Syed Taqi Hasnain
Adam Stuckert
BIOL 4324

Integrative Analysis of Highly Expressed and Variable Genes in Lung Adenocarcinoma: Unveiling Key Molecular Insights from RNA-Seq Data

Abstract

Lung adenocarcinoma (LUAD), a major subtype of non-small cell lung cancer (NSCLC), remains a leading cause of cancer-related mortality around the world due to its molecular complexity and therapeutic resistance. This study employs a bioinformatics-driven approach to analyze LUAD gene expression data, focusing on identifying genes with high mean expression and high variance to elucidate their roles in tumor biology. Gene expression data were obtained from the UCSC Xena database and analyzed using R Studio, revealing six key genes of interest: IGKC, MT-CO3, FTL, SFTPB, IGKV4-1, and IGKV3-20. These genes highlight critical processes, including immune modulation, metabolic reprogramming, iron homeostasis, and lung-specific functionality, with several mapping to the chromosomal region 2p11.2, a locus associated with immune dysregulation. Methodological challenges, such as variance threshold selection and reliance on Ensembl IDs for annotation, were addressed through rigorous filtering and cross-referencing. Visualizations, including heatmaps and bar plots, revealed distinct expression patterns, emphasizing tumor heterogeneity and the interplay of genetic and environmental factors. The findings underscore the importance of integrating mean expression and variance metrics to uncover biologically relevant pathways in LUAD. This study provides a foundation for further experimental validation and multi-omic analyses to advance diagnostic and therapeutic strategies for lung cancer.

Introduction

Lung adenocarcinoma (LUAD), a predominant subtype of non-small cell lung cancer (NSCLC), accounts for a significant proportion of cancer-related deaths worldwide. Smoking remains the most significant risk factor, accounting for 90% of lung cancer cases, with passive smoking also increasing risk by 20–30%. Other environmental and occupational factors, such as exposure to asbestos, radon, and specific metals, further compound the risk (Siddiqui 2023). LUAD metastases account for over 70% of LUAD-related deaths, emphasizing the critical need to understand metastatic progression and its clinical implications (Dingbiao et al. 2021). LUAD is characterized by diverse genetic and molecular alterations, contributing to its complexity and resistance to conventional therapies. Single-cell RNA sequencing demonstrated that the tumor microenvironment of lung adenocarcinoma consists of diverse cell populations, including immune and stromal cells, which significantly influence tumor progression and therapeutic outcomes (Bischoff et al. 2021). Understanding the molecular mechanisms driving LUAD progression and heterogeneity is critical for developing effective diagnostic, prognostic, and therapeutic strategies. High-throughput sequencing technologies, such as RNA sequencing, have

revolutionized cancer research by enabling comprehensive analysis of gene expression profiles across various samples and conditions. These datasets provide an opportunity to explore key genes involved in LUAD biology by identifying patterns of high expression and variability.

Gene expression studies in LUAD offer a dual advantage: identifying genes consistently overexpressed across samples can point to potential biomarkers or drivers of the disease, while analyzing expression variability can reveal genes associated with the tumor microenvironment or heterogeneity among patient populations. Highly expressed genes often represent molecular hallmarks of cancer, such as those involved in cell proliferation, apoptosis evasion, or angiogenesis. Conversely, genes with high variance may highlight pathways involved in diverse phenotypic adaptations, resistance mechanisms, or cellular crosstalk. This study utilizes a bioinformatics approach to analyze LUAD gene expression data with a focus on identifying both highly expressed and highly variable genes. Using R Studio, we calculated the mean expression and variance of each gene to pinpoint candidates for further investigation. Visualizations, including bar plots and heatmaps, were employed to summarize the top candidates. The top three highly expressed genes were further explored for their potential biological significance through literature review. This approach aims to provide a "big picture" perspective, highlighting genes that warrant deeper investigation.

While this analysis focuses on computational insights derived from publicly available data, it underscores the importance of scalable and interpretable methods in modern cancer research. By emphasizing key genes and their potential roles in LUAD, this project contributes to the growing body of research aiming to elucidate the molecular underpinnings of lung cancer and pave the way for improved clinical outcomes.

Methods

https://github.com/taqihasnain/LUAD_Project_Repo

(Repository includes Workbooks, which detail the code and scripts used throughout the project)

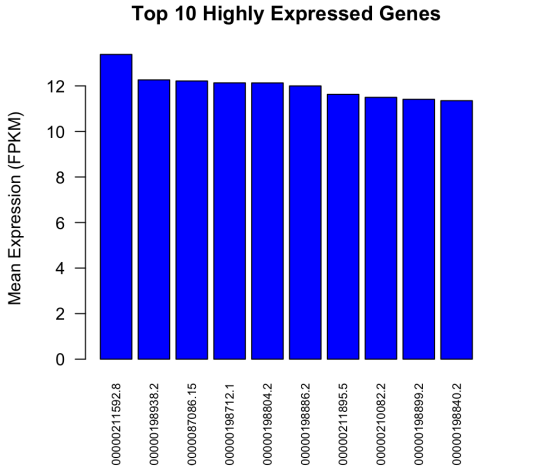
Gene expression data for LUAD was downloaded from UCSC Xena database and was analyzed in R Studio to identify key candidates for further investigation. The data was already processed, and some additional processing was done in R Studio to make the large dataset more manageable for the purposes of this project. Mean expression levels were calculated for each gene, and genes with low expression were filtered out based on a threshold of >5 FPKM to focus on biologically relevant candidates. Variance across samples was computed to pinpoint genes with high variability, as these may indicate differential regulation or heterogeneity, which are hallmarks of cancer biology. This heterogeneity could reflect variations in tumor microenvironments, the presence of driver mutations, or differences in treatment responses.

The top candidates were visualized using bar plots and heatmaps to highlight their expression patterns and variability. Genes with high variance were considered potential

biomarkers due to their capacity to reveal significant correlations with disease states or progression. Conversely, genes with low variance could be investigated for their potential as internal controls or for comparative analyses. These analyses are aimed at uncovering any insights into LUAD pathophysiology and the identification of genes of interest for further exploration.

Results

The combined analysis of mean expression and variance highlights immune system genes, metabolic pathways, and lung-specific proteins as central players in LUAD. Figure 1 shows the top 10 most highly expressed genes (marked by their Ensembl IDs) based on mean expression. Figure 2 is an Expression Heatmap of those same top 10 most highly expressed genes.



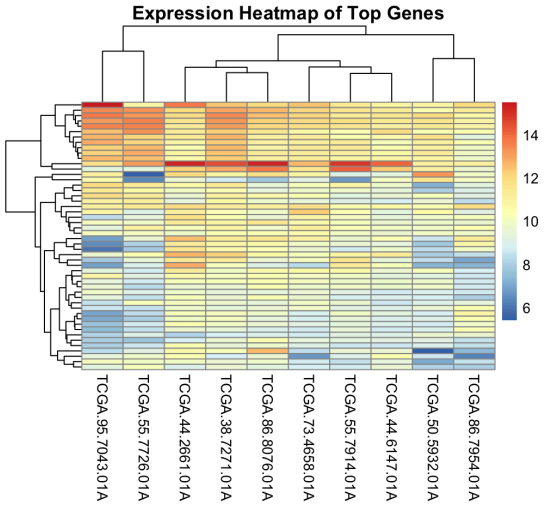
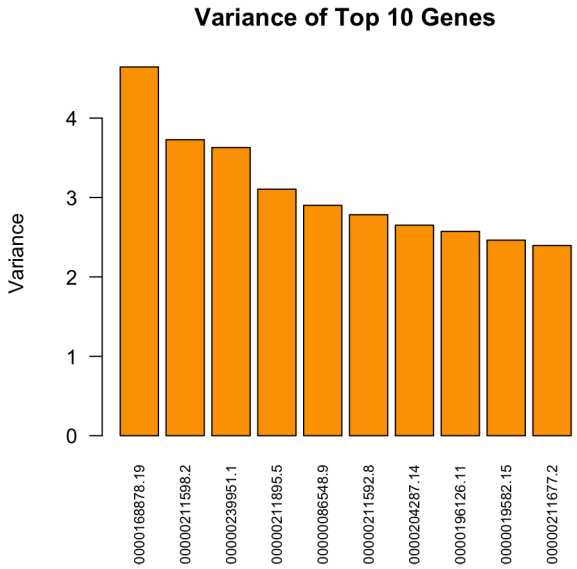
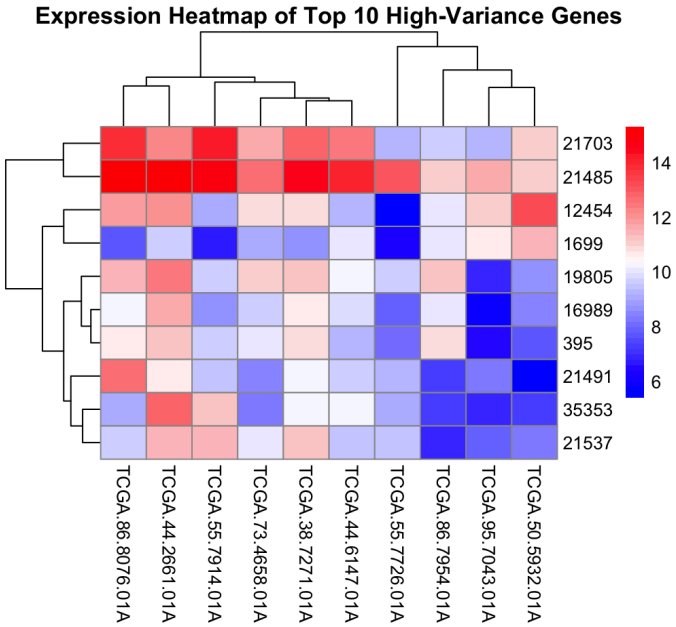


Figure 3 shows the top 10 “high-variance” genes (marked by their Ensembl IDs) from the dataset. Figure 4 is an Expression Heattmap of those top 10 “high-variance” genes.





Based on the mean expression and variance calculations, six genes of interest were identified from the LUAD dataset (using their Ensembl IDs). Notably, multiple genes share significant chromosomal locations or biological functions, offering potential insights into LUAD pathophysiology:

1. Gene A (ENSG00000211592.8):
   - Immunoglobulin kappa constant (IGKC)
   - Chromosomal location: 2p11.2

2. Gene B (ENSG00000198938.2):
   - Mitochondrially encoded cytochrome c oxidase III (MT-CO3)
   - Chromosomal location: mitochondria

3. Gene C (ENSG00000087086.15):
   - Ferritin light chain (FTL)
   - Chromosomal location: 19q13.33

4. Gene D (ENSG00000168878.19):
   - Surfactant protein B (SFTPB)
   - Chromosomal location: 2p11.2

5. Gene E (ENSG00000211598.2):
   - Immunoglobulin kappa variable 4-1 (IGKV4-1)
   - Chromosomal location: 2p11.2

6. Gene F (ENSG00000239951.1):
   - Immunoglobulin kappa variable 3-20 (IGKV3-20)
   - Chromosomal location: 2p11.2

Genes with high mean expression, such as Gene A (ENSG00000211592.8), Gene B (ENSG00000198938.2), and Gene C (ENSG00000087086.15), highlight critical pathways in immune function, metabolic activity, and cellular homeostasis. Gene A (IGKC) is integral to antibody structure, emphasizing its role in the immune response, while Gene B (MT-CO3), a mitochondrial gene, underscores metabolic reprogramming, a hallmark of cancer. Gene C (FTL), involved in iron storage, reflects dysregulated iron metabolism, which is commonly associated with tumor growth and progression.

Variance analysis further identified Gene D (ENSG00000168878.19), Gene E (ENSG00000211598.2), and Gene F (ENSG00000239951.1), which suggest differential regulation or heterogeneity across LUAD samples. Gene D (SFTPB), essential for pulmonary surfactant function, connects directly to lung-specific physiology and LUAD development.

Genes E (IGKV4-1) and F (IGKV3-20), alongside Gene A, all map to the chromosomal region 2p11.2, a locus repeatedly implicated in immune response modulation. This shared location suggests a potential hotspot for immune-related dysregulation in LUAD.

Discussion

This analysis highlights the significance of high-variance genes in reflecting the heterogeneity characteristic of LUAD. Genes such as SFTPB and immunoglobulin kappa-related genes (IGKC, IGKV4-1, IGKV3-20) underscore the interplay between lung-specific processes and immune modulation, while MT-CO3 and FTL emphasize metabolic reprogramming and iron homeostasis, respectively. The repeated identification of the chromosomal locus 2p11.2 further suggests a critical role in immune-related dysregulation within LUAD. Methodologically, challenges included selecting appropriate variance thresholds to capture biologically relevant genes while filtering out noise, as well as navigating reliance on Ensembl IDs for gene annotation. These were addressed through careful thresholding and cross-referencing with the Ensembl database to ensure accurate identification and contextualization of genes. Overall, this study illustrates the potential of integrating expression and variance data to uncover pathways central to LUAD biology and identifies loci warranting further investigation.

Conclusion

This study successfully identified six genes of interest in LUAD based on mean expression and variance, providing valuable insights into the disease's molecular underpinnings. The findings reveal a convergence of immune-related processes (IGKC, IGKV4-1, IGKV3-20), metabolic reprogramming (MT-CO3), iron homeostasis (FTL), and lung-specific functions (SFTPB), with the chromosomal region 2p11.2 emerging as a potentially significant locus for immune-related dysregulation. The presence of surfactant protein B (SFTPB) alongside immune-related genes underscores the complex interplay between lung-specific functions and immune modulation in LUAD, aligning with findings that suggest targeted therapies must address both intrinsic tumor factors and the surrounding tumor microenvironment to enhance therapeutic efficacy (Madeddu et al. 2022). The project's exploratory nature, leveraging computational tools to analyze gene expression variability, underscores the importance of variance in understanding tumor heterogeneity and identifying biomarker candidates.

Strengths of this work include its focus on combining expression levels with variance metrics to capture genes most likely to influence cancer progression and its reliance on rigorous database cross-referencing for gene annotation. Future research could expand on these findings by incorporating experimental validation of identified genes, exploring their roles in tumor microenvironments, and analyzing their interactions in larger, multi-omic datasets. This approach lays the groundwork for integrating computational and biological insights to drive advancements in LUAD research and therapeutic development.

Citations

Bischoff, Philip, et al. “Single-Cell RNA Sequencing Reveals Distinct Tumor Microenvironmental Patterns in Lung Adenocarcinoma.” *Nature News*, Nature Publishing Group, 18 Oct. 2021, www.nature.com/articles/s41388-021-02054-3.

Dingbiao, Li, et al. “Genomic Landscape of Metastatic Lung Adenocarcinomas from Large-Scale Clinical Sequencing.” *Neoplasia*, Elsevier, 1 Nov. 2021, www.sciencedirect.com/science/article/pii/S1476558621000865.

Madeddu, Clelia, et al. “EGFR-Mutated Non-Small Cell Lung Cancer and Resistance to Immunotherapy: Role of the Tumor Microenvironment.” *MDPI*, Multidisciplinary Digital Publishing Institute, 10 June 2022, www.mdpi.com/1422-0067/23/12/6489.

Siddiqui, Faraz. “Lung Cancer.” *StatPearls [Internet]*., U.S. National Library of Medicine, 8 May 2023, www.ncbi.nlm.nih.gov/books/NBK482357/.

Wang, Yuanyuan, et al. “Identification of Key Genes and Biological Pathways in Lung Adenocarcinoma via Bioinformatics Analysis - Molecular and Cellular Biochemistry.” *SpringerLink*, Springer US, 1 Nov. 2020, link.springer.com/article/10.1007/s11010-020-03959-5.

Databases

- UCSC Xena (xena.ucsc.edu)
- Ensembl (www.ensembl.org)
- HGNC (www.genenames.org)

Software/Coding

- R Studio
    - “library(pheatmap)” package

(Repository includes Workbooks, which detail the code and scripts used throughout the project)