Peer-reviewed version available at *J. Pers. Med.* **2019**, *9*, 18; doi:10.3390/jpm9020018

Article

Aligning the aligners: comparison of RNA sequencing data alignment and gene expression quantification tools for clinical breast cancer research.

Isaac D. Raplee [1], Alexei V. Evsikov [2], and Caralina Marín de Evsikova [1,2,*]

[1] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA; iraplee@health.usf.edu (I.D.R.), cmarinde@health.usf.edu (C.M.d.E.)

[2] Epigenetics & Functional Genomics Laboratory, Department of Research and Development, Bay Pines Veteran Administration Healthcare System, Bay Pines, FL 33744, USA; alexei.evsikov@va.gov (A.V.E.), cmarinde@health.usf.edu (C.M.d.E.)

* Correspondence: cmarinde@health.usf.edu; Tel.: +1-813-974-2248

Abstract: The rapid expansion of transcriptomics and affordability of next-generation sequencing (NGS) technologies generate rocketing amounts of gene expression data across biology and medicine, including cancer research. Concomitantly, many bioinformatics tools were developed to streamline gene expression and quantification. We tested the concordance of NGS RNA sequencing (RNA-seq) analysis outcomes between two predominant programs for read alignment, HISAT2 and STAR, and two most popular programs for quantifying gene expression in NGS experiments, edgeR and DESeq2, using RNA-seq data from breast cancer progression series, which include histologically confirmed normal, early neoplasia, ductal carcinoma *in situ* and infiltrating ductal carcinoma samples microdissected from formalin fixed, paraffin embedded (FFPE) breast tissue blocks. We identified significant differences in aligners' performance: HISAT2 was prone to misalign reads to retrogene genomic loci, STAR generated more precise alignments, especially for early neoplasia samples. edgeR and DESeq2 produced similar lists of differentially expressed genes, with edgeR producing more conservative, though shorter, lists of genes. Gene Ontology (GO) enrichment analysis revealed no skewness in significant GO terms identified among differentially expressed genes by edgeR vs DESeq2. As transcriptomics of FFPE samples becomes a vanguard of precision medicine, choice of bioinformatics tools becomes critical for clinical research. Our results indicate that STAR and edgeR are well-suited tools for differential gene expression analysis from FFPE samples.

Keywords: atypia, breast neoplasms, ductal carcinoma *in situ* (DCIS), gene expression profiling, high-throughput nucleotide sequencing, infiltrating ductal carcinoma (IDC), paraffin embedding, sequence alignment, transcriptome

1. Introduction

After next-generation sequencing (NGS) technology was introduced in 2005, development of many high-throughput bioinformatics tools ensued, such as Bowtie, Tophat, Cufflinks, and CuffDiff "Tuxedo Suite" [1]. One of the most rapidly adopted NGS applications, RNA sequencing (RNA-seq), was introduced in 2008 and captures the transcriptome from cells or tissue samples. Bioinformatics tools have been developed in order to ease read mapping, splice junction, novel gene structure and differential expression analysis of RNA-seq output. Sequence reads generated by RNA-seq can be assessed for single nucleotide polymorphisms (SNPs), splice variants, fusion genes, and individual transcript abundance in samples for differential expression analysis. Even in clinical settings, transcriptomics has been embraced for its potential for precision medicine in diagnoses, prognoses, and therapeutic decisions [2-4]. In comparison to the preceding popular technology for gene expression, microarrays, RNA-seq provides substantially increased transcriptome coverage and is

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

more amenable to discovery science, as the identification of expressed loci or alternative splicing is not limited to the probes present on the array [5,6]. At the same time, gene expression analysis becomes computationally more challenging due to the requirement to correctly classify all sequence read outputs in RNA-seq datasets. As "wet lab" NGS technologies and "dry lab" high performance computers supporting NGS technology became more economical, RNA-seq and other NGS datasets expanded exponentially producing massive amounts of data [7]. Many researchers and clinicians, who utilize RNA-seq in their experiments, rely on outsourcing to cores or companies to generate and analyze their RNA-seq data. This practice is common with many niche experiments, especially in "omics"-based studies.

For an ideal RNA-seq experiment, researchers and clinicians would acquire freshly frozen tissue samples with minimal inclusion of non-target tissues, such as blood and fat, the most common contaminants. In reality, majority of clinical research relies on either archived tissue specimens, mainly as formalin-fixed, paraffin-embedded (FFPE) biopsies, which have increased RNA degradation and decreased poly(A) binding affinity [8-10], or punch biopsies for discrete tissue collection, which severely limits sample material [11]. We analyzed RNA-seq data from bore-dissected samples from FFPE breast tissue, which are known to be highly variable for quality and depth of sequencing results. These transcriptome datasets correspond to the typical quality that biomedical researchers will encounter using transcriptome studies for precision medicine, experiments, or re-analysis. The most common experimental question addressed in RNA-seq transcriptome studies is identifying differential expression among normal stages and stages of disease progression, and sometimes among treatment groups, to pinpoint pathways and putative molecular mechanisms underlying disease or its pathophysiology. To determine differential expression, RNA-seq reads need to be assessed for quality and then aligned to a reference genome. This underscores the essential importance to investigate the impacts of bioinformatics tools for sequence alignment and differential expression on accurate results and interpretation of transcriptome studies collected from FFPE specimens, which was the goal of our study. This paper also serves to point out the pragmatic shortcomings of universally applying a rigid "standard 'omics pipeline" set of bioinformatics tools to RNA-seq data, by demonstrating the impacts and limitations of biological conditions, especially in sample processing and choice of programs.

We focused on two of the most popular sequence aligners, HISAT2 [12], and STAR [13], which superseded the once ubiquitous TopHat aligner program because of their superior computational speed. For differential expression, we used two differential gene expression testing tools, DESeq2 [14] and edgeR [15]. According to citation reports, currently edgeR and DESeq2 are the leading programs for RNA-seq data alignment, with 10,013 and 8,147 citations as of February 22, 2019 [16]. These popular bioinformatics tools are available for users via the open-source Galaxy platform [17], a portal designed to support a wide range of researchers from those with little or no experience or training, to professional bioinformaticians. We also investigated the strengths and weaknesses of the two most widely used differential expression analysis tools, and present a straightforward bioinformatics pipeline from raw data to downstream analysis of gene expression, suitable even for researchers with minimal bioinformatics expertise. We performed differential gene expression analysis using the series of breast cancer progression RNA-seq data from micro punched FFPE samples and assessed similarities and differences in the results to detect the effects of different aligners upon read mapping affecting gene expression counts. Furthermore, we tested the impacts of different algorithms used to detect differential gene expression upon the list of statistically significant transcripts, as well as pathway analysis using Gene Ontology [18] enrichment.

2. Materials and Methods

2.1. Breast cancer samples

RNA-seq data used in this study were deposited and reported (BioProject ID: PRJNA205694 [19]. The dataset represents 72 RNA sequencing experiments from biopsies at different stages of breast cancer: 24 normal tissue samples, 25 early neoplasia (Atypia), 9 ductal carcinoma *in situ* (DCIS) and 14 infiltrating ductal carcinoma (IDC), from 25 patients. Briefly, RNA was extracted from

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

core punches of FFPE specimens after histological confirmation of the cancer stage by a board-certified pathologists and only samples that possessed >90% of luminal cells with the appropriate diagnosis were used for sequencing. Directional cDNA libraries were constructed and sequenced using Illumina GAIIx to obtain 36-base single-end reads.

2.2. RNA-seq reads alignment

We used two programs, STAR and HISAT2, which utilize different strategies to align the RNA-seq reads to the genome assembly. To improve alignment accuracy, both programs use a dataset of known splice sites to correctly identify potential spliced sequencing reads among RNA-seq data. This dataset, in "gene transfer format" (gtf), was obtained from ENSEMBL (release 87, 12/8/2016). Reads were aligned to the human reference genome assembly (hg19).

2.2.1. STAR.

STAR's algorithm [13] uses a two-step approach. STAR aligns the first portion, referred to as "seed", for a specific read sequence output against a reference genome to the maximum mappable length (MML) of the read. Next, STAR aligns the remaining portion, called the "second seed", to its MML. After the read is completely aligned, STAR joins the two or more "seeds" together and scores the aligned reads based on a user-defined penalty for mismatches, insertions, and deletions. The joined seeds with the highest score are chosen as the correct alignment for a specific read sequence output. This approach allows for quick and easy annotation of multi-mapping reads with their own alignment scores. In our analysis, STAR was used with the following parameters:
–seedSearchStartLmax 50 –seedSearchStartLmaxOverLread 1.0 –seedSearchLmax 0 –seedMultimapNmax 10000 –seedPerReadMax 1000 –seedPerWindowMax 50 –seedNonoLociPerWindow 10 –alignIntronMin 21 –alignIntronMax 0 –alignMatesGapMax 0 –alignSJoverhangMin 5 –alignSJDBoverhangmin 3 –alignSpliceMateMapLmin 0 –alignSplicedMateMapeLminOverLmate 0.66 –alignWindowsPerReadNmax 10000 –alignTranscriptsPerWindowNmax 100 –alignTranscriptsPerReadNmax 10000 –alignEndsType Local

2.2.2. HISAT2.

HISAT2 uses the Bowtie2 [20] algorithm to construct and search Ferragina-Manzini (FM) indices [21]. HISAT2 employs two types of indices for aligning: firstly, a whole-genome FM index to anchor alignments and secondly, numerous overlapping local FM indices for alignment extension. HISAT was used for our analysis with the following parameters:
–mp MX=6, MN=2 –sp MX=2,MN=1 –np 1 –rdg 5,3 –rfg 5,3 –score-min L,0,-0.2 –pen-cansplice 0 –pen-noncansplice 12 –pen-canintronlen G,-8,1 –pen-noncanintronlen G,-8,1 –min-intronlen 20 –max-intronlen 500000

2.3. Gene expression counts.

The simplest method for estimating transcript expression is to count the raw reads for each annotated gene locus in the genome assembly. This approach uses a gtf file containing genomic coordinates, such as gene nomenclature, positions in the chromosome for each exon, transcription start site, transcription termination site, etc. We used FeatureCounts [22], to extract information from the binary format for storing sequencing data, i.e., BAM files reads overlapping with genomic features in an input gtf file containing exon coordinates for all transcripts in the genome assembly using the following parameters: *–t 'exon' –g 'gene_id' –M –fraction –Q 12 –minOverlap 30*.

2.4. Data normalization and quality control

2.4.1. Data normalization

To account for the depth of sequencing impacts, which affect read numbers of individual transcripts, we use normalized expression data, specifically counts per million (CPM) [23], to perform quality comparison of datasets. To calculate CPM values, we used the following formula:

$$CPM_i = R_i / T_{Ra} \times 1{,}000{,}000 \quad (1)$$

where CPM_i is a CPM value of a gene in a biological replicate; R_i is the number of reads mapping to all exons of this gene in this biological replicate; T_{Ra} is the total number of reads aligned (anywhere in the genome) from this biological replicate (i.e., the number of aligned reads in either STAR or HISAT2 output "binary alignment map" bam files). This procedure also transforms data from counts to a continuous scale.

2.4.2. Quality control

For quality control, we used ClustVis [24], a statistical tool for clustering of complex data such as RNAseq, based on principal component analysis and visualization of results. Any samples that fell outside of the initial 95% confidence interval on the two-dimensional PCA plot were flagged as outliers and removed before further analysis. ClustVis [24] has an intuitive user interface and was built using several R software packages to provide Principal Component Analysis and heat map plots of high-dimensional data from a data matrix. Data may be uploaded as text file, or manually entered into ClustVis text box. Rows (e.g., genes) and columns (e.g., samples) may contain multiple annotations to be detected automatically, or input manually, to provide additional features (e.g. color for groups, confidence intervals) applied to the PCA and heat map plots.

2.5. Differential gene expression analysis.

Both tools are R packages and require raw read counts in a data matrix which are normalized to account for differences in sequencing depth, and low count variability. Both tools assume RNA-seq data display overdispersion with variance greater than expected for random sampling. Both programs also assume RNA-seq data conform to a negative binomial distribution and employ Bayesian methods to fit raw gene expression counts into this distribution. To display differential expression outputs uniformly, we used the R software package Visualization of Differential Gene Expression using R (ViDGER) [25].

2.5.1. DESeq2

This program, DESeq2, was used to detect differential expression in RNA-seq data. DESeq2 normalizes the counts of each gene employing a generalized linear model [26]. Afterwards, DESeq2 uses an empirical Bayes shrinkage to detect and correct for dispersion and $\log_2$-fold change (LFC) estimates.

2.5.2. edgeR

The other popular program to detect differential expression, edgeR, uses a default method of normalization called trimmed mean of M-values, (TMM), which is obtained with the function: calcNormFactors. This method of normalization estimates the ratio of RNA production through a weighted trimmed mean of the log expression ratios. There are alternative normalization methods available in edgeR to account for data that fail to conform to a negative binomial distribution, which is assumed with TMM. To control for false discovery rate (FDR) we applied the estimateDisp function.

2.6. Gene enrichment analysis using Visual Annotation Display (VLAD)

VLAD [27], accessible via MGI web portal, is a powerful tool to find common functional themes in the lists of genes by analyzing statistical over- or underrepresentation of ontological annotations. Currently, users can choose among Gene Ontology (GO) [18] annotations for human genes, Gene Ontology and Mammalian Phenotype Ontology (MP) [28] annotations for mouse genes, or upload a file of own annotations (in open biomedical ontology [29] 'obo' format). Unlike other packages for ontological enrichment, VLAD uniquely allows simultaneous analysis and visualization of more

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

than one query (i.e., several lists of genes may be analyzed and visualized simultaneously), as well as permits user to provide own "universe set", i.e. gene list to test queries.

3. Results

3.1 Bioinformatics Pipeline & Quality control

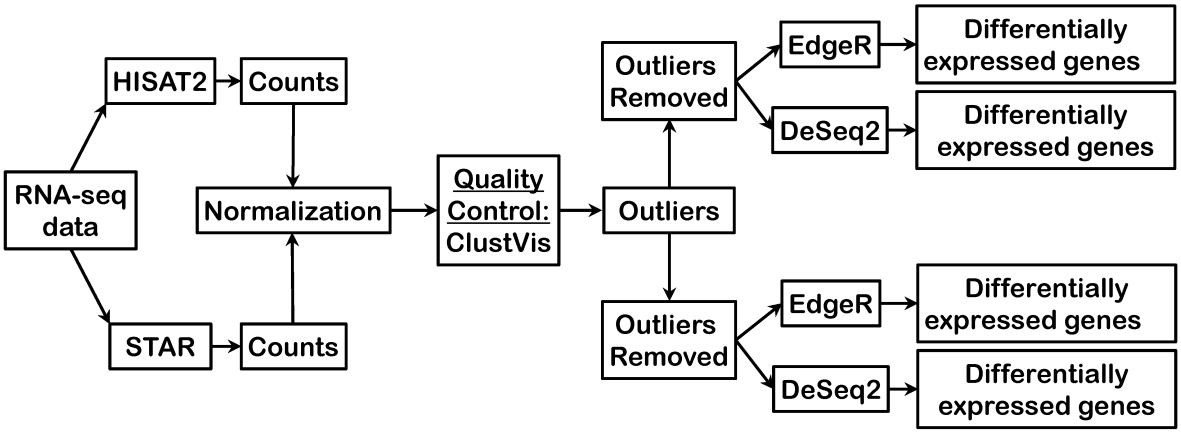

Figure 1. Bioinformatics pipeline employed in the current study.

RNA-seq data were analyzed using two different read aligners, STAR and HISAT2, to compare potential impacts of read mapping to the genome assembly upon the ultimate outputs of differential gene expression. Briefly, RNA-seq output reads were aligned to the genome assembly (hg19) by either HISAT2 or STAR, reads mapped to genes were counted for each individual gene to yield raw counts, subsequently count data were normalized, assessed for quality control using Principal Component Analysis using ClustVis, and sample outliers removed before performing differential gene expression analysis by two different programs, edgeR and DESeq2, to identify significant gene expression changes across breast cancer stages (Figure 1).

To eliminate sample outlier biases, we performed Principal Component Analysis (PCA) of gene expression counts for each sample by each stage for both aligners. Gene expression counts were collected using featureCounts and normalized to the total number of aligned reads for each sample, and PCA completed on these data using ClustVis large edition (Figure 2). For all subsequent analysis, any samples that fell outside the 95% confidence ellipse in their respective stages (Normal, Atypia, DCIS and IDC) were removed. For both HISAT2 and STAR, the same samples fell outside of the 95% confidence ellipse in each stage. In total, we identified six outlier samples in the RNA-seq dataset, which were: SRX286949 (normal tissue), SRX286945 and SRX286964 (atypia), SRX286961 (DCIS) and SRX286951 (IDC). Overall, Atypia stage presented more heterogeneity than any other stage, irrespective of the aligner. Unexpectedly, the PCA plots for all stages in HISAT2 data clustered atypia stage samples separately from all other stages (Figure 2A, 2B).

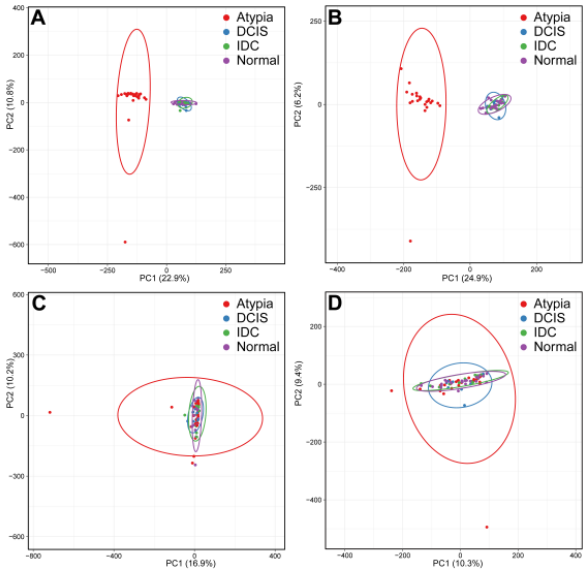

Figure 2. PCA visualization of gene expression data from HISAT2 and STAR alignments. **A, B**: Clustering of HISAT2 samples on the first two principal components before (A) and after (B) outlier removal. **C, D**: Clustering of STAR data before (C) and after (D) outlier removal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

3.2. Outputs of aligners.

All reads for all samples were aligned to the human genome assembly (hg19). Overall, STAR significantly outperformed HISAT2 in aligning the FASTQ reads to the genome (Figure 3). The generally low proportion of aligned reads to all input reads for both programs is likely due to the quality of the libraries, as a significant number of input reads were poly(A) sequences, Illumina adapter sequences, and reads corresponding to the very 3'-ends of mRNAs, which are too uninformative for correct mapping (Supplemental Table 1).

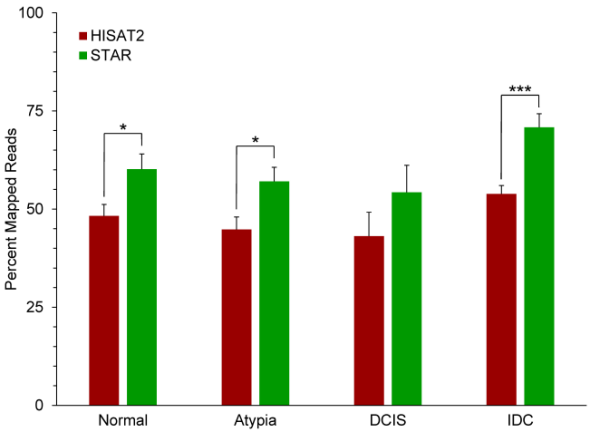

Figure 3. Performance of HISAT2 and STAR aligners on the breast cancer series data.

3.3. Gene expression profiling.

3.3.1. Highly expressed genes

To determine how concordant the alignment tools were in mapping the reads to the genome, we compared the highest expressed genes that correspond to 50% of all reads mapped to exons. In the normal samples 50% of the mapped reads came from 330 and 305 genes for STAR and HISAT2 respectively and they shared 263 of those genes (Figure 4). In atypia samples, 50% of the mapped

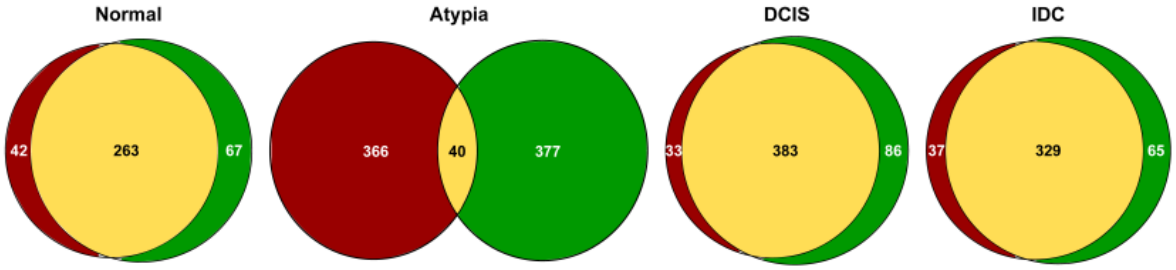

Figure 4. Overlap between the highest expressed genes in the breast cancer datasets aligned by HISAT2 or STAR. HISAT2-identified genes are in red; STAR genes are in green; overlapping genes are in yellow.

reads came from 417 and 406 genes for STAR and HISAT2 respectively and they shared only 40 of those genes. In DCIS samples, 50% of the exon-mapped reads came from 469 and 416 genes for STAR and HISAT2, respectively; of those, 383 genes were shared. In IDC samples, 50% of the exon-mapped reads came from 384 and 366 genes for STAR and HISAT2, respectively, and the lists shared 319 of those genes. The high amount of discrepancies in atypia convinced us to look further into what were the major differences in alignment.

3.3.2. Alignment to pseudogenes

Retrogenes are intronless gene copies produced by reverse transcription of an original "parent" gene mRNA and insertion of the resulting cDNA copy elsewhere in the genome. Retrogenes are often non-functional and are generally assigned to the category of "pseudogenes", i.e. genomic loci harboring similarity to a protein-coding gene but not having any recognized biological function. The sequence similarity among retrogenes and their parent genes poses a problem for aligners, whose algorithms have to decide when assigning a read to a particular locus in the genome. To determine what the differences in alignments were, we analyzed the numbers of reads mapped to pseudogenes by HISAT2, and STAR. Between two aligners tested, HISAT2 consistently had significantly higher amounts of reads aligned to pseudogenes when compared to STAR (Figure 5A). Furthermore, for Atypia stage, HISAT2 had drastically higher amounts of reads aligned to pseudogenes than the other stages. To determine what portion of the top 50% of mapped reads were pseudogenes, we obtained a list of pseudogenes from the hg19 gtf annotation file we used and compared this list with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

the top 50% of mapped genes for each stage and each aligner. A single pseudogene was in the gene list for each stage which represented the top 50% of mapped reads for STAR. Conversely, HISAT2 consistently had higher amounts of pseudogenes represented in the top 50% of mapped reads (Figure 5B).

3.3. Differential gene expression analysis

The differential expression comparison on data from different alignment tools was done to further explore the consequences of previously described alignment tool biases, as well as to compare the two popular tools used for the purpose of identification and quantification of gene expression differences between conditions, edgeR and DESeq2.

3.3.1. edgeR following HISAT2 or STAR

Overall patterns of differential gene expression performed by edgeR on HISAT2 or STAR data were similar for all pairwise stage comparisons, except atypia *vs* any other stage (Figure 6). HISAT2 consistently had the atypia stage comparisons produce > 15,000 statistically significant differentially expressed transcripts with a $\log_2$ fold change (LFC) ≥ 1 (i.e., at least two-fold difference in expression between conditions) (Figure 6, top panels). Conversely, differential expression pairwise comparisons with STAR atypia stage vs each other stage identified 350 to 2,496 differentially expressed genes (Figure 6, bottom panels).

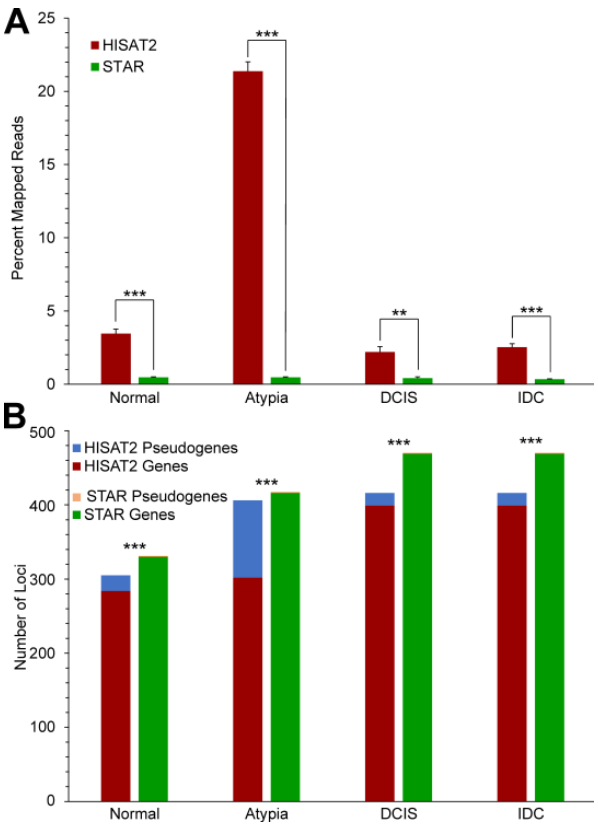

Figure 5. Expression of retrogenes inHISAT2 and STAR alignment data. **A**: Percent of all reads, by stage, aligned to annotated pseudogene loci by HISAT2 (red) and STAR (green). **B**: Number of retrogenes among highest-expressed genes by stage, and aligner.

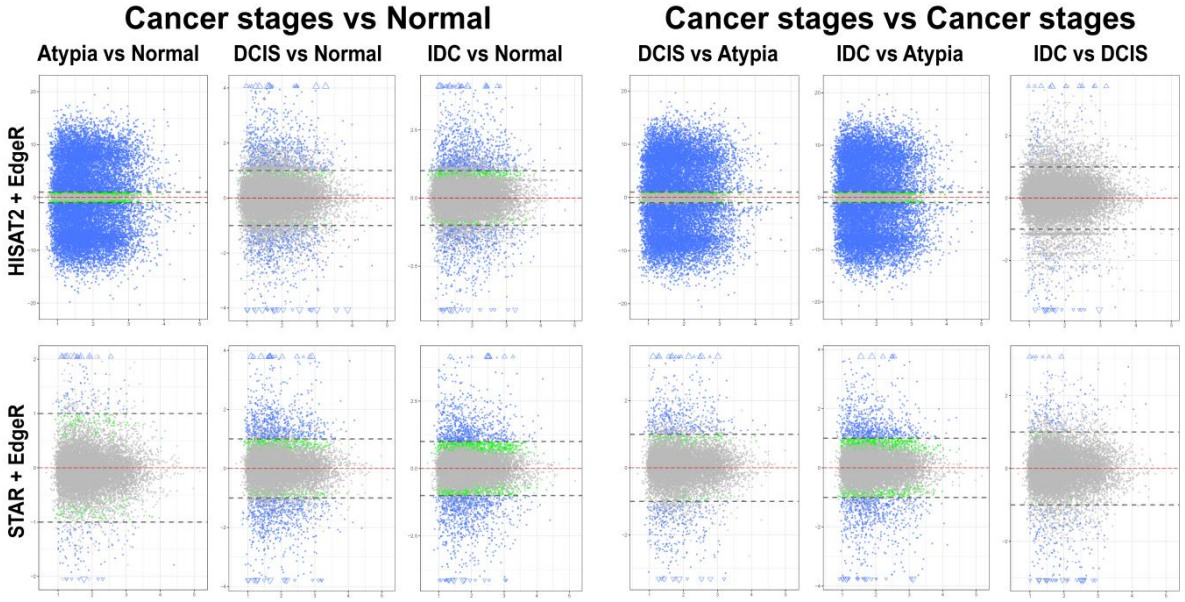

Figure 6. MA plots of pairwise comparisons of all stages using edgeR. HISAT2 (top) and STAR (bottom) gene counts for all samples were analyzed to identify differentially expressed genes. Each gene is represented by a single dot. Blue dots represent genes whose expression difference is both significant and at least two-fold. Green dots represent genes whose expression difference is significant, but less than two-fold; grey dots represent genes whose expression difference is not statistically significant. Y-axis (all plots): $\log_2$ of expression fold change; X-axis (all plots): log of gene expression mean value.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

3.3.2. DESeq2 following HISAT2 and STAR

Differential expression analysis using DESeq2 on pairwise comparisons of STAR alignments revealed 255 transcripts having LFC > 1 and 177 transcripts with LFC < 1 in normal vs atypia comparisons. Normal vs DCIS and Normal vs IDC analysis revealed 1,677 LFC > 1, 482 LFC < 1, and

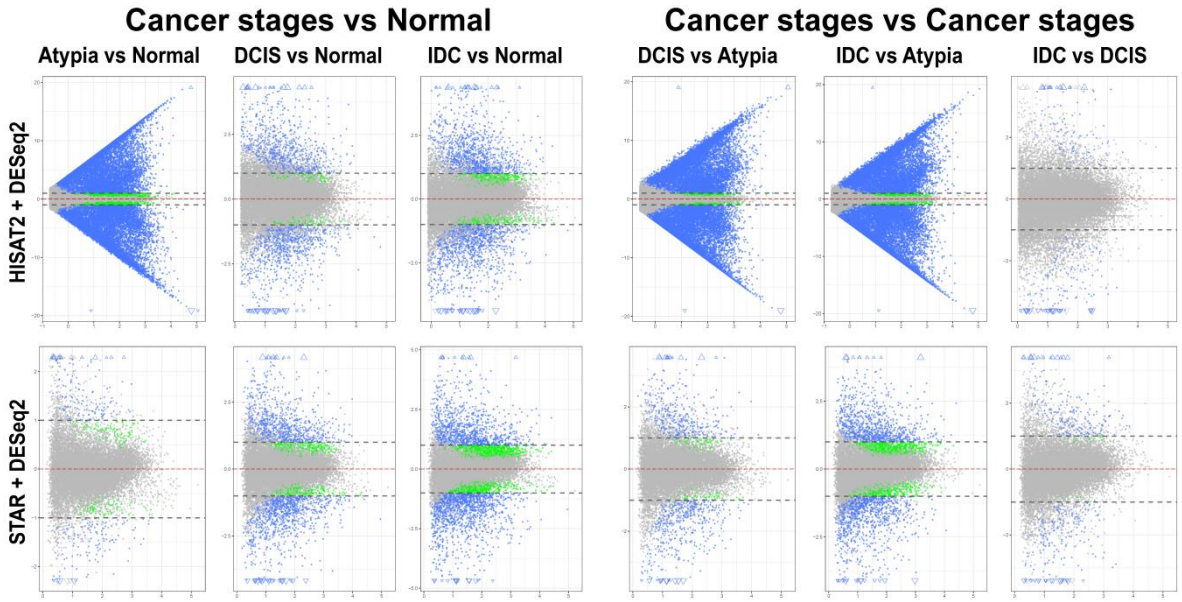

Figure 7. MA plots of pairwise comparisons of all stages using DeSeq2. HISAT2 (top) and STAR (bottom) gene counts for all samples were analyzed to identify differentially expressed genes. Color scheme, axis labels same as Figure 6.

2,304 LFC > 1, 1,417 LFC <1 DE transcripts, respectfully (Figure 7, bottom panels). Similarly to edgeR, differential expression analysis of HISAT2 with DESeq2 consistently produced high numbers of statistically significant differentially expressed transcripts in atypia pairwise comparisons, > 19,000 transcripts with LFC > 1 (Figure 7, top panels). Normal vs Atypia, Normal vs DCIS, and Normal vs IDC analysis revealed 19,419 LFC > 1, 1,212 LFC < 1, 1,585 LFC > 1, 190 LFC < 1, and 2,196 LFC > 1, 732 LFC < 1 DE transcripts, respectfully.

3.3.3. Comparison of DESeq2 and EdgeR results

The total number of statistically significant differentially expressed genes in pairwise comparisons, by aligner and DE tool, is summarized in Figure 8. Overall, DESeq2 produced more inflated lists comparing to edgeR. Next, to compare how similar these most common differential gene expression analysis tools, edgeR and DESeq2, calculate expression differences on the same data, we compared lists of all differentially expressed genes generated by these programs and produced Venn diagrams, for each pairwise cancer stage comparison to normal samples. DESeq2 and EdgeR shared 14,220, 1,433, and 2,137 of the differentially expressed genes on the HISAT2 alignment pairwise comparisons for Normal vs Atypia, Normal vs DCIS, and Normal vs IDC,

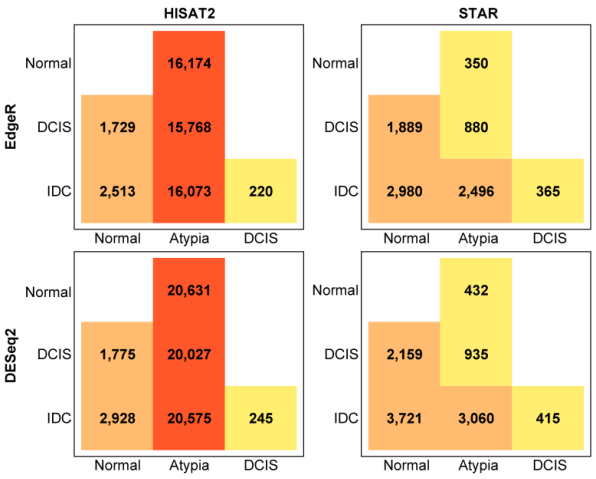

Figure 8. Total number of differentially expressed genes in pairwise comparisons, by aligner (HISAT2 or STAR), and quantification program (edgeR or DESeq2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

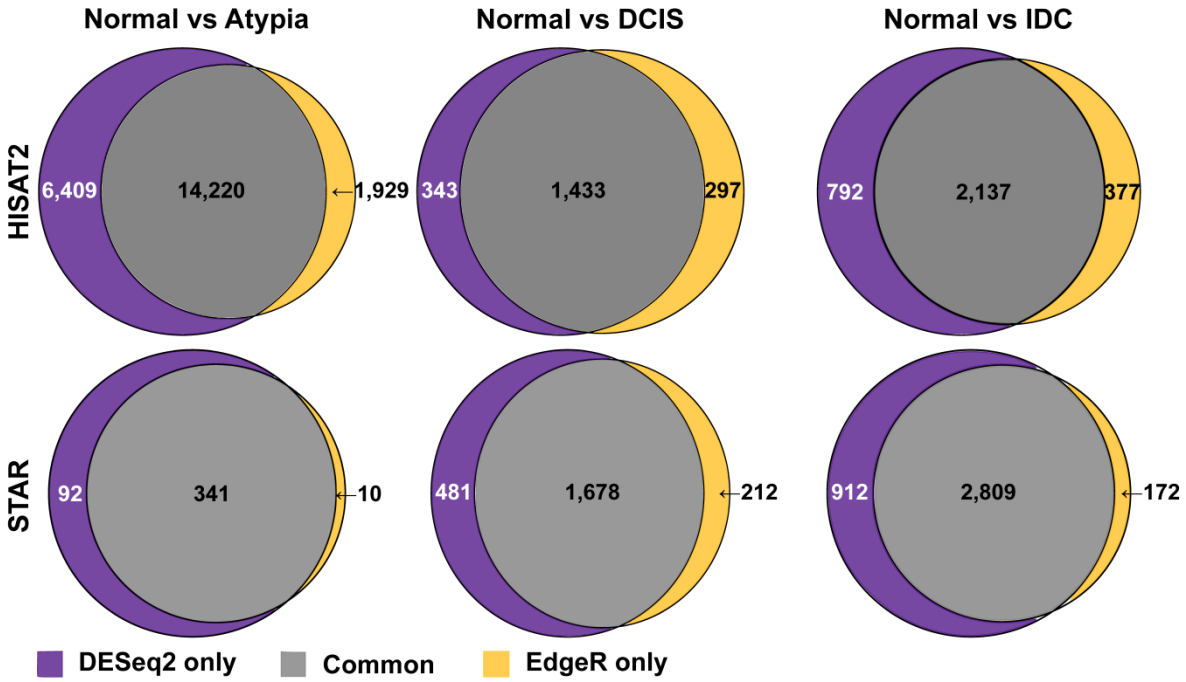

Figure 9. Overlap among genes identified as differentially expressed by either DESeq2 or edgeR in HISAT2 or STAR-aligned RNA-seq data.

respectfully (Figure 9, top row). DESeq2 and edgeR shared 341, 1,678, and 2,809 of the differentially expressed transcripts on the STAR alignment pairwise comparisons for Normal vs Atypia, Normal vs DCIS, and Normal I------DC, respectfully (Figure 9, bottom row).

3.3.4. Downstream analysis of gene expression

VLAD results: To test if downstream analysis of differentially expressed genes is affected by the type of software used to identify these genes, we performed Gene Ontology enrichment studies with the Visual Annotation Display tool (VLAD). Here we corroborate the data of the previously published work with this clinically relevant data. The pathways represented for the GO term molecular function, with data from significantly decreased expression when compared to normal samples, show phosphatidylinositol kinase pathways (Figure 10A). Furthermore, we provide evidence of developmental pathways being decreased in all stages when compared to normal samples (Figure 10B). Significant decrease of extracellular matrix genes, such as basement membrane, was found in cancer stages comparing to normal (Figure 10C). Overall, there was excellent concordance among overrepresented Gene Ontology terms identified among either DESeq2 or EdgeR-identified genes significantly downregulated in cancer (Supplemental Table 2). Similar concordance in significantly enriched GO terms between DESeq2 or edgeR-identified genes (not shown) was observed for genes upregulated in cancer (Supplemental Table 3).

4. Discussion

Personalized medicine is a data science approach that promises focused treatments based on an individual's sequenced genome to precisely target pathways underlying disease or its symptoms. To date only a few genotypes are robustly identified in breast cancer [30], such as *BRCA1* and *BRCA2*, which is inadequate to achieve the full promise of personalized medicine. A promising alternative to this genetic approach to personalized medicine is transcriptome-based identification of altered pathways of the diseased tissue to target its underlying disrupted cellular and molecular machinery. Identification of affected pathways using bioinformatics pipelines for RNA-seq data empowers clinicians to make a focused and informed decision on specific altered genes and pathways as potential therapeutic targets in precision medicine for individual patients. The goal of our studies was to unveil potential hidden biases arising from different bioinformatics analysis pipelines, and to find practical solutions to circumvent and mitigate these biases that impact the downstream

analyses in transcriptomics experiments. Our results indicate that STAR and edgeR are robust and well-suited bioinformatics tools for transcriptomics pipelines of the most commonly used clinical specimens, biopsies from FFPE tissues, to precisely and accurately align and detect differential gene expression.

After initial quality control checks on the raw output from the sequencer, alignment is the first step in RNA-seq analysis and all subsequent analysis relies profoundly upon this initial step [31]. Typically, reads obtained from sequencing will be mapped and aligned to a reference genome, which is particularly prone to errors for RNA-seq data because of the presence of output reads spanning exon-exon splice junctions. The most common software platforms available for mapping to a reference genome, TopHat [32], HISAT2 [12], and STAR [13], identify splice junctions. These platforms differ in computational speed and memory usage, and in their algorithms for handling base and splice junction alignment precision. TopHat is currently becoming obsolete and has been superseded by HISAT2 due to relative computational inefficiency. TopHat and HISAT2 are built on the short read mapping program Bowtie2 [20]. While all three aligners are considered fast, HISAT2 and STAR consistently outperform TopHat with respect to computational speed [13,33,34]. Although all three aligners performed well in aligning a read onto the respective genomic locus, notable discrepancies and deficiencies were found for TopHat yielding insufficient genomic mapping for reliable downstream analysis. Given these reasons, and the fact that TopHat performance has been evaluated previous [34,35], our studies assessed the alignment performance between STAR and HISAT2.

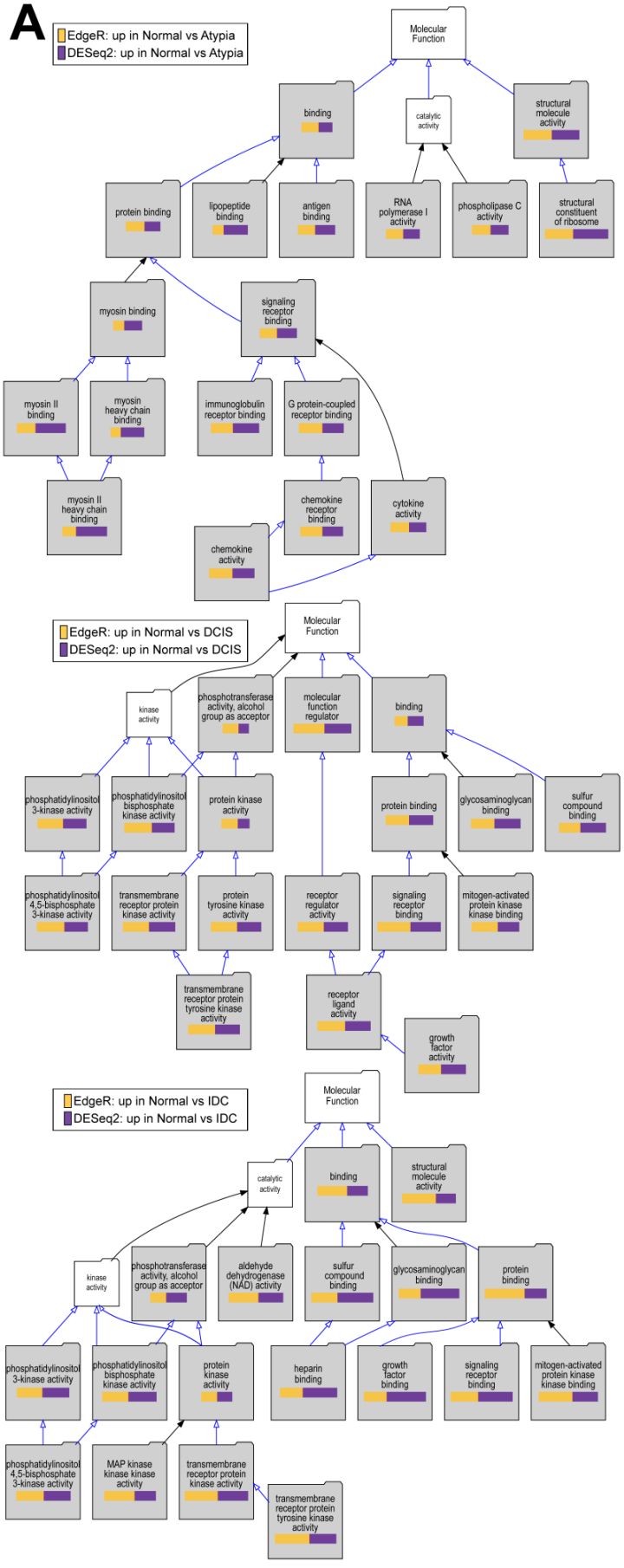

Figure 10. Overrepresented Gene Ontology terms among genes upregulated in normal samples comparing to cancer stages. **A**: Molecular Function, **B**: Biological Process, and **C**: Cellular Component GO categories; Top to bottom: Atypia vs Normal, DCIS vs Normal and IDC vs normal comparisons. Significant GO terms are shaded; Size of the bar represents significance of p-value, and p-value ratios between edgeR and Deseq2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

The next step in bioinformatics pipelines, quantification of gene expression, is commonly performed with DESeq2 or EdgeR programs [14,15]. A common drawback during this step is that difference in sequencing depth between samples or groups by itself can skew results when estimating differences in gene expression levels. The relative expression level of genes is often estimated based on the number of mapped reads. These counts are subjected to statistical tools to assess significant differences between groups. However, there is much confusion in the literature when reporting relative expression level units in RNA-seq data. The confusion stems from the different forms of normalization required for within vs. between sample comparisons. Many methods for within sample comparison attempt to correct for sequencing depth and gene length. These methods produce the most frequently reported unit of expressions for RNA-seq data, which are read per kilobase of exon per million reads (RPKM), fragments per kilobase of exon per million of reads (FPKM), and transcripts per million (TPM) [36]. The order in which RPKM and FPKM normalize the read counts causes differences within samples that should not be ignored. Instead, when comparing within samples one should use TPM values which eliminates the invariance [36]. A relationship among RPKM, FPKM and TPM is further discussed elsewhere [37]. In this study, to account for the impact of sequencing depth, we normalized expression data to counts per million (CPM) [23], which equal TPM values in single-end sequencing RNA-seq datasets, to perform quality comparison between the sample datasets. This normalization has a number of advantages for FFPE samples, where RNA quality is low, captured cDNA is not sheared, and there is no need to control for transcript length. Additional advantage of normalization is conversion

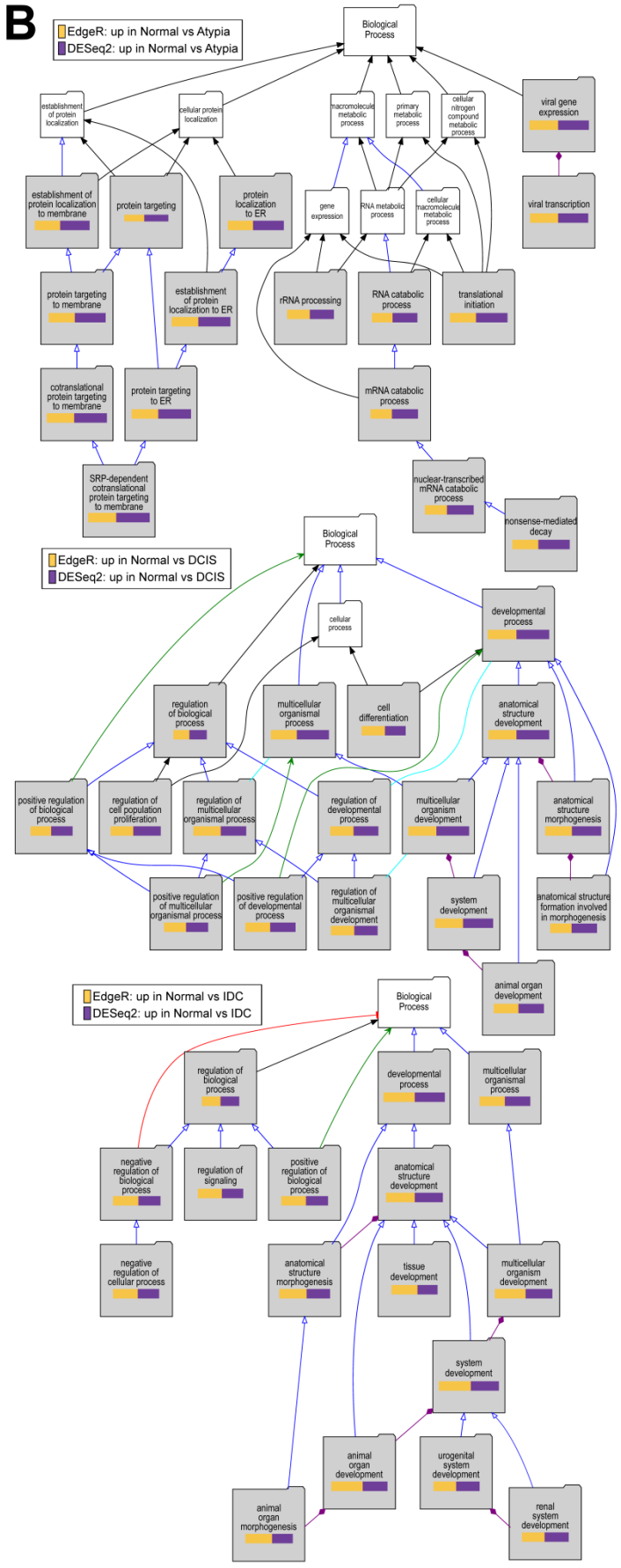

Figure 10. *Continued.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0036.v1

Peer-reviewed version available at J. Pers. Med. 2019, 9, 18; doi:10.3390/jpm9020018

of gene expression data from counts to a continuous scale.

In our study, STAR had the highest average rates of mapped reads for each stage, whereas HISAT2 aligned fewer reads and importantly, had increased rates of alignment to pseudogenes instead of genes, which clearly compromised alignment fidelity, leading to skewed gene expression counts and likely erroneous outputs with either edgeR or DESseq2 programs, albeit the latter produced more expanded lists of differentially expressed genes. For these FFPE specimens, STAR alignment obtained more precise and accurate results than HISAT2 with the fewest misalignments to pseudogenes. Our results clearly demonstrate alignment impacts and play a large role in bioinformatics analysis outcomes, especially to detect and identify differential gene expression. Our data, together with other comparative tests of different programs [34,35]; indicate a single aligner program cannot be applied universally to RNA-seq datasets. It is possible that short output nucleotide sequences may have contributed to the FM index generation utilized by HISAT2 with the annotated settings, propagating misalignments to pseudogenes. While notable improvements in sequencing technology have increased nucleotide output read length to greater than 300 nucleotides, which may increase alignment accuracy and mitigate misalignment to pseudogenes, Chhanawala and co-authors [38] reported that output reads > 25 nucleotides had negligible impacts in detecting differential expression.

A recent study [39] compared the performance and accuracy of the most commonly used differential expression tools available for RNA-seq analysis and discovered DESeq2 and edgeR outperformed all other tools with the lowest false discovery rate (FDR) and highest true discovery rate (TDR). It appeared DESeq2 slightly outperformed edgeR with respect to FDR in datasets with large number

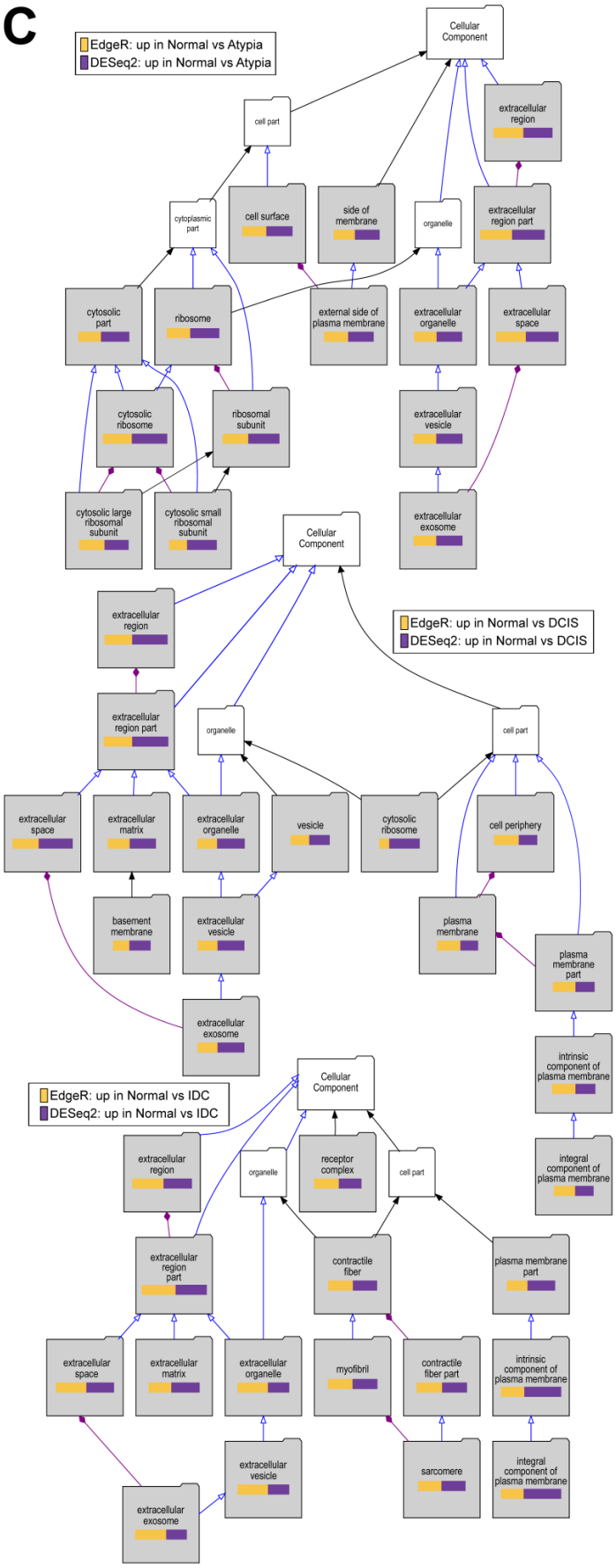

Figure 10. *Continued.*

(>12) of biological replicate samples, however we observed DESeq2 over-predicted differentially expressed genes. This differs from earlier reports of edgeR's propensity to a higher FDR at higher number of biological replicates. The addition of the estimateDisp function may have applied heavier weighted likelihood empirical Bayes methods to obtain the posterior dispersion estimates [40]. Therefore, in studies with a large cohort of replicates, both DESeq2, and edgeR are recommended to be used with the estimateDisp function to control FDR. More commonly for biomedical research and personalized medicine, i.e. in studies with fewer than 12 replicates, edgeR has advantage in reducing false negative rates (FNR).

Our study clearly demonstrates the need for heedful intent and meticulous review executing a bioinformatics pipeline to assess differential expression of RNA-seq data for clinical diagnoses, prognosis, and choice of treatment. Bioinformaticians need to be aware of the biological and clinical impacts, and limitations, of the specimen collection and preservation techniques for samples used for RNA-seq when creating custom transcriptomic pipelines, especially employing an SOP "standard 'omics pipeline" in a genomics core setting. Due to the increased use of RNA-seq to diagnose, prognose, and generate therapeutic options clinicians and biomedical researchers need to more closely obtain a stronger foundation and understanding of the bioinformatics tools and pipelines analyzing RNA-seq data and generating the results. This study highlights possible limitations of this version of HISAT2 for some RNA-seq read generation technologies, poor quality samples, and short RNA seq reads, thus providing clinicians with insights for choosing the right bioinformatics tools for the job.

Application of transcriptomics can facilitate the exploration of underlying pathogenic mechanisms, identification of genetic variants, determination of treatment effects, including screening for molecular biomarkers. Importantly, expression signatures in diseased phenotypes may pinpoint precise interventions required to alleviate the disease state, a goal of precision medicine, without a need for the cost-prohibitive "personalized" assembly and analysis of patient's own genome. Thus, transcriptomics can classify individuals, while simultaneously facilitating discovery, testing, and validation of new therapeutics for breast cancer patients, defined at the cellular and molecular levels.

5. Conclusions.

Transcriptomics is an effective tool for both diagnostics and discovery science, exposing novel cellular and molecular mechanisms in clinical and translational models to yield robust targets for drug discovery to identify and test novel therapeutics. The cost and time required for transcriptome analysis have been greatly reduced by the development of next generation sequencing. Our results underscores the essential importance to investigate the impacts of bioinformatics tools for sequence alignment and differential expression on accuracy of results and interpretation of transcriptome studies collected from FFPE specimens, which was the goal of our study. This paper also serves to point out the pragmatic shortcomings of universally applying a rigid "standard operating procedure" set of bioinformatics tools to RNA-seq data, by demonstrating the impacts and limitations of biological conditions, especially in sample processing and choice of programs.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1: Overrepresented reads in RNA-seq datasets used in this study, www.mdpi.com/xxx/s2 Table S2: Statistically significant genes upregulated in normal samples comparing to cancer stages, www.mdpi.com/xxx/s3 Table S3: Statistically significant genes upregulated in cancer stages comparing to normal samples.

Author Contributions: Conceptualization, A.V.E. and C.M.d.E.; methodology, I.D.R., A.V.E. and C.M.d.E.; validation, I.D.R., and A.V.E.; formal analysis, I.D.R., A.V.E. and C.M.d.E.; writing — original draft preparation, I.D.R.; writing — review and editing, I.D.R., A.V.E. and C.M.d.E.; writing – final draft C.M.d.E.; visualization, I.D.R., A.V.E. and C.M.d.E.; funding acquisition, C.M.d.E., A.V.E.

Funding: This research was funded by Impact Assets (Fund for Science grant to A.V.E. and C.M.d.E.), and the USF Graduate Student Success Fellowship (I.D.R.).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hawkins, R.D.; Hon, G.C.; Ren, B. Next-generation genomics: An integrative approach. *Nature Reviews Genetics* **2010**, *11*, 476.
2. Senkus, E.; Kyriakides, S.; Ohno, S.; Penault-Llorca, F.; Poortmans, P.; Rutgers, E.; Zackrisson, S.; Cardoso, F. Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **2015**, *26*, v8-v30.
3. Coates, A.S.; Winer, E.P.; Goldhirsch, A.; Gelber, R.D.; Gnant, M.; Piccart-Gebhart, M.; Thürlimann, B.; Senn, H.-J.; Members, P.; André, F. Tailoring therapies — improving the management of early breast cancer: St Gallen international expert consensus on the primary therapy of early breast cancer 2015. *Annals of Oncology* **2015**, *26*, 1533-1546.
4. Byron, S.A.; Van Keuren-Jensen, K.R.; Engelthaler, D.M.; Carpten, J.D.; Craig, D.W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics* **2016**, *17*, 257.
5. Marioni, J.C.; Mason, C.E.; Mane, S.M.; Stephens, M.; Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **2008**.
6. Zhang, W.; Yu, Y.; Hertwig, F.; Thierry-Mieg, J.; Zhang, W.; Thierry-Mieg, D.; Wang, J.; Furlanello, C.; Devanarayan, V.; Cheng, J. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome biology* **2015**, *16*, 133.
7. NCBI. SRA Database Growth. https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/ (accessed on 22 February 2019)
8. Ben-Ezra, J.; Johnson, D.A.; Rossi, J.; Cook, N.; Wu, A. Effect of fixation on the amplification of nucleic acids from paraffin-embedded material by the polymerase chain reaction. *Journal of Histochemistry & Cytochemistry* **1991**, *39*, 351-354.
9. Masuda, N.; Ohnishi, T.; Kawamoto, S.; Monden, M.; Okubo, K. Analysis of chemical modification of rna from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Research* **1999**, *27*, 4436-4443.
10. Srinivasan, M.; Sedmak, D.; Jewell, S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *The American Journal of Pathology* **2002**, *161*, 1961-1971.
11. Buckingham, L. *Molecular diagnostics: Fundamentals, methods and clinical applications*. FA Davis: 2011.
12. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature Protocols* **2016**, *11*, 1650-1667.
13. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. Star: Ultrafast universal rna-seq aligner. *Bioinformatics* **2013**, *29*, 15-21.
14. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol* **2014**, *15*, 550.
15. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139-140.
16. Google Scholar. https://scholar.google.com/ (accessed on 22 February 2019)
17. Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy, T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **2010**, *11*, R86.
18. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T., *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **2000**, *25*, 25.

19. Brunner, A.L.; Li, J.; Guo, X.; Sweeney, R.T.; Varma, S.; Zhu, S.X.; Li, R.; Tibshirani, R.; West, R.B. A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions. *Genome Biol* **2014**, *15*, R71.
20. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012**, *9*, 357.
21. Simpson, J.T.; Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **2010**, *26*, i367-i373.
22. Liao, Y.; Smyth, G.K.; Shi, W. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923-930.
23. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **2008**, *5*, 621.
24. Metsalu, T.; Vilo, J. ClustVis: A web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic acids research* **2015**, *43*, W566-570.
25. McDermaid, A.; Monier, B.; Zhao, J.; Ma, Q. VidgeR: An R package for integrative interpretation of differential gene expression results of RNA-seq data. *bioRxiv* **2018**.
26. McCullagh, P.; Nelder, J.A. *Generalized linear models*. CRC press: 1989; Vol. 37.
27. Richardson, J.E.; Bult, C.J. Visual annotation display (VLAD): A tool for finding functional themes in lists of genes. *Mammalian Genome* **2015**, *26*, 567-573.
28. Smith, C.L.; Goldsmith, C.-A.W.; Eppig, J.T. The mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology* **2004**, *6*, R7.
29. Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L.J.; Eilbeck, K.; Ireland, A.; Mungall, C.J., *et al.* The obo foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **2007**, *25*, 1251.
30. PDQ Cancer Genetics Editorial Board. Genetics of breast and gynecologic cancers (PDQ®). In *PDQ cancer information summaries [internet]*, National Cancer Institute (US): 2018.
31. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X., *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **2016**, *17*, 13.
32. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with Tophat and Cufflinks. *Nature Protocols* **2012**, *7*, 562-578.
33. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* **2015**, *12*, 357.
34. Baruzzo, G.; Hayer, K.E.; Kim, E.J.; Di Camillo, B.; FitzGerald, G.A.; Grant, G.R. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* **2016**, *14*, 135.
35. Engström, P.G.; Steijger, T.; Sipos, B.; Grant, G.R.; Kahles, A.; The, R.C.; Alioto, T.; Behr, J.; Bertone, P.; Bohnert, R., *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **2013**, *10*, 1185.
36. Li, B.; Dewey, C.N. Rsem: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **2011**, *12*, 323.
37. Pachter, L. Models for transcript quantification from RNA-seq. *arXiv preprint arXiv:1104.3889* **2011**.
38. Chhangawala, S.; Rudy, G.; Mason, C.E.; Rosenfeld, J.A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology* **2015**, *16*, 131.

39. Schurch, N.J.; Schofield, P.; Gierliński, M.; Cole, C.; Sherstnev, A.; Singh, V.; Wrobel, N.; Gharbi, K.; Simpson, G.G.; Owen-Hughes, T., *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **2016**, *22*, 839-851.
40. Chen, Y.; Lun, A.T.; Smyth, G.K. Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequencing Data*, Springer: 2014; pp 51-74.