Differential Gene Expression Analysis of Healthy and Early-Stage Lung Adenocarcinoma (LUAD) to Identify Potential Early Detection Biomarkers Using Bioinformatics

ABSTRACT

Background

Lung adenocarcinoma is the most common subtype of lung cancer and a leading cause of cancer death, largely because most cases are diagnosed only after it has advanced. Early-stage tumors rarely cause symptoms and are often found only by accident, creating a real gap in early detection. This study aims to identify differentially expressed genes between healthy and early-stage lung adenocarcinoma tissue and examine their functions and pathways.

Methods

We used the NCBI GEO dataset GSE31210, comparing gene expression in healthy lung tissue and early-stage lung adenocarcinoma tumor tissue. Using GEO2R, we identified differentially expressed genes and selected 50 genes using adjusted p value and log fold change. These genes were analyzed with SR Plot for KEGG and GO enrichment.

Results

From 34,753 genes, we found 19,922 significantly differentially expressed genes, narrowed to the top 25 upregulated and 25 downregulated genes. KEGG analysis showed enrichment in the Adherens Junction pathway, with VE-Cadherin and Rasip1 found significant. GO analysis revealed enrichment in vasculogenesis, ameboidal-type cell migration, stem cell proliferation, and smooth muscle cell differentiation, with CDH5, MNX1, SFN, and RASIP1 standing out.

Conclusion

In early-stage lung adenocarcinoma tissue, VE-Cadherin, Rasip1, and CDH5 were differentially expressed and linked to blood vessel formation and tumor angiogenesis. The GO and KEGG results together suggest that even at the earliest detectable stage, tumors may already be building

the blood supply needed to grow. These results may contribute to blood-based early detection tools and targeted anti-angiogenic treatments for lung adenocarcinoma.

INTRODUCTION

In lung cancer, cells within the lungs begin dividing without control, and the disease remains one of the top causes of cancer related death across the United States [1]. The disease is made up of two broad subtypes, small cell lung cancer and non-small cell lung cancer (NSCLC), and NSCLC is responsible for roughly 80% of all cases [2, 3]. Lung adenocarcinoma (LUAD), a subtype of NSCLC, is itself the single most common form of lung cancer, and it represents close to 40% of cases overall [3]. It tends to develop in the lung's outer, peripheral regions, and it carries a molecular signature distinct from other subtypes [2]. LUAD stands out for another reason too, since it shows up disproportionately in never-smokers, women, younger patients, and people of East Asian descent, and this risk profile breaks from the smoker dominated pattern seen in most other lung cancers [2].

For LUAD and lung cancer more broadly, the central problem is timing, since most patients are not diagnosed until the disease has already advanced. This happens largely because early stage tumors rarely cause symptoms, and they are usually spotted only by accident, on scans ordered for something unrelated [2]. The numbers reflect this. Nationally, just 27.4% of lung cancer cases are caught early, a stage where five-year survival sits near 64%, while 43% are not identified until late stage, when that survival figure falls to only 9% [1]. Screening does exist in the form of low-dose CT, yet only around 16% of eligible people nationally actually get scanned, and the test itself is not perfect, since it carries an estimated 25% false-positive rate within the first year [1, 2]. Together, these gaps mean a large share of LUAD cases, particularly among people who do not fit the standard high-risk screening profile, go undetected until it is too late to catch them early.

This study asks whether healthy lung tissue and early-stage LUAD tumor tissue differ at the level of gene expression, and whether any such differences are already detectable at the earliest stage disease is typically caught clinically, before a patient would ever show symptoms. More specifically, it looks at which genes differ most significantly in expression between the two tissue types, and whether those genes cluster within biological pathways tied to cancer development, a pattern that would suggest their promise as early detection biomarkers.

Molecularly, LUAD is far from a single disease, since the driver mutations behind it vary considerably across patient populations. Clinically actionable alterations, including EGFR, ALK, KRAS G12C, and ROS1, show up often enough to matter, and this genetic variety has reshaped how LUAD is treated. What was once managed almost entirely with chemotherapy is now among the most genomically stratified cancers in oncology, with numerous FDA-approved targeted therapies matched to specific genomic subtypes [2]. Even with these treatment gains, though, LUAD still ranks among the deadliest cancers, largely because so many cases are not

caught until they have already advanced [2]. Beyond the tumor cells themselves, the surrounding microenvironment, including immune cells and other cell populations nearby, also shapes how the cancer grows, spreads, and responds to treatment [3].

Prior work in this space has leaned heavily on public gene expression repositories, with NCBI's Gene Expression Omnibus (GEO) chief among them, to pinpoint differentially expressed genes and candidate prognostic biomarkers in LUAD. Bioinformatics methods in particular have been applied to trace genes tied to immune cell infiltration, interactions within the tumor microenvironment, and signaling pathways connected to how LUAD progresses and responds to treatment [3]. Researchers commonly use tools such as GEO2R to run statistical comparisons of gene expression across sample groups, and they then turn to functional enrichment analysis to figure out which biological pathways those genes belong to.

Despite a substantial body of bioinformatics research mapping gene expression signatures and cellular interactions in LUAD, mostly in advanced or metastatic disease, much less is known about which genes are already differentially expressed at the earliest stage where clinical action is possible. Complicating things further, LUAD is usually discovered by chance rather than through deliberate screening, and a large fraction of patients do not match the traditional high-risk profile screening programs are built around. That combination makes it genuinely difficult to pin down which molecular signals could realistically extend earlier detection to a wider range of patients [1].

This research aims to pinpoint genes whose expression differs significantly between healthy individuals and patients with early-stage LUAD, and to characterize the biological functions and pathways those genes are enriched in, ultimately to assess how promising they are as candidate early detection biomarkers.

My hypothesis is that a meaningful set of genes will already be significantly differentially expressed at the early stage of LUAD relative to healthy tissue. I also expect that these genes could point toward biomarkers useful for earlier diagnosis, not as a replacement for current screening methods like low-dose CT, but as a complement to them.

The significance here comes down to a real gap in lung cancer detection today. Nationally, only about 27% of cases are caught early, and only 1 in 5 eligible people actually get screened [2,4]. If certain genes turn out to be reliably differentially expressed even at LUAD's earliest stage, that finding could eventually support new detection tools to work alongside imaging, extending earlier identification of cancer to a broader group of patients, including those who do not currently meet high-risk screening criteria.

METHODS

Data Collection and Analysis of GEO2R Data

To start my investigation, I went to the National Center for Biotechnology Information and looked for GEO Datasets about lung adenocarcinoma. GEO2R is an online tool that is offered through the National Center for Biotechnology Information. It helps analyze gene expression across various samples [5]. Eventually, I selected the dataset GSE31210 which is called "Gene expression data for pathological stage I-II lung adenocarcinomas." This dataset contains microarray gene expression data for healthy lung tissue and early-stage lung adenocarcinoma tumor tissue samples. Figure 1 summarizes the methods and bioinformatics tools used in this study.

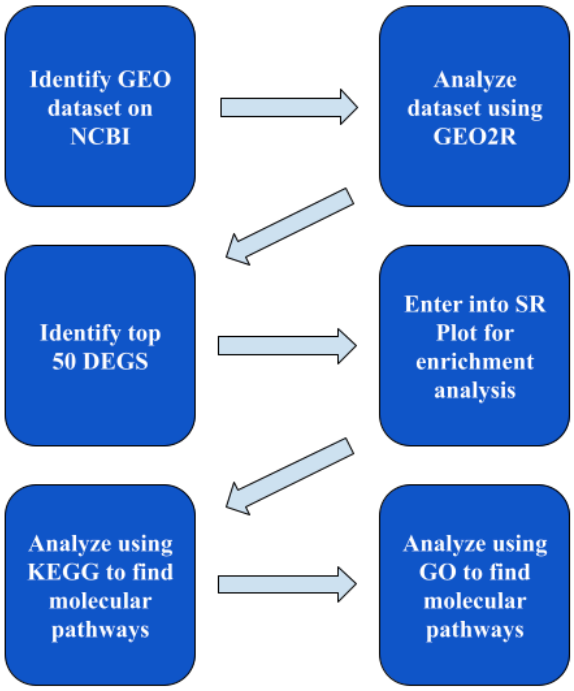

Figure 1: Research Methodology: The steps and bioinformatics used in this study.

Identification of the Top Differentially Expressed Genes

To identify the top 50 most significant differentially expressed genes, statistical analysis was applied. This process used both the adjusted p value and the Log Fold Change (logFC) value. First, genes were sorted by adjusted p value so that the most statistically significant genes were prioritized. Next, 25 genes with the highest logFC value and 25 genes with the lowest logFC value were selected. This process helped find the 50 most significant differentially expressed genes.

Functional and Enrichment Analysis Using SRPlot, KEGG, and GO Bioinformatics Tools

Then SR Plot, KEGG, and GO bioinformatics tools and databases were utilized to analyze the functions of these top genes [6,7,8]. These tools helped uncover the potential roles of the genes in specific molecular pathways, biological processes, and disease processes. First, the gene symbols and logFC values from the top 50 differentially expressed genes were entered into SR Plot for enrichment analysis. Through SR Plot, I used KEGG to identify 4 specific pathways. The molecular pathways found were Adherens Junction, Melanogenesis, Leukocyte Transendothelial Migration, and Glycosaminoglycan Degradation. Through the Adherens Junction pathway, the most relevant out of the four, I was able to identify the VE-Cadherin and Rasip1 genes as being the most significant. Using the GO biological process, I was able to identify several specific pathways, with vasculogenesis, ameboidal-type cell migration, and stem cell proliferation being the most significant. A cnet plot identified specific genes connected to these processes, including CDH5, MNX1, and SFN.

RESULTS

Identification of Differentially Expressed Genes

To identify differentially expressed genes I used GEO2R. From my first two results, 19,922 genes out of the 34,753 total genes expressed differently between the two groups of healthy lung tissue and early-stage LUAD tumor tissue.

From my volcano plot (Figure 2a), the red dots represent upregulated genes, or genes with higher expression levels, in early-stage LUAD tissue compared to healthy tissue, while the blue dots represent downregulated genes, or genes with lower expression levels, in the tumor tissue as opposed to healthy tissue. The black dots represent genes that were not significantly differentially expressed between the two groups.

From the Venn diagram (Figure 2b), in total the study produced 34,753 genes. Out of these genes, there were 19,922 differentially expressed genes that overlapped between healthy lung tissue and early-stage LUAD tumor tissue. This large overlap shows that even at the earliest clinically detectable stage, LUAD tumor tissue already shows a strong, widespread difference in gene activity compared to healthy tissue.

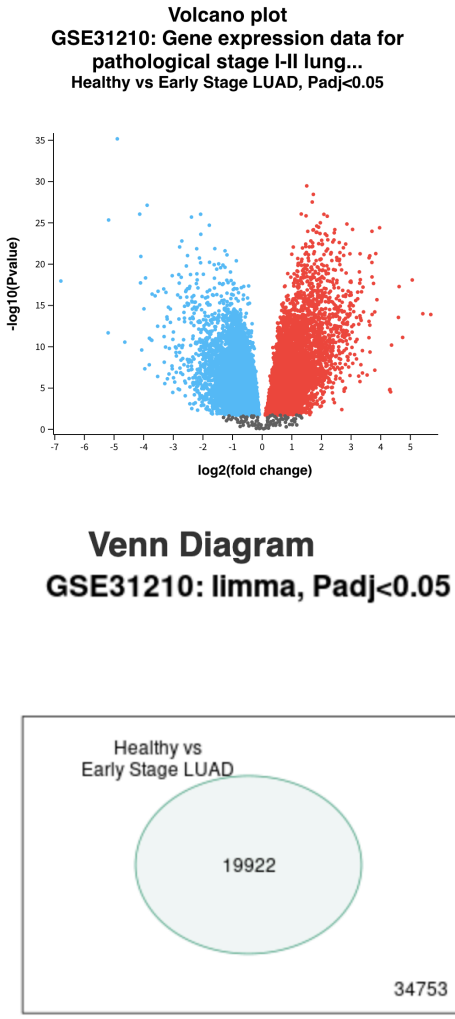

Figure 2: Differentially Expressed Genes: The volcano plot shows the distribution of the differentially expressed genes. The genes in red are the upregulated genes and the genes in blue are the down regulated genes. The venn diagram shows that 19,922 genes out of the 34,753 total genes are significantly differentially expressed between healthy and early-stage LUAD tissue.

Identification of 50 Statistically Significant Differentially Expressed Genes (DEGs)

To narrow down my genes, I used both the adjusted p value and the fold change value. Using these, I narrowed my list to 50 differentially expressed genes. I chose 25 upregulated genes and 25 downregulated genes. (Top 50 DEGs)

Potential Functions and Enrichment of the Identified Genes and/or pathways

I used SR Plot to determine the potential functions of the genes. From the KEGG results, I identified Adherens Junction as a significantly enriched pathway. Within this pathway, the genes that stood out were VE-Cadherin and Rasip1 (Figure 3a).

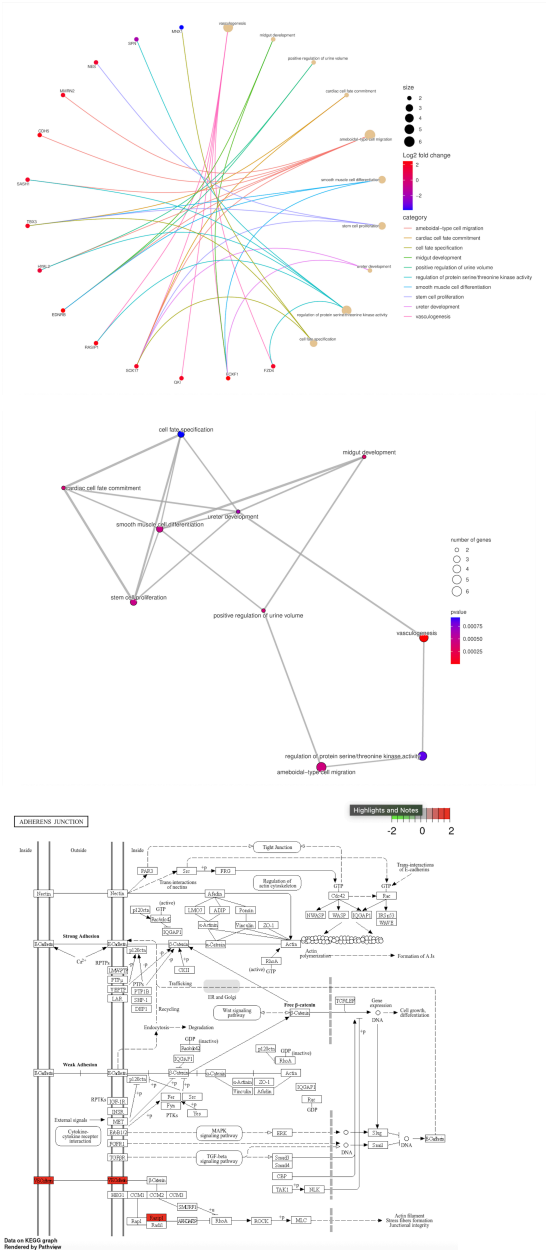

Figure 3: KEGG Functional and Enrichment Analysis. This figure shows the Adherens Junction pathway, the most significantly enriched KEGG pathway identified for the top 50 DEGs, with the genes VE-Cadherin and Rasip1 (highlighted in red) shown as upregulated.

From the GO results, I identified several significant biological processes across all three ontology categories, Biological Process, Cellular Component, and Molecular Function (Figure

4a). Within the Biological Process category specifically, vasculogenesis, ameboidal-type cell migration, and stem cell proliferation were among the most enriched pathways (Figure 4b). A cnet plot identified specific genes connected to these processes, including CDH5, MNX1, and SFN (Figure 4c), and a network map showed how these biological process terms relate to and overlap with each other (Figure 4d).

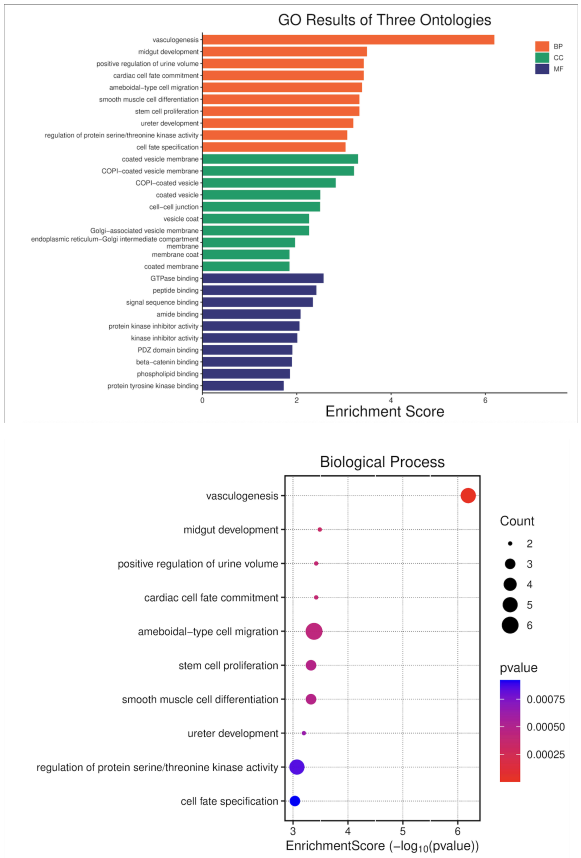

Figure 4: GO Functional and Enrichment Analysis. a) This figure shows the significant Gene Ontology terms across all three ontology categories, Biological Process (BP), Cellular Component (CC), and Molecular Function (MF), with vasculogenesis showing the highest enrichment score overall. b) This figure shows the top enriched Biological Process terms in more detail, with dot size representing gene count and color representing p-value, with vasculogenesis being the most significant.

Table 1: Summary of Key Pathways and Genes Identified in this Study

Key Pathways	**Genes**	**Connection to Early Stage Lung Adenocarcinoma and Future Research**
Adherens Junction	VE-Cadherin, Rasip1	These genes are tied to blood vessel formation and stability, suggesting early-stage LUAD tumors may already be building the vascular infrastructure needed to grow. Future research should test whether VE-Cadherin and Rasip1 could serve as blood-based biomarkers for earlier detection.
Vasculogenesis, ameboidal-type cell migration, stem cell proliferation, smooth muscle cell differentiation	CDH5, MNX1, SFN, RASIP1	These pathways are all connected to blood vessel development and tumor angiogenesis, reinforcing that vascular changes occur even before a tumor is large enough to detect through imaging. Future research should explore whether these genes remain differentially expressed as cancer progresses to later stages, and whether they could support development of anti-angiogenic treatments targeted at early-stage patients.

DISCUSSION

Summary of Findings

The goal of this research study was to identify differentially expressed genes between healthy lung tissue and early-stage lung adenocarcinoma (LUAD) and analyze their biological functions using bioinformatics tools. From the GEO dataset GSE31210, a total of 34,753 genes were analyzed (Figure 2b). Using a Venn diagram, it was seen that 19,922 of these genes were significantly differentially expressed between the healthy and early-stage LUAD groups (Figure 2b). After using statistical analysis such as adjusted p value and fold change to eliminate insignificant genes, the list was narrowed down to the top 50 differentially expressed genes. With

SR Plot, a number of significantly enriched pathways were obtained from the KEGG and GO Biological Process pathways. The top enriched KEGG pathway was Adherens Junction, and within this pathway, the genes VE-Cadherin and Rasip1 stood out (Figure 3). The top enriched GO Biological Process pathways were vasculogenesis, ameboidal-type cell migration, stem cell proliferation, and smooth muscle cell differentiation (Figure 4).

Interpretation of Results

GEO2R analysis results indicate significant gene expression changes in early-stage lung adenocarcinoma (LUAD) tumor tissue compared to healthy lung tissue. The top 50 differentially expressed genes indicate that they could be of significance in the earliest stages of cancer development, before a tumor would typically be caught through imaging. The KEGG analysis indicated that these genes were primarily associated with the Adherens Junction pathway (Figure 3). This pathway showed the highest enrichment score out of the pathways identified. Adherens junctions are structures that hold cells together, and disruption of these junctions is a well known mechanism that allows cancer cells to break away from surrounding tissue and begin to spread [3]. The occurrence of VE-Cadherin and Rasip1 within this pathway is notable since both genes are specifically tied to blood vessel formation and stability, not just general cell adhesion. The GO biological process results supported this same theme, identifying vasculogenesis as the single most enriched process, along with ameboidal-type cell migration and smooth muscle cell differentiation, both of which are also connected to blood vessel development (Figure 4). This consistent pattern across both KEGG and GO results suggests that even at the earliest stage, LUAD tumor tissue may already be building the blood supply infrastructure needed to grow, a process known as tumor angiogenesis, which is one of the well established hallmarks of cancer.

Comparison with Previous Studies

Other studies have also had similar findings. A study that examined vasculogenic mimicry, a process related to tumor vasculogenesis, in lung adenocarcinoma found that genes involved in blood vessel formation were significantly associated with patient prognosis and the tumor immune landscape [9]. This is in line with the GO enrichment results, which show vasculogenesis as the most significantly enriched biological process. Other studies have also highlighted the role of cadherin family genes specifically in lung adenocarcinoma angiogenesis. One study found that a related cadherin gene promoted angiogenesis in tumor-derived endothelial cells and had prognostic significance specifically in adenocarcinoma tissue compared to other lung cancer subtypes [10]. Another study found that cadherin expression in the tumor microenvironment was associated with patient outcomes in EGFR-mutant lung adenocarcinoma specifically [11]. This supports the KEGG enrichment results, which show VE-Cadherin as one of the significant genes within the Adherens Junction pathway.

Implications

This study provides insight on how gene expression changes in lung tissue during the earliest stage of lung adenocarcinoma, before a tumor would typically be caught through imaging. Practical applications of these findings may include identifying biomarkers for earlier detection of LUAD and helping identify which patients are at higher risk of aggressive tumor growth. By specifically identifying genes like VE-Cadherin and Rasip1, which are tied to how tumors build their own blood supply, anti-angiogenic therapies, treatments that work by cutting off a tumor's blood supply, become a more targeted possibility for early-stage patients rather than only late-stage ones. These findings may contribute to the development of a blood-based screening test that works alongside current imaging methods like low-dose CT, potentially helping catch LUAD earlier in a wider range of patients, including those who fall outside current high-risk screening criteria.

Limitations

In our case, since we performed secondary research using bioinformatics datasets from microarray experiments conducted by other researchers, one limitation is that the identified genes and biomarkers will need to be further studied in a laboratory or clinical environment before they could be developed into an actual diagnostic test or treatment.

Future Directions

The identified genes can be tested in the laboratory by scientists in the laboratory or clinical trials to determine if they play a direct role in the early development of lung adenocarcinoma. Through further study, researchers can deduce whether these genes can serve as accurate biomarkers for early detection and their relevance in the development of targeted treatments. Further research could also explore whether these genes remain differentially expressed as the cancer progresses from early to late stage, whether they differ between smoking-related and pollution-related cases of LUAD, and whether true Stage 0 samples show the same gene expression signature as the early-stage samples used in this study.

REFERENCES

1. American Lung Association. (2024, November 19). New report: Lung cancer survival rate improves, but gaps in biomarker testing and lack of screening hinder progress.
2. Kanchustambham, V., & Sharma, S. (2026). Lung adenocarcinoma. In StatPearls. StatPearls Publishing.
3. Han, M., Wan, F., Xiao, B., Du, J., Peng, C., & Peng, F. (2024). Cell components of tumor microenvironment in lung adenocarcinoma: Promising targets for small-molecule compounds. Chinese Medical Journal, 138(8), 905–915.
4. American Cancer Society. (2025, November 19). Latest ACS lung cancer data: Only 1 in 5 eligible adults in U.S. screened for lung cancer; 62,000 lives over 5 years could be saved if all eligible screened.
5. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H.,

Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets—update. Nucleic Acids Research, 41(D1), D991–D995.

6. Tang, D., Chen, M., Huang, X., Zhang, G., Zeng, L., Zhang, G., Wu, S., & Wang, Y. (2023, November 9). SRplot: A free online platform for data visualization and graphing. PLOS ONE.
7. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 27(1), 29–34.
8. Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261.
9. Yang, W., Li, Z., Wang, W., Wu, J., Li, J., Huang, X., Zhang, X., & Ye, X. (2023). Vasculogenic mimicry score identifies the prognosis and immune landscape of lung adenocarcinoma. Frontiers in Genetics, 14, Article 1206141.
10. Zhuo, H., Zhao, Y., Cheng, X., Xu, M., Wang, L., Lin, L., Lyu, Z., Hong, X., & Cai, J. (2019). Tumor endothelial cell-derived cadherin-2 promotes angiogenesis and has prognostic significance for lung adenocarcinoma. Molecular Cancer, 18(1), 34.
11. Chang, Y. P., Huang, G. K., Chen, Y. C., Huang, K. T., Chen, Y. M., Lin, C. Y., Huang, C. C., Lin, M. C., & Wang, C. C. (2023). E-cadherin expression in the tumor microenvironment of advanced epidermal growth factor receptor-mutant lung adenocarcinoma and the association with prognosis. BMC Cancer, 23(1), Article 570.