Allegheny College
Department of Computer Science

Senior Thesis

DEG-PI: A Computational Tool that Identifies Differentially Expressed Genes and Enriched Pathways in Cancer Tissues

ALLEGHENY COLLEGE by

Abigail Waryanka

COMPUTER SCIENCE

Project Supervisor: **Becky Dawson**
Co-Supervisor: **Oliver Bonham-Carter**

28 November 2022

Abstract

Cancer is a devastating disease affecting millions of people. It is caused by mutations in DNA that causes cells to grow uncontrollably in the body forming tumors. Bladder cancer is a common cancer being the tenth most diagnosed in 2020 with approximately 570,000 new cases. There are specific risk factors that increase the risk of developing bladder cancer. Some of these include gender, age, smoking, and occupational hazards such as aromatic amines exposure. Since there are a variety of different genetic and environmental causes of bladder cancer, further research has been done to try to identify differnetially expressed genes and enriched pathways. A differentially expressed gene (DEG) is a gene whose expression level is different in cancer tissue than normal tissue. An enriched pathway is a biological pathway that contains a statistically significant number of DEGs compared to what would be expected by chance. DEG-PI uses both R and Python to identify DEGS and enriched pathways in cancer tissue.

Table of Contents

Introduction **3**
- Cancer Basics . . . 4
- Bladder Cancer . . . 5
- DEGs and Enriched Pathways . . . 8
- Current State of the Art . . . 8

Related work **9**
- Cancer Research . . . 9
- Databases . . . 10
- Tools . . . 11

Method of approach **13**
- Algorithm . . . 15

Experiments **18**
- Results . . . 19
- Experimental Design and Evaluation . . . 24
- Threats to Validity . . . 24

Conclusion **26**
- Future Work . . . 27
- Future Ethical Implications and Recommendations . . . 28
- Final Thought . . . 29

References **30**

List of Figures

1 The algortihm shows how the program goes back and forth between Python and R to create a final project. The yellow square is displays what was completed in Python and the blue diamond is what was completed in R. 15

2 A table containing the Differntially Expressed Genes. 19

3 The tool also produces a volcano plot. A volcano plot is a type of scatter plot that analyzes statistical significance (x-axis) versus magnitude of change (y-axis). 20

4 The tool displays the results of gene ontology analysis. 21

5 A table with the 20 most enriched pathways is shown as a result of the pathway analysis. 21

6 The algortihm shows how the program goes back and forth between Python and R to create a final project. The yellow square is displays what was completed in Python and the blue diamond is what was completed in R. 23

Introduction

Cancer is a devastating disease affecting the lives of millions of people. In 2023, it is projected to cause the death of 600,000 individuals and give rise to nearly two million new cases in the United States alone [18]. Cancer is a disease where cells do not behave or multiply like normal cells. They grow uncontrollably in the body forming masses or tumors. As the tumors mature, cells can break off and travel via the blood stream or lympatic system to other parts of the body. There, they can form secondary tumors [19].

A cancer cell is altered in such a way that it is able to bypass the traditional mechanisms that control cell growth. Its DNA has been modified to either stop apoptosis (controlled cell death) [7] or stimulate pathways that increase cell growth [12]. Because cells are everywhere in the body, cancer can occur anywhere. Some of the most common types are lung, bladder, skin, and liver cancers. Additionally, there are a variety of subtypes. For example, there are at least six different breast cancers and seven different lung cancers. Cancer names are based on the cell type and the body region they originate from.

Due to the vastness and variety in cancer types, cancer reasearch id diffuclt and extensive. Consequently, data may not always be fully analyzed. The 3.2 billion base pairs that make up DNA contain only a small section where cancer-causing mutations occur, making it extremely challenging for humans to identify patterns among tissue samples even of the same cancer type [19]. One solution to the challenge of finding patterns in cancer-causing mutations is to utilize computer technology for research. By analyzing large amounts of genetic data with advanced algorithms and machine learning techniques, computers can help identify subtle patterns and genetic changes that may be difficult or impossible to detect with human analysis alone. This approach allows researchers to gain a more comprehensive understanding of the underlying mechanisms of cancer and may ultimately lead to more effective

treatments and prevention strategies.

Cancer Basics

Deoxyribonucleic acid, or DNA, is a polymer containing genetic code that holds all of the information needed for life. It is shaped in a double helix structure with two sugar-phosphate backbones connected by bases. The four bases, adenine, guanine, cytosine, and thymine, store genetic information similar to the way a computer uses 0's and 1's to store data [19]. The bases in DNA are used to create amino acids which then form chains to create proteins. Proteins are responsible for almost all functions inside a cell and consequently the body. The order of the bases in DNA is extremely important. The order determines the amino acid which is essential for the shape and function of the resulting protein [19]. A single base pair change can make a protein completely ineffective.

A gene is a segment of the DNA that is responsible for making you who you are. It contains the instructions for creating proteins or other molecules. Genes are often passed down from parents to their offspring. Each gene contains a specific sequence of nucleotides that determines the genetic code that is used to build proteins [19]. Different combinations of genes can lead to variations in physical traits, such as eye color or height, and can also contribute to the development of diseases or other health conditions.

Cells are not indestructible. If you were to get a cut on your hand, skin cells would inevitably die and you would lose blood cells. Cells die even without outside influence [19]. For example, red blood cells that carry oxygen have a lifecycle of about 120 days [2] and heart muscle cells have a lifecycle of a couple years [19]. In order to ensure that humans do not run out of cells, the body continuously creates new cells through a process known as mitosis. Mitosis is a form of cell division where one parent cell is multiplied and divided into two new child cells. The DNA and organelles are duplicated and then split in half. The child cells contain all of the organelles and genetic material needed to survive and perform the function required by the body for the specific cell type [19]. The child cell is an exact replica of the parent. Thus, if the parent cell is cancerous containing a mutation, the child cells will also be defective.

Cancer is a genetic disease meaning that ultimately cancer is caused by a change in the genes of an affected individual [19]. When a mutation occurs

in the DNA, it results in cancer. These mutations will often occur in three main gene types proto-oncogenes, tumor-suppressing genes, and DNA repair genes. Oncogenes, a potentially cancer-causing gene, is the activated version of proto-oncogenes. When normal, these genes play a role in cell differentiation and proliferation. However, when mutated, proto-oncogene turned oncogene no longer responds to growth restrictions. This allows the cell to multiply at an unnatural rate [28]. P53, a tumor-suppressing gene, has been found to maintain the microenvironment of the tissue suppressing cancer growth. However, when this gene and other tumor-suppressing genes mutate, they can no longer control the rate at which cancer cells multiply. This causes the formation of tumors. An advantage is given to the now cancerous cells allowing them to multiply before the immune system can dispose of them [5]. Finally, mutations will occur in DNA repair genes such as PRKDC. This gene encodes DNA-dependent protein kinase which plays a role in DNA double-strand break repair and recombination [8]. When damaged DNA is unable to repair itself, it has the potential to multiply unimpeded. Cancer cells differ from normal cells in three distinct ways. Cancer cells grow uncontrollably, do not die like normal cells, and are able to spread throughout the body. The reason cancer cells are so dangerous is they grow and divide uncontrollably taking resources and blood flow from healthy cells and tissues.

Bladder Cancer

Bladder cancer is a common cancer being the tenth most diagnosed in 2020 with approximately 570,000 new cases [**source1?**]. The bladder is an organ located in the lower region of the abdominal cavity that is responsible for holding and disposing urine. The bladder wall is composed of four layers of tissue: lining epithelium, lamina propria, muscularis propria, and the serosa/adventitia. The innermost layer is the lining epithelium which is made up of specialized cells called the urothelium [3]. These cells change from being five to seven layers thick in an empty bladder to two to three layers in a distended bladder [3]. The layering allows the bladder to modify its shape without being damaged. The lamina propria is a thin layer of connective tissue that surrounds the urothelium and contains nerve endings, smooth muscle, and blood vessels [3]. This layer acts as a barrier between the urothelium and the muscle. It also acts as a communication hub between

the bladder and the nervous system. The muscularis propria is also known as the detrusor muscle. The sympathetic nervous system relaxes this muscle until the bladder is full. Then the parasympathetic nervous system contracts the detrusor muscle during urination [3]. The most outer layer of the bladder is the serosa and adventitia. Both are connective tissues. The serosa covers the dome of the bladder, and the adventitia covers the remaining areas of the bladder [3].

There are three main types of bladder cancer: urothelial carcinomas, squamous cell carcinomas, and adenocarcinomas. The most common is urothelial cell carcinomas which account for approximately 90% of all bladder cancer cases [10]. These tumors occur in the urothelium tissue layer in the bladder [10]. Subsequently, squamous cell carcinomas make up five percent of bladder cancer occurrences [10]. This type also occurs in the urothelium layer; however, as the name suggests only squamous cells are affected. Squamous cell carcinomas are often related to bladder issues that irritate such as the parasitic infection, schistosomiasis [14]. Ultimately, adenocarcinomas account for one percent of bladder cancer cases. These tumors also originate in the urothelium but affect the glands of the bladder [6].

The most common symptom of bladder cancer is hematuria or blood in the urine [26]. Hematuria, which is the presence of blood in the urine, may occur in bladder cancer patients but it can vary in severity and frequency, ranging from microscopic to gross. Not every instance of urination for a bladder cancer patient results in hematuria, and it may only occur a few times. Approximately 85% of bladder cancer patients will experience hematuria at least once [26]. It is caused by a rupture of tumors in the bladder. There are other conditions such as urinary tract infections that cause hematuria, meaning blood in the urine is not a sure sign of bladder cancer. Other symptoms include painful and/or frequent urination, an intermittent stream, and abdominal or pelvic pain [25]. Tumors in the bladder may obstruct flow or cause the wall to thicken decreasing bladder capacity.

There are specific genetic risk factors that increase the chance of developing bladder cancer. One of them is that it disproportionately affects males over females. In 2022, there are expected to be about 61,700 cases in males versus only 19,480 cases in females [28]. One possible explanation for this disparity comes from the liver. This organ is responsible for metabolizing bladder carcinogens which are then excreted through the bladder. This means that an inconsistency in the metabolism due to a variant of UGTs and SULTs means the bladder is exposed to different levels of carcinogens in males

versus females [31]. There is also evidence to support that an individual's nationality affects the risk of developing bladder cancer. More developed countries such as Europe and North America have higher cases of bladder cancer versus less developed countries such as Central America, Asia, and Africa [9]. Individuals in developed countries have greater exposure to carcinogens both in the workplace and at home. Even in the same country, race affects the risk of developing bladder cancer. When looking at bladder cancer rates in the United States, non-Hispanic whites had the highest incidence rate at 23.09 incidences per 100,000 people compared to Asian and Pacific Islanders who only had 9.51 incidences per 100,000 people [30]. Again, this distinction can be explained by increased exposure to carcinogens in the non-Hispanic white population. The perhaps greatest risk factor for developing bladder cancer is age. The median age for bladder cancer diagnosis is 70 years old [23]. The simple explanation for this occurrence is the older population has more exposure over a longer period to carcinogens. While this is true, further research has shown there may be more complicated reasoning. As individuals age, gene expression changes. Some proto-oncogenes may have an increase in expression causing a build-up of cells to form [22]. In other cases tumor suppressor genes may have a decrease in expression, meaning the body is unable to remove the build-up of cells as efficiently. Finally, it is more difficult for an older cell to repair damaged DNA [22]. This is because as cells divide over time, their telomeres, or the protective caps on the ends of chromosomes, become shorter, until they can no longer protect the DNA during replication. Additionally, the accumulation of other types of damage, such as oxidative damage from free radicals, can also interfere with the cell's ability to repair itself.

In addition, there are environmental factors that increase the risk of developing bladder cancer. The biggest environmental risk factor is smoking. All smoking increases the risk of developing bladder cancer. However, the risk is higher in individuals who smoke cigarettes over cigar smokers or pipe smokers [24]. The microbiome of the bladder is damaged in individuals who smoke. It has been found that smokers lack bacteria in their stomachs such as Bacteroidaceae, Erysipelotrichales, Lachnospiraceae, and Bacteroides [13]. This correlation suggests that a change in the microbiome can cause bladder cancer. Additonally, the bladder of smokers is exposed to the carcinogenic chemicals present in cigarettes. The chemicals enter the bloodstream and are filtered by the kidneys, which concentrate them in the urine stored in the bladder. There are also occupational hazards that can cause bladder cancer.

One is the exposure to aromatic amines which are commonly used in various industries and agriculture [11]. Another hazard is exposure to polycyclic aromatic hydrocarbons which can be found in gasoline and oil. Other environmental risk factors include high levels of arsenic in drinking water and urinary tract infections [11].

DEGs and Enriched Pathways

A differentially expressed gene (DEG) is a gene that is expressed differently two or more groups. In molecular biology and genetics, the term "expression" refers to the process by which the information encoded in a gene is used to produce a functional protein or RNA molecule [21]. DEGs are important in cancer research because they can help identify genes that are involved in cancer development or progression, and can potentially serve as biomarkers or therapeutic targets.

An enriched pathway is a biological pathway that contains a statistically significant number of DEGs compared to what would be expected by chance. These pathways are identified by analyzing gene expression data to determine which pathways are overrepresented among the DEGs. The identification of enriched pathways can help researchers to understand the underlying biological processes that are altered in a particular disease or condition.

Current State of the Art

DEG-PI is a bioinformatics tool that aims to identify differentially expressed genes (DEGs) and enriched pathways using gene expression data from the Geo database. The database provides information on gene expression levels in cancer tissue and normal tissue. DEG-PI compares the gene expression levels between the two groups to identify genes that are differentially expressed. The identified DEGs are then analyzed using a pathway analyzer to find enriched pathways. Finally, the results are displayed in an easy-to-use interface for the user to view.

Related work

Cancer Research

Due to the vastness of the cancer research industry, the projects can be organized into different focuses. Cancer types vary significantly when looking at the cause, how it spreads, and the severity. In view of this, cancer research projects are usually based on one specific type. More complicated research is centered around subtypes. For example, researchers might be investigating HER2-positive breast cancer or small cell lung cancer in particular, rather than only examining breast cancer or lung cancer in general [19]. This is because cancers can vary greatly even within the same type, and subtyping can help identify differences in the way the disease develops and progresses, as well as differences in treatment options and outcomes. By studying subtypes, researchers can gain a better understanding of the underlying biology and genetics of cancer to develop more targeted and personalized treatments.

The understanding of the fundamental aspects of cancer is well-established. Hence, cancer research has shifted towards more specialized studies. With the knowledge that cancer is a complex disease with multiple subtypes, researchers have begun to focus on studying specific types of cancer to better understand the underlying mechanisms of the disease and develop targeted treatments. There are also two main camps of research: prevention and treatment [29]. Cancer is a progressive disease meaning as time goes on, cancer and prognosis get worse. Therefore, a branch of cancer research focuses on ways to diagnose cancer earlier. This includes both trying to find causes of cancer and biomarkers that mark someone as being predisposed to developing cancer. For example, individuals with the BRCA1 or BRCA2 gene mutation have a higher risk of developing breast and ovarian cancer compared to individuals without the mutation [19]. BRCA genes normally produce proteins that help suppress abnormal cell growth

[17]. Mutations in these genes can lead to a loss of this tumor-suppressing function and increase the likelihood of cancer development.

The other branch of research focuses on finding treatments once an individual develops cancer. There is currently research being done on the development of immuno-oncology drugs. Immuno-oncology is the study of using the immune system to fight cancer cells. A study looked to build guidlines about immuno-oncology drugs in response to found incidences of pseudoprogression and hyperprogression. They acknowledge a lack of data sharing a problem for this field. [20] There is also research being done on the development of precision oncology drugs. Precision oncology is the study of using an individual's cancer cells to create a tumor-specific treatment plan. However, patients need access to advanced diagnostic tools in order to receive precision oncology, and it is very specific. A perspective study attempted to address some of these issues. It highlighted the importance of equal access to genomics tests, robust evidence for new drugs and technologies. They stress that a multi-stakeholder approach is necessary for the successful implementation of precision oncology to benefit patients with cancer worldwide [15].

Databases

Due to the extensiveness of cancer data and research, tools and databases have been and continue to be developed to create a better understanding of cancer and how to treat it. There are currently free public databases with cancer information that give researchers access to cancer data. One of the largest being The Cancer Genome Atlas (TCGA). This database has "over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data." The data spreads across 33 different types of cancer and looks at specific cancer information from over 20,000 primary cancers. A sample of this database was first released in 2006, only containing data related to lung, ovarian, and glioblastoma cancer types. It was such a success that in 2009 it was approved to collect data on even more cancer types. This database is currently used in prognostic research of different cancer types. Oftentimes data from TCGA is used in conjunction with another database: the Gene Expression Omnibus (GEO). This database contains information about gene expression and RNA methylation. It also has visualization and analysis tools to help a researcher analyze the data. One study specifically looked at a model that predicted

the prognosis in bladder cancer patients. A study published in November of 2022 looked at cohorts from TCGA. After running a pathway enrichment analysis, it found seven genes that were used to give each sample a score. The higher scores correlated to a worse prognosis. The GEO database was used to verify that a different cohort produced similar results using the model in the study. A different study used the two databases to identify prognostic factors in hepatocellular carcinomas. The researchers looked at 424 samples from TCGA and three expression profiles from GEO. Using this data, they found two prognostic genes that suggested a patient would survive longer if these genes were active. Both TCGA and GEO are useful in cancer research to provide data to researchers all over the world.

Tools

In addition to the databases that contain genetic information about cancer, there are computational tools that have been created to help analyze the data. On tool is called the WEB-based Gene Set AnLysis Toolkit (WebGestalt). This tool contains information about 12 organisms, 342 genes, and 155,000 functional categories [26]. It has three different analysis features: over-representation analysis (ORA), gene set enrichment analysis (GSEA), and network topology-based analysis (NTA). ORA is used to identify pathways that are over-represented in a set of genes. It compares the input gene set to a reference gene set and calculates the statistical significance between the two. GSEA takes in a gene list and outputs pathways that are identified more than would be expected by chance. A study on lung adenocarcinoma used Webgestalt to identify hub genes. A hub gene is a gene that has more connections in a gene interaction network. In this study, researchers collected genes related to lung adenocarcinomas from three different databases. They created a gene list out of the common genes and used the Webgestalt ORA function to help identify the hub genes. Using this tool, the researchers were able to identify three hub genes related to lung adenocarcinomas. In a different study, researchers examined Topoisomerase 2-alpha (TOP2A) and its role in papillary thyroid cancer (PTC). They also looked at common pathways associated with PTC. Over 20,000 co-expressed related to TOP2A were found using a database. Then a pathway analysis was performed on these genes. The five most common pathways were identified and three of them were positively associated with the TOP2A expression. Overall, Webgestalt,

and other bioinformatic tools allow cancer researchers to analyze genes in a way that would not be possible otherwise.

Another bioinformatic tool is STRING. This program is used to create protein-protein interaction networks and functional enrichment analysis. It contains data from 14094 organisms and over 67.6 million proteins. It is an interactive database that predicts functional associations between genes using comparative genomics methods, co-expression patterns, and previous knowledge mined from PubMed literature abstracts. It can help identify genes involved in the same pathway or function as the query gene, and its functional association data is freely available. In a select study, STRING was used to show significant deregulation of CDKs in breast cancer with high expression of CDK1 in metastatic tumors. This allowed the researchers to suggest targeting CDKs may be a promising approach for breast cancer treatment [16]. Additionally, STRING was used to generate the HAMP-binding protein network that allowed the researchers to conclude HAMP gene expression is related to the infiltration levels of immune cells, and that HAMP affects the survival rate in gastric cancer through the immune pathway.

Method of approach

Although there are existing tools available, researchers lacking coding skills or a computer science background often find them challenging to use and comprehend. Usually, data needs to be organized in a specific way, or the results are displayed without an explanation of what the graphs or results mean. Before researchers can even start analyzing their data, they have to teach themselves a new skill. The aim of this tool is to integrate the scientific capabilities of existing tools and algorithms with a user-friendly interface that presents all the data in a visually accessible manner.

DEG-PI will retrieve a dataset from the Gene Expression Omnibus (GEO) data repository. The data comes from real cancer samples in the form of high-throughput gene expression. PICTURE The data is displayed as a list of genes with their corresponding intensity, which is measured by looking at mRNA quantity. The expression of cancer tissue is evaluated against that of normal tissue. The mean expression value for each gene in each category is determined and compared. If a statistically significant difference is observed, the gene is considered differentially expressed and can be either over-expressed or under-expressed. The differently expressed genes (DEGs) are then run through a pathway analysis. A pathway is a series of interactions and reactions that allow for regulation of cellular processes. Each step involves at least one distinct gene that gives rise to a specific protein, enzyme, or other biological component. A pathway analysis looks to identify known biological pathways that are overrepresented by a gene list. These pathways are said to be "enriched".

Both Python and R were used to create DEG-PI. Python is a high-level, interpreted programming language that is widely used for web development, data analysis, artificial intelligence, and scientific computing. It was chosen for its versatility and ability to create a user interface. An aim of this project was to make a tool as user-friendly as possible, especially for those without

a computer science background. This means it is important to keep the code hidden behind a user-friendly interface. Python's library Streamlit is especially useful in making a dashboard to get input from the user and display results. Streamlit is an open-source Python library. It allows developers to build web applications with interactive user interfaces for machine learning and data science projects. However, it does not possess the data analysis libraries that are available in R.

R is a programming language and open-source software that is especially useful for statistical analysis and graphics. R was selected for its statistical libraries that are specific to biological research. Its used in various fields for data analysis, visualization, and modeling. R also has a vast library of packages. These allow the users to perform an advanced statistical analysis. Using these libraries, researchers can import large amounts of biological data such as gene expression, protein interaction, or sequencing data and then manipulate it, test it, or display it.

Some useful R libraries that were required for this project include: - Bioconductor: A project that store and publishes over 2,500 free open source software libraries related to biological analysis. - DESeq2: A library for differential gene expression analysis of RNA-sequencing data. It is useful for comparing gene expression levels by using a negative binomial distribution. - ggplot2: A library for creating graphs and charts to help visualize data. - GEOquery: A library the Bioconductor software that is used to simply connect R and GEO. - WebgestaltR: A library for functional enrichment analysis that identifies enriched pathways and gene ontology terms

Python is the main driver of this program. It utilizes the subprocess library to call different R programs for conducting statistical analyses. This library creates new processes meaning there can be multiple programs running at once in separate languages. This allows DEG-PI to run an R program while a main Python program is already running. The languages pass variables and data amongst each other by using text files or comma-separated value (CSV) files. The files are all saved in the same directory as the program. This allows R to access the parameters for data collection and analysis and Python to get the data and results.

Algorithm

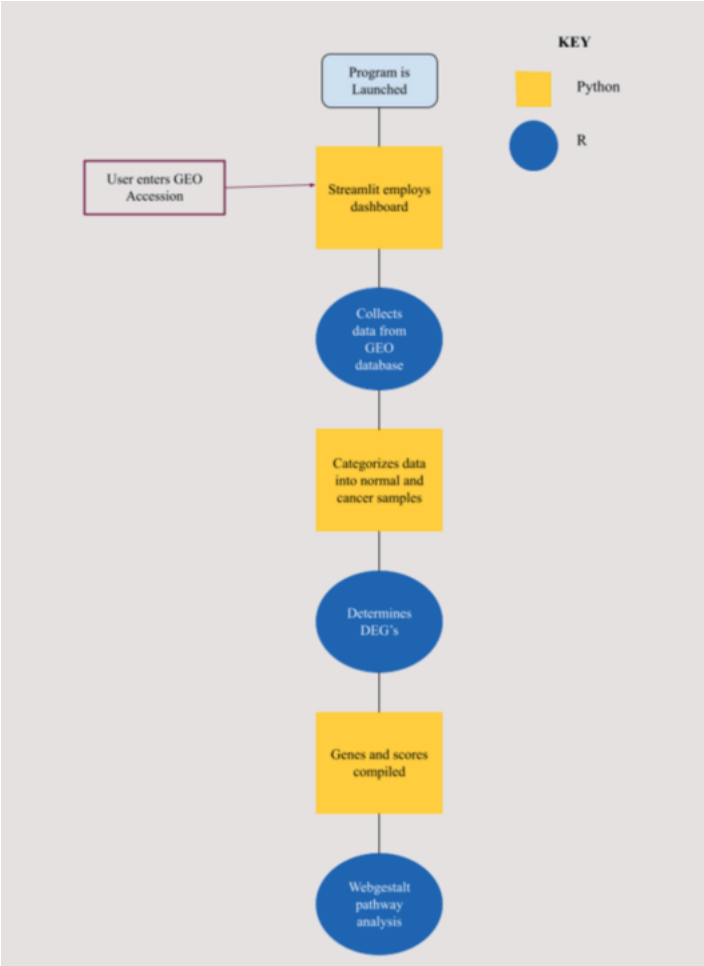

Figure 1: The algortihm shows how the program goes back and forth between Python and R to create a final project. The yellow square is displays what was completed in Python and the blue diamond is what was completed in R.

When the program is run, Python uses Streamlit to build and deploy the dashboard. The dashboard has a home page explaining the necessary libraries and dependencies needed to successfully analyze the dataset. Also, on the home page, the user is prompted to enter a GSE accession number. The GSE accession number corresponds to the data from the GEO database.

This accession number is written to a text file in Python to be read by R.

The R library GEOquery is used to access the GEO database. Because the database is not exclusive to cancer data, accurate results may not be obtained from all datasets. To ensure accurate results, it is crucial for the dataset to only consist of normal and cancer tissue samples, which are separated into two categories using a "normal" or "control" keyword. The normal sample must be labeled. If a sample is unlabeled, the program assumes it is a cancer sample. R reads the GSE accession number from the file and uses the getGEO function to collect the data which is saved into an array. The phenotypic data from this array is written into a CSV file.

Python opens the CSV file containing the high throughput gene expression data. To compare the two groups and find the differentially expressed genes, the program uses the sample names to assign them to groups. Group 0 represents normal samples, while group 1 represents cancerous samples. Python creates a new CSV file containing the title, the GEO accession number, and the group number of each sample. Python also adds the group number to a string. The string is going to be passed to R so the genes in the group types can be compared. The order of the string will match the order of the samples in R, so each sample will be properly assigned a group. After all of the samples have been sorted, Python will write this string into a text file. It will then use the subprocess library to call a second R program.

Because this is a new R program file, it does not have the raw data from GEO. It must reread the data from the database using the GSE accession number similarly to the first R program. The program then transforms the raw data into usable data by removing samples that did not belong to either group and performing a log transformation. By taking the log of the data, the range of values is compressed and outliers are less emphasized making it easier to visualize the data and identify patterns. The samples are then grouped together and put into a matrix to be analyzed. A table containing the top 20,000 genes that are differentially expressed between the two groups is generated and saved into a CSV file.

Python then accesses the CSV file containing information about the genes identified in the samples, as well as their statistical significance and degree of change. It sorts through the samples to identify genes that would be considered differentially expressed. In order to determine this, the program looks at the adjusted p-value and the logFC value. The p-value represents the probability of observing the observed difference in the samples even if there was no actual difference between them. The adjustment takes into account

multiple testing or false positives, and it is generally accepted to be more accurate. The logFC value or Log2-fold change, determines whether a gene is upregulated or downregulated and the respective magnitude. The python program looks to identify genes has a p-value lower than 0.05 and a logFC value greater than 0.5 and less than -0.5. This would make a gene both statistically significant and practically significant. Both of these thresholds are standard (SOURCE) as well as the manipulation of the fold change value (SOURCE). Also, this sorting of data excludes samples that do not have a gene name as the name is needed to perform the pathway analysis. The program writes the information about the genes that pass both tests into a new file. It also adds a score. The score is calculated by taking the negative log of the p-value. This is used to rank the genes by how differentially expressed they are. A gene with a smaller p-value is more statistically significant. This means the chance that the difference is real is high. By taking the negative log, it has a higher score and is ranked higher on the list.

The R library, WebGestaltR, is used for the pathway analysis. This is based on the WebGetstalt online tool. This tool, more formally known as, WEB-based GEne SeT AnaLysis Toolkit, " is designed for functional genomic, proteomic and large-scale genetic studies from which large number of gene lists are continuously generated" (SOURCE). WebGestaltR performs similar functions to the online version. Using the file from Python, the library performs the pathway analysis. It performs enrichment analysis by comparing the input list to various functional annotation databases, such as Gene Ontology (GO) and KEGG pathways, to identify the biological functions and pathways that are overrepresented in the input list.

Python finally gets back the data from the Webgestalt pathway analysis. It uses Streamlit features to display the data in an easy-to-understand format making sure the user does not have to know or understand any code to get results. It also provides explanation for those results.

Experiments

The tool successfully identifies differentially expressed genes and enriched pathways. The DEG's are displayed in a table and a volcano plot shows the distribution of genes that were determined to be differentially expressed and the genes that were not. The tool also shows the results of the Gene Ontology (GO) Slim summary which is also preformed my WebgestaltR. It displays three bar graphs showing information about the biological processes, the cellular component, and the molecular function of the affected genes. They display the number of genes that fit into each subcategory. The tool show the most enriched pathways as well as further information about the pathways. These include the pathway ID and name, enrichment score, p-value, enrichment plot, and the DEGs in that pathway. Finally, the tool has an additional page that explains the meaning behind all of the variables and graphs that the tool shows.

Results

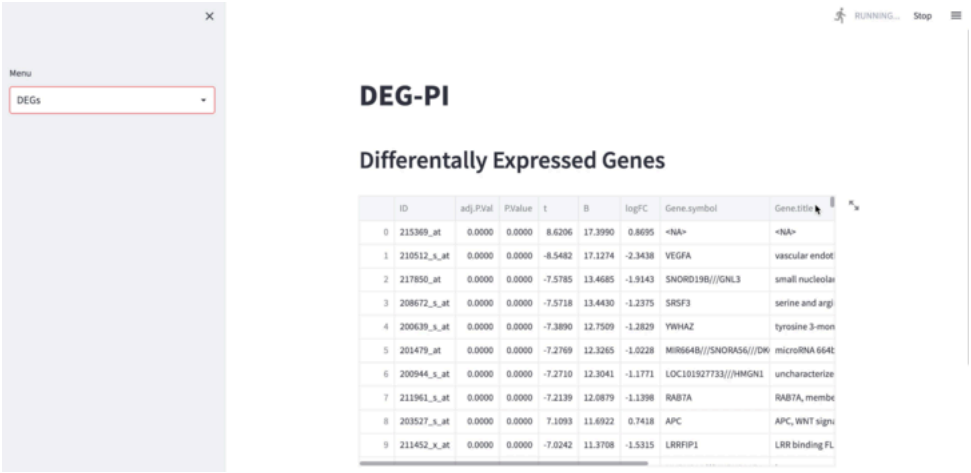

Figure 2: A table containing the Differntially Expressed Genes.

The tool produces a table listing the DEGs with additional information. It has the gene ID, adjusted p-value, p-value, t, B, log FC, gene symbol, and gene title.

The gene ID is the identification number that comes from the raw data [1].

The adjusted p-value is the p-value after adjustment for false positives. This is generally accepted to be more accurate than the normal p-value [1].

The p-value is the raw probability that the difference occurred simply due to chance [1].

t is the moderated T-statistic. This is similar to the normal t statistic except it takes into account the gene expression variability [1].

B is the B-statistic or log-odds that the gene is differentially expressed [1].

logFC is the Log2-fold change. The fold change determines whether a gene is upregulated or downregulated and the respective magnitude [1].

Gene symbol is a short abbreviated version of the gene title [1].

Gene title is the full name of the gene [1].

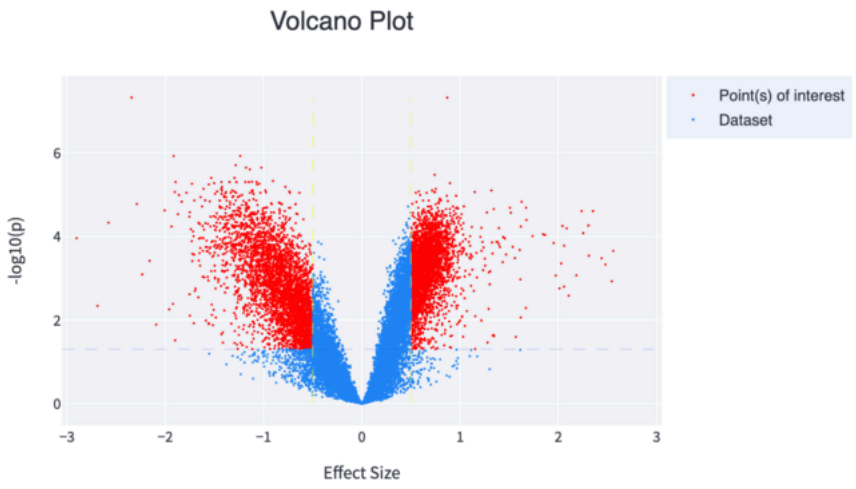

Figure 3: The tool also produces a volcano plot. A volcano plot is a type of scatter plot that analyzes statistical significance (x-axis) versus magnitude of change (y-axis).

The upregulated genes have a positive magnitude of change, and the more upregulated ones are farther away from 0. Similarly, the downregulated genes have a negative magnitude of change, and the more downregulated are farther away from 0. The most statistically significant genes are towards the top of the graph.The red color signifies the genes that are both statistically significant (adjusted p-value < 0.5) and have a fold value between negative 0.5 and 0.5. These are the DEGs. The blue represents the genes that do not meet the criteria and thus are not considered a DEG.

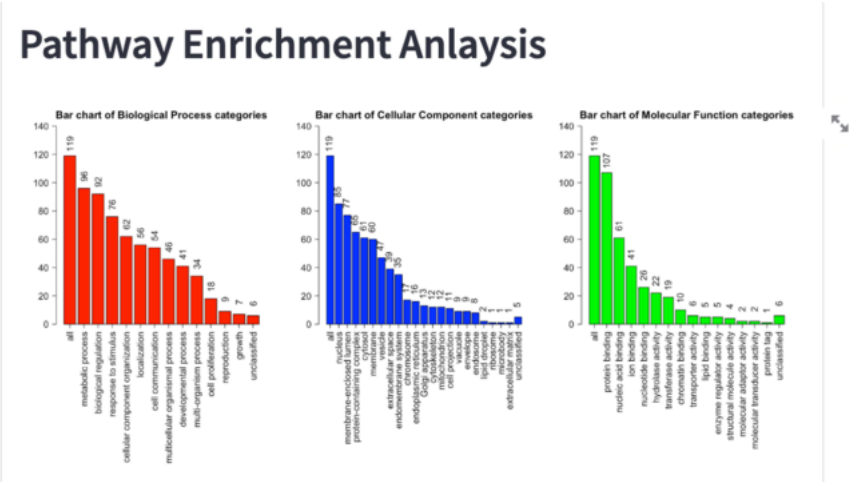

Figure 4: The tool displays the results of gene ontology analysis.

Gene ontology is the a database that collects and stores data related to gene function and attempts to categorize them into three main categories: biological processes, the cellular component, and the molecular function. Each category is divided into subcategories. The number of genes that fall in each subcategory are tallied and displayed in a bar graph [4].

	geneSet	description	normalizedEnrichmentScore	pValue
10	hsa04080	Neuroactive ligand-receptor interaction	2.2562	0.0000
11	hsa04614	Renin-angiotensin system	2.0454	0.0017
12	hsa04060	Cytokine-cytokine receptor interaction	1.9741	0.0038
13	hsa04072	Phospholipase D signaling pathway	2.1194	0.0000
14	hsa00982	Drug metabolism	1.9832	0.0016
15	hsa04610	Complement and coagulation cascades	2.0462	0.0015
16	hsa05225	Hepatocellular carcinoma	1.9296	0.0055
17	hsa00350	Tyrosine metabolism	1.9039	0.0033
18	hsa04740	Olfactory transduction	1.8802	0.0072
19	hsa04510	Focal adhesion	1.8143	0.0078

Figure 5: A table with the 20 most enriched pathways is shown as a result of the pathway analysis.

The table contains the gene set, description, normalized enrichment score, and the p-value:

The gene set is the KEGG identifcation number for the enriched pathway.

The description explains what the pathway does.

The normalized enrichment score describes how enriched the pathway is. It is calculated by the sum of the scores of all DEGs that are in that pathway. A higher enrichment score corresponds to better enrichment. This does not necessarily mean it has the most genes. An enrichment score can be higher if it has fewer higher scoring genes compared to a pathway that may have more lower scored genes [27].

The p-value is the raw probability that the pathway was enriched simply due to chance.

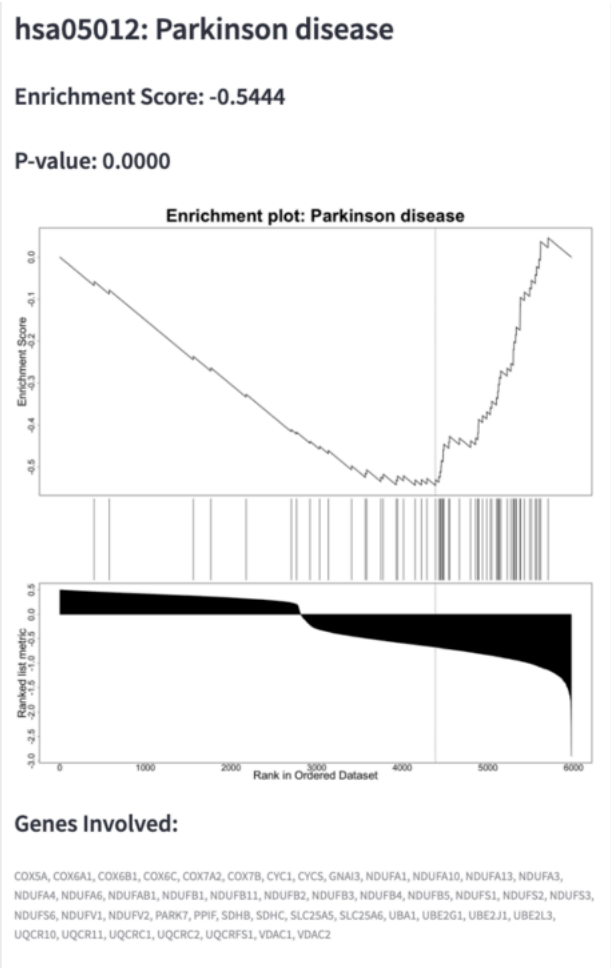

Figure 6: The algortihm shows how the program goes back and forth between Python and R to create a final project. The yellow square is displays what was completed in Python and the blue diamond is what was completed in R.

It has the gene set and description, the enrichment score, the p-value, and the enrichment plot

The enrichment plot has three main sections. The top section shows the cumulative enrichment score as the pathway moves down the ranked list. The genes with a higher score are on the left and the genes with the lower score are on the right. The middle portion contains ticks when a gene from the ranked list is found in the pathway. The bottom portion contains the the value of the ranking metric as you move down the list of ranked genes.

Experimental Design and Evaluation

The algorithm for the tool was based on a study that identified hub genes in bladder cancer, Identification of hub genes and pathways in bladder cancer using bioinformatics analysis. Therefore, comparing results from the tool to the results from the study helped to show the that my tool was successful. This study looked at the dataset "GSE3167" from the GEO database. I was able to identify the same DEG's as the study as we both used the geo2R library to get our results and the same data. The study then used DAVID to complete a functional enrichment analysis or a pathway analysis. The tool used WebgestaltR. Both are commonly used to identify enriched pathways. While the pathways were not exactly the same, common pathways were identified. Both the study and the tool identified the hsa03040: Spliceosome, hsa05012: Parkinson's disease, hsa00830: Retinol metabolism, and hsa05200: Pathways in cancer. Despite using different enrichment pathway analyzers, both were able to identify common pathways. This makes sense because the DEG's were almost identical, so even with different algorithms the pathways should be similar. The tool also identified hsa05033:Nicotime addiction as an enriched pathway. This makes sense because smoking is one of the most common causes of bladder cancer. Even if this pathway does not directly cause bladder cancer, but rather this pathway causes the individual to start smoking. It shows my tool found a relationship between the DEGs in bladder cancer that is already known to exist.

Threats to Validity

The biggest threat to the validity of this tool is the fact that not all of the genes that are determined to be differentially expressed are included in the pathway analysis. WebgestaltR needs the gene name in order to complete the analysis. The gene ID was identified to be differentially expressed, but the ID is not matched to a gene. For example, the gene with the ID of 215369_at is thought to be a "hypothetical gene supported by AK023162". This cannot be mapped by Webgetstalt, so it is excluded. Knowing there are genes that have been identified but have not been included means that our results will never be 100% accurate. Despite the fact that the pathway analysis is incomplete, this tool is still functional. Scientists are still studying genes and their interactions, and they are still adding this information to public databases.

Therefore, as information about the unknown genes become available the program will become more accurate. There is currently enough information available to constitute conducting cancer research including this tool.

Conclusion

DEGs can give researchers answers about the biological mechanisms and genes in cancer development. DEGs in cancer cells can be either upregulated or downregulated. Each provides clues about the mechanisms and pathways involved in cancer growth. A gene that is consistently differentially expressed suggests the gene may be a result of cancer-causing mutations or a consequence of a mutation in its pathway. This gives researchers a place to start to better understand the molecular mechanisms behind cancer subtypes. Understanding how a gene functions in the cell, helps researchers identify biomarkers. Biomarkers provide a more distinct starting point that may help cancer diagnosis and prognosis. The more we understand about cancer subtypes, the better treatment and prevention can be.

Additionally, DEG identification may help find targets for drug therapy by exposing genes that are consistently expressed differently in cancer tissue. They help to understand the pathway and mechanism behind cancer development. Researchers can then create drugs that stop the unnatural cell behavior in cancer cells. For example, if a gene is found to be upregulated, a drug that stops its production may be useful to stop or slow cancer progression. On the other hand, if a gene is found to be downregulated, a drug that increases its prodcution may be most useful.

Studying enriched pathways has similar benefits to studying DEGs. By understanding which pathways are more targeted, researchers can better understand the molecular mechanisms behind the cancer, and the relationship between different genes. If a pathway, is commonly identified as enriched then that means the pathway has multiple genes that are differentially expressed (either upregulated or downregulated) in the cancer tissue. Knowing the pathways that are expressed differently allow the researcher to attempt to draw connections and a picture of how specfic cancer types develop and the mechanisms behind them. Enriched pathways can also help in identifying

biomarkers, which are molecules that acts as a signal for a certain disease or condition. Therefore, if a biomarker is detected, it is a sign the individual could have cancer. This makes earlier diagnosis easier, as you develop tests that check for one molecule.

Additionally, enriched pathways can help to identify visible phenotypes that may indicate a higher probability for an individual to develop certain cancer. For example, if the Alzheimer's pathway is found to be enriched in bladder cancer, then it could be assumed that individuals with Alzheimer's may have an increased risk of developing bladder cancer even if we do not understand why. This means that individuals with Alzheimer's, which is a phenotypically visible characteristic, can be more closely monitored for bladder cancer. This helps early detection and early treatment of certain cancers if phenotypically visible pathways can be identified.

Future Work

As both cancer research and computer technology continue to advance, more complex tools are being created to analyze more complex data. In comparison, DEG-PI is very basic compared to some of the other biology tools that exist currently. This means the tool has a lot of room for improvement:

- Incorporating the DAVID library: DAVID, or the Database for Annotation, Visualization, and Integrated Discovery, is another tool that looks to analyze large lists of genes. Similarly to Webgestalt, DAVID is able to perform pathway analysis. However, the R library that was used to connect DAVID to R has been discontinued. The results of the DAVID pathway analysis could be compared to the Webgestalt pathway analysis since both use slightly different algorithms.

- Allowing the user to group the data: Currently, the tool only compares normal tissue to cancer tissue. Further work could include allowing the user to select the groups. Allowing the user to select the groups, means more than two distinct samples could be analyzed. It could also mean that the user could compare cancer types to each other. Simply comparing cancer tissue to normal tissue is only one small facet of cancer research. There is great variability both among and between cancer types, so it is also beneficial to compare cancer types. The only cancer-specific code in the algorithm is in the sorting of the data into

the groups. Therefore, if the user is able to select the groups, this tool could be used to analyze any sort of data the user wanted to compare against another.

- Other databases, TCGA, upload data: One of the compilations of this tool is it gets data from a database that is not cancer specific. Therefore, not all of the datasets are applicable to the tool analysis. The Cancer Genome Atlas (TCGA) is a database that contains "primary cancer and matched normal samples" for 33 different cancer types. If the data could be formatted in the same way the data from the GEO is, the rest of the algorithm would not need to be changed.

- Way to save data: Currently, the tool only compares one set of cancer tissues to one set of normal tissues. However, adding a save feature would allow researchers to compare the pathways distinguished as enriched in two different trials. This would help to facilitate more comprehensive and meaningful analysis by allowing researchers to compare results from multiple studies and potentially uncover common pathways and mechanisms involved in cancer development and progression.

Future Ethical Implications and Recommendations

Cancer research in general is a sensitive topic because of how many people have been affected. It is important for researchers to collect the data ethically and in a standardized manner. This means that the methods used to gather data should be transparent, reliable, and respectful of the individual providing his or her data. Without standardized techniques, comparing different samples that are collected in unstandardized ways will result in incorrect conclusions. These conclusions could derail future research wasting resources and time. It is also important to remember the data came from real humans who really suffered. Therefore, handling this data with sensitivity and ethical considerations is essential. The data the tool accesses has very little personal information connected to it. Therefore, it becomes easy to forget where the data really came from and what people had to go through as a result of this disease. Sharing results is a way to show respect to the individuals who were willing to share their data with the public. It shows

the collective effort towards ending the crisis that is cancer.

Another suggestion is to avoid over-analyzing the results. The purpose of the tool is to identify differentially expressed genes and enhanced pathways between two sets of data. Currently, the tool analyzes gene expression from normal tissues and cancer tissues. There currently is no standardized way to identify differentially expressed genes and enhanced pathways because there is no perfect algorithm. Different libraries will yield different results. In other words, the identified genes and pathways through the use of this tool may not provide a definite diagnosis. The tool is intended to be used primarily for research purposes to aid in understanding the molecular mechanisms of cancer. It is not a diagnostic tool, meaning that it cannot provide a definitive diagnosis or treatment plan for patients with cancer. It is important to keep this in mind when interpreting the results generated by the tool.

Final Thought

Cancer is a pervasive disease that affects people from over the world. It affects all countries, races, genders, and ages. One in three individuals is expected to develop some type of cancer during their. Therefore, almost everyone, if everyone has been affected by cancer in some way. Despite all the research and resources that have been poured into curing cancer, it remains a common issue across the world. Cancer cases have even increased over the last couple of years as individuals are being exposed to more toxins and carcinogens. Also, as medicine continues to increase life expectancy, cancer cases will continue to increase. Cancer affects older individuals more often than young individuals. The more cancer there is, the more data we have to try to understand it. This includes information specific to the cancer, the individual, and the outcome. Analyzing this data will provide new ways to treat and prevent it. However, cancer continues to be complex as each cancer is unique to the individual. There is no "one-size-fits-all" cause of cancer, so there is not going to be a "one-size-fits-all" solution. Mutations can occur in different kinds of genes including oncogenes, tumor suppressor genes, and DNA repair genes. Understanding cancer with the intention of curing it is going to take experts from all over the world in different fields to collaborate and share results. This makes cancer research a challenging but critical area of study that has the potential to impact the lives of millions of people around the world.

References

[1] GEO2R: Interactive analysis of gene expression data in GEO.

[2] NI Berlin, TA Waldmann, and SM Weissman. 1959. Life span of red blood cell. *Physiological Reviews* 39, 3 (1959), 577–616.

[3] Jetti R. Bolla SR Odeluga N. 2020. Histology, bladder. *StatPearls [Internet]* (2020).

[4] Gene Ontology Consortium. 2004. The gene ontology (GO) database and informatics resource. *Nucleic acids research* 32, suppl_1 (2004), D258–D261.

[5] Rodier F Coppé JP Patil CK. 2008. Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. *PLoS Biol* 6, 12 (2008), 2853–2868.

[6] Guo CC Dadhania V Czerniak B. 2015. Adenocarcinoma of the urinary bladder. *Am J Clin Exp Urol* 3, 2 (2015), 51–63.

[7] Khaniani Goldar S. 2008. Molecular mechanisms of apoptosis and roles in cancer development and treatment. *Asian Pac J Cancer Prev* 16, 6 (2008), 2129–2144.

[8] Robinson DR Grasso CS Wu YM. 2020. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* 487, 7406 (2020), 239–243.

[9] Prasad SM Greiman AK Rosoff JS. 2018. Association of human development index with global bladder, kidney, prostate and testis cancer incidence and mortality. *BJU Int* 120, 6 (2018), 799–807.

[10] Feldman A Kaufman D Shipley W. 2009. Bladder cancer. *The Lancet* 374, 9685 (2009), 239–249.

[11] Prasad A Kiriluk K Sandip M. 2012. Bladder cancer risk from occupational and environmental exposures. *Urologic Oncology: Seminars and Original Investigations.* 30, 2 (2012), 199–211.

[12] Pouyssegur J Kroemer G. 2020. Tumor cell metabolism: Cancer's achilles' heel. *Cancer Cell* 13, 6 (2020), 472–482.

[13] Shen L Ma W Zhang W. 2021. Can smoking cause differences in urine microbiome in male patients with bladder cancer? *Front Oncol.* 11, (2021).

[14] Ahmed A Martin JW Carballido EM. 2016. Squamous cell carcinoma of the urinary bladder: Systematic review of clinical characteristics and therapeutic approaches. *Arab J Urol* 14, 3 (2016), 183–191.

[15] Joaquin Mateo, Lotte Steuten, Philippe Aftimos, Fabrice André, Mark Davies, Elena Garralda, Jan Geissler, Don Husereau, Iciar Martinez-Lopez, Nicola Normanno, and others. 2022. Delivering precision oncology to patients with cancer. *Nature Medicine* 28, 4 (2022), 658–665.

[16] Sofi Mehraj Umara. 2022. Expression pattern and prognostic significance of CDKs in breast cancer: An integrated bioinformatic study. *Cancer Biomarkers* 34, 4 (2022), 505–519.

[17] Conleth G Murphy and Mary Ellen Moynahan. 2010. BRCA gene structure and function in tumor suppression: A repair-centric perspective. *The Cancer Journal* 16, 1 (2010), 39–47.

[18] World Heath Organization. 2020. Estimated number of new cases in 2020, world, both sexes, all ages (excl. NMSC). *International Agency for Research on Cancer* (2020).

[19] Buckman R. 1997. What you really need to know about cancer a comprehensive guide for patients and thier families. (1997).

[20] Jorge Luis Ramon-Patino, Sabine Schmid, Sally Lau, Lesley Seymour, Pierre-Olivier Gaudreau, Janice Juan Ning Li, Penelope Ann Bradbury, and Emiliano Calvo. 2022. iRECIST and atypical patterns of response to immuno-oncology drugs. *Journal for Immunotherapy of Cancer* 10, 6 (2022).

[21] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 3 (2003), 368–375.

[22] Droller M Shahrok F Shariat M. 2019. Bladder cancer in the elderly. *Urologic Oncology: Seminars and Original Investigations* 27, 6 (2019), 653–667.

[23] Droller MJ Shariat SF Sfakianos JP. 2010. The effect of age and gender on bladder cancer: A critical review of the literature. *BJU Int* 105, 3 (2010), 300–308.

[24] Ksheersagar Sharma S. 2009. Diagnosis and treatment of bladder cancer. *American family physician* 80, 7 (2009), 717–723.

[25] Sharma P Sharma S Ksheersagar P. 2009. Diagnosis and treatment of bladder cancer. *Am Fam Physician* 80, 7 (2009), 717–723.

[26] Lokeshwar VB. Shirodkar SP. 2008. Bladder tumor markers: From hematuria to molecular diagnostics–where do we stand? *Expert Rev Anticancer Ther.* 8, 7 (2008), 1111–1123.

[27] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J. P. Mesirov. 2007. GSEA user guide.

[28] Cooper GM Torry DS. 1991. Proto-oncogenes in development and cancer. *Am J Reprod Immunol* 25, 3 (1991), 129–132.

[29] Harold Varmus. 2006. The new era in cancer research. *Science* 312, 5777 (2006), 1162–1165.

[30] Li Y Wang Y Chang Q. 2018. Racial differences in urinary bladder cancer in the united states. *Sci Rep* 8, 1 (2018), 12521.

[31] Zhang Y. 2013. Understanding the gender disparity in bladder cancer risk: The impact of sex hormones and liver on bladder susceptibility to carcinogens. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 31, 4 (2013), 287–304.