An Introductory Genomics Workflow for Exploring Publicly Available Infectious Disease Data

by

Praveer S. Vyas

Bachelor of Science, Carnegie Mellon University, 2018

Submitted to the Graduate Faculty of the

Department of Infectious Diseases and Microbiology

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Public Health

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This essay is submitted

by

Praveer S. Vyas

on

April 29, 2022

and approved by

Essay Advisor: Jeremy J. Martinson, DPhil, Assistant Professor, Department of Infectious Diseases and Microbiology and Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Essay Reader: Jean B. Nachega, MD, PhD, MPH, Associate Professor, Department of Epidemiology and Department of Infectious Diseases and Microbiology, Graduate School of Public Health, University of Pittsburgh

An Introductory Genomics Workflow for Exploring Publicly Available Infectious Disease Data

Praveer S. Vyas, MPH

University of Pittsburgh, 2022

Abstract

Advances in computational and gene sequencing technology provide an avenue for public health students and professionals who are interested in gaining exposure to biological research. Differential expression (DE) analysis can be performed using publicly available tools as well as data to learn more about the biological differences between samples from humans, animals or pathogens. There is a vast amount of publicly available gene expression data that can be searched to find a dataset related to a topic of interest. As an example, infectious disease epidemiology students could use their own computer to perform a DE analysis on an existing dataset related to a trend they studied or observed, without the need to enter a laboratory. DE analyses can be performed quickly on personal computers using pseudoalignment software, which is less computationally intensive and faster than alignment of RNA-seq reads to a reference genome. An algorithm for performing a DE analysis on an infectious disease topic utilizing pseudoalignment will be provided. Basic requirements for using this algorithm are a working knowledge of the statistical programming language R, familiarity with executing shell scripts and a general understanding of the central dogma of biology. This approach will provide the user with experience performing complex genomics analyses and further their professional development. An illustrative analysis related to the public health issue of progression of disease in individuals with latent tuberculosis infection will be provided. Direct applications of the results from this example will

be discussed in addition to how individuals in public health may benefit from utilizing this algorithm and expanding their genomics skillset.

Table of Contents

List of Tables

List of Figures

Preface

I would like to acknowledge my advisor, Jeremy Martinson, DPhil, for his valuable guidance and feedback throughout this project. I would also like to thank Jean Nachega, MD, PhD, MPH, for his valuable review of the application of the workflow presented here in the public health context of tuberculosis disease progression.

1.0 Introduction

Genomics, or the analysis of data related to the genome of an organism, is widely used in biomedical research that informs public health. Some genomics analyses include genome-wide association studies and the use of whole genome sequencing (WGS) to detect variants in genes, the sequencing and analysis of samples containing multiple genomes (metagenomics) and differential gene expression analysis. There are many examples of how research utilizing these techniques has the potential to benefit public health. WGS can be used to effectively predict drug resistance in *Mycobacterium tuberculosis* isolates from infected patients.[1] This has been implemented in healthcare settings in Western countries, and in the future, may be used in low and middle income countries (LMIC) that have a larger burden of tuberculosis as the costs of sequencing and analysis decline.[2] Phylogenetic analysis of whole genome sequences of pathogens has also been used to aid investigations of outbreaks, such as during the outbreak of *E. coli* O104:H4 in Germany in 2011.[3] Data from exome sequencing of human samples have been used to screen populations historically underrepresented in genomics research for conditions such as hereditary breast and ovarian cancer risk.[4]

Large datasets produced from genomics research are often available online to the public as a requirement of funding by the National Institutes of Health (NIH). As part of the *All of Us* research program, the NIH has itself released thousands of whole genome sequences that have been de-identified.[5] There are both free and commercial tools that can be used to analyze various types of genomics data including whole genome sequences and RNA-sequencing (RNA-seq) reads.

1.1 Overview of Differential Expression Analyses

Differential expression (DE) analyses are performed to understand how gene expression changes across biological conditions. Data obtained from RNA-seq experiments can be used in these analyses. In RNA-seq, RNA is isolated from samples, enriched for mRNA (as an example) and fragmented. The mRNAs are then converted to a cDNA library that is amplified and sequenced using a next generation sequencing technique.[6]

Gene expression data from RNA-seq experiments are available online in three major repositories: Gene Expression Omnibus (GEO), Array-express and ENCODE. However, the raw reads themselves from RNA-seq experiments are available online in repositories such as the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA). Decreases in the cost of sequencing and analysis along with the requirement of federal funding to make datasets publicly available online have resulted in a vast amount of data that is available for analysis.[7] Data can be downloaded from these repositories and analyzed using either free or commercial tools.

Prior to performing a DE analysis, quality of RNA-seq reads can be assessed by performing various quality checks, such as looking at the average per-base quality for all the base positions along reads generated from short-read sequencing. The general approach following this step is to compare the reads to a reference and quantify the genes that are present. Traditionally, this is done by aligning the reads to a reference genome (available online) and then counting genes. This can be computationally intensive. A newer technique for quantification is pseudoalignment. In this approach, reads are quantified using a reference transcriptome, which is converted into a special graph of *k*-mers (nucleotide strings of length *k*) derived from transcripts. *k*-mers in the read being mapped are then compared to this graph to determine related transcripts. Pseudoalignment with Kallisto, a tool that is freely available online, has been shown to be at least comparably accurate

to other methods of read mapping but is significantly faster than other approaches.[8] Kallisto outputs information regarding transcript abundances that can be used to perform DE testing in order to determine which transcripts or genes are differentially expressed between two conditions. In addition to identifying genes that are differentially expressed, functional enrichment can be performed to identify clusters of genes or pathways that are differentially represented across samples. A basic DE analysis is appropriate for an introductory genomics analysis due to the abundance of publicly available RNA-seq data as well as the feasibility of performing the analysis on a personal computer.

1.2 A Public Health Context for a DE Analysis: Tuberculosis Pathogenesis

Tuberculosis (TB) provides an interesting public health context for walking through an introductory genomics analysis. TB is an infectious disease caused by the bacterium *M. tuberculosis* (Mtb) and is transmitted through the air by individuals with an active pulmonary infection when they cough.[9] Symptoms associated with pulmonary infection include coughing, fever, weight loss and fatigue.[10] Most individuals infected with Mtb have a latent TB infection (LTBI). While LTBI patients do not transmit disease, they can develop active disease spontaneously due to a variety of causes such as immunosuppression or coinfection with HIV.[11]

There is a significantly large burden of disease for TB. Over a fifth of the world's population is thought to be infected with Mtb (most of whom have LTBI) and approximately 1.6 million people die annually from TB worldwide. The vast majority of cases occur in Africa, South-east Asia and India with incidence rates highest in countries such as South Africa, the Central African Republic and the Philippines.[12]

Two important aspects of global TB control are identification of active cases of disease and prevention of the development of active disease in those already infected. This latter approach is especially important as reactivation of LTBI is the main cause of active disease and because treating cases of both LTBI and active infection is necessary to meet global targets for the reduction of TB incidence.[13] While treatment regimens exist for both LTBI and active infections, there are diagnostic challenges. For example, a diagnosis of active TB is made through a clinical assessment of symptoms and a chest X-ray supplemented by tests such as sputum smear microscopy and sputum culture. However, sputum smear microscopy is only sufficiently sensitive when a large number of bacteria are present in the sputum sample and culture-based tests can be expensive and take weeks to provide results, during which time cases may actively transmit disease unless treated for active infection.[14] A robust and inexpensive test for detecting progression of LTBI to active infection would thus be helpful for reducing incidence. A DE analysis of publicly available RNA-seq data from patients with LTBI and active infection could be used to identify potential biomarkers for a blood test using certain peripheral blood mononuclear cells (PBMCs) that might be more sensitive than traditional microscopy and culture-based tests. This essay aims to first present an introductory workflow for analyzing publicly available RNA-seq data on a personal computer and then to provide an illustrative example in the context of progression of LTBI to active infection.

2.0 Methods

The general steps of a differential expression analysis of publicly available data include obtaining RNA-seq data from an online repository, assessing the quality of the reads, mapping the reads to either a reference genome or transcriptome, quantifying abundance of genes or transcripts and performing differential expression testing to see which genes are differentially expressed. A workflow including these steps was developed.

Because traditional alignment of reads to a genome is computationally intensive and not feasible on a personal computer, pseudoalignment was selected as the approach for read mapping and quantification. Kallisto, a widely utilized, simple-to-use, free and cross-platform pseudoalignment tool was preselected for the read mapping step of the workflow. The software to use for the remaining steps of the workflow was selected according to the following criteria: ease of access to the software, simplicity of software environment and ease of use. Overall, the workflow was developed to be as simple as possible to execute and to be usable with any data available on the SRA. An additional criterion for the workflow was that the analysis be performed at the level of genes and not transcripts.

A simple set of scripts for performing analysis using the workflow was prepared for Debian-based Linux distributions and macOS. This was created as an archive containing three scripts (two shell scripts and one R script) for performing the analysis as well as directories for storing data that are downloaded and results that are outputted. Instructions and scripts for performing this analysis were additionally prepared for Windows but require manual execution of programs from the command-line by the user.

Publicly available RNA-seq data from patients with LTBI and active infection were then analyzed using the workflow to provide an illustrative example. The results from this analysis are discussed with regard to how they can be applied in a public health and medical context. Additionally, the opportunity for professional development from performing analyses like these for a public health student or professional is discussed.

3.0 Results

A simple workflow for differential expression analyses was developed according to the criteria discussed earlier (Figure 1). Archives containing scripts needed to run an analysis are provided for Debian-based Linux distributions, macOS and Windows as supplementary materials. Links for these materials are provided in Appendix B. The methods detailed in this section are presented for Linux and macOS. Specific instructions for carrying out this analysis are provided in the README included in the archive provided for each of the three major operating systems that are supported.

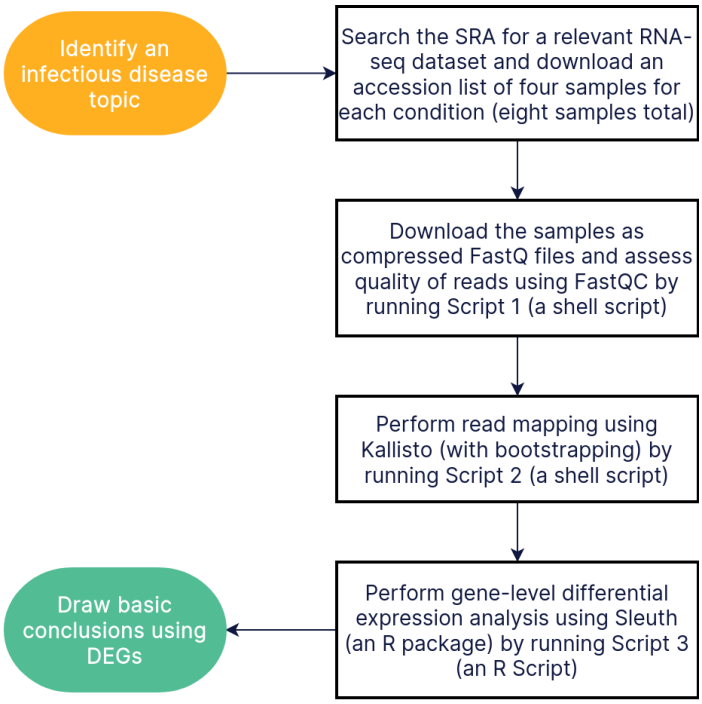

Figure 1 Diagram of Workflow on Linux/macOS

Extracting the archive provides a project directory in which the analysis is performed. The README in this directory contains instructions for installing the required software (SRA toolkit, FastQC, Kallisto, R and RStudio) and for performing the analysis.[8,15–17] One important note is that Kallisto is downloaded as an executable and placed in the directory named "programs". While a system-wide installation would be more appropriate for a pipeline, this is a simple way to run Kallisto on any platform. In the workflow, an infectious disease topic of interest is first chosen. The SRA is then searched for a study with RNA-seq runs of samples from *Homo sapiens* that relates to this topic. Four samples for both a control group (e.g. non-infected samples) and experimental group (e.g. infected samples) are then selected in the SRA Run Selector. An Accession List, or text file containing a list of the run IDs, can then be downloaded on this page. The Accession List is placed in the root of the project directory. Script 1, a shell script, is then run to first download gzip-compressed FastQ files from the SRA and then perform quality checks of reads using FastQC. Quality can be manually assessed by opening the HTML file outputted by FastQC for each set of reads in a browser and looking at per-base sequence quality as well as the summary of quality checks on the left-hand side. Samples that exhibit poor quality (for example having low per-base quality or a poor distribution of GC content) can be discarded before proceeding. Discarded samples should be removed from the data folder and deleted from Script 2.

Script 2 (another shell script) uses Kallisto to first create an index from a reference transcriptome for *Homo sapiens* (available online from Ensembl) and then perform read mapping using the reads and this index. For read mapping, Kallisto is run for each sample with arguments for performing bootstrapping. When bootstrapping is performed, a random subset of reads is repeatedly sampled from each set of reads and then mapped to estimate uncertainty on transcript abundances, even when there is only one replicate for the sample. Running Kallisto without

bootstrapping is fast; mapping is complete within minutes on a typical personal computer. Bootstrapping increases the processing time per sample but is required for the downstream analysis that was selected for this workflow. The output of Kallisto is information regarding transcript abundances as well as statistical information about abundance uncertainty. Samples that do not have at least 75% of reads mapping back to a reference should be discarded at this stage. This information is available in logs generated by running Kallisto using the provided scripts.

The next step in the workflow is to perform differential expression testing. This is done by running Script 3, an R script, from within the RStudio IDE. This script utilizes sleuth, an R package for differential expression testing, to perform two separate tests using abundance data. Sleuth was selected for this step because it automatically filters out low abundance transcripts and normalizes abundances, which reduces the number of steps necessary for the overall analysis. First, a likelihood ratio test is performed to determine which genes are differentially expressed at a significant level. This test compares a full model containing the experimental condition as a variable and a reduced model that does not containing the experimental condition as a variable. Performing this test provides adjusted p-values (or q-values) for each transcript that can be automatically aggregated to gene-level q-values.[18] A Wald test using sleuth in gene-mode is then performed to determine the extent to which genes are differentially expressed. Here, gene abundances are aggregated to determine gene-level expression changes. The beta coefficients from this model are estimates of log2 fold-change (log2fc) for each gene. The log2fc value represents the log-transformed ratio of abundance of a gene in the experimental condition to abundance in the control condition. While it is cumbersome to perform two separate tests, this allows for the determination of both gene-level p-values and gene-level estimates of fold-change using sleuth.

The results from these tests can be used to explore which genes are differentially expressed. Script 3 provides a formatted table of the topmost differentially expressed genes (DEGs). A principal component analysis (PCA) plot (showing sample clustering), MA plot (of log2fc estimate versus mean abundance) and volcano plot (of significance versus log2fc estimate) are provided. The final step of the workflow is to interpret these results. An example of this workflow, including interpretation of results, is provided in the following section.

3.1 Illustrative Example: TB Disease Progression

The algorithm presented in the previous section was used to analyze data related to the chosen public health issue of tuberculosis disease progression. This analysis was performed on Debian using a Lenovo ThinkPad X1 Carbon Gen 9 (Intel Core i5 with 8 threads, 16GB RAM). Instructions for performing this analysis on Linux, macOS and Windows are provided as supplementary materials.

An RNA-seq dataset was identified in the SRA for an experiment performed at University College London that sequenced the transcriptome of CD14+ monocytes isolated from PBMCs of patients with LTBI and active TB infection. These data were obtained from paired-end sequencing of samples using an Illumina NextSeq 500. Although obtained from the SRA, this dataset was originally submitted to the EBI database. The Accession Code for this project is ERP116604. The data from this study do not appear to have been used in any published study. Four samples for each condition were selected for differential expression analysis. For all eight samples, monocytes had been incubated with purified protein derivative (PPD) in order to stimulate an immune response.

The Accession List for these samples was downloaded and Script 1 was run to download the data and assess quality of the reads. All of the samples exhibited good quality, and none were discarded.

Read mapping was then performed using Script 2. The average percentage of reads mapped to the reference transcriptome across all samples was 86.3% and the range was 83.1%-88.5%. No samples were discarded at this stage. Differential expression analysis was performed using Script 3. A PCA plot showing how LTBI (control) and active infection (experimental) samples cluster with regard to gene expression is shown in Figure 2. LTBI samples clustered much more closely together as compared to active infection samples, which were more variable. There are many potential reasons for this difference in variation. It is unclear whether all the samples from active patients were taken at a similar time point in disease progression or treatment. Additionally, there might be variation in host gene expression due to variation in the strain of Mtb.

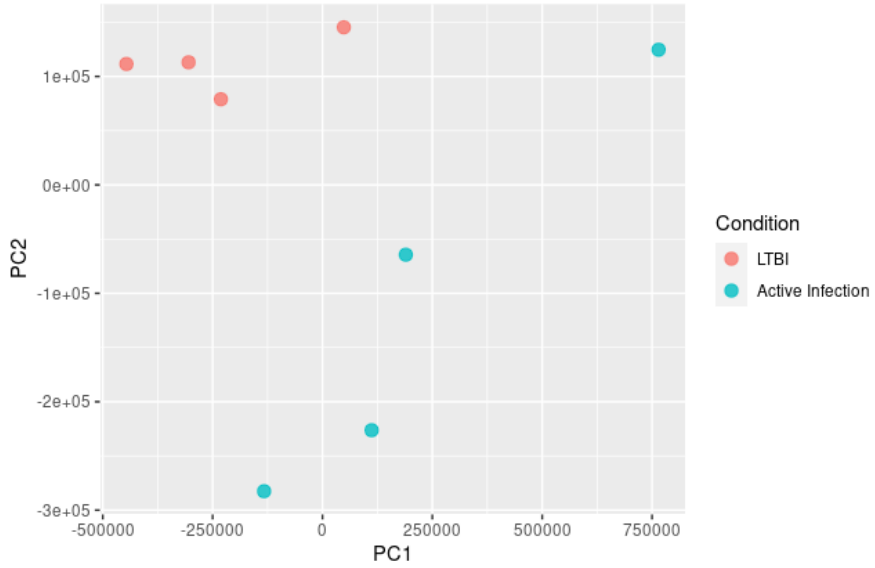

Figure 2 PCA Plot of Samples by Condition

Using a standard significance threshold of 0.05 for q-values, 526 genes were found to be significant. A list of DEGs was determined by filtering these genes for beta values (or log2fc estimates) with an absolute value greater than one. Table 1 shows the top DEGs (sorted by increasing q-value). Some of the DEGs include HLA-C (upregulated in active infection), caspase-1 (CASP1, upregulated in active infection), ATF3 (downregulated), clusterin (CLU, upregulated), interleukin-6 (IL6, upregulated), SLAMF7 (upregulated) and HLA-DQB1 (downregulated). A complete list of the 112 DEGs that were identified is shown in Table 2 (Appendix A).

Table 1 Top DGEs

Gene	q	beta	Gene	q	beta
HLA-C	1.617312e-23	6.885859	EPSTI1	4.965700e-05	1.086535
CASP1	6.649265e-14	1.142736	SLCO4A1	6.107973e-05	1.229731
SLC7A7	3.376703e-12	1.073208	GPR84	1.423342e-04	1.058340
CARD16	2.477350e-07	1.092130	STK38L	1.784130e-04	-1.160653
SLAMF7	4.833854e-07	1.241225	TFPI2	2.091147e-04	1.129504
CLU	2.482188e-06	2.670332	GBP5	2.091147e-04	1.448773
ATF3	2.482188e-06	-1.019846	PLAC8	2.627277e-04	1.557786
IL6	5.949102e-06	1.102287	GBP1	2.738659e-04	1.175454
FPR2	1.465902e-05	1.294572	FPR1	3.301515e-04	1.453429
LIMK2	3.300002e-05	1.221194	NCF1	3.506456e-04	1.705306
PSME2	4.207968e-05	1.432369	BCL2A1	3.738795e-04	1.034614
HLA-DQB1	4.576339e-05	-2.498258			

A volcano plot with top DEGs labeled is shown in Figure 3. This is a plot of significance (calculated as the negative base-ten log of the q-value) versus log2fc estimate. Genes are colored according to significance and differential expression; genes that are black are not significant and/or differentially expressed whereas genes that are blue or red are significantly downregulated or upregulated in active infection, respectively. Genes present in the top-right or top-left of the plot are the most differentially expressed and significant genes. Based on this plot, additional genes of

interest include MITF (downregulated), STK38L (downregulated), HLA-B (upregulated), FPR2 (upregulated) and NTSR1 (upregulated).

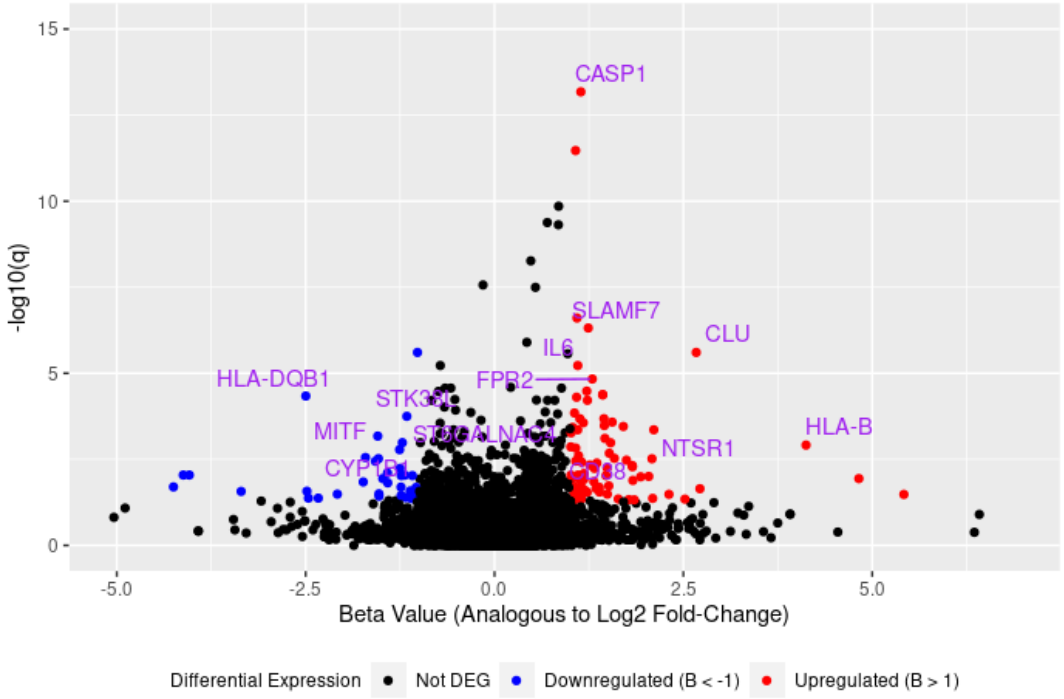

Figure 3 Volcano Plot

An MA plot of log2fc estimate versus mean gene abundance (across all samples) is presented in Figure 4. For this analysis, this plot is simply used to show that many of the important DEGs identified above were present in relatively high abundances (toward the right of the chart) and are thus less likely to be the result of noise. It also appears that more genes are upregulated in active infection than are downregulated.

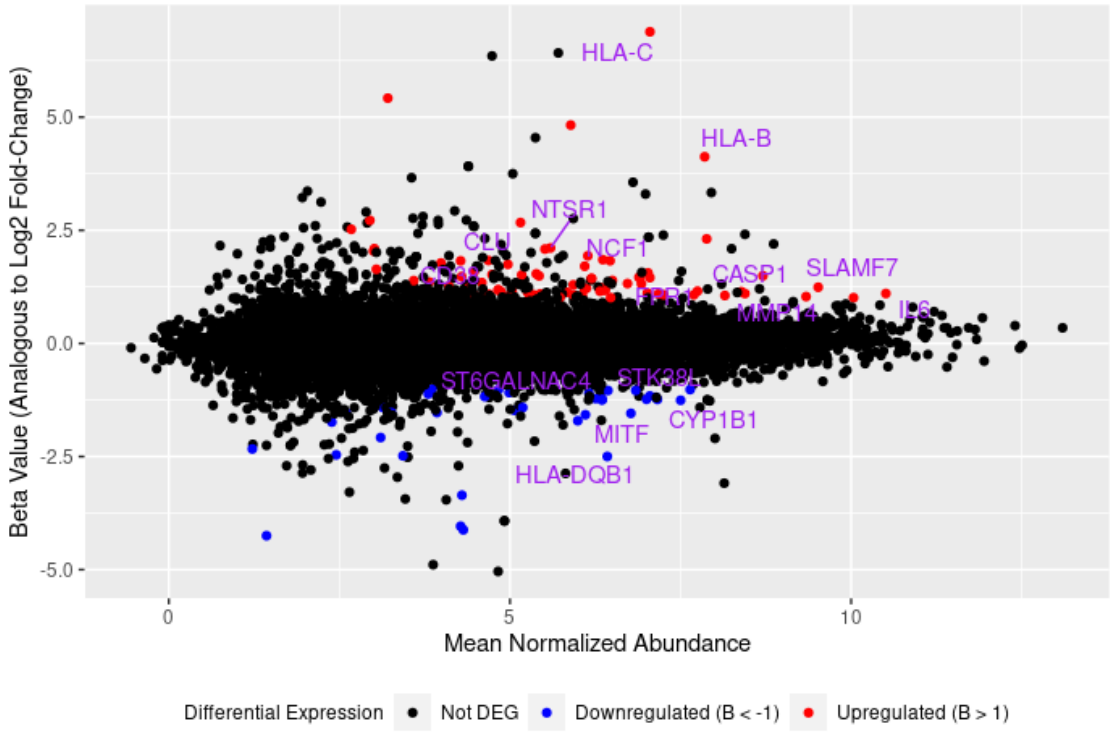

Figure 4 MA Plot

4.0 Discussion

4.1 Application of Results from the DE Analysis on TB Progression

Performing the workflow resulted in the identification of several genes that are differentially expressed by PPD-stimulated monocytes obtained from patients with active infection as compared to patients with LTBI. Some of these DEGs can be categorized. MHC class I genes such as HLA-B and HLA-C were found to be considerably upregulated. While some MHC class II genes such as HLA-DQA1 were downregulated, HLA-DRB4 was found to be highly upregulated. This may warrant further analysis as macrophages use class II presentation for TB antigens and Mtb has been shown to inhibit transcription of MHC class II molecules.[19] Clusterin, found here to be upregulated in active infection, has been previously shown to be upregulated in patients with active TB as compared to non-infected patients following a Mtb antigen stimulus.[20]

Results from analyses like this one could be used to inform the monitoring and diagnosis of active TB. Some of the DEGs identified here could be evaluated as markers of disease progression. A combination of robust markers could be utilized in a blood test that incubates monocytes from patients with PPD to identify cases of active disease. A test like this could be used to monitor patients suspected to have LTBI but who begin to develop symptoms. This test would have to be both highly predictive of active infection as compared to LTBI and similar in cost to sputum smear tests (which are used to inexpensively identify infectious cases in LMIC) to justify its use. This likely presents a challenge due to the cost of isolating certain leukocytes from human blood and measuring a signature of markers. However, preventing progression of disease with timely treatment for active TB may be valuable in locations where TB is not endemic and where

there is the potential for an outbreak. Additionally, timely treatment of active TB can prevent the spread of drug-resistant TB.[21]

There is an abundance of publicly available RNA-seq data for a variety of topics (and organisms). Searching the SRA with keywords such as “tuberculosis” and “rna-seq” yields many RNA-seq datasets for *Homo sapiens*. The analysis presented here could be performed on datasets related to either the same topic (to validate DEGs that are discovered) or a different TB topic, such as drug-resistance or coinfection with HIV. In addition to manually reviewing DEGs, more advanced analyses such as functional enrichment can be performed to see which pathways are differentially represented across biological conditions to further explore these topics.

Results from transcriptomics analyses like the one presented here are not generally conclusive on their own. Results for genes that are differentially expressed can be confirmed using techniques like RT-PCR (to quantify mRNA) or antibody-based assays (to quantify the protein product of the gene). DEGs can be knocked down, for example, to assess functional importance.

4.2 Opportunities for Professional Development

Gaining experience with this workflow can be beneficial from an educational standpoint. Performing this analysis provides opportunities to gain experience with R (for use in statistical analysis, plotting and genomics), command-line interface tools and interpretation of results. These skills are directly applicable in many fields other than genomics. More broadly, this analysis serves as an introduction to powerful genomics analyses that can be performed on personal computers using publicly available data. Without needing access to a lab, one can analyze existing data and obtain potentially important results that could be further explored in a lab. This is a low-barrier

approach to gaining biological research-related experience for someone who may not work in the field of genomics but who is interested in expanding their skillset. Additionally, this experience is increasingly useful and relevant with the growing abundance of genomics data and use of genomics in medicine and public health.

The cost of sequencing genetic material continues to decline. Sequencing with the Oxford Nanopore Technologies MinION (a third-generation sequencing platform) is relatively inexpensive compared to next generation sequencing platforms, which are still the most used platforms for generating RNA-seq reads. A starter pack for sequencing with the MinION costs approximately $5000.[22] RNA-seq reads generated from the MinION can be analyzed for differential expression using the workflow presented here. Investigators could use this workflow to gain familiarity with this type of RNA-seq analysis pipeline before they begin collecting and sequencing more data.

This workflow requires some manual effort (such as executing shell scripts) and provides only a basic differential expression analysis. Despite this, it is a good starting point for those interested in exploring genomics. This experience can be expanded upon by setting up a proper pipeline for differential expression analyses. There are several courses and tutorials available online that walk through this process.[23,24]

4.3 Evaluation of Approach and Future Directions

This workflow was used to identify a handful of potentially important genes that are differentially expressed in monocytes from patients with active infection as compared to LTBI. Some potential applications of these DEGs as well as future analyses were discussed. While this

workflow allowed for a walkthrough of differential gene expression analysis, there were a few limitations. Sleuth was selected for differential gene testing because it automatically handles filtering of low abundance transcripts and normalization of abundances. It has also been recommended for analysis of data outputted by Kallisto for statistical and performance reasons.[25] While sleuth simplified part of the analysis, the downstream analysis was more complicated than it is with alternative software. It was necessary to perform two separate tests: one to determine gene-level p-values and another to determine gene-level log2fc estimates. Additionally, only estimates for log2fc and not the actual values for log2fc are provided. For educational or illustrative purposes, it may be preferable to use an alternative to sleuth, such as edgeR and limma.[26,27] Additionally, using an alternative to sleuth to analyze abundance data from Kallisto removes the need to perform bootstrapping, allowing for a much faster analysis that can be performed more feasibly on a larger number of samples. Despite these drawbacks, the algorithm presented here introduces genomics analyses and can provide the familiarity needed to set up a proper pipeline for analysis.

This workflow is best carried out on Linux or macOS, but general instructions for Windows are provided. The two shell scripts that are executed for Linux and macOS could be executed from an R script to further simply the workflow. For Windows, downloading data from the SRA, checking the quality and performing read mapping are currently executed through manual commands which could be placed in batch scripts. In addition to simplifying this workflow for the target audience, a better illustrative analysis could be performed using a dataset that clusters better on a PCA plot and that yields results that are more easily interpreted (or validated with existing literature).

5.0 Conclusion

A workflow for analyzing publicly available RNA-seq data related to an infectious disease topic was presented. This workflow was illustrated using a dataset related to tuberculosis pathogenesis. While basic applications of genes found to be differentially expressed in monocytes from active patients as compared to those from patients with LTBI were discussed with regard to monitoring disease progression, further analysis is needed to draw useful conclusions about changes in gene expression across conditions. While there are certain drawbacks to this workflow, it introduces complex genomics analyses that can be performed on a personal computer with publicly available data related to a public health topic of interest or to expand one's genomics skillset.

Appendix A List of All Differentially Expressed Genes

Table 2 Complete List of DEGs

Gene	q	beta	Gene	q	beta	Gene	q	beta
HLA-C	1.62E-23	6.885859	KREMEN1	0.004622	1.260569	FFAR2	0.031795	1.15546
CASP1	6.65E-14	1.142736	DLL1	0.004805	1.823705	MEIKIN	0.031795	-1.52992
SLC7A7	3.38E-12	1.073208	CSF3	0.005209	1.825888	GNG11	0.032422	1.162129
CARD16	2.48E-07	1.09213	NCF1B	0.005652	1.491374	PDK4	0.032504	-2.08057
SLAMF7	4.83E-07	1.241225	APOBEC3A	0.005787	1.322295	TNFAIP6	0.032504	1.493944
CLU	2.48E-06	2.670332	LPAR1	0.005798	-1.25084	ACOD1	0.032934	2.309641
ATF3	2.48E-06	-1.01985	ARL4C	0.007889	-1.41709	IRF4	0.032934	-1.03619
IL6	5.95E-06	1.102287	DUSP16	0.008421	1.182409	HLA-DRB4	0.033141	5.419455
FPR2	1.47E-05	1.294572	UBE2L6	0.009058	1.106188	KCNE1	0.035476	-1.23619
LIMK2	3.3E-05	1.221194	BRD2	0.009058	-4.1204	MAP1LC3A	0.035955	1.093917
PSME2	4.21E-05	1.432369	PSMB10	0.00914	1.011261	SPRED1	0.037249	-1.52857
HLA-DQB1	4.58E-05	-2.49826	ME1	0.009167	-1.18862	ZNF304	0.040736	-1.17551
EPSTI1	4.97E-05	1.086535	CRTAM	0.009403	-1.23531	DDX11L2	0.04251	-2.46348
SLCO4A1	6.11E-05	1.229731	IL1RL2	0.009403	-1.08743	ARHGAP23	0.04251	-2.33272
GPR84	0.000142	1.05834	FCGR1B	0.009403	1.476783	NEURL3	0.043228	2.093606
STK38L	0.000178	-1.16065	WNT5B	0.009927	2.041729	SLC26A2	0.043901	-1.09808
TFPI2	0.000209	1.129504	SIGLEC5	0.010188	1.934865	SLPI	0.044953	1.637855
GBP5	0.000209	1.448773	GLIS3	0.011485	-1.47881	TIMP4	0.045782	-1.01856
PLAC8	0.000263	1.557786	RIPOR2	0.012134	1.182985	SECTM1	0.045941	1.076528
GBP1	0.000274	1.175454	CLEC2B	0.012194	1.029523	FCGR1A	0.046829	1.146591
FPR1	0.00033	1.453429	GYPC	0.012966	1.143129	ANKRD22	0.046829	1.778761
NCF1	0.000351	1.705306	VAMP5	0.012966	1.831899			
BCL2A1	0.000374	1.034614	IFITM3	0.014135	1.314693			
NTSR1	0.000439	2.108529	ALPK2	0.014394	-1.73783			
MMP14	0.000439	1.101101	ST6GALNA(	0.015282	-1.41369			
ADCY3	0.000538	1.013593	PLAAT4	0.018681	1.513228			
MITF	0.00067	-1.54574	SLC2A3P1	0.020051	-4.24986			
CTSC	0.00078	1.461136	LRRC8B	0.0203	-1.23792			
ST6GALNA(	0.001036	-1.22024	ISG15	0.020584	1.077859			
C1QTNF1	0.001045	1.536859	TMEM255E	0.02068	1.388			
HLA-B	0.001222	4.124176	IL15RA	0.020745	1.15268			
S100A8	0.001375	1.009528	RCAN1	0.020745	-1.03778			
ADGRB1	0.001489	1.074354	NKX3-1	0.022252	1.343542			
CYP1B1	0.001657	-1.25645	APOBEC3B	0.022613	2.719265			
CD38	0.002105	1.519911	NKG7	0.024451	1.041			
IFI6	0.00249	1.112589	TCN2	0.025542	1.199145			
RAB7B	0.002801	-1.7086	DCUN1D4	0.025556	-1.01766			
CLEC4D	0.002961	1.586334	PIKFYVE	0.025556	-1.04412			
TINAGL1	0.003021	-1.53657	APOL3	0.025662	1.054998			
TREML4	0.003049	2.084673	PKD1P4	0.026246	-1.11651			
C2	0.003346	1.746451	HLA-DQA1	0.027073	-3.35413			
GNG2	0.003571	1.074492	IL36G	0.028271	1.390844			
AKR1C1	0.003611	-1.57696	IL6ST	0.028881	-1.04023			
PPP1R18	0.003822	1.153248	MCTP2	0.028881	1.226225			
GBP4	0.004054	1.354444	DNAJC15	0.031667	1.052965			

Appendix B Scripts to Run Illustrative DE Analysis

1. Archive of scripts for Linux:

   http://d-scholarship.pitt.edu/42898/1/de_example_linux_v0.1.tar.gz

2. Archive of scripts for macOS:

   http://d-scholarship.pitt.edu/42898/2/de_example_macOS_v0.1.tar

3. Archive of scripts for Windows:

   http://d-scholarship.pitt.edu/42898/3/de_example_win_v0.1.zip

Bibliography

1. Lam C, Martinez E, Crighton T, et al. Value of routine whole genome sequencing for Mycobacterium tuberculosis drug resistance detection. *Int J Infect Dis*. 2021;113:S48-S54. doi:10.1016/j.ijid.2021.03.033

2. Rivière E, Heupink TH, Ismail N, et al. Capacity building for whole genome sequencing of Mycobacterium tuberculosis and bioinformatics in high TB burden countries. *Brief Bioinform*. 2021;22(4):bbaa246. doi:10.1093/bib/bbaa246

3. Mellmann A, Harmsen D, Cummings CA, et al. Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS ONE*. 2011;6(7):e22751. doi:10.1371/journal.pone.0022751

4. Abul-Husn NS, Soper ER, Braganza GT, et al. Implementing genomic screening in diverse populations. *Genome Med*. 2021;13(1):17. doi:10.1186/s13073-021-00832-y

5. National Institutes of Health. NIH's All of Us Research Program Releases First Genomic Dataset of Nearly 100,000 Whole Genome Sequences. National Institutes of Health (NIH). Published March 17, 2022. Accessed April 24, 2022. https://www.nih.gov/news-events/news-releases/nih-s-all-us-research-program-releases-first-genomic-dataset-nearly-100000-whole-genome-sequences

6. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *WIREs RNA*. 2017;8(1). doi:10.1002/wrna.1364

7. Muir P, Li S, Lou S, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol*. 2016;17(1):53. doi:10.1186/s13059-016-0917-0

8. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

9. Churchyard G, Kim P, Shah NS, et al. What We Know About Tuberculosis Transmission: An Overview. *J Infect Dis*. 2017;216(Suppl 6):S629-S635. doi:10.1093/infdis/jix362

10. Loddenkemper R, Lipman M, Zumla A. Clinical Aspects of Adult Tuberculosis. *Cold Spring Harb Perspect Med*. 2016;6(1):a017848. doi:10.1101/cshperspect.a017848

11. Kiazyk S, Ball T. Latent tuberculosis infection: An overview. *Can Commun Dis Rep*. 2017;43(3-4):62-66.

12. Global tuberculosis report 2021. Accessed April 24, 2022. https://www.who.int/publications-detail-redirect/9789240037021

13. Chee CBE, Reves R, Zhang Y, Belknap R. Latent tuberculosis infection: Opportunities and challenges. *Respirol Carlton Vic*. 2018;23(10):893-900. doi:10.1111/resp.13346

14. Desikan P. Sputum smear microscopy in tuberculosis: Is it still relevant? *Indian J Med Res*. 2013;137(3):442-444.

15. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Accessed April 29, 2022. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

16. The Comprehensive R Archive Network. Accessed April 29, 2022. https://cran.r-project.org/

17. RStudio | Open source & professional software for data science teams. Accessed April 29, 2022. https://www.rstudio.com/

18. Yi L, Pimentel H, Bray NL, Pachter L. Gene-level differential analysis at transcript-level resolution. *Genome Biol*. 2018;19(1):53. doi:10.1186/s13059-018-1419-z

19. Harding CV, Boom WH. Regulation of antigen presentation by Mycobacterium tuberculosis: a role for Toll-like receptors. *Nat Rev Microbiol*. 2010;8(4):296-307. doi:10.1038/nrmicro2321

20. Tanaka T, Sakurada S, Kano K, et al. Identification of tuberculosis-associated proteins in whole blood supernatant. *BMC Infect Dis*. 2011;11:71. doi:10.1186/1471-2334-11-71

21. Fox GJ, Schaaf HS, Mandalakas A, Chiappini E, Zumla A, Marais BJ. Preventing the spread of multidrug-resistant tuberculosis and protecting contacts of infectious cases. *Clin Microbiol Infect*. 2017;23(3):147-153. doi:10.1016/j.cmi.2016.08.024

22. Product comparison. Oxford Nanopore Technologies. Accessed April 28, 2022. http://nanoporetech.com/products/comparison

23. DIY.transcriptomics – RNAseq course. Accessed April 24, 2022. https://diytranscriptomics.com/

24. Skidmore Z. Intro to RNAseq Analysis. Griffith Lab. Published January 1, 6AD. Accessed April 24, 2022. https://pmbio.org/module-06-rnaseq/0006/01/01/Intro_to_RNAseq_Analysis/

25. A sleuth for RNA-Seq. Bits of DNA. Published August 17, 2015. Accessed April 24, 2022. https://liorpachter.wordpress.com/2015/08/17/a-sleuth-for-rna-seq/

26. Bioconductor - edgeR. Accessed April 29, 2022. https://bioconductor.org/packages/release/bioc/html/edgeR.html

27. Smyth G, Hu Y, Ritchie M, et al. *Limma: Linear Models for Microarray Data*. Bioconductor version: Release (3.15); 2022. doi:10.18129/B9.bioc.limma