Gene5150 Emily White

Transcriptomic profiling of mouse lung endothelial cells after SARS-CoV-2 infection

GOAL: Identify differentially expressed genes between control vs. SARS-CoV2 infected mouse epithelial lung cells and infer relevant biological pathways

Data Source & Purpose

Data Generation

To demonstrate clear pathophysiology (eliminate the confounding variable of age), samples were stratified between young and mid-aged mice. Tissue samples were biopsied from healthy controls and "treatment groups", i.e samples extracted post-viral infections.

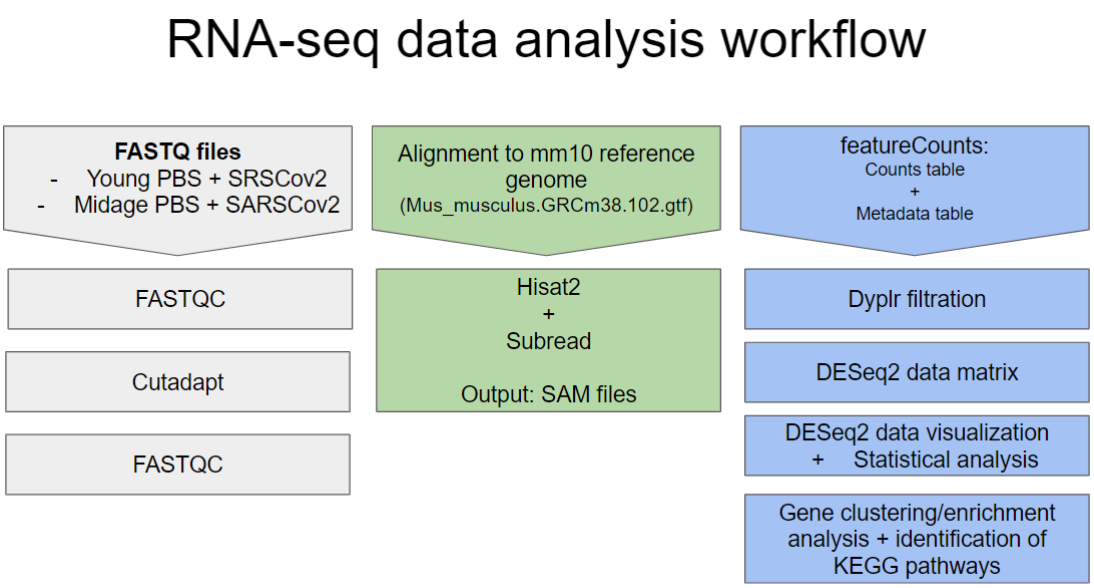
Overall, the study aimed to provide evidence of comparative gene expression profiling for the RNA-seq data between four different samples (uninfected/infected, young/mid-aged) of isolated, mouse lung endothelial cells (GEO Accession Viewer).

Potential challenges

An advantage of using the DEseq2 software package is that it internally corrects for library size; this means that it will normalize the RNA-seq data to accurately estimate read count distribution and quantification of gene expression. While useful, genes that demonstrate significant differential expression may be difficult to validate because the statistical inference is done internally by the software. Furthermore, there are many transcripts that were removed from the data due to having low (<10) read count; while this is statistically accurate, the assumption must be made that the gene is indeed absent across all four samples, and that the low read count wasn't caused by sequencing errors/artifacts. However, if the replicate of the initial experiment is known, the confidence of this assumption could increase significantly.

Potential outcomes & future directions

Due to the nature of SARS-Cov2 and viral infections, I hypothesize upregulation in RNA-seq counts from the post-SARSCov2 infection samples within genes/pathways involved in immune response.



Data Analysis

Preprocessing raw RNA sequences

Originally, 15 bp were removed from the beginning of every read to rid of low-quality reads; the FASTQ file indicated (Figure 1A the presence of adaptor sequences still attached to the beginning (5' end) of the reads). At this step, according to the FASTQC report (figure 1A), the reads were averaging the following metrics:

- sequence length of 300
- %GC content of 50
- Pre base sequence quality 35-36
- Total sequences: 5,000,000

However, when attempting to run the HISAT2 alignment followed by the generation of a counts table, there were many transcripts with low coverage (<10) read counts across all samples. This suggested that the alignment was requiring too high of read counts to pass the default threshold, or that the reads were too long and thus aligning poorly. Therefore, the parameters of cutadapt were changed, to remove 200 bp from the beginning of each RNA seq read. Once this was complete, the FASTQC report (Figure 1B) provided the following metrics:

- sequence length of 0-85
- %GC content of 51
- Pre base sequence quality 36
- Total sequences: 5,000,000

The Phred quality score of 36 qualifies this RNA-seq dataset to have >99.9% accuracy and ~ 0.055% error. In other words, there is less than a 1/1000 change of the incorrect base being called.

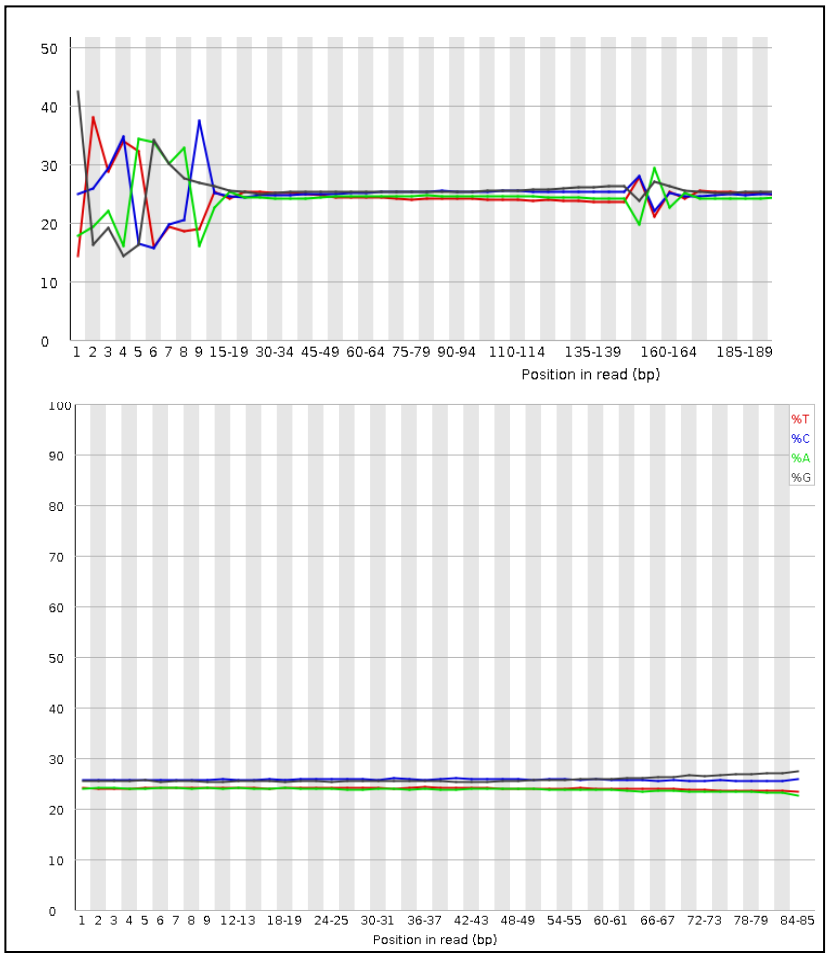


Figure 1A (top left): depicts the FASTQC per base sequence content for the Young SARSCov2 sample prior to pre-processing. The high variability between the base content in the first 1-20 base pairs indicates the presence of primer adaptor sequences.

Figure 1B (bottom left): demonstrates the consistent and narrow quality score of by base content (%T,C,A, or G); the narrow range in quality score indicates the samples demonstrate more consistence sequence content across all bases of the same read.

Alignment

HISAT2 alignment generated SAM files for each sample; the alignment summaries confirmed improved alignment rates. For example, after the cutadapt -200 bp processing step, the SAM alignment for the Midage_SARSCov2 sample improved from a 0.56% overall alignment (300 bp average read length) to 49.4% "successfully assigned alignments" (0-85 bp read lengths). Overall, this suggests that the HISAT2 aligner is more effective on shorter read lengths; this is likely because longer read lengths are likely to generate more "distinct" alignments, that do not pass the threshold quality or read count filters.

RNA-Seq Readcount filtration

The following transcripts passed each (read count) filtering step:

- Raw RNA-seq alignments to transcripts: 55,487
- After filtering for transcripts >0 read counts: 21,353
- After filtering for transcripts >10 read counts: 14,376

Gene annotation

GO ontology enrichment applied *m. musculus* gene annotation to each transcript in the dataset, using org.Mm.eg ENSEMBL, an R object that contains mappings between Entrez Gene identifiers and GenBank accession numbers.

Gene Expression Analysis

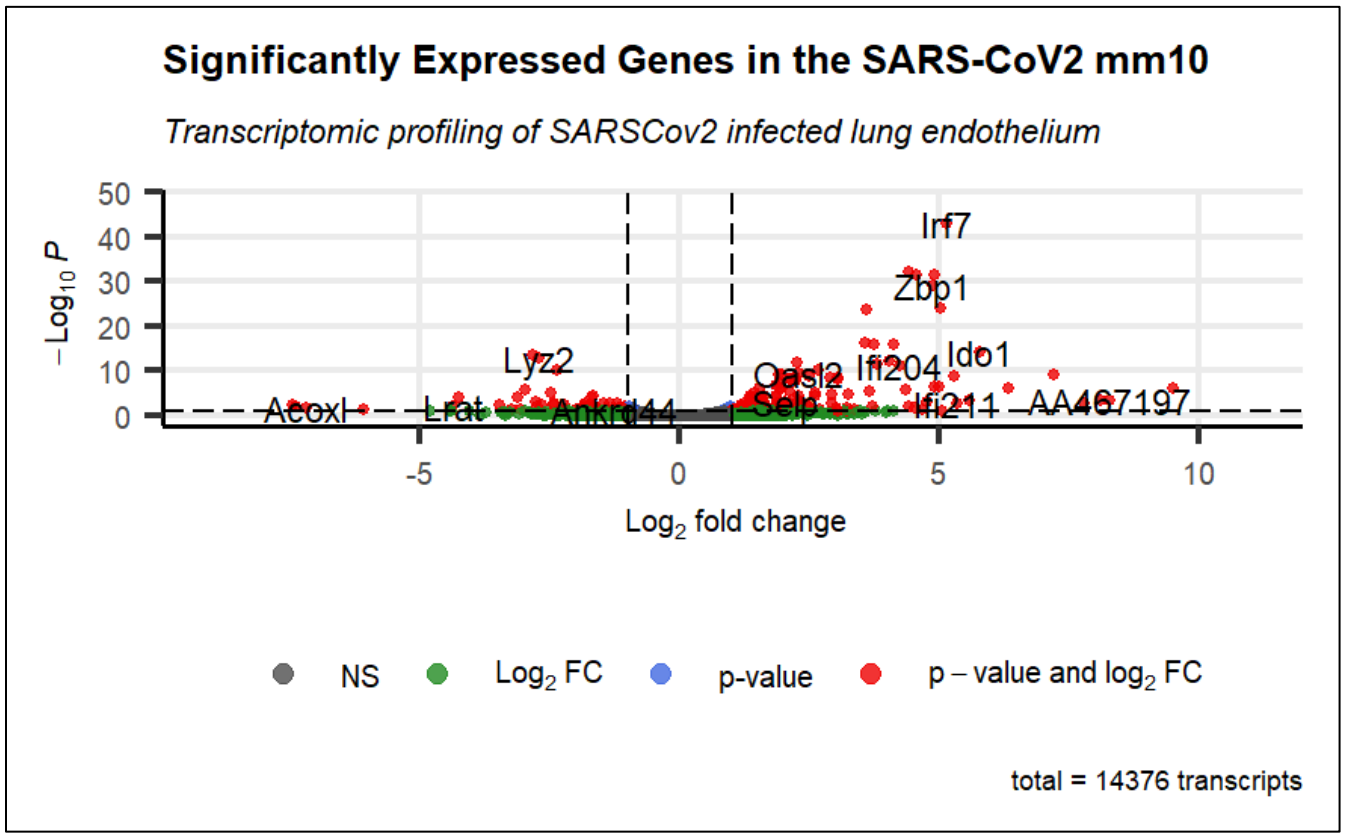


Figure 2: Significantly up-and down- regulated genes in the SARS-Cov2 data in comparison to the control. Genes that fall to the right of the dotted line, and above the horizontal line (P-value < 0.05), are significantly upregulated, while genes that fall to the left of the dotted line and above -log(0.05) are significantly downregulated.

Hierarchical clustering

Fig 3A. demonstrates that differential expression patterns are correlated across all samples, to an extent, regardless of treatment. By selecting for genes with significant differential expression, Fig. 3B.) identified genes with the most significant expression p-values (< $5x10^{-9}$).

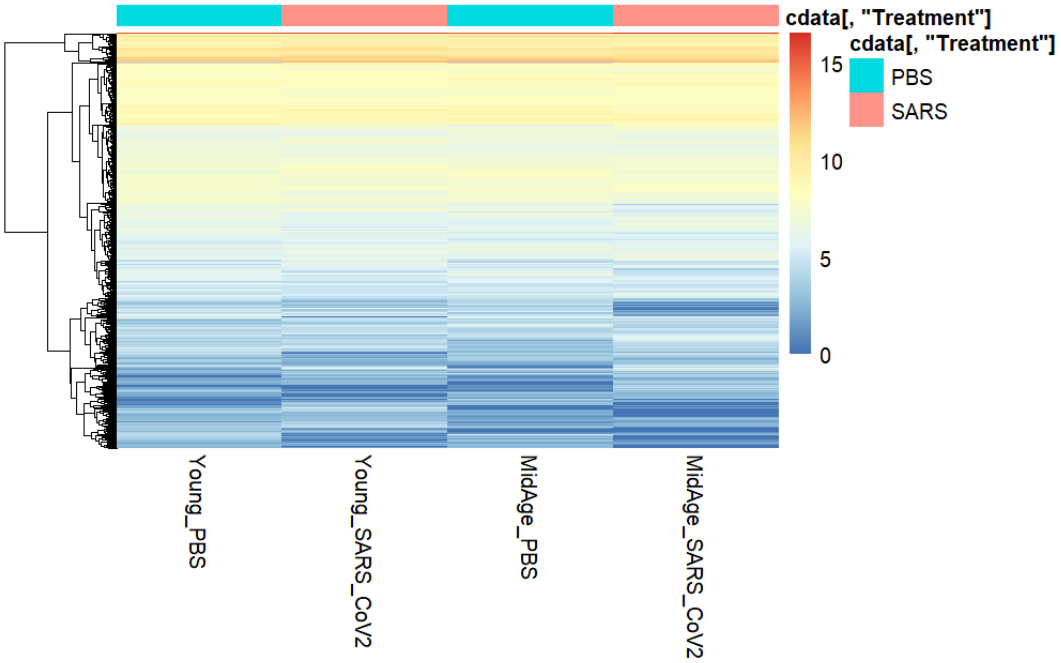


Figure 3A (top left): A global view of gene expression levels cross all four samples, prior to filtering. The expression data is sorted to be clustered with rows (genes) displaying similar expression patterns.

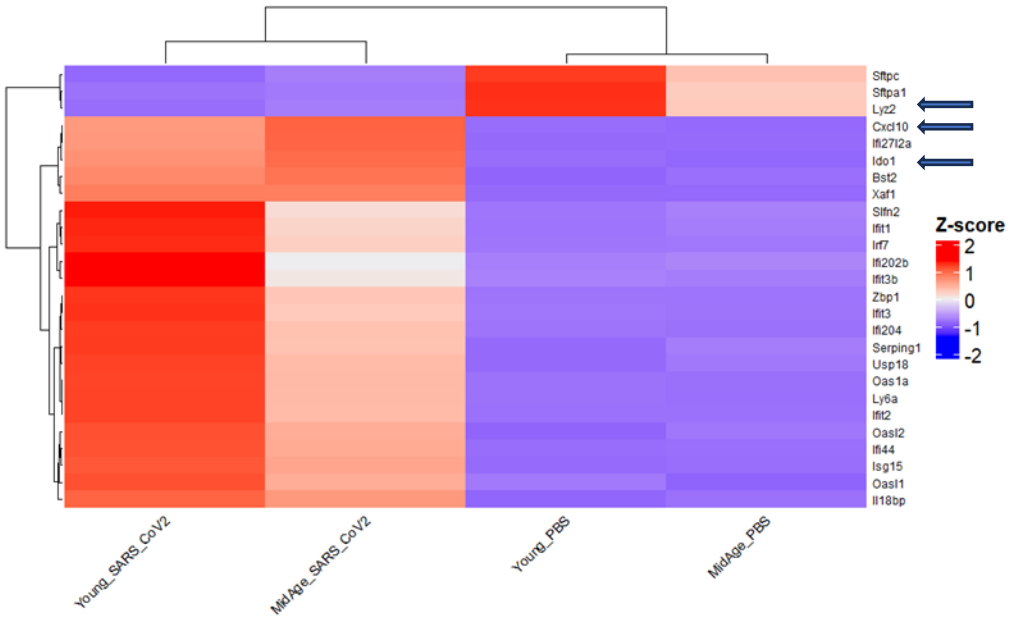


Figure 3B (bottom left): Clustered gene expression profiling after filtering for the most significant differentially expressed genes (adjusted p-values < $5x10^{-9}$). For downstream analysis, Lyz2 was selected for significant downregulation in SARS CoV2 samples, while cxcl10 and Ido1 were selected for significant upregulation.

Gene Ontology Enrichment

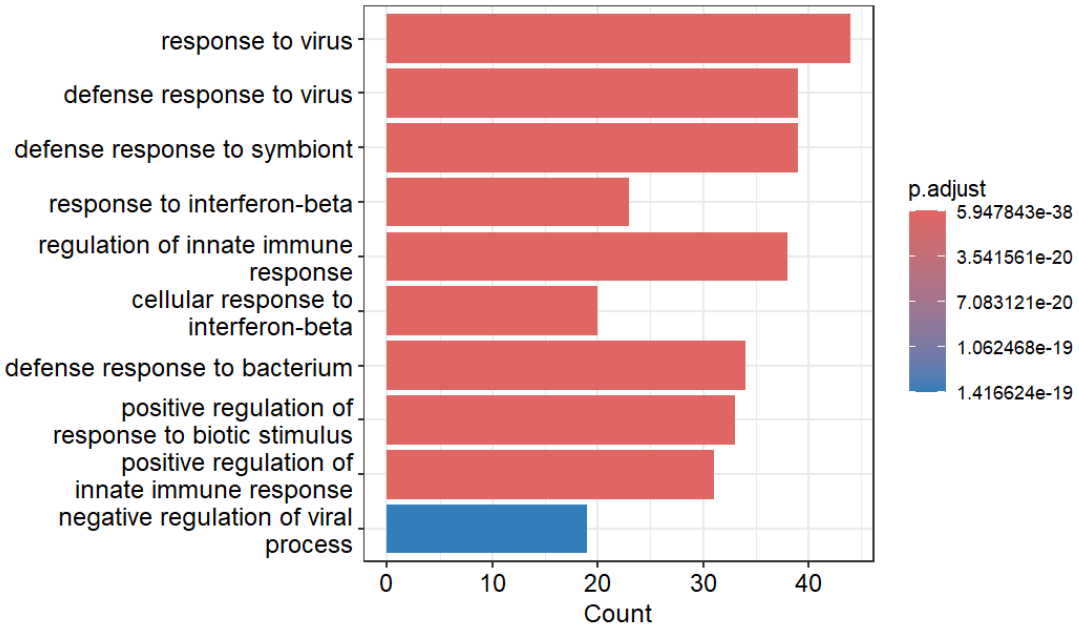


Figure 4: Identification of overrepresented gene ontology terms across significantly expressed genes. The filtered and scaled RNA seq. data was sorted by ascended order of adjusted p-values. The genes with the highest levels of expression were selected and quantified; those that appeared more frequently in the dataset were associated with higher RNA counts/overexpression. The Gene Enrichment analysis compared the prominent genes with the appropriate organism database package (org.Mm.eg.db) and identified the top 10 relevant biological pathways.

Gene Set Enrichment Analysis (GSEA)

Aligning possible gene networks with relation to the differentially expressed genes in SARS-Cov2-infected (lung endothelial) phenotypes.

Significant differenential expression was classified by a 2-fold log change as assessing how important a gene plays a role in a pathway.

GSEA can also identify and downregulated and lower expressed genes that are more subtle in a dataset, but whose expression change is still consistent with a gene cluster.

My goals of GSEA were to:

1. Identify gene enrichment pathway related to 2-fold upregulation of cxcl10 and IDO1
   - IDO1 is normally an inducible enzyme and its most important inducer is the cytokine interferon-γ (IFN-γ).
2. Identify gene enrichment pathway to explain significant downregulation of Lyz2
   - Lyz2 enables lysozyme function

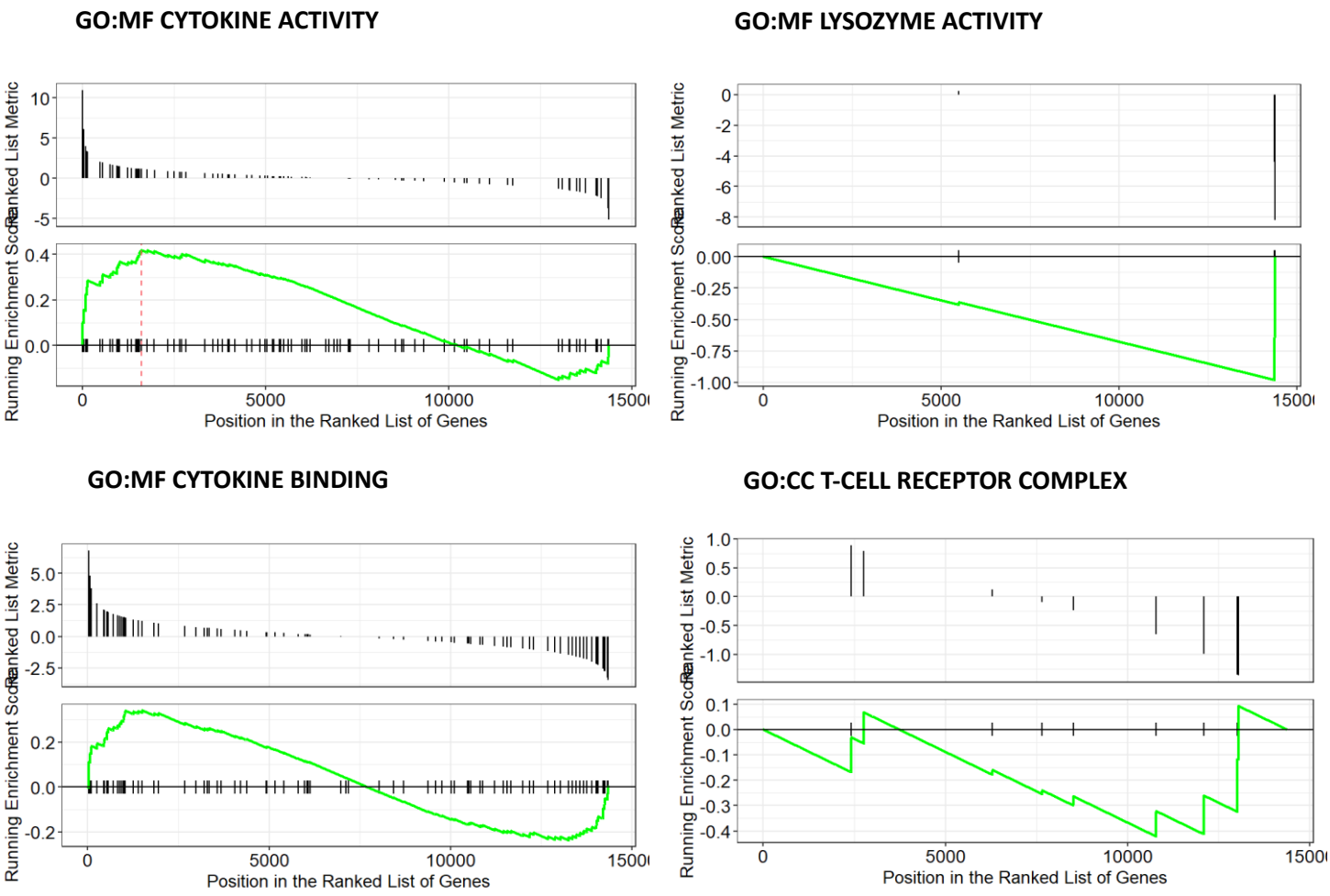


Figure 5: Enrichment Scores by gene rank position, with respect to differential expression data. Peaks in the ES (enrichment score) plot, depicted by the green line, indicate genes (by location within the ranked list), that are significantly enriched. The ranked list was put in descending order by the gene's respective log2fold change score (higher stat values indicate stronger differential expression). Due to the shape of the curves, the enriched gene sets correlate most strongly with the cytokine activity (top left) and binding (bottom left) Molecular Function (MF) pathways. In comparison to the poor correlation with the T-cell receptor complex (bottom right); the phenotypic traits observed in the gene expression analysis, are related to cytokine activity and not and T-cell activity, even though both cell-types are known to interact and are involved in overlapping immunological processes. The Lysozyme activity (top right) was plotted, due to the significant downregulation of Lyz2; this plot does not confirm significance of the transcriptome data, or of Lyz2 in this pathway.

References

"GEO Accession Viewer." n.d. Accessed April 20, 2024. https://www-ncbi-nlm-nih-gov.ccl.idm.oclc.org/geo/query/acc.cgi?acc=GSE230022.

"Org.m.ed.db". (V.3.19.1). Accessed May 3, 2025). https://www-ncbi-nlm-nih-gov.ccl.idm.oclc.org/gene. https://bioconductor.org/packages/release/data/annotation/html/org.Mm.eg.db.html#:~:text=https%3A//bioconductor.org/packages/org.Mm.eg.db/.

Blighe, K, S Rana, and M Lewis. 2018. "EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling." https://github.com/kevinblighe/EnhancedVolcano.

Gene Set Enrichment Analysis (GSEA 4.3.3). Subramanian, Tamayo, et al. (2005, PNAS) and Mootha, Lindgren, et al. (2003, Nature Genetics).Accessed May 3, 2024. https://www.gsea-msigdb.org/gsea/index.jsp.

Differential Expression Analysis of RNA Seq. Data (1.44.0). https://bioconductor.org/packages/DESeq2/.

Molecular Signatures Database (MSigDB 2023.2) website. Accessed May 5, 2024. https://www.gsea-msigdb.org/gsea/index.jsp.