



**Federal Aviation
Administration**

DOT/FAA/AM-23/02
Office of Aerospace Medicine
Washington, DC 20591

RNA-Seq Alignment and Differential Expression Software Comparison

Susan K. Munster
Scott J. Nicholson
Hilary A. Uyhelji

Civil Aerospace Medical Institute (CAMI)
Federal Aviation Administration
Oklahoma City, OK 73169

January 2023

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications website (www.faa.gov/go/oamtechreports) and at the National Transportation Library's Repository & Open Science Access Portal (<https://rosap.ntl.bts.gov/>)

Technical Documentation Page

1. Report No. DOT/FAA/AM-23/02	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle RNA-Seq Alignment and Differential Expression Software Comparison		5. Report Date January, 2023	
		6. Performing Organization Code AAM-612	
7. Author(s) Munster SK, Nicholson SJ, Uyhelji HA		8. Performing Organization Report No. DOT/FAA/AM-23/02	
9. Performing Organization Name and Address Civil Aerospace Medical Institute (CAMI) Federal Aviation Administration Oklahoma City, OK 73169		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591		13. Type of Report and Period Covered Technical Report	
		14. Sponsoring Agency Code	
15. Supplementary Notes Author ORCIDs: Munster (0000-0002-3223-0076), Nicholson (0000-0002-2201-744X), Uyhelji (0000-0002-3433-8320) Technical report DOI: https://doi.org/10.21949/1524443			
16. Abstract The twofold goals for this study were to determine an optimum choice for ribonucleic acid sequencing (RNA-Seq) alignment software and to determine which differential expression software packages produced consistent and accurate results. RNA was extracted from blood and pooled to produce homogenous sample material to ensure that any differential expression between samples was attributable to characteristics of downstream processing or software choice. Also, simulated sequence data were produced with a known rate of differential expression. After RNA-Seq, all datasets had alignments (or pseudoalignments) performed by Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR. Feature counts were tabulated and analyzed for differential expression using ALDEx2, baySeq, DEGseq, DESeq2, edgeR, limma, NOISeq, PoissonSeq, and SAMseq (samr), and results were compared. Findings indicated that kallisto, Salmon, and STAR provided superior mapping performance, were quickest, and had the smallest output file size compared to the others tested. The differential expression software DESeq2, edgeR, and limma had the most accurate true positive rate with simulated data and consistently performed as expected with real datasets.			
17. Key Word RNA-Seq, alignment, differential expression		18. Distribution Statement Document is available to the public through the National Transportation Library: https://ntl.bts.gov/ntl	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 41	22. Price NA

Author Note

Funding: This research was funded by the Federal Aviation Administration.

Conflicts of Interest: The authors declare that they have no competing interests.

Author Contributions:

SKM, SJN, and HAU designed the study. SKM performed analyses and writing with input from SJN and HAU. All authors have read and approved this manuscript.

Data Availability: Genetics datasets are stored at NIH dbGaP accession (phs003001.v1.p1).

Acknowledgements

The authors wish to acknowledge the feedback and guidance provided by our reviewers, Christopher Tracy, Ph.D., Kyle Copeland, Ph.D., and Michael Ding, Ph.D.

Table of Contents

Author Note	iii
Acknowledgements.....	iii
Table of Contents	iv
List of Figures	v
List of Tables	v
List of Abbreviations	v
Abstract.....	1
Introduction.....	2
Methods & Materials	3
Sample Preparation	3
Sample Quality Assurance/Quality Control and RNA-Seq	4
Simulated Sample Production	4
Sample Alignment.....	5
Differential Expression Analysis	5
Simulated Data Analysis	7
Results.....	8
Alignment Program Comparisons.....	8
Differential Expression Comparisons	12
Differential Expression Determined by FDR Alone	17
Differential Expression Determined by LFC and FDR Together	18
ROC Curves Plotted Against FDR-Only Simulated Data.....	19
True and False Positive Detection in Simulated Data Comparisons.....	20
Discussion.....	26
Alignment Software	26
Differential Expression Software	26
Conclusion	30
References.....	32

List of Figures

Supplementary Figure 1. ROC Curves for All Alignment/DE Program Combinations.....	38
---	----

List of Tables

Table 1. Alignment, Counts, and DE Organization.....	6
Table 2. Alignment Time Comparison in Minutes.....	9
Table 3. File Storage Space Requirements for Each Alignment Software Program Output File, in GB.....	9
Table 4. Alignment Rates of Each Alignment Software Program.....	10
Table 5. Count Tabulation Rates for Each Alignment/Count Program.....	11
Table 6. Differential Expression Results for Real and Simulated Data.....	13
Table 7. Counts of DE Genes Detected, True Positive Counts, False Positive Counts, and Numbers of Incorrectly Aligned Genes Given for Each Alignment Program & DE Package Combination Given Based on Simulated Data.....	21
Supplementary Table 1. AUC Values for Each ROC Curve.....	41

List of Abbreviations

AUC	area under the curve
DE	differential expression
FAA	Federal Aviation Administration
FDR	false discovery rate
GLM	generalized linear models
HC	high-concentration
KW	Kruskal-Wallis test
LFC	log ₂ fold change
NASA	National Aeronautics and Space Administration
(pseudo)alignment	alignment or pseudoalignment, as appropriate
QA/QC	quality assurance/quality control
RINe	RNA Integrity Number
ROC	receiver operating characteristic
RNA-Seq	ribonucleic acid sequencing

Abstract

The twofold goals for this study were to determine an optimum choice for ribonucleic acid sequencing (RNA-Seq) alignment software and to determine which differential expression software packages produced consistent and accurate results. RNA was extracted from blood and pooled to produce homogenous sample material to ensure that any differential expression between samples was attributable to characteristics of downstream processing or software choice. Also, simulated sequence data were produced with a known rate of differential expression. After RNA-Seq, all datasets had alignments (or pseudoalignments) performed by Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR. Feature counts were tabulated and analyzed for differential expression using ALDEx2, baySeq, DEGseq, DESeq2, edgeR, limma, NOISeq, PoissonSeq, and SAMseq (samr), and results were compared. Findings indicated that kallisto, Salmon, and STAR provided superior mapping performance, were quickest, and had the smallest output file size compared to the others tested. The differential expression software DESeq2, edgeR, and limma had the most accurate true positive rate with simulated data and consistently performed as expected with real datasets.

Introduction

Currently, there are many different alignment programs and even more differential expression (DE) analysis programs and packages for use with ribonucleic acid sequencing (RNA-Seq) data. There have been many comparisons (Baruzzo et al., 2017; Simoneau et al., 2021; Williams et al., 2017), but no general consensus across the field about which aligners and differential expression analysis programs provide the greatest accuracy. A typical workflow to detect DE from RNA-Seq data is to trim low-quality reads and/or adapters from the raw data from each sample, align sequence reads, tabulate gene counts, and then perform analyses to detect DE. This approach often is used to compare levels of expression among samples representing different biological conditions or timepoints. It would be reasonable to expect that because many programs perform similar mapping and DE analysis functions, they should produce similar results. In reality, the differing alignment and analysis packages/programs can vary widely in the time required to complete the work and data storage footprint, alignment performance, and DE results. Ultimately this may impact downstream applications of RNA-Seq studies, such as the discovery of RNA biomarkers (biological indicators) whose expression levels could serve as a surrogate metric for performance, safety, and health conditions.

New approaches for translating raw sequence reads into gene expression counts are continuously being developed and improved. For model or otherwise well-studied species such as humans, traditional approaches generally align sequence reads to a reference genome or transcriptome, then count the number of alignments to each 'feature' of interest (typically, a gene or exon). Innovations include efforts to enable the use of different programming environments such as R (Liao et al., 2019) and approaches to enhance the speed with new search strategies or even pseudoalignment or quasi-mapping that bypass time-consuming traditional alignment steps (Bray et al., 2016; Kim et al., 2019; Patro et al., 2016). While some efforts exist to develop and publish standardized pipelines, as by the National Aeronautics and Space Administration (NASA) GeneLab (Overbey et al., 2021), other studies seek to apply the latest advances in reference dataset availability to improve alignment accuracy (Kaminow et al., 2022). In the present study, a suite of commonly employed alignment programs, including Bowtie2 (Langmead & Salzberg, 2012), HISAT2 (Kim et al., 2019), RSEM (Li & Dewey, 2011), Rsubread (Liao et al., 2019), STAR (Dobin et al., 2013), and pseudoalignment strategies Salmon (Patro et al., 2016) and kallisto (Bray et al., 2016) were evaluated.

Once read locations are mapped, the number of reads mapped to specific regions of the genome can be compared between samples, and numerous software packages exist for this purpose. Many commonly used approaches are based on transforming raw count data and using linear modeling approaches or applying non-normal distributions such as the negative binomial to account for the discrete RNA-Seq raw count datasets. Limma implements a linear model originally developed for differential analysis of continuous expression from microarray data (Ritchie et al., 2015). Substantial updates have occurred over time to integrate the analysis of discrete counts from RNA-Seq data. Limma offers two approaches for transforming count data

for use with linear models; these are referred to as limma-trend and voom (Law et al., 2014). BaySeq uses an empirical Bayes approach and offered an early advancement beyond many of the original RNA-Seq analysis programs limited to pairwise comparisons (Hardcastle & Kelly, 2010; Hardcastle, 2021). Among the generalized linear modeling strategies, a Poisson, binomial, or (more commonly) a negative binomial distribution is applied. Examples of this strategy include PoissonSeq (Li et al., 2012), DEGseq (Wang et al., 2010), DESeq2 (Love et al., 2014), and edgeR (Robinson et al., 2010). Non-parametric modeling approaches are offered by programs including NOISeq (Tarazona et al., 2011, 2015) and the SAMseq command from the samr package (Li & Tibshirani, 2013). ALDEx2 presents yet another approach based on compositional data analysis (Fernandes et al., 2013, 2014; Gloor et al., 2016; Quinn et al., 2018). ALDEx2 allows for multiple statistical tests for DE, including t-tests; Kruskal-Wallis; general linearized models; and Pearson, Spearman, and Kendall correlations combined with resampling data using Monte Carlo simulation.

This study aimed to examine various alignment and DE analysis packages to determine which combination of alignment program and DE analysis provided the most reliable DE results. While previous comparisons have been published by other laboratories (e.g., Williams et al., 2017), constant modifications to software algorithms warrant further analysis. Also, the present study provides a combination of simulated and "real" datasets, with the latter generated from a homogenized human sample source that should result in little, if any, biologically meaningful differential expression among samples. While generating real samples with perfectly known true and false positive rates is nearly impossible, this approach provided the potential for general expectations for assessing computational pipelines. Alignment programs/packages Bowtie2 v. 2.4.1, HISAT2 v.2.2.1, kallisto v. 0.44.0, RSEM v.1.2.28, Rsubread v. 2.4.2, Salmon v. 1.4.0, and STAR v. 2.7.8 were run on each dataset, and counts were tabulated. Each alignment was then analyzed for DE using ALDEx2 v. 1.26.0, baySeq v. 2.28.0, DEGseq v. 1.48.0, DESeq2 v. 1.34.0, edgeR v. 3.36.0, limma v. 3.50.1, NOISeq 2.38.0, PoissonSeq v. 1.1.2, and SAMseq (samr package v. 3.0) for a total of 112 combinations of alignment programs/packages, DE packages, and statistical methods. No one method is anticipated to provide the most reliable results in all conditions, as different datasets and study aims may necessitate tailored approaches. The findings of this work provide insights into relative software performance that may be used to guide method selection in DE experiments for biomarker discovery.

Methods & Materials

Sample Preparation

In sample handling and preparation for RNA-Seq, the overall goal was to test samples with minimal differences between them other than those attributable to technical variation, such as handling or stochastic differences between samples. Sample preparation and sequencing were accomplished under a companion study evaluating RNA purification and concentration techniques (Munster et al., 2022); methods are summarized below. Multiple samples were

collected from three volunteers with informed consent and Federal Aviation Administration (FAA) Institutional Review Board approval. Blood was drawn into PAXgene Blood RNA tubes (BD Biosciences, P/N: 762165). Total RNA was extracted from samples using a PAXgene Blood miRNA kit (QIAGEN, P/N: 763134). During the extraction procedure, after the tubes had been incubated overnight, spun down, washed with RNase-free water, and spun down again, buffer BM1 was added, and then all pellets were resuspended, pooled, and mixed. Aliquots from the pooled mixture were used for extraction, which continued as directed by the kit. One-third of samples were eluted, as directed, in BR5, and two-thirds were eluted in RNase-free water, hereafter referred to as water. After extraction, all water-eluted samples were combined and mixed again, as were BR5-eluted samples. These combined samples had similar concentrations and should not have any differential expression observed. Samples eluted in BR5 were designated HC-BR. Some of the high-concentration water-eluted sample (HC) was diluted with water to a lower concentration and designated as LOW-15. Simulated data were also produced with a known rate of differential expression, as a method of verifying the accuracy of DE predicted by software algorithms.

Sample Quality Assurance/Quality Control and RNA-Seq

All samples were checked for quality assurance/quality control (QA/QC) using Qubit 3.0 (Invitrogen Life Technologies, P/N: Q33216) with a Qubit RNA BR Assay Kit (QIAGEN, P/N: 763134) and Tapestation 4200 (Agilent, P/N: G2991BA) using RNA screentapes (Agilent, P/N: 5067-5576) to measure concentration and the RNA Integrity Number (RIN^e) for each sample. All samples were found to have similar high-quality RIN^e scores of roughly 7.5. The higher concentration samples HC-BR and HC had RNA concentrations of approximately 74 ng/μL (n=3 each), while the diluted LOW-15 samples (n=2) had a concentration of 19.9 ng/μL.

Illumina's TruSeq Stranded Total RNA kit with Ribo-Zero Globin (P/N: 20020612) was used to perform RNA-Seq on 9 μL of each sample using a NovaSeq 6000 (Illumina). QA/QC of fastq files was first performed with FastQC v. 0.11.8 (Andrews, 2010); all files were trimmed with Trimmomatic v. 0.39 (Bolger et al., 2014) and rechecked with FastQC.

Simulated Sample Production

The R package polyester v.1.30.0 (Frazee et al., 2021) was used to simulate RNA-Seq data with a known rate of DE. Data were simulated for all of chromosome 22 using the built-in chr22.fa data from the polyester package with the simulate_experiment() command, and the rate of DE was set at 10% using a count matrix. Twenty samples were simulated, with ten in the control group and ten in the treatment group. There are nearly 45,000 coding and noncoding genes in the Ensembl annotation of the human genome (Ensembl, 2022b). There are 1,056 coding and noncoding genes found on chromosome 22, or approximately 2% of the human genome (Ensembl, 2022a). For this study, an RNA-Seq file of 70 million reads was chosen as a model for our simulated data based on the average read count observed previously in our lab.

Proportionately, a chromosome 22 sequencing depth of 367.5x was chosen to produce simulated fastq files that would contain roughly 1.4 million reads, or 2% of our average fastq file.

Sample Alignment

Over 24 alignment programs were initially investigated before deciding which alignment programs/packages would be used in this study. The programs/packages Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR were chosen to assess alignment. Alignment programs were selected based on factors including number of citations, suitability for handling RNA-Seq data, and public availability. Default settings were used in all instances for each alignment program, and, to maintain consistency, the tabulation of aligned counts was carried out using the Rsubread `featureCounts()` (Liao et al., 2019) function to tabulate results for Bowtie2, HISAT2, Rsubread, and STAR alignments. Gene alignments and counts were completed using `GRCh38.primary_assembly.genome.fa.gz` and `gencode.v36.primary_assembly.annotation.gtf.gz` (Gencode Project, 2020a, 2020b). Salmon and kallisto require both transcript and genome assemblies to align reads to transcripts and then produce counts for genes. The file `gencode.v36.transcripts.fa.gz` (Gencode Project, 2020c) was used for both programs. Kallisto also requires chromosome lengths, which were derived from `GCA_000001405.28_GRCh38.p13_assembly_report.txt` (NCBI, 2020). Count tabulation was performed for kallisto and Salmon using the R packages `tximeta` (Love et al., 2020), `tximport` (Soneson et al., 2016), and `DESeq2`, as outlined by Love et al. (2022). HISAT2 and Bowtie2 both output `.sam` files, so `samtools` (Danecek et al., 2021) was used with each to convert `.sam` files to `.bam` files, and then to convert `.bam` files to sorted `.bam` files for tabulation using `featureCounts()`. RSEM has its own command, `rsem-generate-data-matrix`, that was used instead of `featureCounts()` to generate count tables. All alignments and subsequent analyses were run on a workstation with eight 4.00 GHz cores and 128 GB RAM, operating with RedHat Enterprise Linux version 8.4.

Differential Expression Analysis

A review of existing literature showed more than 20 DE packages available for use. Those used were chosen based on factors including theoretical basis for the analysis, public availability, and the number of citations. Selection criteria of packages was also limited to ensure representation of the differing analysis algorithms available while avoiding duplication of highly similar methods. Differential expression packages `ALDEx2`, `baySeq`, `DEGseq`, `DESeq2`, `edgeR`, `limma`, `NOISeq`, `PoissonSeq`, and `samr`, which includes the `SAMseq` command, were selected for analysis. Default settings were used for all DE analyses (Table 1). The `ALDEx2` offers six statistical test options that were tested: Pearson correlation, Spearman correlation, Kendall correlation, generalized linear models (GLM), Kruskal-Wallace tests (KW), and t-tests. In `limma`, both `voom` and `trend` were assayed. `NOISeq` has options to analyze data as biological replicates or as technical replicates, both of which were tested.

Table 1. Alignment, Counts, and DE Organization.

Alignment program	Counts method	Results of each alignment is run through all 16 DE analysis variations, for a  total of 112 combinations of alignment program and DE analysis	DE analysis package	Specific test used within method (if more than one is available)
Bowtie2	featureCounts		ALDEx2	T-test
Hisat2	featureCounts		ALDEx2	Kruskal-Wallace
RSEM	rsem-generate-data-matrix		ALDEx2	Generalized Linear Model
Rsubread	featureCounts		ALDEx2	Pearson Correlation
STAR	featureCounts		ALDEx2	Spearman Correlation
Salmon	tximeta, tximport, DESeq2		ALDEx2	Kendall Correlation
Kallisto	tximeta, tximport, DESeq2		baySeq	N/A
			DEGSeq	N/A
			DESeq2	N/A
			edgeR	N/A
			limma	trend
			limma	voom
			NOISeq	Biological replicates
			NOISeq	Technical replicates
			PoissonSeq	N/A
			SamSeq	N/A

Note. N/A is given for the specific test used within method if there is only one statistical model used by the method. N/A = not applicable; DE = differential expression.

When possible, differential expression was determined using a false discovery rate (FDR) cutoff of ≤ 0.05 , and a log₂ fold change (LFC) threshold of greater than $|\pm 1|$ was included as an additional selection factor. Not all differential expression software packages included both FDR and LFC as part of their standard output. ALDEx2 outputs FDRs for Pearson correlation, Spearman correlation, Kendall correlation, generalized linear model, and Kruskal-Wallis methods, but LFC is not provided. This also happens with output from the baySeq package. Differential expression counts could only be tabulated for the previously mentioned ALDEx2 methods and baySeq based on FDR alone.

Conversely, output from the SAMseq command only gives values for genes that meet both LFC and FDR thresholds, so no counts from the SAMseq comparison could be reported for DE counts determined only by FDR. The remaining analyses, ALDEx2 t-test, DEGseq, DESeq2, edgeR, limma-trend, limma-voom, NOISeq biological replicates, NOISeq technical replicates, and PoissonSeq, did allow consideration of both FDR and LFC values. For these programs, DE counts were determined using FDR alone as well as using FDR with LFC. Unless otherwise specified, default settings and the program's recommended approach from review of user guides and/or manuals were employed.

Simulated Data Analysis

As part of simulating data, the package polyester outputs a file containing information on each transcript used in the simulation, including the DE multiplier applied to each transcript. All simulated transcripts were assigned an expression level based on a normal distribution. The expression levels of transcripts which are not differentially expressed have a DE multiplier of one, and those which are differentially expressed have a multiplier of ten. There were 918 simulated transcripts used in the simulated dataset used for this study. Transcripts were converted to genes using biomaRt (Durinck et al., 2005; Durinck et al., 2009), producing a list of 561 genes from chromosome 22, including 96 differentially expressed genes and 465 non-differentially expressed genes. Gene assignments were verified using the UCSC Genome Browser (University of California Santa Cruz Genomics Institute, 2022).

The list of known chromosome 22 genes was compared to data from DE comparisons using the simulated dataset. Using the pROC package (v. 1.18.0; Robin et al., 2011), ROC curves were plotted, and area-under-the-curve (AUC) values were determined for each alignment / DE combination, except for comparisons using the SAMseq DE analysis software. Receiver operating characteristic (ROC) curves require a label for each gene to indicate whether it is 'true' DE; differentially expressed genes were labeled as "1," and non-differentially expressed genes were labeled as "0". The label for each gene was taken from the list of known chromosome 22 genes produced by the polyester package, which described each transcript used in the simulation and its status as differentially expressed or not. ROC curves also require a prediction about whether a gene is differentially expressed. Because "true" DE genes were labeled with "1," the genes predicted to be DE should also be near one. Thus, $1 - \text{FDR}$ (or adjusted p-value) was used for this prediction. The SAMseq command outputs LFCs and q-values, this package's analog for

FDR, for DE genes alone. No values are given for non-DE genes, so SAMseq results were not included in the ROC curve or AUC analysis.

Data produced by the polyester package enabled the evaluation of true positive rates for DE genes detected in the simulation. For each combination of alignment software and differential expression analysis package, two lists of DE genes were made. One of the lists was all of the DE genes, based solely on an FDR of ≤ 0.05 . The second list of DE genes contained only those with an FDR of ≤ 0.05 and a minimum LFC of $|\pm 1|$. Each compiled DE list was then compared to the list of known true positive genes and also to the list of all possible genes from chromosome 22 used in the simulation. Numbers of true positive, false positive, and incorrectly aligned genes were reported for each alignment and DE software combination.

Results

Alignment Program Comparisons

Average times for (pseudo)alignment varied by nearly tenfold from the fastest to the slowest program (Table 2). Averages were taken across all samples in the study and included index building, alignment, any conversion needed to produce sorted .bam files, and count tabulation. Rsubread had the greatest average at 312.4 minutes per file. Kallisto, Salmon, and STAR had the lowest averages at 34.0, 41.8, and 42.7 minutes per file, respectively. Salmon, STAR, and kallisto have the lowest data storage footprint at an average of 0.01, 10.0, and 9.52 GB per output file respectively (Table 3). HISAT2 and Bowtie2 have the greatest averages, although this is due at least in part to the need to convert .sam to .bam files and then to convert .bam to .bamsorted files for use with featureCounts. This requires three output files per sample rather than a single output file for each sample from alignment. Alignment rates were recorded for programs that generated unique output alignment rates and are all similar, ranging from roughly 80% to 90% for HISAT2, STAR, and Rsubread (Table 4). Bowtie2's alignment rate was distinctly lower, around roughly 50%. Salmon, kallisto, and RSEM do not output alignment rates before tabulation, but do output alignment rates at the point of gene count tabulation. As this is more similar to the percentage of reads assigned during featureCounts(), these rates were listed with the other percentages of reads assigned (Table 5). The percentage of reads assigned in tabulation were similar for Rsubread, STAR, HISAT2, and Bowtie2. These were all tabulated using featureCounts() from the Rsubread package. RSEM uses a command within that program, rsem-generate-data-matrix, to produce a counts table and displays a consistently lower percentage of aligned genes than other software packages. Salmon and kallisto counts were tabulated using a published method using the tximeta and tximport packages and have slightly higher percentages of alignment than is shown by the other five alignment programs used.

Table 2. Alignment Time Comparison in Minutes.

Sample	# reads being aligned (millions)	Rsubread	STAR	HISAT2 + SAMTOOLS (HISAT2 + sam->bam, bam sort)	Salmon	kallisto	RSEM	Bowtie2 + samtools (sam->bam, bam sort)
Index building	NA	57.7	54.2	31.8	25.1	4.6	30	27.0
HC-1	117.4	368.2	41.5	150.0	46.7	42.1	248	240.5
HC-2	91.7	297.3	33.0	114.0	38.5	32.5	205	188.1
HC-3	84.4	272.6	29.5	102.0	35.6	29.5	182	173.8
HC-BR1	96.5	336.3	32.7	105.0	41.3	33.8	201	201.7
HC-BR2	123.5	442	45.2	156.0	52.5	43.9	265	257.6
HC-BR3	94.3	327.6	32.2	119.0	39.8	33.3	200	197.9
LOW-15A	68.6	200.3	26.4	82.0	28.3	26.0	148	126.7
LOW-15B	65.2	183.8	24.6	78.0	26.5	25.3	138	139.1
Count tabulation	NA	46.9	77.8	93.9	1.0	1.0	10	69.7
TOTAL	741.6	2499.2	341.4	964.6	334.6	272.0	1589.9	1572.3
Average per sample	92.7	312.4	42.7	120.6	41.8	34.0	198.7	196.5

Note. Three replicate samples eluted in water (HC), three eluted in BR-5 buffer (HC-BR), and two samples eluted and diluted with water to a low concentration (LOW-15)

Table 3. File Storage Space Requirements for Each Alignment Software Program Output File, in GB.

Sample	Bowtie2 (sam -> bam -> bamsort)	HISAT2 (sam -> bam -> bamsort)	kallisto	RSEM	Rsubread	Salmon	STAR	# Reads before alignment (millions)
HC-1	118	140	12.2	20.2	18.7	0.01	12.6	117.4
HC-2	92.6	110	9.38	15.8	14.3	0.01	10.2	91.7
HC-3	85.8	99	8.68	14.7	13.3	0.01	9.1	84.4
HC-BR1	97.4	113	9.76	16.8	15	0.01	10.3	96.5
HC-BR2	124	147	12.5	21.7	19.3	0.01	13.2	123.5
HC-BR3	95.2	113	9.58	16.5	14.7	0.01	10.2	94.3
LOW-15A	69.9	79.2	7.18	11.8	10.9	0.01	7.4	68.6
LOW-15B	66.3	74.8	6.85	11.2	10.4	0.01	7.1	65.2
AVERAGE	93.65	109.50	9.52	16.09	14.58	0.01	10.0	92.71

Note. Data are given for three replicate samples eluted in water (HC), three eluted in BR-5 buffer (HC-BR), and two samples eluted and diluted with water to a low concentration (LOW-15).

Table 4. Alignment Rates of Each Alignment Software Program.

Sample	Rsubread % uniquely mapped in alignment	STAR % uniquely mapped in alignment	HISAT2 % uniquely mapped in alignment	Bowtie2 % uniquely mapped in alignment
HC-1	89.33	85.89	80.40	51.10
HC-2	88.35	84.94	80.84	50.12
HC-3	89.34	86.29	82.94	50.30
HC-BR1	89.41	85.59	81.32	48.66
HC-BR2	89.13	85.14	80.26	48.48
HC-BR3	89.51	85.73	80.01	48.38
LOW-15A	87.24	84.53	81.72	49.60
LOW-15B	87.05	84.20	81.79	50.74
AVERAGE	88.67	85.29	81.16	49.67

Note. Data are given for three replicate samples eluted in water (HC), three eluted in BR-5 buffer (HC-BR), and two samples eluted and diluted with water to a low concentration (LOW-15).

Table 5. Count Tabulation Rates for Each Alignment/Count Program.

Sample	Rsubread % assigned in tabulation (fC)	STAR % assigned in tabulation (fC)	HISAT2 % assigned in tabulation (fC)	Bowtie2 % assigned in tabulation (fC)	RSEM % aligned (rsem-generate- data-matrix	Salmon (tximeta, tximport, DESeq2)	kallisto (tximeta, tximport, DESeq2)
HC-1	41.8	42.9	39.8	37.0	24.7	47.9	49.4
HC-2	41.9	42.0	38.2	37.0	24.7	49.3	50.7
HC-3	42.1	43.2	39.6	37.3	25.7	49.7	51.1
HC-BR1	43.1	43.6	40.8	37.9	25.5	51.8	52.9
HC-BR2	43.5	44.1	39.6	38.1	25.8	51.8	53.0
HC-BR3	43.4	43.8	41.4	38.2	25.7	51.9	53.0
LOW-15A	39.8	41.8	37.7	35.4	24.5	47.4	48.9
LOW-15B	39.9	41.7	37.6	35.4	24.4	47.2	48.7
AVERAGE	41.9	42.9	39.3	37.0	25.8	49.6	50.9

Note. Data are given for three replicate samples eluted in water (HC), three eluted in BR-5 buffer (HC-BR), and two samples eluted and diluted with water to a low concentration (LOW-15). Tabulation method listed in parentheses at the top of each column. fC = feature counts.

Differential Expression Comparisons

The number of DE genes was determined using FDR and also using FDR and LFC together for each combination of alignment program and DE analysis. Some comparisons, such as those using baySeq and most of the ALDEx2 analyses, did not produce LFC values, so FDR alone was used to identify DE genes for those comparisons. Cutoffs of $>|\pm 1|$ for LFC and < 0.05 for FDR were chosen. All DE counts, based on FDR alone as well as LFC and FDR, are listed in Table 6. It is known from the files used to create the simulated data set that there were a total 561 genes used in the simulation from chromosome 22, and 96 of them were differentially expressed, producing an overall 'true' DE rate of 17.1%. This true rate was compared to the rate of DE detection for each alignment and analysis method comparison using simulated datasets to determine which had detected rates of DE close to the 'true' rate. The simulated dataset was created using transcripts from chromosome 22, but was aligned using the entire annotation genome. Some of the simulated reads were predicted by certain programs to align with sequences not located on chromosome 22, which is why some alignments produced more "aligned" genes than was anticipated based on the transcripts originally used to formulate the simulated dataset.

Table 6. Differential Expression Results for Real and Simulated Data.

Alignment method	DE analysis method	Specific statistical test used	DE GENES BY FDR & LFC				DE GENES BY FDR			
			HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)	HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)
Bowtie2	ALDEx2	GLM	NA	NA	NA	NA	0	0	477	85.03%
HISAT2	ALDEx2	GLM	NA	NA	NA	NA	0	0	435	77.54%
kallisto	ALDEx2	GLM	NA	NA	NA	NA	0	0	424	75.58%
RSEM	ALDEx2	GLM	NA	NA	NA	NA	0	0	351	62.57%
Rsubread	ALDEx2	GLM	NA	NA	NA	NA	0	0	476	84.85%
Salmon	ALDEx2	GLM	NA	NA	NA	NA	0	0	443	78.97%
STAR	ALDEx2	GLM	NA	NA	NA	NA	0	0	467	83.24%
Bowtie2	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	556	99.11%
HISAT2	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	526	93.76%
kallisto	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	522	93.05%
RSEM	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	530	94.47%
Rsubread	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	530	94.47%
Salmon	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	519	92.51%
STAR	ALDEx2	Kendall Corr.	NA	NA	NA	NA	0	0	529	94.30%
Bowtie2	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	557	99.29%
HISAT2	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	527	93.94%
kallisto	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	523	93.23%
RSEM	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	527	93.94%
Rsubread	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	530	94.47%
Salmon	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	517	92.16%
STAR	ALDEx2	Kruskal-Wallace	NA	NA	NA	NA	0	0	528	94.12%
Bowtie2	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	560	99.82%
HISAT2	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	530	94.47%
kallisto	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	526	93.76%
RSEM	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	528	94.12%
Rsubread	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	531	94.65%
Salmon	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	521	92.87%
STAR	ALDEx2	Pearson Corr.	NA	NA	NA	NA	0	0	529	94.30%
Bowtie2	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	556	99.11%

Alignment method	DE analysis method	Specific statistical test used	DE GENES BY FDR & LFC				DE GENES BY FDR			
			HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)	HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)
HISAT2	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	528	94.12%
kallisto	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	524	93.40%
RSEM	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	536	95.54%
Rsubread	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	530	94.47%
Salmon	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	520	92.69%
STAR	ALDEx2	Spearman Corr.	NA	NA	NA	NA	0	0	529	94.30%
Bowtie2	ALDEx2	t-test	0	0	95	16.93%	0	0	558	99.47%
HISAT2	ALDEx2	t-test	0	0	88	15.69%	0	0	528	94.12%
kallisto	ALDEx2	t-test	0	0	93	16.58%	0	0	525	93.58%
RSEM	ALDEx2	t-test	0	0	100	17.83%	0	0	522	93.05%
Rsubread	ALDEx2	t-test	0	0	90	16.04%	0	0	530	94.47%
Salmon	ALDEx2	t-test	0	0	83	14.80%	0	0	519	92.51%
STAR	ALDEx2	t-test	0	0	85	15.15%	0	0	530	94.47%
Bowtie2	baySeq		NA	NA	NA	NA	5	660	551	98.22%
HISAT2	baySeq		NA	NA	NA	NA	6	764	534	95.19%
kallisto	baySeq		NA	NA	NA	NA	30	987	378	67.38%
RSEM	baySeq		NA	NA	NA	NA	25	966	627	111.76%
Rsubread	baySeq		NA	NA	NA	NA	10	702	527	93.94%
Salmon	baySeq		NA	NA	NA	NA	18	1022	512	91.27%
STAR	baySeq		NA	NA	NA	NA	8	782	530	94.47%
Bowtie2	DEGseq		170	748	115	20.50%	4990	12485	638	113.73%
HISAT2	DEGseq		212	899	122	21.75%	6085	14576	611	108.91%
kallisto	DEGseq		329	446	115	20.50%	7841	4696	607	108.20%
RSEM	DEGseq		244	439	177	31.55%	5171	6191	716	127.63%
Rsubread	DEGseq		183	737	108	19.25%	5579	12487	608	108.38%
Salmon	DEGseq		228	353	97	17.29%	6452	4383	581	103.57%
STAR	DEGseq		192	841	98	17.47%	5361	13919	579	103.21%
Bowtie2	DESeq2		0	46	109	19.43%	150	1100	232	41.35%
HISAT2	DESeq2		0	48	100	17.83%	190	1093	212	37.79%
kallisto	DESeq2		3	71	105	18.72%	195	1135	181	32.26%
RSEM	DESeq2		2	34	185	32.98%	165	1002	283	50.45%

Alignment method	DE analysis method	Specific statistical test used	DE GENES BY FDR & LFC				DE GENES BY FDR			
			HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)	HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)
Rsubread	DESeq2		0	30	105	18.72%	167	1090	226	40.29%
Salmon	DESeq2		1	41	90	16.04%	165	1021	167	29.77%
STAR	DESeq2		0	35	95	16.93%	170	1089	208	37.08%
Bowtie2	edgeR		2	133	107	19.07%	304	2095	170	30.30%
HISAT2	edgeR		2	112	95	16.93%	262	1930	141	25.13%
kallisto	edgeR		2	135	97	17.29%	14	635	137	24.42%
RSEM	edgeR		10	132	93	16.58%	142	1632	140	24.96%
Rsubread	edgeR		0	117	102	18.18%	283	2038	158	28.16%
Salmon	edgeR		0	73	91	16.22%	6	471	124	22.10%
STAR	edgeR		1	136	95	16.93%	272	2055	136	24.24%
Bowtie2	limma	trend	0	41	107	19.07%	0	89	167	29.77%
HISAT2	limma	trend	0	24	95	16.93%	0	40	142	25.31%
kallisto	limma	trend	0	95	96	17.11%	0	190	132	23.53%
RSEM	limma	trend	1	32	93	16.58%	1	85	144	25.67%
Rsubread	limma	trend	0	30	102	18.18%	0	65	159	28.34%
Salmon	limma	trend	2	58	88	15.69%	2	186	121	21.57%
STAR	limma	trend	0	45	95	16.93%	0	101	136	24.24%
Bowtie2	limma	voom	2	101	107	19.07%	364	1464	170	30.30%
HISAT2	limma	voom	2	95	95	16.93%	254	1275	141	25.13%
kallisto	limma	voom	5	138	96	17.11%	166	1271	139	24.78%
RSEM	limma	voom	13	117	93	16.58%	192	1148	143	25.49%
Rsubread	limma	voom	0	85	102	18.18%	268	1342	159	28.34%
Salmon	limma	voom	2	95	88	15.69%	83	1151	125	22.28%
STAR	limma	voom	1	98	95	16.93%	257	1308	139	24.78%
Bowtie2	NOISeq	biorep	0	3448	144	25.67%	235	9300	706	125.85%
HISAT2	NOISeq	biorep	134	42	168	29.95%	1638	786	691	123.17%
kallisto	NOISeq	biorep	11	107	227	40.46%	154	548	973	173.44%
RSEM	NOISeq	biorep	2883	52	741	132.09%	29864	689	2231	397.68%
Rsubread	NOISeq	biorep	1644	546	135	24.06%	8310	3162	659	117.47%
Salmon	NOISeq	biorep	89	2276	132	23.53%	1348	5129	671	119.61%
STAR	NOISeq	biorep	7	482	122	21.75%	378	2734	645	114.97%

Alignment method	DE analysis method	Specific statistical test used	DE GENES BY FDR & LFC				DE GENES BY FDR			
			HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)	HC vs. HC-BR	HC vs. LOW15	SIM Control vs. Treatment	SIM DE percent of total (17.1%)
Bowtie2	NOISeq	techrep	71	289	486	86.63%	167	456	486	86.63%
HISAT2	NOISeq	techrep	47	228	469	83.60%	175	492	469	83.60%
kallisto	NOISeq	techrep	42	225	535	95.37%	214	505	535	95.37%
RSEM	NOISeq	techrep	27	132	81	14.44%	82	274	81	14.44%
Rsubread	NOISeq	techrep	57	220	473	84.31%	193	424	473	84.31%
Salmon	NOISeq	techrep	21	124	530	94.47%	90	220	530	94.47%
STAR	NOISeq	techrep	47	221	482	85.92%	186	437	482	85.92%
Bowtie2	PoissonSeq		19	0	0	0.00%	1183	0	0	0.00%
HISAT2	PoissonSeq		22	0	0	0.00%	1150	0	0	0.00%
kallisto	PoissonSeq		39	0	0	0.00%	1030	0	0	0.00%
RSEM	PoissonSeq		38	0	0	0.00%	980	0	0	0.00%
Rsubread	PoissonSeq		15	0	0	0.00%	1112	0	0	0.00%
Salmon	PoissonSeq		27	0	0	0.00%	934	0	0	0.00%
STAR	PoissonSeq		21	0	0	0.00%	1203	0	0	0.00%
Bowtie2	SAMseq		0	0	137	24.42%	NA	NA	NA	NA
HISAT2	SAMseq		410	0	131	23.35%	NA	NA	NA	NA
kallisto	SAMseq		0	0	128	22.82%	NA	NA	NA	NA
RSEM	SAMseq		0	0	196	34.94%	NA	NA	NA	NA
Rsubread	SAMseq		0	0	133	23.71%	NA	NA	NA	NA
Salmon	SAMseq		0	0	111	19.79%	NA	NA	NA	NA
STAR	SAMseq		0	0	123	21.93%	NA	NA	NA	NA

The cutoff for FDR < 0.05 and LFC > |±1|. The true rate of differential expression in the simulated dataset is 17.1%, for comparison to the SIM DE percent of total column. Analyses with rates of differential expression within ± 2% of the known rate of DE appear in bold. SIM = simulated dataset, GLM = generalized linear model, DE = differential expression, Corr. = correlation. NA indicates that LFC was not available from an analysis and could not be used as criteria for DE. If more than one statistical test was available within a method, then each is listed with its results. If left blank, then only one statistical method was available for that DE analysis method.

Differential Expression Determined by FDR Alone

Only one comparison using FDR alone to detect DE had rates close to 17.1%. The RSEM aligned dataset using NOISeq analysis, set to utilize technical replicates, had a DE rate of 14.44% (Table 6). The same number of DE genes were detected in the RSEM aligned NOISeq analyzed comparison using technical replicates when differential expression was determined by both FDR and LFC for the simulated dataset.

ALDEx2 statistical models using generalized linear models, Kendall correlation, Kruskal-Wallis, Pearson correlation, and Spearman correlation had 62% or greater DE rates for the simulated data (Table 6). BaySeq also had high rates of DE genes found in the comparison for the simulated dataset, with rates of 67% to 112% DE observed. BaySeq and all the statistical tests in the ALDEx2 package (other than the t-test) were based solely on FDR. DEGseq analyses using only FDR to determine differential expression had rates of 103.21% or more DE genes detected in the simulated datasets and more than the total number of genes included in the dataset. Results using FDR alone to determine DE for DESeq2, edgeR, limma-trend, and limma-voom, were closer to the expected DE rate of 17.1%, ranging from 21.57% to 50.45%. NOISeq analyses set for biological replicates had much higher rates, ranging from 114.97% to 397.68% DE genes detected, similar to results observed from the DEGseq analysis. NOISeq analyses set for technical replicates ranged from 83.60% to 95.37% DE genes detected, with one comparison (RSEM) having 14.44% DE genes detected. PoissonSeq analyses did not produce DE genes using simulated datasets for any alignment software programs/packages tested. No data was reported for SAMseq using FDR alone to determine differential expression.

While the ALDEx2 Kruskal-Wallis testing, generalized linear model, Pearson correlation, Spearman correlation, and Kendall correlation demonstrated high rates of DE for the simulated dataset comparisons, no DE genes in the HC vs. HC-BR or HC vs. LOW-15 comparisons were detected. The HC vs. HC-BR comparison using the baySeq package produced low numbers of differentially expressed genes, ranging from five to 30 DE genes. The HC vs. LOW-15 baySeq comparison indicated between 660 and 1022 DE genes. For all other alignment combinations and DE analysis methods, the rate of DE detected with the simulated dataset varied but was notably higher than the known number (i.e., 96) of DE genes or a 17.1% rate of differential expression. When comparing numbers of DE genes detected with the real data comparisons between HC vs. HC-BR and HC vs. LOW-15, the number of DE genes generally were, as would be expected, much higher when using FDR alone to determine DE as compared to using both FDR and LFC to determine differential expression. Because FDR, as a means to determine differential expression, was only successful in one instance using simulated data out of the 112 combinations analyzed. Because these results were duplicated using both FDR and LFC to determine DE, the use of FDR alone to determine differential expression DE was not utilized further in analyses.

Differential Expression Determined by LFC and FDR Together

For all remaining comparisons using the simulated dataset's known rate of DE of 17.1%, only those alignment and analysis combinations providing both LFC and FDR were considered. The only ALDEx2 model tested that provides both LFC and FDR is the t-test comparison. It is also the only ALDEx2 method that consistently approaches the expected rate of DE for the simulated data. The Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR alignments of simulated data all resulted in DE rates of 16.93%, 15.69%, 16.58%, 17.83%, 16.04%, 14.80%, and 15.15%, respectively, when processed with the ALDEx2 t-test DE method (Table 6). ALDEx2 t-test did not detect any differential expression in either of the real dataset comparisons HC vs. HC-BR or HC vs. LOW-15.

DEGseq had rates of differential expression using the simulated dataset near the expected rate of 17.1% (Table 6). Data aligned using Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR had rates of DE of 20.50%, 21.75%, 20.50%, 31.55%, 19.25%, 17.29% and 17.47%, respectively. Only the RSEM aligned dataset (31.55%) did not have a differential expression rate close to the expected 17.1% rate known from the creation of the simulated dataset. The true rates for the real data (HC vs. HC-BR and HC vs. LOW-15) comparisons are not known, but are expected to be lower than that observed for the simulated data comparison. Very little or no differences are anticipated in the HC vs. HC-BR dataset, as these differ solely in their elution buffer, yet the numbers of DE genes estimated with DEGSeq range from 170 to 329. A low number of DE genes could be expected in the HC vs. LOW-15 dataset, which only differ in RNA concentration. DEGseq detected between 353 and 899 DE genes in the HC vs. LOW-15 comparison.

DESeq2, edgeR, limma-voom, and limma-trend all demonstrated similar rates of DE detection across alignment options when analyzing the simulated dataset (Table 6). DESeq2 had simulated data DE detection rates varying between 16.04% and 19.43%, except for a rate of 32.98% when using the RSEM aligned dataset. EdgeR, limma-trend, and limma-voom had identical DE rates of 19.07%, 16.93%, 16.58%, 18.18%, and 16.93%, respectively for simulated aligned with Bowtie2, HISAT2, RSEM, Rsubread, and STAR. In the kallisto pseudoalignment, edgeR, limma-trend, and limma-voom had DE rates of 17.29%, 17.11% and 17.11%, respectively. For the Salmon pseudoalignment, their DE rates were 16.22%, 15.69%, and 15.69% for edgeR, limma-trend, and limma-voom, respectively.

The performance of DESeq2, edgeR, limma-voom, and limma-trend was also similar for detecting DE in the real data comparisons (Table 6). As expected, very little to no differential expression was detected in the HC vs. HC-BR comparison, with numbers of DE genes ranging from zero to five (except in RSEM alignments, with DE ranging from two genes up to 13). The HC vs. LOW-15 comparison did demonstrate a greater number of DE genes detected and more variability in the number of DE genes detected by each analysis method. DESeq2 detected between 30 and 71 differentially expressed genes in the HC vs. LOW-15 comparison. EdgeR detected from 73 to 136 DE genes in this comparison. Limma-trend and limma-voom produced

similar numbers of DE genes for the HC vs. LOW-15 comparison, ranging from 24 to 138 DE genes for both analyses. Kallisto mapping generated the highest number of DE genes in DESeq2, limma-voom, and limma-trend analyses.

NOISeq biological replicate testing of the simulated data had rates of DE detection ranging from 21.75% to 132.09% and technical replicates ranging from 14.44% to 94.47%, demonstrating considerable variability across differing alignments (Table 6). Real data DE analyses had similar variabilities, with biological replicate analysis ranging from zero to 2883 DE genes and technical replicate analysis ranging from 21 to 71 for the HC vs. HC-BR comparison. There were between 42 and 3448 DE genes detected with the biological replicates analysis, and 124 to 289 DE genes detected in the technical replicate analysis for the HC vs. LOW-15 comparison. In most of the NOISeq analyses using the simulated dataset, the numbers of DE genes were unexpectedly high. Based on the data used to create the simulated dataset, the total number of genes that should have been aligned was 561 genes on chromosome 22, with 96 differentially expressed. Most of the NOISeq analyses performed using the simulated dataset and using the technical replicates settings had a DE rate of over 80%. NOISeq analyses performed using biological replicates on the simulated dataset had DE rates closer to the expected rate of 17.1%, falling between 21.75% and 40.46%, albeit the RSEM aligned dataset produced a DE rate of 132.09%.

PoissonSeq and SAMseq each displayed consistent but inverse patterns in their differential expression detection. PoissonSeq detected 15 to 39 DE genes in the HC vs. HC-BR comparison but detected none in any HC vs. LOW-15 or simulated data comparisons where differential expression detection was expected (Table 6). SAMseq detected no differential expression in the HC vs. HC-BR or in the HC vs. LOW-15 comparisons of real data, with one exception. There were 410 DE genes detected in the HC vs. HC-BR comparison of the HISAT2 aligned data. SAMseq did detect DE in the simulated data comparison, ranging from 19.79% to 34.94%.

ROC Curves Plotted Against FDR-Only Simulated Data

ROC curves were plotted for all alignment and differential expression package combinations (Supplementary Figure 1), and AUC values were computed (Supplementary Table 1) using results from simulated data alignments and DE analyses and using FDR values to predict differential expression. An ideal ROC curve will increase quickly from zero to one along the y-axis and then continue in a relatively straight line at the top of the graph along the x-axis. For the Bowtie2, HISAT2, Rsubread, and STAR alignments, the results of DE tests with ALDEx2 Pearson correlation, ALDEx2 generalized linear model, ALDEx2 t-test, DESeq2, edgeR, limma-trend, limma-voom, and PoissonSeq all demonstrate this behavior. NOISeq with technical replicates has curves that are close to ideal. Kallisto, RSEM, and Salmon have these results for all DE packages except for DEGseq and NOISeq with biological replicates. AUC values are poor, near 0.5 for all of the DEGseq comparisons and are mostly poor and variable for both the NOISeq biological and technical replicates. Regardless of the alignment algorithm, all

other DE packages besides NOISeq and DEGseq had AUC ratios from approximately 0.83 to nearly 1.0.

True and False Positive Detection in Simulated Data Comparisons

In addition to comparing rates of DE observed from simulated datasets, lists of DE and non-DE genes, as determined by FDR and by FDR and LFC together, were compared to determine if DE genes identified through our analyses matched the original lists of DE genes (herein called "true positives"), genes on chromosome 22 simulated to not show DE (herein "false positives"), or genes incorrectly aligned. The results are given in Table 7. The known number of true positives in the simulated dataset is 96. For the DE gene lists determined by FDR alone, the numbers of false positives are more than double the number of true positives and the counts of incorrectly aligned (not aligned to any of the genes in the chromosome 22 simulation list) genes are the same or greater than those detected using DE lists determined by both FDR and LFC. In this study, using FDR as the sole threshold to determine differential expression does not appear to provide enough benefits to outweigh the disadvantages of producing DE lists containing more false positive genes than true positive genes.

Table 7. Counts of DE Genes Detected, True Positive Counts, False Positive Counts, and Numbers of Incorrectly Aligned Genes Given for Each Alignment Program & DE Package Combination Given Based on Simulated Data.

Alignment program	DE analysis program	Specific statistical test used	Determined by FDR < 0.05 only				Determined by FDR < 0.05 and LFC > ±1			
			# DE genes detected	# True Positive	# False Positive	# aligned incorrectly	# DE genes detected	# True Positive	# False Positive	# aligned incorrectly
Bowtie2	ALDEx2	GLM	477	83	394	63	NA	NA	NA	NA
HISAT2	ALDEx2	GLM	435	83	352	35	NA	NA	NA	NA
kallisto	ALDEx2	GLM	424	86	338	27	NA	NA	NA	NA
RSEM	ALDEx2	GLM	351	82	269	18	NA	NA	NA	NA
Rsubread	ALDEx2	GLM	476	83	393	44	NA	NA	NA	NA
Salmon	ALDEx2	GLM	443	86	357	23	NA	NA	NA	NA
STAR	ALDEx2	GLM	467	83	384	43	NA	NA	NA	NA
Bowtie2	ALDEx2	Kendall Corr.	556	83	473	87	NA	NA	NA	NA
HISAT2	ALDEx2	Kendall Corr.	526	83	443	63	NA	NA	NA	NA
kallisto	ALDEx2	Kendall Corr.	522	86	436	46	NA	NA	NA	NA
RSEM	ALDEx2	Kendall Corr.	530	83	447	65	NA	NA	NA	NA
Rsubread	ALDEx2	Kendall Corr.	530	84	446	59	NA	NA	NA	NA
Salmon	ALDEx2	Kendall Corr.	519	86	433	38	NA	NA	NA	NA
STAR	ALDEx2	Kendall Corr.	529	83	446	62	NA	NA	NA	NA
Bowtie2	ALDEx2	Kruskal-Wallace	557	83	474	88	NA	NA	NA	NA
HISAT2	ALDEx2	Kruskal-Wallace	527	83	444	63	NA	NA	NA	NA
kallisto	ALDEx2	Kruskal-Wallace	523	86	437	47	NA	NA	NA	NA
RSEM	ALDEx2	Kruskal-Wallace	527	83	444	63	NA	NA	NA	NA
Rsubread	ALDEx2	Kruskal-Wallace	530	84	446	59	NA	NA	NA	NA
Salmon	ALDEx2	Kruskal-Wallace	517	86	431	38	NA	NA	NA	NA
STAR	ALDEx2	Kruskal-Wallace	528	83	445	61	NA	NA	NA	NA
Bowtie2	ALDEx2	Pearson Corr.	560	84	476	90	NA	NA	NA	NA
HISAT2	ALDEx2	Pearson Corr.	530	83	447	64	NA	NA	NA	NA
kallisto	ALDEx2	Pearson Corr.	526	86	440	48	NA	NA	NA	NA
RSEM	ALDEx2	Pearson Corr.	528	83	445	62	NA	NA	NA	NA
Rsubread	ALDEx2	Pearson Corr.	531	84	447	60	NA	NA	NA	NA
Salmon	ALDEx2	Pearson Corr.	521	86	435	38	NA	NA	NA	NA
STAR	ALDEx2	Pearson Corr.	529	83	446	62	NA	NA	NA	NA
Bowtie2	ALDEx2	Spearman Corr.	556	83	473	87	NA	NA	NA	NA
HISAT2	ALDEx2	Spearman Corr.	528	83	445	63	NA	NA	NA	NA
kallisto	ALDEx2	Spearman Corr.	524	86	438	46	NA	NA	NA	NA
RSEM	ALDEx2	Spearman Corr.	536	83	453	70	NA	NA	NA	NA

Alignment program	DE analysis program	Specific statistical test used	Determined by FDR < 0.05 only				Determined by FDR < 0.05 and LFC > ±1			
			# DE genes detected	# True Positive	# False Positive	# aligned incorrectly	# DE genes detected	# True Positive	# False Positive	# aligned incorrectly
Rsubread	ALDEx2	Spearman Corr.	530	84	446	59	NA	NA	NA	NA
Salmon	ALDEx2	Spearman Corr.	520	86	434	39	NA	NA	NA	NA
STAR	ALDEx2	Spearman Corr.	529	83	446	62	NA	NA	NA	NA
Bowtie2	ALDEx2	t-test	558	83	475	89	95	71	24	21
HISAT2	ALDEx2	t-test	528	83	445	63	88	73	15	12
kallisto	ALDEx2	t-test	525	86	439	46	93	74	19	16
RSEM	ALDEx2	t-test	522	83	439	56	100	71	29	27
Rsubread	ALDEx2	t-test	530	84	446	59	90	72	18	16
Salmon	ALDEx2	t-test	519	86	433	37	83	75	8	5
STAR	ALDEx2	t-test	530	83	447	62	85	73	12	9
Bowtie2	baySeq		551	83	468	92	NA	NA	NA	NA
HISAT2	baySeq		534	83	451	68	NA	NA	NA	NA
kallisto	baySeq		378	86	292	31	NA	NA	NA	NA
RSEM	baySeq		627	83	544	166	NA	NA	NA	NA
Rsubread	baySeq		527	84	443	61	NA	NA	NA	NA
Salmon	baySeq		512	86	426	36	NA	NA	NA	NA
STAR	baySeq		530	83	447	62	NA	NA	NA	NA
bowtie2	DEGseq		638	84	554	151	115	58	57	39
HISAT2	DEGseq		611	85	526	122	122	59	63	46
kallisto	DEGseq		607	86	521	103	115	60	55	34
RSEM	DEGseq		716	83	633	240	177	57	120	113
Rsubread	DEGseq		608	85	523	119	108	59	49	35
Salmon	DEGseq		581	86	495	75	97	60	37	15
STAR	DEGseq		579	85	494	90	98	60	38	20
Bowtie2	DESeq2		232	83	149	39	109	80	29	26
HISAT2	DESeq2		212	83	129	29	100	81	19	16
kallisto	DESeq2		181	86	95	24	105	83	22	19
RSEM	DESeq2		283	83	200	110	185	80	105	102
Rsubread	DESeq2		226	83	143	32	105	81	24	20
Salmon	DESeq2		167	86	81	12	90	82	8	6
STAR	DESeq2		208	83	125	24	95	81	14	11
Bowtie2	edgeR		170	83	87	33	107	80	27	23
HISAT2	edgeR		141	83	58	21	95	81	14	11
kallisto	edgeR		137	86	51	13	97	83	14	10
RSEM	edgeR		140	83	57	15	93	80	13	10

Alignment program	DE analysis program	Specific statistical test used	Determined by FDR < 0.05 only				Determined by FDR < 0.05 and LFC > ±1			
			# DE genes detected	# True Positive	# False Positive	# aligned incorrectly	# DE genes detected	# True Positive	# False Positive	# aligned incorrectly
Rsubread	edgeR		158	83	75	26	102	81	21	17
Salmon	edgeR		124	86	38	6	91	83	8	4
STAR	edgeR		136	83	53	20	95	81	14	11
Bowtie2	limma	trend	167	83	84	34	107	80	27	23
HISAT2	limma	trend	142	83	59	22	95	81	14	11
kallisto	limma	trend	132	86	46	13	96	83	13	10
RSEM	limma	trend	144	83	61	16	93	80	13	10
Rsubread	limma	trend	159	83	76	27	102	81	21	17
Salmon	limma	trend	121	86	35	6	88	83	5	3
STAR	limma	trend	136	83	53	21	95	81	14	11
Bowtie2	limma	voom	170	83	87	33	107	80	27	23
HISAT2	limma	voom	141	83	58	21	95	81	14	11
kallisto	limma	voom	139	86	53	12	96	83	13	10
RSEM	limma	voom	143	83	60	15	93	80	13	10
Rsubread	limma	voom	159	83	76	27	102	81	21	17
Salmon	limma	voom	125	86	39	6	88	83	5	3
STAR	limma	voom	139	83	56	21	95	81	14	11
Bowtie2	NOISeq	biorep	706	84	622	221	144	58	86	68
HISAT2	NOISeq	biorep	691	85	606	205	168	59	109	92
kallisto	NOISeq	biorep	973	86	887	472	227	60	167	146
RSEM	NOISeq	biorep	2231	83	2148	1756	741	57	684	677
Rsubread	NOISeq	biorep	659	85	574	171	135	59	76	61
Salmon	NOISeq	biorep	671	86	585	167	132	60	72	50
STAR	NOISeq	biorep	645	85	560	159	122	60	62	44
Bowtie2	NOISeq	techrep	486	78	408	54	486	78	408	54
HISAT2	NOISeq	techrep	469	79	390	31	469	79	390	31
kallisto	NOISeq	techrep	535	82	453	41	535	82	453	41
RSEM	NOISeq	techrep	81	57	24	15	81	57	24	15
Rsubread	NOISeq	techrep	473	80	393	33	473	80	393	33
Salmon	NOISeq	techrep	530	82	448	35	530	82	448	35
STAR	NOISeq	techrep	482	81	401	37	482	81	401	37
Bowtie2	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
HISAT2	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
kallisto	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
RSEM	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA

Alignment program	DE analysis program	Specific statistical test used	Determined by FDR < 0.05 only				Determined by FDR < 0.05 and LFC > ±1			
			# DE genes detected	# True Positive	# False Positive	# aligned incorrectly	# DE genes detected	# True Positive	# False Positive	# aligned incorrectly
Rsubread	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
Salmon	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
STAR	PoissonSeq		0	NA	NA	NA	0	NA	NA	NA
Bowtie2	SAMseq		NA	NA	NA	NA	139	0	139	139
HISAT2	SAMseq		NA	NA	NA	NA	132	0	132	132
kallisto	SAMseq		NA	NA	NA	NA	130	0	130	130
RSEM	SAMseq		NA	NA	NA	NA	190	3	187	179
Rsubread	SAMseq		NA	NA	NA	NA	134	0	134	134
Salmon	SAMseq		NA	NA	NA	NA	112	0	112	112
STAR	SAMseq		NA	NA	NA	NA	124	0	124	124

Note. The known values are 96 DE genes, 465 non-DE genes, for a total of 561 genes possible on chromosome 22 in the simulated dataset.

ALDEx2 Pearson correlation, Spearman correlation, Kendall correlation, generalized linear model (GLM), Kruskal-Wallace (KW), and baySeq did not have LFC as part of their comparisons and are not included in the FDR and LFC comparison. SAMseq's output filters DE genes only based on FDR and LFC, so no data was reported for determining DE gene counts by FDR. PoissonSeq detected no significant DE genes in any of the simulated data analyses. If more than one statistical test was available within a differential expression analysis program, each are listed with their results. If left blank, then only one statistical test was available for that DE analysis program. FDR = false discover rate; LFC = log₂ fold change; DE = differential expression, NA = not applicable.

Comparisons between the known list of DE genes, the list of all potential chromosome 22 genes used in the study, and lists of DE genes from the simulated dataset comparisons, show fewer false positives detected and similar or fewer incorrectly aligned genes across all alignment and differential expression package combinations when determined by both FDR and LFC. True positive counts from the ALDEx2 t-test comparison are lower, ranging from 71 to 75, from the FDR and LFC differential expression list, than the counts from the FDR-only differential expression list, but are still close to 96, the known number of DE genes in the simulated dataset. DEGseq analyzed comparisons show the same pattern, with smaller numbers of true positives detected in the FDR and LFC DE list, ranging from 57 to 60, as compared to the FDR-only differential expression list, with counts of true positives falling between 83 and 86. Similarly, analyses using NOISeq with biological replicates demonstrate the same pattern, with true positive counts from FDR-only DE lists ranging from 83 to 86, and true positive counts from both FDR and LFC differential expression lists dropping to a range of 57 to 60. NOISeq with technical replicates had identical results for the number of DE genes detected, number of true positives, number of false positives, and number of incorrectly aligned genes that were identical, when comparing FDR-only and FDR and LFC differential expression lists, with true positive counts ranging from 57 to 82.

DESeq2, edgeR, limma-voom, and limma-trend all detected similar numbers of true positives. DESeq2, edgeR, limma-trend, and limma-voom all had true positive counts ranging between 83 and 86 for the FDR-only DE list and ranging between 80 and 83 for the FDR and LFC differential expression list. Kallisto pseudoaligned comparisons produced 83 true positives in the FDR and LFC differential expression list for DESeq2, edgeR, limma-voom, and limma-trend. Salmon comparisons also produced 83 true positives in the FDR and LFC differential expression list for edgeR, limma-voom, and limma-trend, but not for DESeq2, which had 82 true positives.

Salmon and kallisto both had the highest rates of true positives of all the (pseudo)alignment programs tested. Differential expression analysis carried out using ALDEx t-test, DESeq2, limma-trend, limma-voom and edgeR also had higher numbers of true positives than results from other DE programs used. Results for false positives and incorrectly aligned genes were mixed. Simulated data analyzed with DESeq2 and differential expression determined by both LFC and DE produced a variable number of false positives, ranging from eight to 105 genes, and a number of incorrectly aligned genes ranging from six to 102. Analyses of simulated data using ALDEx2, edgeR, limma-voom, and limma-trend had less variation observed in the number of false positives. DEGseq, NOISeq using biological replicates, NOISeq using technical replicates and SAMseq counts of false positives and incorrectly aligned genes were higher, ranging from 15 to 120, 44 to 684, 15 to 453, and 112 to 187, respectively. SAMseq only had one instance of detecting any true positive genes, where three true positive genes were detected in the RSEM/SAMseq combination. PoissonSeq did not detect any DE genes in any of the analyses.

Discussion

Alignment Software

Considerations of study aims and resource availability may influence the choice of alignment software. Storage of large .bam files, especially in large studies with dozens or hundreds of subjects, can become problematic in terms of file storage capacity. Time consumed performing alignments is another consideration. Alignment programs on a moderately sized workstation can take from approximately 30 minutes to more than 6 hours per sample, pending factors including the number of reads to be aligned. In contrast, many DE packages complete their analyses in minutes or seconds. With a large study, days, weeks, or months could be devoted solely to aligning fastq files, unless multiple machines or a high-performance computing environment is available. While the accuracy of the alignment is of paramount importance, time and memory considerations can be critical in selecting an alignment program, depending on the computational resources available.

Based on the datasets analyzed, it is clear (Table 2) that kallisto, Salmon, and STAR have the lowest average times. On average, all other alignment programs require two or more hours for each alignment performed. File storage capacity can also be a concern when storing large numbers of .bam files. STAR, kallisto, and Salmon are much lower in their file storage footprint compared to other alignment programs. STAR does require a large amount of temporary operating memory while aligning samples, which is a potential detraction. RSEM, and Rsubread require roughly 15 to 16 GB of file storage per sample for .bam (or other output) files, considerably less than Bowtie2 or HISAT2, which each needed close to 100 GB.

Williams et al. (2017) found that the alignment program and tabulation method had less influence than the DE package/program on the results of RNA-Seq data. This is congruent with results from the present study, with the number of DE genes detected and percent of simulated DE genes varying more based on the differential expression analysis method than the alignment method utilized (Table 6). Although the differences are modest, when looking at the percent of DE simulated genes (Table 6) detected by ALDEx2 t-test, DESeq2, edgeR, limma-voom, and limma-trend, kallisto is the alignment software that has the percentages for all of those tests closest to the known rate of DE (i.e., 17.1%). Most of the DE packages tested appear to have high AUC values, even when the true positive rate from the simulated data is far different than expected. When looking at true positive rates, in Table 7, kallisto and Salmon have the highest rates of true positives, even if the differences between alignment programs are small.

Differential Expression Software

The comparison between HC vs. HC-BR was expected to produce little to no differential expression. Other than their differing elution buffers of BR5 and water, these samples were processed identically. The HC vs. LOW-15 comparison was expected to produce a modest number of DE genes, based on input RNA concentration differences between the two samples (Bhargava et al., 2014; Wang et al., 2019). The LOW-15 sample was a dilution of the HC

sample, from 74.2 ng/ μ L to 19.9 ng/ μ L. Any observed differences would be expected to be due to stochastic or technical effects, such as differences due to the dilution of the LOW-15 sample or inefficient amplification of lower-expressing transcripts in low-concentration samples. If any DE is observed, it should be at a minimal rate. Simulated samples were known to have a rate of DE equal to 17.1%, so any DE detected should be close to this rate. Ideally, differential expression results would produce little to no DE genes detected in the HC vs. HC-BR comparison, low numbers of DE genes detected in the HC vs. LOW-15, and a rate of DE expression at approximately 17.1% for the simulated data comparison.

DEGseq detected surprisingly high numbers of differential expression. In the HC vs. HC-BR comparison, 170 to 329 DE genes (Table 6) were detected in analyses from all algorithms. Higher numbers of DE genes, ranging between 353 and 899, were detected in the HC vs. LOW-15 contrast. Somewhat lower rates of DE genes were detected when testing the simulated data, with 97 to 177 genes indicated to be differentially expressed. Most DE levels with simulated data were similar to the anticipated 17.1% (96 genes), with the Salmon and STAR aligned datasets being closest to the known. The RSEM aligned dataset was notably higher, with a DE rate of 31.55%. Overall, performance of DEGseq was close to the expected values for most of the simulated data analyses, but higher than anticipated for both real data comparisons.

ALDEx2 tests using Kruskal-Wallis, generalized linear models, Pearson correlation, Spearman correlation, and Kendall correlation exhibited the same general results across all alignment programs used. All five statistical tests found no differential expression in either the HC vs. HC-BR or the HC vs. LOW-15 comparisons, which is ideal for the HC vs. HC-BR comparison (Table 6). Very low rates of DE detection were anticipated for the HC vs. LOW-15 comparison. There were unusually high rates of DE in the simulated dataset comparison, with DE rates of roughly 63% to almost 100%, where the expected rate is 17.1%. These five variations of the ALDEx2 test performed acceptably with the real data comparison between HC vs. HC-BR but performed very poorly in DE detection with simulated data.

SAMseq exhibited similar behavior in the HC vs. HC-BR and HC vs. LOW-15 comparisons, with no differential expression observed in nearly all comparisons (Table 6). An exception is seen in the HC vs. HC-BR comparison aligned by HISAT2, with 410 DE genes detected. SAMseq has higher than expected rates of DE genes observed with the simulated dataset, ranging from 20% to 35%. In the Bowtie2, HISAT2, kallisto, Rsubread, Salmon, and STAR aligned datasets, none of the DE genes detected by SAMseq were found on the chromosome 22 list used to make the simulated dataset (Table 7). The RSEM dataset had three of 190 DE genes detected that were matched with the list of DE genes from the simulated dataset. The remaining 187 DE genes detected were not located on the chromosome 22 list and thus appear to have been incorrectly aligned.

BaySeq had relatively low numbers of DE detected in the HC vs. HC-BR contrast, from five to 30 DE genes (Table 6). It also had consistently high numbers of DE genes detected in the

HC vs. LOW-15 contrast and in the simulated data contrast. BaySeq does not perform as anticipated with either the HC vs. LOW-15 or the simulated data comparisons.

Similarly, NOISeq had generally lower rates of DE genes detected in the HC vs. BR technical replicates contrast, ranging between 21 and 71 DE genes detected (Table 6). NOISeq displays considerable variability in the numbers of DE genes detected using the biological replicates setting to test real samples in the HC vs. HC-BR and HC vs. LOW-15 comparisons. As all these samples are made from the same mixture of RNA combined from three individuals, these samples would be considered technical replicates, and less variable DE counts were seen in the real data technical replicates data. DE counts from the simulated dataset, which would be more akin to biological replicates, are less variable but still higher than anticipated. The rates of DE gene detection, expected to be near 17.1%, range from 21.75% to 40.46%, with one outlier at 132.09% for the RSEM aligned data. There is less variability observed from the real data sample comparisons HC vs. HC-BR and HC vs. LOW-15 using NOISeq's technical replicates. These samples are technical replicates, so this result could be anticipated. Yet rates of DE are still higher than might be expected, especially in the HC vs. HC-BR contrast, where rates of DE detection would be expected to be closer to zero. Simulated data analyzed using the technical replicate setting were much higher than expected, with rates of DE detection ranging from 83.60% to 95.37%, with one outlier at 14.44% for the RSEM aligned data.

The results from PoissonSeq analysis were unanticipated. PoissonSeq detected low rates of DE (from 15 to 39) in the HC vs. HC-BR comparison, where little to no differential expression was expected (Table 6). There was no DE observed for the HC vs. LOW-15 comparison, where it was anticipated there would be low rates of DE. However, since there was some DE detected in the HC vs. HC-BR comparison, it would also be expected that more DE would be detected in the HC vs. LOW-15 comparison. The difference in RNA concentration in the latter comparison was anticipated to have a greater impact on results than the difference in elution buffer between HC and HC-BR. The simulated data comparison also produced no DE genes observed, in contrast to the expected DE rate of 17.1%. PoissonSeq did not produce results consistent with expectations.

DESeq2, edgeR, limma-voom, limma-trend, and ALDEx2 t-test produced results closer to expectations. ALDEx2 t-test produced no DE genes for the HC vs. HC-BR or the HC vs. LOW-15 comparison (Table 6) in any of the alignments, and may not be as sensitive to gene expression differences. DESeq2, edgeR, limma-voom, and limma-trend produced very low rates of DE for the HC vs. HC-BR comparison, ranging from zero to 13, and higher rates of DE genes observed for the HC vs. LOW-15 comparison, ranging from 24 to 138. Limma-trend produced the lowest number of DE genes. EdgeR produced the largest number of DE genes for most real data comparisons. The DE genes observed in the simulated data comparison demonstrated great consistency between DESeq2, edgeR, limma-voom, and limma-trend. Limma-voom and limma-trend produced identical rates of DE genes for all simulated data comparisons and were identical to edgeR results for alignments performed with Bowtie2, HISAT2, RSEM, Rsubread, and

STAR. DESeq2 had rates of DE genes that were slightly higher than edgeR, limma-voom, and limma-trend for the Bowtie2, HISAT2, kallisto, Rsubread, and Salmon aligned datasets. The STAR aligned dataset analyzed using DESeq2 had the same percentage of DE genes detected from the simulated dataset as those analyzed using edgeR, limma-voom, and limma-trend. The DESeq2 analyzed and RSEM aligned dataset had a distinctly higher rate of DE detection at 32.98%. Salmon and kallisto produced a slightly higher rate of DE genes in the edgeR analysis of simulated data. Rates of DE simulated genes for DESeq2 (excluding the RSEM outlier), limma-trend, limma-voom, and edgeR were between 15.69% and 19.43%, all very close to the expected 17.1% rate of DE genes. ALDEx2 t-test analysis had rates of DE detection of 14.80% to 17.83% in the simulated data comparison.

Further comparisons of programs evaluated the identity of DE genes detected in simulated datasets, as compared to the known DE and non-DE genes. The ALDEx2 t-test consistently had counts of between 71 and 75 out of 96 possible true positives being identified across all alignment programs. DESeq2, edgeR, limma-trend, and limma-voom had identical numbers of true positives detected by each alignment program, with true positive counts of 80, 81, 83, 80, 81, (82 for DESeq2) 83, and 81 for Bowtie2, HISAT2, kallisto, RSEM, Rsubread, Salmon, and STAR, respectively. EdgeR, limma-trend, and limma-voom also generally had lower numbers of false positives than DESeq2 or ALDEx2. When Salmon pseudoaligned data were analyzed with edgeR, limma-trend, or limma-voom, the lowest numbers of false positives and the lowest number of incorrectly aligned genes were observed.

The composition of chromosome 22 may explain how some reads were incorrectly aligned to other parts of the genome. Chromosome 22 was the first completely sequenced human chromosome and is one of the few acrocentric chromosomes, which have all been observed to have repeat sequences and ribosomal RNA genes in tandem on the short arm of the chromosome (Dunham et al., 1999). Chromosome 22 has also been found to have interspersed repeats in as much as 42% of its sequence and to also have roughly 17% of its sequence comprised from *Alu* repeats (Holste et al., 2001). *Alu* elements are found in primate genomes, encompass 11% of the human genome, and are highly successful mobile elements within the human genome (Deininger, 2011). Most of the pseudogenes found on chromosome 22 are thought to have arisen from gene duplication and have both intron and exon structure (Collins et al., 2003). Simulated reads created based on transcripts from chromosome 22 could potentially align with other portions of the full genome, due to factors such as gene duplication and transposition. Hence, the reports of incorrect alignments reported here could be interpreted as over-estimating errors in alignment, if these are correct alignments where a simulated sequence is shared by chromosome 22 and another location.

Conclusion

Based on the counts of 'true positives' in the simulated data, Salmon and kallisto would be the best choices for (pseudo)aligners, although the difference is modest compared to STAR, Rsubread, or HISAT2. Salmon also demonstrated the lowest numbers of false positives and incorrectly mapped genes. Salmon, kallisto, and STAR have the smallest data storage footprint of all output files, and were the quickest at performing (pseudo)alignments. Time can be a significant factor in mapping sequences when processing hundreds or thousands of files. After considering a combination of time requirements, memory footprint, and DE analysis performance, the study supports kallisto, Salmon, and STAR as providing an optimal balance of these qualities for the datasets and mapping approaches tested.

Testing of differential expression software indicates that the ALDEx2 t-test, DESeq2, edgeR, limma-trend, and limma-voom produce results that were the closest to expected values for real data comparisons of HC vs. HC-BR and HC vs. LOW-15 datasets and are most similar to the known data for the simulated data comparison. Time and memory are not generally as much of a consideration for differential expression analyses vs. mapping, as all these analyses are relatively quick and typically do not produce large output files. While there is great consistency between the edgeR, limma-trend, and limma-voom results, this could be expected as edgeR's quasi-likelihood functions similarly moderate dispersion estimates to how limma moderates its variances (Chen et al., 2016). The statistical basis used by DESeq2 is similar to that used in edgeR, which may explain its similar performance. ALDEx2 t-test does not perform as well as edgeR, DESeq2, or either of the limma variations, but it could be used in combination with edgeR, DESeq2, limma-trend, or limma-voom to determine if the same DE genes are detected by analyses using differing statistical models. Further, ALDEx2 could not detect the more subtle DE anticipated in the real data samples. Overall, edgeR, limma-trend, limma-voom, and DESeq2 may be superior to ALDEx2 t-test.

The limitations of simulated data must be acknowledged in any study. Here, it was repeatedly observed that simulated data based on chromosome 22 were predicted by several mapping approaches to contain features aligned to other parts of the genome. It is possible that some of these simulated reads were similar enough to other sequences in the genome to be aligned to other portions of the genome and thus may not have been an ideal test of mapping approaches. It is impossible to know the true expression rates from a real data set, but it is illustrative to compare results to simulated data, indicating the differences between real datasets and simulated data. Any package that behaves similarly with both simulated and real data, where the genes detected as differentially expressed in a simulated dataset are those which truly were DE genes, and where differential expression rates observed in real datasets occur as could be expected for a well understood set of samples, could be thought of as more reliable in its results.

Numerous studies have assessed the performance of differing alignment and differential expression software packages in combination. Williams et al. (2017) and Corchete et al. (2020) both found that the program selected for differential expression analysis had a greater impact

than the choice of alignment program. Some studies do not make a specific recommendation regarding the most optimal alignment and differential expression software options to use (Conesa et al., 2016; Huang et al., 2015; Palajev, 2017). Other studies do make recommendations for a subset of the DE packages tested. Costa-Silva et al. (2017) recommend DESeq2 and limma for differential expression analysis. Ching et al. (2014) and Lin & Pang (2019) both recommend DESeq2 and edgeR. Corchete et al. (2020) noted a high degree of similarity and stability in results for DESeq2, limma, and edgeR. The results observed in this study concur with those general conclusions and tend to support recommending the use of limma, edgeR, and to a lesser extent DESeq2 for DE analysis, along with kallisto, Salmon, or STAR (pseudo)aligners. We also suggest that DE software may be more critical than alignment software, yet both can impact findings.

The final selection of an analytical pipeline should consider the specifics of the study design and modeling needs. For example, the linear models in limma allow one to apply mixed effects models with random effects, whereas the programs edgeR and DESeq2 have been described as requiring alternative strategies (Law et al., 2020). Findings here are designed to guide the selection of (pseudo)alignment and differential expression software, using not only simulated data but the use of real datasets designed to have minimal (technical) variation.

References

- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data (version 0.11.8). [computer software]. Babraham Institute.
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-Seq aligners. *Nature Methods*, 14(2), 135-139. <https://doi.org/10.1038/nmeth.4106>
- Bhargava, V., Head, S. R., Ordoukhanian, P., Mercola, M., & Subramaniam, S. (2014). Technical variations in low-input RNA-Seq methodologies. *Scientific Reports*, 4, 3678. <https://doi.org/10.1038/srep03678>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-Seq quantification. *Nature Biotechnology*, 34(5), 525-527. <https://doi.org/10.1038/nbt.3519>
- Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees: 5 approved]. *F1000Research*, 5, 1438. <https://doi.org/10.12688/f1000research.8987.2>
- Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, 20(11), 1684-1696. <http://www.rnajournal.org/cgi/doi/10.1261/rna.046011.114>
- Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., Bye, J. M., Beare, D. M., & Dunham, I. (2003). Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Research*, 13(1), 27-36. <http://www.genome.org/cgi/doi/10.1101/gr.695703>
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cerveza, A., McPherson, A., Szczęśniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-Seq data analysis. *Genome Biology*, 17(1), 1-19. <https://doi.org/10.1186/s13059-016-0881-8>
- Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-Seq procedures for gene expression quantitative analysis. *Scientific Reports*, 10(1), 1-15. <https://doi.org/10.1038/s41598-020-76881-x>

- 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. <https://doi.org/10.1371/journal.pone.0190152>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.48550/arXiv.2012.10295>
- Deininger, P. (2011). *Alu* elements: know the SINEs. *Genome Biology*, 12(12), 236. <https://doi.org/10.1186/gb-2011-12-12-236>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*, 29(1), 15-21. <https://doi.org/10.1093/bioinformatics/bts635>
- Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., ... O'Brien, K. P. (1999). The DNA sequence of human chromosome 22. *Nature*, 402(6761), 489-495. <https://doi.org/10.1038/990031>
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439-3440. <https://doi.org/10.1093/bioinformatics/bti525>
- Durinck, S., Spellman, P., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8), 1184–1191. <https://doi.org/10.1038/nprot.2009.97>
- Ensembl. (Jul. 2022a). *Human (GRCh38.p13) Chromosome 22: 42,302,381-42,303,381*. http://uswest.ensembl.org/Homo_sapiens/Location/Chromosome?r=22:42302381-42303381
- Ensembl. (Jul. 2022b). *Human (GRCh38.p13) Whole genome*. http://uswest.ensembl.org/Homo_sapiens/Location/Genome?r=Y:1-1000.
- Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., & Gloor, G. B. (2013). ANOVA-Like Differential Gene Expression Analysis of Single-Organism and Meta-RNA-Seq. *PLoS ONE*, 8(7), e67019. <https://doi.org/10.1371/journal.pone.0067019>
- Fernandes, A. D., Reid, J., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, B. G. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-Seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, (2)15. <https://doi.org/10.1186/2049-2618-2-15>

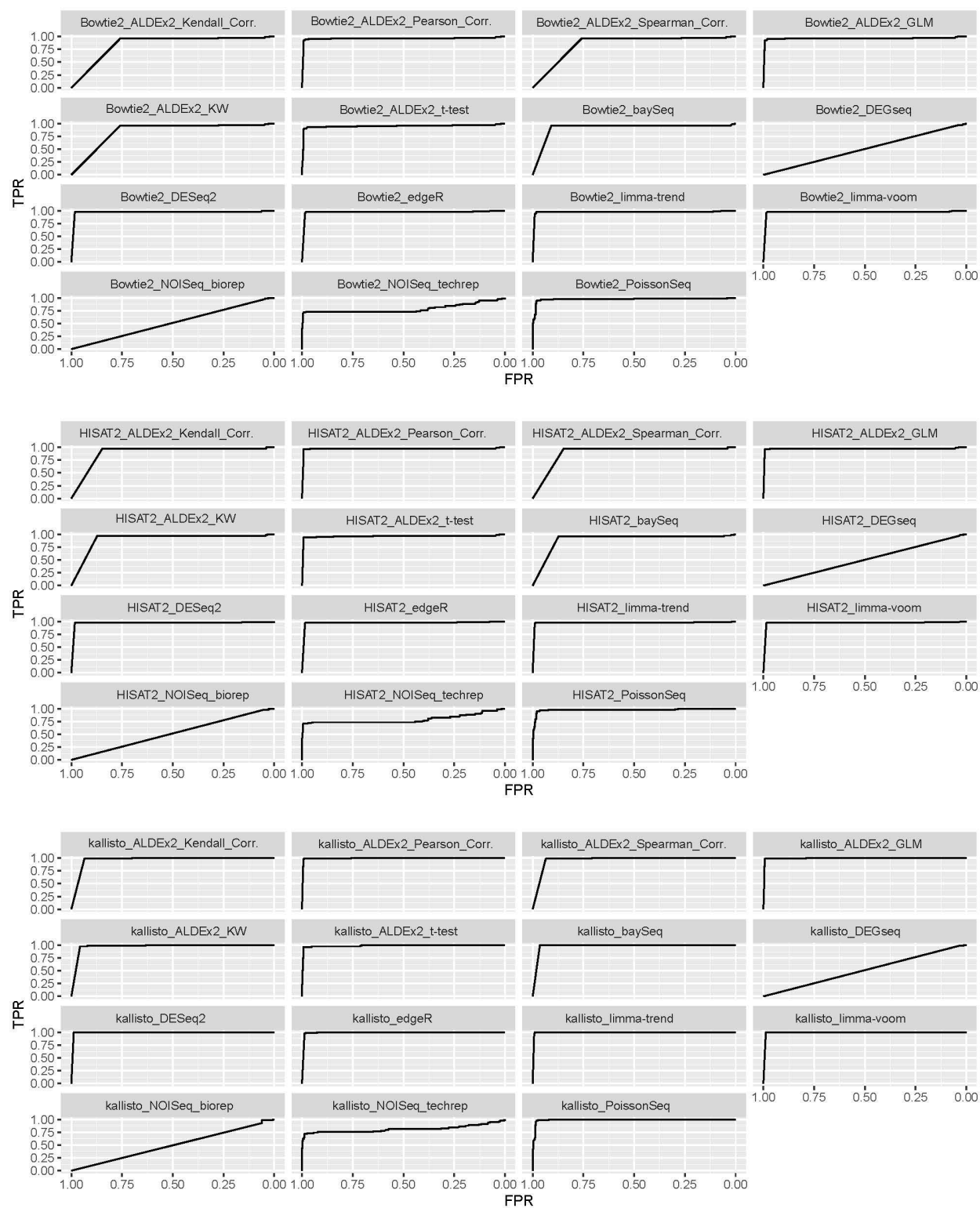
- Frazee, A. C., Jaffe, A. E., Kircher, R., & Leek, J. T. (2021). Polyester: Simulate RNA-Seq reads. R package version 1.32.0. <https://doi.org/doi:10.18129/B9.bioc.polyester>
- Gencode Project. (2020a, November). *GRCh38.primary_assembly.genome.fa.gz* [Data file]. https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/
- Gencode Project. (2020b, November). *gencode.v36.primary_assembly.annotation.gtf.gz* [Data file]. https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/
- Gencode Project. (2020c, November). *gencode.v36.transcripts.fa.gz* [Data file]. https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/
- Gloor, G. B., Macklaim, J. M., & Fernandes, A. D. (2016). Displaying variation in large datasets: plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, 25(3), 971-979. <https://doi.org/10.1080/10618600.2015.1131161>
- Hardcastle, T. J. (2021). baySeq: empirical Bayesian analysis of patterns of differential expression in count data. R package version 2.28.0. <https://doi.org/doi:10.18129/B9.bioc.baySeq>
- Hardcastle, T. J., & Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1), 1-14. <https://doi.org/10.1186/1471-2105-11-422>
- Holste, D., Grosse, I., & Herzog, H. (2001). Statistical analysis of the DNA sequence of human chromosome 22. *Physical Review E*, 64(4), 041917. <https://doi.org/10.1103/PhysRevE.64.041917>
- Huang, H. C., Niu, Y., & Qin, L. X. (2015). Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software. *Cancer Informatics*, 14(S1), 57-67. <https://doi.org/10.4137/CIN.S21631>
- Kaminow, B., Ballouz, S., Gillis, J., & Dobin, A. (2022). Pan-human consensus genome significantly improves the accuracy of RNA-Seq analyses. *Genome Research*, 32(), 738-749. <https://doi.org/10.1101/gr.275613.121>
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907-915. <https://doi.org/10.1038/s41587-019-0201-4>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-359. <https://doi.org/10.1038/nmeth.1923>
- Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-Seq read counts. *Genome Biology*, 15(2), 1-17. <https://doi.org/10.1186/gb-2014-15-2-r29>

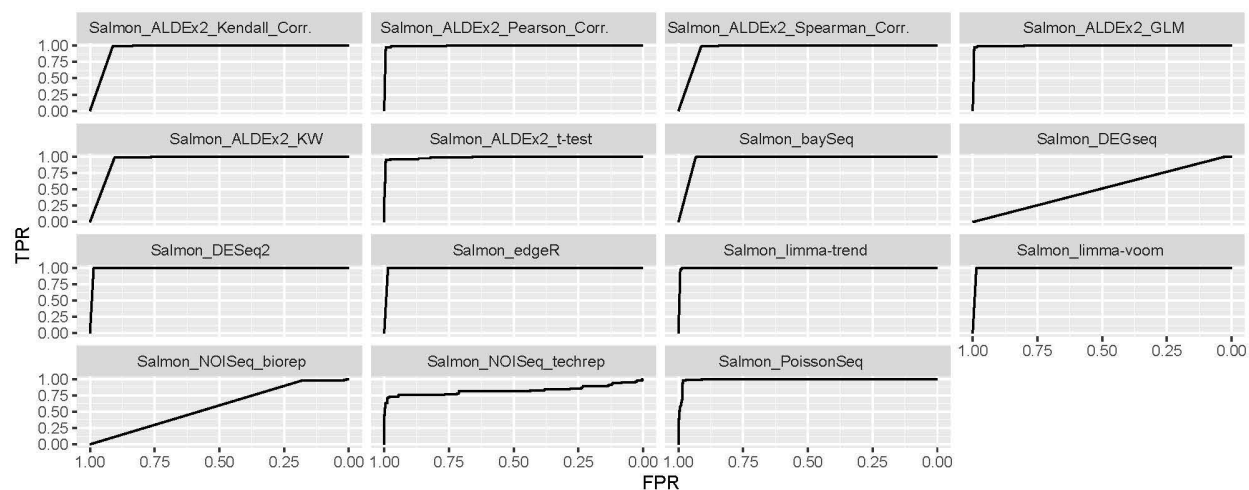
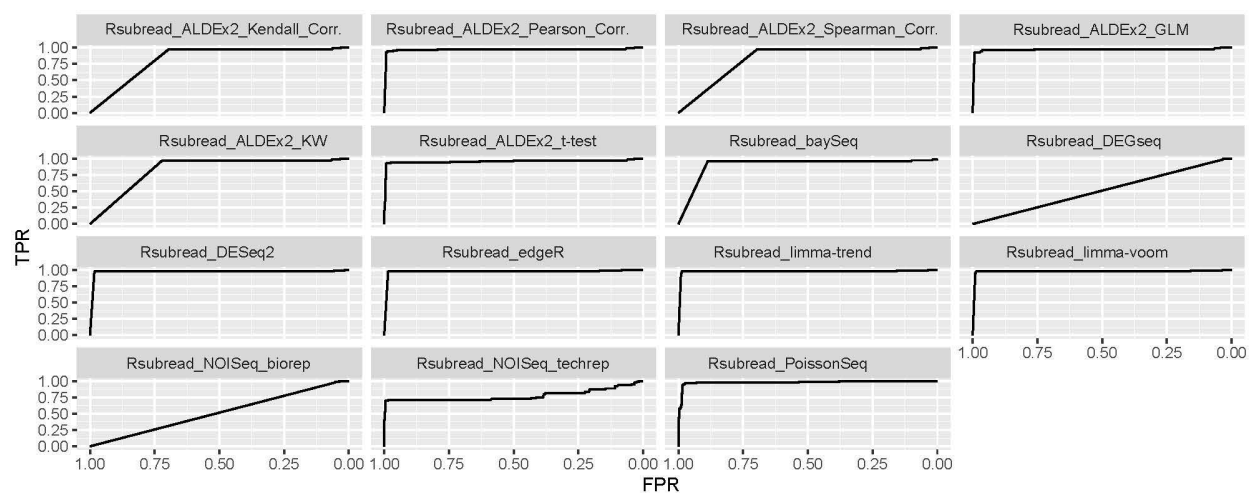
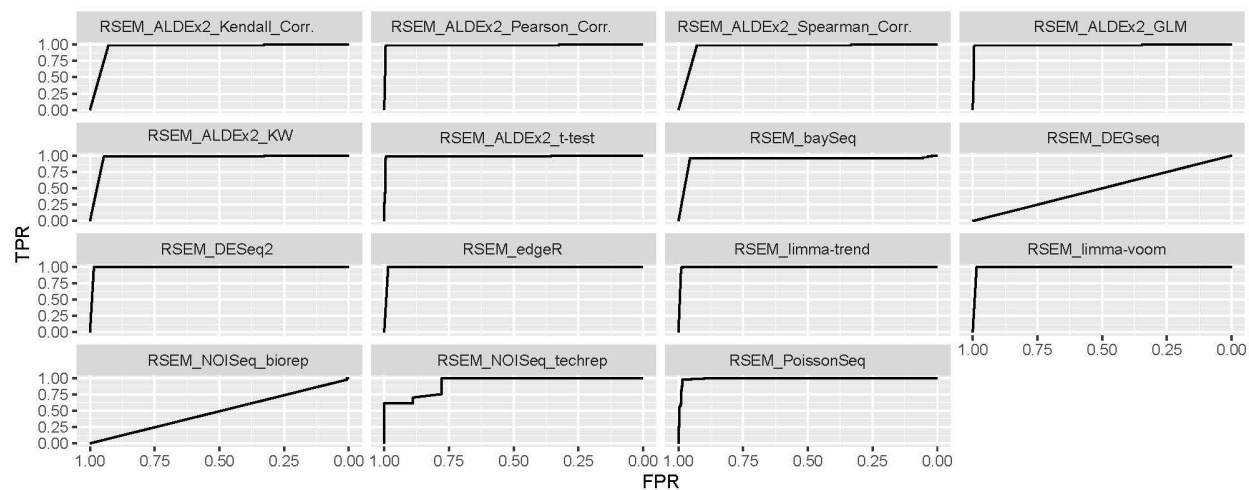
- Law, C. W., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G. K., & Ritchie, M. E. (2020). A guide to creating design matrices for gene expression experiments [version 1; peer review: 2 approved]. *F1000Research*, 9, 1444. <https://doi.org/10.12688/f1000research.27893.1>
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), 1-16. <https://doi.org/10.1186/1471-2105-12-323>
- Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-Sequencing data. *Biostatistics*, 13(3), 523-538. <https://doi.org/10.1093/biostatistics/kxr031>
- Li, J., & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22(5), 519-536. <https://doi.org/10.1177/0962280211428386>
- Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8), e47. <https://doi.org/10.1093/nar/gkz114>
- Lin, B., & Pang, Z. (2019). Stability of methods for differential expression analysis of RNA-Seq data. *BMC Genomics*, 20(1), 1-13. <https://doi.org/10.1186/s12864-018-5390-6>
- Love, M. I., Anders, S., & Huber, W. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15(550). <https://doi.org/10.1186/s13059-014-0550-8>
- Love, M. I., Soneson, C., Hickey, P. F., Johnson, L. K., Pierce, N. T., Shepherd, L., Morgan, M., & Patro, R. (2020). Tximeta: reference sequence checksums for provenance identification in RNA-Seq. *PLoS Computational Biology*, 16(2), e1007664. <https://doi.org/10.1371/journal.pcbi.1007664>
- Love, M. I., Soneson, C., & Robinson, M. D. (2022, Nov. 1). *Importing transcript abundance with tximport*. Bioconductor.org. <https://bioconductor.org/packages/release/bioc/vignettes/tximport/inst/doc/tximport.html#DESeq2>
- Munster, S. K., Uyhelji, H. A., & Nicholson, S. J. (2022). *An evaluation of the downstream effects of purification methods on RNA-Seq differential expression*. [Manuscript in preparation]. Functional Genomics, Federal Aviation Administration.
- NCBI. (2020, September 2). GCA_000001405.28_GRCh38.p13_assembly_report.txt [Data file]. https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.28_GRCh38.p13/

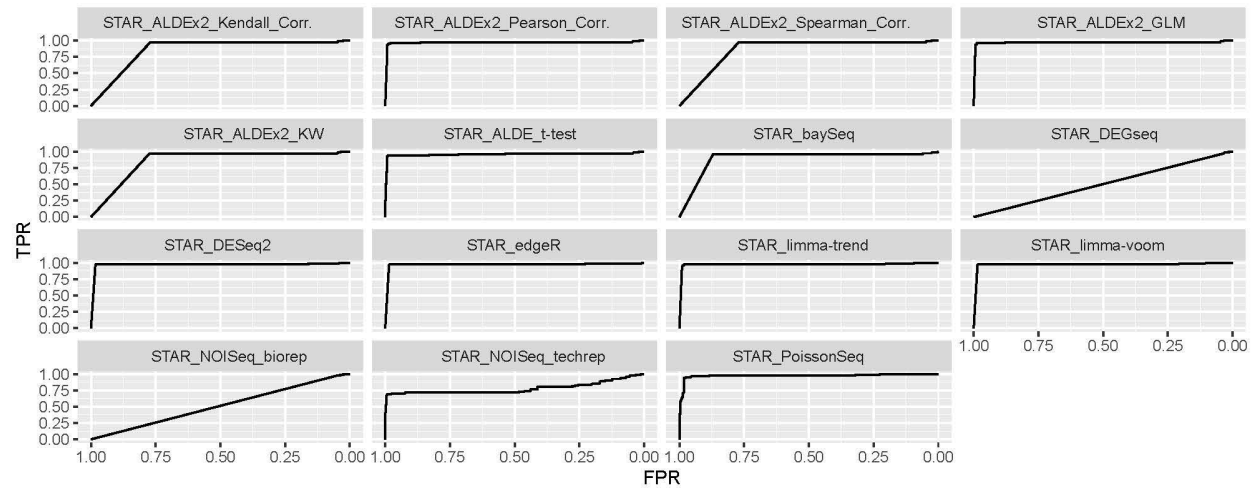
- Overbey, E. G., Saravia-Butler, A. M., Zhang, Z., Rath, K. S., Fogle, H., da Silveira, W. A., Barker, R. J., Bass, J. J., Beheshti, A., Berrios, D. C., Blaber, E. A., Cekanaiciute, E., Costa, H. A., Davin, L. B., Fisch, K. M., Gebre, S. G., Genzita, M., Gilbert, R., Gilroy, S., ... & Galazka, J. M. (2021). NASA GeneLab RNA-Seq consensus pipeline: standardized processing of short-read RNA-Seq data. *iScience*, 24(4), 102361. <https://doi.org/10.1016/j.isci.2021.102361>
- Palajev, D. (2017). Comparison of RNA-Seq differential expression methods. *Cybernetics and Information Technologies*, 17(5), 60-67. <https://doi.org/10.1515/cait-2017-0055>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2016). Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *BioRxiv*, 021592. <https://doi.org/10.1101/021592>
- Quinn, T. P., Crowley, T. M., & Richardson, M. F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics*, 19(1), 1-15. <https://doi.org/10.1186/s12859-018-2261-8>
- Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-Sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47. <https://doi.org/10.1093/nar/gkv007>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(77). <https://doi.org/10.1186/1471-2105-12-77>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140. <https://doi.org/10.1093/bioinformatics/btp616>
- Simoneau, J., Dumontier, S., Gosselin, R., & Scott, M. S. (2021). Current RNA-Seq methodology reporting limits reproducibility. *Briefings in Bioinformatics*, 22(1), 140-145. <https://doi.org/10.1093/bib/bbz124>
- Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-Seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. *F1000Research*, 4, 1521. <https://doi.org/10.12688/f1000research.7563.2>
- Tarazona, S., Furio-Tari, P., Turra, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-Seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21), e140. <https://doi.org/10.1093/nar/gkv711>

- Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-Seq: a matter of depth. *Genome Research*, 21(12), 4436. <http://www.genome.org/cgi/doi/10.1101/gr.124321.111>
- University of California Santa Cruz Genomics Institute. (2022). *Genome Browser Gateway*. <https://genome.ucsc.edu/cgi-bin/hgGateway>
- Wang, L., Felts, S. J., Van Keulen, V. P., Pease, L. R., & Zhang, Y. (2019). Exploring the effect of library preparation on RNA sequencing experiments. *Genomics*, 111(6), 1752-1759. <https://doi.org/10.1016/j.ygeno.2018.11.030>
- Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data. *Bioinformatics*, 26(1), 136-138. <https://doi.org/10.1093/bioinformatics/btp612>
- Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2017). Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 18(1), 1-12. <https://doi.org/10.1186/s12859-016-1457-z>

Supplementary Figure 1. ROC Curves for All Alignment/DE Program Combinations.







Note. Corr. = correlation; DE = differential expression; GLM = generalized linear model; KW = Kruskal-Wallis; biorep = biological replicate; ROC = receiver operating characteristic; techrep = technical replicate; TPR = True positive rate; FPR = false positive rate.

Supplementary Table 1. AUC Values for Each ROC Curve.

		Bowtie2	HISAT2	kallisto	RSEM	Rsubread	Salmon	STAR
ALDEx2	T-test	0.9535	0.9663	0.9894	0.989	0.9603	0.9899	0.963
ALDEx2	KW	0.8521	0.9147	0.9739	0.9655	0.8409	0.9503	0.8666
ALDEx2	GLM	0.964	0.9737	0.9939	0.9889	0.9706	0.9944	0.9724
ALDEx2	Pearson Corr.	0.9622	0.9727	0.9939	0.9888	0.9702	0.994	0.9712
ALDEx2	Spearman Corr.	0.852	0.9027	0.9648	0.9568	0.8288	0.9548	0.8652
ALDEx2	Kendall Corr.	0.8521	0.9027	0.9647	0.9567	0.8288	0.9547	0.8652
baySeq		0.9212	0.9045	0.9823	0.9448	0.9117	0.9665	0.9034
DEGSeq		0.5054	0.5045	0.512	0.5013	0.5115	0.5133	0.5045
DESeq2		0.9695	0.9699	0.994	0.9924	0.969	0.9929	0.9704
edgeR		0.9717	0.9729	0.993	0.9924	0.9721	0.9927	0.972
Limma	Trend	0.9734	0.9742	0.9963	0.9948	0.9738	0.9962	0.9749
Limma	Voom	0.9709	0.9721	0.994	0.9924	0.9738	0.9929	0.9729
NOISeq	Biological reps	0.5121	0.5157	0.4967	0.4919	0.5159	0.5775	0.5133
NOISeq	Technical reps	0.7899	0.792	0.813	0.9269	0.7753	0.8269	0.7791
PoissonSeq		0.9757	0.9766	0.9936	0.9926	0.9809	0.9926	0.975

Note. AUC = area under the curve; GLM = generalized linear model; KW = Kruskal-Wallace; ROC = receiver operating characteristic.