12-3-2020

Regulation of Transposable Elements by Tumor Suppressor Protein 53

Andrew J. Rosato
ajr7533@rit.edu

Regulation of Transposable Elements by Tumor Suppressor Protein 53

Authored by Andrew J. Rosato

Advised by Dr. Feng Cui

Committee Members: Dr. Gregory Babbitt, Dr. Feng Cui, Dr. Leslie Kate Wright

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Thomas H. Gosnell School of Life Sciences

College of Science

Rochester Institute of Technology

Rochester, NY

December 3, 2020

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Andrew Rosato, a candidate for the Master of Science degree in Bioinformatics, has submitted his thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	**Date**
Feng Cui, Ph.D. Thesis Advisor	
Gregory Babbitt, Ph.D.	
Leslie Kate Wright, Ph.D.	

Acknowledgements

I would like to thank Julia Freewoman of the Rochester Institute of Technology for performing the Nutlin and DMSO chemical treatment work on the IMR90 RNA-Seq samples.

I would like to thank the University of Rochester Genomics Research Center for sequencing and performing quality filtering on the IMR90 RNA-Seq data.

I would like to thank my thesis committee members Dr. Gregory Babbitt and Dr. Leslie Kate Wright for their guidance throughout the project and especially for their help with shaping the manuscript pushing me to communicate my work more effectively.

I would especially like to thank my thesis advisor Dr. Feng Cui for his continued support and guidance. I have worked in Dr. Cui's lab since my sophomore year and his constant encouragement and driven pursuit of knowledge has inspired me to become the scientist I am today.

Table of Contents

A. Abstract:

Multiple transposable elements have been identified by colocalization analysis that display a strong predicted regulatory relationship with p53 associated peaks. RNA-Seq was used to identify differentially expressed transposable elements. ChIP-Seq was used to identify peaks representing transcription factor binding sites in p53 activated cells. The results of both experiments were combined in a colocalization analysis identifying transposable element locations that were both differentially regulated and located near p53 associated peaks. The colocalization of ChIP-Seq and RNA-Seq analyses allows for the verification of p53's regulatory role in the expression of transposable elements across the genome. A Monte Carlo simulation was performed verifying that the frequency of the colocalizations observed occurred more frequently than due to random chance.

B. Introduction:

Cancers are an extremely complex and deadly disease in which tumor suppressor protein 53 (p53) plays an integral role in preventing. Understanding the equally complex regulatory mechanism of p53 enables the development of novel treatment strategies and therapies for cancers. Previous studies have shown up to 40% of p53 associated ChIP-Seq fragments located inside of transposable elements. This project identifies and investigates a proposed regulatory relationship that p53 has to various transposable elements throughout the human genome. Transposable elements regulated by p53 were identified in both normal IMR90 lung fibroblasts and cancerous HCT116 colorectal p53 wild type and p53 knockout cells. This was accomplished by identifying instances of genomic colocalization between the results of RNA-Seq and ChIP-Seq experiments. RNA-Seq was used to identify differentially expressed transposable elements. ChIP-Seq was used to identify significant peaks representing transcription factor binding sites in p53 activated cells compared against control cells. The results of both experiments were then combined to identify transposable element locations that were both differentially regulated and located near peaks in tissue samples with activated p53. The comparison of ChIP-Seq and RNA-Seq data allows for the verification of p53's regulatory role in the expression of transposable elements across the genome. This better understanding of the p53 regulatory network may lead to the advent of new treatments for destructive diseases like cancer that are kept in check by this mechanism.

Transposable Elements

Transposable elements are thought to make up over 50% of the human genome (SanMiguel, 1996). Transposable elements also called transposons, or "jumping genes," possess the ability to replicate themselves and insert themselves in other locations across the genome

(McClintock, 1950). There are two major classes of transposable elements. Approximately 90% of transposable elements in humans are class I retrotransposons, which move across the genome with the help of RNA intermediaries, unlike class II DNA transposons, which do not employ an RNA intermediary during transposition (Pray, 2008). This "copy and paste" process of replication and insertion is thought to be why transposons comprise such a large portion of the human genome (Kazazian, 2004). This process, however, can have detrimental effects if a transposon is inserted in the middle of a gene or if two exons are spliced together after a transposon is excised from its location in the genome (Chuong, 2017). The cell has many mechanisms to silence otherwise intact functional transposons from employing epigenetic changes such as DNA methylation to the use of miRNAs to degrade RNA transcripts and chromatin remodeling to inactivate large regions of the chromosome. There are multiple distinct groups of transposable elements, members of which have been found to be actively involved in cellular regulatory networks (Rebollo, 2012).

Alu elements, originally identified in the mechanism of an endonuclease in the Arthrobacter luteus (ALU) bacteria, are a type of primate-specific repetitive element belonging to the short-interspersed element (SINE) order of retroelements. Alus are non-autonomous, but active elements, that still possess the ability to replicate throughout the genome with the help of trans-activating factors from LINEs (Dewannieux, 2003). The active nature of Alus makes them one of the most common mobile elements making up roughly 11% of the human genome (Lander, 2001). Alu elements have been observed influencing gene expression in multiple ways with noted effects on gene splicing, polyadenylation, and in adenosine deaminase that acts on RNA (ADAR) editing (Chen, 2009; Shen, 2011; Dominissini, 2011).

Mammalian-wide interspersed repeats (MIRs) are tRNA-derived members of the SINE order of retroelements and make up roughly 2.5% of the total human genome (Lander, 2001). MIRs are the oldest family of transposable elements and as such one of the most highly conserved. MIRs have been found to be some of the most functionally relevant transposable elements to gene expression, providing the human genome with microRNAs, enhancer sequences, and transcription factor binding sites (Piriyapongsa, 2007; Jjingo, 2014; Teng, 2011).

Long interspersed element 1 (LINE-1 or L1) is a member of the LINE order of retrotransposons and is the most prevalent mobile element in the human genome, comprising approximately 17% of the total genome (Lander, 2001). Most L1 elements are no longer active in the human genome, however, some are still capable of transposition. L1 elements move using an RNA intermediate created by an L1-encoded reverse transcriptase and other L1 associated proteins. These regions are mutated in many instances, inactivating many L1 elements (Scott, 2013). L1 elements have been found to mostly impact gene expression by disrupting gene expression due to transposition into protein-coding regions.

Human endogenous retroviruses (HERVs) are members of the long terminal repeat (LTR) order of retrotransposons. HERVs are remnants of retroviruses that have inserted themselves into the human genome, accounting for roughly 8% of the total genome (Lander, 2001). HERVs contain a provirus derived structure of open reading frames typically flanked by two LTR regions (Hurst, 2017). These LTR regions have some functionality as promoters of the HERVs, containing transcription factor binding sites (Manghera, 2013). HERVs have been implicated in the expression of protein-coding genes as well as with the expression of regulatory long non-coding RNAs (Dunn, 2006; Laurent, 2013). A majority of HERVs are inactive due to the gradual accumulation of mutations and from targeted epigenetic silencing preventing the

expression of unwanted protein-coding genes. Transposable elements have been found to play an important, but understudied role in cellular regulatory mechanisms. The regulation of these mechanisms can influence the occurrence of various human diseases, such as cancers.

Tumor Protein 53

Tumor Suppressor Protein 53 (p53), also referred to as the "master guardian of the genome," is a transcription factor responsible for inducing cell cycle arrest or apoptosis in response to potential DNA damaging stressors including hypoxia, ultraviolet radiation, radioactive compounds, heat-shock, and cellular expression of viral vectors (Lane, 1994). P53 has an essential role in preventing cancer with more than half of all primary tumors carrying a p53 mutation. Monomeric p53 has a relatively short half-life in human cells, only remaining stable for about 38 minutes in normal cells under normal conditions (Baresova, 2014). In a cell not under genotoxic stress, negative regulators MDM2 and MDM4 bind to the transcriptional activation domains (TAD) of p53 monomers, preventing transactivation (Momand 1992; Davoni 2004). MDM2 has an additional function as an E3 ubiquitin ligase, which induces proteasome-mediated degradation of p53, maintaining low p53 levels in normal cells (Haupt, 1997). When a cell experiences a double-stranded DNA break event, CHK1 and CHK2 kinases phosphorylate the TAD of p53 preventing MDM2 from inhibiting transactivation and allowing p53 to accumulate in its stable tetrameric confirmation (Kastan, 2004). Once stable p53 can transcriptionally activate a variety of signaling pathways to pause cell cycle progression, initiate DNA repair mechanisms, cause cellular senescence, or trigger apoptosis depending on the type and extent of DNA damage.

The chemical treatments Dimethyl sulfoxide and Nutlin-3a can be used to experimentally regulate the expression of TP53 in a sample of exposed cells. Nutlin-3a is a small molecule

inhibitor that binds to MDM2 which results in increased p53 expression across cell types (Kumamoto, 2008). Dimethyl sulfoxide, also known as DMSO, is a solvent Nutlin-3a is commonly dissolved in. During an experimental investigation of p53, cells treated with Nutlin-3a dissolved in DMSO could be compared to a control group of cells treated only with DMSO to identify p53 associated genetic elements and transcription factor binding.

P53 has demonstrated a significant ability to influence and repress retrotransposon activity defending the cell from additional retrotransposon replication events (Wylie, 2016). In previous analyses, it has been shown that around 40% of p53 ChIP fragments are found inside transposable elements such as LTR Class I ERVs, SINEs, and LINEs (Bao, 2017, Wang, 2007). This further suggests a regulatory role of p53 to retrotransposons. Further analysis to identify which specific sites are functionally relevant and how those sites influence retrotransposon activity would further our understanding of the p53 retrotransposon relationship. Investigators have studied the regulatory mechanism of p53 using various traditional strategies, including RNA-Seq and CHIP-Seq. However, the target elements of this analysis are in repetitive regions of DNA which have proved uniquely difficult to align using traditional RNA-Seq and ChIP-Seq methods.

Overcoming Alignment Challenges

Reads that originate from repetitive regions of DNA present a unique challenge for traditional read alignment programs. Shorter length reads have an increased potential to map to multiple locations across the genome, this is especially true in reads from transposable elements due to the repetitive nature of their base pair sequence. This phenomenon of a read mapping to multiple locations is referred to as a read being multi-mapped, or simply as a multi-read. In many standard bioinformatics data analysis protocols, multi-reads are recommended to be excluded

from the analysis, rather than attempt to determine their origin. All alignment programs have parameters that allow the user to select if and how many alternative alignment sites should be included in the alignment analysis. However, it can be computationally expensive to allow a read aligner to search for all potential alignment sites. Many protocols, in fact, suggest disregarding multi-reads for this reason. The exclusion of multi-reads from an analysis represents the loss of a substantial amount of data and should be reconsidered with the increasing availability of tools designed to identify a multi-read's location of origin. Multi-read location assignment tools have been developed for both RNA-Seq and ChIP-Seq analyses. In all cases, these multi-read assignment tools work by determining the best possible mapping location from multiple alignment locations. Alignment programs needed to be instructed to identify and retain multiple possible alignment locations for each read to be used in these tools. Telescope, a software tool developed for RNA-Seq data, works using an expectation-maximization algorithm to directly assign multi-reads to their most likely location of origin on a reference genome (Bendall, 2019). CSEM, a software tool developed for ChIP-Seq data, works by implementing an expectation-maximization algorithm to score multi-reads based on their likelihood of being correctly mapped and then retaining only the highest scoring multi-read (Chung, 2011). In this work, RNA-Seq and ChIP-Seq technologies used in combination with these tools were employed to identify differential expression and transcription factor binding in these difficult to map repetitive element regions.

C. Materials and Methods:

Data analysis was performed on transcriptional data obtained from normal and cancerous cells treated with DMSO or Nutlin. Genetic information was isolated from samples using both RNA-Seq and ChIP-Seq sequencing technologies. Sequenced HCT116 colorectal cancer cell samples were obtained from the Espinosa Lab at the University of Colorado Anschutz Medical Campus and processed in this workflow with both ChIP-Seq and RNA-Seq analysis (Espinosa, 2017). Samples of IMR90 normal noncancer cells generated for ChIP-Seq analysis were obtained from the Sammons Lab at the University of Pennsylvania (Sammons, 2015). Additional IMR90 normal noncancerous cell samples were generated for RNA-Seq analysis by the Cui Lab at the Rochester Institute of Technology. Table 1 contains summary information about each sample analyzed including, sample ID, cell type, sequencing method, and chemical treatment. Espinosa and Sammons Lab RNA-Seq and ChIP-Seq data were downloaded as FASTQ files of total RNA from samples sequenced in an Ion Torrent Proton sequencer and Illumina HiSeq 2000 sequencer, respectively. Raw FASTQ files were downloaded from the Sequence Read Archive (SRA) using the *prefetch* command from the SRAtoolkit software suite with each sample's SRA ID.

Figure 1 shows a visual overview of the analysis steps used in this project, the analysis was performed on both the HCT116 and IMR90 cell lines. First, raw reads were obtained, and quality control was performed. Reads were then aligned to the hg19 reference genome. Subsequently, both the unique and multiple mapping RNA-Seq reads were assigned to repetitive elements with the Telescope program producing a counts table. Differential expression analysis was then performed on the counts table identifying differentially expressed repetitive elements between the DMSO and Nutlin sample groups. Then CSEM program was used to create a ChIP-

Seq read set containing unique reads and the highest scoring multi-reads. Peak calling was then performed on these reads using MACS2 to identify significant peaks between the Nutlin and the control group samples. These peaks and repetitive elements were then colocalized, identifying where the two datasets overlapped. A Monte Carlo simulation was performed verifying that the frequency of the colocalizations observed occurred more frequently than due to random chance.

Flowchart:

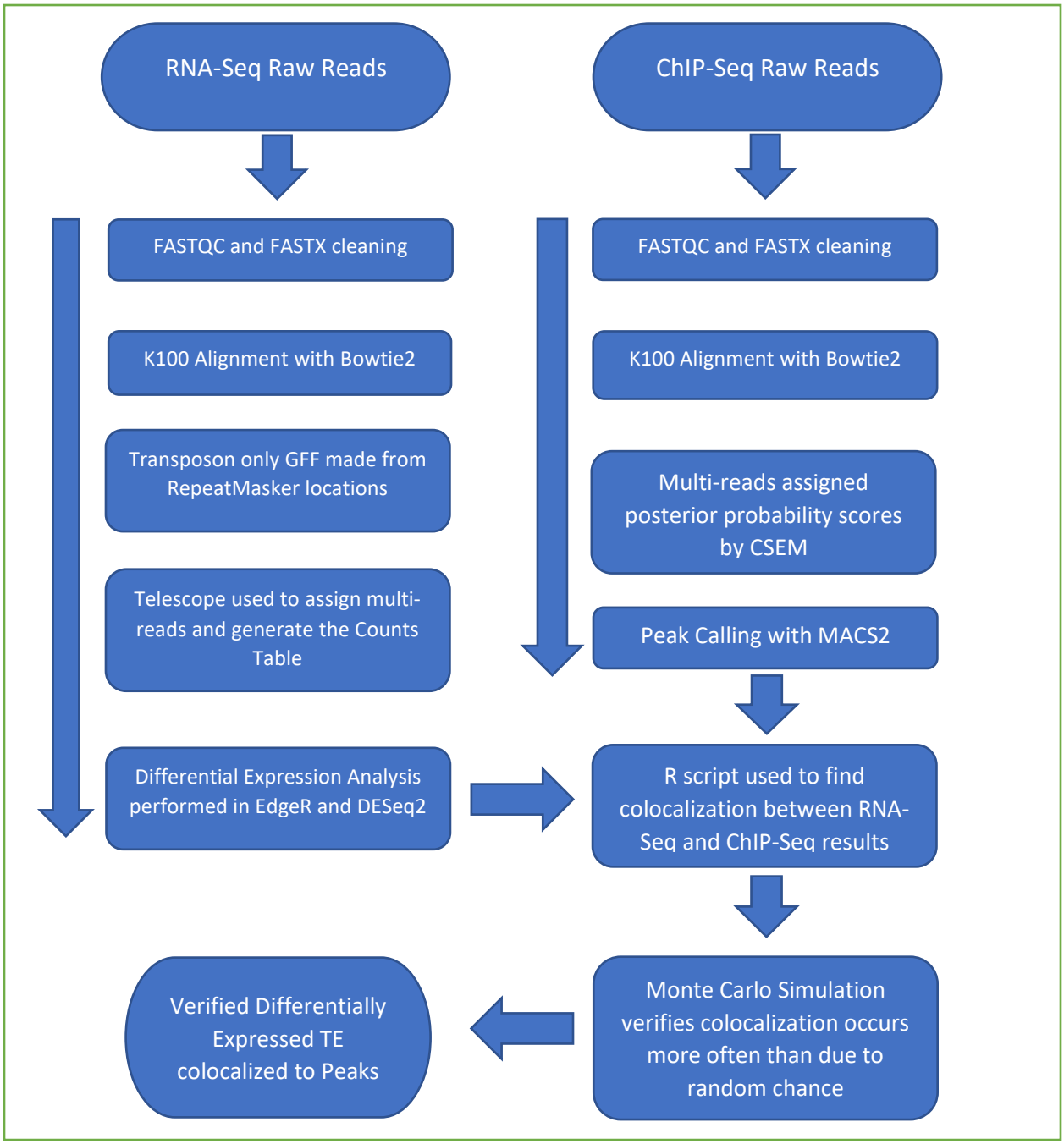

Figure 1: Flowchart depicting the analysis workflow of the project from raw sequenced reads to the identification of colocalized transposable elements.

Table 1: Experimental Sample Overview

Sample ID	Cell Line	Treatment	Replicate	Analysis Type	Cancerous
SRR4098426	p53ko_HCT116	DMSO	1	Total RNA-Seq	Cancer
SRR4098427	p53ko_HCT116	DMSO	2	Total RNA-Seq	Cancer
SRR4098428	p53ko_HCT116	Nutlin	1	Total RNA-Seq	Cancer
SRR4098429	p53ko_HCT116	Nutlin	2	Total RNA-Seq	Cancer
SRR4098430	p53wt_HCT116	DMSO	1	Total RNA-Seq	Cancer
SRR4098431	p53wt_HCT116	DMSO	2	Total RNA-Seq	Cancer
SRR4098432	p53wt_HCT116	Nutlin	1	Total RNA-Seq	Cancer
SRR4098433	p53wt_HCT116	Nutlin	2	Total RNA-Seq	Cancer
SRR4090089	HCT116	N/A	1	ChIP-Seq	Cancer
SRR4090090	HCT116	DMSO	1	ChIP-Seq	Cancer
SRR4090091	HCT116	Nutlin	1	ChIP-Seq	Cancer
SRR1448786	IMR90	DMSO	1	ChIP-Seq	Normal
SRR1448787	IMR90	Nutlin	1	ChIP-Seq	Normal
SRR1448788	IMR90	DMSO	1	ChIP-Seq	Normal
SRR1448789	IMR90	Nutlin	1	ChIP-Seq	Normal
clt_Control1	IMR90	N/A	1	Total RNA-Seq	Normal
clt_Control2	IMR90	N/A	2	Total RNA-Seq	Normal
clt_DMSO1	IMR90	DMSO	1	Total RNA-Seq	Normal
clt_DMSO2	IMR90	DMSO	2	Total RNA-Seq	Normal
clt_Nutlin1	IMR90	Nutlin	1	Total RNA-Seq	Normal
clt_Nutlin2	IMR90	Nutlin	2	Total RNA-Seq	Normal

Table 1: A table that lists all samples used in the analysis and indicates each sample's SRR ID, cell line, treatment chemical, experimental replicate, sequencing data type, and if the tissue is cancerous or normal.

RNA-Seq Analysis

RNA-Seq Quality Control

Quality control for the eight HCT116 RNA-Seq samples obtained from the Espinosa Lab was performed manually using the FASTQC and FASTX toolkit. Quality control for the six IMR90 RNA-Seq samples generated by the Cui Lab at RIT was performed by the University of Rochester Genomics Research Center (URGRC) using FastP, an all in one FASTQ file processing program. The command *fastp 0.20.0, --in1 ../${SAMPLE}_R1.fastq.gz --out1 clt_${SAMPLE}_R1.fastq.gz --length_required 35 --cut_front_window_size 1 --cut_front_mean_quality 13 --cut_front --cut_tail_window_size 1 --cut_tail_mean_quality 13 --cut_tail -w 8 -y -r -j ${SAMPLE}_fastp.json* was used to process all six raw read files. This command removes all sequences shorter than 35bps, and all bases lower than an average Phred quality score of 13 in an 8bp sliding window.

The sequence quality of the eight HCT116 raw reads was first assessed using the FASTQC tool. FASTQC was run to obtain multiple quality metrics namely, per base sequence quality score, base-pair distribution, percent GC content, and the presence of any overrepresented sequences. FASTQC was run on each of the raw FASTQ input data files and an HTML report with the previously listed metrics was generated for each file. FASTQC was run using the command *fastqc*, specifying the input raw FASTQ file's location, and the location the quality report should be written to. The reports for each of the input FASTQ files were visually inspected to determine how the data needed to be cleaned before alignment.

The eight raw HCT116 RNA-Seq samples contained mixed quality reads of varying lengths. FASTQC flagged the Per base sequence quality, Per base sequence content, and k-mer

content to be unacceptably high in addition to flagging several other metrics as being potentially problematic.

The FASTX toolkit version 0.13.2 was selected to perform the functions necessary to clean the data. The Per base sequence content indicated that base pair distributions at the start and end of the reads were irregular and needed to be trimmed. The outside regions of the reads, base positions 0-9 and 150+, were removed with the FASTX trimmer. The trimmer ran using the command *fastx_trimmer -i {input_filename} -o {output_filename} -f 10 -l 150 -Q33*. After removing the problematic regions at both ends of the sequences, the remaining reads with low-quality scores needed to be removed. In addition, the dataset had an abnormally high amount of short length reads that needed to be removed. The FASTX quality filter was used to both remove reads with an average quality score of less than 20 and to remove reads that were shorter than 10 base pairs in length. The quality filter ran using the command *fastq_quality_trimmer -i {input_filename} -o {output_filename} -t 20 -l 10 -Q33*

FASTQC was run again on the RNA-Seq data after cleaning and showed significantly improved data quality. The data cleaning process removed between x and x reads from each sample. This brought all the Per base sequence quality scores at each read position above the poor reference metric average quality score of 20 and into the medium to high-quality range.

RNA-Seq Processed Read Alignment

After cleaning, the sequenced reads contained each sample's FASTQ file needed to be aligned to a human reference genome. The Bowtie2 read aligner was selected to align the raw reads to a reference genome. The repetitive element locations used in the downstream analysis were derived from human genome version GRCh37/hg19, as such hg19 was selected as the

template genome. Index files for hg19 were precompiled before alignment to accelerate the alignment process. Index files were generated using the command *build index hg19_ind* where hg19_ind was the index file prefix to be given to Bowtie2 during alignment. Alignment with Bowtie was run on each of the cleaned input FASTQ files using the command *bowtie2 -p 35 -x hg19_ind -U {input.fastq} -S {output.sam} -k 100*, where -k 100 instructs bowtie to identify up to 100 possible alternative mapping sites where a given read could map. The aligned reads from each sample were outputted in a sequence alignment format (SAM) format file. The Samtools program was then used to convert the SAM files of each sample into BAM files, which take up less space.

RNA-Seq Read Assignment to Repetitive Elements

After alignment to a reference genome, the reads needed to be assigned to specific genomic features to tally how many reads in each sample were associated with each feature. TE assignment was performed using the Telescope program developed by the Brenner Lab at George Washington University. Telescope was run using the command *Telescope assign* provided with an input SAM file of aligned reads and a gene transcript format (GTF) file of the genomic element locations where reads can be assigned.

In this project, only regions of transposable elements (TEs) are of interest. To further simplify the analysis, the genome needed to be restricted to only those TE regions. A text file of all TE names and locations in the hg19 reference genome was downloaded from the RepeatMasker website. This text file was converted to a general feature format 3 (GFF3) file using the rmOutToGFF3.pl Perl script, included in the RepeatMasker software suite. The GFF3 file of TEs was then converted to GTF format using a custom R script GFF3_to_GTF.R. This script created a unique TE name by combining the original RepeatMasker name with the

chromosome number, start, and stop positions of the RE. This new unique TE name allowed for the quantification of the expression of specific individual instances of the TE. The GTF file was created by combining all GFF3 data columns, except for the "Phase," and including a custom attribute in the attribute column named "Locus," containing the unique TE name. The Telescope assign program selects the Locus attribute by default as the name of the genomic feature where reads are assigned.

Telescope assign was run on each aligned read SAM file and a report file containing alignment statistics for each file was generated. A custom R script was used to combine the data in the unique count column from each of the reports into a counts table of all unique TEs in the experiment. Any columns with TEs that had no data recorded for them were marked as NA by the Telescope program, and these features were manually reassigned a count value of 0.

RNA-Seq Differential Expression Analysis

Differential expression analysis was then performed on these counts tables in R using both the EdgeR package and the DESeq2 package. Transposable elements were filtered before analysis to remove any transcript that was not present at least once in four of the eight samples. In addition, TEs were removed if they did not contain at least five counts per one million transformation counts. The respective analyses were run according to their respective protocols using supplied default variable values. This resulted in two lists of differentially expressed transcripts located in TE regions along with their p-value and log fold change value. The results of the programs were compared to identify transcripts mapped to TEs that overlapped in both analyses which were then exported for further analysis.

ChIP-Seq Analysis

ChIP-Seq Quality Control

The ChIP-Seq data were prepared for analysis in a similar way to the RNA-Seq data, identifying and removing problematic reads. FASTQ files from each ChIP-Seq experiment were downloaded from the Sequence Read Archive using the prefetch command from the SRAtoolkit software suite. Once downloaded, the reads contained in each sample's FASTQ file were aligned to a reference genome. The three raw ChIP-Seq samples of the HCT116 dataset contained mostly high-quality reads that did not require filtering.

ChIP-Seq Processed Read Alignment

The cleaned ChIP-Seq reads followed a similar alignment protocol to the RNA-Seq reads. The Bowtie2 read aligner was used to align the raw ChIP-Seq reads to the GRCh37/hg19 reference genome. Index files were precompiled with the prefix hg19_ind which was given to Bowtie2 during alignment. Alignment with Bowtie2 was run on each of the cleaned input FASTQ files using the command *bowtie2 -p 35 -x hg19_ind -U {input.fastq} -S {output.sam} -k 100,* where *-k 100* instructs bowtie2 to identify up to 100 possible alternative mapping sites where a given read could map. The aligned reads from each sample were outputted in a SAM format file and then converted to a BAM file.

ChIP-Seq Multiread Processing

In a typical peak calling protocol, aligned reads, which map to more than one location, called multi-reads, are first removed from the analysis. Multireads typically complicate the peak calling analysis because their location of origin is difficult to determine. The ChIP-Seq multi-read allocation using Expectation-Maximization (CSEM) tool was used to score the probability

of how likely an individual reported multi-read was the read's actual location of origin. CSEM version 2.4 was downloaded from the Dewey Lab at the University of Wisconsin. Read probability scores were calculated for every read in each sample file using the command *run-csem --sam -p 10 input_name.sam 51 output_name_csem*. In the run-csem command, *-sam* specifies the input file format, -p specifies the number of cores to use, *input_name.sam* specifies the name of the input sam file, 51 specifies the average read length of reads in the input sam file, and *output_name_csem* specifies the name of the output bed file. Once probability scores were assigned to each read, multi-reads were sampled, meaning all multireads associated with a specific read were removed and only the multiread with the highest probability score was retained. This new set of uni-reads and the highest scoring multireads was produced using the CSEM data processing command *csem-generate-input --sampling --bed input_name.bam output_name_csem_sampling.* In the *csem-generate-input* command, the *--sampling* flag tells the program to only retain uni-reads and multi-reads with the highest probability score, the *--bed* flag specifies the output format of the program, *input_name.bam* specifies the name of the input CSEM processed read bam file, and *output_name_csem_sampling* specifies the name of the sampling processed output file.

ChIP-Seq Differential Peak Calling Analysis with MACS2

Peak calling was then performed using the Model-based Analysis of ChIP-Seq 2 (MACS2) software suite from the University of California San Diego. MACS2 was used to identify differentially expressed peaks between two control samples and one treatment sample in the HCT116 dataset and three control and one treatment sample in the IMR90 dataset. Peaks were called for the samples in the HCT116 dataset with the command *macs2 callpeak -t SRR4090091_1_k100_csem_sampling.bed -c SRR4090089_1_k100_csem_sampling.bed*

SRR4090090_1_k100_csem_sampling.bed -f BED --outdir /shared/rc/fxcsbi/Tom_Andrew/Andrew/MACS/HCT116/csem_sampling/ -n HCT116_csem_sampling. Peaks were called for the samples in the IMR90 dataset with the command *macs2 callpeak -t filtered_trimmed_SRR1448787_1_k100_csem_sampling.bed -c SRR1448786_1_k100_csem_sampling.bed SRR1448788_1_k100_csem_sampling.bed SRR1448789_1_k100_csem_sampling.bed -f BED --outdir /shared/rc/fxcsbi/Tom_Andrew/Andrew/MACS/IMR90/csem_sampling/ -n IMR90_csem_sampling.* In the MACS2 *callpeak* command, every input file following the -t flag is loaded as a set of treatment reads, while every input file following the -c flag is loaded as a set of control reads. In the MACS2 *callpeak* command, -f specifies the file type of the input read data, and -n specifies the root name of all output files generated by peak calling.

Multiple Sequencing Data Colocalization

A script was drafted in R designed to compare the results of the ChIP-Seq data and the RNA-Seq data. The script compares the results of both experiments and looks for overlapping genomic coordinates between differentially expressed repetitive element regions and significantly identified peak regions. This comparison was made by first using the unique repetitive elements GTF file as a map to map identified DE repetitive element names to specific genomic locations. The genomic locations of the significant peaks were increased by 2k, 5k, 10k, and 20k bases in both directions and written as four subset peak files. Repetitive element locations were then investigated for overlaps with the locations in the significantly identified peak files using the intersect() and findOverlaps() commands from the R GenomicRanges package. This produced datasets of differentially expressed repetitive elements colocalized with significant peaks.

Monte Carlo Simulation

A Monte Carlo simulation was constructed to determine the natural rate of occurrence of overlapping genomic locations between the DE peaks and the DE reads. Synthetic datasets were generated representing both the DE RNA-Seq reads and the DE ChIP-Seq peaks. First, the genomic locations of both the read and peaks were isolated and turned into BED format files. Then all BED files were sorted for further analysis with the Bedtools suite using the command *bedtools sort -i $filename > $filename_sorted.bed*. Next, synthetic DE read and peak datasets were generated using the bedtools shuffle command *bedtools shuffle -i in_dataset_sorted.bed -g hg19.chrom.sizes -chrom.* The Bedtools shuffle command used the hg19 chromosome size file to randomly generate a set of individual read locations that retained the same read size and chromosome distribution as the original dataset. The synthetic read set was then compared to the synthetic peak set using the command *bedtools intersect -a $peakset.bed -b $readset.bed.* This command counts the number of intersections between the genomic coordinates of the two BED files given as input. If the number of intersections between the sets is greater than the number of observed interactions in the experiment a counter is incremented by 1. These individual commands were connected to make the simulation, which was performed 1000 times shown in Figure 2. After 1000 iterations of the simulation the total of the counter was divided by 1000 to derive the p-value describing how likely it would be to see what is observed compared to the null model, the result occurring by chance. The empirical P-value formula the simulation used was $\boldsymbol{P=(r + 1)/(n + 1)}$ where $\boldsymbol{r}$ is the counter and $\boldsymbol{n}$ is 1000 (Davison and Hinkley, 1997). This simulation was repeated for all the DE read FC/FDR cutoff data subsets and run against all four different peak ranges (2k, 5k, 10k, and 20k). This allows for the generation of a Monte Carlo observance likelihood table of the same format as the RE-Peak colocalization summary tables.

Figure 2: Monte Carlo Code Snippet

```
for i in {1..1000};
do
  bedtools shuffle -i peak_file -g hg19.chrom.sizes -chrom > temp_peaks.bed
  Null=`bedtools shuffle -i read_file.bed -g hg19.chrom.sizes -chrom | bedtools
intersect -a temp_peaks.bed -b - | wc -l`
  if [ $Null -ge $Observed ]; then
    r_statistic=$((r_statistic + 1))
  fi
done
text="P-value = (r_statistic+1)/(n_statistic+1) =
(${r_statistic}+1)/(${n_statistic}+1) = "
awk -v n_statistic=$n_statistic -v r_statistic=$r_statistic -v t="$text" 'BEGIN
{ans=(r_statistic+1)/(n_statistic+1); print t " " ans}'
```

Figure 2: Bash code snippet from the Monte Carlo simulation testing if the frequency of observed Repetitive Element-Peak colocalization events is greater than that of colocalization events observed due to random chance.

D. Results:

Sequence Quality Assessment

Figure 3: IMR90 RNA-Seq Trimmed Reads

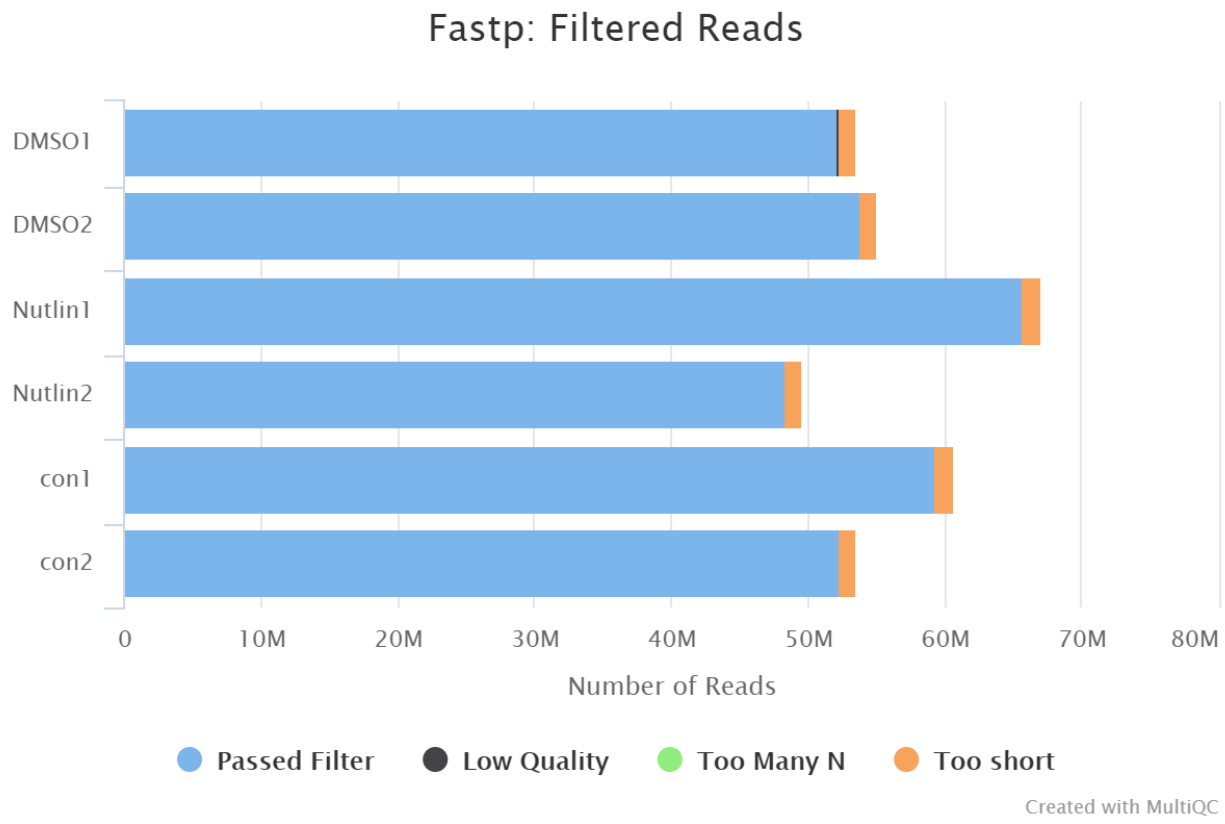

Figure 3: Bar chart showing which reads were removed from each of the six IMR90 RNA-Seq samples after quality control was performed with the FastP cleaning tool.

Data quality control was performed on all experimental samples. The IMR90 RNA-Seq samples were visualized and cleaned using the FastP read preprocessing utility shown in Figure 3. FastP removed fewer than 2 million reads from each IMR90 RNA-Seq sample, most reads were removed for being too low quality or for having a shorter than desired read length. The HCT116 RNA-Seq reads were manually cleaned through an iterative process of visual inspection of FASTQC read summary reports and cleaning with multiple tools from the FASTX toolkit in

addition to the Cutadapt program used to trim overrepresented adapter sequences. Figure 4 visualizes the two to four million reads trimmed from the HCT116 RNA-Seq samples. All seven raw ChIP-Seq samples were surprisingly clean and did not require much cleaning. Figure 5 shows fewer than one million reads were removed from 6 of 7 ChIP-Seq samples. The IMR90 ChIP-Seq samples contained a significantly larger number of reads than the HCT116 ChIP-Seq samples, with three of the four samples containing more than 120 million reads.

Figure 4: HCT116 RNA-Seq Trimmed Reads

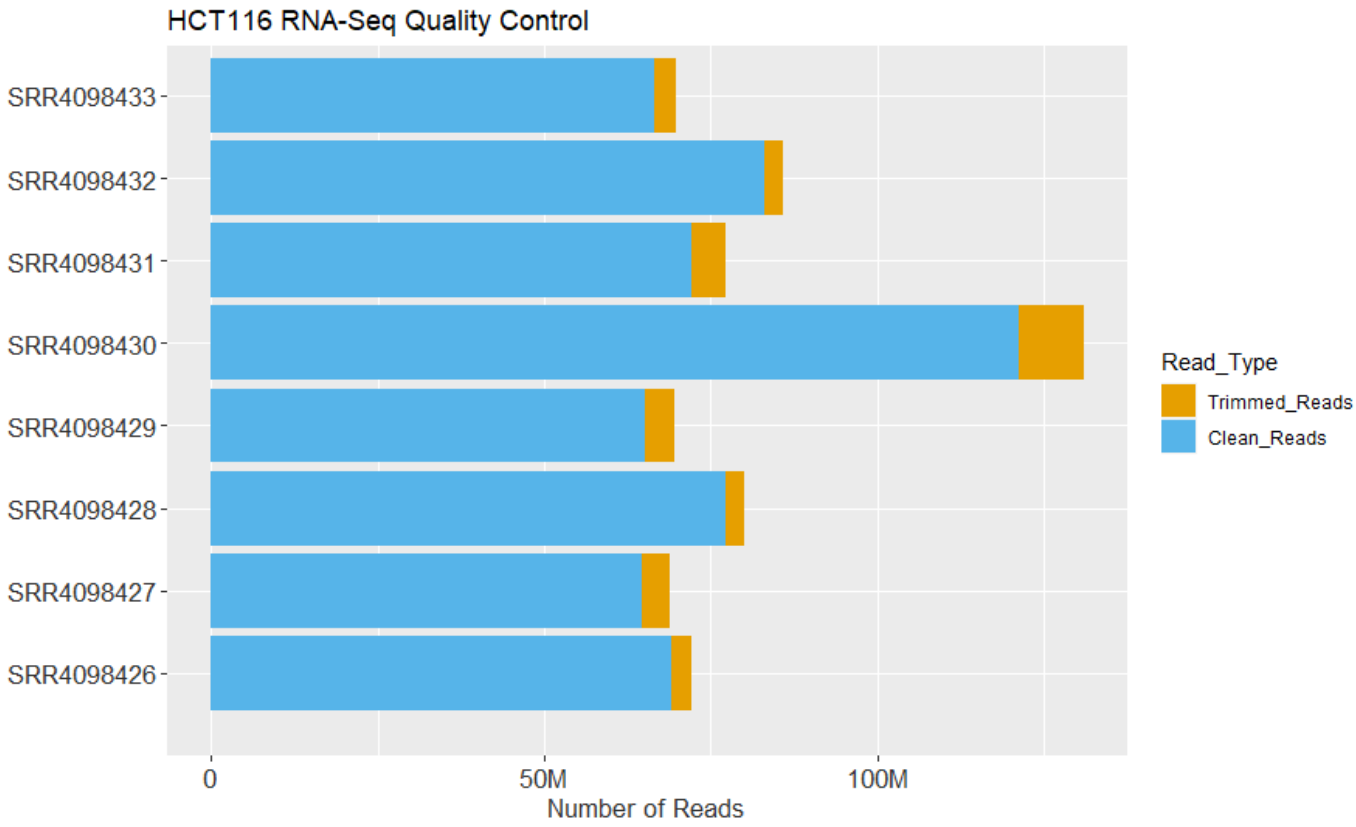

Figure 4: Bar chart showing how many reads were removed from each of the eight HCT116 RNA-Seq samples after quality control was performed with the FASTX.

Figure 5: ChIP-Seq Trimmed Reads

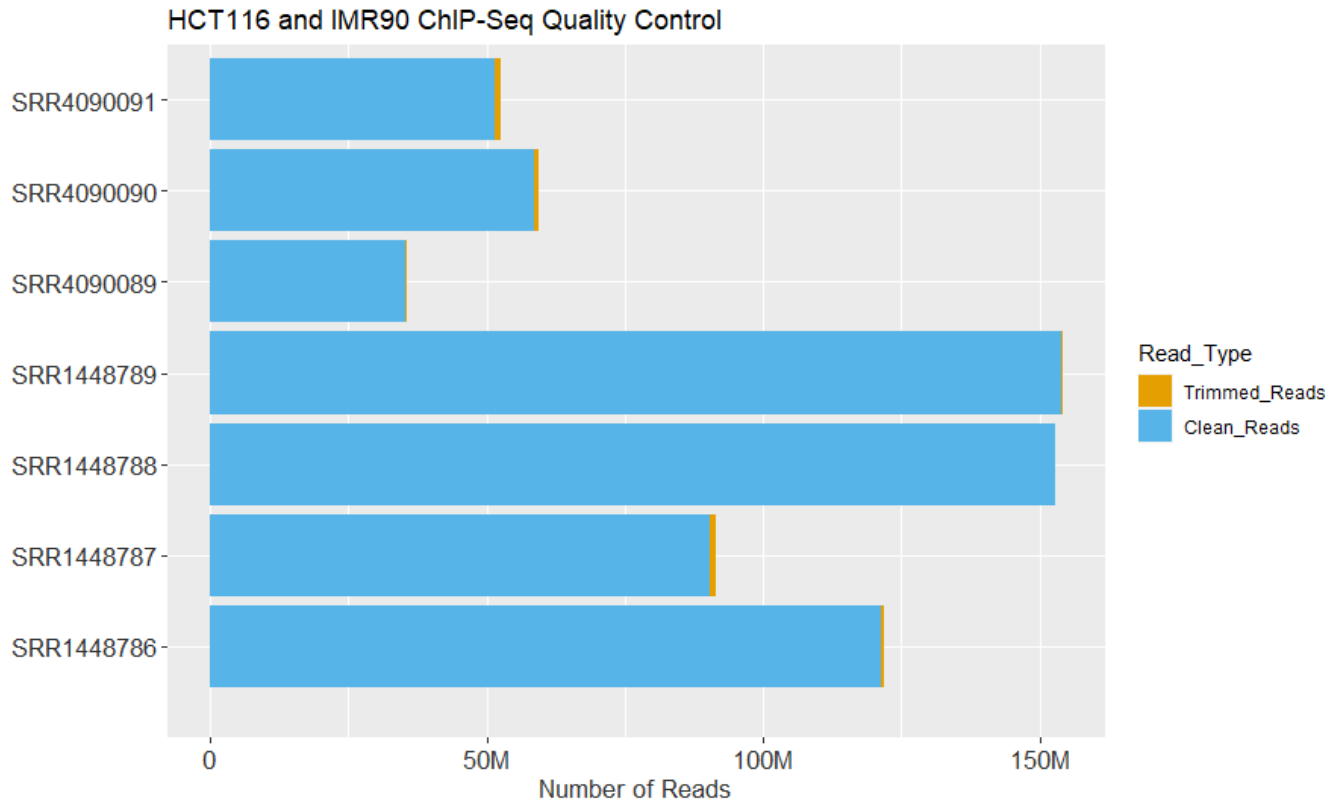

Figure 5: Bar chart showing how many reads were removed from each of the three HCT116 and four IMR90 ChIP-Seq samples after quality control was performed with the FASTX.

Sequence Alignment Assessment

Figure 6: HCT116 RNA-Seq Alignment

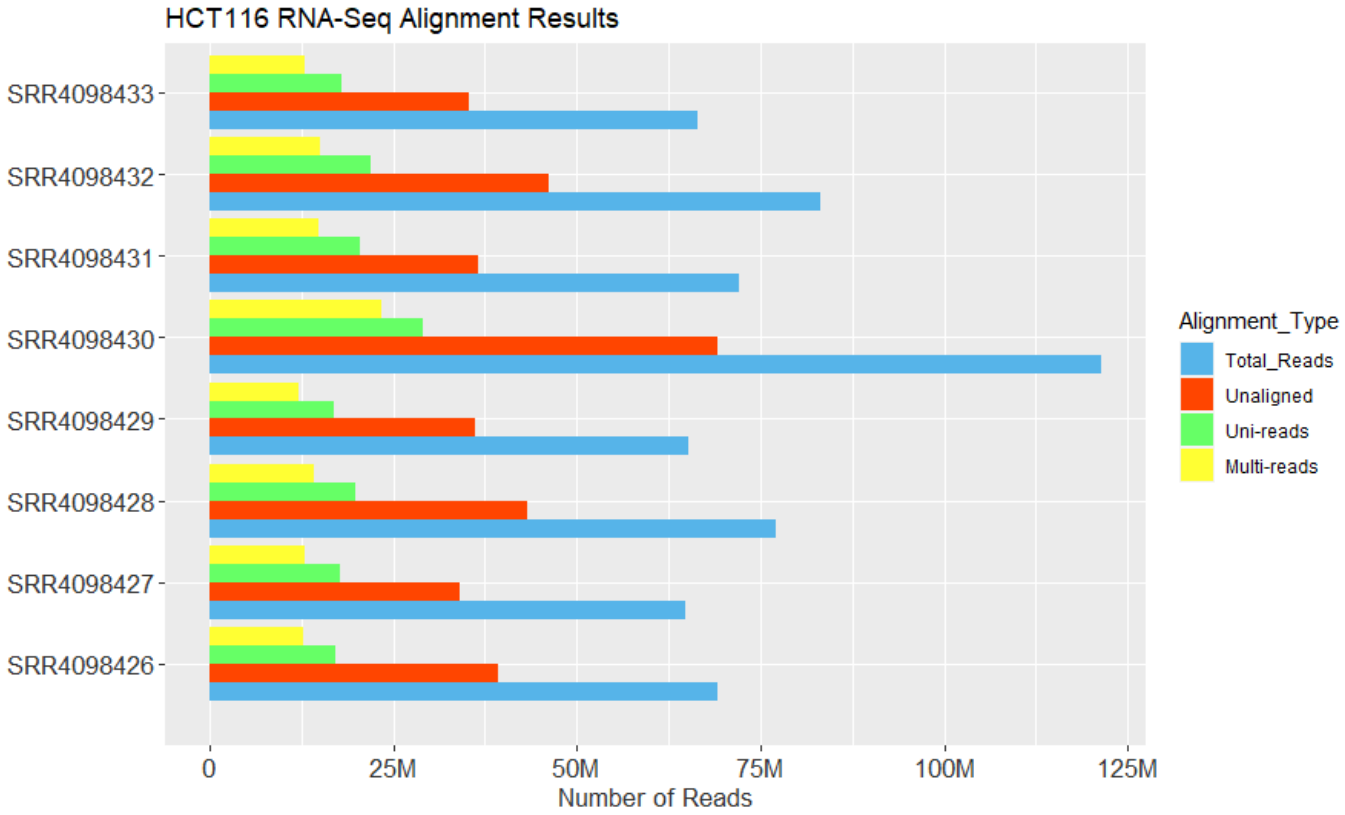

Figure 6: Bar chart of Bowtie2 read alignment rates for HCT116 RNA-Seq samples. The total number of reads before alignment is shown in blue, while the number of reads that remain unaligned after alignment is shown in red. The number of reads mapping to one location is shown in green and the number of reads mapping to two or more possible locations is in yellow.

Figure 7: IMR90 RNA-Seq Alignment

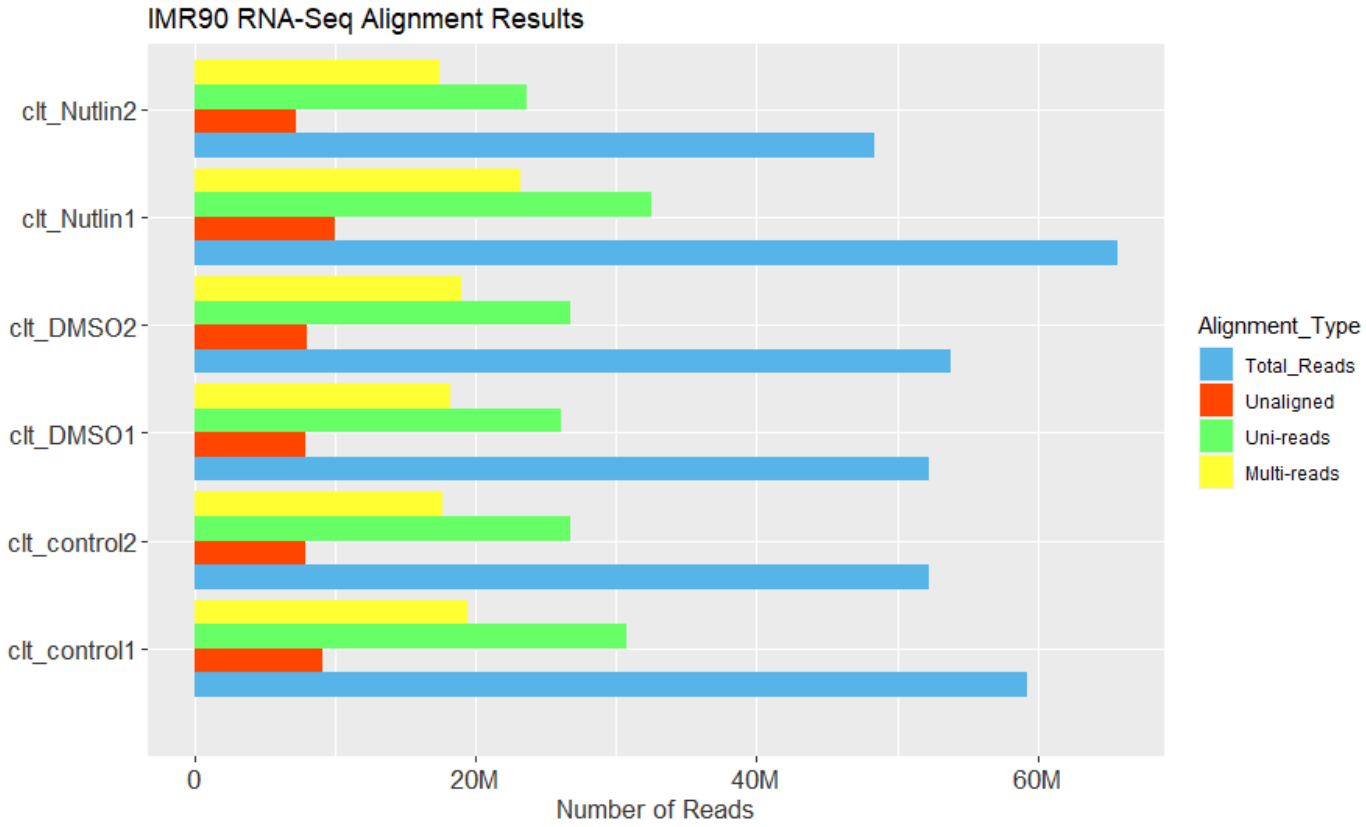

Figure 7: Bar chart of Bowtie2 read alignment rates for IMR90 RNA-Seq data. The total number of reads before alignment is shown in blue, while the number of reads that remain unaligned after alignment is shown in red. The number of reads mapping to one location is shown in green and the number of reads mapping to two or more possible locations is in yellow.

Figure 8: IMR90 ChIP-Seq Alignment

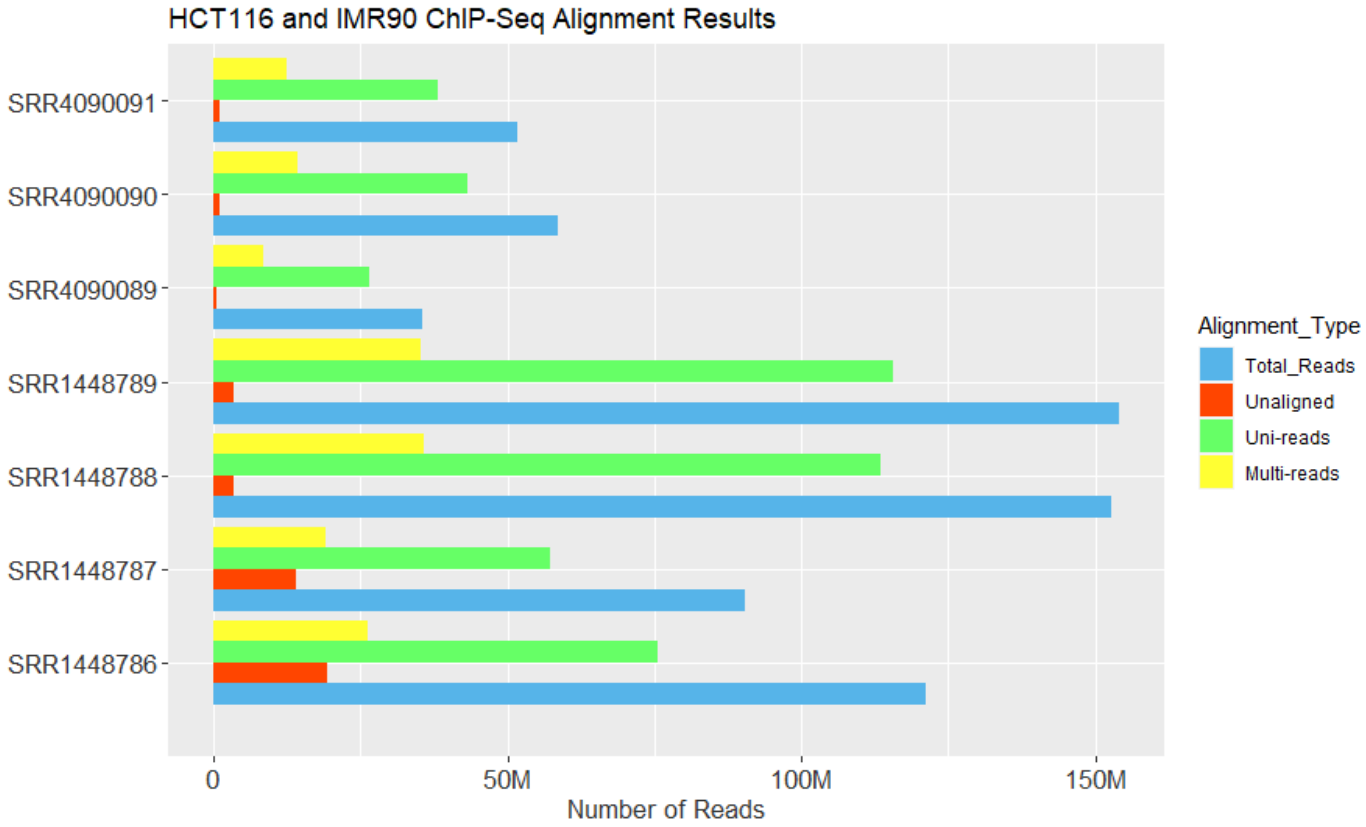

Figure 8: Bar chart of Bowtie2 read alignment rates for HCT116 and IMR90 ChIP-Seq data. The total number of reads before alignment is shown in blue, while the number of reads that remain unaligned after alignment is shown in red. The number of reads mapping to one location is shown in green and the number of reads mapping to two or more possible locations is in yellow.

Alignment to the hg19 reference genome was performed on all samples with alignment rates varying widely between the sample groups. The HCT116 RNA-Seq samples had difficulty aligning, with an average of just under half of all processed reads not aligning across the eight samples. An average of 30% of HCT116 RNA-Seq reads mapped to a single location while on average 20% of reads mapped to more than one location across the genome (Figure 6). The IMR90 RNA-Seq alignment was more successful with 10% of the total reads unaligned. Additionally, the ratio of uni-reads to multi-reads was the same as with the HCT116 data,

however, both groups made up larger fractions of the whole, totaling approximately 50% and 40% of total reads, respectively (Figure 7). The ChIP-Seq alignments were similarly successful with the HCT116 samples leaving around 1% of reads unaligned while the IMR90 samples left less than 10% unaligned. Multi-reads comprised 25% of total aligned reads while uni-reads made up 75% of the total aligned reads (Figure 8).

RNA-Seq: Read Assignment with Telescope

Figure 9: HCT116 RNA-Seq Read Assignment

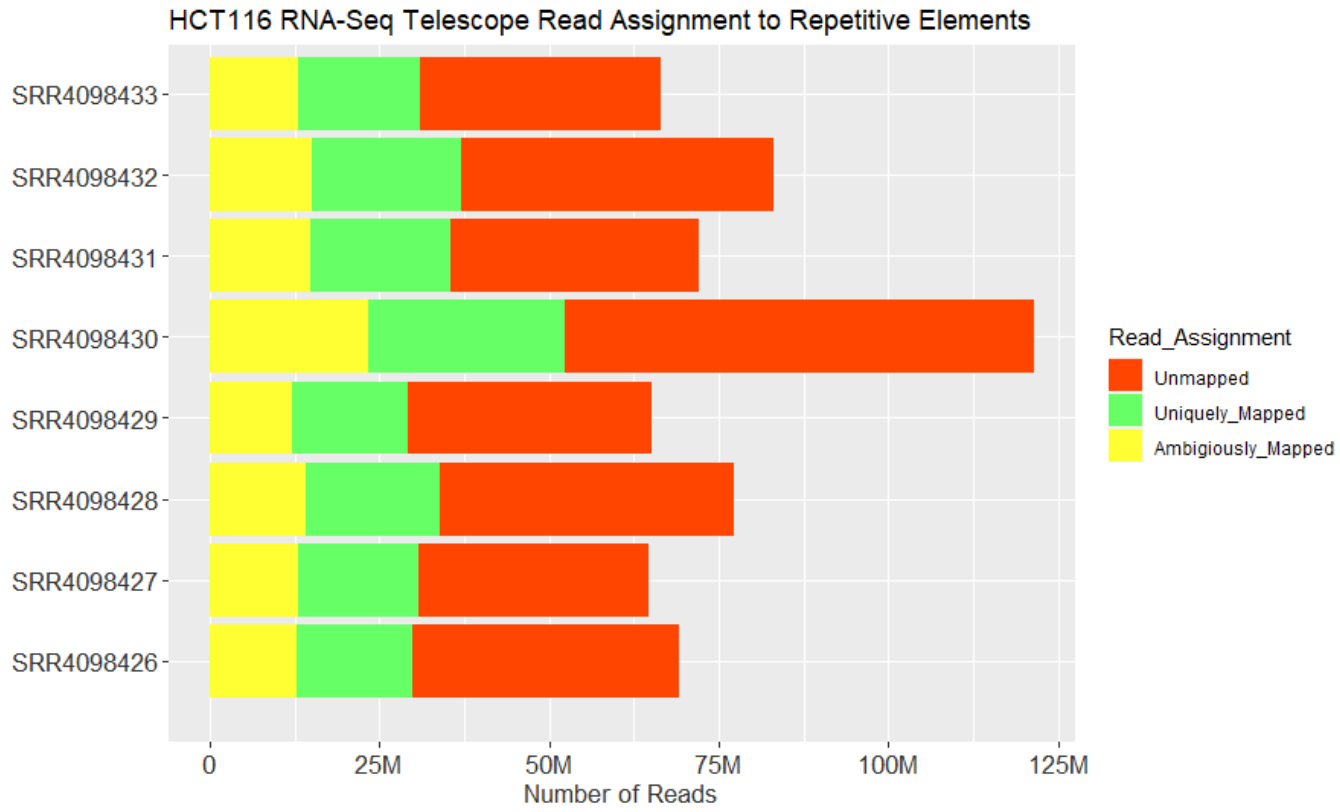

Figure 9: Bar chart depicting the success of read assignment to repetitive element locations with Telescope in the HCT116 RNA-Seq samples. Reads that Telescope is unable to assign are shown in red. Uni-reads that are assigned to a repetitive element are shown in green. Multi-reads that are estimated and successfully assigned to a repetitive element are shown in yellow.

Figure 10: IMR90 RNA-Seq Read Assignment

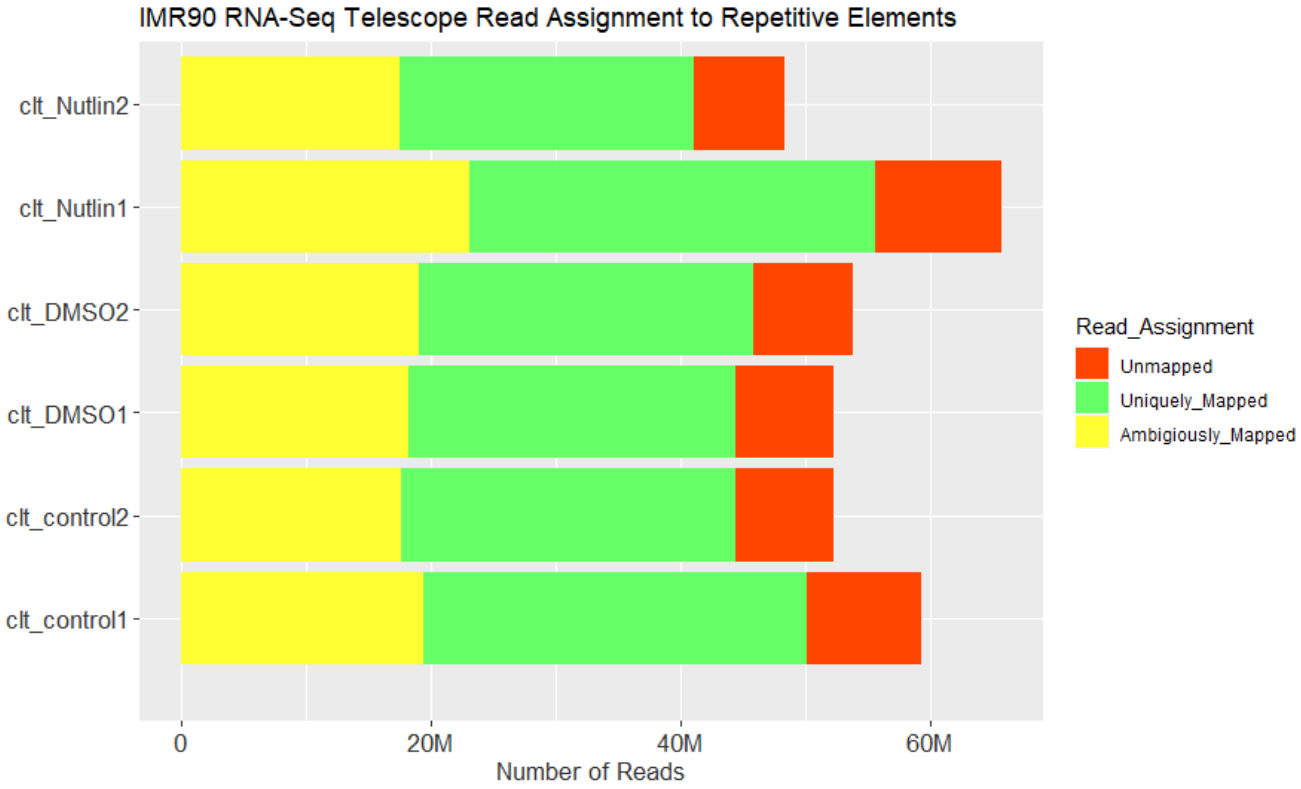

Figure 10: Bar chart depicting the success of read assignment to repetitive element locations with Telescope in the IMR90 RNA-Seq samples. Reads that Telescope is unable to assign are shown in red. Uni-reads that are assigned to a repetitive element are shown in green. Multi-reads that are estimated and successfully assigned to a repetitive element are shown in yellow.

After read alignment with Bowtie2 the read sets contained both multi-reads and uni-reads, each multi-read retaining up to 100 alternative alignment sites. The telescope program attempted to assign all aligned reads to documented repetitive elements locations. The reads of the HCT116 cells were assigned with slightly under a 50% success rate, of those reads assigned more than 40% were multi-reads whose ambiguous mapping had been resolved and ultimately assigned to a repetitive element (Figure 9). The IMR90 RNA-Seq read assignment was much more successful with about 80% of reads successfully being assigned. Like the HCT116 reads, approximately 40% of all assigned reads were multi-reads (Figure 10).

RNA-Seq: Differential Expression Analysis

A counts table of unique repetitive elements was filtered using a CPM threshold of 5 ensuring only elements with enough total counts were included in the Differential Gene expression analysis. IMR90 elements were filtered to 5,368 elements from 2,054,903 and HCT116 elements were filtered to 4768 from 4,507,199.

Table 2: Differential Expression Analysis Summary Table

Differentially Expressed Repetitive Elements Identified: DMSO vs Nutlin			
Cutoff Criteria	HCT116 p53 WT	HCT116 p53 KO	IMR90
PV: 0.05 L2FC: 1	330	78	339
PV: 0.05 L2FC: 1.5	77	13	97
PV: 0.05 L2FC: 2	29	5	33
PV: 0.05 L2FC: 2.5	10	2	22
PV: 0.05 L2FC: 3	6	2	17
PV: 0.05 L2FC: 3.5	5	2	10
PV: 0.05 L2FC: 4	3	1	7
FDR: 0.05 L2FC: 1	208	11	331
FDR: 0.05 L2FC: 1.5	76	7	93
FDR: 0.05 L2FC: 2	28	4	33
FDR: 0.05 L2FC: 2.5	9	1	22
FDR: 0.05 L2FC: 3	5	1	17
FDR: 0.05 L2FC: 3.5	4	1	10
FDR: 0.05 L2FC: 4	2	1	7

Table 2: Table depicting the differentially expressed repetitive elements identified between the DMSO and Nutlin sample groups in the HCT116 p53 WT, HCT116 p53 KO, and IMR90 datasets. The table shows the number of elements identified that meet the p-value, FDR, and log2 foldchange thresholds listed in the left-hand column.

Figure 11: Venn Diagram of Overlapping DE REs IMR90 and HCT116

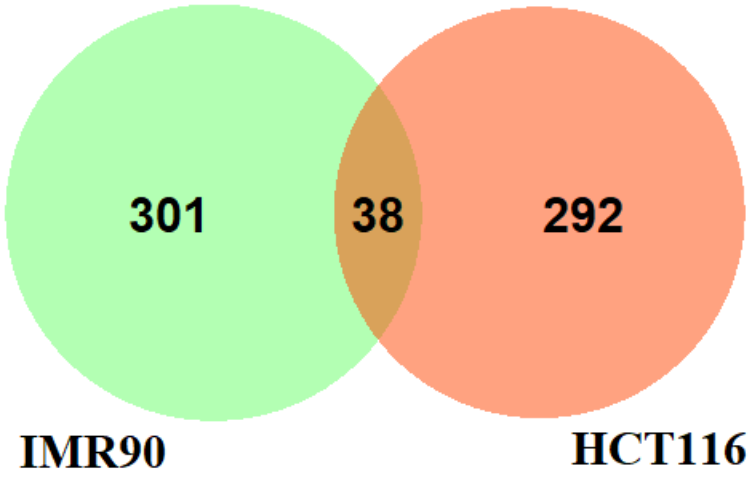

Figure 11: Venn diagram showing the number of differentially expressed repetitive elements that overlap between the IMR90 and HCT116 p53 WT datasets.

Figure 12: HCT116 p53 WT Differential Expression Heatmap

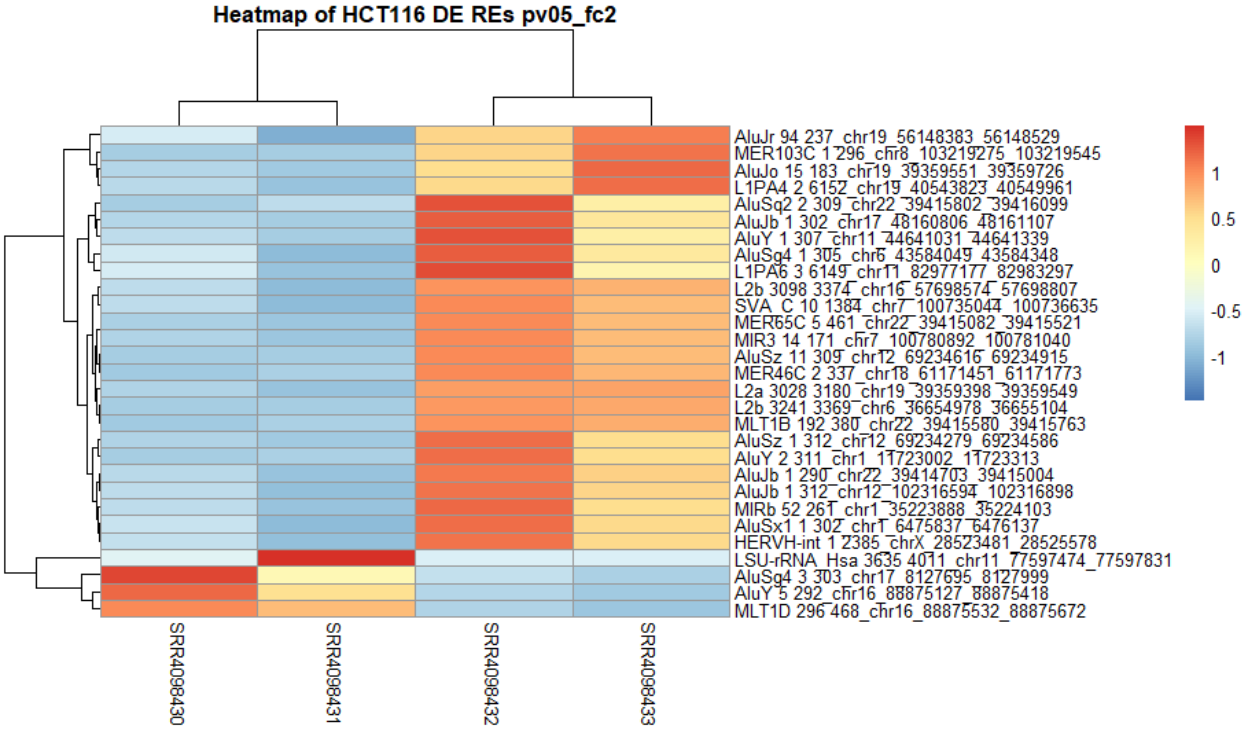

Figure 12: A Heatmap of the differentially expressed repetitive elements in the HCT116 p53 WT dataset, where red indicates increased expression and blue indicates decreased expression.

Figure 13: IMR90 Differential Expression Heatmap

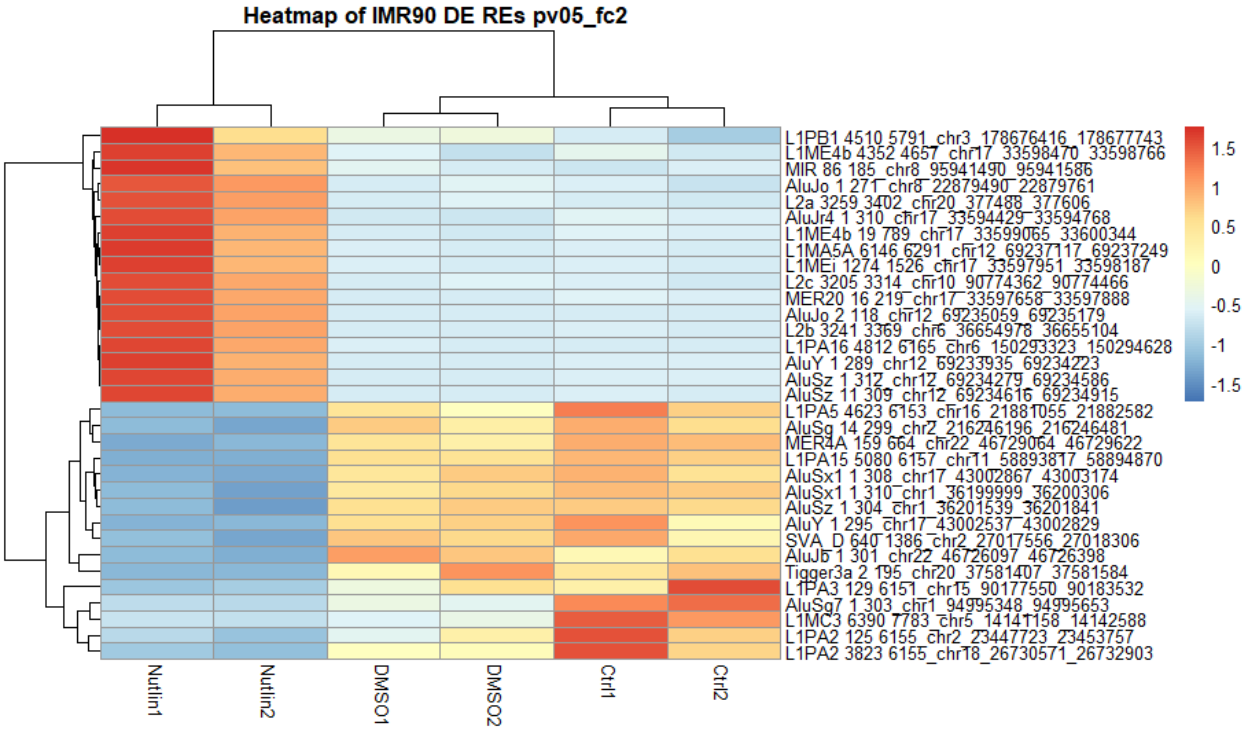

Figure 13: A Heatmap of the differentially expressed repetitive elements in the IMR90 dataset, where red indicates increased expression and blue indicates decreased expression.

Differential Expression analysis was run in both EdgeR and DESeq2 and a consensus set of REs that met a specific fold change and p-value cutoffs were retained. This method was applied to the IMR90, HCT116 p53 WT and p53 KO data, at the lowest cutoff threshold 330, 78, and, 339 DE REs were identified, shown in Table 2. 38 REs between those sets overlapped, shown in Figure 11. Heatmaps clustering relative expression levels based on RE counts across all experimental samples, were also produced as visualization of the differential expression of repetitive elements between the two different treatment groups (Figure 12, Figure 13).

ChIP-Seq: Analysis of Peak Calling Methods

Figure 14: Overlapping HCT116 Peaks

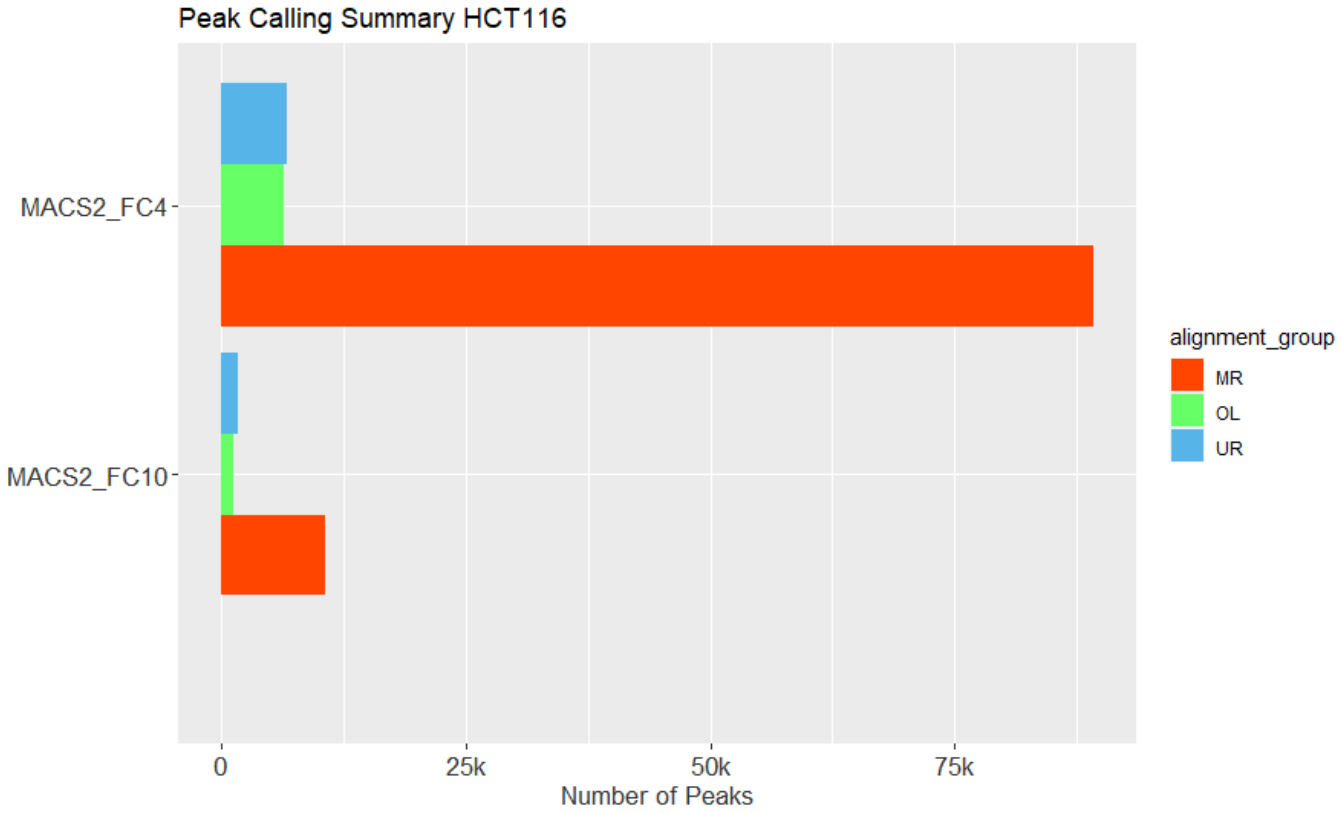

Figure 14: Bar chart showing overlapping MACS2 peaks between multi-read and unique-read peak sets in HCT116 cells. The total number of peaks called from the multi-read dataset are shown in red. The total number of peaks called from the uni-read dataset are shown in blue. The total number of peaks that overlap between the multi-read and uni-read dataset are shown in green.

Figure 15: Overlapping IMR90 Peaks

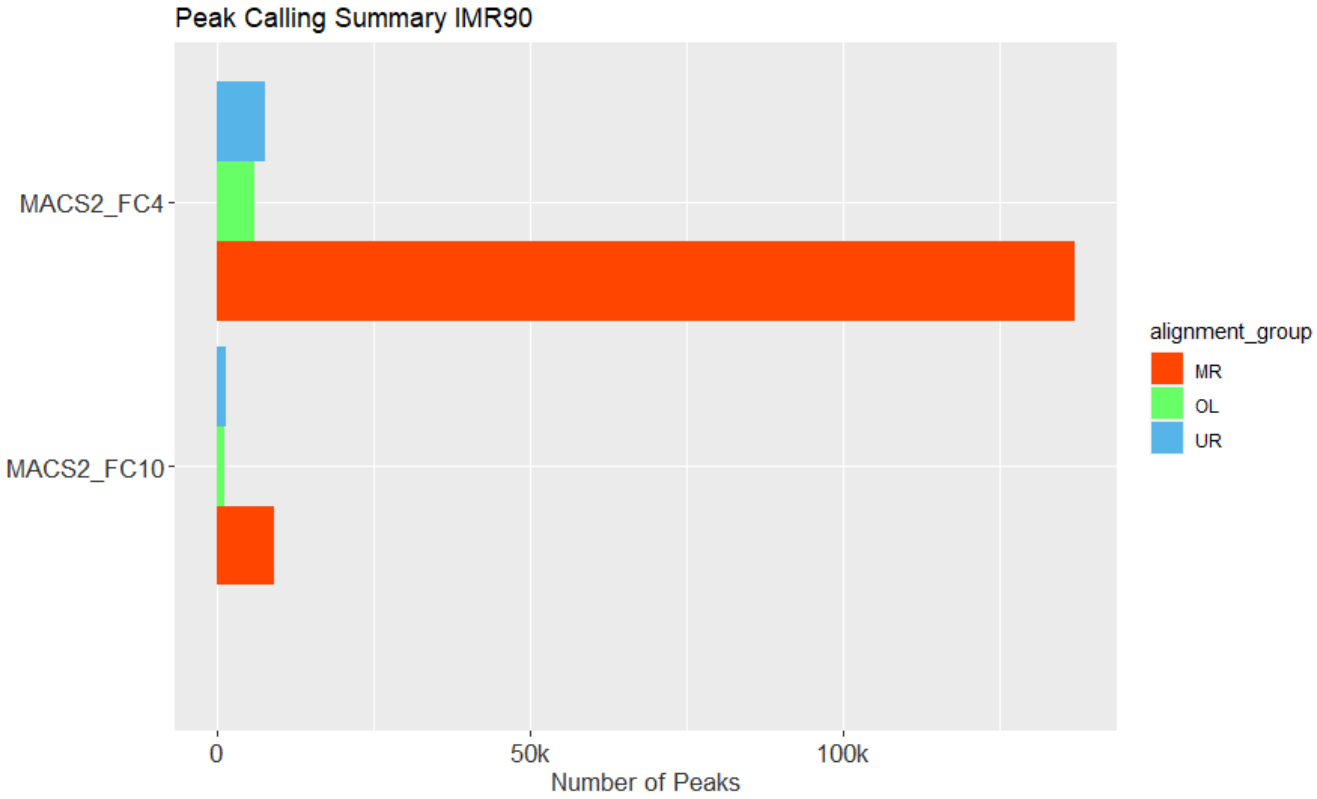

Figure 15: Bar chart showing overlapping MACS2 peaks between multi-read and unique-read peak sets in IMR90 cells. The total number of peaks called from the multi-read dataset are shown in red. The total number of peaks called from the uni-read dataset are shown in blue. The total number of peaks that overlap between the multi-read and uni-read dataset are shown in green.

Before Peak Calling was performed on the aligned ChIP-Seq data, the data was processed by CSEM, a program designed to score the likelihood of alignment at all alternative alignment locations in a ChIP-Seq multi-read. Once all multi-reads were scored a CSEM companion program was used to generate two datasets. The unique read (UR) dataset completely excluded all non-uni-read peak data. The multi-mapped read (MR) dataset included all uni-read data but also included the highest-scoring alignment location of every multi-read, essentially converting all multi-reads to uni-reads. These datasets were run through both the HOMER and MACS2 peak callers and the resulting peak sets overlapped to create an overlapping (OL) peak set as

shown in Figures 14 and 15. This peak set was found to overlap with roughly half of the peaks in the UR set.

ChIP-Seq: Analysis of Peaks with Annotated Consensus Set

Figure 16: Peak Verification with Bao Consensus Set

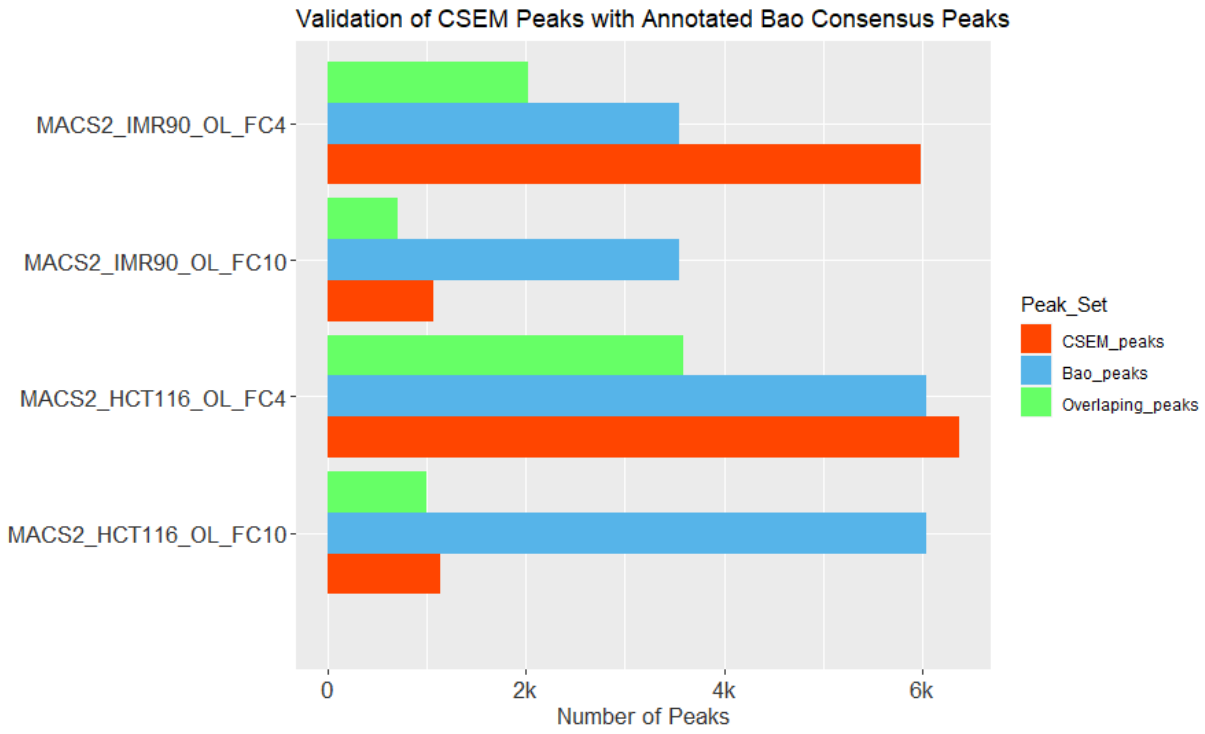

Figure 16: Bar chart showing overlaps between the CSEM OL peaks and the Bao annotated consensus peak sets. The total umber of CSEM OL peaks for each subset are shown in red. The total number of normal p53 or cancer p53 associated peaks in the Bao consensus set are shown in blue. The total number of peaks that overlap between the CSEM peaks the corresponding Bao consensus peaks set are shown in green.

The overlapping (OL) HCT116 and IMR90 peak sets were compared to their corresponding normal cell and cancer cell p53 consensus peak sets generated by Bao et al. 2017. This consensus set serves as a robust reference peak set of externally validated data which can further validate the OL peak set identified by this workflow. Figure 16 shows the overlapping elements between the OL peaks and the Bao consensus peaks. Among the FC > 10 OL peak sets,

greater than 90% of the peaks overlap with the Bao peak set. Among the FC > 4 OL peak sets, greater than 70% of the peaks overlap with the Bao peak set. These overlaps are strong indicators that peaks in the OL sets are likely genuinely p53 associated peaks.

Repetitive Element-Peak Colocalization

The RE-Peak colocalization summary tables are shown below; these tables summarize the total number of colocalizations that occur between a given set of REs and a peak range. The peak range takes the original peak start and stop locations and increases it in both directions by 2000, 5000, 10000, or 20000bp. These ranges are important as p53 could bind upstream or downstream of the RE element's start or stop coordinates but still have a significant regulatory effect on the RE. The program checks to see if the genomic coordinates of the DE REs are within the peak range specified, and if they are it is tallied as a colocalization. One RE can be recorded as a colocalization multiple times if multiple different peaks are found to interact with it. Tables 3 and 4 show RE-Peak colocalization summary tables between the DE RE sets, the HCT116 p53 WT and the IMR90 RE set, and the overlapping (OL) peak sets called by the MACS2 peak caller. Additionally, RE-Peak colocalization sets with an FDR of 0.05 and fold change of 2 have been overlapped with the 20kbp peak range in the HCT116 p53 WT and IMR90 cell lines and selected to be displayed in full. The resulting interaction has been highlighted in yellow on the colocalization summary tables and are shown in full in Tables 5 and 6.

Table 3: HCT116 RE-Peak Colocalization Summary Table

Data Source	Peaks_2k	Peaks_5k	Peaks_10k	Peaks_20k
HCT116_REs_FDR05_fc1	7	23	43	62
HCT116_REs_FDR05_fc1.5	3	11	20	24
HCT116_REs_FDR05_fc2	1	6	8	9
HCT116_REs_FDR05_fc2.5	0	1	2	3
HCT116_REs_FDR05_fc3	0	1	2	2
HCT116_REs_FDR05_fc3.5	0	1	2	2
HCT116_REs_FDR05_fc4	0	1	2	2

Table 3: Summary table counting the number of colocalizations events occurring between the HCT116 p53 WT DE RE read sets at varying thresholds and the OL MACS2 called HCT116 peaks with ranges increased by 2k, 5k, 10k, and 20k. The yellow highlighted cell indicates the colocalization selected to be displayed in Table 5

Table 4: IMR90 RE-Peak Colocalization Summary Table

Data Source	Peaks_2k	Peaks_5k	Peaks_10k	Peaks_20k
IMR90_REs_FDR05_fc1	2	13	48	115
IMR90_REs_FDR05_fc1.5	1	7	21	35
IMR90_REs_FDR05_fc2	1	2	8	10
IMR90_REs_FDR05_fc2.5	0	1	5	7
IMR90_REs_FDR05_fc3	0	1	5	7
IMR90_REs_FDR05_fc3.5	0	1	5	6
IMR90_REs_FDR05_fc4	0	0	0	0

Table 4: Summary table counting the number of colocalizations events occurring between the IMR90 DE RE read sets at varying thresholds and the OL MACS2 called IMR90 peaks with ranges increased by 2k, 5k, 10k, and 20k. The yellow highlighted cell indicates the colocalization selected to be displayed in Table 6

Table 5: HCT116 p53 Wildtype RE-Peak Colocalization Table

RE_Name	logFC	FDR	Chr	Start_RNA	End_RNA	Strand	overlap	start_Peak	end_Peak	PeaktoRead
L2b 3241 3369_cl	4.317262	1.89E-32	chr6	36654978	36655104	+	127	36644883	36645500	9478
L2b 3241 3369_cl	4.317262	1.89E-32	chr6	36654978	36655104	+	127	36650192	36650615	4363
MIR3 14 171_chr	2.658478	9.93E-17	chr7	100780892	100781040	-	149	1.01E+08	1.01E+08	10566
AluSx1 1 302_chr	2.383674	1.03E-09	chr1	6475837	6476137	-	301	6474407	6475008	829
MER65C 5 461_cl	2.324965	2.50E-13	chr22	39415082	39415521	+	440	39410605	39410710	4372
MIRb 52 261_chr	2.319364	2.56E-10	chr1	35223888	35224103	+	216	35220935	35221160	2728
AluJb 1 290_chr2	2.094689	4.67E-11	chr22	39414703	39415004	+	302	39410605	39410710	3993
MLT1B 192 380_(	2.024527	1.58E-06	chr22	39415580	39415763	-	184	39410605	39410710	4870
AluSq2 2 309_chr	2.019919	5.70E-07	chr22	39415802	39416099	-	298	39410605	39410710	5092

Table 5: RE-Peak colocalization of HCT116 DE REs at the FDR 0.05 FC 2 cutoff with the 20kb range MACS2 HCT116 peak set. This table displays the name of the repetitive element, its log fold change, FDR, strand, chromosome number, start position and stop position. In addition, the table shows the non-peak range augmented peak start and stop positions, the distance from the peak to the RE, and the total length the overlapping coordinates.

Table 6: IMR90 RE-Peak Colocalization Table

RE_Name	logFC	FDR	Chr	Start_RNA	End_RNA	Strand	overlap	start_Peak	end_Peak	PeaktoRead
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36634699	36635311	19667
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36643727	36645569	9409
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36645624	36646808	8170
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36647057	36647464	7514
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36648117	36648431	6547
L2b 3241 3369_c	3.799748	8.64E-92	chr6	36654978	36655104	+	127	36650158	36650716	4262
MIR 86 185_chr8	3.033492	8.54E-28	chr8	95941490	95941586	-	82	95961505	95962572	-19919
AluJo 1 271_chr8	2.042188	1.61E-34	chr8	22879490	22879761	+	272	22871827	22872277	7213
MER4A 159 664_	-2.00147	6.24E-12	chr22	46729064	46729622	+	559	46731481	46731651	-1859
AluJb 1 301_chr2	-2.49869	3.87E-14	chr22	46726097	46726398	+	302	46731481	46731651	-5083

Table 6: RE-Peak colocalization of IMR90 DE REs at the FDR 0.05 FC 2 cutoff with the 20kb range MACS2 IMR90 peak set. This table displays the name of the repetitive element, its log fold change, FDR, strand, chromosome number, start position and stop position. In addition, the table shows the non-peak range augmented peak start and stop positions, the distance from the peak to the RE, and the total length the overlapping coordinates.

Table 7: Differentially Expressed HCT116 p53 Knockout Repetitive Elements

RE_Name	logFC	logCPM	FDR	baseMean	log2FoldChange	lfcSE	stat	padj
FLAM_C 1 133_ch	6.457436	10.17158	5.20E-48	4603.75	6.455933098	0.283137	22.80144	2.00E-111
AluJr 5 294_chr7_	2.152874	4.172105	1.17E-05	70.44043	2.156896298	0.42777	5.042187	0.000345
AluY 1 303_chr2_1	-1.07129	7.585443	0.00839	765.6709	-1.074200245	0.24258	-4.42824	0.00475
AluSg7 1 303_chr1	-1.07453	6.490484	0.010422	357.2708	-1.076940963	0.279581	-3.85198	0.047931
AluSp 2 293_chr1_	-1.30911	4.920619	0.001749	118.9102	-1.314631116	0.333597	-3.94078	0.036548
AluSx3 1 294_chr1	-1.43266	5.365303	5.78E-05	162.6189	-1.437336296	0.308734	-4.65558	0.002077
L1PA5 4405 6147_	-1.59031	6.072488	6.57E-07	266.9186	-1.593297943	0.279275	-5.70511	1.47E-05
AluSz 1 300_chr17	-1.61196	4.541451	8.83E-05	91.01226	-1.616750616	0.361609	-4.47099	0.004379
AluSz6 1 298_chr1	-1.72179	5.951469	6.57E-07	245.2103	-1.725247834	0.303471	-5.68505	1.47E-05
BC200 1 200_chr2	-2.18735	11.11146	4.32E-07	8843.375	-2.18972001	0.219541	-9.9741	4.45E-20
AluSg4 1 288_chr6	-2.29942	4.646159	4.47E-07	97.98595	-2.306530532	0.411629	-5.60343	1.89E-05

Table 7: Differentially expressed repetitive elements identified in the HCT116 p53 knockout differential expression analysis. This table show fold change increase and significance statistics generated by both the EdgeR (left) and DESeq2 (right) differential expression analysis software.

Tables 5 and 6 show summary statistics about the RE-Peak interactions identified in each dataset selected at the set FDR of 0.05 and absolute value fold change criteria of two. Each of the tables shows the differentially expressed repetitive element's name, log fold change value, FDR significance values, strand, and genomic coordinates. In addition, each table shows the coordinates of the peak the RE interacted with, the total length of the RE-Peak overlap, and the distance from the non-range adjusted peak location to the RE location with a negative value indicating that a RE is upstream of the peak. The results of the HCT116 p53 Knockout differential expression analysis are shown in Table 7 and show RE's that may be differentially expressed due to the treatment conditions and are unlikely related to p53. Table 7 serves as a form of control RE set that can be used to eliminate non p53 associated REs discovered in the HCT116 p53 Wildtype and IMR90 datasets.

Validation of RE-Peak Colocalizations with Monte Carlo Simulation

Table 8: Monte Carlo Simulation HCT116

Data Source	Peaks_2k	Peaks_5k	Peaks_10k	Peaks_20k
HCT116_REs_FDR05_fc1	0.000999001	0.000999001	0.000999001	0.000999001
HCT116_REs_FDR05_fc1.5	0.0679321	0.000999001	0.000999001	0.000999001
HCT116_REs_FDR05_fc2	0.25974	0.000999001	0.000999001	0.002997
HCT116_REs_FDR05_fc2.5	1	0.203796	0.0729271	0.0689311
HCT116_REs_FDR05_fc3	1	0.125874	0.028971	0.0749251
HCT116_REs_FDR05_fc3.5	1	0.0879121	0.018981	0.042957
HCT116_REs_FDR05_fc4	1	0.0509491	0.004995	0.024975

Table 8: Table showing the Monte Carlo Simulation generated p-values representing the significance of the total colocalization events occurring between the HCT116 p53 WT DE RE read sets at varying thresholds and the OL MACS2 called HCT116 peaks with ranges increased by 2k, 5k, 10k, and 20k.

Table 9: Monte Carlo Simulation IMR90

Data Source	Peaks_2k	Peaks_5k	Peaks_10k	Peaks_20k
IMR90_REs_FDR05_fc1	0.862138	0.0539461	0.000999001	0.000999001
IMR90_REs_FDR05_fc1.5	0.664336	0.00999001	0.000999001	0.000999001
IMR90_REs_FDR05_fc2	0.325674	0.215784	0.000999001	0.002997
IMR90_REs_FDR05_fc2.5	1	0.410589	0.00599401	0.00799201
IMR90_REs_FDR05_fc3	1	0.358641	0.001998	0.002997
IMR90_REs_FDR05_fc3.5	1	0.218781	0.000999001	0.001998
IMR90_REs_FDR05_fc4	1	1	1	1

Table 9: Table showing the Monte Carlo Simulation generated p-values representing the significance of the total colocalization events occurring between the IMR90 DE RE read sets at varying thresholds and the OL MACS2 called IMR90 peaks with ranges increased by 2k, 5k, 10k, and 20k.

A Monte Carlo simulation has been designed and run to investigate the significance of the results of the RE-Peak colocalization summary tables. The simulation randomly reshuffles the RE and Peak dataset maintaining the same size and chromosomal distributions and tallies colocalization events that occur, this is the estimation of nature, the null model. The simulation then runs 1000 times and tallies the total number of times the total null model colocalization events is greater than what was observer in the RE-Peak colocalization summary table. This tally become the **r** value in the empirical p-value formula. The **r** value plus one is divided by the total number of simulations run plus one to obtain the empirical p-value; ***P=(r + 1)/(n + 1)***. Tables 8 and 9 show the empirical p-values calculated for each colocalization summary. In both the IMR90 and HCT116 datasets the colocalization made between the 20k peaks and the fc1 REs are found to have a significant p-value. The relative sizes of the these datasets, less than three hundred REs and less than ten thousand peaks in comparison to the total size of the genome illustrate that the dataset with the highest p-value has not been cherry pick for validation but actually represents significant colocalization events. While the peak ranges have been increased by 40k base pairs in length, 40k is still an incredibly small percent of the size of an individual chromosome, 40k is 0.016% of 248,956,422, the length of chromosome one. The Monte Carlo simulations validate the colocalizations as occurring more frequently than due to random chance.

Transposable Element Representation Analysis

Figure 17: HCT116 p53 Wild Type TE Representation Analysis

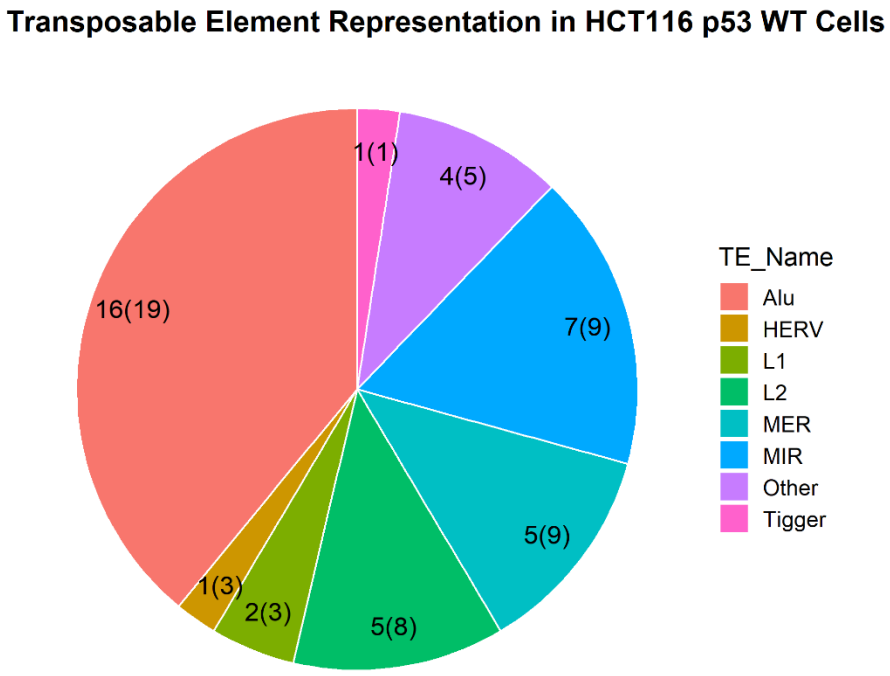

Figure 17: Pie chart showing the representation of transposable elements in the colocalization results of the HCT115 p53 WT data. The pie chart is color coded by which type of transposable element is present and in what quantity. The count labels show the number of colocalization events of that occur in a TE group irrespective of a TE colocalizing with multiple peaks, while showing in parenthesis the number of colocalization events of that occur in a TE group including colocalizations of a TE with multiple peaks.

Figure 18: IMR90 TE Representation Analysis

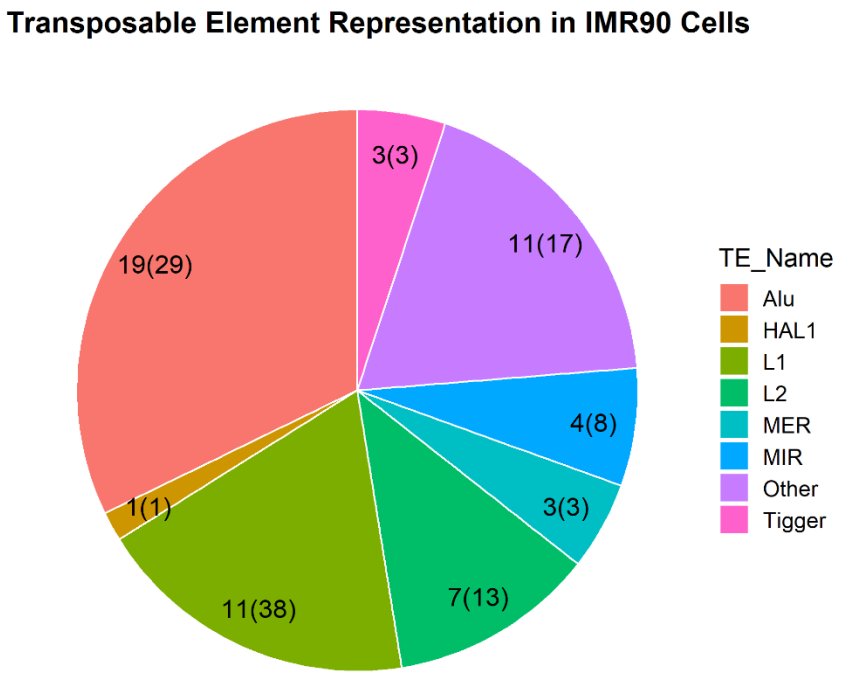

Figure 18: Pie chart showing the representation of transposable elements in the colocalization results of the IMR90 data. The pie chart is color coded by which type of transposable element is present and in what quantity. The count labels show the number of colocalization events of that occur in a TE group irrespective of a TE colocalizing with multiple peaks, while showing in parenthesis the number of colocalization events of that occur in a TE group including colocalizations of a TE with multiple peaks.

The pie charts created show a breakdown of the different types of TEs found in RE-Peak colocalization datasets at an FDR of 0.05, with a fold change of 1, and a peak range of 20kb. A pie chart breaking down TE representation was generated for the RE-Peak colocalization tables of both the IMR90 and HCT116 p53 WT datasets generated with MACS2 peaks. The HCT116 p53 WT chart showed that the most represented TEs were the Alus, with 16 unique elements, followed by the MIRs with 7 unique elements identified in the colocalization with the MACS2 peak dataset (Figure 17). The IMR90 pie chart showed that the most represented TEs were the

Alus, with 19 unique elements, followed by the L1s with 11 unique elements identified in the colocalization with the MACS2 peak dataset (Figure 18).

In the HCT116 data the TEs with the top five highest positive expression (FC > 2.0) values identified in the colocalization datasets were L2b 3241 3369, MIR3 14 171, AluSx1 1 302, MER65C 5 461, MIR 52 261. While the TE's with the lowest negative expression (FC < -1.5) value identified in the colocalization dataset was the AluSx1 1 312 element. In the IMR90 data, the TEs with the highest positive expression (FC > 2.0) values identified in the colocalization datasets were L2b 3241 3369, MIR 86 185, and AluJo 1 271. While the TE's with the lowest negative expression (FC < -1.5) values identified in the colocalization datasets were the AluJb 1 301, MER4A 159 664, AluJb 1 302, AluSz 1 302, and AluJr 1 301 element.

RE L2b 3241 3369 is not only present in both the HCT116 p53 WT and IMR90 datasets, but it is also the most significantly expressed element (FC > 3) in both datasets, this would suggest a significant relationship to p53, however this specific element is located right next to the p21 gene which is a major downstream p53 target. The colocalization that has occurred with this specific element is likely a false positive, illustrating that all colocalized transposable elements need to be examined carefully before deciding to proceed with further experimental validation. The full list of colocalized elements discovered in both the HCT116 p53 WT and IMR90 datasets can be viewed in supplementary Table S1 and S2.

E. Discussion:

The objective of this study was to investigate a hypothesized regulatory relationship p53 has on various transposable elements. To achieve that goal this study has overcome alignment and mapping challenges associated with transposable element work by including and mapping multi-reads to a custom transcriptome of annotated repetitive elements. In a typical analysis multi-reads would have been thrown out, while this analysis successfully included multi-reads using the Telescope and CSEM programs. Telescope specifically was able to successfully map between 20 and 30% of total reads, which would have otherwise been excluded.

This study has identified 330 and 339 differentially expressed repetitive elements (FDR less than 0.05), as a response to the p53 activating Nutlin chemical treatment, in p53 wildtype cancer and normal cells, respectively. Additionally, the differential expression analysis performed on the p53 knockout cancer cells yielded comparatively very few significant reads (FDR less than 0.05) identifying only 11 DE REs. This DE results was expected as the chemical treatment groups compared in the DEA rely on the experimental activation of p53, without p53 the treatments would be expected to have little effect on gene expression. The DE REs identified in the HCT116 p53 knockout cells also had use as a filter to remove non-p53 associated DE REs from the results of the other DEAs.

This study has identified a robust set of peaks associated with p53 activation in cancer and normal cells. This peak set has been validated via a plurality of overlap during comparison with the Bao p53 consensus normal cell and cancer cell peaks sets. RE-Peak colocalization sets have been generated characterizing genomic proximity suggesting an association between differentially expressed REs and p53 associated peaks. Monte Carlo simulations have been performed and verify that the rate of these RE-Peak colocalizations are unlikely a byproduct of

random chance. Transposable elements with high levels of differential expression that have colocalized with multiple peaks are the strongest candidates for regulatory relationships with p53 or a p53 controlled protein. Multiple transposable elements have been identified in the RE-Peak colocalization tables that look to have strong interactions with p53 associated peaks.

However, this study was an exploratory analysis combining multiple datasets and was limited in several ways. The colocalization analysis of the IMR90 ChIP-Seq and RNA-Seq may have been impacted by differences in IMR90 expression. While both analyses were performed using samples of the IMR90 cell line the analyses were run five years apart and by different labs. The IMR90 cells used in the ChIP-Seq may have accumulated a small number of mutations in the five years before the IMR90 cells were used in the RNA-Seq analysis. Providing the cell line has been maintained correctly by the manufacturer this problem should pose little risk to impacting the outcome of the experiment.

Similarly, use of the IMR90 cell line itself presents a limitation of the analysis. In the analysis IMR90 cells are used to represent normal tissue, however the process of immortalizing a cell line comes at a cost. The IMR90 cells have been mutated in some way to make them immortal and no longer truly represent normal cells. The focus of this study is related to p53 expression and p53 with its outsized role in cell cycle progression may have been affected by the immortalization process. These results can only be used as an approximation of the activity of p53 in normal tissue.

All differential expression analyses were performed with only one experimental replicate sample. While one experimental replicate is the minimum number of sample replicates needed to perform a differential expression analysis, more replicates are always suggested. Including multiple replicates increases the statistical power of the model and overall confidence in the

results of the differential expression analysis. This analysis would have benefited from including more than one replicate of each sample.

Future work should be done to verify the colocalization analysis performed in the study. Analyses with additional experimental replicates performed on the same tissue types would provide greater evidence supporting the existence of p53 regulated repetitive elements. Alternative tissue types should be investigated with both cancer and normal cells samples to determine if these colocalizations are tissue specific or are representative of a shared mechanism. A replication of the analysis in samples of truly normal lung fibroblasts and a comparison with the IMR90 results would also shed insight on both the colocalizations observed and the drawbacks associated with using an immortalized cell line to represent normal cells.

The results of the colocalization and Monte Carlo analyses suggest a genuine association between p53 and specific transposable elements. Wet lab work should be done to verify the existence of the most significant transposable elements identified by the analysis. The RT-PCR method could be used to detect the expression of specific transposable elements and additional knockout experiments could be performed to investigate if the transposable elements have some greater functional role in the cell.

Cancers are the second most common cause of death in the United States, taking the lives of almost 600,000 people each year. P53 has an essential role in preventing cancer with more than half of all primary tumors carrying a p53 mutation. A better understanding of the p53 regulatory network may lead to the advent of novel therapeutics for the improved treatment of these devastating diseases. This project has provided evidence that the regulation of transposable elements is another part of the p53 mechanism and warrants continued investigation.

F. References:

1. SanMiguel, P., et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765–768 (1996)
2. Pray, L. A. (2008). Transposons: The jumping genes. Nat Educ, 1(1), 204.
3. Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2), 71.
4. McClintock B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci.* 1950;36(6):344–55
5. Wylie, A., Jones, A. E., D'Brot, A., Lu, W. J., Kurtz, P., Moran, J. V., ... & Amatruda, J. F. (2016). p53 genes function to restrain mobile elements. *Genes & development*, 30(1), 64-77.
6. Andrysik, Z., Galbraith, M. D., Guarnieri, A. L., Zaccara, S., Sullivan, K. D., Pandey, A., ... & Espinosa, J. M. (2017). Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome research, 27(10), 1645-1657.
7. Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664), 1626-1632.
8. Bao, F.*, LoVerso, P.R.*, Fisk, J.N.*, Zhurkin, V.B. and Cui, F. (2017) p53 binding sites in normal and cancer cells are characterized by distinct chromatin context. *Cell Cycle* 16, 2073-2085. (*student authors)
9. Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci.* USA 104, 18613-18618.
10. Sammons, M. A., Zhu, J., Drake, A. M., & Berger, S. L. (2015). TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. *Genome research*, 25(2), 179-188.
11. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
12. Bao, F., LoVerso, P. R., Fisk, J. N., Zhurkin, V. B., & Cui, F. (2017). p53 binding sites in normal and cancer cells are characterized by distinct chromatin context. *Cell Cycle*, 16(21), 2073-2085.
13. Koike, M., Ishino, K., Kohno, Y., Tachikawa, T., Kartasova, T., Kuroki, T., & Huh, N. H. (1996). DMSO induces apoptosis in SV40-transformed human keratinocytes, but not in normal keratinocytes. *Cancer letters*, *108*(2), 185-193.
14. Kumamoto, K., Spillare, E. A., Fujita, K., Horikawa, I., Yamashita, T., Appella, E., ... & Harris, C. C. (2008). Nutlin-3a activates p53 to both down-regulate inhibitor of growth 2 and up-regulate mir-34a, mir-34b, and mir-34c expression, and induce senescence. *Cancer research*, *68*(9), 3193-3203.
15. Baresova, P., Musilova, J., Pitha, P. M., & Lubyova, B. (2014). p53 tumor suppressor protein stability and transcriptional activity are targeted by Kaposi's sarcoma-associated

herpesvirus-encoded viral interferon regulatory factor 3. *Molecular and cellular biology*, *34*(3), 386-399.
16. Lane, D. P. (1992). Cancer. p53, guardian of the genome. Nature, 358, 15-16.
17. Momand, J., Zambetti, G. P., Olson, D. C., George, D. & Levine, A. J. The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 69, 1237–1245 (1992).
18. Danovi, D. et al. Amplification of Mdmx (or Mdm4) directly contributes to tumor formation by inhibiting p53 tumor suppressor activity. *Mol. Cell. Biol.* 24, 5835–5843 (2004)
19. Haupt, Y., Maya, R., Kazaz, A. & Oren, M. Mdm2 promotes the rapid degradation of p53. *Nature* 387, 296–299 (1997).
20. Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* 432, 316–323 (2004).
21. Todd, E. V., Black, M. A., & Gemmell, N. J. (2016). The power and promise of RNA-seq in ecology and evolution. Molecular ecology, 25(6), 1224-1241.
22. Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, *12*(12), e0190152.
23. Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, *10*(10), 669-680.
24. Kim, T. H., & Dekker, J. (2018). Formaldehyde cross-linking. *Cold Spring Harbor Protocols*, *2018*(4), pdb-prot082594.
25. Nelson, J. D., Denisenko, O., & Bomsztyk, K. (2006). Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nature protocols*, *1*(1), 179.
26. Thomas, R., Thomas, S., Holloway, A. K., & Pollard, K. S. (2017). Features that define the best ChIP-seq peak calling algorithms. Briefings in bioinformatics, 18(3), 441-450.
27. Rebollo, R., Romanish, M. T., & Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual review of genetics*, *46*, 21-42.
28. Dewannieux M, Esnault C, Heidmann T: LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet.* 2003, 35: 41-48. 10.1038/ng1223.
29. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al: Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature.* 2001, 409: 860-921. 10.1038/35057062.
30. Chen C, Ara T, Gautheret D: Using *Alu* elements as polyadenylation sites: A case of retroposon exaptation. *Mol Biol Evol.* 2009, 26: 327-334. 10.1093/molbev/msn249.
31. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y: Widespread establishment and regulatory impact of *Alu* exons in human genes. *Proc Natl Acad Sci USA.* 2011, 108: 2837-2842. 10.1073/pnas.1012834108.
32. Dominissini D, Moshitch-Moshkovitz S, Amariglio N, Rechavi G: Adenosine-to-inosine RNA editing meets cancer. *Carcinogenesis.* 2011, 32: 1569-1577. 10.1093/carcin/bgr124.

33. Piriyapongsa J, Marino-Ramirez L, Jordan IK: Origin and evolution of human microRNAs from transposable elements. *Genetics* 2007,**176**(2):1323-1337.
34. Jjingo, D., Conley, A. B., Wang, J., Mariño-Ramírez, L., Lunyak, V. V., & Jordan, I. K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA*, *5*(1), 1-12.
35. Teng L, Firpi HA, Tan K: Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic Acids Res* 2011,**39**(17):7371-7379. 10.1093/nar/gkr476
36. Scott, E. C., & Devine, S. E. (2017). The role of somatic L1 retrotransposition in human cancers. *Viruses*, *9*(6), 131.
37. Hurst, T. P., & Magiorkinis, G. (2017). Epigenetic control of human endogenous retrovirus expression: focus on regulation of long-terminal repeats (LTRs). *Viruses*, *9*(6), 130.
38. Manghera, M., & Douville, R. N. (2013). Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors?. *Retrovirology*, *10*(1), 16.
39. Dunn, C. A., Romanish, M. T., Gutierrez, L. E., van de Lagemaat, L. N., & Mager, D. L. (2006). Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene*, *366*(2), 335-342.
40. St Laurent, G., Shtokalo, D., Dong, B., Tackett, M. R., Fan, X., Lazorthes, S., ... & Xiao, W. (2013). VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome biology*, *14*(7), R73.
41. Bendall, M. L., De Mulder, M., Iñiguez, L. P., Lecanda-Sánchez, A., Pérez-Losada, M., Ostrowski, M. A., ... & Ormsby, C. E. (2019). Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS computational biology*, 15(9), e1006453.
42. Chung, D., Kuan, P. F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E. H., ... & Keleş, S. (2011). Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput Biol*, *7*(7), e1002111.
43. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048.
44. Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application (No. 1). Cambridge university press.