Module_3:

Team Members:

Jack O'Hearn, Wyatt Young

Project Title: Angiogenesis in Lung Cancer

Module 3 Hallmarks of Cancer

Project Goal:

This project seeks to understand the mechanism and associated genes of angiogenesis in lung cancer and determine if there are specific genes or markers related to this hallmark.

Disease Background:

- Cancer hallmark focus: evading angiogenesis
- Overview of hallmark: Evading apoptosis is the ability for nearly all cancer cells to circumvent the cells natural process of programed cell death. Without cell death, malignant cells can continue to reproduce and spread their altered genes that allows cancer to continue to grow harmful tumors within the body. Evading this normal regulation is a hallmark of cancer growth and exponential reproduction. The main pathway of evading apoptosis is through inactivation of the p53 protein. This protein can upregulate proapoptotic expression through BAX after DNA damage. BAX in turn stimulates mitochondrial release of cytochrome C which promotes apoptosis. Without this p53 activated circuit, apoptosis can be down regulated and evaded.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate): TP53 gene which is what produces the p53 protein is the gene and protein most associated with evading angionesis. This protein works by upregulating the expression of the BCL2 gene which is responsible for producing BAX which promotes apoptosis after DNA damage.
- Prevalence: 654,000 cases of lung cancer are in the US today.
- Incidence: 226,000 is the estimated number of cases of lung cancer for 2025 in the US.
- Risk factors (genetic, lifestyle) & Societal determinants: Smoking cigarettes is a major risk factor for developing lung cancer, as well as living or working in areas with poor air

quality in general. Family history and mutations in the EGFR, KRAS, and ALK genes increase the risk for lung cancer. Limitied heathcare, environment, and poverty are all societal determinants for lung cancer as well.

- Standard of care treatments (& reimbursement): Surgery, chemotherapy, and radiation are the primary treatments for lung cancer. Other options include a lung transplant but this is only an option when the patient has been cancer free for a period of time.

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology): Lung cancer prevents the lungs from its normal process of gas exchange and starts in the bronchi or alveoli. The mechanism includes tumor formation that immune invasion, angiogenesis, and spreading of the cancer throughout the lungs. Often the tumor metastasize quickly and spreads to other organs and bone.

Data-Set:

Once you decide on the subset of data you want to use (i.e. only 1 cancer type or many; any clinical features needed?; which genes will you look at?) describe the dataset. There are a ton of clinical features, so you don't need to describe them all, only the ones pertinent to your question. (Describe the data set(s) you will analyze. Cite the source(s) of the data. Describe how the data was collected -- What techniques were used? What units are the data measured in? Etc.)

The data for our analysis come from The Cancer Genome Atlas (TCGA) RNA-sequencing dataset, which was re-processed by Rahman et al.. Expression levels are comparable across cancer types. The dataset includes 24 different cancers, with RNA-seq values expressed as $\log_2(\text{TPM} + 1)$ (Transcripts Per Million).

The dataset was subsetted to include the 3,000 most variable protein-coding genes out of 15,000 total genes, across 1,802 tumor samples (out of 9264 total). This is typically 50–100 per cancer type. The metadata file provides approximately 70 clinical and molecular variables for each sample, such as patient age, tumor stage, histological subtype, as well as survival outcomes. The metadata codebook explains what each variable means, as well as variable type, and the metadata percent non-NA file shows how complete each variable is across different cancer types.

Survival data was added from the Pan-Cancer Clinical Data Resource (Liu et al., 2018), which provides several outcome measures. Some examples are overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI), and disease-free interval (DFI).

The two cancer types we will be focused on in our analysis are LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma). The variables for both of these are of similar completeness, both having about 50% of variables being complete enough for analysis. This includes general clinical background like race, gender, ehtnicity. It also includes radiation treatment status, tumor status, and survival related variables.

These files specifically will be used for our analysis:

GSE62944_subsample_log2TPM.csv GSE62944_subsample_topVar_log2TPM.csv GSE62944_metadata.csv subsampled_TCGA_CDR_survival.csv

Data Analyis:

Methods

The machine learning technique I am using is: *fill in and describe*

What is this method optimizing? How does the model decide it is "good enough"?

**

Analysis

Uploaded as part of main.py/main.pdf

Verify and validate your analysis:

Pick a SPECIFIC method to determine how well your model is performing and describe how it works here.

Code for validation using training/test splituploaded in main.py. Analysis of success and literture validation:

- TP53 mutations are more prevalent in LUSC compared to LUAD (reported ranges: ~75-82% in LUSC vs. 47-60% in LUAD)
- Both VEGFA and ANGPT2 are higher in lung cancer compared to normal, but their protein levels are associated with tumor size and lymph node metastasis only in LUAD, not in LUSC.
- Increased VEGFA and ANGPT2 expression is associated with lower survival in LUAD patients, but this connection is not seen in LUSC.
- Studies examining direct comparisons of expression levels between LUAD and LUSC emphasize the stronger associations in LUAD for HIF1A.
- This literature validation reinforces our finding that it is practical to use certain genes (TP53, VEGFA, HIF1A, ANGPT2, FLT1) as a trustworthy validation to distinguish LUAD and LUSC.
- With an AUC and AP values of 0.93 and 0.95 respectively, our validation suggests success when using these 5 genes to differentiate lung cancers.
- In addition, our confidence matrix shows both labels to be correct over 80% of the time.

(Describe how you checked to see that your analysis gave you an answer that you believe (verify). Describe how your determined if your analysis gave you an answer that is supported by other evidence (e.g., a published paper).

Conclusions and Ethical Implications:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

NOTES FROM YOUR TEAM:

Possible questions:

- Due to the lungs already having significant vasculature, is angiogenesis "easier" for the tumor to achieve in lung cancer?
- How does mutation of the TP53 affect the survival rate and progression of lung cancer?

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

We would like to know if the correct validation methods were utilized specifically for an analysis like using gene expression to differentiate between two types of cancer.

In [ ]: