IMPACT: A NOVEL ALGORITHM FOR GENE EXPRESSION QUANTIFICATION IN TAGSEQ ANALYSES

By

Bradley Nagel Jenner

Signature work, Thesis in Major, submitted for completion of the University Honors Program and Bachelor of Science in Biotechnology

December 3, 2021

University Honors Program
University of California, Davis

APPROVED

Daniel Runcie, Ph.D
Department of Plant Sciences

Lolita Adkins, Interim Associate Director of the University Honors Program
Undergraduate Education

Table of Contents

Acknowledgements

I would like to sincerely thank my thesis mentors Dr. Matthew Settles, Dr. Samuel Hunter, and Dr. Jie Li for their guidance during the course of my Undergraduate Honors Thesis. Their years of experience in the field of Bioinformatics was essential to the projects direction and completion. Gratitude is also owed to Dr. Blythe Durbin-Johnson for contributing her statistical expertise by advising on data analysis methods. I would also like to thank my faculty advisor, Dr. Daniel Runcie, for his editorial contributions and advisorship as well as the UC Davis Bioinformatics Core and High Performance Computing Core facilities for their educational and technical support. Lastly, I would also like to thank UC Davis principal investigators Dr. Arta Monjazeb, Dr. Alison Berry, and Dr. Smita Iyer for contributing TAGseq datasets.

Abstract

TAGseq is a powerful tool for profiling the transcriptome of an organism. It serves as a cheaper alternative to traditional RNAseq that is less susceptible to transcript length bias and equally effective at capturing gene expression. Unfortunately, current analysis methods for TAGseq often have difficulty quantifying expression for genes with poorly annotated untranslated regions. Our algorithm, impact, attempts to alleviate this issue by quantifying gene expression using read clusters— or groups of overlapping reads— that have a higher probability of overlapping with gene annotations. We tested the efficacy of this new algorithm using real TAGseq datasets in mice, rhesus macaques, and alfalfa. Our results demonstrate that this method is effective at increasing the number of reads captured in poorly annotated genomes while preserving gene expression profiles and recreating the results of differential expression analyses. This method also creates future opportunities for transcript level expression, which was previously not possible in 3’ RNA sequencing methods.

Introduction

The transcriptome of an organism is fundamental to its biology. Nuanced regulation of gene expression and transcript isoform abundance govern a variety of cellular processes that affect development, disease, and responses to environmental conditions (Wang *et al.,* 2009). For this reason, profiling the transcriptome has been the primary goal for countless genomics studies, many of which rely on RNA sequencing (RNAseq). While the term RNAseq can refer to a number of experiments that sequence different components of the transcriptome, a majority of RNAseq experiments emphasize quantification of messenger RNA (mRNA) and the identification of differentially expressed genes.

RNA sequencing experiments begin with the extraction of total RNA from a tissue sample. This process can vary greatly based on the organism, experimental design, and tissue type. Once total RNA has been extracted, the sample must be enriched for mRNA, as total RNA is disproportionately high in ribosomal RNA. This is accomplished by ribo-depletion, or, in eukaryotes, mRNA selection via hybridization to the polyA tail of mRNA transcripts. The remaining mRNA molecules are then fragmented and undergo size selection to remove RNA molecules that are too long or too short for sequencing. Next, the RNA is converted into cDNA via reverse-transcription and Illumina adapters are ligated onto the ends of the reads before being amplified via PCR. The amplified libraries are then ready to be sequenced on Illumina sequencing systems to a depth of 10-30 million reads per sample (Kukurba and Montgomery, 2015).

In order to convert these sequencing reads to gene expression data for differential expression analysis, cleaning the raw sequencing data via a preprocessing pipeline is necessary. This can remove unwanted sequences such as contaminants and PCR duplicates while also

removing sequences that can impact mapping quality and speed, like adapters, low quality sequences, and polyAT sequences. After the raw data has been sufficiently preprocessed, the reads are aligned to a reference genome and overlapped with known genes. At this point, gene expression is estimated for each gene by simply counting the number of reads that overlap to a gene. While simple and effective, this method of quantification with traditional RNAseq is susceptible to biases like transcript length bias, where longer transcripts have more reads and are thus over-represented in the detection of differentially expressed genes (Oshlack and Wakefield, 2009; Moll *et al.,* 2014). There are many complicated statistical methods that attempt to correct this bias, yet this issue can be overcome more easily and inexpensively by using a different library preparation technique.

3' PolyA sequencing methods— more specifically, TAGseq and Lexogen's QuantSeq 3' mRNA-Seq Kit— are methods of library preparation where only a single fragment from the 3' untranslated region (UTR) of an mRNA molecule is sequenced. This is accomplished by fragmenting total RNA and capturing only the 3' polyadenylated tail of mRNA transcripts using oligo polyT priming and reverse transcription (Moll *et al.,* 2014). The advantage of TAGseq is in the single fragment per transcript libraries, as it allows for simple quantification of gene expression due to the lack of susceptibility to transcript length bias (Tandonnet and Torres, 2016; Meyer *et al.,* 2011). Additionally, reducing the number of reads per transcript translates to less sequencing necessary for capturing gene expression (Moll *et al.,* 2014; Meyer *et al.,* 2011), which reduces the cost of sequencing compared to traditional RNAseq. However, one drawback of TAGseq is that because it captures the 3'-most region of the transcript, it is required that organisms have a reference genome with sufficiently annotated UTRs. Otherwise, traditional counting methods discard many reads because there is no feature in the annotation to which they

overlap. This results in significant loss of data for experiments using nonmodel organisms. However, the library preparation method in TAGseq allows us to make assumptions about groups of reads that can alleviate this issue. Because reads originate from random priming along the polyA-containing fragment, we expect most reads belonging to the same transcript isoform to map to the same region in the UTR (Moll *et al.,* 2014). These reads should form distinct groups, or clusters, that correspond to a unique polyadenylation sites (Wang *et al.,* 2009) and theoretically, represent all expressed transcripts that are similarly polyadenylated. Therefore, it is reasonable to regard groups of overlapping reads as a single read cluster during gene expression quantification and assign *read clusters* to genes, even if not all reads explicitly overlap with the feature.

Here, we propose an algorithm, impact (identifies multiple peaks and counts transcripts), that attempts to identify groups of overlapping reads and then assign their combined read counts to genes based on their positions within the exons and UTRs of annotated genes (https://github.com/bnjenner/impact). This aims to retain the benefits of the TAGseq protocol while alleviating issues with traditional counting methods that result in high numbers of unassigned reads in organisms with poor genome annotations. We demonstrate that this method preserves patterns of gene expression across an experiment and identifies significantly differentially expressed genes.

Methods

Impact requires only a genome annotation file in GTF format and a sorted and indexed alignment file in BAM format as inputs. The algorithm begins by first parsing the annotation file and creating a doubly linked list of genes. Each gene is represented as a node in the list, which contains the start and stop positions of every exon listed for that gene. Next, the alignment records in the bam file are processed sequentially and overlapped with each other based on their position in the genome. By default, impact only consider uniquely mapped reads with a mapping quality score greater than 0. Overlapping groups of reads are regarded as one cluster and the boundaries of these clusters are recorded. As with the genes from the annotation file, each cluster is represented as a node in a doubly linked list. This data structure was chosen because it is optimized for sequential traversal and insertion/deletion operations take constant time, which is essential for efficiently creating non-redundant sets of read clusters. Since RNAseq aligners are splicing aware, gapped alignments are also considered. In this case, a cluster is divided into smaller subclusters based on these gapped alignments that represent continuous regions of coverage in the genome. Once all reads are grouped together, impact attempts to assign these clusters to genes by overlapping them with the boundaries recorded in the gene annotations. If a read cluster overlaps with a gene, the total number of reads in that cluster are assigned to that gene (Supplementary Figure 1).

However, since impact deals with read clusters instead of individual reads, a different assignment scheme from traditional read counting algorithms was implemented to maximize the number of reads assigned to genes and reduce loss of signal due to read clusters overlapping with multiple genes. Standard read counting algorithms only consider how the alignment overlaps with individual genes. There are many variations on this approach, but many give a gene

assignment preference if the entire read is within the bounds of an exon for that gene. The impact overlapping scheme, however, considers both the number of reads in the cluster or subcluster with which the exon overlaps as well as an overlap score that is determined by how the read cluster overlaps with the exon. If a read cluster is entirely within the bounds of an exon, it is assigned an overlap score of 2; otherwise, it is assigned a 1. A read cluster is ultimately assigned to the gene with which most overlapping reads; however, if the number of reads between two genes is equal, it then considers the overlap score and assigns the read counts to the gene with the higher score. If both the number of reads and overlap score are equal, the read cluster is considered ambiguous and no reads are counted towards either gene. Additionally, read clusters that do not overlap with a gene are considered to have no feature and are disregarded along with multimapping reads and PCR duplicates. The final results is a counts table for each gene listed in the annotation file and, if desired, an annotation file of the identified read clusters in GTF format.

In order to validate our algorithm, we first created a reduced sample dataset from a mouse TAGseq dataset comprised of 15 regions of varying expression levels along the first chromosome of the mouse genome. These regions contained 45 read clusters, 30 of which overlap with genes (assigned), 5 have ambiguous assignment (ambiguous), and 10 do not overlap with genes (unassigned) given our gene assignment criteria. Our first validation test was to prove that read clusters were being properly created from the alignment data. Since the goal of the read clustering algorithm is to create a non-redundant, non-overlapping set of read clusters, the efficacy of our algorithm can be tested by taking the annotation of the read clusters output by impact and then using featureCounts (Liao *et al.,* 2014), a standard tool for read counting, to count the number of reads overlapping with the annotated read clusters. If all reads clusters were correctly assembled, the counts for these clusters should be equal to the total read count reported

by impact and no reads should be marked as ambiguous or unassigned. We also aimed to test the efficacy of the gene assignment algorithm. To do this, we manually viewed each region in our reduced dataset using IGV (Thorvaldsdóttir *et al.*, 2013) and recorded the bounds of any cluster or subcluster that was inside of or spanning these regions. We then recorded their expected gene assignment based on impact's gene assignment algorithm and ran impact on this dataset to validate that the program accurately assigns read clusters to the proper genes.

We also sought to validate the impact algorithm on real-world TAGseq datasets and assess its effect on differential expression analysis. We quantified gene expression and performed differential expression analysis on 3 unpublished datasets in mice (*Mus musculus)*, alfalfa (*Medicago sativa*), and rhesus macaque (*Macaca Mulatta*). The mice dataset came from a study of differentially expressed genes in tumors between mice with different diets and sexes. The analysis was performed on the GRCm39 genome (accession: GCA_000001635.9) with the corresponding gencode annotation and was chosen due to the popularity of mice in genomics research. The alfalfa dataset sought to identify differentially expressed genes in alfalfa roots after interactions with different microorganisms. This analysis was performed on a version of the alfalfa genome and genome annotation generated by Li *et al.* (2020) and this dataset was chosen because of the incomplete annotation of the alfalfa genome that resulted in large numbers of the reads being unassigned to genes. It was also chosen because there was a significant batch effect present in the data that resulted in very few differentially expressed genes, meaning this dataset will serve as our negative control for the impact algorithm. Lastly, we chose a rhesus macaque dataset consisting of differentially expressed genes in 4 brain tissues— abbreviated as S, H, P, and C— in macaques that were infected with SIV. This macaque dataset also showed a high number of reads being marked as unassigned by traditional counting algorithms, making it an

ideal dataset for impact to attempt to recover those reads. This analysis was performed using the Mmul 10 (accession: GCA_003339765.3) genome and annotation from ensembl (version 104).

We began our bioinformatic analyses with cleaning the raw data using a preprocessing pipeline built with HTStream, a preprocessing toolkit (https://s4hts.github.io/HTStream/). This consisted of removing adapter sequences, low quality regions, screening for residual ribosomal RNA (mouse samples only) and PhiX sequences, and removing reads lower than 50 bp in length. The raw reads also contained UMIs, which were excised from the read and added to the read ID using a custom script to later be used with alignment data by umi-tools for PCR deduplication (Smith *et al.,* 2017). After preprocessing, the remaining reads were aligned to their respective genomes using STAR (Dobin *et al.,* 2013) and deduplicated using the umi-tools dedup function. At this point, gene expression was quantified using two programs: featureCounts and our novel impact program.

After gene expression was quantified, differential expression analysis was performed on each dataset using R (R Core Team, 2017) and the limma-voom pipeline (Ritchie *et al.,* 2015; Law *et al.,* 2014). This first consists of calculating normalization factors according to the TMM method and filtering low expressed genes where the max log-normalized read count for the gene across all samples was less than 1. After filtering, the counts were log2 transformed and variance weights for use in the weighted least squares algorithm were obtained using voom. Empirical Bayes smoothing was used to calculate improved estimates of the standard errors of log fold changes, and finally, multiple testing correction was performed using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate. Correlation of the p-values between the two methods was also computed using Spearman's rank correlation coefficient (Spearman, 1987).

Another question of interest we investigated was whether or not unassigned, ambiguous, and assigned read clusters had distinct characteristics. If this is the case, these metrics could potentially be used to identity read cluster arising from expression of novel genes or exons, which is a goal for future iterations of this program. To investigate this possibility, we observed the expression levels and width in basepairs of the read clusters identified in the mice dataset, as its genome annotation is the most complete.

Results

Testing on our reduced validation dataset allowed us to assess the behavior of both our read clustering and gene assignment algorithms. The results of our read clustering test, which used featureCounts to overlap reads with the read cluster annotation created by impact, showed all reads being assigned to read clusters with the read counts reported by featureCounts matching those reported by impact. Read clusters were also confirmed visually by inspecting each of the 15 regions in our reduced dataset using IGV. This revealed that impact was successful at establishing the bounds of read clusters in regions with few distinct read clusters and low gene expression (Figure 1) as well as more complicated regions with many different read clusters of differing expression levels (Figure 2). As for assessing gene assignment, we expected 30 read clusters to be assigned to genes, 5 to have ambiguous assignment, and 10 to have no overlapping gene. The gene assignment reported by impact showed assignment of all the read clusters that matched our expectations based on impact's gene assignment criteria.

Figure 1: Low Complexity Read Clusters

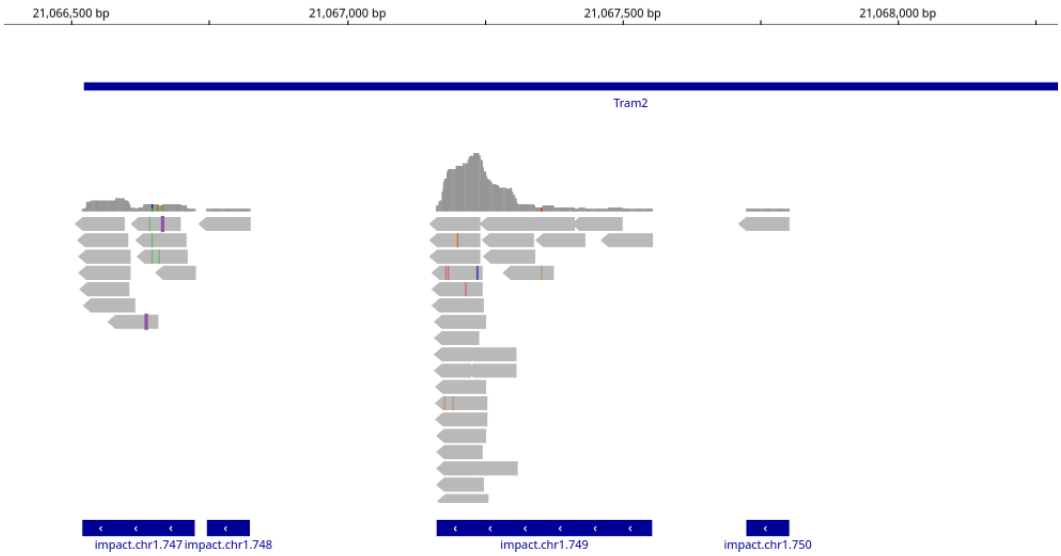
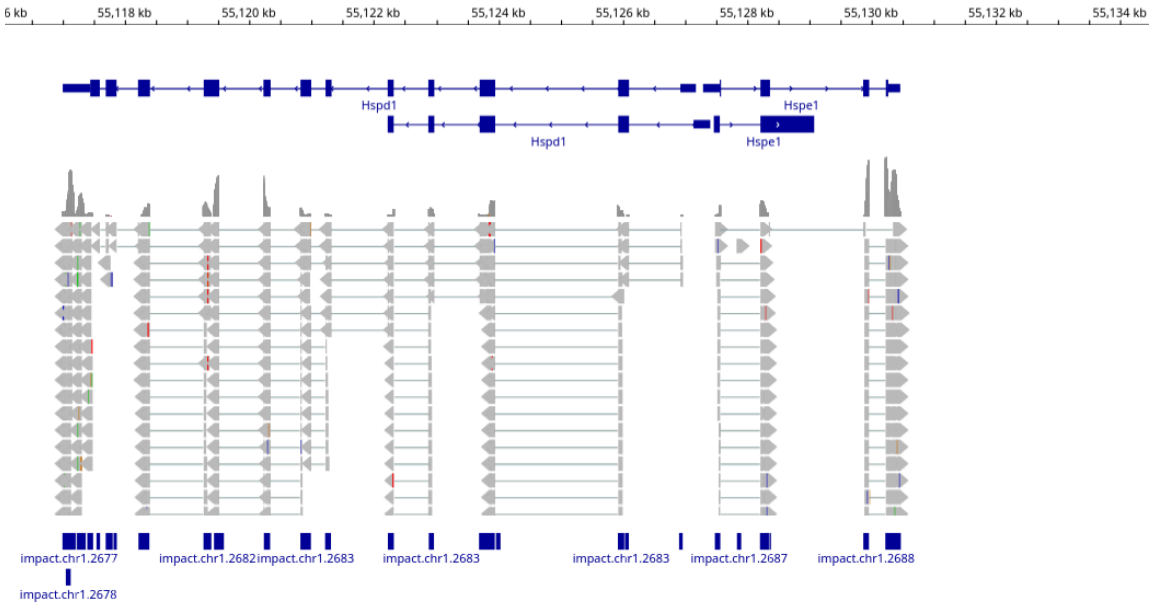
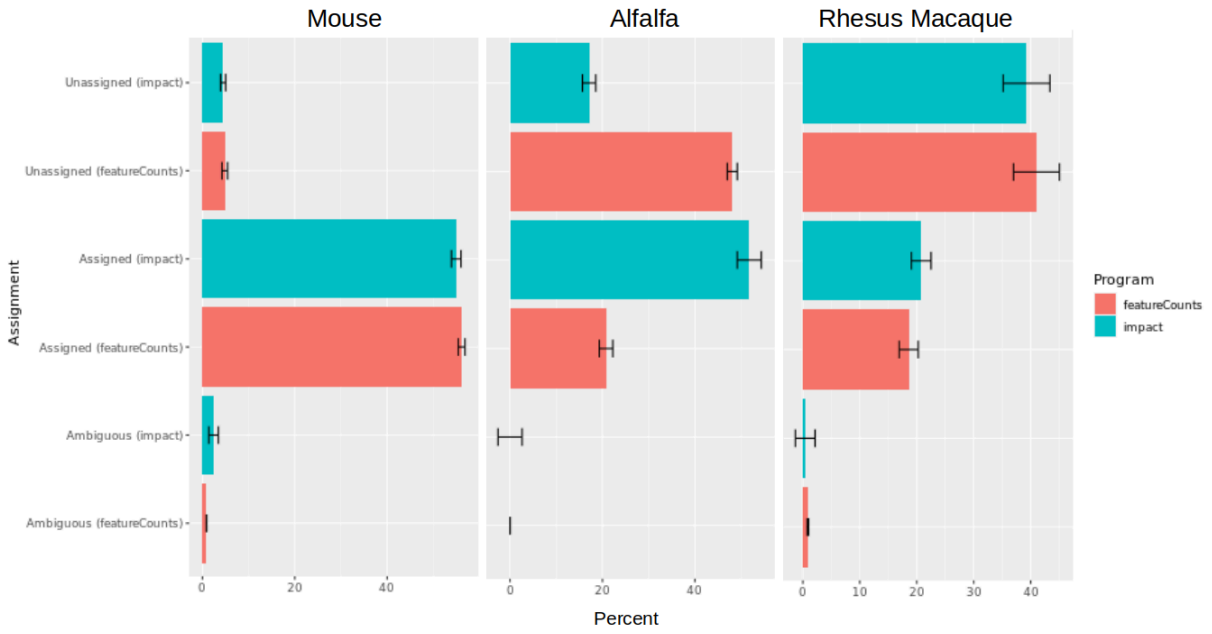


Figure 2: High Complexity Read Clusters



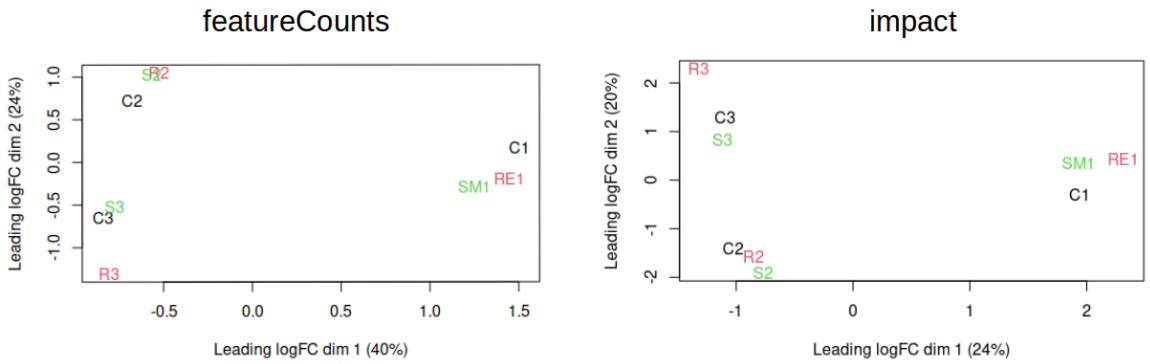
Comparisons between impact and featureCounts began with assessing read assignment statistics. Impact was successful in reducing the number of unassigned reads in two of the three datasets (Figure 3). This change was most substantial in the alfalfa dataset, as quantification with impact resulted in the number of unassigned reads decreasing by 31.0% and the number of assigned reads increasing by 31%. The change in ambiguously assigned reads was negligible, as it only increased by 0.003%. The rhesus macaque dataset saw only incremental improvements to read assignment statistics, increasing the number of assigned reads by 2.2% and reducing unassigned reads by 1.7%. Interestingly, the number of reads that were ambiguous saw a slight decrease in the rhesus macaque dataset, which was about 0.5% of the total reads. The mouse dataset saw slightly worse results when impact was used for quantification. While the number of unassigned reads did see a slight improvement, this was accompanied by a decrease in the number of assigned reads by 1.2% and increase in the number of ambiguous reads by 1.5%.

Figure 3: Read Assignment Statistics



Our next goal was to assess impact's ability to preserve global patterns of gene expression. This was assessed by creating a multidimensional scaling (MDS) plot for each dataset using counts tables generated from featureCounts and impact. An MDS plot is a distance-preserving two dimensional representation of gene expression profiles and is a popular tool for assessing gene expression patterns across and entire experiment. Beginning again with the alfalfa dataset, we see that the MDS plots are very similar between the two methods, as sample clusters are conserved (Figure 4). Additionally, we see that the batch effect present in this dataset is equally distinct in the two MDS plots (Supplementary Figure 2). However, we do see the scale of the values on the y-axis in the two MDS plots differ considerably. The rhesus

Figure 4: MDS Plots (Alfalfa)



macaque dataset also saw similar results, as the MDS plots were almost exact mirror images of each other but with a larger scale on the y-axis (Figure 5). Lastly, the mice dataset yielded mixed results. The MDS plot from impact had somewhat similar clustering patterns between samples with similar placement along the x-axis, yet distances between samples representing different sexes of mice was less present (Figure 6, Supplementary Figure 3). Additionally, impact seemed to generate an outlier in the data, sample TS25_4_2, which was not present in the featureCounts MDS plot.

Figure 5: MDS Plots (Rhesus macaque)

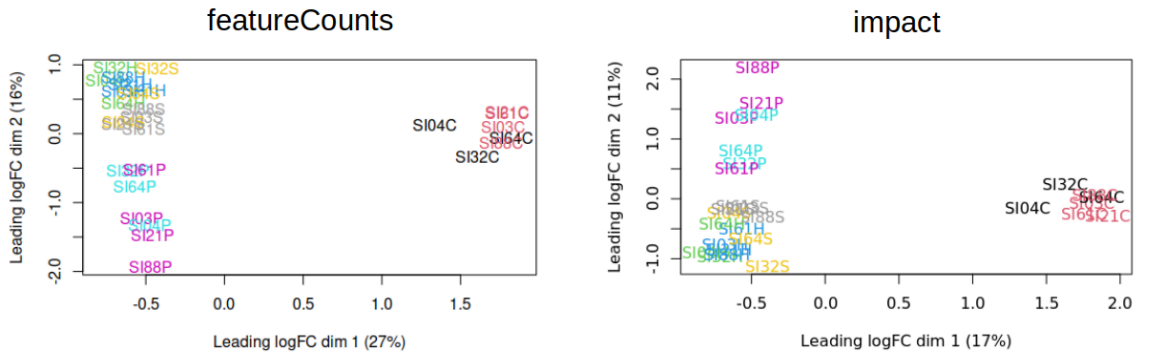
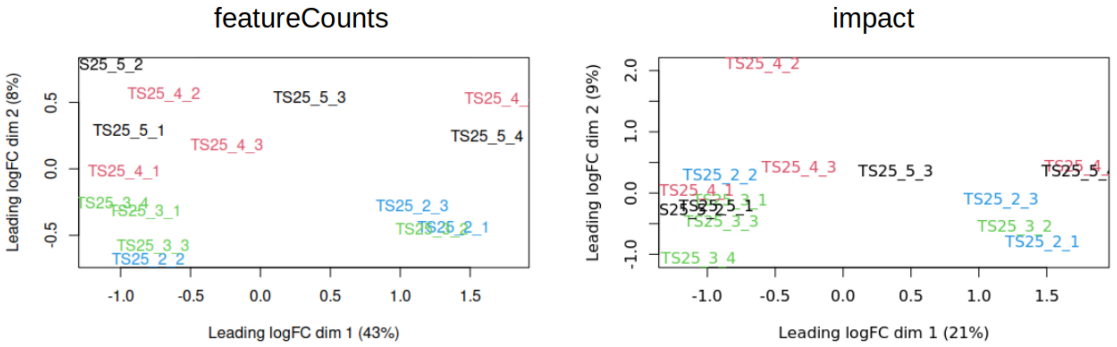


Figure 6: MDS Plot (Mouse)



Comparing the results of differential expression analysis revealed that impact detected many of the same differentially expressed genes identified by featureCounts (Supplementary Table 1). In the rhesus macaque dataset, where we compared tissues S, H, and P to tissue C in the infected macaques, all contrasts yielded similar lists of differentially expressed genes (Figure 7). When contrasting tissue P with tissue C, 75.5% of the total DE genes were shared and 74.5% of the 1000 most DE genes were shared. Impact, however, identified 314 fewer differentially expressed genes. For tissue H versus tissue C, the shared genes increased to 79% when comparing all the DE genes and 81.7% when comparing the 1000 most DE genes. Impact again detected less genes (180). The last contrast was between tissue S and tissue C and showed 77.5% of the total DE genes were shared with 82% of the top 1000 DE genes being shared. Here, impact identified 214 less differentially expressed genes. In the mouse dataset, we examined contrasts between the sexes for the two diets. These contrasts had similar results to the rhesus macaque dataset, but with higher dissimilarity and greater reduction in detection of differentially expressed genes with impact (Figure 8). The first contrast between males and females in diet 1 resulted in featureCounts identifying 2,010 significantly differentially expressed genes whereas impact only identified 1,302. Of the total set of DE genes, 61.3% were shared and 69.5% of the

Figure 7: Differentially Expressed Genes Identified by featureCounts and Impact (Rhesus macaque)

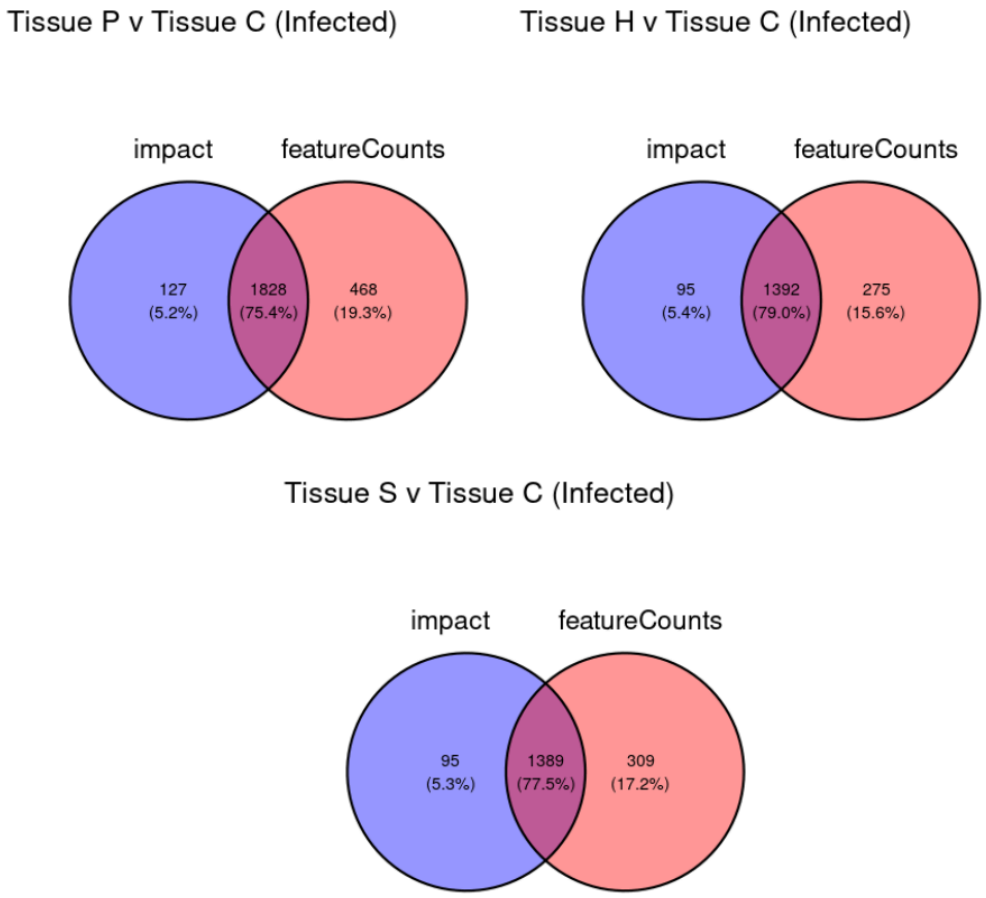


Figure 8: Differentially Expressed Genes Identified by featureCounts and Impact (Mice)

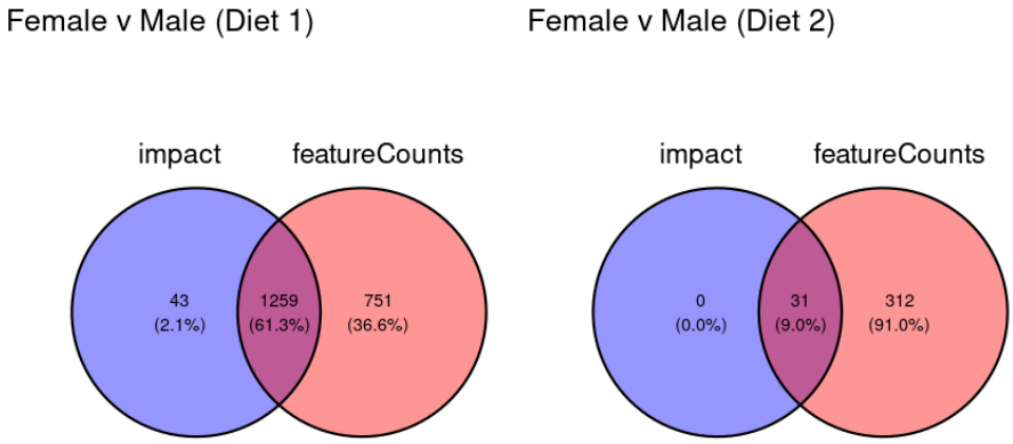
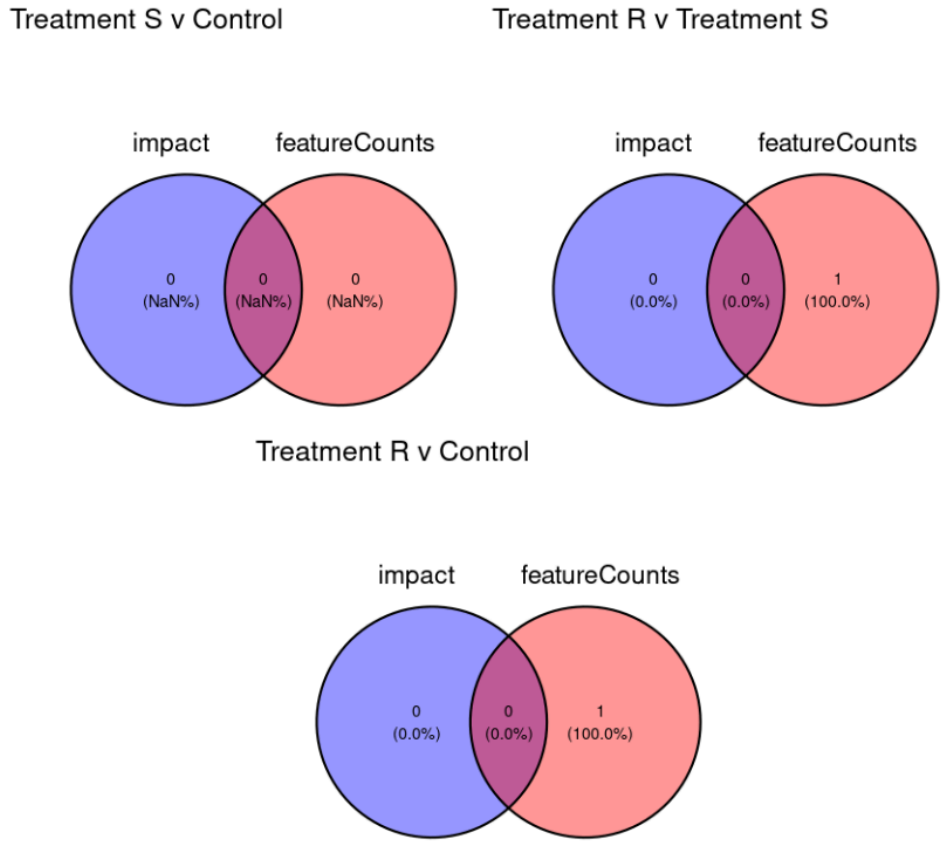


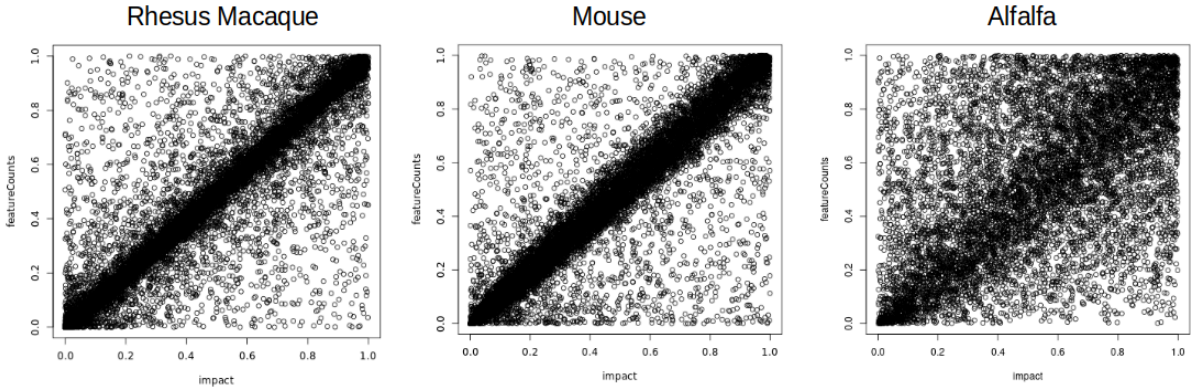
Figure 9: Differentially Expressed Genes Identified by featureCounts and Impact (Alfalfa)



1000 most differentially expressed genes were shared. Contrasting the sexes in diet 2 yielded only 343 statistically significant differentially expressed genes using featureCounts, while impact was only able to detect 31. All of these 31 differentially expressed genes were also identified by featureCounts. The results of differential expression of the alfalfa dataset also yielded similar results between impact and featureCounts. As previously stated, this experiment was to be used as a negative control, as we expect to identify few, if any, differentially expressed genes due to the significant batch effect present in this experiment. This result was preserved using the impact algorithm, as zero differentially expressed genes were identified using the impact method for all comparisons. FeatureCounts, on the other hand, was able to identify 1 differentially expressed gene each in two of the contrasts (Figure 9).

We also assessed similarities in differential expression results by observing the correlation between the list of p-values for each contrast. This was accomplished using dot plots to visualize the correlation and by calculating the average Spearman correlation coefficient for all contrasts. When we compared the p-values for the contrasts in the rhesus macaque dataset, we found strong positive correlations. This was also supported by the contrasts having an average correlation coefficient of 0.93. Similar results were observed in the mice datasets, as the dotplots for all comparisons had equally tight correlations. The average correlation coefficient for this dataset was also 0.91. Interestingly, when the p-values for the contrasts in the alfalfa dataset were compared, they showed substantially less correlation between the two methods, yielding an average correlation coefficient of 0.52 (Figure 10). When the raw gene counts were considered, the mouse dataset showed the tightest correlation, with the alfalfa and rhesus macaque datasets exhibiting more variation between the two methods (Supplementary Figure 4).

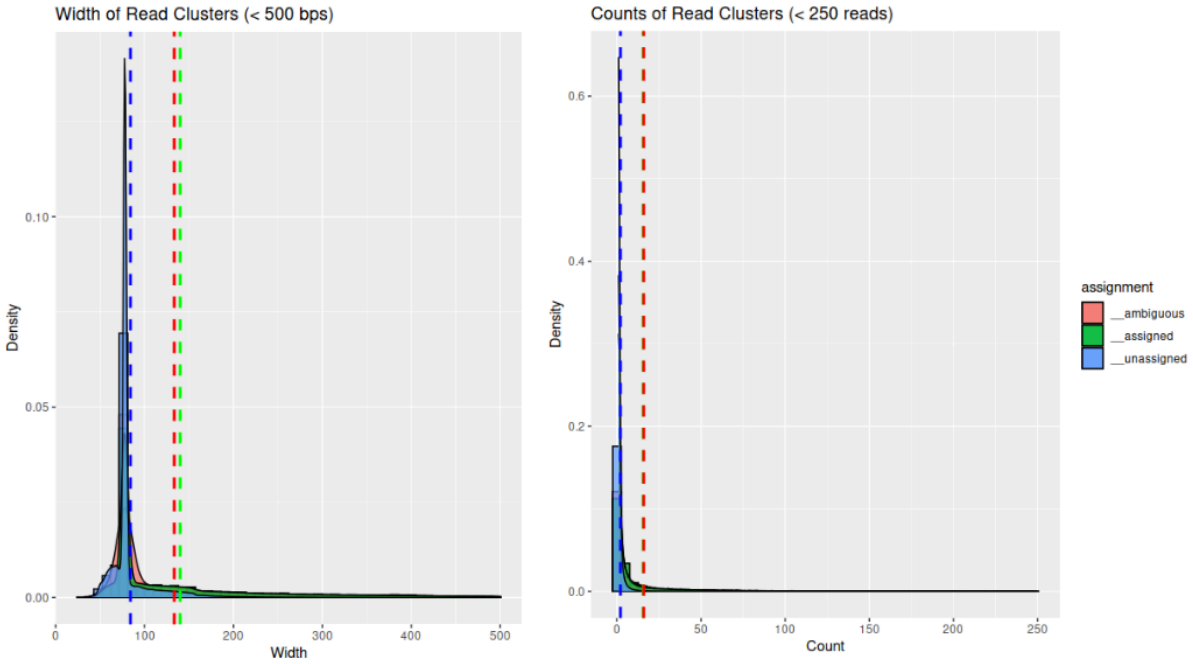
Figure 10: Example P-Value Dot Plots for Rhesus Macaque, Mouse, and Alfalfa



Lastly, we attempted to identify differences in the characteristics of read clusters that were differentially assigned. The two attributes we investigated were read cluster width and read counts, which showed distinct differences based on their read assignment (Figure 11). On average the assigned read clusters contained more reads and spanned greater regions than the read clusters that were unassigned. Additionally, the difference between cluster width and read

counts were far less distinct, if present at all, between assigned read clusters and those marked as ambiguous.

Figure 11: Histogram of Cluster Widths and Counts by Assignment



Discussion

The results of our read assignment analysis shows that, depending on the completeness of the genome annotation, impact is successful in reducing the number of unassigned reads and increasing the total signal captured in 3'-based RNAseq methods. One consequence, however, is that impact can also increase the number of ambiguous reads. This is an expected result, as overlapping reads and treating them as a read cluster increases the bounds of the reads, thereby increasing the probability the entire cluster overlaps with multiple genes. This probability also seems to increase with completeness of genome annotation, as evident from the more substantial increase in ambiguously assigned reads in mice relative to alfalfa. By the same logic, the number of ambiguously assigned reads is also likely related to gene density as well, as the more gene rich the genome is, the more likely a read cluster is to overlap with multiple genes.

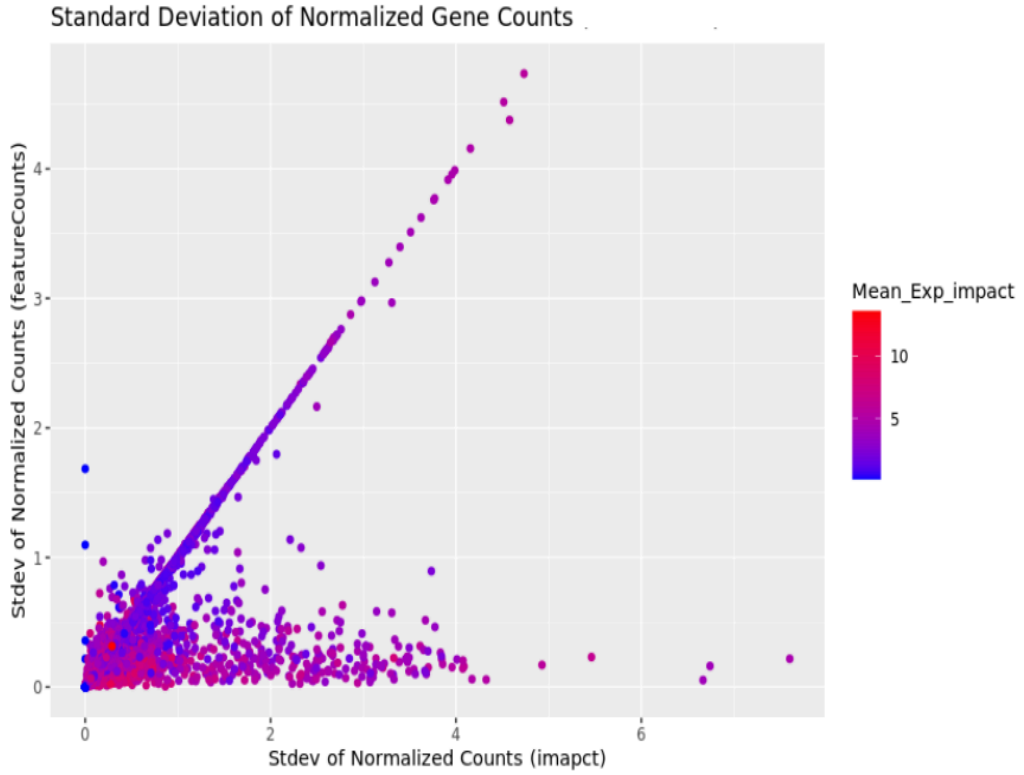
Overall, the results of our differential expression analysis for the three datasets shows that impact is capable of preserving patterns of gene expression identified by traditional RNAseq methods and detecting significantly differentially expressed genes. This was most apparent in the rhesus macaque datasets, where the MDS plots showed similar clustering of samples and the list of differentially expressed genes were highly similar. Additionally, the p-values generated between the two methods had high correlation coefficients for this dataset, indicating similar results in gene quantification. In the case of the alfalfa dataset, despite picking up significantly more signal, impact did not identify differentially expressed genes. This is what we would expect given the significant batch effect in the data and is evidence that our method is capturing true signal that is representative of the experimental data and not generating false positives by capturing technical noise. The results from the mice dataset indicate that there is room for

improvement with impact's algorithm, as our program was moderately successful in identifying most differentially expressed genes, yet global patterns of expression were not well preserved.

Assessing the characteristics of read clusters based on their assignment to genes revealed marked differences in read counts and cluster widths. Both assigned and ambiguous read clusters had much larger widths and greater read counts than read clusters who were unassigned. These results suggest that read counts and cluster width may be useful for identifying read clusters that correspond to novel genes or polyadenylation sites in future iterations of this algorithm.

One puzzling aspect of these results was that the number of differentially expressed genes was lower when gene expression was quantified using impact compared to featureCounts. Originally, we suspected the number of lower number of differentially expressed genes to be the result of more genes being filtered due to low expression. However, the datasets that yielded lower numbers of differentially expressed genes (mouse and rhesus macaque) showed a higher number of expressed genes for impact relative to featureCounts. We then investigated the possibility that increased variation in gene counts was responsible for the decrease in sensitivity for differentially expressed genes, as an increase in variability within the groups would decrease the number of differentially expressed genes. This was investigated in the mouse dataset where the standard deviations within groups for the normalized gene counts was compared between the two methods. The observed trends are best exemplified by figure 12, which shows

Figure 12: Standard Deviation of Normalized Gene Counts (Mouse)

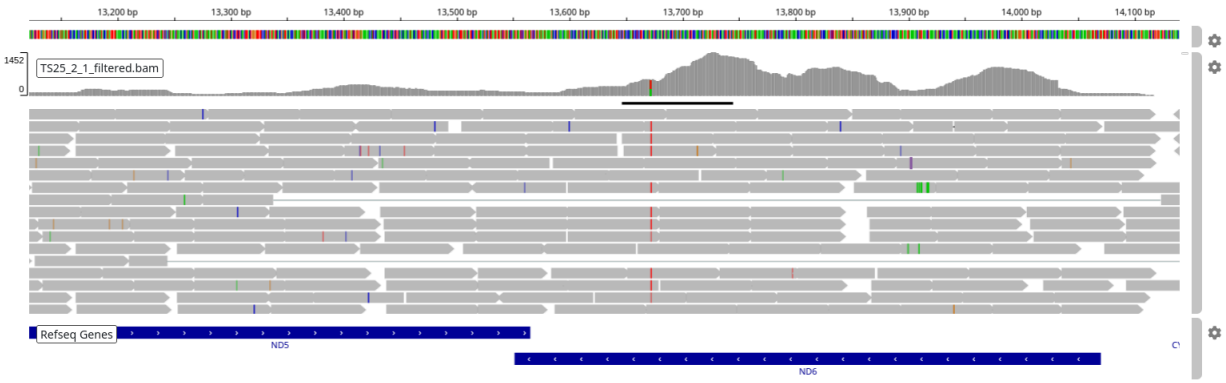


gene expression standard deviations for the males of diet 2. The figure shows a positive linear correlation between many of the genes, yet there seemed to be a distinct cluster of genes that were moderately to highly expressed that had little variation in the featureCounts table but increased group variability for the gene counts generated from impact. Upon further investigation of these genes, a majority of them contained instances where a few samples in the group had extremely high or low values compared to the rest of the group, indicating that true gene expression was being lost in some samples. These regions were observed in IGV, where we learned that many of these genes did in fact have overlapping read clusters; however, they contained likely misaligned reads with large gaps in the alignment. These reads often spanned multiple read groups, bridging them together so that they are erroneously treated as a single read

cluster by the impact algorithm. When these read clusters overlapped multiple genes in some samples, it resulted in the read cluster being missassigned and the gene counts being lost, thus increasing the variability of the entire group and decreasing detection sensitivity for differentially expressed genes.

Investigation of the increased group variability using impact revealed one significant shortcoming of the impact method— its inability to resolve overlapping read clusters originating from two different polyadenylation sites. The impact algorithm erroneously treats these clusters as a single group, which prevents it from accurately assigning the distinct read clusters to the appropriate genes. An example of this issue can be seen in figure 13 where we see a single continuous read cluster that spans multiple genes and contains multiple peaks in the coverage.

Figure 13: Overlapping Read Clusters



This situation will either artificially increase the gene expression of one gene if the overlap score is better for one gene, or result in the complete loss of signal by marking it as ambiguous. More work is currently being done to resolve this issue, as we are developing a peak detection method to detect and separate these individual read clusters. We are also working to use this peak detection functionality to add transcript level expression quantification; although, due to the inherent limitations of 3’ sequencing methods, these results will be only be able to differentiate transcripts that use different terminal exons or are differentially polyadenylated.

Conclusion

The combined results of our analysis showed impact is an effective but imperfect method for quantifying gene expression for TAGseq analysis. The read clustering approach is successful in rescuing lost reads in organisms with incomplete annotations while offering similar read assignment results to traditional methods in sufficiently annotated organisms. It preserves global patterns of gene expression across the experiment and mostly reproduces results of differential expression, although with less sensitivity to differentially expressed genes due to an increases in gene count variability from erroneous read cluster assignment. Overall, these results suggest that the best use case for this algorithm in its current form is in organisms with poor gene annotations. Future iterations of this program will seek to implement peak detection methods to improve cluster identification, resulting in the ability to separate overlapping read clusters as well as identify novel genes that are not in the annotation. This will also aid in reducing the number of ambiguously assigned reads in well studied organisms while also allowing for transcript level expression quantification, which was previously not possible using 3’-based sequencing methods.

References

[1] Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B 57:289–300

[2] 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. Jan 1; 29(1): 15-21.

[3] Kukurba, K.R., and Montgomery, S.B. (2015). RNA Sequencing and Analysis. Cold Spring Harb Protoc *2015*, 951–969.

[4] Law, C.W., Chen, Y., Shi, W. et al. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15, R29.

[5] Li, A., Liu, A., Du, X., Chen, J.-Y., Yin, M., Hu, H.-Y., Shrestha, N., Wu, S.-D., Wang, H.-Q., Dou, Q.-W., et al. (2020). A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. Hortic Res *7*, 1–12.

[6] Liao Y., Smyth G.K., and Shi W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30

[7] Meyer, E., Aglyamova, G.V., and Matz, M.V. (2011). Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Molecular Ecology *20*, 3599–3616.

[8] Moll, P., Ante, M., Seitz, A., and Reda, T. (2014). QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods *11*, i–iii.

[9] Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct *4*, 14.

[10] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[11] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7),e47

[12] Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res *27*, 491–499.

[13] Spearman, C. (1987). The Proof and Measurement of Association between Two Things. The American Journal of Psychology *100*, 441–471.

[14] Tandonnet, S., and Torres, T.T. (2016). Traditional versus 3′ RNA-seq in a non-model species. Genom Data *11*, 9–16.

[15] Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*, 178–192.

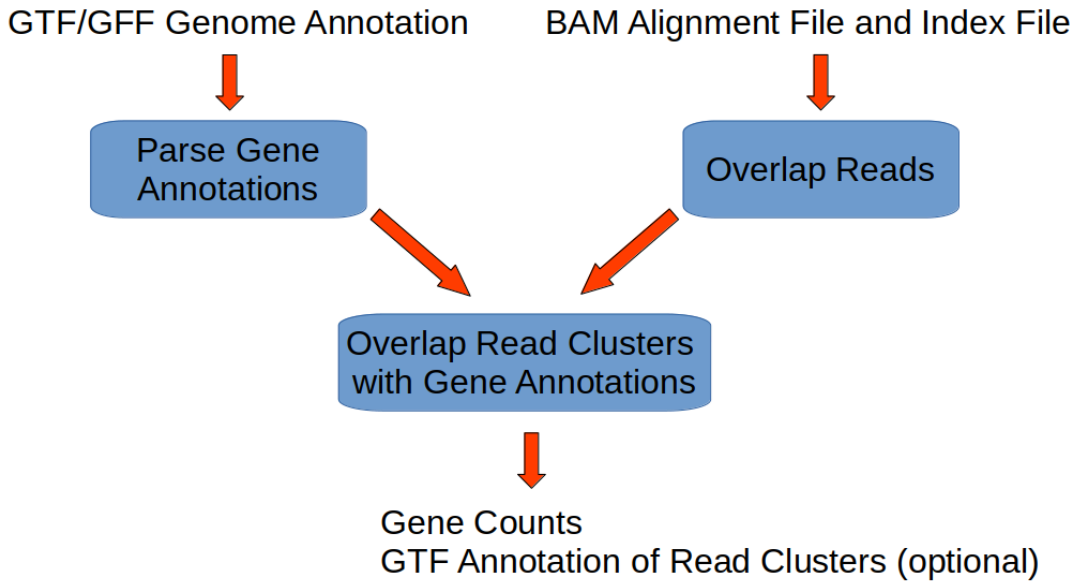
[16] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57–63.

Supplementary Material

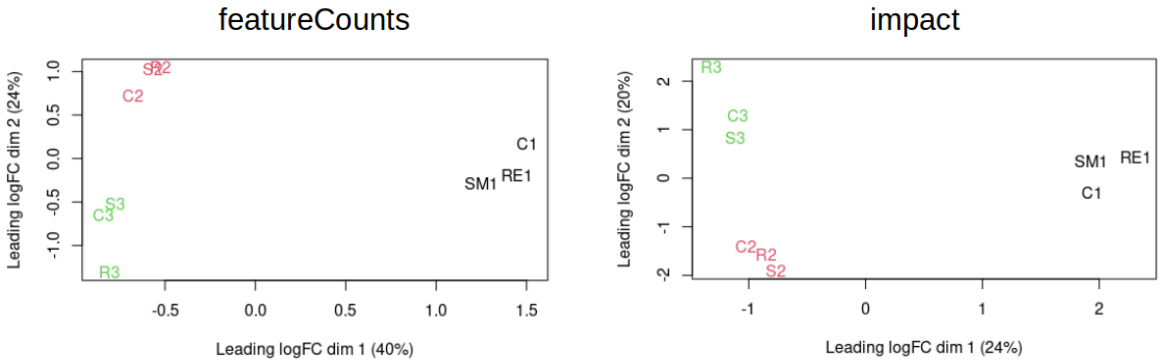
Supplementary Table 1: Comparison of Differential Expression Analysis Results

Contrast	featureCounts	impact	Percent Shared
Female vs Male (Diet 1)	2,010	1,302	61.3%
Female vs Male (Diet 2)	343	31	9.0%
Treatment S vs Control	0	0	NA
Treatment R vs Control	1	0	0%
Treatment R vs Treatment S	1	0	0%
Tissue P vs Tissue C (Infected)	2,296	1,955	75.4%
Tissue H vs Tissue C (Infected)	1,667	1,487	79.0%
Tissue S vs Tissue C (Infected)	1,698	1,485	77.5%

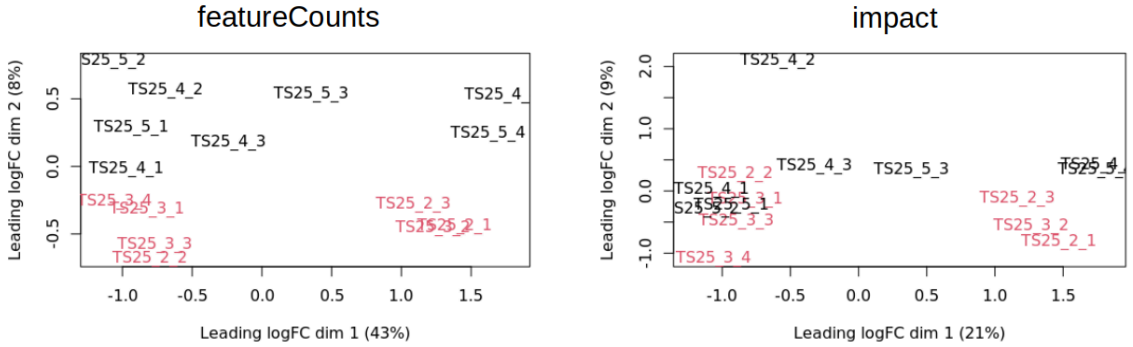
Supplementary Figure 1: Algorithm Flowchart



Supplementary Figure 2: Alfalfa MDS Plot by Batch



Supplementary Figure 3: Mouse MDS Plot by Sex



Supplementary Figure 4: Dot Plot of Log2(Raw Counts)