**Identifying Discrepancies between RNA-Seq Methods in *Mycobacterium Tuberculosis* Research**

Mirza Khan, Katie McCormack, Meiling Melzer

**Abstract**

RNA-seq analysis is a way of determining differential gene expression using differences in RNA transcript abundance. Broadly, this consists of mapping reads to a reference genome, counting the number of reads that map to each locus, and quantifying differential expression. There are numerous methods available for conducting this type of analysis, which differ in both speed and accuracy, such that using different types of methods may produce biased results. The purpose of our project was to determine the extent to which the method utilized alters which genes are identified as differentially expressed. To this end, we selected a recently published article that used the RNA-seq methods Tophat2 and Cufflinks to identify 750 differentially expressed genes between murine bone marrow derived macrophages (BMDMs) infected with virulent *Mycobacterium tuberculosis* and BMDMs infected with an attenuated strain. Because this is a recent paper aimed at solving a serious global health problem, it provides an excellent case study for highlighting the importance of ensuring the robusticity of RNA-seq analysis methods. We obtained the datasets used in the original article and reanalysed them with the quasi-mapping method, Salmon. We hypothesized that although there will be overlap between our results and those published, some of the top differentially expressed genes identified between methods will differ. We found that 14 of the top 20 differentially expressed genes identified in our analysis were also identified by the original paper, suggesting relatively high robusticity between these RNA-seq methods. Finally, we conducted gene enrichment analysis in order to explore the biological relevance of our results.

**Introduction**

The widespread potential applications of RNA-seq analysis has lead to subsequent development of dozens of methods which aim to improve the speed and accuracy of their workflow (Pachter 2017). The generalized RNA-seq workflow consists of mapping RNA transcript reads to the appropriate reference genome, counting the number of reads that map to each locus of interest, and finally using this count data to identify genes that are differentially expressed between the conditions of interest (Zhang et al. 2014). Comparative studies have demonstrated the existence of significant differences between RNA-seq methods that increase with the dissimilarity of the approach used (Sahraeian, et al., 2017). It is therefore imperative to understand the degree to which the use of different RNA-seq analysis methods may affect the genes that are identified as differentially expressed and how robust studies that utilize RNA-seq methods are to these biases.

Our study aims to address this problem by comparing the results of analyzing the same RNA-Seq datasets between two different methods to determine how differentiated the results are from one another and whether there are biologically-relevant differences between these results. To this end, we chose a 2019 paper published in *Scientific Report* that used Tophat2 and Cufflinks to analyze and identify differentially expressed genes between BMDMs infected with a virulent strain of *Mycobacterium tuberculosis* (H37Rv) and BMDMs infected with an attenuated version of this strain (H37Ra) (Lee et al. 2019). Assessing the validity of these results remains particularly valuable because of the high significance of tuberculosis research, which costs billions of dollars each year and aims to combat a leading cause of infectious disease death globally (WHO 2018). This study also provided us sequences which were available online, allowing us to conduct our own RNA-seq analysis of the two using the quasi-mapping method Salmon. The results of this analysis were then compared with the original results.We found that 14/20 of our top differentially expressed genes (DEGs) overlapped with the DEGs identified in the original paper. Finally, we conducted a gene enrichment analysis in order to further explore the biological relevance of our results.

**Results**

**Quantification**

Of our four total samples, two were treated with the non-virulent strain and two were treated with the virulent strain. Counts using Salmon yielded 20,533,789 and 12,370,920 reads for each of the non-virulent samples; 23,455,676 and 15,172,309 reads for each of the virulent samples (Table 1).

**Differential Expression**

Of the original 35,663 genes, filtering to include only those genes with greater than one count yielded 21,704 genes. In other words, 13,959 genes had null or only one read. Differential expression analysis revealed 189 genes as significant at a Benjamini-Hochberg adjusted p-value threshold of < 0.001 (Table 2). Comparing virulent to non-virulent strains, the log-fold change was greater than zero for 43 genes (0.2% of the 21,704 genes) and less than zero for the remaining 146 (0.67%) differentially expressed genes.

We visualized the differential expression findings using an average vs mean–difference (MA) plot of these results where red points are significant differentially expressed genes and grey points are not differentially expressed genes (Figure 1). The *x*-axis reflects the average expression of each gene such that genes that are higher on the *x*-axis are more highly transcribed. The *y*-axis is the relative difference between the virulent and non-virulent samples based upon log fold-change. The shape of the MA-plot in the region with low average expression levels is due to the high degree of variability of the log ratios for genes with lower counts. For further analysis, we were able to remedy this high variability using a shrinkage estimation procedure (figure not shown).

Following shrinkage estimation, twenty differentially expressed genes (DEGs) we identified with the lowest adjusted p-values (Table 3). We generated for comparison to DEGs reported in the Lee et al. study. The gene with the lowest adjusted p-value was Matrix Metallopeptidase 13 (MMP13; Ensembl ID: ENSMUSG00000050578). Figure 2 shows the normalized counts for MMP13 between the virulent and non-virulent strains. This highlights the difference in observed counts between the virulent and non-virulent strains for this differentially expressed gene. Although represented in normalized counts, we find that the non-virulent strains express more MMP13 than their virulent counterparts.

**Gene Enrichment Analysis**

We conducted gene enrichment analysis as another means of comparison to the Lee et al. (2019) study that directly addresses the biological relevance of the results. We found 931 gene ontology (GO) categories that were over-represented using an Benjamini-Hochberg adjusted p-value threshold of < 0.05 for significance. We examined those GO categories with the twenty lowest adjusted p-values (Table 4). Our findings show that many over-represented processes are related to immune and inflammatory response. With respect to those categories related to molecular function, we found cytokine activity and cytokine receptor binding to be among the most enriched GO pathways.

**Discussion**

When comparing virulent and non-virulent strains, Lee et al. reported 155 differentially expressed genes based on their analysis using the same significance threshold. By contrast, we found 189 differentially expressed genes between both groups. Upon comparing these 155 and 189 genes, we find that only 98 (63.2%) of the genes are common between both groups, which was below our *a priori* tolerance threshold of 75%.

In examining the output from the Lee et al. analysis, each of their DEGs have an adjusted p-value < 0.001, but all documented as having an adjusted p-value of exactly 0.000574564. It is unclear why each of their DEGs is listed as having the same adjusted p-value. The other genes listed that are not DEGs do not suffer from this issue. We were thus unable to identify those genes with the lowest adjusted p-values to compare to our own adjusted p-values for DEGs, since all of their DEGs comparing virulent and non-virulent strains are listed as having the same p-value. Given this constraint, we discussed that we would expect at least each of our twenty most significant DEGs, with the lowest adjusted p-values, would be present among the 155 DEGs they identified.

Of the twenty most significant DEGs, only fourteen of these were identified among the 155 DEGs reported by Lee et al. The six DEGs not found by Lee et al. are: CXCL3, CSF3, OLR1, LIF, CAR2, and AA467197. Each of these genes appears to play some role in immune or inflammatory response. For example, CXCL3 serves as a chemoattractant for neutrophils and CSF3 in granulocyte formation and function [GeneCards]. Of note, LIF, CAR2 and AA467197 are the eighteenth to twentieth members on our list of twenty most significant DEGs. We suspect that these being less significant than the other seventeen may have some correlation with these genes not being deemed significant in the Lee et al. study.

Our results indicate that there can be meaningful differences in the results obtained via the RNA-seq methods Tophat2/Cufflinks and Salmon/DESeq2. It is promising, however, that both analyses identified many of the same genes and that the gene enrichment analysis showed a connection between these genes and immune response. These results are consistent with what would be expected given that Lee et al.’s study focused on murine immune responses to TB infection, underscoring the importance of the differences between analysis methodologies and potentially using a combination of perspectives to improve the scope of the results. For example, Sahraeian et al. (2017) proposed a protocol, called RNACocktail, which claims to be more comprehensive than using a single set of methods alone. An exploration of this protocol and other RNA-seq methods would have allowed us to better address the degree to which different methods produce different results, as would the inclusion of datasets from other contexts. It is unclear how generalizable our findings are to other studies or RNA-seq methods, but we were successful in demonstrating the continued relevance of this problem for contemporary research.

**Methods**

We obtained published data used by Lee et al. for their RNA-seq from the NCBI database BioProject PRJNA50644 (Lee, J. et al., 2019). The reference mouse transcriptome was obtained from Ensembl (ftp://ftp.ensembl.org/pub/release-98/fasta/mus\_musculus/cdna/). The SRA toolkit was used to generate FASTQ files from the original compressed SRA files (SRA Toolkit Development Team). An index using the reference transcriptome was built using Salmon. Quantification and mapping were then performed using Salmon taking the index and FASTQ files as inputs (Patro et al., 2017). The subsequent steps were performed using Bioconductor and R version 3.6.0 on the Vanderbilt University computing cluster. The tximeta Bioconductor package was used to read the transcript abundance files generated from Salmon. This generates gene-level count matrices and provides automated annotation metadata, which assists with transcript-to-gene mapping (Love et al., 2019a; Love et al., 2019b). The resulting output was then read using the DESeq2 package, which provides tools for differential gene expression analysis (Love et al., 2014). Our study compared virulent and non-virulent strains. We excluded genes with counts that were null or less than or equal to one. We used an adjusted p-value of < 0.001 as our threshold for significance, which was the significance threshold used by Lee et al. Gene enrichment analysis of the RNA-seq analysis was performed using the goseq package (Young et al., 2010).

We established *a priori* that concordance of ≥ 75% of our total population of differentially expressed genes in comparison with the total list of DEGs identified by Lee et al. for their comparison of virulent and non-virulent strains would be considered satisfactory. Additionally, we planned to compare our twenty DEGs with the lowest adjusted p-value to those from Lee et al.; we expected to find 100% concordance between both these groups. On examination of the reported p-values provided by Lee et al., all of the DEGs for their comparison of virulent and non-virulent strains were documented as having the same adjusted p-value. Thus, in lieu of our previously defined mode of comparison, we determined that we would expect to find all of the twenty DEGs with the lowest adjusted p-value from our analysis to be among the total population of DEGs found by Lee et al. comparing virulent and non-virulent strains.

Code is available upon request.

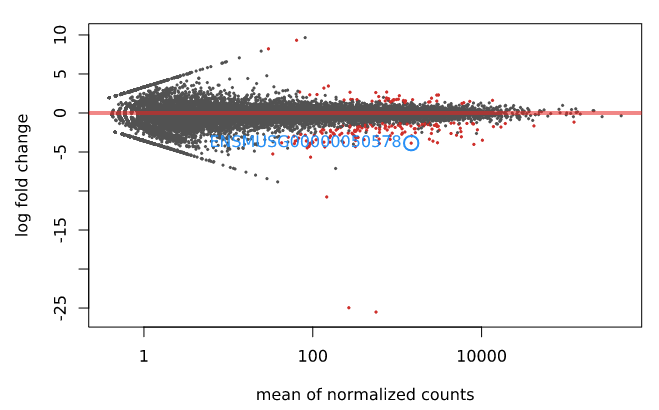
**Figures and Tables**

**Table 1.** Individual reads for virulent and non-virulent samples

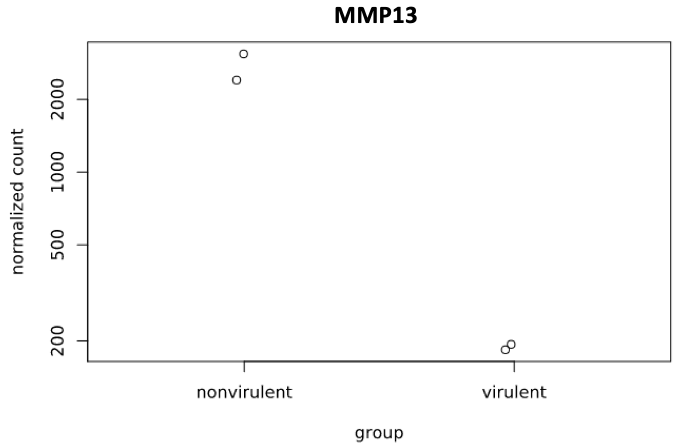
|  |  |  |
| --- | --- | --- |
|  | **Sample 1** | **Sample 2** |
| **Virulent Strain** | 23,455,676 | 15,172,309 |
| **Non-Virulent Strain** | 20,533,789 | 12,370,920 |

**Table 2.** Number of genes based on number of reads

|  |  |
| --- | --- |
|  | **Genes** |
| **Total** | 35,663 |
| **Null or Only One Count** | 13,959 |
| **Greater than One Count** | 21,704 |
| **Differentially Expressed (p<0.001)** | 189 |

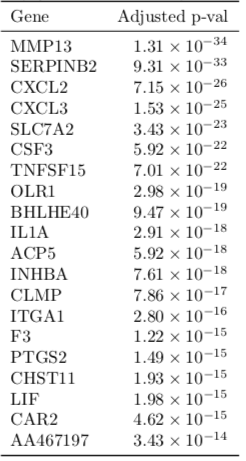
****

**Figure 1.** Average vs mean–difference plot. Each gene is represented by a point in the graph. Red points are those genes with an adjusted p-value < 0.001. The light blue circle highlights the Matrix Metallopeptidase 13 gene. This gene had the lowest adjusted p-value in this study.

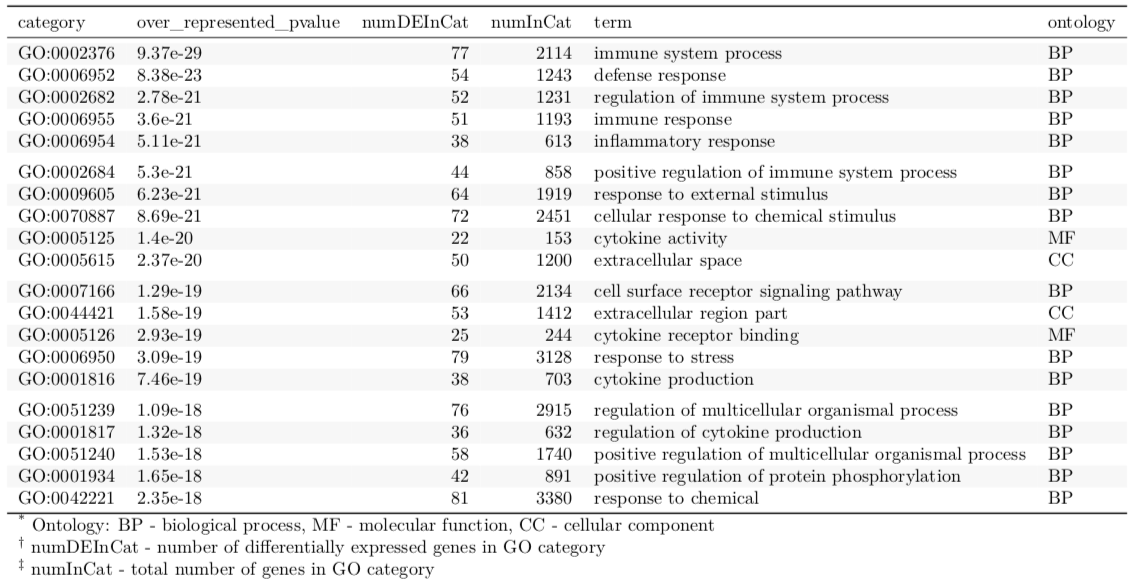
****

**Figure 2.** Normalized count information for the Matrix Metallopeptidase 13 gene between virulent and non-virulent strains of Mycobacterium tuberculosis.

**Table 3.** Twenty differentially expressed genes with the lowest adjusted p-value



**Table 4.** Gene Enrichment Analysis - categories with the twenty lowest adjusted p-values



**Individual Contributions**

**Mirza Khan** - I was involved with data acquisition, downloading the required software (Salmon, Bioconductor libraries, etc.), performing quantification using Salmon, differential expression and gene enrichment analyses. I also analyzed the differential expression output and gene enrichment analyses from the Lee et al. supplemental materials, and compared these with our own findings. With respect to the presentation, I generated results and figures and assisted with slide preparation. For the report, I documented the results and methods, and prepared the figures and tables.

**Katie McCormack**- I conducted a literature review in order to identify relevant papers for our analysis and to understand the current literature that exists for the comparison of RNA-seq analysis. I also researched the tuberculosis aspect of our project and was involved with the interpretation of the results of our analysis. I worked to create the slides for our presentation and presented on the background and future direction components of the presentation. I worked with the other members of our group to write the final paper for the project, especially the introduction,abstract, and discussion sections.

**Meiling Melzer** - I researched current RNA-Seq analysis methods, to which I then created slides for and presented as background information in our presentation. I also assisted in researching current literature on the MMP-13 protein in tuberculosis research.

**Acknowledgements**

We would like to acknowledge Professor Capra and the BSCI 3272 students who peer-reviewed our proposal for their suggestions and assistance with this project.

**References**

[1] Pachter, Lior.. “How Not to Perform a Differential Expression Analysis (or Science).” *Bits of*

*DNA*,( 2017).

[2] Zhang, H. Z., et al. “A Comparative Study of Techniques for Differential Expression Analysis

on RNA-Seq Data.” Plos One 9(8), e103217. (2014).

[3] Sahraeian, S., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P., Au, and K., Lam, H.

Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun*, 8(1), 59–59. (2017).

[4] Lee, Junghwan, et al. “Characterisation of Genes Differentially Expressed in Macrophages by

Virulent and Attenuated Mycobacterium Tuberculosis through RNA-Seq Analysis.” Scientific Reports, 9,1,1–9. (2019).

[5] WHO Global TB Report (2018).

[6] CXCL3 Gene - GeneCards | CXCL3 Protein | CXCL3 Antibody. Available at: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=CXCL3>.

[7] CSF3 Gene - GeneCards | CSF3 Protein | CSF3 Antibody. [Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=Csf3](https://www.genecards.org/cgi-bin/carddisp.pl?gene=Csf3).

[8] SRA Toolkit Development Team. *ncbi/sra-tools*. (NCBI - National Center for Biotechnology Information/NLM/NIH, 2019). https://github.com/ncbi/sra-tools.

[9] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419 (2017).

[10] Love, M. I. *et al.* *Tximeta: reference sequence checksums for provenance identification in RNA-seq*. http://biorxiv.org/lookup/doi/10.1101/777888 (2019a) doi:10.1101/777888.tximeta

[11] Love, M. I. *et al.* RNA-seq workflow: gene-level exploratory analysis and differential expression.<https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html>. (2019b)

[12] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).

[13] Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11, R14 (2010).goseq