Differential Gene Expression analysis in PM2.5-Exposed Bronchial Epithelial Cells

Prasanth Kumar Thuthika.

Content Table

Introduction

Overview of Differential Gene Expression Analysis

Differential gene expression (DGE) analysis is a powerful approach to identify genes whose expression levels change significantly under specific conditions or treatments. By comparing gene expression profiles between experimental groups, DGE analysis reveals key molecular responses and pathways involved in biological processes. This method is widely used to study cellular responses to environmental stressors, such as pollutants, and can provide insights into disease mechanisms and potential therapeutic targets.

Impact of PM2.5 on Bronchial Epithelial Cells

PM2.5, or particulate matter with a diameter of less than 2.5 μm, is a critical air pollutant linked to numerous adverse health effects, particularly in the respiratory system. Bronchial epithelial cells, as the first line of defence against airborne pollutants, exhibit substantial transcriptomic changes when exposed to PM2.5. These changes often involve the activation of inflammatory pathways, oxidative stress responses, and disruption of cellular homeostasis. Prolonged exposure to PM2.5 has been associated with the development of respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD), as well as systemic inflammatory conditions.

Relevance of Transcriptomic Analysis in Environmental Studies

Transcriptomic analysis using RNA sequencing (RNA-Seq) provides a comprehensive and high-resolution view of gene expression changes caused by environmental exposures, including PM2.5. Differential gene expression analysis allows researchers to identify key genes, pathways, and biological processes that are disrupted, offering valuable insights into the molecular mechanisms underlying pollutant-induced cellular responses. This knowledge contributes to a better understanding of the health risks posed by environmental pollutants and informs strategies for prevention and intervention.

Objectives of the Study

Identify Differentially Expressed Genes (DEGs) in Response to PM2.5 Exposure

- Use RNA-Seq data to perform differential gene expression analysis, pinpointing genes with significant changes in expression levels under PM2.5 exposure.

Determine Key Pathways and Biological Processes Affected by PM2.5

- Conduct pathway enrichment analysis on the identified DEGs to reveal critical biological processes and molecular pathways disrupted by PM2.5 exposure.

Materials and Methods

Data

The data used in this study explores transcriptomic changes in bronchial epithelial cells upon exposure to fine particulate matter (PM2.5). The cell line utilized in the study is **BEAS-2B**, a widely used model for human bronchial epithelial cells, selected for its relevance in studying respiratory health. These cells were subjected to varying doses and durations of PM2.5 exposure to mimic both acute and chronic exposure scenarios encountered in polluted environments (Huang et al., 2021).

In terms of experimental design, the dataset includes both acute and chronic PM2.5 exposure conditions. Acute exposure was simulated by treating cells with either a high dose of PM2.5 (30 μg/cm²) or a low dose (1 μg/cm²) for 24 hours. Chronic exposure involved repeated low-dose treatments (1 μg/cm²) over seven days, allowing the study of long-term effects. Vehicle-treated cells were used as controls to establish baseline transcriptomic profiles. All conditions were performed in biological triplicates to ensure robust and reliable results.

The RNA-Seq data provides genome-wide transcriptomic profiles, enabling the identification of differentially expressed genes (DEGs) under PM2.5 exposure. This approach facilitates the exploration of disrupted pathways and biological processes associated with PM2.5-induced cellular responses.

The raw sequencing data is publicly available in the **NCBI SRA** under the accession ID **SRP275645**. The **Human reference genome (GRCh38)** and annotation files were downloaded from **Ensembl** for alignment and downstream analysis.

The RNA-Seq dataset serves as a valuable resource for understanding the molecular mechanisms underlying PM2.5 exposure, offering insights into the cellular responses of bronchial epithelial cells.

Setting Up the Computational Environment

A dedicated **Conda environment** was created to efficiently manage and install all required software. This isolated environment ensured compatibility and minimized software conflicts throughout the analysis. The **Bioconda channel** was utilized for installing bioinformatics tools, providing a comprehensive and reliable repository for the necessary packages. The environment supported tools for data retrieval, quality control, alignment, and differential expression analysis.

To handle the computational demands of processing RNA-Seq data, the analysis was performed on a **high-performance cluster (HPC)**. The **Slurm workload manager** facilitated job scheduling, enabling the parallel execution of tasks such as sequence alignment, quantification, and differential expression analysis. This setup optimized processing efficiency, significantly reducing runtime and ensuring scalability for large datasets.

By leveraging this robust computational setup, the pipeline ensured seamless integration of tools and reproducibility of results across all stages of the RNA-Seq workflow, from raw data preprocessing to generating differential gene expression insights.

Tools and Setting Up the Environment

The analysis of RNA-Seq data for **Differential Gene Expression (DGE)** requires a robust computational environment and a combination of specialized bioinformatics tools. This study utilizes a streamlined pipeline to ensure reproducibility, accuracy, and efficiency. The computational environment was established using **Conda**, a widely used package and environment manager, ensuring compatibility across all software tools. Below is a summary of the tools used at various stages of the analysis:

1. **SRA Toolkit**: Retrieval of raw sequencing data.
2. **FastQC**: Quality control checks on raw FASTQ files.
3. **Trim Galore**: Adapter trimming and removal of low-quality reads.
4. **HISAT2**: Alignment of cleaned reads to the GRCh38 genome.
5. **Samtools**: Sorting and indexing of aligned reads.
6. **Subread (featureCounts)**: Quantification of gene-level read counts.
7. **DESeq2**: Differential expression analysis to identify differentially expressed genes (DEGs).
8. **R and Bioconductor Packages**: Custom scripts for statistical analysis and visualization.
9. **ClusterProfiler**: Pathway enrichment and functional annotation of DEGs.
10. **MultiQC**: Aggregated quality control reporting across all samples.

RNA-Seq Analysis Workflow

- **Data Retrieval**: RNA-Seq data was downloaded using the **SRA Toolkit** from publicly available repositories.
- **Quality Control**: **FastQC** was employed to evaluate the quality of the raw reads, followed by **Trim Galore** for adapter trimming and quality filtering.
- **Alignment**: High-quality reads were aligned to the human reference genome (GRCh38) using **HISAT2**, a splice-aware aligner optimized for RNA-Seq data.

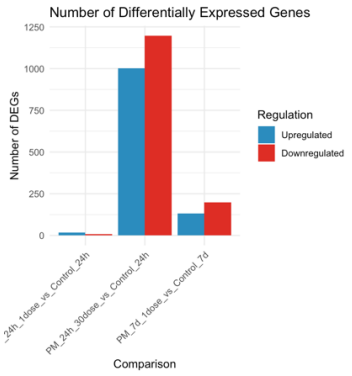
- **Quantification**: Gene-level read counts were generated using **featureCounts** from the Subread package.
- **Differential Expression Analysis**: The **DESeq2** package in R was used to identify DEGs between experimental conditions. It provided log fold change, statistical significance, and adjusted p-values for each gene.
- **Pathway Enrichment**: Identified DEGs were annotated using **ClusterProfiler**, revealing enriched biological pathways and functional categories.

Data Visualization

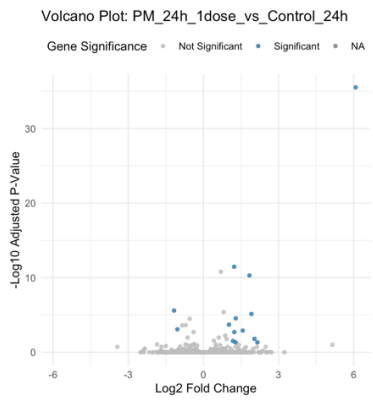
Data visualizations were created using the following tools to ensure clear and interpretable results:

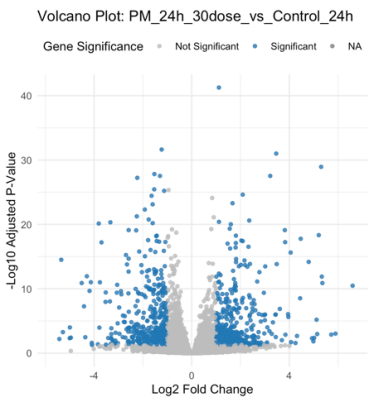
- **ggplot2**: For publication-ready plots, including volcano plots and bar charts.
- **MultiQC**: For consolidated quality reports of sequencing data.

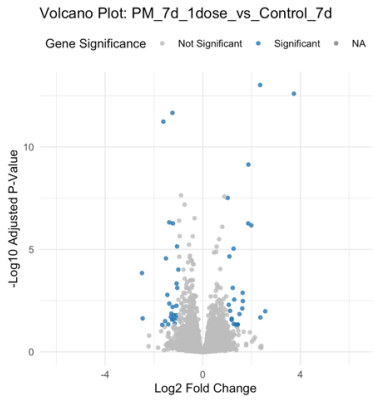
Results

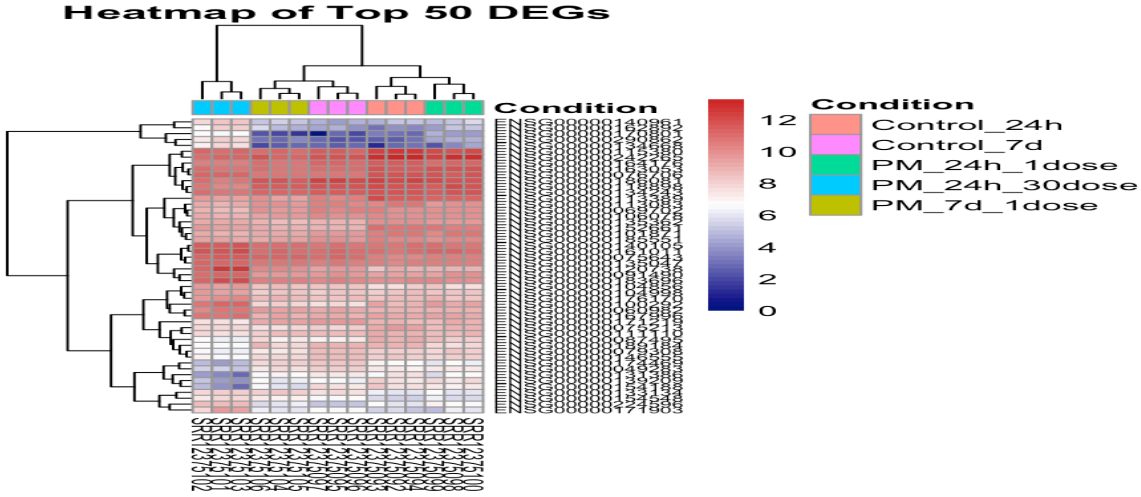


The PM_24h_30dose vs. Control_24h group exhibited the most substantial transcriptional changes, with 1,003 upregulated and 1,197 downregulated genes, indicating a strong dose-dependent response. In contrast, the PM_24h_1dose vs. Control_24h comparison showed minimal changes, with only 18 upregulated and 7 downregulated genes, while the PM_7d_1dose vs. Control_7d group demonstrated moderate gene expression alterations, suggesting sustained but less pronounced effects over time. The volcano plot for the PM_24h_30dose vs. Control_24h comparison highlights a widespread transcriptional disruption, with numerous genes exhibiting significant upregulation and downregulation (log2 fold change > ±1, adjusted p-value < 0.05).

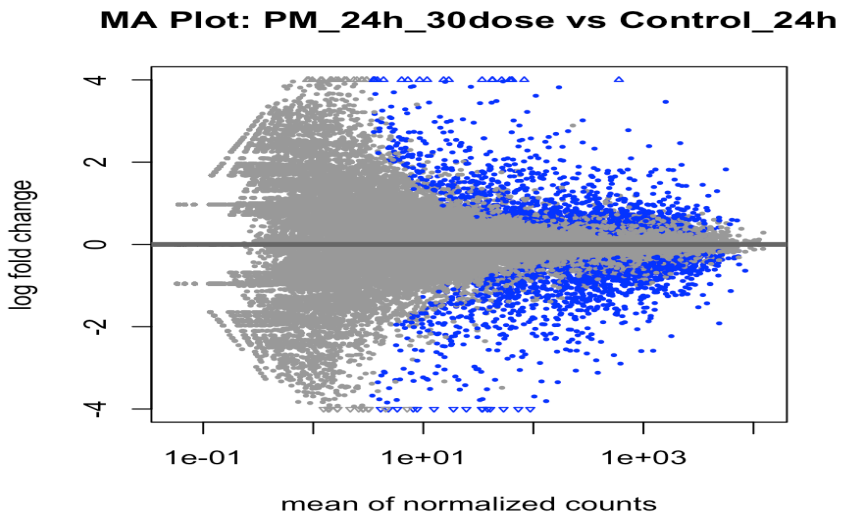




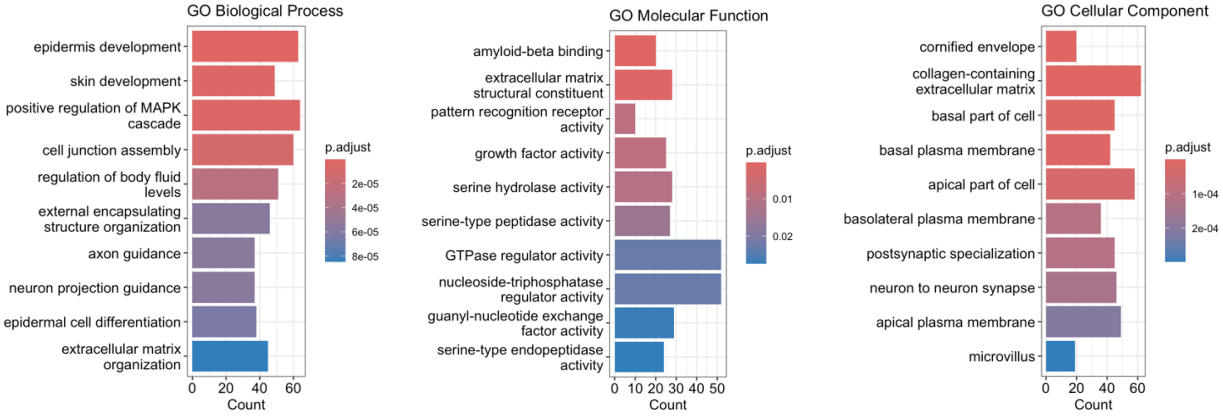




The heatmap of the top 50 differentially expressed genes (DEGs) reveals distinct clustering patterns across experimental conditions. Notably, the PM_24h_30dose group showed pronounced upregulation of most DEGs, reflecting a strong dose-dependent transcriptional activation. In contrast, the PM_24h_1dose and PM_7d_1dose groups displayed more moderate gene expression changes, suggesting a subdued or delayed regulatory response to lower or prolonged exposure.



The MA plot further illustrates significant transcriptional alterations in the PM_24h_30dose vs. Control_24h group, with marked upregulation and downregulation patterns. These changes, especially in highly expressed genes, emphasize the dose-dependent impact on gene regulation. Principal Component Analysis (PCA) revealed distinct clustering of samples, with PC1 and PC2 explaining 58.3% and 22.8% of the total variance, respectively. The PM_24h_30dose group clustered separately from all other groups, reflecting a unique and pronounced transcriptional response to high-dose exposure. In contrast, the Control_24h, PM_24h_1dose, Control_7d, and PM_7d_1dose groups clustered more closely, indicating relatively similar gene expression profiles under lower or prolonged exposure.



Gene Ontology (GO) enrichment analysis identified key biological processes and molecular functions disrupted by PM exposure. Biological processes such as epidermis development, skin development, and cell junction assembly were significantly enriched, indicating compromised tissue integrity. Activation of the MAPK signaling pathway suggests enhanced stress and inflammatory responses. Molecular function analysis highlighted enrichment in extracellular matrix structural constituents and growth factor activity, reflecting disruptions in cellular structure and signaling. Additionally, enrichment in collagen-containing extracellular matrix and plasma membrane components points to extensive extracellular remodeling and impaired cell communication. Collectively, all these demonstrate that particulate matter exposure profoundly affects pathways involved in tissue structure, cellular signaling, and homeostasis.

References

1. Burnett, R., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences*, 115(38), 9592-9597. https://doi.org/10.1073/pnas.1803222115

2. Huang, S. K., Tripathi, P., Koneva, L., Cavalcante, R., et al. (2021). Effect of concentration and duration of particulate matter exposure on the transcriptome and DNA methylome of bronchial epithelial cells. *Environmental Epigenetics*, 7(1), dvaa022. https://doi.org/10.1093/eep/dvaa022

3. Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), 559. https://doi.org/10.1186/1471-2105-9-559

4. Tripathi, P., Huang, S. K., et al. (2021). DNA methylation changes induced by PM2.5 exposure. *Environmental Health Perspectives*, 129(1), 017002. https://doi.org/10.1289/EHP7723

5. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 83. https://doi.org/10.1186/s13059-017-1215-1