
SAPIENZA
UNIVERSITÀ DI ROMA

Data Science Master's Degree

# Differential Analyses of Gene Expression TGCA-LUAD

Digital Epidemiology And Precision Medicine

**Students:**
Concari Laura, 1890490
Prinzi Giuliana, 1952137

---

Academic Year 2023/2024

# Abstract

Lung adenocarcinoma (LUAD), the most common subtype of non-small cell lung cancer, remains a leading cause of cancer-related deaths. This study leverages transcriptomic and mutational data from The Cancer Genome Atlas (TCGA) to uncover key molecular mechanisms in LUAD. Differentially expressed genes (DEGs) reveal upregulated cancer-promoting processes and downregulated tumor-suppressor pathways. Co-expression networks highlight tumor-specific hubs involved in mitosis and DNA repair, while differential co-expression analysis identifies cancer-specific interactions driving protein synthesis and cell division.

A Patient Similarity Network (PSN) was constructed, identifying four molecularly distinct patient subgroups using the Louvain algorithm. Mutational analysis revealed frequent alterations in key genes such as TTN and TP53, linked to immune responses and genomic instability. An enrichment analysis was performed to identify biological pathways and processes associated with DEGs, providing functional insights into the molecular mechanisms driving LUAD progression.


# Introducton

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for 28% of all cancer fatalities. In 2010, approximately 222,520 people in the U.S. were diagnosed with lung cancer, leading to 157,300 deaths. The prognosis remains poor, with a five-year survival rate of only 16% due to late-stage diagnoses.

Among the various subtypes of lung cancer, lung adenocarcinoma (LUAD) is the most common form of non-small cell lung cancer (NSCLC), representing a significant portion of cases. LUAD has been extensively studied by The Cancer Genome Atlas (TCGA). Research has identified key genetic mutations (e.g., EGFR, KRAS, TP53) that drive LUAD progression, but the molecular mechanisms underlying this disease remain incompletely understood, highlighting the need for further research.

This study leverages TCGA transcriptomic data to analyze LUAD through:

- Identifying differentially expressed genes (DEGs) between tumor and normal tissues.
- Constructing co-expression networks to identify central genes and characterize network structures.
- Developing differential co-expression networks to reveal cancer-specific interactions.
- Creating patient similarity networks to identify molecular subgroups and biomarkers.

The goal is to deepen the understanding of LUAD pathogenesis, a major subtype of lung cancer, and discover potential molecular targets to improve patient outcomes.

# Materials And Methods

## Data Collection, Preprocessing, and Normalization

RNA-Seq data for lung adenocarcinoma were retrieved from The Cancer Genome Atlas (TCGA) using the `TCGAbiolinks` package. The dataset included raw gene expression counts from tumor (primary tumor) and normal (solid tissue normal) samples. Clinical data were also downloaded to provide additional information about the patients.

The preprocessing steps included:

- Removing duplicate tumor samples to retain only one sample per patient.

- Excluding normal samples without a matched tumor sample, ensuring a dataset of paired tumor and normal samples.

This resulted in a dataset with matched samples, suitable for comparative analyses. The raw count data were normalized using the `DESeq2` package. This process corrected for differences in sequencing depth across samples and ensured that the data were comparable. Genes with low expression across samples were filtered out to retain only those with sufficient expression for downstream analyses.

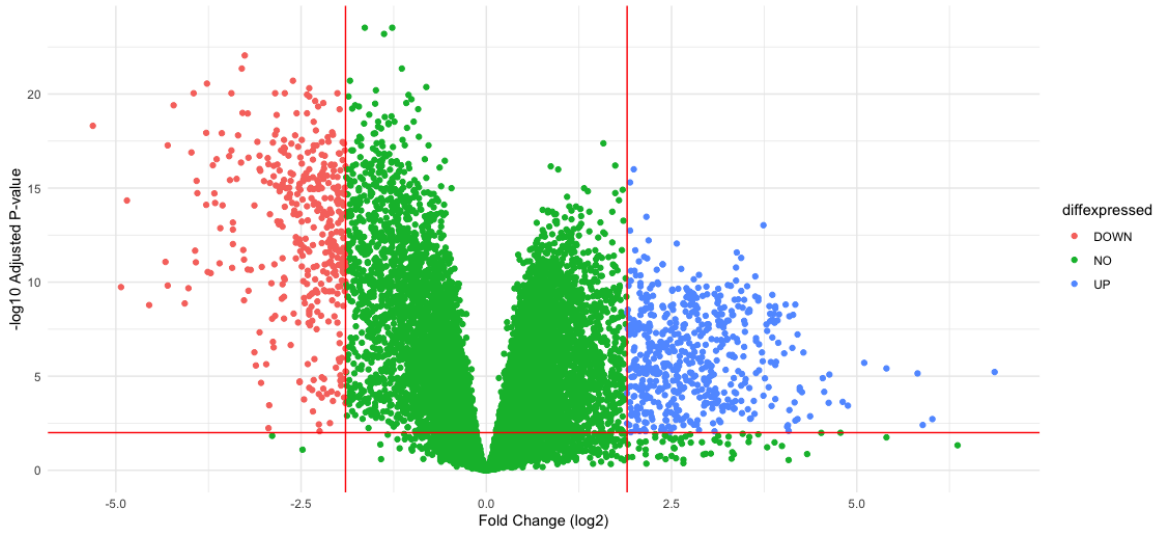
## Differential Gene Expression in Lung Adenocarcinoma

To investigate gene expression differences between tumor and normal tissues in lung adenocarcinoma, a combination of fold change analysis and statistical testing was applied. The $\log_2$ fold change (FC) for each gene was calculated to measure the magnitude of expression differences, and paired t-tests were performed to evaluate their statistical significance. To control for false positives, p-values were adjusted using the False Discovery Rate (FDR) method. Genes with $|\text{FC}| \geq 1.9$ and an FDR-adjusted p-value $\leq 0.01$ were classified as differentially expressed.

The analysis identified 898 differentially expressed genes (DEGs), including:

- 541 upregulated genes, which were more highly expressed in tumor samples. These genes likely drive key cancer processes, such as cell proliferation, survival, and metabolic reprogramming.
- 357 downregulated genes, which showed lower expression in tumor samples. These are often linked to tumor suppressor functions, immune signaling, or tissue maintenance, whose loss may enable cancer progression.

The results were visualized using a volcano plot, highlighting the predominance of upregulated genes. This reflects the activation of cancer-related pathways and provides potential insights into biomarkers and therapeutic targets for LUAD.

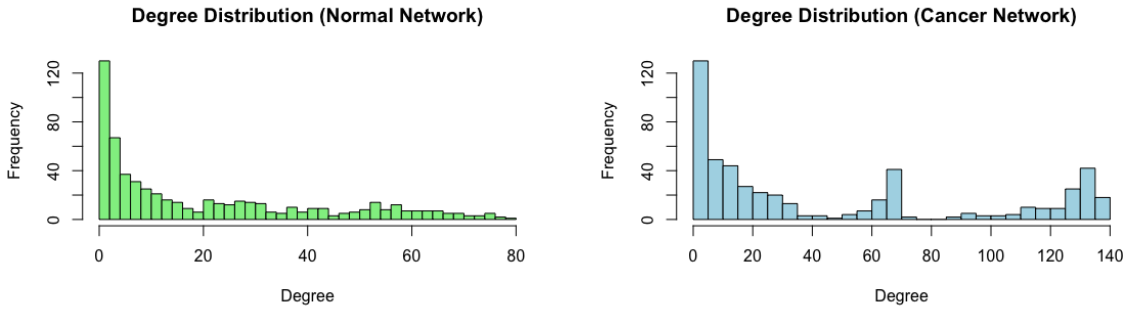



The cancer can be categorized into three distinct subtypes based on gene expression data: Proximal proliferative, proximal inflammatory, and terminal respiratory unit. In order to identify these subtypes we analyzed the differentially expressed genes (DEGs) and mapped them to known subtype-specific markers. This process revealed IL6 as associated with the Inflammatory subtype and SFTPC with the TRU subtype, enabling us to classify patients based on their expression profiles and link these classifications to enriched biological pathways for further validation.

## Co-expression networks

The study of co-expression networks in lung adenocarcinoma highlights how gene interactions are reorganized during cancer progression. Using Pearson correlation, networks were constructed for tumor and normal tissues, where genes are nodes and significant co-expression relationships ($|\text{correlation}| \geq 0.7$) are edges. This analysis revealed distinct connectivity patterns between the two conditions, with the tumor network showing more highly connected genes, suggesting a reorganization of interactions to support cancer-specific processes such as proliferation and survival.

To explore the structure of these networks, we examined whether they followed a scale-free topology, a common property of biological systems. Degree distributions were analyzed, and histograms were created to visualize the connectivity patterns. These histograms provide a visual indication of whether the networks exhibit a scale-free structure. Despite the R-squared values of 0.35 for the tumor network and 0.68 for the normal network in the log-log plots of degree distributions—indicating that neither network strongly adheres to a scale-free topology—we identified the hubs based on the observation of peaks of similar height in the histograms, which suggest the presence of highly connected genes.




It is plausible that the co-expression network does not follow a scale-free topology because LUAD is biologically heterogeneous, characterized by a complex tumor microenvironment and a high mutational burden, both of which can disrupt typical scale-free structures. Nevertheless, the distinct hubs identified between the tumor and normal conditions underline critical differences in network organization and highlight genes that may play pivotal roles in LUAD pathogenesis.

We analyzed the unique hubs of tumor and normal co-expression networks in lung cancer, using the mapping of Ensembl identifiers to gene symbols to interpret their biological significance. Our analysis highlighted significant differences between the two conditions. In the tumor network, hubs such as BUB1, TTK, and RAD54L regulate critical processes such as mitosis, DNA repair, and cytoskeletal dynamics, promoting uncontrolled proliferation and genomic instability. In the normal network, on the other hand, hubs like CCNA2 and HMMR maintain homeostasis and tissue stability, functions that are lost in the tumor context.

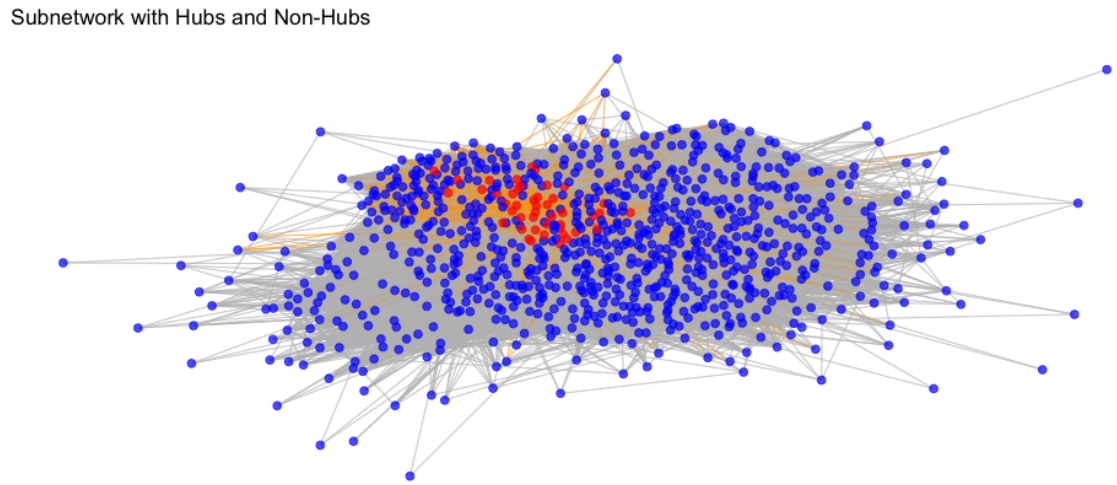
These findings highlight how cancer reorganizes genetic networks to support growth, making tumor hubs potential therapeutic targets. Conversely, preserving normal hubs could help prevent disease progression.

## Differential Co-expression Networks in Lung Adenocarcinoma

To investigate how gene interactions change between tumor and normal conditions in lung adenocarcinoma , we constructed a Differential Co-expression Network. This network highlights gene pairs

with significantly altered co-expression relationships between the two conditions. Using Fisher's Z-transformation, we compared the Pearson correlation coefficients from tumor and normal samples and calculated differential Z-scores. Gene pairs with an absolute Z-score greater than 3 were identified as significantly altered and included in the DCN.

The DCN revealed a distinct reorganization of gene interactions. Analysis of the degree distribution showed that the network does not strongly follow a scale-free topology (R-squared $< 0.8$).




Despite this, the most connected nodes (top 5% by degree) were identified as hubs, representing genes with the most altered interactions. This hubs reveal significant molecular shifts. Many hubs are ribosomal proteins (RPL13, RPL13A, RPS6), reflecting increased protein synthesis to support rapid tumor growth. Other key hubs include KIF20A, highlighting dysregulated cell division, and TMEM132A, indicating changes in cellular adhesion and the tumor microenvironment. Additionally, AOC1 and RPS6 point to altered metabolism and signaling. These findings suggest a reorganization of gene interactions in LUAD, providing potential therapeutic targets and emphasizing the role of translational dysregulation and cell division in tumor progression.

## Patient Similarity Network

The Patient Similarity Network (PSN) is used to identify relationships between patients based on similarities in their gene expression profiles. In this context, "similarities" mean that patients share similar expression patterns of differentially expressed genes (DEGs). These similarities may indicate that patients have common biological characteristics, such as molecular alterations, tumor subtypes, or potential responses to specific therapies.

We constructed a PSN to analyze how the gene expression profiles of LUAD patients are similar to each other. In this network, patients are represented as nodes, and the connections between them (edges) indicate a significant correlation in their DEG expression profiles. To build the PSN, we calculated a correlation matrix based on the DEG expression profiles using Pearson's correlation method, which quantifies the similarity between each pair of patients. We included only relevant connections by applying a threshold (only correlations with an absolute value greater than or equal to 0.7), creating a binary adjacency matrix that retains only strong correlations. This approach allowed us to focus on meaningful relationships within the network.

The Patient Similarity Network highlights molecular relationships among patients. The network

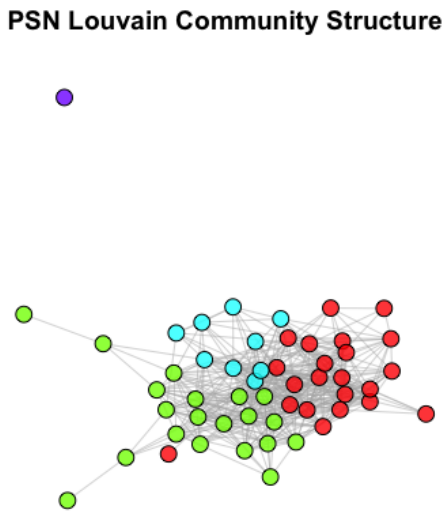
has 472 connections and a density of 0.355, indicating a moderately dense structure consistent with LUAD's biological heterogeneity.

Key hubs (*TCGA-55-6979, TCGA-55-6982, TCGA-55-6985*) represent patients with highly similar gene expression profiles, suggesting shared tumor subtypes or molecular features. The central core of the network reflects groups of biologically similar patients, while peripheral nodes highlight unique profiles that may require further investigation.

This network supports the identification of patient subgroups and provides insights into LUAD's molecular organization, aiding in personalized medicine and targeted therapy development.

The Louvain community detection algorithm applied to the Patient Similarity Network (PSN) identified 4 distinct communities of LUAD patients, with sizes of 22, 20, 9, and 1, reflecting the molecular heterogeneity of the disease. These communities represent groups of patients whose gene expression profiles are more similar within the group than with others, driven by shared molecular mechanisms such as genetic mutations, pathway activation, or tumor subtypes.

The larger communities (22 and 20 patients) correspond to more common LUAD subtypes, characterized by prevalent molecular patterns linked to processes like cell proliferation or immune responses. In contrast, the smaller community (9 patients) and the single-node community highlight rare molecular profiles or unique tumor characteristics, potentially revealing novel subtypes or biological behaviors.




This clustering reflects the natural variability in LUAD's gene expression, shaped by genetic differences and tumor microenvironment diversity. The analysis underscores the PSN's ability to group patients into biologically meaningful subgroups, supporting personalized medicine by identifying distinct tumor subtypes and enabling tailored therapeutic strategies.

The mutational analysis of LUAD patients revealed that 45 out of 52 samples (86.54%) had significant genomic alterations, with frequent mutations in key genes such as TTN (44%), linked to high tumor mutational burden (TMB), MUC16 (37%), associated with immune evasion and cancer progression, and TP53 (37%), a critical tumor suppressor driving genomic instability. Other notable genes included RYR2 (37%), involved in calcium signaling, and LRP1B (27%), associated with genomic stability and immune checkpoint therapy responses.

The mutational data were processed into a binary mutation matrix, capturing whether specific genes were mutated in each patient. Using this matrix, a Jaccard similarity index was calculated to measure mutational similarity between patients. This mutational similarity network was normalized and integrated with a gene expression similarity network using the Similarity Network Fusion (SNF) algorithm. SNF iteratively combined the two data layers, creating a fused network that reflects comprehensive patient relationships based on both mutational and expression profiles.

## Additional Analysis

### Betweenness Centrality Analysis in Cancer and Normal Networks

Betweenness centrality identifies nodes that act as bridges in a network, critical for communication and connectivity. We analyzed the top 5% of nodes by betweenness centrality in Cancer and Normal networks and compared them with degree-based hubs. In the Cancer network, 3 hubs overlapped, while in the Normal network, 6 hubs were shared. These results highlight nodes that are both highly connected and pivotal for network flow, reflecting their biological significance and the reorganization of molecular interactions in cancer.

### Differential Co-Expression Network: Positive and Negative Subnetworks Analysis

The analysis identified the most connected nodes (hubs) in two subnetworks: a positive subnetwork, representing genes with strong positive relationships, and a negative subnetwork, representing genes with strong negative relationships.

Positive hubs, such as *ENSG00000171049.9*, highlight genes that are highly connected through aligned activity patterns, likely reflecting co-regulated pathways or cooperative functions. Negative hubs, including *ENSG00000175832.13*, represent genes with strong inverse relationships, potentially involved in opposing biological processes or regulatory mechanisms. These findings provide valuable insights into the network's structure and reveal key genes driving synergistic or antagonistic interactions.

### Enrichment Analysis

The analysis begins with identifying patients who exhibit high expression levels of IL6 and SFTPC, two critical markers for lung adenocarcinoma subtypes. These patients were then mapped to communities within the Patient Similarity Network (PSN) and clusters in the Fused Network, which integrate gene expression and mutational data. The goal was to uncover enriched biological pathways that explain the grouping of patients and their connection to the known molecular roles of IL6 and SFTPC.

IL6 is a key marker for the proximal inflammatory subtype, and the enriched pathways associated with it include inflammatory and cytokine signaling pathways. These pathways highlight IL6's pivotal role in promoting inflammation, driving tumor progression, and modulating the tumor microenvironment. Pathways like the regulation of cytokine production are directly linked to IL6's function in immune response and its involvement in creating a pro-inflammatory tumor microenvironment. These findings reinforce the connection between IL6 expression and the proximal inflammatory subtype of lung adenocarcinoma, characterized by aggressive behavior and immune interactions.

SFTPC, on the other hand, is a marker for the terminal respiratory unit subtype, a group associated with more differentiated and less aggressive tumors. Enriched pathways for SFTPC include processes related to cell differentiation, lung epithelial development, and terminal respiratory function. For example, pathways like "negative regulation of cell division" and "cellular response to oxidative stress" suggest the involvement of SFTPC in maintaining the integrity and functionality of mature respiratory cells. These pathways underline the less aggressive and more specialized characteristics of this subtype, correlating with its role in terminal respiratory units.

Interestingly, shared pathways such as spindle assembly checkpoint signaling and mitotic spindle checkpoint signaling were found enriched in both IL6 and SFTPC groups in certain communities and clusters. These pathways are crucial for cell cycle regulation and mitotic fidelity, which are often dysregulated in cancer. Their presence in both gene groups indicates overlapping mechanisms of tumor progression, such as unchecked cell proliferation. This shared enrichment suggests that even

though the biological functions of IL6 and SFTPC differ, some patient groups exhibit convergent molecular pathways related to tumor growth and maintenance.

The relevance of these pathways lies in their ability to distinguish between subtypes of lung adenocarcinoma and provide a molecular framework for understanding tumor behavior. Pathways enriched in IL6-dominated groups, such as those related to inflammation, suggest potential targets for immunotherapy or cytokine-modulating treatments. Similarly, pathways enriched in SFTPC-dominated groups highlight opportunities for therapies aimed at preserving lung functionality and preventing aggressive transitions.

Overall, the enriched pathways not only clarify the biological significance of IL6 and SFTPC expression but also offer actionable insights into subtype-specific tumor biology. These results bridge molecular data with clinical implications, providing a pathway-centric view of tumor heterogeneity that could guide personalized therapeutic strategies for lung adenocarcinoma.

# References


[1] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 211(2):543–550, 2014.

[2] National Cancer Institute. Lung adenocarcinoma study - the cancer genome atlas. `https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/lung-adenocarcinoma-study`, 2024. Accessed on December 12, 2024.

[3] GeneCards. Rad54l gene - genecards summary. `https://www.genecards.org/cgi-bin/carddisp.pl?gene=RAD54L`, 2024. Accessed on December 12, 2024.