Comparing RNA-seq alignment tools and how they affect downstream DEG analysis

Matthew Dukeshire, Sawyer Lehman, William Teske

December 10, 2021

1 Introduction

1.1 Background

RNA-seq Technology "RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies" [13]. The development of modern RNA-seq methodologies followed the introduction of Next Generation Sequencing (NGS). "Next generation sequencing (NGS) technologies provide a high-throughput means to generate large amount of sequence data" [9]. One of the resulting challenges of NGS is quality control and the downstream analysis of read data. Alignment tools can calculate scores for each read, based on alignment metrics, in order to perform an early quality control on the reads. The determination of whether certain reads are of high enough quality is done by the alignment tool, and can severely affect any further analysis. RNA-seq data specifically, is used to analyze the expression level of genes at a specific point in time. The expression levels can be quantified with fragments per kilobase of exon model per million mapped reads (FPKM), which can be used to compare the expression of different genes within a sample. However, another analysis that can be done with expression data is to look for differentially expressed genes (DEGs) between two samples. In order to compare expression levels from two different samples, another normalization method is needed, transcripts per million (TPM). TPM normalizes the expression of each feature based on total transcripts, allowing for comparisons between samples. This creates an experimental opportunity where a sample or group can be treated in some way, and the expression can be compared to that of a control [2].

RNA-seq Aligners The purpose of using aligner tools after the RNA-seq read is to figure out what gene the transcript originated from. RNA-seq reads can be unpaired or paired, and modern alignment tools can deal with both. The benefit of paired-end reads is the decrease in the number of reads mapped to multiple locations, as well as a more accurate profile of expression and transcriptome assembly [4]. Additionally, paired-end sequencing has an increased ability for the detection of isoforms and splice junctions.

A problem that arises when aligning an RNA transcript to a genome is that when aligning reads coming from alternatively spliced genes, there will be a very large gap between the exons, thus resulting in low alignment scores for the reads. Several different aligner algorithms that are able to handle short, unspliced RNA-seq reads which align the reads to a reference transcriptome, thus eliminating all intronic regions. Some alignment tools that are meant for unspliced reads are BWA and Bowtie. However, in order to align RNA-seq reads to a reference genome, splice-aware aligners such as STAR and TopHat are required. Aligning to a reference genome is preferred when comparing DEGs as it allows a more accurate analysis of sequence data and allows for the analysis of differential splicing [5]. STAR and TopHat differ immensely in the amount of time and memory required. TopHat internally uses the unspliced Bowtie alignment algorithm but performs a further analysis to identify the splice junctions

[11]. TopHat assigns reads across splice junctions via the split read method. Conversely, STAR has a built in functionality to identify splice junctions and correctly place reads. The STAR algorithm contains two main steps, the seed searching step and the clustering/stitching/scoring step. This method is a more accurate way of placing reads spanning splice junctions [5].

DEG Analysis Following alignment, the coordinates of all reads in the form of a SAM/BAM file are inputted into a DEG, or differentially expressed gene calling tool. Many tools that exist for calling DEGs and they all have small differences. All the tools essentially normalize the transcript counts per feature (gene, ncRNA, etc.), resulting in a TPM value or transcript per million. TPM is used when comparing the relative expression of features between samples. Other RNA-seq analyses may require comparison within a sample between genes, in which FPKM values will be used [15].

Following normalization, these DEG tools will calculate the fold change (FC) between samples for each gene. The tool then can identify genes that are differentially expressed if they exceed a certain level of FC as well as a P-value threshold. The p-value is calculated with the null hypothesis that no difference exists between a gene from two samples. The P-values are adjusted (q-value) to account for multiple testing using the Benjamini and Hochberg method which controls for FDR during multiple testing [1]. When performing DEG analysis, it is important to ensure results are accurate by using multiple biological replicates for each sample. Replicates are just multiple separate samples all treated with the same conditions in order to show the reported results are reproducible. Uniformity within both sets replicates indicates a more robust analysis. The importance of using specific DEGs calling tools is explained by the fact that there is no absolute level of significance and FC that will result in a gene being called, as it varies for each program and experiment. Cuffdiff uses an internally normalized metric for calling DEGs [12], as does DEseq2 [8]. Comparisons of these tools may elucidate any specific behaviors unique to one of the tools. Any found differences can be taken into account in further experiments when selecting a DEG calling tool. The major difference between these tools in regards to DEI (differentially expressed isoform) analysis is that cufflinks can account for the varying expression levels of isoforms and therefore can accurately normalize isoform counts, whereas DEseq2 cannot.

1.2 Thesis

Research Problem While each step of any pipeline intended for identifying DEGs between samples is critical for a robust analysis, alignment tools are incredibly important especially when identifying differential isoforms. This is due to the fact that the placement of reads determines expression levels, which are used to call DEGs. If reads are misplaced it will affect the reported expression levels of each gene, as well as the reported expression levels of the true origin gene. In addition, the number of known isoforms for any particular gene is perpetually increasing as more is discovered, highlighting the importance of identifying specific splice sites to analyze isoforms as NGS data grows. If alignment tools misplace a read that spans a splice junction, it will affect any splice variant analysis.

Due to this importance of splice-aware aligners in a pipeline of DEG identification, we aimed to compare the currently available splice-aware alignment tools to determine what differences may lie between the two in regards to differentially expressed isoform (DEI) analysis, if any. The metrics we focused on were: the number of DEGs called, number of DEIs, uniformity between replicates, computational resources needed and run time. The combined results of the metrics should elucidate which pipeline is recommended when doing DEI analysis.

Expected Results For the alignment tools, we anticipate that the STAR aligner will perform better in identifying splice junctions and thus isoforms than TopHat will. We also confidently predict that the STAR aligner will be faster than TopHat, as it is designed to be a faster, more memory intensive alignment tool [5] For the DEG analysis tools, we believe that cuffdiff will provide a more conservative approach to calling DEGs when compared to DEseq2, and therefore have less numbers of DEGs and DEIs. We hypothesize that the pipeline that will produce the largest number of DEGs will be TopHat followed by DEseq2, regardless of accuracy.

2 Methods

2.1 Data

Datasets Used to Test Methods All raw RNA-seq data is downloaded from the ENCODE project (Illumina-HiSeq 2500, single-end 100nt protocol on total RNA) [3]. The RNA-seq data that is used in this study is originally obtained from the liver tissue. One sample is of a 4-year-old female child (ENCBS401URL) and the other is a 34-year-old adult male with non-obstructive coronary artery disease (ENCBS046RNA). It is not necessarily important that we address causal factors for any deferentially expressed genes (DEGs) that are identified since the scope of this study is to analyze the impact of specific alignment tools on downstream DEG analysis. These specific data sets were chosen because we hypothesized that they would show a large number of DEGs and allow for a better comparison of the alignment tools with respect to the variation in that number. In addition, the data is single-end read. Single-end read data is used since it would result in a much harder task for alignment tools and thus exacerbate the difference between alignment tools. The annotation file used is from ENSEMBLE and the reads from two replicates of each sample are aligned to the Human Genome build GRCh38 [14].

2.2 Software

Tools, Packages, and Pipelines The alignment tools used are STAR and TopHat (Figure 1). These tools are both spliced read aligners, meaning they can detect the presence of different isoforms by taking known splice junctions into account. SAMtools is used whenever converting alignment file types [7]. For the DEG calling tools, cufflinks and DEseq2 are used (Figure 1). Both tools use an internalized normalization method, rather than just TPM. All data visualization is done using the R Bioconductor package (Figure 1). The described programs are put into four separate pipelines, one for each alignment tool, and are analyzed in this study.

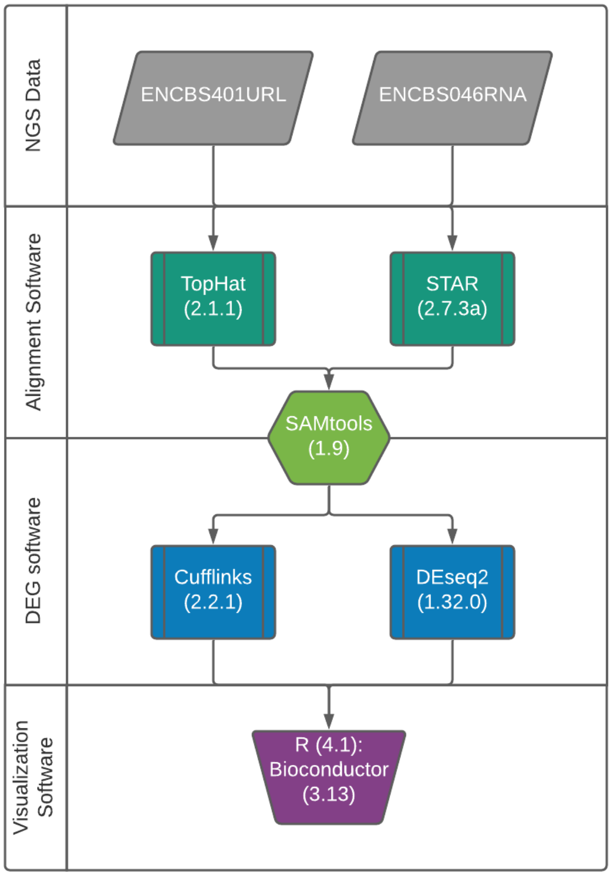


Figure 1: NGS data and Software with versions.

2.3 Experimental design

Overall Approach The following metrics are to be gathered in order to compare the pipelines: number of DEGs called, uniformity between replicates, alignment quality, and run time. This study is not particularly focused on analyzing the biological accuracy of DEG calls as some previous studies have done, but rather the tendencies of each alignment and DEG calling tool in relation to downstream DEG analysis. These tendencies may be that a tool is very conservative in alignment and thus is better for DEG analysis, but it may take more time. Despite this, any DEGs that are found to be unique to a specific pipeline are analyzed further.

RNA-seq Alignment Two separate runs of the alignment tools are done on each RNA-seq read sample allowing for the inclusion of pseudo replicate comparisons. This to see how the alignment tools perform when aligning reads of two separate sequencing runs from the same sample. Ideally, these tools would see near-identical alignments for the two runs. In a true RNA-seq study aiming to gain any functional understanding from a DEG analysis, actual biological replicates would be used rather than just two runs of the same sample. For the intended purposes of the analysis done in this study, running the same biological sample for each sample is sufficient.

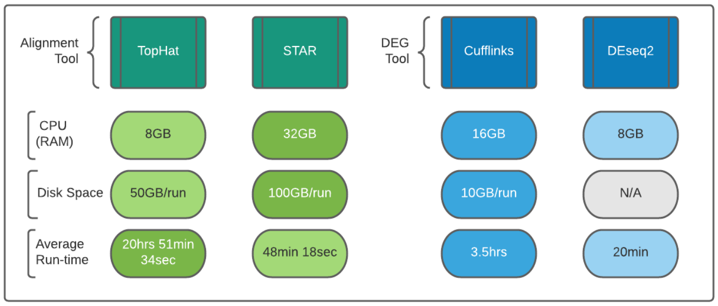
DEG Calling SAM/BAM files are outputted from the alignment tools, and then are used as the input to cufflinks cuffdiff command for detecting DEGs, along with the GFF file containing features and their coordinates. The cutoff q-value (FDR) used by cufflinks in calling DEGs from the cuffdiff output is 0.05. For all pipelines, the following log2 FC threshold is used: —log2(FC)— $\geq$ 1.5. Cuffdiff's differential splicing output is then used for determining the number of isoforms detected. DEseq2 does not provide a feasible method for analyzing isoform counts. As a result, the isoform count metric is only used to compare the alignment tools when using cuffdiff downstream in the pipeline.

Gathering the Metrics The pipelines are run on the carbonate cluster using UNIX bash scripts due to the computational intensity of the tasks (at least 32GB RAM, 100GB storage). The total run time for each pipeline is recorded as reported by the cluster, as well as the total memory required. The presence of differential splicing variants (isoforms) is recorded from analysis of all DEG outputs using R. Uniformity between replicates is manually observed using a genome viewer (IGV) [10] as well as searching for any significant DEGs between replicates. The metrics found from these four pipelines are then visualized using R.

Pipelines using cufflinks are not manually normalized or filtered as it is done within the cufflinks algorithm. Pipelines using DEseq2, however, were normalized using the Bonferonni correction and the results were filtered for sequences with log2 fold change greater than 1.5 and adjusted p-values of less than 0.05.

Additionally, for calculating the number of differentially expressed isoforms cufflinks uses an internal algorithm for compiling the isoforms and the metrics were reported. For DEseq2, isoforms were not automatically filtered and any manual filtering would lead to inaccuracies due to variablility in isoform counts. Cufflinks internally accounts for this variability. In order to get accurate isoform counts and thus the differentially expressed isoforms using DEseq2, other tools must be utilized in order to account for the variability of isoform counts. We report the estimated differentially expressed isoforms from the DEseq2 pipelines, however the accuracy of these specific results is likely to be low.

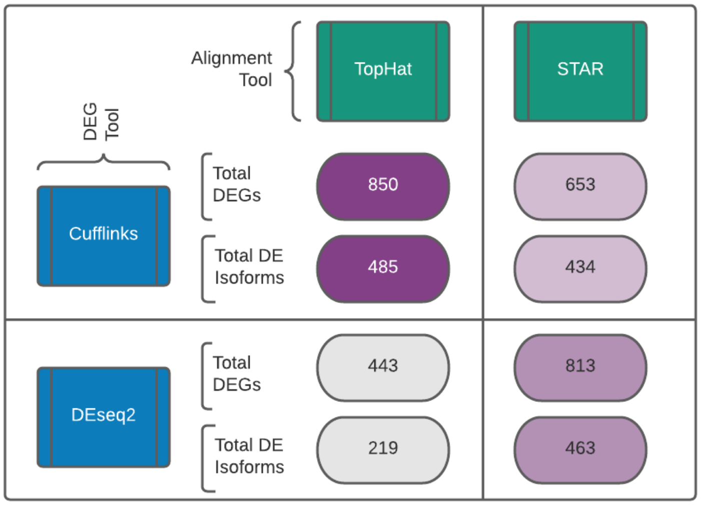
3 Results



	Alignment Tool		DEG Tool	
	TopHat	STAR	Cufflinks	DEseq2
CPU (RAM)	8GB	32GB	16GB	8GB
Disk Space	50GB/run	100GB/run	10GB/run	N/A
Average Run-time	20hrs 51min 34sec	48min 18sec	3.5hrs	20min

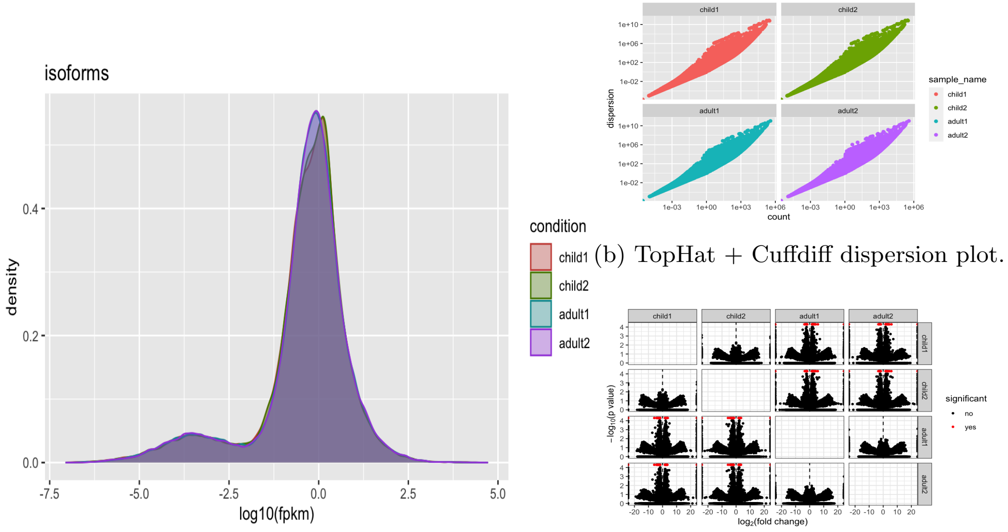
Figure 2: Diagram of memory requirements for optimal usage and average run time for each tool.

Uniformity Across Replicates For all 4 pipelines analyzed, the replicates of each group were paired together. In the two cufflinks pipelines, each set of replicates were clustered together when viewing the dendrogram and the distances between each replicate was shorter than the distance between groups/conditions. Additionally, for the two DEseq2 pipelines, the replicates were clustered closer to each other than to a replicate from the other group. These results simply indicate two things, that our replicates were good enough to show similar expression patterns and that cufflinks and DEseq2 have thresholds high enough, or normalization powerful enough to filter out between replicate differences.



		TopHat	STAR
Cufflinks	Total DEGs	850	653
	Total DE Isoforms	485	434
DEseq2	Total DEGs	443	813
	Total DE Isoforms	219	463

Figure 3: Diagram of DEGs and DE isoforms called in each pipeline.

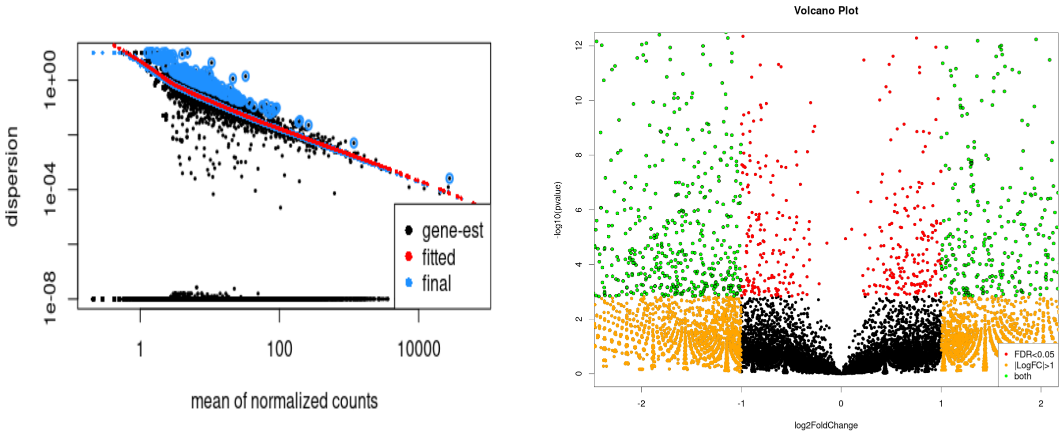


(b) TopHat + Cuffdiff dispersion plot.

(a) TopHat + Cuffdiff density plot.

(c) TopHat + Cuffdiff volcano plot matrix.

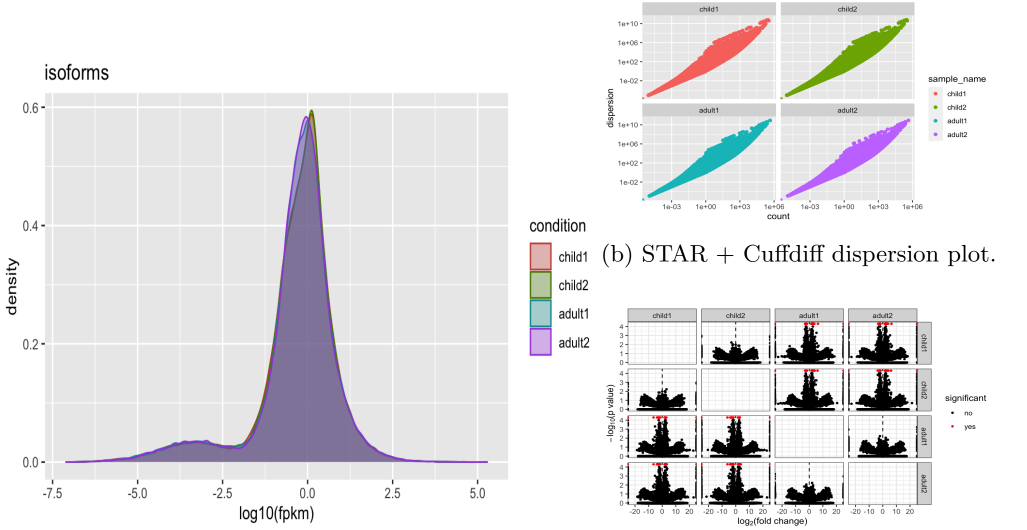
Figure 4: TopHat and Cuffdiff DEG result plots.



(a) TopHat + DEseq2 dispersion plot.

(b) TopHat + DEseq2 volcano plot.

Figure 5: TopHat and DEseq2 DEG result plots.

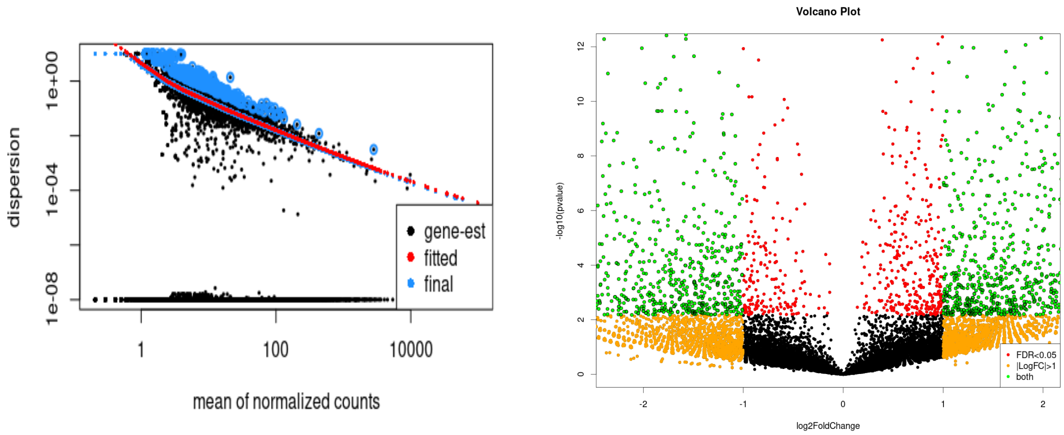


(b) STAR + Cuffdiff dispersion plot.

(a) STAR + Cuffdiff density plot.

(c) STAR + Cuffdiff volcano plot matrix.

Figure 6: STAR and Cuffdiff DEG result plots.



(a) STAR + DEseq2 dispersion plot.

(b) STAR + DEseq2 volcano plot.

Figure 7: STAR and DEseq2 DEG result plots.

4 Conclusion

Alignment tools, Which is better? The alignments of reads to the genome is a crucial step in any analysis which uses sequencing data. From our study, we observed little technical difference between the tools other than the fact that STAR better accounts for splice junctions while TopHat does not. From our results, it can be seen that TopHat actually had higher numbers of both DEGs and DEIs when using the same DEG calling tool. Cufflinks has an internally method for accounting for variation in isoform counts, however it does not move and placed reads following alignment. This means that the relative counts of isoforms are directly correlated to where the alignment tool places reads. Knowing this, the DEIs identified by STAR are likely of higher accuracy. This is also supported by previous studies which looked at the accuracy of read placement, and found that STAR better dealt with isoform identification [6].

Overall, our analysis was not focused solely on which alignment tool is better for isoform detection but also what tools produce the most number of isoforms. Our results indicate that TopHat is the preferred tool when looking to identify the most number of isoforms, but their quality is likely lower on average. However, TopHat uses a fraction of the computing power of STAR making it more accessible to those without high powered computing. The main trade off for computing power is time, a lot of time. TopHat took over 20hours, on average, to align reads from one replicate to the genome. Since 4 samples were used, TopHat was running for over 80 hours to generate 4 BAM files. Conversely, STAR took less than an hour per sample allowing all alignments to be created in just one day. This difference in time and computation power seem to be the most considerable differences, with very little difference in recorded metrics other than STAR having slightly higher accuracy of read placement around splice junctions, as reported by previous studies.

DEG-calling Tools Cufflinks and DEseq2, both used as DEG calling tools in this study, vary greatly in their execution and methods. DEseq2 is an R package primarily used for DEG analysis. One of the main benefits of using DEseq2 over cufflinks, is how customized it can be. Since it is implemented using R functions the user can edit normalization methods, thresholds, and even see which entries are being filtered out. While this does mean the user must know R coding, it can lead to a much more study-specific analysis. Unfortunately, DEseq2 also suffers from the fact that it does not have internal normalization methods for isoform counts. It can normalize all counts, but it does not contain a process for accounting between the extreme variance between isoform counts. Essentially, this means the differentially expressed isoform counts reported by DEseq2 are just DEGs which match to isoform entries in the GTF. Cufflinks, on the other hand, does have the ability to normalize isoform counts and is the preferred DEG calling tool when studying isoforms.

Unlike isoform counts, the total number of DEGs reported by Cufflinks and DEseq2 can be effectively compared as both are reliable methods for calling basic DEGs. Our results indicate that cufflinks has a slightly lower number of DEGs when compared to DEseq2, and this is what we expected. From previous studies, it is known that cufflinks has a more robust count normalization method and therefor consistently reports lower number of DEGs, which are often of higher accuracy [12]. Therefore, cufflinks is the standard for looking at isoform expression, however DEseq2 was included due to it's widespread popularity. Overall, with regards to time and memory requirements, DEseq2 tends to be much faster and requires less memory, making it more accessible and likely contributing to it's popularity despite it's lack of normalization methods.

Choosing DEG Analysis Pipelines In summary, our results confirm our initial hypothesis that cufflinks pipelines will contain less DEGs due to the more stringent normalization methods and thresholds. However, we were surprised to see that the highest number of both DEGs and DEIs came from the TopHat+cufflinks pipeline. Additionally, STAR produces better alignments, specifically around splice junctions. When comparing the total number of DEGs, there wasn't much variation. Therefore, the software chosen for any specific pipeline should reflect the goal of the study. STAR and cufflinks should be used when looking at isoform expression, whereas TopHat could be used when computational resources are limited. In terms of DEG calling tools, Cufflinks is easier to use and produces more consistent and accurate results according to previous studies. Additionally, DEseq2 should primarily be used as a way of further analyzing the counts of specific genes as it's implementation with R allows for easier data wrangling. Overall, for the identification of differentially expressed isoforms, TopHat+Cufflinks produced the greatest amount while STAR+Cufflinks produced slightly less, but also accounts for the greatly varying expression of isoforms leading to more accurate results. DEseq2 is not recommended for use when identifying isoform expression levels as it cannot account for the variance.

5 Appendix

5.1 Unix Code

Alignment and cufflinks: Bash files using TopHat and STAR aligners with cufflinks DEG calling tool.

```bash
#!/bin/bash
#SBATCH --job-name=INTROSTAR
#SBATCH -p general
#SBATCH --mail-user=mdukeshi@iu.edu
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --time=0-08:00:00
#SBATCH --mem=64G
#SBATCH --output=%x.o%j
#SBATCH --error=%x.e%j


module load star/2.7.3a

#do for each replicate

STAR --runThreadN 12 --genomeDir ~/DEG_research/hg38genome/ --sjdbGTFfile ~/DEG_research/Homo_sapiens.GRCh38.79.gtf --sjdbOverhang 99 --readFilesIn /N/slate/mdukeshi/introbioinfo/ --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --outFileNamePrefix /N/slate/mdukeshi/introbioinfo/child1/

#!/bin/bash
#SBATCH --job-name=INTROTOP
#SBATCH -p general
#SBATCH --mail-user=mdukeshi@iu.edu
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=12
#SBATCH --time=0-08:00:00
#SBATCH --mem=128G
#SBATCH --output=%x.o%j
#SBATCH --error=%x.e%j

module load tophat2/2.1.1

tophat2 -p 12 -o /N/slate/mdukeshi/introbioinfo/child1_top/ -G ~/DEG_research/Homo_sapiens.GRCh38.79.gtf --transcriptome-index ~/DEG_research/hg38genome/ /N/slate/mdukeshi/introbioinfo/child1/ENCFF049EBF.fastq.gz

#!/bin/bash
#SBATCH --job-name=star_cuff_a1
#SBATCH -p general
#SBATCH --mail-user=mdukeshi@iu.edu
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=8
#SBATCH --time=0-24:00:00
```

```
#SBATCH --mem=64G
#SBATCH --output=%x.o%j
#SBATCH --error=%x.e%j

module load cufflinks

# do for each replicate

cufflinks -p 8 -g ~/DEG_research/Homo_sapiens.GRCh38.79.gtf /N/slate/mdukeshi/introbioinfo/adult1/Aligned.sortedByCoord.out.bam -o /N/slate/mdukeshi/introbioinfo/adult1/

#!/bin/bash
#SBATCH --job-name=Cuff_Diff
#SBATCH -p general
#SBATCH --mail-user=mdukeshi@iu.edu
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=8
#SBATCH --time=0-24:00:00
#SBATCH --mem=64G
#SBATCH --output=%x.o%j
#SBATCH --error=%x.e%j


module load cufflinks

#cuffdiff with replicates grouped

cuffdiff -o /N/slate/mdukeshi/introbioinfo/ -p 8 -L child,adult /N/u/mdukeshi/Carbonate/merged_asm/merged.gtf \
/N/slate/mdukeshi/introbioinfo/child1/Aligned.sortedByCoord.out.bam, /N/slate/mdukeshi/introbioinfo/child2/Aligned.sortedByCoord.out.bam \
/N/slate/mdukeshi/introbioinfo/adult1/Aligned.sortedByCoord.out.bam, /N/slate/mdukeshi/introbioinfo/adult2/Aligned.sortedByCoord.out.bam
```

5.2 R code

DEseq2: R code used for DEseq2. Example shown is for STAR. Input was changed for TopHat files, but the rest was kept the same:

```
knitr::opts_chunk$set(echo = TRUE)
library(Rsamtools)
library(DESeq2)
library(GenomicAlignments)
library(dplyr)
library(tidyr)

star_bams <- BamFileList(c("/N/slate/mdukeshi/introbioinfo/child1/child1sorted.bam", "/N/slate/mdukeshi/introbioinfo/child2/child2sorted.bam", "/N/slate/mdukeshi/introbioinfo/adult1/adult1sorted.bam", "/N/slate/mdukeshi/introbioinfo/adult2/adult2sorted.bam"))

```

```r
star_gtf = data.table::fread(file = "/N/slate/mdukeshi/introbioinfo/merged_stargtf/merged.gtf")

tail(star_gtf$seqnames)

star_gtf_granges <- makeGRangesListFromDataFrame(star_gtf)

star_gtf <- star_gtf%>%rename(end = V5)
star_gtf <- star_gtf%>%rename(start = V4)
star_gtf <- star_gtf%>%rename(seqnames = V9)
star_gtf <- star_gtf%>%rename(strand = V7)
star_gtf <- star_gtf%>%rename(chr = V1)

star_gtf$seqnames <- as.character(star_gtf$seqnames)
star_gtf <- star_gtf%>%separate(seqnames, sep = "; ", into = c(’gene’,’transcript’, ’exon’, ’oId’, ’extra’))

expdesign = read.csv("design_intro.csv", row.names=1, sep=",")

se_star = summarizeOverlaps(features=star_gtf_granges,
    reads=star_bams, mode="Union", singleEnd=TRUE,
    ignore.strand=TRUE )

counts = assay(se_star)

countgenenames = gsub("[.][1234567890]", "", row.names(counts))
rownames(counts)=countgenenames

medianCountByGroup = t(apply(counts, 1, tapply,
     expdesign, median))

maxMedian=apply(medianCountByGroup, 1, max)

counts_filtered = counts[maxMedian >= 10,]

cds = DESeqDataSetFromMatrix(countData=counts,
 colData=expdesign,
 design= ~ condition)

cds = estimateSizeFactors(cds)

cds = estimateDispersions(cds)

plotDispEsts(cds)

pre_result <- DESeq(cds)

results_STAR <- results(pre_result)

# NUMBER OF GENES WITH PVALUE<0.05 for STAR
sum(results_STAR$padj < 0.05 , na.rm=T)

sum(results_STAR$log2FoldChange > 1.5 & results_STAR$padj < 0.05, na.rm = T)

# STAR MA PLOT
plotMA(results_STAR, ylim=c(-10,10))
```

```
plotCounts(pre_result, gene=which.min(results_STAR$padj), intgroup="condition")

library(pheatmap)

ntd <- normTransform(pre_result)

STARdf = as.data.frame(results_STAR)

STARdf["gene_names"] <- star_gtf$gene
STARdf["transcript_names"] <- star_gtf$transcript
STARdf["exon_number"] <- star_gtf$exon

STARdf%>%drop_na()

volcanoplot <- function (res, lfcthresh=2, sigthresh=0.05, main="Volcano Plot", legendpos="bottomright", labelsig=TRUE, textcx=1, ...) {
  with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main=main, ...))
  with(subset(res, padj<sigthresh ), points(log2FoldChange, -log10(pvalue), pch=20, col="red", ...))
  with(subset(res, abs(log2FoldChange)>lfcthresh), points(log2FoldChange, -log10(pvalue), pch=20, col="orange", ...))
  with(subset(res, padj<sigthresh & abs(log2FoldChange)>lfcthresh), points(log2FoldChange, -log10(pvalue), pch=20, col="green", ...))
  if (labelsig) {
    require(calibrate)
    with(subset(res, padj<sigthresh & abs(log2FoldChange)>lfcthresh), points(log2FoldChange, -log10(pvalue), cex=textcx, ...))
  }
  legend(legendpos, xjust=1, yjust=1, legend=c(paste("FDR<",sigthresh,sep=""), paste("|LogFC|>",lfcthresh,sep=""), "both"), pch=20, col=c("red","orange","green"))
}
png("diffexpr-volcanoplot.png", 1200, 1000, pointsize=20)
volcanoplot(results_STAR, lfcthresh=1, sigthresh=0.05, textcx=.8, xlim=c(-2.3, 2), ylim=c(0,12))
dev.off()
```

Plotting results: R code used for visualization of cufflinks output. Example shown is for TopHat. Input was changed for STAR + cufflinks output, but the rest was kept the same:

```
library(cummeRbund)
setwd("~/Desktop/MS IUPUI FALL 2021/Python")

all_difft <- readCufflinks('topcuffdata')

csScatter(genes(all_difft), 'child1', 'adult1')

samples(all_difft)

```

```
#Below command is equivalent to looking at the gene_exp.diff file that we spent a lot of time parsing yesterday
gene_diff_datat <- diffData(genes(all_difft))
#Do gene_diff_data followed by tab to see all the variables in this data object

sig_gene_datat  <- subset(gene_diff_datat, (significant ==  ’yes’))

#Count how many we have
nrow(sig_gene_datat)

#Below command is equivalent to looking at the isoform_exp.diff file

isoform_diff_datat <-diffData(isoforms(all_difft))
sig_isoform_datat <- subset(isoform_diff_datat, (significant == ’yes’))

#Count how many we have
nrow(sig_isoform_datat)

dispersionPlot(isoforms(all_difft))


csDensity(isoforms(all_difft))

csDendro(isoforms(all_difft))

MAplot(genes(all_difft),"child1","adult1")

csVolcanoMatrix(isoforms(all_difft))

csDistHeat(isoforms(all_difft))
```

References

[1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)

[2] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al.: A survey of best practices for rna-seq data analysis. Genome biology **17**(1), 1–19 (2016)

[3] Consortium, E.P., et al.: An integrated encyclopedia of dna elements in the human genome. Nature **489**(7414), 57 (2012)

[4] Corley, S.M., MacKenzie, K.L., Beverdam, A., Roddam, L.F., Wilkins, M.R.: Differentially expressed genes from rna-seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. BMC genomics **18**(1), 1–13 (2017)

[5] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)

[6] Jia, C., Guan, W., Yang, A., Xiao, R., Tang, W.W., Moravec, C.S., Margulies, K.B., Cappola, T.P., Li, C., Li, M.: Metadiff: differential isoform expression analysis using random-effects meta-regression. BMC bioinformatics **16**(1), 1–12 (2015)

[7] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009)

[8] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**(12), 1–21 (2014)

[9] Patel, R.K., Jain, M.: Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. PloS one **7**(2), e30619 (2012)

[10] Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**(2), 178–192 (2013)

[11] Trapnell, C., Pachter, L., Salzberg, S.L.: Tophat: discovering splice junctions with rna-seq. Bioinformatics **25**(9), 1105–1111 (2009)

[12] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols **7**(3), 562–578 (2012)

[13] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**(1), 57–63 (2009)

[14] Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al.: Ensembl 2018. Nucleic acids research **46**(D1), D754–D761 (2018)

[15] Zhao, Y., Li, M.C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., McShane, L.M.: Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. Journal of Translational Medicine **19**(1), 1–15 (2021)