Project Component 5: Final Written Report

Microarray-based Transcriptomic Analysis and Signature Extraction in Mice Lungs following 4 and 6 months of Cigarette Smoke Exposure

Meena Easwaran

McWilliams School of Biomedical Informatics

The University of Texas Health Science Center at Houston

BMI 5333 Systems Medicine: Principles and Practice

Dr. Xiaobo Zhou

Fall 2024

December 3, 2024

Specific Aims

Cigarette smoke (CS) exposure poses serious health risks and significantly contributes to the development of various respiratory diseases. Understanding the molecular mechanisms behind these diseases is crucial for developing effective therapeutic strategies. This project aims to investigate the transcriptomic changes in gene expression in mouse lung tissue resulting from prolonged exposure to CS at different time intervals of four and six months. The central hypothesis of this research is that extended exposure to CS leads to specific changes in the gene expression of lung tissue in mice, which are closely linked to the progression of lung disease over time. To thoroughly explore this hypothesis, the study will focus on three main objectives:

Aim 1: Characterize the differentially expressed genes (DEGs) in the lung tissue of mice after four and six months of exposure to CS. Identifying DEGs at each time point will reveal gene expression changes linked to prolonged CS exposure. Data will be obtained using the "GEOquery" R package (1), and Limma (2) will be used for differential expression analysis to highlight significant DEGs that may influence lung pathology.

Aim 2: Conduct a functional enrichment analysis of DEGs to identify key pathways involved in CS-induced lung pathology at both time points. Analyzing affected biological pathways over time will clarify disruptions from CS exposure. Metascape (3) will be used for Gene Ontology (GO) (4) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (5–7) enrichment analyses to categorize DEGs into functional pathways, highlighting the modulation of molecular and cellular processes over four and six months.

Aim 3: **Identify pathway-specific regulatory networks and hub genes in response to four and six months of CS exposure.** Mapping DEGs from specific pathways to interaction networks will reveal key regulatory elements and hub genes that drive disease progression, potentially

uncovering new therapeutic targets. StringDB (8) and Cytoscape (9) will be used for network analysis and visualization, identifying central regulators relevant to CS-induced lung pathology at each time point.

A. Background and Significance

CS exposure remains a major global health burden, contributing to various respiratory diseases. The detrimental effects associated with CS become increasingly apparent as they extend deeper into the airway passages, particularly within the lungs (10–12). Chronic exposure to CS can lead to chronic obstructive pulmonary disease (COPD) and lung cancer (**Fig. 1**), both of which significantly impact global morbidity and mortality. These diseases result from cumulative molecular and cellular changes in lung tissue, leading to progressive and often irreversible damage. Despite advancements in lung transcriptomic studies of CS exposure, a critical gap exists in understanding the temporal dynamics and regulatory mechanisms driving CS-induced lung pathology at distinct stages of disease progression.

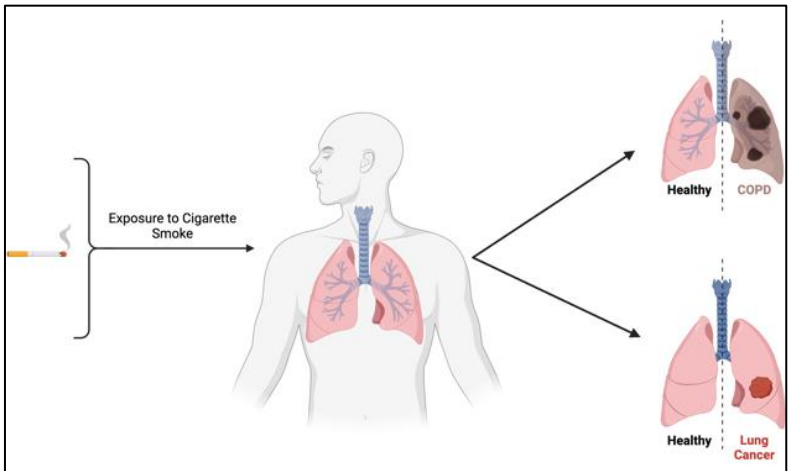

Fig. 1. Cigarette smoking and lungs. Prolonged exposure to cigarette smoke can lead to chronic obstructive pulmonary disease (COPD) and lung cancer. Created in https://BioRender.com

This research aims to fill this existing gap by examining transcriptomic profiles in mouse lungs after four and six months of exposure to CS. By identifying the specific molecular signatures and regulatory networks associated with each time point using various bioinformatic tools, this study seeks to provide essential insights into the development of lung pathology related to CS exposure. This knowledge may help discover novel therapeutic targets.

Additionally, this work supports the broader goal of enhancing precision medicine by identifying molecular markers that can guide early intervention strategies. The integration of these findings with my ongoing research on the effects of CS on the larynx at Stanford University will further illuminate tissue-specific responses, contributing to a comprehensive systems medicine approach for addressing CS-related health issues.

B. Innovation

B.1. Temporal Transcriptomic Profiling. By analyzing changes in gene expression after four and six months of CS exposure, the study offers a clearer understanding of how lung pathology evolves over time, highlighting disease dynamics often missed in cross-sectional studies.

B.2. Integration of Multi-level Bioinformatic Analyses. By integrating differential gene expression, functional enrichment, and network analysis, this study provides a thorough understanding of the molecular basis of CS-induced lung pathology.

B3. Identification of Novel Therapeutic Targets. By conducting network analysis, this study seeks to identify key regulatory hub genes from specific pathway mechanisms, which could serve as potential targets for precision therapies to alleviate lung diseases induced by CS.

B.4. Systems Medicine Framework. The project combines findings from mouse lungs with my ongoing research on the effects of CS on the larynx, providing a systems medicine perspective to identify specific and shared disease mechanisms in different tissues.

C. Preliminary Data and Approach

C.1. Differential expression of genes in mouse lung tissue following four and six months of CS exposure. Transcriptomic data with the accession ID GSE52509 was retrieved from the NCBI Gene Expression Omnibus (GEO) (13) using the GEOquery R package (1). The differential expression was assessed using linear models with empirical Bayes moderation implemented in the

Limma package (2) in R. This approach ensured the robust identification of significantly differentially expressed genes (DEGs) while accounting for variability among biological replicates. Specifically, significant DEGs at each time point were identified by applying criteria of an adjusted p-value < 0.05 via the Benjamini-Hochberg method and a |fold change| >1.5. Data was visualized by volcano plots, as shown in **Fig. 2,** using the EnhancedVolcano R package (14). This analysis was conducted to identify genes significantly influenced by exposure to CS. It highlighted the molecular changes specific to different time points and their relevance to lung-related diseases. The results of the differential expression analysis indicated that after six months of exposure to CS, the gene expression changes were markedly greater than that observed after four months, suggesting a progressive impact of prolonged CS exposure on lung pathology.

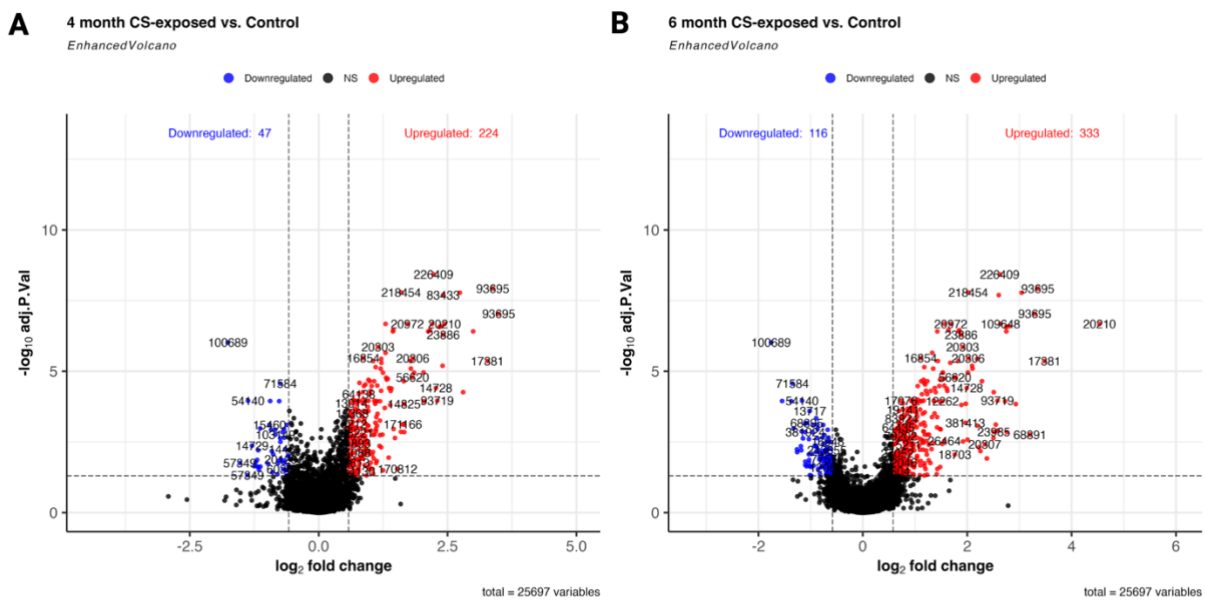

Fig. 2. Differential expression analysis. Differential expression analysis revealed 271 DEGs between four-month CS-exposed and control groups (**A**). In particular, 224 genes showed increased expression while 47 genes exhibited decreased expression. At six months, 449 DEGs with 333 upregulated and 116 downregulated genes were identified between both groups (**B**). The volcano plots here indicate the significant upregulated and downregulated genes as red and blue dots respectively. Created in https://BioRender.com

C.2. Functional enrichment analysis of DEGs to identify critical pathways involved in CS-induced lung pathology. A temporal comparison will be conducted for the pathways enriched at four and six months via clustering techniques to identify overlapping and time-specific biological processes using Metascape, a web-based gene annotation bioinformatics tool (3). This analysis focused on GO and pathways from the KEGG database. Hypergeometric tests were used to assess the results, and corrections for multiple comparisons were applied using the Benjamini-Hochberg method to control the false discovery rate (FDR). In terms of GO enrichments, all GO subcategories, biological process (BP), cellular component (CC), and molecular function (MF), were included to understand how molecular and cellular processes changed over time. Enriched terms with a minimum gene overlap of 3, an enrichment ratio of 1.5,

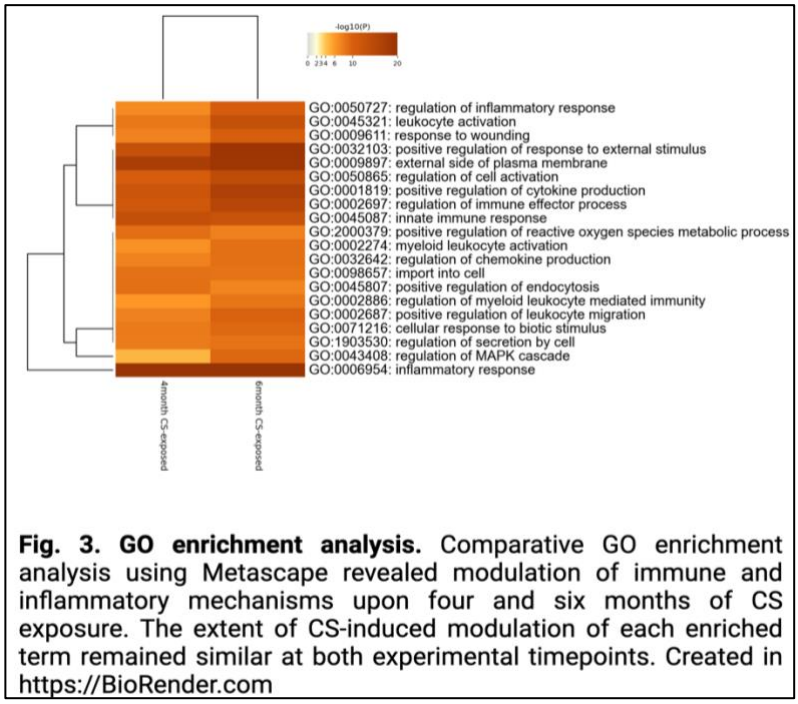

Fig. 3. GO enrichment analysis. Comparative GO enrichment analysis using Metascape revealed modulation of immune and inflammatory mechanisms upon four and six months of CS exposure. The extent of CS-induced modulation of each enriched term remained similar at both experimental timepoints. Created in https://BioRender.com

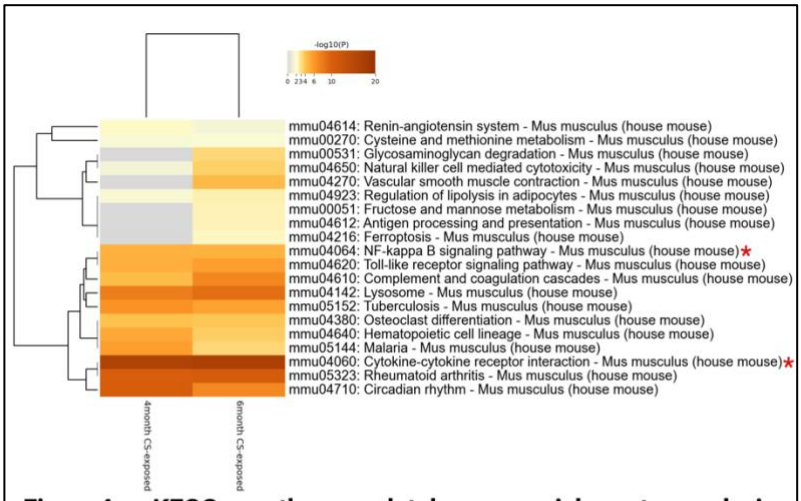

Fig. 4. KEGG pathway database enrichment analysis. Comparative KEGG database enrichment analysis using Metascape revealed modulation of cytokine-cytokine receptor interaction and NF-kappa beta immuno signaling pathway upon four and six months of CS exposure. The extent of CS-induced modulation of each enriched term remained similar at both experimental timepoints. Red colored asterisks indicate these terms in the heatmap. Created in https://BioRender.com

and FDR < 0.05 were reported as significant. These enriched terms are illustrated as bar plots, as demonstrated in **Fig. 3** and **Fig. 4**. Both the GO and KEGG enrichment analyses revealed strong enrichments in immune and inflammatory pathways at both time points following CS exposure. Specifically, inflammatory responses and cytokine-cytokine receptor interaction mechanisms were the most enriched at both time points. These heightened immune and inflammatory responses observed could significantly influence disease progression beyond four months post-exposure. Further functional validation is necessary to elucidate this relationship fully.

C.3. Identify key pathway-specific regulatory networks and hub genes over four and six months of CS exposure. Around 144 and 36 DEGs compiled from both exposure time points were identified to be enriched in the inflammatory responses and cytokine-cytokine receptor interaction mechanisms, respectively. Medium confidence protein-protein interaction (PPI) networks (score > 0.4) were constructed using the identified DEGs from inflammatory responses and cytokine-cytokine receptor interaction mechanisms via StringDB, as seen in **Fig. 5** (8). The networks were visualized and analyzed for hub genes with Cytoscape (9). Specifically, the Maximal Clique Centrality (MCC) algorithm (15) on the cytoHubba plugin (15) on Cytoscape was used to identify the top ten key regulatory hub genes in these specific pathway mechanisms as visualized in **Fig. 6**. Statistical analysis included permutation testing to confirm the significance of hub genes in driving network connectivity in these pathways. The top 10 hub genes associated with inflammatory responses (**Fig. 5**) included granulocyte-macrophage colony-stimulating factor (CSF2), interleukin-1 alpha (IL1A), C-X-C motif chemokine ligand 10 (CXCL10), C-C chemokine receptor type 5 (CCR5), C-X-C motif chemokine ligand 12 (CXCL12), C-C motif chemokine ligand 2 (CCL2), C-C motif chemokine ligand 3 (CCL3), C-X-C motif chemokine ligand 1 (CXCL1), C-C motif chemokine ligand 4 (CCL4), and tumor necrosis factor (TNF).

Additionally, the top 10 hub genes associated with cytokine-cytokine receptor interaction mechanisms (**Fig. 6**) were C-X-C motif chemokine ligand 5 (CXCL5), C-X-C motif chemokine ligand 1 (CXCL1), C-X-C motif chemokine ligand 16 (CXCL16), C-C motif chemokine ligand 12 (CCL12), C-C motif chemokine ligand 2 (CCL2), C-X-C motif chemokine ligand 13 (CXCL13), C-C motif chemokine ligand 3 (CCL3), C-C chemokine receptor type 5 (CCR5), C-X-C motif chemokine ligand 9 (CXCL9), and C-X-C motif chemokine ligand 10 (CXCL10). The identified genes are key regulators in the inflammatory and cytokine signaling pathways involved in CS-induced lung pathology, which may drive disease progression and inform potential therapeutic targets.

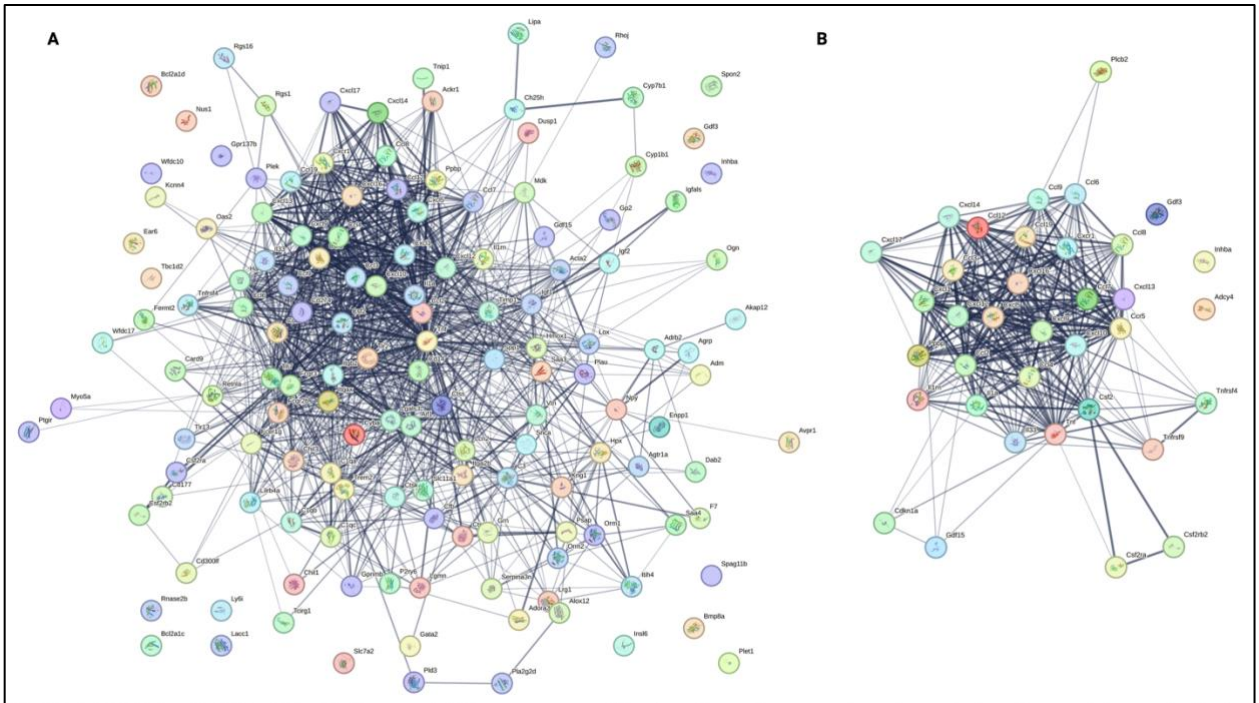

Fig. 5. StringDB PPI. A total of 144 and 36 enriched genes under inflammatory response **(A)** and cytokine-cytokine receptor interaction mechanisms **(B)** respectively, were visualized as a medium confidence network in StringDB, as indicated by the multi-colored circles in the panels. The connections between the circles represent gene associations, with thicker lines indicating stronger associations. Created in https://BioRender.com.

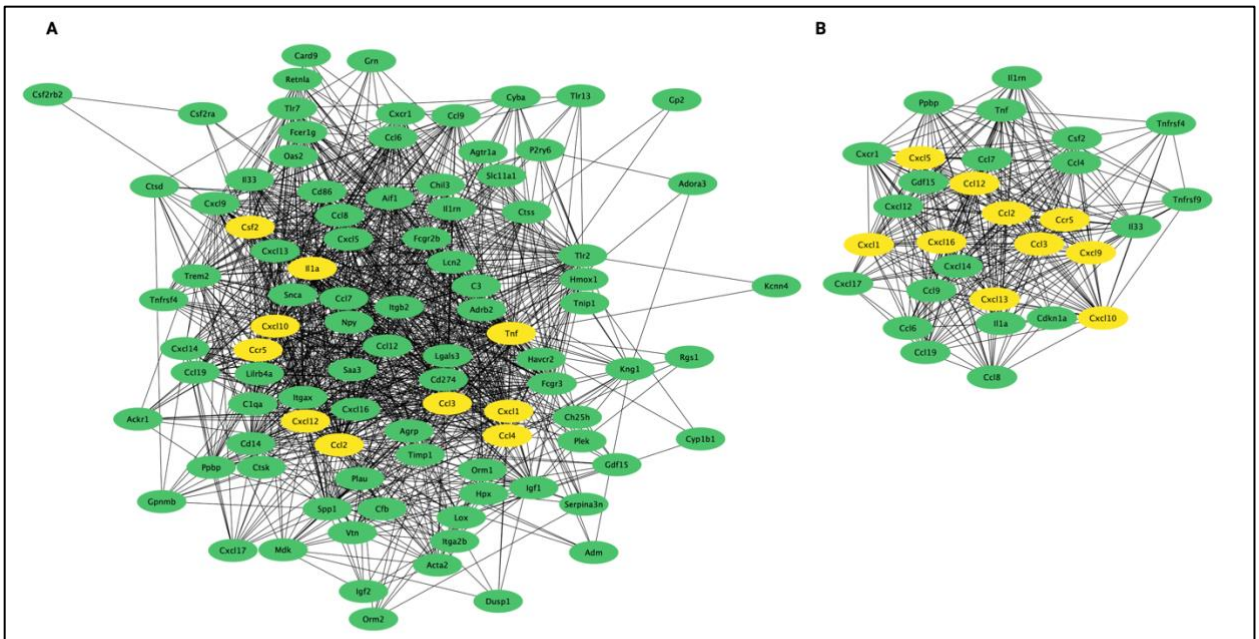

Fig. 6. Top ten hub genes in the inflammatory and cytokine-cytokine receptor interaction PPI. Csf2, Il1a, Cxcl10, Ccr5, Cxcl12, Ccl2, Ccl3, Cxcl1, Ccl4, and Tnf were the top ten hub genes associated with inflammatory response as indicated by the yellow-colored ellipses (**A**). Cxcl5, Cxcl1, Cxcl16, Ccl12, Ccl2, Cxcl13, Ccl3, Ccr5, Cxcl9, and Cxcl10 were the top ten hub genes associated with the cytokine-cytokine receptor interaction as indicated by the yellow-colored ellipses (**B**). Created in https://BioRender.com.

D. Discussion and Conclusion

This research uncovered significant regulatory networks and central hub genes linked to inflammatory responses and the mechanisms of cytokine-cytokine receptor interactions in lung tissue after exposure to CS. A comparison of the major hub genes within both pathways highlighted several shared genes, including CXCL1, CXCL10, CCR5, CCL2, and CCL3. These genes play crucial roles in the recruitment of immune cells, chemotaxis, and inflammation, all of which are vital processes in lung pathology following exposure to CS. For example, CXCL1 and CXCL10 are chemokines that facilitate the recruitment of neutrophils and T-cells to the injury site (16–18). CCR5 functions as a receptor that aids in recruiting macrophages and other immune cells (19). Furthermore, CCL2 and CCL3 also partake in the activation of macrophages (20). Identifying these genes as crucial regulators highlights their potential as therapeutic targets for treating lung diseases caused by CS, particularly in the context of precision medicine.

The findings of this study align with my previous research (21) concerning laryngeal tissue exposed to CS, wherein similar inflammatory and immune-related genes, such as CXCL1 and CCL2, emerged as significant hub genes associated with CS-induced laryngeal inflammation. This observation suggests that the inflammatory pathways identified in lung tissue may be broadly relevant to various respiratory tissues, highlighting a conserved mechanism of injury and repair in response to CS exposure. From a systems medicine perspective, the correlation of hub gene findings between lung and laryngeal tissues underscores the importance of viewing these anatomical sites as part of a larger respiratory system, especially alluding to the concept of unified airway theory (22,23). According to this theory, responses in one airway region can affect responses in another. This implies that cells in the lung tissue may behave similarly to other cellular types found in other airway regions after CS exposures and vice versa. Furthermore, analyzing temporal changes in gene expression gives a deeper understanding of these processes' dynamic nature over time. Identifying hub genes across different inflammatory pathways illustrates the complexity of CS-induced pathology and the potential for targeted interventions focusing on specific regulatory nodes within these networks. This systems-level approach provides a comprehensive perspective on how environmental factors, such as CS, elicit multifaceted molecular changes that contribute to the development of lung disease, thereby linking the pathological processes of the lung and larynx within a unified framework of respiratory health.

This project for the course centered around the analysis of genomic/omics data and the extraction of signatures from publicly accessible data repositories, reflecting the broader topics of bioinformatics and systems medicine covered in the course. A significant skill I acquired was the ability to perform differential gene expression analysis using the Limma package in R, which

offered insights into how CS exposure affects lung tissue at a molecular level. This involved managing transcriptomic data and applying various bioinformatics tools for functional enrichment analysis, including Metascape for identifying pathways and StringDB for mapping protein-protein interactions. Moreover, I gained expertise in integrating and visualizing intricate datasets using volcano plots, bar plots, and network visualizations, which were crucial for effectively communicating the results.

Throughout the project, I encountered several challenges related to data analysis using R scripts. I faced issues with handling missing values, which complicated the accuracy of my results. Additionally, optimizing the performance of my scripts for large datasets proved to be a significant hurdle. I also struggled with effectively visualizing the results, prompting me to explore various libraries and techniques within R to improve my data presentation. A significant lesson learned was the importance of validating computational findings with experimental data, as bioinformatics analyses alone may not fully capture the complexity of biological systems. The project also raised questions about the temporal dynamics of gene expression in response to prolonged exposure to environmental stressors, which could inform future studies in lung disease and airway research. Overall, this project provided a comprehensive understanding of the power of bioinformatics in exploring complex biological questions and further emphasized the value of these approaches in studying disease progression and therapeutic targets.

References

1. Sean D, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics [Internet]. 2007 Jul 15 [cited 2023 Dec 7];23(14):1846–7. Available from: https://dx.doi.org/10.1093/bioinformatics/btm254

2. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Apr 20 [cited 2023 Dec 7];43(7):e47–e47. Available from: https://dx.doi.org/10.1093/nar/gkv007
3. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications 2019 10:1 [Internet]. 2019 Apr 3 [cited 2024 Jan 25];10(1):1–10. Available from: https://www.nature.com/articles/s41467-019-09234-6
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics 2000 25:1 [Internet]. 2000 May [cited 2024 Jan 25];25(1):25–9. Available from: https://www.nature.com/articles/ng0500_25
5. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science [Internet]. 2019 Nov 1 [cited 2022 Oct 16];28(11):1947–51. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3715
6. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2022 Oct 16];28(1):27–30. Available from: https://academic.oup.com/nar/article/28/1/27/2384332
7. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res [Internet]. 2023 Jan 6 [cited 2024 Jan 25];51(D1):D587–92. Available from: https://pubmed.ncbi.nlm.nih.gov/36300620/
8. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization

of user-uploaded gene/measurement sets. Nucleic Acids Res [Internet]. 2021 Jan 1 [cited 2024 Jan 25];49(D1):D605. Available from: /pmc/articles/PMC7779004/

9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res [Internet]. 2003 Nov [cited 2024 Jan 25];13(11):2498–504. Available from: https://pubmed.ncbi.nlm.nih.gov/14597658/

10. Curtin GM, Higuchi MA, Ayres PH, Swauger JE, Mosberg AT. Lung tumorigenicity in A/J and rasH2 transgenic mice following mainstream tobacco smoke inhalation. Toxicological Sciences [Internet]. 2004 Sep [cited 2021 Mar 21];81(1):26–34. Available from: https://pubmed.ncbi.nlm.nih.gov/15159525/

11. Mouronte-Roibás C, Leiro-Fernández V, Fernández-Villar A, Botana-Rial M, Ramos-Hernández C, Ruano-Ravina A. COPD, emphysema and the onset of lung cancer. A systematic review. Cancer Lett. 2016 Nov 28;382(2):240–4.

12. Walser T, Cui X, Yanagawa J, Lee JM, Heinrich E, Lee G, et al. Smoking and Lung Cancer: The Role of Inflammation. Proc Am Thorac Soc [Internet]. 2008 Dec 12 [cited 2024 Jan 25];5(8):811. Available from: /pmc/articles/PMC4080902/

13. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res [Internet]. 2013 Jan 1 [cited 2023 Oct 8];41(D1):D991–5. Available from: https://dx.doi.org/10.1093/nar/gks1193

14. Bioconductor - EnhancedVolcano [Internet]. [cited 2024 Apr 24]. Available from: https://bioconductor.org/packages/release/bioc/html/EnhancedVolcano.html

15. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: Identifying hub objects and sub-networks from complex interactome. BMC Syst Biol [Internet]. 2014 Dec 8 [cited 2022 May 1];8(4):1–7. Available from: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-S4-S11

16. Peperzak V, Veraar EAM, Xiao Y, Bąbała N, Thiadens K, Brugmans M, et al. CD8+ T Cells Produce the Chemokine CXCL10 in Response to CD27/CD70 Costimulation To Promote Generation of the CD8+ Effector T Cell Pool. The Journal of Immunology [Internet]. 2013 Sep 15 [cited 2024 Dec 2];191(6):3025–36. Available from: https://dx.doi.org/10.4049/jimmunol.1202222

17. Tecchio C, Cassatella MA. Neutrophil-derived chemokines on the road to immunity. Semin Immunol [Internet]. 2016 Apr 1 [cited 2024 Dec 2];28(2):119. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7129466/

18. Sawant K V., Poluri KM, Dutta AK, Sepuru KM, Troshkina A, Garofalo RP, et al. Chemokine CXCL1 mediated neutrophil recruitment: Role of glycosaminoglycan interactions. Scientific Reports 2016 6:1 [Internet]. 2016 Sep 14 [cited 2024 Dec 2];6(1):1–8. Available from: https://www.nature.com/articles/srep33123

19. Oppermann M. Chemokine receptor CCR5: Insights into structure, function, and regulation. Cell Signal [Internet]. 2004 Nov [cited 2024 Dec 2];16(11):1201–10. Available from: https://pubmed.ncbi.nlm.nih.gov/15337520/

20. Bagheri-Hosseinabadi Z, Kaeidi A, Rezvani M, Taghipour Khaje Sharifi G, Abbasifard M. Evaluation of the serum levels of CCL2, CCL3, and IL-29 after first and second administrations of the COVID-19 vaccine (Oxford–AstraZeneca). Immunobiology. 2024 Mar 1;229(2):152789.

21. Easwaran M, Martinez JD, Kim JB, Erickson-DiRenzo E. Modulation of mouse laryngeal inflammatory and immune cell responses by low and high doses of mainstream cigarette smoke. Scientific Reports 2022 12:1 [Internet]. 2022 Nov 4 [cited 2024 Dec 2];12(1):1–19. Available from: https://www.nature.com/articles/s41598-022-23359-7

22. Yii ACA, Tay TR, Choo XN, Koh MSY, Tee AKH, Wang DY. Precision medicine in united airways disease: A "treatable traits" approach. Allergy [Internet]. 2018 Oct 1 [cited 2024 Jan 25];73(10):1964–78. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/all.13496

23. Bachert C, Luong AU, Gevaert P, Mullol J, Smith SG, Silver J, et al. The Unified Airway Hypothesis: Evidence From Specific Intervention With Anti–IL-5 Biologic Therapy. J Allergy Clin Immunol Pract. 2023 Sep 1;11(9):2630–41.