Transcriptomics in R practical (Report) | UVic-UCC, 2023-24

Vadim Leonov

2024-01-18

Introduction

Tuberculosis (TB), instigated by the bacterium Mycobacterium tuberculosis (M.tuberculosis), stands as a prominent threat among infectious diseases. It predominantly impacts the respiratory system, extending its reach to other bodily organs, leading to instances of extrapulmonary tuberculosis. Macrophages (Mp) provide a niche for the persistent infectious agent, often referred to as intracellular or intramacrophage infection. This practical exercise report focuses on analyzing the dataset GSE162729 obtained from the research project with the title 'Comparative Transcriptomic Analysis of THP-1 Derived Mp infected with M. tuberculosis H37Rv, H37Ra, and BCG,' published on Dec 06, 2020, available at https://www.ncbi.nlm.nih.gov. The study investigates the immune responses induced by different genotypes of M.tuberculosis strains in macrophages through transcriptomic analysis (Pu et al., 2021).

Materials and Methods

Summary of the Study from Gene Expression Omnibus (GEO): RNA sequencing was employed to investigate global transcriptome changes in Mp during different strains of M. tuberculosis infection. THP-1 cell-derived Mp were exposed to the virulent M. tuberculosis strain H37Rv, the avirulent M. tuberculosis strain H37Ra, and the M. tuberculosis BCG vaccine strain as a control. cDNA libraries were prepared from M. tuberculosis-infected Mp and subsequently sequenced. Detailed laboratory procedures and RNA sequencing analyses are available in the original paper by Pu et al., 2021. Here, we focus on bioinformatics and statistical analysis.

Dataset Download: Gene expression data were collected from the GEO database (GSE162729) based on the GPL11154 Illumina HiSeq 2000 platform (Homo sapiens). The dataset comprised gene expression profiles from 8 samples and 4 controls.

Data Processing: The raw count matrix, initially containing 19965 rows and 12 columns, underwent filtering to exclude genes with fewer than 10 reads in at least 6 samples. Gene information was integrated into the filtered count matrix by matching gene symbols obtained from Ensembl, including HGNC symbols. Duplicate entries were removed, and missing values related to read counts were replaced with zeros. Following these procedures of filtering, annotation, and trimming, the dataset had dimensions of 13196x12. Gene expression data were normalized for read depth, scaled by gene length to obtain fragment per kilobase (FPKM), and transcripts per million (TPM) were calculated. Total counts of the TPM-normalized data were checked, followed by TMM (Trimmed Mean of M-values) normalization for aggregation analysis.

Differential Expression Analysis: The preprocessed data were transformed for linear modeling, allowing the comparison of different experimental conditions. Specific comparisons (contrast matrix) were defined and applied, and Bayesian adjustment was employed. Genes with an absolute log-fold change (logFC) > 1 and a p-value < 0.01 were considered differentially expressed. Upregulated and downregulated DEGs were identified based on logFC thresholds. Volcano plots and heatmaps were created to visualize 20 significant

DEGs and their expression in macrophages under different conditions, with colors representing up- and downregulated genes. The heat maps represent the differential gene expression patterns between uninfected and infected Mp with BCG (con1), H37Ra (con2), or H37Rv (con3) strains. Each row in the heat map corresponds to a gene, and the color scale indicates the magnitude of expression changes. Warmer colors represent upregulated genes in the condition groups compared to the control, while cooler colors indicate downregulated genes.

Software and Packages: No private code was used; all software utilized was publicly or commercially available. R (v3.4.1)/BioConductor packages EdgeR (v2.20.2) and Limma (v3.34.4) were employed to adjust counts for library size differences. Data were loaded using the GEOquery package. DEGs were calculated using the edgeR package. Hierarchical clusterings and Principal Component Analysis (PCA) were performed in R using 'hclust' and 'ggfortify', respectively. Heatmap and scatterplot visualizations were generated using the 'heatmap' and 'ggplot2' libraries in R (Wickham, 2016).

Results and Discussion

RNA-seq analysis of samples from both uninfected negative control and infected cells resulted in an average of 45 million reads, all of which were successfully mapped to the human genomes. PCA unveiled distinct clustering patterns: uninfected Mp formed a separate cluster from infected cells, and samples from BCG-infected cells exhibited unique clustering. Furthermore, samples from infected H37Rv and H37Ra cells displayed notable dissimilarity, with clear segregation between them, as depicted in Figure 1 (Appendix). To further investigate these observed differences, a differential expression analysis was conducted.

All results from the analysis of DEG analysis are presented in volcano plots and heatmaps, which can be found in the appendix (Figure 2-4) to the report. It is important to note that each volcano plot and heat map focuses on the comparison of the controls (Ctr) with a specific condition (TB strain) according to the contrast matrix, which encompasses three conditions. This allows a comprehensive exploration of gene expression changes across various experimental conditions.

In response to M. tuberculosis BCG infection, a comprehensive transcriptomic analysis revealed a total of 4553 differentially expressed genes (DEGs) compared with the untreated control. Among these, 2282 genes were upregulated, while 2271 were downregulated. The top 20 DEGs were highlighted and presented on volcano plot and heatmap (Figure 2). Similarly, in response to M. tuberculosis H37Ra infection, a total of 4883 DEGs were identified, with 2801 genes upregulated and 2082 downregulated. The top 20 DEGs were illustrated on a volcano plot and heatmap (Figure 3). For M. tuberculosis H37Rv infection, 3863 DEGs were identified, including 1984 upregulated and 1879 downregulated genes. The top 20 DEGs were visualized on a volcano plot and heatmap (Figure 4).

In terms of percentages, the Mp DEGs profiles show an equal proportion of 17.3% upregulated and downregulated genes under BCG infection. For the virulent strain H37Ra, there are 21.2% upregulated and 15.7% downregulated genes in Mp, while the H37Rv strain upregulates 15.0% and downregulates 14.2% of genes in Mp. By comparing the down- and upregulated genes between infected and uninfected Mp, we found that 6 (IL4I1, PSAT1, PHGDH,MSR1, PLSCR1, DDIT4L) and 2 (MT2A, STAC2) genes were shared between the conditions. Due to limitation of the data for our report and absence of pathway analysis definitive conclusion cannot be drawn. However, in contrast to BCG, H37Ra and H37Rv infection revealed a more extensive transcriptional pattern of several factors likely associated with virulence. Indeed, even without pathway enrichment analysis, it is evident that H37Rv predominantly deactivates inflammatory genes (downregulated genes) such as CXCL16, TGTBI, and IL4I1 to enhance its survival (virulence). In addition, all mycobacterial strains stimulate the STAC2 genes associated with phagocytosis.

In summary, we scrutinized the transcriptional responses of Mp infected with two widely recognized laboratory strains of M. tuberculosis (H37Rv and H37Ra), along with a vaccine strain of M. bovis (BCG). The obtained results delineate distinctive transcriptional features associated with DEGs induced by virulent strains of M. tuberculosis. To identify specific pathways for identified DEGs an enrichment analysis approach should be employed as well as compassion analysis of expression patterns of Mp between virulent and vaccine

strains of M.tuberculosis (monoconditional design matrix) might be useful for better understanding cellular parthenogenesis of mycobacterial infection.

References

Dataset Link: Pu, W., Zhao, C., Wazir, J., Su, Z., Niu, M., Song, S., … Wang, H. (2021). Comparative transcriptomic analysis of THP-1-derived macrophages infected with Mycobacterium tuberculosis H37Rv, H37Ra and BCG. Gene Expression Omnibus (GEO) accession GSE162729.

Journal Article: Pu, W., Zhao, C., Wazir, J., Su, Z., Niu, M., Song, S., … Wang, H. (2021). Comparative transcriptomic analysis of THP-1-derived macrophages infected with Mycobacterium tuberculosis H37Rv, H37Ra and BCG. Journal of Cell and Molecular Medicine, 25(22), 10504-10520. https://doi.org/10.1111/jcmm.16980

Book Chapter: Smyth, G. K. (2005). Limma: Linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, & W. Huber (Eds.), Bioinformatics and Computational Biology Solutions using R and Bioconductor (pp. 397–420). New York: Springer. https://doi.org/10.1007/0-387-29362-0_23 Book Chapter (EdgeR):

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edgeR. In S. Datta & D. S. Nettleton (Eds.), Statistical Analysis of Next Generation Sequence Data (pp. 51–74). New York: Springer. https://doi.org/10.1007/978-1-4939-0709-0_3

Book: Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4.

Appendix

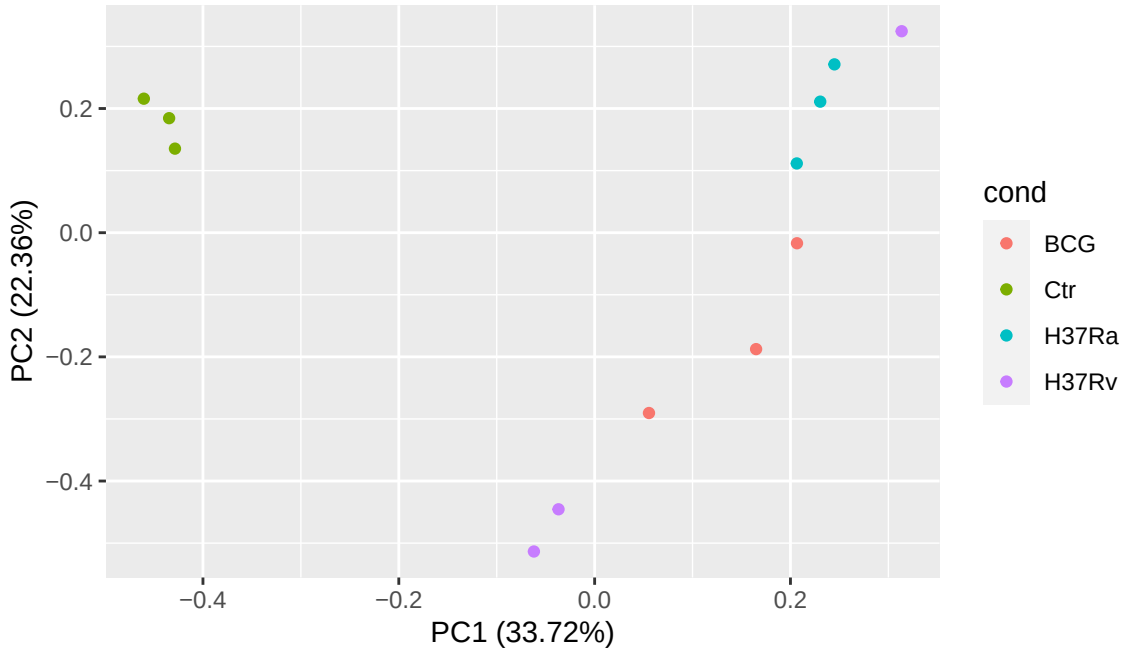

Figure 1. Results of Preincipal component analysis for uninfected and M. tuberculosis-infected THP-1 Macrophages

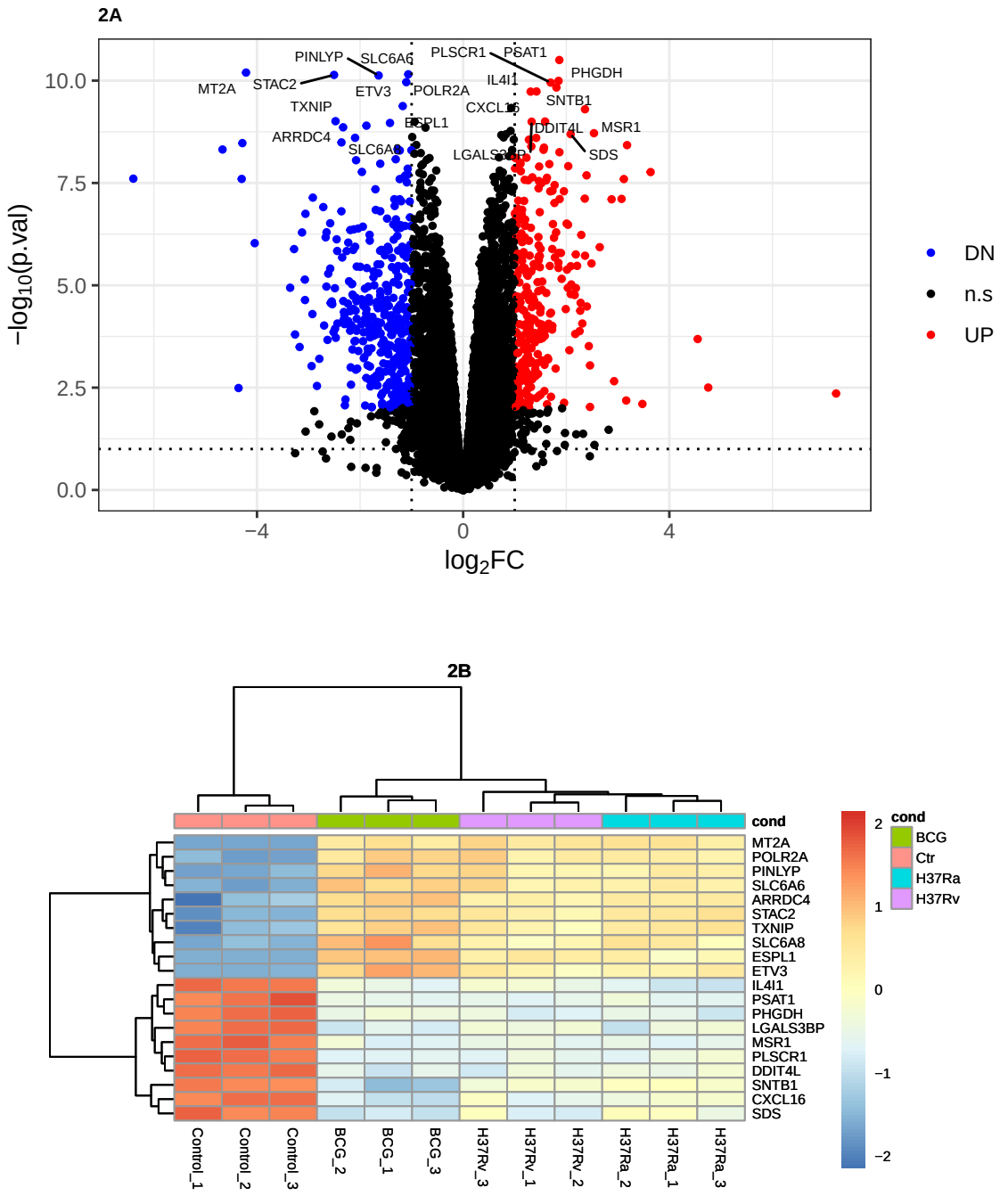

Figure 2. Differential Gene Expression Patterns for uninfected and BCG M. tuberculosis-infected THP-1 Macrophages. (2A) Volcano Plots: visualizing markedly up-regulated genes (depicted in red) and down-regulated genes (depicted in blue). (2B) Heatmaps: lllustrating the expression profiles of the 20 most significant differentially expressed genes.

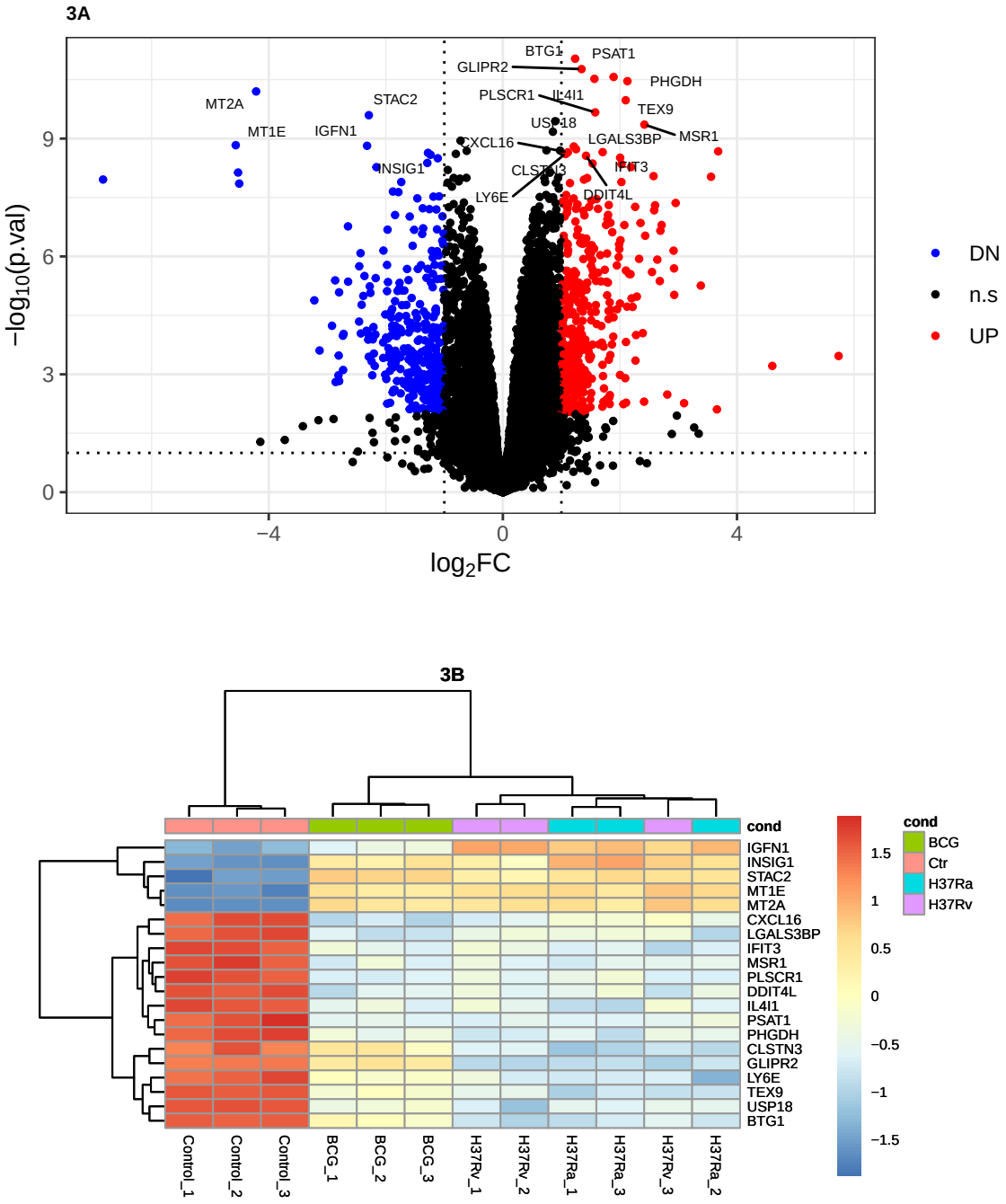

Figure 3. Differential Gene Expression Patterns for uninfected and H37Ra M. tuberculosis-infected THP-1 Macrophages. (3A) Volcano Plots: visualizing markedly up-regulated genes (depicted in red) and down-regulated genes (depicted in blue). (3B) Heatmaps: lllustrating the expression profiles of the 20 most significant differentially expressed genes.

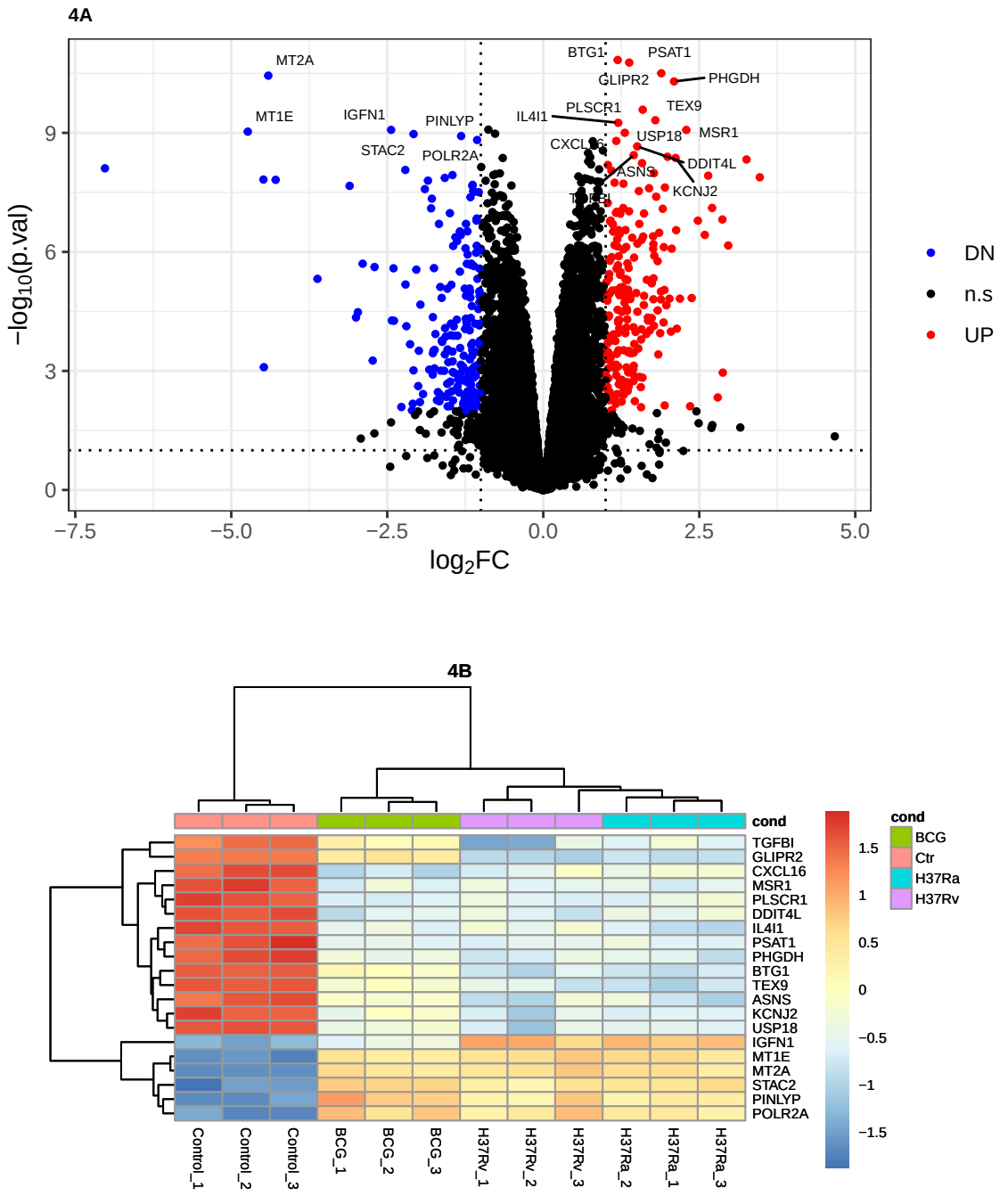

Figure 4. Differential Gene Expression Patterns for uninfected and H37Rv M. tuberculosis-infected THP-1 Macrophages. (4A) Volcano Plots: visualizing markedly up-regulated genes (depicted in red) and down-regulated genes (depicted in blue). (4B) Heatmaps: lllustrating the expression profiles of the 20 most significant differentially expressed genes.