A Graph-Based Analysis of Gene Activity in Lung Cancer based on Protein-Interaction Networks

Simone Bergmann
simone.bergmann@uni-bielefeld.de

December 18, 2024

Abstract

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat,sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse

1 Introduction

Lung cancer is one of the leading causes of mortality worldwide, with significant implications for public health and individual well-being. This thesis aims to contribute to the ongoing efforts to combat this disease through identifying potential biomarkers using graph databases and algorithms.

1.1 Motivation

Lung cancer poses a significant global health threat, with multiple factors contributing to its development, including smoking, air pollution, and genetic factors [15]. It is the leading cause of cancer-related deaths worldwide [15]. Despite advances in medical care, the prognosis for patients remains relatively low, with only 30.2% of women and 22.1% of men (2016) surviving beyond five years after diagnosis [57], often due to late-stage detection. Therefore, early detection is crucial to initiate timely treatment and reduce mortality rates.

To address this challenge, we leverage gene expression data to identify potential biomarkers for lung cancer, enabling us to better understand the underlying molecular mechanisms driving this disease. Graph databases provide a powerful toolset, as they offer ways to represent complex relationships between large datasets and analyze them efficiently using graph-based algorithms. By analyzing these data, we want to gain valuable insights that can be translated into improved early diagnosis and treatment strategies.

1.2 Goals

The main goal of this thesis is to identify potential biomarkers for lung cancer by analyzing gene expression data. These Biomarkers could be used for early detection or personalized treatment. This will be done by uncovering genes with altered cancer activity that are closely connected to other genes. To achieve this goal, we have three strategic objectives:

1. **Objective - Find significant changes in gene activity:**
   We must first decide on an appropriate measurement for cancer-related changes in gene activity. Next, we will compare the expression levels of healthy and cancerous tissues to identify genes with substantial differences. These genes will serve as our focus for further analysis.

2. **Objective - Application of a graph algorithm:**
   We will construct a graph database containing genetic information with connections between genes and proteins. By applying a graph algorithm, we aim to identify the most central genes in the network, which may have the highest impact on other genes.

3. **Objective - Validation of results through study comparison:**
   We will assess the relevance and novelty of our findings by comparing them with existing research on lung cancer. By validating our results against established knowledge, we can ensure that our identified genes are indeed potential biomarkers.

These objectives provide a clear roadmap for achieving our main goal. By determining significant changes in gene activity, applying graph algorithms, and validating results through comparisons with existing research, we aim to contribute meaningfully to lung cancer research.

1.3 Structure

The thesis starts with a closer look at the biological and computational background of the project (section 2), which includes a comprehensive overview of lung cancer, gene expression analysis, and graph databases in cancer research.
Next, we describe the experimental setup and methodology employed to collect and analyze data (section 3), including the creation of a graph database and the application of a graph algorithm.
The result of our investigation is presented in chapter 4.
We then delve into the implications and significance of these findings, providing a comprehensive analysis of their relevance to lung cancer research (section 5).
Finally, we conclude by summarizing the key takeaways from our study and provide an outlook on future research directions (section 6).

1.4 Limitations

Our study focused on lung cancer due to its high incidence rate and the availability of relevant datasets. However, this focus may not be representative of other types of cancers or diseases.
We only considered protein-coding genes in our analysis due to technical constraints related to our graph database setup.
Rather than collecting new data, we utilized existing datasets to expedite the analysis and focus on method development. But our experimental setup may also be used for other datasets ins future.

In conclusion, this thesis aims to find novel biomarkers on lung cancer by investigating the potential of graph databases and algorithms in the context of lung cancer research. We will focus on three objectives to achieve this goal: analyse gene expression data, build a graph database and apply a graph algorithm, and validate the results. The following chapters will provide a detailed explanation of the methods used and the results obtained.

2 Background

In this chapter we will take a closer look at the biological and computational background of our project. Also, we will discuss related work in the field cancer research with graph databases.

2.1 Biological background

To have a better understanding of our project, we need to have a closer look at some biological topics and concepts.
Cancer is a complex, multifactorial disease characterized by uncontrolled cell growth. Research shows that widespread metastases are the primary cause of death from cancer [41] and that cancer comes in more than 100 different types [24].
The five most common forms of cancer worldwide in 2022 were lung cancer ($>$ 2.4 million), breast cancer ($>$ 2.2 million), colorectal cancer ($>$ 1.9 million), prostate cancer ($>$ 1.4 million) and stomach cancer ($>$ 0.9 million). In 2022, cancer was one of the leading causes of death globally, with approximately 10 million fatalities [15].
It can be caused by a combination of genetic and environmental factors, such as diet, radiation, age, exposure to certain chemicals or viruses [22]. Early detection and the right treatment can significantly improve the chances of curing many types of cancer.

In this study we will focus on **Lung cancer**, a type of cancer that affects the lung organ. There are two subtypes: non-small cell carcinoma (NSCLC) and small cell carcinoma (SCLC) [21]. The primary risk factor for developing lung cancer is smoking, but passive smoking and environmental pollution also significantly increase the risk [21].
Symptoms of lung cancer often resemble those of common colds or other minor illnesses, such as coughing and fatigue. This makes it challenging for patients to receive timely treatment, often leading to a diagnosis at an advanced stage [42].

Gene expression is a crucial factor in the investigation of cancer. It describes the process of transcribing DNA into RNA molecules that code for proteins, which are then translated and regulated through complex interactions to control the production and function of gene products [25]. Alterations in gene expression can contribute to the uncontrolled multiplication and abnormal behavior of cancer cells, which ultimately leads to the development and progression of cancer [16].

To accurately compare gene expression data across different experiments, it is essential to normalize and quantify the expression levels. The most commonly used metrics for this purpose include Reads per million mapped reads (RPM), Reads Per Kilobase Million (RPKM), Fragments Per Kilobase Million (FPKM), and Transcripts Per Kilobase Million (TPM) [18]. Among these measures, **TPM** is widely adopted in available datasets. Therefore, we will utilize it as our metric of choice for quantifying gene expression levels.

A common way to identify genes is through the use of unique identifiers known as **ENSEMBLE IDs**. These are unique ids for genes, proteins and other genetic elements collected in the Ensemble database from 1999 by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute [5]. These make it easier to compare and analyze gene data across different datasets.

Omics is a collection of fields in biology that end in -omics, such as genomics, proteomics, metabolomics, etc. [55]. **Multi-omics data**, the combination of data from different omics fields, are used to gain a better understanding of associations between biological molecules (e.g., genes and proteins) and their interactions [56]. Since we will be using genes and proteins in our study, we also have a multi-omics approach.

2.2 Computational background

Graph databases are an effective way to model and analyze graph-like data, which is in contrast to traditional relational databases well-suited for applications involving relationships between entities. For instance, social networks and biological systems can be effectively represented using these graph structures [27]. In particular, graph databases have been widely used in biological data analysis, such as modeling Protein-Protein-Interaction (PPI) networks that illustrate the interaction between proteins to control cellular processes. In this thesis, we use a graph database to put our collected cancer data into a meaningful context.
Just like a graph, a graph database consists of two basic components: nodes and edges between nodes. These components form the foundation for storing and analyzing large amounts of data.
Nodes are the units that store the basic information of a certain type of entity [27]. In our case, genes and proteins are represented as two different node types in the database. Each gene is, therefore, a node with properties, such as an ID, a *name* or a $TPMvalue$ for its activity.
Relationships between two nodes are represented by an edge, which also can be of different types. In our graph database, for example, there are "interactions" as edges between proteins and "connections" as edges between a gene and a protein. As in common graph theory, edges can represent structures like one-to-many or many-to-many relationships. Edges can also have weights, directions, or properties like *name* or *type*[44]. We do not use these additional metadata in our setup, but they could be used to model more complex relationships.
We will implement our graph database using Neo4J, since it is a widely used graph database management system.

Graph database algorithms enable the efficient analysis and extraction of meaningful insights from large-scale network data. They can be broadly categorized into three primary groups, which provide different perspectives on graph structure.
Traversal and Pathfinding Algorithms enable the identification of shortest or most optimal paths between nodes in the graph, thereby facilitating the exploration of network topology. Examples include Depth-First-Search, Breadth-First-Search, Shortest Path, and Max-Flow-Min-Cut[6].
Centrality Algorithms are instrumental in understanding which nodes within a graph hold significant importance. By evaluating centrality measures such as degree centrality, closeness centrality, betweenness centrality, and PageRank, valuable insights into the underlying network structure is gained[6].
Community Detection Algorithms, also known as clustering or partitioning algorithms, are essential for identifying groups of nodes within a graph that share similar characteristics or exhibit extensive connectivity. Examples include Label Propagation, Louvain Modularity, and Strongly Connected Components [6].

In our analysis, we utilize the **PageRank algorithm** to identify influential nodes within the network based on their connectivity and centrality. The PageRank algorithm was originally a method for evaluating and prioritizing websites on the internet. It was developed by Larry Page and Sergey Brin in 1988 at Stanford University [43], the founders of Google. The general idea behind the algorithm was that the more links pointing to a website, the more important it is. This concept is extended in the PageRank algorithm, which assigns a node's score by iteratively considering not only its direct connections but also the indirect relationships through its linked neighbors.
The algorithm calculates a PageRank score per node using the following formula:

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)})$$

where $PR(A)$ is the PageRank score of node A, T_1 to T_n are nodes with edges to A, d is the damping factor, and $C(T_1)$ represents the number of edges to node T_1 [6].

2.3 Related work

The investigation of tumors and the identification of biomarkers for diagnosis and treatment are crucial tasks in oncology [11]. Over the past few years, various approaches have been explored to address these challenges. Notably, graph-based methods have gained popularity for analyzing genetic data. This section presents a brief insight of three studies in this field from the last years, focusing on the use of graph databases and their algorithms for biological applications and cancer research [47, 51, 49]. By reviewing existing literature, we aim to provide a comprehensive understanding of the current state of the art in this area.

The recent systematic review by Rout et al. (2024) [47] provides an extensive overview of various graph-based methodologies for analyzing PPI networks, emphasizing best practices and common challenges in integrating multi-omics data. The authors highlight the importance of graph databases in storing and querying large-scale biological networks and essential graph algorithms such as centrality measures and community detection that are relevant to my work using PageRank.

A study by Simpson et al. (2020) [51] investigated the molecular mechanisms underlying various cancer types through the analysis of gene expression data sourced from the TCGA database. The authors employed co-expression networks, using Pearson correlation to establish relationships between genes based on their expression patterns. However, our thesis will take a distinct approach by concentrating specifically on lung cancer and utilizing PPI networks as base instead of co-expression networks. In contrast to Simpson et al.'s comprehensive application of multiple graph algorithms, including PageRank, Louvain community detection and Dijkstra's algorithm in each cancer type, we will focus on a more targeted approach. Specifically, we will apply the Pagerank algorithm to identify key genes within our network.

Shang and Liu (2020) [49] proposed a method for prioritizing cancer genes called iRank, which integrates various biological levels, including gene and protein expression, as well as PPI, to identify hepatocellular carcinoma. In contrast, we will focus on building a PPI network to analyze the interactions between proteins and also add their corresponding genes. Unlike Shang and Liu's approach, which concentrates on the TCGA dataset, our work will utilize the Cell Model Passport dataset. Similar to their approach, we will employ the PageRank algorithm to identify important genes in our analysis.

In summary, the studies mentioned above provide valuable insights into the application of graph-based methods for analyzing biological data. Each with a slightly different focus, they demonstrate the versatility of graph databases and algorithms in cancer research.

Overall, in the course of this background section, we have gained a comprehensive overview of the biological background of lung cancer and the importance of gene expression analysis. In particular, we have looked at TPM values, which are an important measure of gene activity.
We have established that the use of graph databases is an effective way to analyze complex networks, and we have explained corresponding algorithms, especially the PageRank algorithm.
Finally, we have reviewed related work in the field of graph databases in cancer research.

3 Methodology

Here, we will delve into the details of the experimental setup. We will discuss the data sources and their transformation, the creation of the graph database, and the application of algorithms on the graph database.

3.1 Data

Our study consists of two main datasets. One includes healthy tissue data from the Genotype-Tissue Expression (GTEx) project, and the other includes cancerous tissue data from the Cell Model Passport

(CMP) project. While these resources are widely used and well-established in research, they do have limitations in terms of their scope and coverage. The GTEx dataset consists of only adult postmortem donor samples, with 2/3 of them between 50 and 69 years old and 2/3 being male [3]. The Cell Model Passport dataset contains data from donors who are predominantly male (60%) and have an age range that is biased towards older adults (50-69 years old) [1].
The two datasets provide the foundation for our graph database, specifically for the gene nodes. We aim to create a table with genes as rows, where each row contains a unique Ensemble ID along with one value representing healthy TPM and another value representing cancerous TPM.

For the healthy tissue samples used in this study, we utilized the "GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm" dataset from the GTEx portal [8]. The **GTEx portal** is a large-scale, publicly available resource for studying gene activity. The Adult GTEx project aims to characterize the gene expression patterns in healthy tissues across different individuals, providing valuable insights into the underlying biology of human development and disease.
The used dataset A_{orig} is in a .gct file format that contains TPM values for 56,156 genes i identified by Ensemble ID as rows in 17,382 different tissues j as columns (see Table 1.0). The data was initially stored in a file in wide format where each tissue j had its own column. The TPM values for these tissues range between 0 and 747,400. Since there are no missing values in the dataset, we did not need to handle any missing data. To process this data into a suitable format, we employed the following steps:

1. **Reshaping to long format:** We read the data from the original format in chunks of 3,000 rows at a time due to RAM capacity constraints. For each chunk of data, we transformed the columns for each tissue j into individual rows, resulting in a dataset with three columns and 976.103.592 $(i * j)$ rows (see Table 1.1).

2. **Grouping by genes:** Once all chunks had been processed, we separated the combined dataset again for RAM reasons in new chunks of approximately 200 million rows. These chunks have been grouped by gene using an aggregate function that calculated both the sum S_i and count C_i of TPM values for each gene i (see Table 1.2).

3. **Calculating mean TPM:** To handle genes that had been split across multiple chunks, we performed a global aggregation on the genes of the sum of the dataset. Then we calculated the mean TPM value M_i for each gene i by dividing the sum S_i by the count C_i of each observation (see Table 1.3).

0. Original Format	**1. Reshaping to long format**	**2. Grouping by genes**	**3. Calculating mean TPM**
i : number of genes j : number of tissues a_{ij} : TPM for gene i in tissue j	$A_{\text{long}} = (\text{Gen}_i, \text{Tissue}_j, a_{ij})$	$A_{\text{agg}} = (\text{Gen}_i, S_i, C_i)$	$A_{\text{mean}} = (\text{Gen}_i, M_i)$
		$S_i = \sum_{j=1}^{j} a_{ij}, \quad C_i = \sum_{j=1}^{j} 1$	$M_i = \frac{S_i}{C_i}$
$A_{\text{orig}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} \end{bmatrix}$	$A_{\text{long}} = \begin{bmatrix} \text{Gen}_1 & \text{Tissue}_1 & a_{11} \\ \text{Gen}_1 & \text{Tissue}_2 & a_{12} \\ \vdots & \vdots & \vdots \\ \text{Gen}_i & \text{Tissue}_j & a_{ij} \end{bmatrix}$	$A_{\text{agg}} = \begin{bmatrix} \text{Gen}_1 & S_1 & C_1 \\ \text{Gen}_2 & S_2 & C_2 \\ \vdots & \vdots & \vdots \\ \text{Gen}_i & S_i & C_i \end{bmatrix}$	$A_{\text{mean}} = \begin{bmatrix} \text{Gen}_1 & M_1 \\ \text{Gen}_2 & M_2 \\ \vdots & \vdots \\ \text{Gen}_i & M_i \end{bmatrix}$
$A_{\text{orig},i} = \begin{bmatrix} 8.764 & 0.07187 & \dots & 3.215 \end{bmatrix}$	$A_{\text{long},i} = \begin{bmatrix} \text{ENSG...938} & \text{GTEX-111...} & 8.764 \\ \text{ENSG...938} & \text{GTEX-112...} & 0.07187 \\ \text{ENSG...938} & \text{GTEX-113...} & 3.215 \end{bmatrix}$	$A_{\text{agg},i} = \begin{bmatrix} \text{ENSG...938} & 12.051 & 3 \end{bmatrix}$	$A_{\text{mean},i} = \begin{bmatrix} \text{ENSG...938} & 4.017 \end{bmatrix}$

Table 1: Data transformation pipeline for the GTEx dataset: Formular and example data per gene

The resulting dataset (see Figure 1) with 56,156 genes i and a mean TPM value M_i was saved as a CSV file for further processing.

	Gene ID	healthy TPM
0	ENSG00000000003	15.765183
1	ENSG00000000005	3.568990
2	ENSG00000000419	48.419258
3	ENSG00000000457	5.825362
4	ENSG00000000460	2.375547

Figure 1: Example data of processed Genotype-Tissue Expression dataset

For the purpose of the analysis of gene activity in lung cancer, we utilized data from the **CMP project**, a comprehensive resource for studying cancer-related gene expression.
We obtained the dataset from the CMP portal [2] with the name `rnaseq_all_data_20220624`. This dataset contains data from the Sanger Institute and the Broad Institute and consists of a file containing genes associated with diverse cancer types, including lung cancer. Initially, the data was stored in long format with columns for gene identifiers, tissues, TPM values, and additional information.
However, the CMP dataset lacked Ensemble IDs, which were crucial for our analysis. So we needed to add the Ensemble IDs to the dataset to ensure consistency and accuracy in our analysis. To focus on lung cancer-specific data, we loaded an additional file containing tissue-specific annotations. [7] Then we filtered the CMP dataset to include only tissues from the annotation file with lung cancer as the cancer type. Specifically, we filtered the dataset to include only lung cancer-specific models labeled as Small Cell Lung Carcinoma, Non-Small Cell Lung Carcinoma or Squamous Cell Lung Carcinoma in the *cancer_type* column.
The resulting dataset comprises 7,564,389 rows containing genes and tissues with associated TPM values for lung cancer (see Table 2.0). Specifically, the dataset includes information on 37,262 unique genes i across 203 distinct tissue types j. The resulting dataset contains no missing values and the TPM values span a range of 0 and 132,676.
To prepare the data for further processing, we performed the following steps:

1. **Grouping by genes:** We grouped the dataset by genes to obtain a mean TPM value for every gene. This step involved aggregating the data by gene names, resulting in a new dataset with a mean TPM value for each gene (see Table 2.1).

2. **Adding Ensembl ID:** The original dataset contained own IDs for the genes but lacked the universal Ensembl ID required for matching genes across datasets. We overcame this challenge by adding the required Ensembl IDs to our dataset using gene names as a reference point. For this purpose we downloaded an Ensemble file from biomart [4], which contains the Ensembl ID and corresponding gene name.

   By analyzing the Ensemble file, we encountered an issue where some gene name were not unique within the file. To resolve this problem, we dropped all rows with duplicate gene names. We then merged the Ensemble table with our CMP data on the gene name to retrieve the Ensemble ID for each gene (see Table 2.2).

3. **Removing missing Ensembl ID:** After merging the data, we found that 3,760 of our 37,262 genes still had no Ensembl ID associated with them. Since these genes were likely duplicates or did not exist in the Ensemble file, we removed them from our dataset to ensure consistency and accuracy of our analysis (see Table 2.3).

The resulting dataset (Figure 2) contains 33,502 genes with mean TPM values for lung cancer and was saved as a CSV file for further processing.

3.2 Nodes and Edges

As we progress in our analysis, we now turn our attention to constructing the fundamental building blocks of our network.

0. Original Format	1. Grouping by genes	2. Adding Ensembl ID	3. Removing missing Ensembl ID
i : number of genes j : number of tissues b_{ij} : TPM for gene i in tissue j $B_{\text{orig}} = (\text{ID}_i, \text{Name}_i, \text{Tissue}_j, b_{ij}, \dots)$ $B_{\text{orig}} = \begin{bmatrix} \text{ID}_1 & \text{Name}_1 & \text{Tissue}_1 & b_{11} \\ \text{ID}_1 & \text{Name}_1 & \text{Tissue}_2 & b_{12} \\ \vdots & \vdots & \vdots & \vdots \\ \text{ID}_i & \text{Name}_i & \text{Tissue}_j & b_{ij} \end{bmatrix}$	$M_i = \frac{1}{j}\sum_{j=1}^{j} b_{ij}$ $B_{\text{agg}} = (\text{ID}_i, \text{Name}_i, M_i)$ $B_{\text{agg}} = \begin{bmatrix} \text{ID}_1 & \text{Name}_1 & M_1 \\ \text{ID}_2 & \text{Name}_2 & M_2 \\ \vdots & \vdots & \vdots \\ \text{ID}_i & \text{Name}_i & M_i \end{bmatrix}$	$B_{\text{ens}} = (\text{ID}_i, \text{Name}_i, \text{ENS ID}_i, M_i)$ $B_{\text{ens}} = \begin{bmatrix} \text{ID}_1 & \text{Name}_1 & \text{ENS ID}_1 & M_1 \\ \text{ID}_2 & \text{Name}_2 & \text{ENS ID}_2 & M_2 \\ \vdots & \vdots & \vdots & \vdots \\ \text{ID}_i & \text{Name}_i & \text{ENS ID}_i & M_i \end{bmatrix}$	
$B_{\text{org},i} = \begin{bmatrix} \text{SIDG...16} & \text{CASP10} & \text{SIDM...13} & 14.41 \\ \text{SIDG...16} & \text{CASP10} & \text{SIDM...61} & 0.64 \\ \text{SIDG...16} & \text{CASP10} & \text{SIDM...50} & 3.26 \end{bmatrix}$	$B_{\text{agg},i} = \begin{bmatrix} \text{SIDG...16} & \text{CASP10} & 5.017 \end{bmatrix}$	$B_{\text{ens},i} = \begin{bmatrix} \text{SIDG...16} & \text{CASP10} & \text{ENSG...400} & 5.017 \end{bmatrix}$	

Table 2: Data transformation pipeline for the CMP dataset: Formular and example data per gene

	Gene ID	Gene Name	cancerous TPM
0	ENSG00000121410	A1BG	0.827192
1	ENSG00000268895	A1BG-AS1	4.676305
2	ENSG00000148584	A1CF	1.355369
3	ENSG00000175899	A2M	1.669212
4	ENSG00000245105	A2M-AS1	1.033596

Figure 2: Example data of processed Cell Modell Passport dataset

First we create a basic PPI network with proteins as nodes and their interactions as edges. Then we extend the network with genes as nodes that are translated into proteins. The edges between them are called connections as shown in Figure 3.
In summary the graph database will consist of four components: genes, proteins, gene-protein edges (connections), and protein-protein edges (interactions). For each of these components, we create a table with their necessary attributes.

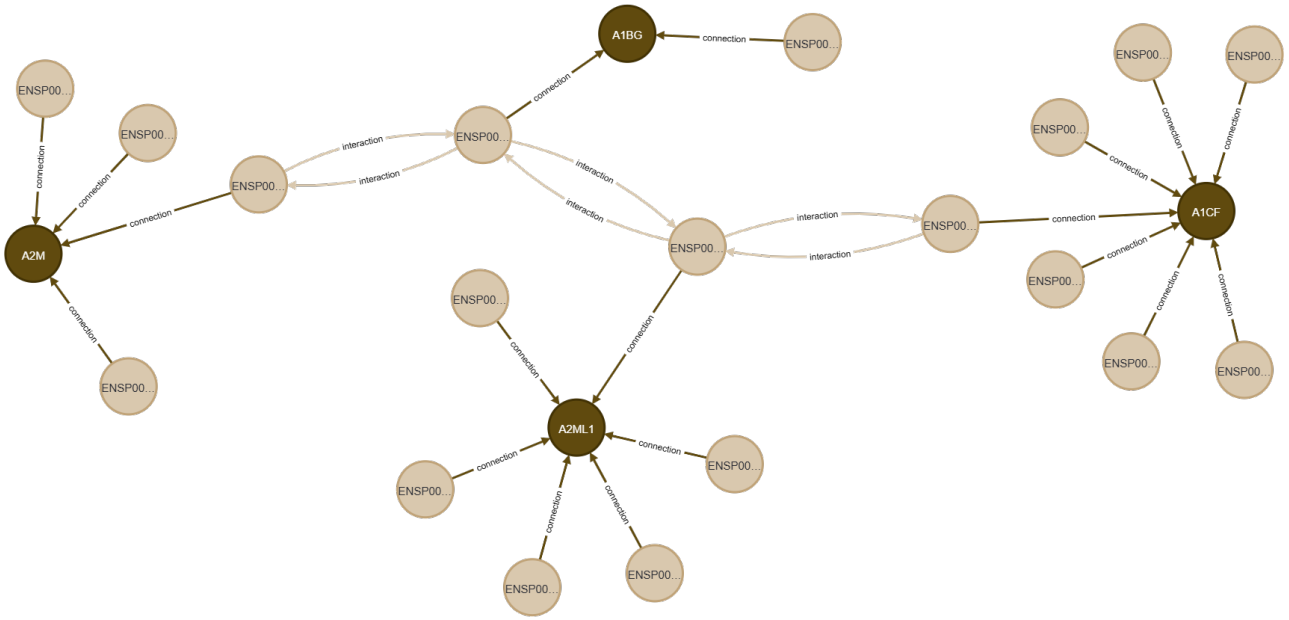

Figure 3: Extract from the graph database setup with genes in dark brown and proteins in light brown including their edges

Protein-protein edges

To create the protein-protein edges we use the data from the String (Search Tool for the Retrieval of Interacting Genes/Proteins) Database. This database is a comprehensive resource for studying protein-protein interactions and often used in research [58]. The protein links file for Homo sapiens, which we download from the database website [9], contains 13,715,404 rows of protein-protein edges (Figure 4). Each row includes two protein IDs, representing the edge between them.
Since no additional attributes are required for this table, we do not need to perform any further editing.

	left Protein ID	right Protein ID
0	ENSP00000000233	ENSP00000356607
1	ENSP00000000233	ENSP00000427567
2	ENSP00000000233	ENSP00000253413
3	ENSP00000000233	ENSP00000493357
4	ENSP00000000233	ENSP00000324127

Figure 4: Example data of protein-protein edges that will be used as interaction between proteins in the graph database

Gene-protein edges

To build the connections between genes and proteins, we need to link each gene to its corresponding protein translation. For this purpose, we downloaded a file from biomart containing Gene IDs and their Protein IDs for the human genome [4] The initial dataset comprised of 185,330 entries.
First we drop any rows where either the Ensembl ID for gene or protein is missing, as these would represent incomplete edges. Next we built the intersection of rows where the gene ID matched an existing gene node (as described in 3.2 - Gene nodes)
The final gene-protein edge table (Figure 5) features 101,731 rows as edges and two columns: Ensembl ID for the gene and Ensembl ID for the protein.

	Gene ID	Protein ID
0	ENSG00000198888	ENSP00000354687
1	ENSG00000198763	ENSP00000355046
2	ENSG00000198804	ENSP00000354499
3	ENSG00000198712	ENSP00000354876
4	ENSG00000228253	ENSP00000355265

Figure 5: Example data of gene-protein edges that will be used as connection between genes and proteins in the graph database

Gene nodes

To create the rows for our table of gene nodes, we use the preprocessed CMP and GTEx datasets, which contain mean TPM values for cancerous and healthy genes. We build the intersection of both datasets on their gene Ensemble ID to get a subset. The new dataset only contains genes with TPM values for both conditions.
Next we want to filter for genes that are translated into proteins. Because other types of genes would have no connection to the protein nodes in the network. Therefor we build a new subset of genes that also occur in the gene-protein edge table (3.2 - Gene-protein edges).
Now we have all rows we need for our gene nodes, and we want to focus on the columns that we will need as attributes. When examining the mean TPM values per dataset, we observe a right-skewed distribution, with most values close to zero and a long tail extending towards higher values. The cancerous TPM values vary from 0 to approximately 41,173 and the healthy TPM values range from 0 to around 36,200.

To fulfill our first objective 1 for these genes, we want to calculate a measure that captures significant changes between cancerous and healthy gene activity. First we normalize the TPM values from both datasets by performing a common log scaling between 0 and 1 for all TPM values.

$$log\_norm(x) = \frac{\log(1+x) - \log(1+x_{min})}{\log(1+x_{max}) - \log(1+x_{min})} \tag{1}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum TPM values across both datasets. After applying the normalization, the distribution of the TPM values is more balanced, as shown in Figure 6.

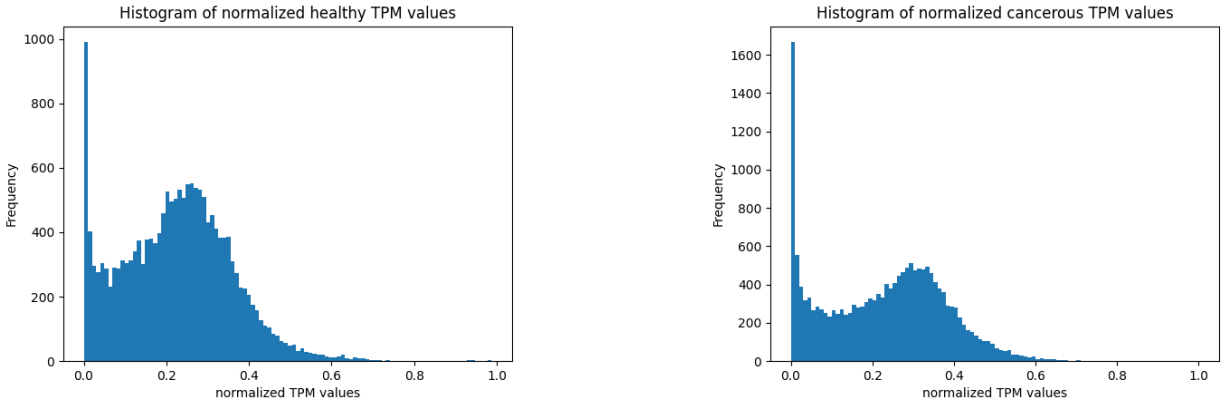

Figure 6: Normalized mean healthy and cancerous TPM values between 0 and 1

Next, we calculate the difference between the normalized mean healthy and cancerous TPM values per gene by subtracting the two values and call it Δ_{TPM}. The distribution of Δ_{TPM} values is shown in Figure 7.

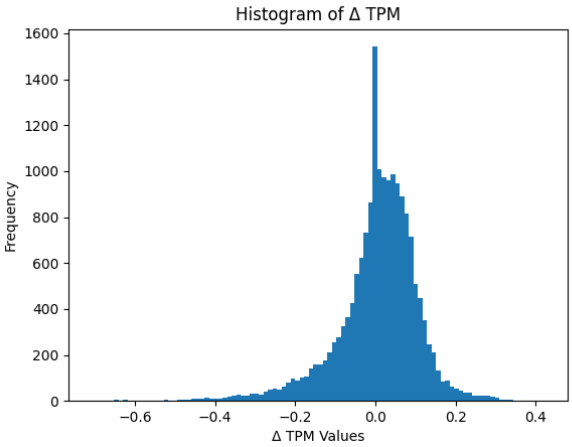

Figure 7: Distribution of the difference between the normalized mean healthy and cancerous TPM values, called Δ_{TPM} values

We then define an Δ_{type} as either *increase* or *decrease*, depending on whether the delta value is positive or negative.
As the final step for the objective, we need to determine if a change in gene activity is relevant. To do this, we use the z score measure, which calculates how many standard deviations a Δ_{TPM} value is away from the mean of all Δ_{TPM} values. The *zscore* is given by:

$$zscore(x) = \frac{x - \mu}{\sigma} \tag{2a}$$

$$\text{where } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2b}$$

$$\text{where } \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{2c}$$

where x is the Δ_{TPM} value, μ is the mean of all Δ_{TPM} values, σ is the standard deviation of all Δ_{TPM} values.
We define a threshold of $zscore = 1.96$ to indicate significant changes in gene activity, which corresponds to a confidence level of 95% (p = 0.05). Genes with Δ_{TPM} values exceeding this threshold will be flagged as *true* in the $\Delta_{TPM\_relevant}$ column. By this we fulfilled our first objective 1 by finding 1,034 genes with significant changes in gene activity.
The resulting table contains 17,627 gene nodes as rows with their associated attributes, including TPM values and derived metrics such as Δ_{TPM} or $zscore$. The head of the table is shown in Figure 8.

	Gene ID	Gene Name	norm healthy TPM	norm cancerous TPM	Δ TPM	Δ type	z score	Δ TPM relevant
0	ENSG00000121410	A1BG	0.220900	0.056729	-0.164171	decrease	-1.612	False
1	ENSG00000148584	A1CF	0.057012	0.080626	0.023614	increase	0.183	False
2	ENSG00000175899	A2M	0.579917	0.092398	-0.487519	decrease	-4.704	True
3	ENSG00000166535	A2ML1	0.332704	0.062953	-0.269751	decrease	-2.622	True
4	ENSG00000184389	A3GALT2	0.021656	0.046758	0.025101	increase	0.198	False

Figure 8: Example data of gene nodes that will be used in the graph database

Protein nodes

Lastly we create the table for the protein nodes. To collect all protein nodes for our database creation we use the gene-protein edges(Section 3.2 - Gene-protein edges) and the protein-protein edges (3.2 - Protein-protein-edges), which we already created.
From the gene-protein edge table, we extract all unique protein Ensembl IDs. From the protein-protein edge table, we concat both columns of protein Ensembl IDs and extract all unique values. We then merge both tables and drop duplicates again to get a list of all unique protein Ensembl IDs from both tables.
The resulting table is a list of 104,235 unique Protein Ensembl IDs as shown in Figure 9.

	index	Protein ID
0	0	ENSP00000354687
1	1	ENSP00000355046
2	2	ENSP00000354499
3	3	ENSP00000354876
4	4	ENSP00000355265

Figure 9: Example data of protein nodes that will be used in the graph database

3.3 Graph Database

As we have created our base data models, our next objective 2 is to create a graph database that can be used to perform the PageRank algorithm as a graph algorithm. With four huge datasets at our disposal, optimizing the generation of the database is crucial. In this section, we describe the queries, called cypher queries, used to create and populate our graph database, including information about creating indexes, constraints, relationships between nodes, and loading data into the database.

To construct the **basic PPI network**, we need to create protein nodes and their interactions as edges. To optimize query execution, we start with indexing the protein property ID. Next we load the data (3.2 - Protein nodes) as a list and then create the nodes in batches for efficient processing. The graph database is populated with 104,235 protein nodes.

Since these nodes are used solely for connections between genes, no additional properties are required. The cypher query for creating protein nodes is:

```
CREATE (p:protein {id: 'Protein ID'})
```

We then create interaction edges by loading the saved data (3.2 - Protein-protein edges) as a list of protein tuples and searching for both protein nodes by their ID in the graph database. The edges are created in batches for efficient processing, resulting in 11,247,242 interactions between protein nodes. The cypher query for creating protein-protein edges is:

```
MATCH (s:protein{id:'left Protein ID'})
MATCH (s:protein{id:'right Protein ID'})
CREATE (s)-[:interaction]->(t)
```

We now have a complete PPI network in place.

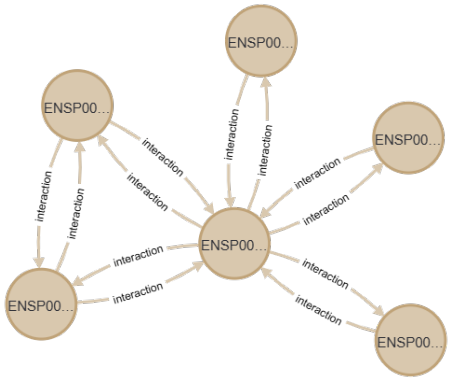

Figure 10: Extract from the graph database setup showing example protein nodes and their interactions

To construct our **extended PPI network** we need to integrate genes into the created protein network. First, we implement an index on the gene property *ID* for faster querying. We load the saved data (3.2 - Gene nodes) as a list and create 17,626 gene nodes in batches with their properties.
The cypher query employed for creating these gene nodes is:

```
CREATE (p:gene {
       id: 'Gene ID',
       gene_name: 'Gene Name',
       norm_healthy_tpm: 'norm healthy TPM',
       norm_cancerous_tpm: 'norm cancerous TPM',
       delta_tpm: 'Δ TPM',
       delta_type: 'Δ type',
       z_score: 'z score',
       delta_tpm_relevant: 'Δ TPM relevant'})
```

We characterize the genes in our network using these properties: *gene_name*, *norm_healthy_tpm*, *norm_cancerous_tpm*, *delta_tpm*, *delta_type*, *z_score*, and *delta_tpm_relevant*.
Among these, the value of *delta_tpm_relevant* stands out as a pivotal factor in our analysis. Although other attributes may be less relevant to our current investigation, they retain potential value for future tasks.

To model relationships between genes and proteins, we establish edges between these nodes by loading the gene-protein interaction data (3.2 - Gene-protein edges) as a list of tuples. This involves matching gene and protein nodes based on their respective IDs, and creating edges in batches to optimize processing efficiency. The result is a comprehensive network of 101,731 connections between gene and protein nodes.
The cypher query for creating gene-protein connection edges is:

```
MATCH (s:protein{id:'Protein ID'})
MATCH (s:gene{id:'Gene ID'})
CREATE (s)-[:connection]-(t)
```

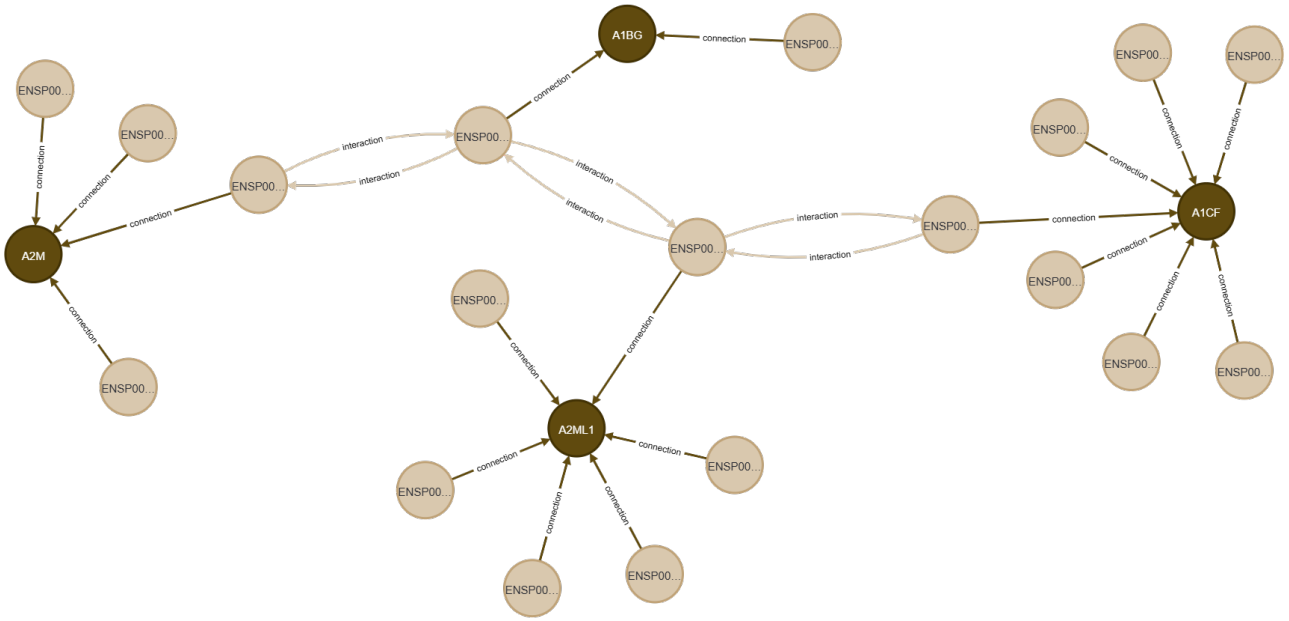

Figure 11: Extract from the graph database setup showing example gene nodes and their connections to protein nodes

Through the execution of these cypher queries, we successfully populate our graph database with the required nodes and edges, enabling us to perform the PageRank algorithm.

3.4 Graph Database Algorithm

As our final step for objective 2, we employ the **PageRank** algorithm to measure the importance of nodes within our graph database. This algorithm is particularly well-suited for identifying genes that play a crucial role in the network, enabling us to identify not only highly connected genes but also those with significant functional relevance.
For efficient computation and scalability, we start with creating a projection of the entire graph prior to applying PageRank analysis to our large-scale graph database.
To perform the PageRank analysis, we use the following cypher query:

```
CALL gds.pageRank.stream('gene_protein_graph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).id AS Gene_ID,
       gds.util.asNode(nodeId).gene_name AS Gene_Name,
       score,
       gds.util.asNode(nodeId).Δ_TPM AS Δ_TPM,
       gds.util.asNode(nodeId).Δ_TPM_relevant AS Δ_TPM_relevant
ORDER BY score DESC
```

This query returns a list of nodes, including genes and proteins, along with their respective PageRank scores. By filtering out the proteins from this list, we isolated the genes and further refined them to include only those exhibiting a significant change in gene activity, indicated by the value for *delta_tpm_relevant*.
The resulting subset of genes represents a group that has not only high connectivity within the network but also exhibit a substantial difference in gene expression.
In conclusion, by successfully completing our first and second objective (Item 1 and **??**, we have created the base data, established a graph database and performed the PageRank algorithm on our network.

This has allowed us to identify key genes within the network that exhibit both high connectivity and significant changes in gene activity.

4 Results

Within this chapter, we present the crucial component of our study, where we distill the essence of the extended PPI network by identifying the top 10 genes that stand out as significant contributors. These genes have been selected based on their high connectivity and substantial change in gene activity within the network.

Before exploring these key genes, it is essential to provide an overview of the distribution of PageRank scores and Δ_{TPM} values across the gene dataset. This will allow us to better understand the characteristics of our data and how they relate to one another.
The PageRank distribution exhibit an initial peak at a score of approximately 67, which stands out as an outlier in comparison to other scores (see Figure 12). This peak is followed by a gradual decline to around 40, with subsequent scores showing progressively smaller differences as they approach the mid-30s range. Notably, once the scores fall to this level, the gaps between them become considerably narrower, suggesting a pattern of logarithmic decay. From the 1033 relevant genes 155 have the lowest Pagerank score of 0.151. These genes consist of only one edge with a single protein that has not that many connections to other proteins.

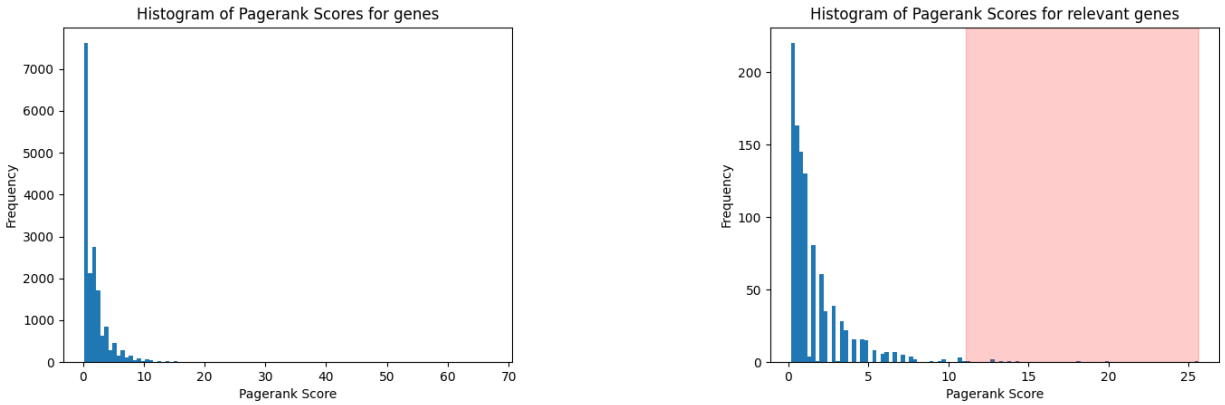

Figure 12: Comparison of the distribution of Pagerank Scores for all genes and genes with significant change in gene activity with highlighted area for the top 10 genes with the highest PageRank scores

The Δ_{TPM} values display a symmetric distribution with a pronounced peak around zero, showing a central tendency near the mean (see Figure 13). The data ranges from -0.709 to 0.423, showing moderate variability, and a normal-like distribution. The significant change in gene activity is defined as a Δ_{TPM} value between -0.709 and -0.201 or 0.210 and 0.423. We observe that 59% of the relevant genes have an increase in gene activity and 41% a decrease of activity.

To identify the most significant genes, we extracted the **top 10 genes** with the highest PageRank scores from the list of relevant genes, as shown in Figure 14. These genes are not only highly connected within the network but also exhibit a substantial change in gene activity. The PageRank scores are between 25.635 and 11.090, while the Δ_{TPM} values span a range from 0.214719 to -0.317776, with only one gene showing a decrease in gene activity.

As an illustrative example of the structure of the top genes in our network, we present a gene with high PageRank score and a significant change in gene activity (see Figure 15).

The 10 genes presented here have been identified as key players in the extended PPI network related

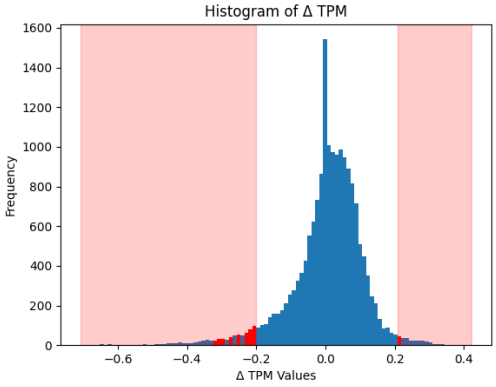

Figure 13: Distribution of Δ_{TPM} values for genes with the area for significant change in gene activity highlighted in red

	Gene_ID	Gene_Name	score	Δ_TPM	Δ_TPM_relevant
0	ENSG00000165795	NDRG2	25.635	-0.317776	True
1	ENSG00000161249	DMKN	19.901	-0.203997	True
2	ENSG00000092529	CAPN3	18.136	-0.310797	True
3	ENSG00000157103	SLC6A1	14.170	-0.267816	True
4	ENSG00000110436	SLC1A2	13.729	-0.233129	True
5	ENSG00000012048	BRCA1	13.289	0.214719	True
6	ENSG00000022267	FHL1	12.850	-0.217644	True
7	ENSG00000197971	MBP	12.849	-0.296146	True
8	ENSG00000049540	ELN	11.119	-0.246766	True
9	ENSG00000172995	ARPP21	11.090	-0.201848	True

Figure 14: Top 10 genes with the highest PageRank scores and a significant change in gene activity

to lung cancer. By extracting these top-scoring genes, we not only highlight potential biomarkers but also shed light on the intricate relationships within this complex biological system. To further validate these findings, we will conduct a detailed analysis of the current literature and compare our results with existing studies.

5 Discussion

The discussion section that follows delves into the practical implications of our research, where we examine how the insights gained from analyzing the top 10 genes can be translated into actionable recommendations for future studies or clinical applications. We also acknowledge the limitations of our approach and highlight potential avenues for improving the accuracy and reliability of predictive models in lung cancer diagnostics.

5.1 Result Analysis

As the final milestone of our study goal we need to fulfill the last objective (see 3) by analyzing and validating the results of the PageRank algorithm.

NDRG2

The gene **NDRG2** (N-Myc downstream-regulated gene 2) plays a critical role as a tumor suppressor, regulating cell growth and preventing the spread of cancer cells to other parts of the body. Its

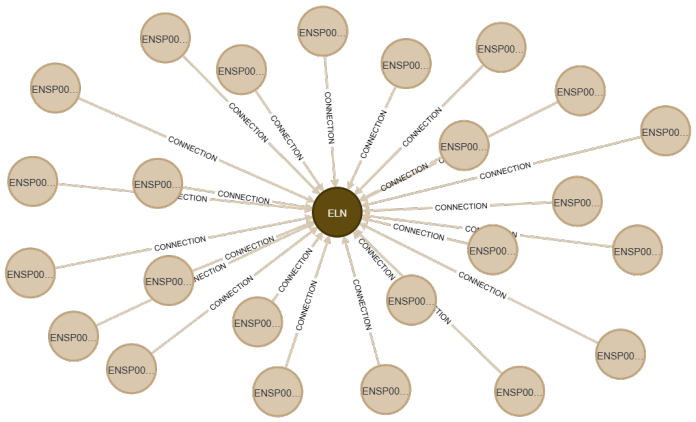

Figure 15: An example extract from the graph database for a gene with high PageRank score and significant change in gene activity

importance is particularly pronounced in tissues that are more likely for tumorigenesis, highlighting its potential as a key regulator of cellular behavior [28]
Recent studies have shed light on NDRG2's role in human lung cancer, revealing that low expression levels of this gene are associated with aggressive tumor behavior and poor prognosis in patients. Specifically, research has demonstrated that NDRG2 acts as a crucial suppressor of tumor cell metastasis and functions as a negative regulator of tumor progression [33]. Furthermore, another study has shown that downregulation of NDRG2, particularly in small-cell lung cancer, inhibits tumor progression, underscoring the gene's importance in this context [35].
Our analysis highlights NDRG2 as a potential key player in the lung cancer landscape, with a PageRank score of 25.635 and a significant downregulation in lung cancer cells (Δ_{TPM} = -0.317). This finding is consistent with previous studies showing that NDRG2 acts as a tumor suppressor, inhibiting metastasis and tumor progression [33, 35].
Notably, it boasts the highest delta TPM value among the top 10 genes, indicating a substantial decrease in expression levels in lung cancer cells.

DMKN

The gene **DMKN**, also known as Dermokine, is a gene involved in skin cell function and regulation. Specifically, it plays a crucial role in maintaining the integrity of the outer layer of the skin by regulating various cellular processes [37].
While DMKN has not been directly implicated in lung cancer, its connections to other cancers offer valuable insights into its potential function. Notably, research has demonstrated that DMKN-α is overexpressed in pancreatic cancer [63] and the β isoform is downregulated in skin cancer and serves as a biomarker for early-stage colorectal cancer [20].
It has a PageRank score of 19.900 and is downregulated in lung cancer cells, with a Δ_{TPM} value of -0.203, indicating further investigation into this gene's role is warranted. Our findings suggest that DMKN's involvement in lung cancer may be linked to its role in maintaining tissue integrity, potentially serving as a tumor suppressor. The gene's downregulation in lung cancer cells could compromise the skin's protective barrier function, facilitating tumor growth and spread.

The gene **CAPN3**, also known as Calpain 3, is an integral component of the Calpain system, a complex network of gene products and proteins that play critical roles in cellular processes. Primarily expressed in skeletal muscle cells, CAPN3 plays a pivotal role in maintaining muscle function and integrity [52]. Although CAPN3 has been extensively studied in the context of muscle diseases, its involvement in cancer, particularly lung cancer, remains poorly understood. However, research has shown that the Calpain system, which includes CAPN3, is implicated in cancer progression and metastasis [54]. Notably, studies have demonstrated that another member of the Calpain system, Calpain 2, is upregulated in

lung cancer [61].
Our analysis suggests that CAPN3 may be a gene of interest in lung cancer research, with a PageRank score of 18.136 and significant downregulation (Δ_{TPM} = -0.310) in lung cancer cells, indicating its potential importance in the disease. This finding could indicate that CAPN3's loss of function contributes to the disruption of muscle integrity and structure, potentially facilitating tumor growth and metastasis. The gene's significant downregulation in lung cancer cells (second-highest decrease in expression levels among the top 10 genes) suggests a potential link between muscle dysfunction and cancer progression. As studies have already shown that the Calpain system is involved in cancer progression and metastasis [54], and especially Calpain 2 is involved in lung cancer [61], there needs to be further research on CAPN3 in lung cancer.

The **SLC6A1** gene encodes a protein that functions as a GABA carrier, playing a crucial role in regulating neurotransmitter levels at synapses between neurons [10].
Research has demonstrated that SLC6A1 promotes cancer growth in certain types of tumors, including clear cell renal cell carcinoma, ovarian cancer, and prostate cancer [10]. This suggests a potential role for SLC6A1 in tumor progression and metastasis.
Our analysis found that SLC6A1 is downregulated in lung cancer cells, with a PageRank score of 14.170 and a Δ_{TPM} value of -0.267. This result appears counterintuitive, given SLC6A1's established role as an oncogene in other types of cancer [10]. Further investigation into the mechanisms underlying SLC6A1's expression in lung cancer is warranted to clarify its role in this disease.

The **SLC1A2** gene encodes a protein that functions as a glutamate transporter, playing a crucial role in regulating glutamate levels in the brain by removing glutamate from synapses [40].
Research has identified SLC1A2 as a partner gene in fusion genes associated with certain types of cancer. Notably, studies have found that CD44-SLC1A2 fusions are present in gastric cancer [59] and primary colorectal cancer [50], indicating its potential role in cancer development. Additionally, an analogous APIP/SLC1A2 fusion has been identified in colon cancer [19]. Notably, these gene fusions were not detected in non-small cell lung cancers [50].
In contrast to SLC1A2's association with oncogenic fusion genes in other cancers, our analysis reveals that it is downregulated in lung cancer cells, with a PageRank score of 13.729 and a Δ_{TPM}value of -0.233. This finding raises questions about the gene's role in lung cancer and warrants further investigation. The downregulation of SLC1A2 in lung cancer may indicate that the tumor suppressive effects of this gene are lost or disrupted, contributing to tumor progression.

The **BRCA1** gene, also known as Breast Cancer Gene 1, is a tumor suppressor gene that produces a protein involved in repairing damaged DNA. This process is essential for maintaining genome stability by facilitating the correct repair of DNA breaks [23].
BRCA1 has been associated with various types of cancer, including breast and ovarian cancers [29]. However, its relationship to lung cancer is more complex. Initial research suggested a potential link between overexpressed BRCA1 and poor survival in non-small cell lung cancer patients [46]. However, subsequent studies have contradicted this notion, saying that BRCA1 does not play a role in NSCLC [17, 29].
Interestingly, our analysis reveals a different pattern for BRCA1 in lung cancer cells. With a PageRank score of 13.288 and a Δ_{TPM} value of 0.214, BRCA1 is upregulated in these cells, making it the only gene in our top 10 list to exhibit this behavior. The upregulation of BRCA1 in lung cancer may indicate a distinct molecular mechanism at play, potentially involving the activation of tumor-promoting pathways or the suppression of normal DNA repair functions. Given BRCA1's role in maintaining genome stability, its overexpression in lung cancer may indicate a compensatory response to genomic stress or other forms of cellular damage.

The **FHL1** (four and a half LIM domains 1) gene encodes a protein involved in regulating muscle cell differentiation and maturation, playing a crucial role in maintaining proper muscle function. It is primarily expressed in striated muscles but also found in other tissues such as the brain and testis [53].

FHL1 has been implicated in various types of cancer, including breast, kidney, prostate and gastric cancer. Research suggests that FHL1 expression is often reduced in these cancers, which may contribute to their development [34, 48]. Specifically, studies have found that FHL1 is downregulated in NSCLC patients [38].
Our analysis reveals that FHL1 has a PageRank score of 12.849 and is downregulated in lung cancer cells, with a Δ_{TPM} value of -0.217. This finding is consistent with previous studies suggesting that FHL1 expression is reduced in various types of cancer, including NSCLC, which may contribute to tumor development. The loss of FHL1's tumor suppressive effects could potentially contribute to the aggressiveness and metastatic potential of NSCLC.

The **MBP** (Myelin Basic Protein) gene plays a crucial role in producing proteins essential for creating and maintaining the protective myelin covering around nerve fibers, as well as facilitating repair of damaged nerve coverings and regulating immune cell responses to infections or inflammation [39].
MBP has been linked to cancer research, particularly in the context of brain tumors. Studies have suggested that MBP levels are elevated in patients with brain cancer and may serve as a potential biomarker for diagnosis. However, increased MBP levels have also been observed in other brain diseases, such as multiple sclerosis [62]. Furthermore, research has found a correlation between MBP expression levels and brain metastasis from lung cancer [36].
Interestingly, our analysis reveals that MBP is downregulated in lung cancer cells with a Δ_{TPM} value of -0.296. This finding contradicts previous research suggesting elevated MBP levels in brain tumors, including those metastasizing from lung cancer. The discrepancy between these findings raises questions about the role of MBP in lung cancer and its relationship to brain metastasis. One possible explanation is that the downregulation of MBP in lung cancer cells may be a result of epigenetic modifications or other regulatory mechanisms that prevent the normal expression of this gene. Further investigation is needed to fully understand the role of MBP in lung cancer biology and its potential as a biomarker for diagnosis.

The **ELN** gene encodes for the protein Elastin, a key component of connective tissue that provides elasticity and flexibility to tissues such as skin, lungs, and blood vessels [12].
According to studies, elastin plays a significant role in cancer development. Research has shown that elastin gene expression is increased in tumors from colorectal cancer patients [31]. In breast cancer, elastosis, a condition characterized by an abnormal increase in elastin fibers, is a common feature that increases with tumor progression [30]. Additionally, studies have found that Elastin in gastric cancer tissues is linked to changes that enable cancer cells to spread more easily [14].
Our findings indicate that ELN expression is downregulated in lung cancer cells, a paradoxical finding given its upregulation in other types of cancers. With a Δ_{TPM} value of -0.246, this unexpected result could suggest a unique role for Elastin in the context of lung cancer. The divergence from the established pattern of elastin's involvement in cancer development implies that lung cancer may exploit distinct pathways to regulate elastin expression.

The **ARPP21** (cAMP regulated phosphoprotein 21) gene is involved in regulating cellular processes, particularly in brain development and function. It plays a crucial role in shaping the complexity of neurons during development [45].
In addition to its direct impact on brain development, research has uncovered an indirect link between ARPP21 and cancer. Specifically, the microRNA produced by ARPP21, miR-128, has been implicated in various types of cancer [32]. Notably, studies have shown that miR-128 is downregulated in prostate cancer [26], ovarian cancer [60], and breast cancer [64].
While the direct connection between ARPP21 and cancer remains unclear, its connection to miR-128 suggests an indirect role for ARPP21 in cancer development through this microRNA. Our analysis provides further insight into this relationship, revealing that ARPP21 is downregulated in lung cancer cells, with a Δ_{TPM} value of -0.201. This result stands in contrast to the upregulation of miR-128 observed in other types of cancer, such as breast and ovarian cancer. Further investigation is needed to elucidate the mechanisms by which ARPP21 expression influences lung cancer biology, and whether its downregulation in lung cancer cells has implications for tumor progression or response to treatment.

In conclusion, our analysis has revealed a diverse set of top-ranked genes for lung cancer biomarkers, each with unique characteristics and relationships to lung cancer. The genes are involved in various cellular processes, including cell growth, muscle function, and neurotransmitter regulation. While some genes, such as NDRG2 and MBP, exhibit direct connections to lung cancer, others display complex or paradoxical roles in the disease.
Interestingly, even among the top-ranked genes, their connection to lung cancer is not always clear-cut. For instance, BRCA1 has a controversial role in lung cancer, with some studies suggesting that it may act as an oncogene, while others propose a tumor suppressor function. Moreover, several genes on our list remain understudied in the context of lung cancer, underscoring the necessity of continued research to elucidate their potential involvement.
Our findings collectively emphasize that lung cancer is a multifaceted disease with a complex genetic underpinning, and underscore the importance of interdisciplinary approaches and continued investigation into the roles of these genes in disease progression. By shedding light on these previously understudied genes, our analysis highlights new avenues for research and potential therapeutic targets in lung cancer.

5.2 Study Evaluation

Evaluation of the Network with cancer-associated Genes

To evaluate the robustness of our network construction and PageRank algorithm, we assess the placement of genes known to be involved in cancer development within our network. Specifically, we examined a set of 9 genes identified as being associated with lung cancer in a study by El-Telbany et al. [13]. We sought to determine whether these cancer-associated genes would emerge as hub nodes or exhibit high PageRank scores, and how their expression levels Δ_{TPM} compare to those of other genes within our network.
We found that all 9 genes are present in our network, which is a positive validation of the comprehensive coverage of our dataset. Notably, the gene LKB1 is listed under the name STK11 in our data, which is a common alias for this gene. 16 provides an illustration of these genes and their respective Δ_{TPM} values and PageRank scores.

	Gene_ID	Gene_Name	score	Δ_TPM	Δ_TPM_relevant
0	ENSG00000171094	ALK	1.560	0.042775	False
1	ENSG00000157764	BRAF	4.093	0.038695	False
2	ENSG00000146648	EGFR	2.386	0.097984	False
3	ENSG00000133703	KRAS	4.959	0.104333	False
4	ENSG00000105976	MET	1.969	0.265047	True
5	ENSG00000121879	PIK3CA	1.969	0.034747	False
6	ENSG00000165731	RET	4.093	0.113269	False
7	ENSG00000047936	ROS1	0.783	0.059812	False
8	ENSG00000118046	STK11	4.525	-0.026488	False

Figure 16: 9 known cancer-associated genes and their respective Δ_{TPM} values and PageRank scores

When examining the placement of these cancer-associated genes within the network, we observed that only one gene (with a Δ_{TPM} value of 0.265) is classified as $deltaTPMrelevant$. None of the others have a notable Δ_{TPM} value, with most being situated in the middle of the distribution and showing no clear pattern or outliers.
Upon examining the placement of these cancer-associated genes, we observed that none of them have a pagerank score greater than 5, which seems to be far away from the top 10 genes. However, as shown in Fig. 17, 4 of the 9 genes are situated within the top 10 percent of the distribution (PR score ¿ 3.67), and all are above the median, indicating that they are relatively well-connected nodes within our network.

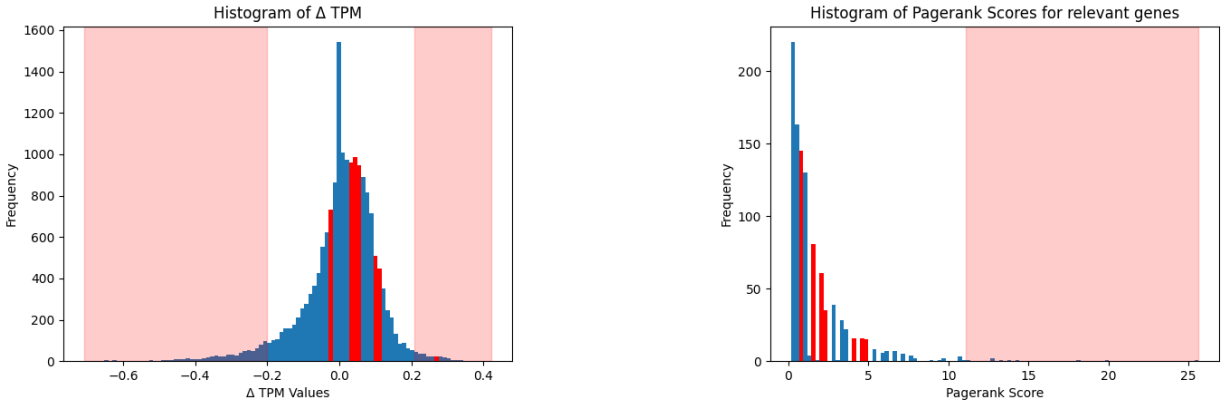

Figure 17: Distribution of the Pagerank score and the Δ_{TPM} with highlighting of values for 9 known cancer-associated genes

Our analysis suggests that while our network construction and PageRank algorithm are robust, the placement of these cancer-associated genes within the network may not be directly indicative of their role in cancer development.

Limitation of the Δ_{TPM} and the $\Delta TPMrelevant$ measure

The calculated Δ_{TPM} values represent the relative difference in gene expression between normalized mean cancerous and healthy states, indicating an increase or decrease in transcript abundance in the tumor compared to normal tissue. We classified the significant 5% of the genes as $\Delta TPMrelevant$ in our study.
The top 10 genes that we looked at all had a connection to cancer based on previous studies - even if not all of them had a connection to lung cancer. However, 10 genes is a rather small group to make a general statement of the quality of the delta TPM measure we developed. By looking at the genes known for lung cancer, only one was classified as having relevant changes in gene activity. This raises questions about the robustness and sensitivity of our approach. While it seems like a solid starting point, there are some limitations and improvements that could be made based on the assumptions we made.
One potential problem may be that our calculation assumes a linear relationship between healthy and cancerous TPM values by subtracting the log-scaled mean TPM values. Additionally, the log-scaled normalization method of TPM values might not be optimal, and could be improved by using a different, e.g. non-linear method such as quantile normalization. Furthermore, our measure may be too general, potentially leading to false negatives for genes that are more sensitive to smaller changes. To address these concerns, it could be a way to investigate the correlation between xxx with methods like fold change and enrichment scores. Another improvement could be to use an additional measure such as FPKM or RPKM to validate the TPM results and not rely on only one measure.

Limitation of the PageRank score

* Pagerank might be too focused on first level Proteins

* Other datasets could be used like TCGA which is common in cancer research
* CMP Dataset matched with ENS ID Names might be wrong - would be better to find a good dataset with ens IDs

5.3 Future Work

* Pagerank is a good way to find important genes - well established / known / typical research
* Clear finding of new genes that may be biomarkers
* Analyse the top 10 genes seems to find good results

* These should be further investigated in future studies
* collective behavior of multiple genes rather than focusing solely on individual genes
* Other datasets may be used - especcialy CMP is a bit clupsy with the ENS ID

6 Conclusion

References

[1] Cell model passport - model list. https://cellmodelpassports.sanger.ac.uk/passports?datasets=expression. Accessed: 25.11.2024.

[2] Cmp - downloads - expression data. https://cellmodelpassports.sanger.ac.uk/downloads. Accessed: 09.10.2024.

[3] A de-identified, open access version of the subject phenotypes available in dbgap. https://storage.googleapis.com/adult-gtex/annotations/v8/metadata-files/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt. TXT-File with GTEx Model anotation, Accessed: 25.11.2024.

[4] Ensembl biomarts. https://www.ensembl.org/biomart/martview/. Accessed: 14.10.2024.

[5] The ensembl project. https://www.ebi.ac.uk/training/online/courses/ensembl-browsing-genomes/what-is-ensembl/ensembl-project/. Accessed: 09.10.2024.

[6] Graph algorithms in neo4j: 15 different graph algorithms and what they do. https://neo4j.com/blog/graph-algorithms-neo4j-15-different-graph-algorithms-and-what-they-do/. Accessed: 23.09.2024.

[7] Gtex portal. https://www.gtexportal.org/home/. Accessed: 09.10.2024.

[8] Gtex portal - rna-seq. https://www.gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression. Accessed: 09.10.2024.

[9] String database for homo sapiens. https://stringdb-downloads.org/download/protein.links.full.v12.0.txt.gz. Accessed: 04.12.2024.

[10] Chaojiang Chen, Zhiduan Cai, Yangjia Zhuo, Ming Xi, Zhuoyuan Lin, Funeng Jiang, Zezhen Liu, Yueping Wan, Yu Zheng, Jianxin Li, Xing Zhou, Jianguo Zhu, and Weide Zhong. Overexpression of slc6a1 associates with drug resistance and poor prognosis in prostate cancer. *BMC Cancer*, 20(1):289, apr 2020. Accessed: 28.10.2024.

[11] Sreyashi Das, Mohan Kumar Dey, Ram Devireddy, and Manas Ranjan Gartia. Biomarkers in cancer detection, diagnosis, and prognosis. *Sensors*, 24(1), 2024.

[12] L. Debelle and A.M. Tamburro. Elastin: molecular description and function. *The International Journal of Biochemistry & Cell Biology*, 31(2):261–272, 1999. Accessed: 29.10.2024.

[13] Ahmed El-Telbany and Patrick C. Ma. Cancer genes in lung cancer: Racial disparities: Are there any? *Genes & Cancer*, 3(7-8):467–480, 2012. PMID: 23264847.

[14] Tianyi Fang, Lei Zhang, Xin Yin, Yufei Wang, Xinghai Zhang, Xiulan Bian, Xinju Jiang, Shuo Yang, and Yingwei Xue. The prognostic marker elastin correlates with epithelial–mesenchymal transition and vimentin-positive fibroblasts in gastric cancer. *The Journal of Pathology: Clinical Research*, 9(1):56–72, 2023. Accessed: 01.11.2024.

[15] J Ferlay, M Ervik, F Lam, M Laversanne, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: Cancer today. https://gco.iarc.who.int/today. Accessed: 27.09.2024.

[16] Tanguy Ferlier and Cédric Coulouarn. Regulation of gene expression in cancer—an overview. *Cells*, 11:4058, 12 2022. Accessed: 17.11.2024.

[17] Mariam Gachechiladze and Josef Skarda. The role of brca1 in non-small cell lung cancer. *Biomedical papers*, 156(3):200–203, 2012. Accessed: 01.11.2024.

[18] CD Genomics. Rpm, rpkm, fpkm, and tpm: Normalization methods in rna sequencing. Accessed: 19.11.2024.

[19] Craig P. Giacomini, Steven Sun, Sushama Varma, A. Hunter Shain, Marilyn M. Giacomini, Jay Balagtas, Robert T. Sweeney, Everett Lai, Catherine A. Del Vecchio, Andrew D. Forster, Nicole Clarke, Kelli D. Montgomery, Shirley Zhu, Albert J. Wong, Matt van de Rijn, Robert B. West, and Jonathan R. Pollack. Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. *PLOS Genetics*, 9(4):1–19, 04 2013. Accessed: 31.10.2024.

[20] Minoru Hasegawa, Kiyoshi Higashi, Chikako Yokoyama, F Yamamoto, Taro Tachibana, Takashi Matsushita, Y Hamaguchi, K Saito, M Fujimoto, and K Takehara. Altered expression of dermokine in skin disorders. *Journal of the European Academy of Dermatology and Venereology : JEADV*, 27, 05 2012.

[21] National Cancer Institute. Lung cancer—patient version. `https://www.cancer.gov/types/lung`. Accessed: 27.09.2024.

[22] National Cancer Institute. Risk factors for cancer. `https://www.cancer.gov/about-cancer/causes-prevention/risk`. Accessed: 27.09.2024.

[23] National Cancer Institute. Brca gene changes: Cancer risk and genetic testing, 2020. Accessed: 29.10.2024.

[24] National Cancer Institute. What is cancer? `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`, 2021. Accessed: 27.09.2024.

[25] National Human Genome Research Institute. Gene expression. `https://www.genome.gov/genetics-glossary/Gene-Expression`. Accessed: 17.11.2024.

[26] Amjad P. Khan, Laila M. Poisson, Vadiraja B. Bhat, Damian Fermin, Rong Zhao, Shanker Kalyana-Sundaram, George Michailidis, Alexey I. Nesvizhskii, Gilbert S. Omenn, Arul M. Chinnaiyan, and Arun Sreekumar. Quantitative proteomic profiling of prostate cancer reveals a role for mir-128 in prostate cancer*. *Molecular & Cellular Proteomics*, 9(2):298–312, 2010. Accessed: 02.11.2024.

[27] Rohit kumar Kaliyar. Graph databases: A survey. In *International Conference on Computing, Communication & Automation*, pages 785–790, 2015. Accessed: 20.11.2024.

[28] K. W. Lee, S. Lim, and K. D. Kim. The function of n-myc downstream-regulated gene 2 (ndrg2) as a negative regulator in tumor cell metastasis. *International Journal of Molecular Sciences*, 23(16):9365, 2022. Accessed: 27.10.2024.

[29] Yen-Chien Lee, Yang-Cheng Lee, Chung-Yi Li, Yen-Ling Lee, and Bae-Ling Chen. Brca1 and brca2 gene mutations and lung cancer risk: A meta-analysis. *Medicina*, 56(5), 2020. Accessed: 01.11.2024.

[30] Arkadiusz Lepucki, Kinga Orlińska, Aleksandra Mielczarek-Palacz, Jacek Kabut, Pawel Olczyk, and Katarzyna Komosinska-Vassev. The role of extracellular matrix proteins in breast cancer. *Journal of Clinical Medicine*, 11:1250, 02 2022. Accessed: 01.11.2024.

[31] Jinzhi Li, Xiaoyue Xu, Yanyan Jiang, Nicole Hansbro, Philip Hansbro, Jincheng Xu, and Gang Liu. Elastin is a key factor of tumor development in colorectal cancer. *BMC Cancer*, 20, 03 2020. Accessed: 01.11.2024.

[32] Molin Li, Weiming Fu, Lulu Wo, Xiaohong Shu, Fang Liu, and Chuangang Li. mir-128 and its target genes in tumorigenesis and metastasis. *Experimental Cell Research*, 319(20):3059–3064, 2013. Accessed: 02.11.2024.

[33] SJ Li, WY Wang, B Li, B Chen, B Zhang, X Wang, CS Chen, QC Zhao, H Shi, and L Yao. Expression of ndrg2 in human lung cancer and its correlation with prognosis. *Med Oncol*, 30(1):421, mar 2013. Accessed: 30.10.2024.

[34] Xun Li, Zhenyu Jia, Yongquan Shen, Hitoshi Ichikawa, Jonathan Jarvik, Robert G. Nagele, and Gary S. Goldberg. Coordinate suppression of sdpr and fhl1 expression in tumors of the breast, kidney, and prostate. *Cancer Science*, 99(7):1326–1333, 2008. Accessed: 01.11.2024.

[35] Zhenchuan Ma, Yuefeng Ma, Jie Feng, Zhengshui Xu, Chuantao Cheng, Jie Qin, Shaomin Li, Jiantao Jiang, and Ranran Kong. Ndrg2 acts as a negative regulator of the progression of small-cell lung cancer through the modulation of the pten-akt-mtor signalling cascade. *Toxicology and Applied Pharmacology*, 485:116915, 2024. Accessed: 30.10.2024.

[36] Hidemitsu M.D. Nakagawa, Masanobu M.D. Yamada, Takuji M.D. Kanayama, Koichiro M.D. Tsuruzono, Yoji M.D. Miyawaki, Koji M.D. Tokiyoshi, Yasusi M.D. Hagiwara, and Toru M.D. Hayakawa. Myelin basic protein in the cerebrospinal fluid of patients with brain tumors. *Neurosurgery*, 34(5):825–833, may 1994. Accessed: 01.11.2024.

[37] Michael F. Naso, Bailin Liang, C. Chris Huang, Xiao-Yu Song, Lillian Shahied-Arruda, Stanley M. Belkowski, Michael R. D'Andrea, Debbie A. Polkovitch, Danielle R. Lawrence, Don E. Griswold, Ray W. Sweet, and Bernard Y. Amegadzie. Dermokine: An extensively differentially spliced gene expressed in epithelial cells. *Journal of Investigative Dermatology*, 127(7):1622–1631, 2007. Accessed: 28.10.2024.

[38] Chang Niu, Chaoyang Liang, Juntang Guo, Long Cheng, Hao Zhang, Xi Qin, Qunwei Zhang, Lihua Ding, Bin Yuan, Xiaojie Xu, Jiezhi Li, Jing Lin, and Qinong Ye. Downregulation and growth inhibitory role of fhl1 in lung cancer. *International Journal of Cancer*, 130(11):2549–2556, 2012. Accessed: 01.11.2024.

[39] Steven H. Nye, Clara M. Pelfrey, Jeffrey J. Burkwit, Rhonda R. Voskuhl, Michael J. Lenardo, and John P. Mueller. Purification of immunologically active recombinant 21.5 kda isoform of human myelin basic protein. *Molecular Immunology*, 32(14):1131–1141, 1995. Accessed: 29.10.2024.

[40] National Library of Medicine. Slc1a2 - solute carrier family 1 member 2. RefSeq Entry, Juni 2017.

[41] World Health Organization. Cancer. `https://www.who.int/news-room/fact-sheets/detail/cancer`. Accessed: 27.09.2024.

[42] World Health Organization. Lung cancer. `https://www.who.int/news-room/fact-sheets/detail/lung-cancer`. Accessed: 27.09.2024.

[43] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[44] Jaroslav Pokorný. Graph databases: Their power and limitations. In Khalid Saeed and Wladyslaw Homenda, editors, *Computer Information Systems and Industrial Management*, pages 58–69, Cham, 2015. Springer International Publishing. Accessed: 20.11.2024.

[45] Frederick Rehfeld, Daniel Maticzka, Sabine Grosser, Pina Knauff, Murat Eravci, Imre Vida, Rolf Backofen, and F. Gregory Wulczyn. The rna-binding protein arpp21 controls dendritic branching by functionally opposing the mirna it hosts. *Nature Communications*, 9(1):1235, 2018. Accessed: 29.10.2024.

[46] Rafael Rosell, Marcin Skrzypski, Ewa Jassem, Miquel Taron, Roberta Bartolucci, Jose Javier Sanchez, Pedro Mendez, Imane Chaib, Laia Perez-Roca, Amelia Szymanowska, Witold Rzyman, Francesco Puma, Grazyna Kobierska-Gulida, Raffaele Farabi, and Jacek Jassem. Brca1: A novel prognostic factor in resected non-small-cell lung cancer. *PLOS ONE*, 2(11):1–7, 11 2007. Accessed: 01.11.2024.

[47] Trilochan Rout, Anjali Mohapatra, and Madhabananda Kar. A systematic review of graph-based explorations of ppi networks: methods, resources, and best practices. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1):29–?, may 2024.

[48] Katsuya Sakashita, Koshi Mimori, Fumiaki Tanaka, Yukio Kamohara, Hiroshi Inoue, Tetsuji Sawada, Kosei Hirakawa, and Masaki Mori. Clinical significance of loss of fhl1 expression in human gastric cancer. *Annals of Surgical Oncology*, 15(8):2293–2300, aug 2008. Accessed: 01.11.2024.

[49] Haixia Shang and Zhi-Ping Liu. Network-based prioritization of cancer genes by integrative ranks from multi-omics data. *Computers in Biology and Medicine*, 119:103692, 2020.

[50] Kazuya Shinmura, Hisami Kato, Hisaki Igarashi, Yusuke Inoue, Satoki Nakamura, Chunping Du, Kiyotaka Kurachi, Toshio Nakamura, Hiroshi Ogawa, Masayuki Tanahashi, Hiroshi Niwa, and Haruhiko Sugimura. Cd44-slc1a2 fusion transcripts in primary colorectal cancer. *Pathology & Oncology Research*, 21, 01 2015. Accessed: 31.10.2024.

[51] Claire M Simpson and Florian Gnad. Applying graph database technology for analyzing perturbed co-expression networks in cancer. *Database*, 2020:baaa110, 12 2020.

[52] Simone Spinozzi, Sonia Albini, Heather Best, and Isabelle Richard. Calpains for dummies: What you need to know about the calpain family. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1869(5):140616, 2021. Accessed: 28.10.2024.

[53] Emily C Storey, Ian Holt, Glenn E Morris, and Heidi R Fuller. Muscle cell differentiation and development pathway defects in emery-dreifuss muscular dystrophy. *Neuromuscular Disorders*, 30(6):443–456, 2020. Accessed: 29.10.2024.

[54] S Storr, N Carragher, and M et al Frame. The calpain system and cancer. *Nat Rev Cancer*, 11:364–374, 2011.

[55] Prabal Subedi, Simone Moertl, and Omid Azimzadeh. Omics in radiation biology: Surprised but not disappointed. *Radiation*, 2(1):124–129, 2022. Accessed: 23.11.2024.

[56] Indhupriya Subramanian, Srikant Prasad Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14, 2020. Accessed: 23.11.2024.

[57] National Cancer Institute Surveillance Research Program. Seer*explorer: An interactive website for seer cancer statistics. `https://seer.cancer.gov/statistics-network/explorer/`. [Internet]. Updated: 2024 Jun 27; Cited: 2024 Sep 29. Available from: https://seer.cancer.gov/statistics-network/explorer/. Data source(s): SEER Incidence Data, November 2023 Submission (1975-2021), SEER 22 registries (excluding Illinois and Massachusetts). Expected Survival Life Tables by Socio-Economic Standards Accessed: 29.09.2024.

[58] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 11 2020.

[59] Jiong Tao, Nian Tao Deng, Kalpana Ramnarayanan, Baohua Huang, Hue Kian Oh, Siew Hong Leong, Seong Soo Lim, Iain Beehuat Tan, Chia Huey Ooi, Jeanie Wu, Minghui Lee, Shenli Zhang, Sun Young Rha, Hyun Cheol Chung, Duane T. Smoot, Hassan Ashktorab, Oi Lian Kon, Valere Cacheux, Celestial Yap, Nallasivam Palanisamy, and Patrick Tan. Cd44-slc1a2 gene fusions in gastric cancer. *Science Translational Medicine*, 3(77):77ra30–77ra30, 2011. Accessed: 31.10.2024.

[60] Ho-Hyung Woo, Csaba Laszlo, Stephen Greco, and Setsuko Chambers. Regulation of colony stimulating factor-1 expression and ovarian cancer cell behavior in vitro by mir-128 and mir-152. *Molecular cancer*, 11:58, 08 2012. Accessed: 02.11.2024.

[61] Fengkai Xu, Jie Gu, Chunlai Lu, Wei Mao, Lin Wang, Qiaoliang Zhu, Zhonghe Liu, Yiwei Chu, Ronghua Liu, and Di Ge. Calpain-2 enhances non-small cell lung cancer progression and chemoresistance to paclitaxel via egfr-pakt pathway. *Int J Biol Sci*, 15:127–137, 2019.

[62] M.G. Zavialova, VE Shevchenko, Evgeny (Eugene) Nikolaev, and Victor Zgoda. Is myelin basic protein a potential biomarker of brain cancer? *European Journal of Mass Spectrometry*, 23:192–196, 08 2017. Accessed: 01.11.2024.

[63] Yan Zhang, He wei Zhang, Xian dong Zhu, Yong qiang Wang, Xiao wu Wang, Bei shi Zheng, Bi cheng Chen, and Zong jing Chen. Overexpression of dermokine-a enhances the proliferation and epithelial-mesenchymal transition of pancreatic tumor cells. *Cellular Signalling*, 99:110439, 2022. Accessed: 30.10.2024.

[64] Yinghua Zhu, Fengyan Yu, Yu Jiao, Juan Feng, Wei Tang, Herui Yao, Chang Gong, Jianing Chen, Fengxi Su, Yan Zhang, and Erwei Song. Reduced mir-128 in breast tumor–initiating cells induces chemotherapeutic resistance via bmi-1 and abcc5. *Clinical Cancer Research*, 17(22):7105–7115, 11 2011. Accessed: 02.11.2024.