Establishing the Cross-Cancer Mutational Landscape through Differential Gene Expression Analysis

Hypothesis

The objective of our analysis is to characterize patterns of differential gene expression across four common types of cancer, to gain insight into the transcriptomic profiles across our diseases of interest. We hypothesize that variation in levels of gene expression across distinct types of cancer contributes to molecular differences in disease onset, progression, and metastasis. Thus, identifying functionally relevant genes across enriched pathways may reveal the underlying biological relationships that govern shared mutational drivers across multiple forms of cancer.

Motivation & Background

Although mutations serve as the genetic basis of cancer development, malignant phenotypes occur beyond the initial modification, primarily resulting from variations in gene expression across cancerous cell and tissue types (Sager, 1997). To reveal the fundamental biological mechanisms associated with tumor initiation and metastasis, previous studies sought to investigate transcriptomic profiles between healthy and diseased states (Rodriguez-Esteban, 2017). However, the majority of existing studies failed to establish concrete signatures of differential gene expression contributing to phenotypic heterogeneity across distinct types of cancer. Thus, we seek to identify similar patterns of differential expression across cancer-associated genes and the corresponding impact on biological pathways, to reveal common genetic risk factors across multiple disease contexts and provide broader insight into the mechanistic underpinnings of cancer development and proliferation (Jiang et al., 2019).

We seek to investigate and compare the gene expression profiles across four major cancer types: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD) (Kandoth et al., 2013). Furthermore, we intend to interrogate differentially expressed genes (DEGs) to characterize coordinately regulated gene modules and reveal enriched biological pathways governed by our genes of interest. Elucidating functional similarities in expression motifs may suggest unifying molecular themes associated with cancer-specific initiation and progression (Martínez et al., 2015), and outlining differences in expression patterns may confirm the influence of tissue-specific mechanisms and context-dependent functions of DEGs on cancer development and proliferation (Frost, 2021). Overall, constructing a representative transcriptomic profile from observed expression patterns of cancer-associated genes establishes a preliminary compendium of candidate therapeutic biomarkers, allowing for increasingly accurate predictions of clinical outcomes and potential advancements toward patient-specific treatments (van't Veer et al., 2002).

Significance

Our cross-cancer analysis of differential expression patterns will reveal DEGs associated with initiation, development, and proliferation, the results of which will establish a preliminary transcriptomic profile and facilitate downstream analyses across a broader range of disease contexts. In addition, identifying DEGs involved in cancer proliferation provides a foundation for functional enrichment analysis, such as Gene Ontology (GO), allowing us to elucidate the common biological pathways governed by our DEGs of interest. Characterizing the cross-cancer mutational landscape through differential expression analysis pinpoints the genetic associations responsible for malignant phenotypes, and may facilitate advancements toward increasingly precise treatment approaches for a greater range of cancer types.

Feasibility

Methodology

Data Description & Pre-Processing

The data for this analysis are sourced from OncoDB, a comprehensive database containing publicly-available gene expression profiles for 33 major types of cancer, aggregated from over 10,000 samples in The Cancer Genome Atlas (TCGA) Program and the Genotype-Tissue Expression (GTEx) Project (Tang et al., 2021). To obtain the initial gene expression metrics for this analysis, a .zip file containing separate .txt files for each cancer type was downloaded from OncoDB 'Curated Expression Data.' Each .txt file contains a list of genes exhibiting the highest levels of differential expression between normal and cancer samples. Based on the overall availability of the data and the number of DEGs in the initial files, we intend to cross-compare DEGs across BRCA, COAD, KIRC, and LUAD. The variables included in each file are listed below, along with their respective descriptions.

Name	**Description**
Cancer Type	TCGA Study Abbreviation
NCBI Gene ID	NCBI Gene ID
FDR Adjusted P-Value	Adjusted student's t-test p-value
Cancer Sample Med	Median expression levels across cancer samples
Normal Sample Med	Median expression levels across normal samples
Log2-Fold Change	Log ratio of expression values across two conditions (normal vs. disease)
P-Value	Student's t-test p-value
Gene Symbol	Gene Symbol

Differential Gene Expression Analysis

To facilitate the identification and cross-comparison of DEGs for our cancer types of interest, we seek to develop the following visualizations: volcano plots, heatmaps, and Manhattan plots. We intend to generate annotated volcano plots for each cancer type using *ggplotly* in R to assess the relationship between log-fold change and inverse FDR adjusted p-value. In a volcano plot, DEGs with lower adjusted p-values are positioned higher on the vertical axis, positive log-fold change denotes genes upregulated in diseased samples, and negative log-fold change denotes genes downregulated in diseased samples. Thus, genes located higher on the vertical axis of each volcano plot are most significant in disease, and may serve as potential therapeutic targets for the respective cancer type.

Functional Enrichment Analysis

Functional enrichment analysis provides insight into the underlying biological mechanisms associated with DEGs, as genes with shared expression patterns are more likely to impact related biological pathways. Furthermore, clustering genes into coordinately regulated modules allows for increasingly accurate identification of gene function and pinpoints key targets for future mechanistic studies. Prior to executing signature analysis, we intend to aggregate DEGs exhibiting similar expression patterns across all cancer types, and convert the list of DEGs to a dataframe. To reveal coordinately regulated gene modules, we seek to leverage *hclust* in R to perform agglomerative hierarchical clustering with the Pearson correlation coefficient. We intend to develop heatmap visualizations using *gplots* in R, allowing us to visually assess genes modules and variations in expression patterns across significant DEGs for all cancer types of interest.

We seek to leverage GO to perform signature analysis, to identify common biological pathways associated with cancer development. GO requires an arbitrary selection of DEGs, thus, we intend to establish a cutoff value to extract a subset of genes from our original sample. We will utilize *gprofiler2* in R to compare the subset sample, corrected for false discovery rate (FDR), to all annotated genes in humans. As GO is redundant, we will manually summarize enrichment results

for all cancer types, allowing us to identify and compare the primary biological pathways shared across the top DEGs. To facilitate the interpretation of enriched GO terms, we will utilize *gostplot* in R to develop Manhattan plots, allowing us to assess adjusted p-values for top DEGs across various GO annotation categories and identify shared biological pathways associated with cancer development.

Leverage

All group members are Bioinformatics majors, with extensive computational experience in Python, R, and C++. All group members have taken a diverse range of coursework in computer science, data science, biology, and statistics. All group members have research experience in academia, and several group members have contributed to bioinformatics projects in the life sciences industry. All group members understand scientific literature and have experience with technical writing.

Expected Results

We expect significant DEGs present across all types of cancer to exhibit a similar degree and pattern of expression. Moreover, we expect shared DEGs to disrupt biological pathways of related function across various forms of cancer. Common DEGs exhibiting variations in expression levels across distinct cancer types may suggest the influence of context-specific mechanisms on disease initiation and proliferation.

Timeline

Week	**Tasking**
[1] 10/31 - 11/04	**DUE: Proposal Peer Review (10/31 at 11:59 PM)** Complete Final Draft of Proposal, Data Collection & Pre-Processing
[2] 11/07 - 11/11	**DUE: Final Proposal (11/07 at 11:59 PM)** Analysis (Identify DEGs, Functional Enrichment), Visualization
[3] 11/14 - 11/18	Analysis (Functional Enrichment), Visualization, Begin Written Report
[4] 11/21 - 11/25	Visualization, PowerPoint Presentation, Revise Written Report
[5] 11/28 - 12/02	Rehearse Presentation, Revise Written Report
[6] 12/05 - 12/09	**DUE: In-Class Presentation (12/05, 12/07)** In-Class Presentation, Complete Final Draft of Written Report
[7] 12/12 - 12/16	**DUE: Report, Peer Evaluation, Self-Reflection (12/13 at 11 AM)** Peer Evaluation, Self-Reflection

References

[1] Frost, H. R. (2021). Analyzing cancer gene expression data through the lens of normal tissue-specificity. PLoS Computational Biology, 17(6), e1009085. https://doi.org/10.1371/journal.pcbi.1009085

[2] Jiang, X., Finucane, H. K., Schumacher, F. R., Schmit, S. L., Tyrer, J. P., Han, Y., Michailidou, K., Lesseur, C., Kuchenbaecker, K. B., Dennis, J., Conti, D. V., Casey, G., Gaudet, M. M., Huyghe, J. R., Albanes, D., Aldrich, M. C., Andrew, A. S., Andrulis, I. L., Anton-Culver, H., ... Lindström, S. (2019). Shared heritability and functional enrichment across six solid cancers. Nature Communications, 10(1), 431. https://doi.org/10.1038/s41467-018-08054-4

[3] Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. Nature, 502(7471). https://doi.org/10.1038/nature12634

[4] Martínez, E., Yoshihara, K., Kim, H., Mills, G. M., Treviño, V., & Verhaak, R. G. (2015). Comparison of gene expression patterns across twelve tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects. Oncogene, 34(21), 2732–2740. https://doi.org/10.1038/onc.2014.216

[5] Rodriguez-Esteban, R., & Jiang, X. (2017). Differential gene expression in disease: A comparison between high-throughput studies and the literature. BMC Medical Genomics. https://pubmed.ncbi.nlm.nih.gov/29020950/

[6] Sager, R. (1997). Expression genetics in cancer: Shifting the focus from DNA to RNA. Proceedings of the National Academy of Sciences of the United States of America, 94(3), 952–955. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC19620/

[7] Tang, G., Cho, M., & Wang, X. (2021). OncoDB: An interactive online database for analysis of gene expression and viral infection in cancer. Nucleic Acids Research, 50(D1). https://doi.org/10.1093/nar/gkab970

[8] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871), Article 6871. https://doi.org/10.1038/415530a