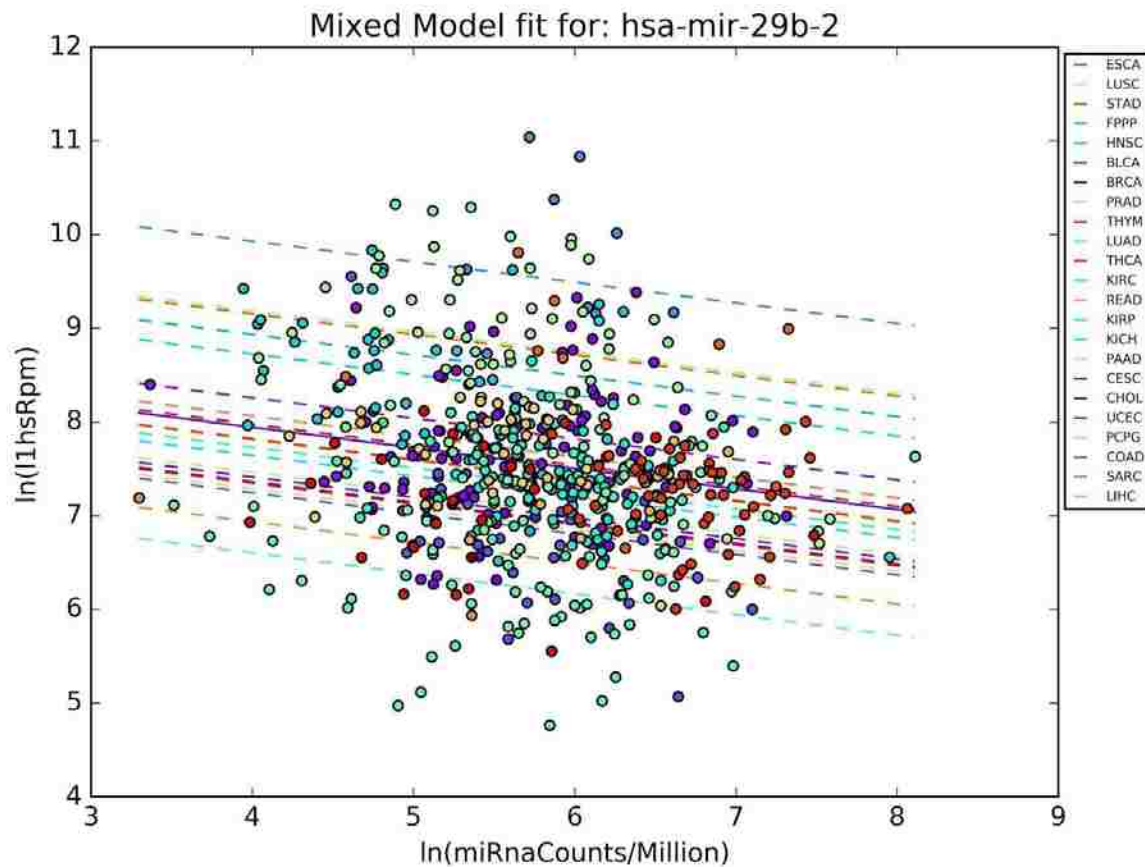


Figure 17: Linear mixed model fit miRNA hsa-mir-29b-2. Sample graph for linear mixed model results for hsa-mir-29b-1 comparing log MiRna Counts / Million versus log L1HS RPM Values. We compute the global regression line, as well as individual cancer type regressions.

We compile a list of MiRNAs, fitted slope coefficient, and p-value. We compute q-values from our compiled list and consider only miRNAs with significant q-values of  $\alpha < 0.05$ . We end up with a list of 38 significant MiRNAs (**TABLE 10**).



*Table 10: List of Significant MiRNAs. MiRNAs with significant correlation based on linear mixed model results.*

<b>Mirna</b>	<b>Slope</b>	<b>P-Value</b>	<b>Q-Value</b>
hsa-mir-29b-2	-0.219	0	0
hsa-mir-29b-1	-0.216	0	0
hsa-mir-29a	-0.2	0	0
hsa-mir-138-2	-0.128	0	0
hsa-mir-142	-0.122	0	0
hsa-mir-3614	-0.117	0	0
hsa-mir-155	-0.112	0	0
hsa-mir-150	-0.099	0	0
hsa-mir-1269	0.046	0	0
hsa-mir-767	0.085	0	0
hsa-mir-105-2	0.092	0	0
hsa-mir-200c	0.098	0	0
hsa-mir-105-1	0.101	0	0
hsa-mir-1266	0.106	0	0
hsa-mir-3677	0.12	0	0
hsa-mir-452	0.126	0	0
hsa-mir-1180	0.129	0	0
hsa-mir-1911	0.137	0	0
hsa-mir-940	0.137	0	0
hsa-mir-320a	0.16	0	0
hsa-mir-27b	0.22	0	0
hsa-mir-423	0.237	0	0
hsa-mir-622	-0.421	0.001	0.0233
hsa-mir-3065	0.08	0.001	0.0233
hsa-mir-653	0.084	0.001	0.0233
hsa-mir-518b	0.102	0.001	0.0233

#### 4.3.3. REC Scores did not predict significant miRNAs

Python and R were run to calculate the REC scores. A successful REC score is considered to be less than -6.2 (Jacobsen et al., 2013).

*Table 11: REC Scores Results in Python. This list contains miRNA with the highest REC scores.*

<b>miRNA</b>	<b>REC Scores</b>	<b>Number of Cancers</b>
hsa-mir-219-2	-2.874	11
hsa-mir-561	-2.368	15
hsa-mir-7-2	-2.164	17
hsa-mir-518e	-1.862	7
hsa-mir-3676	-1.809	20
hsa-mir-548q	-1.802	14
hsa-mir-582	-1.612	20

#### 4.3.4. TargetScan predicts binding site of L1HS in several miRNAs

Likely miRNA target sites have near perfect miRNA sequence matching on base pairs 2-8, known as the seed sequence, and a low context score (Doron Betel et al., 2010). We found several miRNA that meets this criterion.

Table 12: List of miRNA transcripts from the short CDS. Using TargetScan we computed likely miRNA target sites. This list miRNA transcripts 8-mers and 7-mers with good context scores with short CDS.

miRNA Transcript	N-mer Length	Context Score
hsa-miRNA-138-5p	8mer-1a	-0.149
hsa-miRNA-150-5p	8mer-1a	-0.164
hsa-miRNA-2116-3p	7mer-1a	-0.133
hsa-miRNA-3127-3p	8mer-1a	-0.347
hsa-miRNA-6756-3p	8mer-1a	-0.347
hsa-miRNA-331-5p	8mer-1a	-0.237
hsa-miRNA-3657	7mer-1a	-0.102
hsa-miRNA-3667-3p	7mer-m8	-0.155
hsa-miRNA-455-5p	7mer-m8	-0.138
hsa-miRNA-4667-3p	7mer-1a	-0.137
hsa-miRNA-4669	7mer-1a	-0.194
hsa-miRNA-4713-5p	7mer-m8	-0.195
hsa-miRNA-5684	7mer-1a	-0.138
hsa-miRNA-657	8mer-1a	-0.225
hsa-miRNA-675-5p	7mer-1a	-0.146
hsa-miRNA-6826-3p	7mer-1a	-0.155
hsa-miRNA-6840-5p	7mer-1a	-0.261
hsa-miRNA-6841-5p	8mer-1a	-0.404
hsa-miRNA-6887-3p	7mer-m8	-0.149

*Table 13: List of miRNA transcripts from the long CDS. Using TargetScan we computed likely miRNA target sites. This list contains likely miRNA transcripts 8-mers and 7-mers with good context scores using the long CDS.*

<b>miRNA Transcript</b>	<b>N-mer Length</b>	<b>Context Score</b>
hsa-miRNA-3127-3p	8mer-1a	-0.225
hsa-miRNA-6756-3p	8mer-1a	-0.225
hsa-miRNA-331-5p	8mer-1a	-0.115
hsa-miRNA-4669	7mer-1a	-0.156
hsa-miRNA-4713-5p	7mer-m8	-0.135
hsa-miRNA-657	8mer-1a	-0.103
hsa-miRNA-675-5p	7mer-1a	-0.108
hsa-miRNA-6826-3p	7mer-1a	-0.117
hsa-miRNA-6840-5p	7mer-1a	-0.223
hsa-miRNA-6841-5p	8mer-1a	-0.282

Table 14: List of miRNA transcript 8-mer for short CDS. This list all the miRNAs that Targetscan found with the 8-mer matching of the seed sequence for the short CDS.

miRNA Transcript	Context Score
hsa-miRNA-138-5p	-0.149
hsa-miRNA-150-5p	-0.164
hsa-miRNA-3127-3p	-0.347
hsa-miRNA-6756-3p	-0.347
hsa-miRNA-331-5p	-0.237
hsa-miRNA-5580-3p	-0.091
hsa-miRNA-657	-0.225
hsa-miRNA-6841-5p	-0.404

Table 15: List of miRNA transcript 8-mer for long CDS. This list all the miRNAs that Targetscan found with the 8-mer matching of the seed sequence for the long CDS.

MiRNA Transcript	Context Score
hsa-miRNA-138-5p	-0.03
hsa-miRNA-150-5p	-0.042
hsa-miRNA-3127-3p	-0.225
hsa-miRNA-6756-3p	-0.225
hsa-miRNA-331-5p	-0.115
hsa-miRNA-5580-3p	-0.03
hsa-miRNA-657	-0.103
hsa-miRNA-6841-5p	-0.282

#### 4.3.5. Miranda predicts binding site of L1HS in several miRNAs

We identified miRNA target sites matching though L1HS 3' UTR region and CDS. The following miRNAs were predicted by the mixed model results, MiRanda, and TargetScan: hsa-mir-138-2-5p, hsa-mir-150-5p, hsa-mir-331-5p, hsa-mir-3127-3p. Several other miRNAs were predicted by MiRanda in the CDS of L1HS. As TargetScan only checks the 3' UTR these were not found by TargetScan: hsa-mir-29a, hsa-mir29b-1, hsa-mir29b-2.

```
3' gccggaCUAAGUGUUGUGGUCGa 5'
      | : |  ||   : |||||
5' ttagtgGGTGCAGCGCACCAAGCa 3'
```

Forward: Score: 145.000000 Q:2 to 10 R:3264 to 3285

Query: 3' uaugaugaagaaAUGUAAGGu 5'

Ref: 5' tagaagaaatggaTACATTCCt 3'

Energy: -10.130000 kCal/Mol

*Figure 18: Target Site for hsa-mir-138-2-5p. miRNA Transcript target site located at 5930 to 5952 in our L1HS 3' UTR region.*

Table 16: microRNA predicted by the mixed model, Targetscan, Rec score and MiRanda

MiRNA	Mixed Model	Targetscan Context Score	REC Score	MiRanda
hsa-mir-150	-.099	-0.042	-0.4843	<b>148.00 -19.28</b> (5p) Q:2 to 19 R:5886 to 5906
hsa-mir-138-1	-.077	-0.03	-0.7914	(3p) nothing (5p) <b>153 -18.02</b>
hsa-mir-138-2	-.128	-0.03	0.02512	<b>156.00 -14.33</b> (3p) Q:2 to 17 R:5379 to 5400 (3p) Q:2 to 9 R:3018 to 3039 <b>153 -18.02</b> (5p) Q:2 to 18 R:5930 to 5952 (5p) Q:2 to 19 R:1956 to 1978 (5p) Q:2 to 17 R:974 to 994
hsa-mir-622	-.421	N/A	0	<b>153 -23.98</b>
hsa-mir-3614	-.117	N/A	-0.2231	<b>154 -17.93</b> (3p) Q:2 to 19 R:1025 to 1047 (3p) Q:3 to 21 R:2770 to 2792 <b>141 -26.77</b> (5p) Q:2 to 22 R:674 to 696
hsa-mir-3127	.128	-0.225	0.50323	<b>145 -23.37</b> (3p) Q:2 to 19 R:5824 to 5846 (3p) Q:2 to 21 R:422 to 443 (3p) Q:2 to 18 R:5867 to 5888 <b>145.00 -21.23</b> (5p) Q:2 to 20 R:3337 to 3358
hsa-mir-29a	-.2	N/A	0.37903	<b>162.00 -17.08</b> (3p) Q:2 to 20 R:4774 to 4797 <b>160 -12.3</b> (5p) Q:2 to 20 R:2998 to 3019
hsa-mir-29b-1	-.216	N/A	0.57599	<b>Score:166 Energy: -17.99</b> (3p) Q:2 to 22 R:4774 to 4797 <b>150 -27.44</b> (5p) Q:2 to 23 R:3503 to 3526 (5p) Q:2 to 22 R:2668 to 2694 (5p) Q:3 to 23 R:4519 to 4541
hsa-mir-29b-2	-.219	N/A	0.20491	<b>166.00 -17.99</b> (3p) Q:2 to 22 R:4774 to 4797 <b>160 -12.3</b> (5p) Q:2 to 21 R:2796 to 2817



We looked at miRNAs that had strong results through multiple methods of testing. After using regressions to find targets we tried to find binding sites that could be responsible for L1Hs dysregulation. We found the mixed model to best predict L1HS and miRNA correlation. Mixed models do not have the underlying assumption of linear regression that doesn't allow secondary categorization

REC scores are a measure of recurring miRNA relationships in the pan-cancer network. This method is able to find miRNAs that correspond to findings in Targetscan and MiRanda. The REC scores we found don't support our targets.

We found several top-scoring candidate miRNA target sites from TargetScan. Unfortunately, we could not cross check these miRNAs against other methods as no data was available from them in TCGA.

We found a hsa-miR--29b relationship with NREP -had a REC score of -19.62.

## Conclusion

miRNA 150, 138, and 3127 do well in three measures Mixed model, Targetscan, MiRanda. miRNA 150, 138, and 3127 might be strong candidates for regulation of L1HS. Although the REC scores for these miRNAs are considered insignificant, they are among the lowest values closest to -6.2 that were received from the REC score analysis.

## 4.4. Conclusion

Overall, we measured the expression level of retrotransposon in different cancer tissues as well as healthy tissues. L1HS, one of the members of the LINE-1 family, was overexpressed significantly in 15 different cancer tissues compared to corresponding healthy tissues. L1HS expression level was not higher in STAD and THCA compared to corresponding healthy tissue. Also, we compared the L1HS expression level among different healthy tissues. We observed large number variation in terms of L1HS expression in different healthy tissues. STAD and ESCA had a much higher expression of L1HS compared to others. L1HS expression also varied across different individuals.

Next, we sought to find the association of L1HS expression with different factors including genes and miRNA. We predicted different genes and miRNA that might be associated with L1HS expression by different methods including linear regression, linear mixed model, REC score, TargetScan, and MiRanda.

## ***Chapter 5: Reference***

- Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 Elements in Structural Variation and Disease. *Annual Review of Genomics and Human Genetics*, 12(1), 187–215. <https://doi.org/10.1146/annurev-genom-082509-141802>
- Callinan, P. A., & Batzer, M. A. (2006). Retrotransposable Elements and Human Disease. In J.-N. Volff (Ed.), *Genome Dynamics* (pp. 104–115). Basel: KARGER. <https://doi.org/10.1159/000092503>
- Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., ... Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes & Development*, 28(13), 1397–1409. <https://doi.org/10.1101/gad.241661.114>
- Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10), 691–703. <https://doi.org/10.1038/nrg2640>
- Hancks, D. C., & Kazazian, H. H. (2012). Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development*, 22(3), 191–203. <https://doi.org/10.1016/j.gde.2012.02.006>
- Imbeault, M., Helleboid, P.-Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646), 550–554. <https://doi.org/10.1038/nature21683>
- Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., & Sander, C. (2013). Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural & Molecular Biology*, 20(11), 1325–1332. <https://doi.org/10.1038/nsmb.2678>

- Khetchoumian, K., Teletin, M., Mark, M., Lerouge, T., Cerviño, M., Oulad-Abdelghani, M., ...  
Losson, R. (2004). TIF1delta, a novel HP1-interacting member of the transcriptional  
intermediary factor 1 (TIF1) family expressed by elongating spermatids. *The Journal of  
Biological Chemistry*, 279(46), 48329–48341. <https://doi.org/10.1074/jbc.M404779200>
- Mathias, S., Scott, A., Kazazian, H., Boeke, J., & Gabriel, A. (1991). Reverse transcriptase  
encoded by a human transposable element. *Science*, 254(5039), 1808–1810.  
<https://doi.org/10.1126/science.1722352>
- Meroni, G., & Diez-Roux, G. (2005). TRIM/RBCC, a novel class of “single protein RING finger” E3  
ubiquitin ligases. *BioEssays: News and Reviews in Molecular, Cellular and Developmental  
Biology*, 27(11), 1147–1157. <https://doi.org/10.1002/bies.20304>
- Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are  
active in the human genome? *Trends in Genetics*, 23(4), 183–191.  
<https://doi.org/10.1016/j.tig.2007.02.006>
- Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., ... Hughes, T.  
R. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon.  
*Nature Biotechnology*, 33(5), 555–562. <https://doi.org/10.1038/nbt.3128>
- Tallant, C., Valentini, E., Fedorov, O., Overvoorde, L., Ferguson, F. M., Filippakopoulos, P., ...  
Ciulli, A. (2015). Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger  
and Bromodomain of the Chromatin Remodeling Complex NoRC. *Structure*, 23(1), 80–  
92. <https://doi.org/10.1016/j.str.2014.10.017>
- Wolf, D., & Goff, S. P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs.  
*Nature*, 458(7242), 1201–1204. <https://doi.org/10.1038/nature07844>

A YY1-binding site is required for accurate human LINE-1 transcription initiation. (n.d.).

Retrieved May 18, 2018, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC506791/>

Ahluwalia, J. K., Khan, S. Z., Soni, K., Rawat, P., Gupta, A., Hariharan, M., ... Brahmachari, S. K.

(2008). Human cellular microRNA hsa-miR-29a interferes with viral nef protein

expression and HIV-1 replication. *Retrovirology*, 5, 117. [https://doi.org/10.1186/1742-](https://doi.org/10.1186/1742-4690-5-117)

[4690-5-117](https://doi.org/10.1186/1742-4690-5-117)

An, W., Han, J. S., Wheelan, S. J., Davis, E. S., Coombes, C. E., Ye, P., ... Boeke, J. D. (2006). Active

retrotransposition by a synthetic L1 element in mice. *Proceedings of the National*

*Academy of Sciences*, 103(49), 18662–18667. <https://doi.org/10.1073/pnas.0605300103>

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-

throughput sequencing data. *Bioinformatics*, 31(2), 166–169.

<https://doi.org/10.1093/bioinformatics/btu638>

Arjan-Odedra, S., Swanson, C. M., Sherer, N. M., Wolinsky, S. M., & Malim, M. H. (2012).

Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but

not the replication of exogenous retroviruses. *Retrovirology*, 9(1), 53.

<https://doi.org/10.1186/1742-4690-9-53>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using

lme4. *ArXiv:1406.5823 [Stat]*. Retrieved from <http://arxiv.org/abs/1406.5823>

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 Elements in Structural

Variation and Disease. *Annual Review of Genomics and Human Genetics*, 12(1), 187–

215. <https://doi.org/10.1146/annurev-genom-082509-141802>

- Becker, K. G., Swergold, G. D., Ozato, K., & Thayer, R. E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Human Molecular Genetics*, 2(10), 1697–1702.
- Belancio, V. P., Roy-Engel, A. M., Pochampally, R. R., & Deininger, P. (2010). Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Research*, 38(12), 3909–3922.  
<https://doi.org/10.1093/nar/gkq132>
- Belgnaoui, S. M., Gosden, R. G., Semmes, O. J., & Haoudi, A. (2006). Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer Cell International*, 6, 13. <https://doi.org/10.1186/1475-2867-6-13>
- Betel, D., Wilson, M., Gabow, A., Marks, D. S., & Sander, C. (2007). The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(Database), D149–D153.  
<https://doi.org/10.1093/nar/gkm995>
- Betel, Doron, Koppal, A., Agius, P., Sander, C., & Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8), R90. <https://doi.org/10.1186/gb-2010-11-8-r90>
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127–135.  
<https://doi.org/10.1016/j.tree.2008.10.008>
- Bratthauer, G. L., Cardiff, R. D., & Fanning, T. G. (1994). Expression of LINE-1 retrotransposons in human breast cancer. *Cancer*, 73(9), 2333–2336.

- Callinan, P. A., & Batzer, M. A. (2006). Retrotransposable Elements and Human Disease. In J.-N. Volff (Ed.), *Genome Dynamics* (pp. 104–115). Basel: KARGER.  
<https://doi.org/10.1159/000092503>
- Chen, H., Lilley, C. E., Yu, Q., Lee, D. V., Chou, J., Narvaiza, I., ... Weitzman, M. D. (2006). APOBEC3A Is a Potent Inhibitor of Adeno-Associated Virus and Retrotransposons. *Current Biology*, 16(5), 480–485. <https://doi.org/10.1016/j.cub.2006.01.031>
- Chiu, Y.-L., & Greene, W. C. (2008). The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annual Review of Immunology*, 26, 317–353.  
<https://doi.org/10.1146/annurev.immunol.26.021607.090350>
- Clifford, R., Louis, T., Robbe, P., Ackroyd, S., Burns, A., Timbs, A. T., ... Schuh, A. (2014). SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. *Blood*, 123(7), 1021–1031. <https://doi.org/10.1182/blood-2013-04-490847>
- Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10), 691–703. <https://doi.org/10.1038/nrg2640>
- Didiano, D., & Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology*, 13(9), 849–851.  
<https://doi.org/10.1038/nsmb1138>
- Fedoroff, N. V. (2012). Transposable Elements, Epigenetics, and Genome Evolution. *Science*, 338(6108), 758–767. <https://doi.org/10.1126/science.338.6108.758>

- Filipowicz, W., Bhattacharyya, S. N., & Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews. Genetics*, 9(2), 102–114. <https://doi.org/10.1038/nrg2290>
- Gasior, S. L., Wakeman, T. P., Xu, B., & Deininger, P. L. (2006). The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *Journal of Molecular Biology*, 357(5), 1383–1393. <https://doi.org/10.1016/j.jmb.2006.01.089>
- Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., ... Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science*, 320(5879), 1077–1081. <https://doi.org/10.1126/science.1157396>
- Goodier, J. L., Cheung, L. E., & Kazazian, H. H. (2012). MOV10 RNA Helicase Is a Potent Inhibitor of Retrotransposition in Cells. *PLoS Genetics*, 8(10), e1002941. <https://doi.org/10.1371/journal.pgen.1002941>
- Hamdorf, M., Idica, A., Zisoulis, D. G., Gamelin, L., Martin, C., Sanders, K. J., & Pedersen, I. M. (2015). miR-128 represses L1 retrotransposition by binding directly to L1 RNA. *Nature Structural & Molecular Biology*, 22(10), 824–831. <https://doi.org/10.1038/nsmb.3090>
- Hasan, M., & Yan, N. (2014). Safeguard against DNA sensing: the role of TREX1 in HIV-1 infection and autoimmune diseases. *Frontiers in Microbiology*, 5, 193. <https://doi.org/10.3389/fmicb.2014.00193>
- Holmes, R. K., Malim, M. H., & Bishop, K. N. (2007). APOBEC-mediated viral restriction: not simply editing? *Trends in Biochemical Sciences*, 32(3), 118–128. <https://doi.org/10.1016/j.tibs.2007.01.004>



- Hu, S., Li, J., Xu, F., Mei, S., Le Duff, Y., Yin, L., ... Guo, F. (2015). SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation. *PLoS Genetics*, 11(7).  
<https://doi.org/10.1371/journal.pgen.1005367>
- Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *The American Journal of Human Genetics*, 92(6), 841–853. <https://doi.org/10.1016/j.ajhg.2013.04.015>
- Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., & Sander, C. (2013). Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural & Molecular Biology*, 20(11), 1325–1332. <https://doi.org/10.1038/nsmb.2678>
- Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). Tetrascripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, 31(22), 3593–3599. <https://doi.org/10.1093/bioinformatics/btv422>
- Kano, H., Godoy, I., Courtney, C., Vetter, M. R., Gerton, G. L., Ostertag, E. M., & Kazazian, H. H. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes & Development*, 23(11), 1303–1312.  
<https://doi.org/10.1101/gad.1803909>
- Kinomoto, M., Kanno, T., Shimura, M., Ishizaka, Y., Kojima, A., Kurata, T., ... Tokunaga, K. (2007). All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic Acids Research*, 35(9), 2955–2964. <https://doi.org/10.1093/nar/gkm181>
- Kretschmer, S., Wolf, C., König, N., Staroske, W., Guck, J., Häusler, M., ... Lee-Kirsch, M. A. (2015). SAMHD1 prevents autoimmunity by maintaining genome stability. *Annals of the*

- Rheumatic Diseases*, 74(3), e17–e17. <https://doi.org/10.1136/annrheumdis-2013-204845>
- Li, M., Fu, W., Wo, L., Shu, X., Liu, F., & Li, C. (2013). miR-128 and its target genes in tumorigenesis and metastasis. *Experimental Cell Research*, 319(20), 3059–3064. <https://doi.org/10.1016/j.yexcr.2013.07.031>
- Li, X., Zhang, J., Jia, R., Cheng, V., Xu, X., Qiao, W., ... Cen, S. (2013). The MOV10 Helicase Inhibits LINE-1 Mobility. *Journal of Biological Chemistry*, 288(29), 21148–21160. <https://doi.org/10.1074/jbc.M113.465856>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12). <https://doi.org/10.1186/s13059-014-0550-8>
- MacFarlane, L.-A., & Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. *Current Genomics*, 11(7), 537–561. <https://doi.org/10.2174/138920210793175895>
- Martin, S. L., & Branciforte, D. (1993). Synchronous expression of LINE-1 RNA and protein in mouse embryonal carcinoma cells. *Molecular and Cellular Biology*, 13(9), 5383–5392.
- McMillan, J. P., & Singer, M. F. (1993). Translation of the human LINE-1 element, L1Hs. *Proceedings of the National Academy of Sciences*, 90(24), 11533–11537. <https://doi.org/10.1073/pnas.90.24.11533>
- Minakami, R., Kurose, K., Etoh, K., Furuhashi, Y., Hattori, M., & Sakaki, Y. (1992). Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Research*, 20(12), 3139–3145.

- Nagamori, I., Kobayashi, H., Shiromoto, Y., Nishimura, T., Kuramochi-Miyagawa, S., Kono, T., & Nakano, T. (2015). Comprehensive DNA Methylation Analysis of Retrotransposons in Male Germ Cells. *Cell Reports*, 12(10), 1541–1547.  
<https://doi.org/10.1016/j.celrep.2015.07.060>
- Niewiadomska, A. M., Tian, C., Tan, L., Wang, T., Sarkis, P. T. N., & Yu, X.-F. (2007). Differential Inhibition of Long Interspersed Element 1 by APOBEC3 Does Not Correlate with High-Molecular-Mass-Complex Formation or P-Body Association. *Journal of Virology*, 81(17), 9577–9583. <https://doi.org/10.1128/JVI.02800-06>
- Obbard, D. J., Gordon, K. H. J., Buck, A. H., & Jiggins, F. M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1513), 99–115.  
<https://doi.org/10.1098/rstb.2008.0168>
- Reichmann, J., Crichton, J. H., Madej, M. J., Taggart, M., Gautier, P., Garcia-Perez, J. L., ... Adams, I. R. (2012). Microarray Analysis of LTR Retrotransposon Silencing Identifies Hdac1 as a Regulator of Retrotransposon Expression in Mouse Embryonic Stem Cells. *PLoS Computational Biology*, 8(4), e1002486. <https://doi.org/10.1371/journal.pcbi.1002486>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Rodić, N., Sharma, R., Sharma, R., Zampella, J., Dai, L., Taylor, M. S., ... Burns, K. H. (2014). Long Interspersed Element-1 Protein Expression Is a Hallmark of Many Human Cancers. *The*

- American Journal of Pathology*, 184(5), 1280–1286.  
<https://doi.org/10.1016/j.ajpath.2014.01.007>
- Silverman, R. H. (2007). Viral encounters with 2',5'-oligoadenylate synthetase and RNase L during the interferon antiviral response. *Journal of Virology*, 81(23), 12720–12729.  
<https://doi.org/10.1128/JVI.01471-07>
- Skowronski, J., Fanning, T. G., & Singer, M. F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. *Molecular and Cellular Biology*, 8(4), 1385–1397.
- Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4), 272–285. <https://doi.org/10.1038/nrg2072>
- Stetson, D. B., Ko, J. S., Heidmann, T., & Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. *Cell*, 134(4), 587–598.  
<https://doi.org/10.1016/j.cell.2008.06.032>
- Transposable elements and miRNA: Regulation of genomic stability and plasticity: Mobile Genetic Elements: Vol 6, No 3. (n.d.). Retrieved May 23, 2017, from <http://www.tandfonline.com/doi/full/10.1080/2159256X.2016.1175537>
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63. <https://doi.org/10.1038/nrg2484>
- Willems, L., & Gillet, N. A. (2015). APOBEC3 Interference during Replication of Viral Genomes. *Viruses*, 7(6), 2999–3018. <https://doi.org/10.3390/v7062757>
- Wissing, S., Montano, M., Garcia-Perez, J. L., Moran, J. V., & Greene, W. C. (2011). Endogenous APOBEC3B Restricts LINE-1 Retrotransposition in Transformed Cells and Human

- Embryonic Stem Cells. *Journal of Biological Chemistry*, 286(42), 36427–36437.  
<https://doi.org/10.1074/jbc.M111.251058>
- Xiao-Jie, L., Hui-Ying, X., Qi, X., Jiang, X., & Shi-Jie, M. (2016). LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genetics in Medicine*, 18(5), 431–439.  
<https://doi.org/10.1038/gim.2015.119>
- Yan, N., Regalado-Magdos, A. D., Stiggelbout, B., Lee-Kirsch, M. A., & Lieberman, J. (2010). The cytosolic exonuclease TREX1 inhibits the innate immune response to human immunodeficiency virus type 1. *Nature Immunology*, 11(11), 1005–1013.  
<https://doi.org/10.1038/ni.1941>
- Yang, N., Zhang, L., Zhang, Y., & Kazazian Jr, H. H. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Research*, 31(16), 4929–4940.
- Zamudio, N., Barau, J., Teissandier, A., Walter, M., Borsos, M., Servant, N., & Bourc'his, D. (2015). DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes & Development*, 29(12), 1256–1270.  
<https://doi.org/10.1101/gad.257840.114>
- Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., ... Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3), 453–466.  
<https://doi.org/10.1016/j.cell.2012.12.023>
- Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T., Teng, C., ... Jacobsen, S. E. (2015). A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell*, 163(2), 445–455. <https://doi.org/10.1016/j.cell.2015.09.032>

## ***Chapter 6: Curriculum Vitae***

GM Jonaid  
4247 Claymont ST., Apt# 3, Las Vegas, NV 89119.  
Email: gmbd97@gmail.com  
Mobile: 702-502-8832

### **RESEARCH INTERESTS:**

Genomics, functional genomics, bioinformatics, computational biology, system biology, machine learning, algorithms.

### **EDUCATION:**

**Master of Science in Biological Science (Cell and Molecular Biology concentration), 2018**  
University of Nevada, Las Vegas.

**Bachelor of Science in Microbiology, 2012**  
University of Dhaka, Bangladesh

### **RESEARCH EXPERIENCE:**

#### **Graduate research experience:**

**Thesis:** Understanding the expression of retrotransposons in human somatic cells

- Expression of retrotransposons in different cancer types
- Identifying specific miRNAs responsible for controlling retroactivity
- Identifying association between SNPs and retrotransposon expression

**Undergraduate Thesis:** Advanced laboratory detection of *Vibrio cholerae* and ETEC at ICDDR,B, 2012.

- Detection of colonization factors and surface antigens of *Vibrio cholerae* and ETEC
- Detection of toxic gene of ETEC by multiplex PCR

### **KEY GRADUATE COURSES TAKEN:**

Multivariate Analysis, Analysis of Algorithms, Genomics , Molecular Biology, Biostatistics

### **WORK EXPERIENCE:**

Teaching: Teaching assistant, BIOL 189 - Fundamentals of Life Science, University of Nevada, Las Vegas. (Fall 2015 - Current )

Teaching: GRE and SAT , May 2014- 2015

Teaching: Part time instructor, A' level Mathematics, 2008 -2015

### **BASIC SKILLS:**

Programming : Python, R, C++, Unix, Linux, SPSS

Adobe Illustrator, Adobe Photoshop

ELISA, SDS-PAGE, PCR, Western blot

**AWARDS:**

- Graduate Access Funds Scholarship
- Daily Star Award

**PUBLICATIONS AND PRESENTATIONS:**

**G.M. Jonaid**, Sophia Quinton, Nicky Chung, Daphnie Churchill, Adrian Alberto, Cody Clymer, Austin Ross, Omar Navarro Leija, Mira V. Han. Variation in L1HS expression in 17 somatic tissues. 2017. (In preparation)

**G.M. Jonaid**, Sophia Quinton, Daphnie Churchill, Nicky Chung, Cody Clymer, Adrian Alberto, Austin Ross, Omar Navarro Leija, Mira V. Han. (2017). Understanding retrotransposons expression in human somatic cells. ISMB (Intelligent Systems for Molecular Biology) 2017. Prague, Czech Republic.

**G.M. Jonaid**, Austin Ross, Nicky Chung, Elizabeth Park, Omar Navarro Leija, Sophia Quinton, Daphnie Churchill, and Mira Han. (2016). Understanding transposable element dysregulation in human somatic cells. Submission 487. ISMB (Intelligent Systems for Molecular Biology) 2016. Orlando, Florida.

**G.M. Jonaid**. Transposons Genetics, NIPM (Nevada Institute Of Personalized Medicine) Symposium, 2017.

Nicky Chung, Austin Ross, Adrian Alberto, Daphnie Churchill, Cody Clymer, **G.M. Jonaid**, Sophia Quinton, and Mira Han. (2017). Analyses of novel transposable element insertions in cancerous genomes. Spring 2017 Undergraduate Research Forum. University of Nevada, Las Vegas.

Sophia Quinton, Mira Han, Daphnie Churchill, **G.M. Jonaid**, Nicky Chung, Cody Clymer, Adrian Alberto, Omar Navarro Leija, Elizabeth Park. (2017). Identifying Gene and MicroRNA Control of L1HS Retrotransposon expression in Human Somatic Cells. Spring 2017 Undergraduate Research Forum. University of Nevada, Las Vegas.

Sophia Quinton, Daphnie Churchill, Omar Navarro Leija, **G.M. Jonaid**, Elizabeth Park, Nicky Chung, Mira Han. (2016). Identifying MicroRNA Control of L1HS Retrotransposon Expression in Human Cells. Minority Access INC Student Researchers Competition. Washington, D.C.

**VOLUNTARY ACTIVITIES:**

- Secretary, ISCB California and Nevada region student organization
- GMAT and SAT content maker for Higher Study Abroad group
- Volunteering: JAAGO Foundation, Volunteer of Bangladesh
- 1 degree initiative: 100 smile

