DEVELOPMENT OF VISUAL ANALYSIS INTERFACES FOR LARGE BIOLOGICAL DATA AND CHARACTERIZATION OF IMMUNOMODULATORY NONCODING RNA NETWORKS CANCER

A Thesis Submitted to

the Graduate School of Engineering and Sciences of

İzmir Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Molecular Biology and Genetics

by

Muhammet Emre KUŞ

July 2023

İZMİR

We approve the thesis of **Muhammet Emre KUŞ**

Examining Committee Members:

Prof. Dr. Bünyamin AKGÜL
Department of Molecular Biology and Genetics, Izmir Institute of Technology

---

Asst. Prof. Dr. Hüseyin Atakan Ekiz
Department of Molecular Biology and Genetics, Izmir Institute of Technology

---

Asst. Prof. Dr. Melis Kartal Yandım
Department of Basic Medical Sciences, İzmir University of Economics

23 July 2023

Asst. Prof. Dr. Hüseyin Atakan Ekiz
Department of Molecular Biology and Genetics, Izmir Institute of Technology

---

Prof. Dr. Özden Yalçın Özuysal
Head of the Department of Molecular Biology and Genetics

Prof. Dr. Mehtap EANES
Dean of the Graduate School

ACKNOWLEDGMENTS

I would like to sincerely thank my mentor Asst. Prof. Dr. Hüseyin Atakan Ekiz for his ceaseless support, guidance and for being a role model as an investigator. I share my thanks to my family and friends. They have always been there for me and supported me through the good and bad. I also would like to thank the members of Ekiz Lab for creating a family outside of the home.

I would like to express my endless gratitude to Prof. Dr. Bünyamin Akgül, our faculty dean, and to Assoc. Prof. Dr. Alper Arslanoğlu, one of our very valuable department professors, for always being by my side unconditionally and offering their support.

I would like to express my sincere gratitude to my thesis defense juries to Prof. Dr. Bünyamin Akgül, Asst. Prof. Dr. Melis Kartal Yandım, Asst. Prof. Dr. Atakan Ekiz, Asst. Prof. Dr. Ayşe Banu Demir, and Assoc. Prof. Dr. Alper Arslanoğlu.

Many thanks to The Scientific and Technological Research Council of Turkey (TUBITAK) who supported me with the 2210/A National MSc Scholarship.

I would like to express my biggest thanks to veteran Mustafa Kemal Atatürk, who paved the way for us young people to pursue scientific research by establishing the independent and modern Turkish Republic.

ABSTRACT

DEVELOPMENT OF VISUAL ANALYSIS INTERFACES FOR LARGE BIOLOGICAL DATA AND CHARACTERIZATION OF IMMUNOMODULATORY NONCODING RNA NETWORKS CANCER

These days we are collecting data in higher and higher dimensions, processing it, and developing tools that have strong descriptive and predictive powers. Especially in the field of cancer, the processing of data collected from patients has substantial potential in terms of discovering new biomarkers, developing personalized treatment methods, and better prognosticators. However, there are significant difficulties in utilizing and analyzing high-dimensional data. A good level of coding skills is required to bring the data together and apply different analysis methods. With the visual interfaces created in this study, we offer the opportunity to examine and analyze the high-dimensional data of thousands of cancer patients, which are open to the public through The Cancer Genome Atlas initiative, especially for bench scientists who has no prior coding expertise.

The Cancer Genome Explorer, shortly TCGEx, is a robust bioinformatic tool that we developed to facilitate high-throughput cancer data analysis through several sophisticated algorithms. With special features like subset-specific analysis and comparative analysis by using multiple cancer data, TCGEx can contribute to the literature by accelerating the studies, especially in hypothesis-driven research. This study also describes a use-case scenario that demonstrates how hypothesis-driven research can be performed using TCGExplorer for melanoma. In melanoma, elucidating the interactions between the tumor and the immune system at the miRNA level is crucial for developing new therapeutics. In this study, we characterize the properties of potential therapeutic targets that act on tumor and immune cells, which we have identified using various statistical analysis methods including machine learning, dimensionality reduction, and survival modeling using the TCGEx portal.

ÖZET

BÜYÜK BİYOLOJİK VERİLER İÇİN GÖRSEL ANALİZ ARAYÜZLERİNİN GELİŞTİRİLMESİ VE KANSERDE İMMÜNOMODÜLATÖR KODLAMAYAN RNA AĞLARININ KARAKTERİZASYONU

Her geçen gün daha yüksek boyutlarda veri toplamaya başladığımız bu günlerde, toplanan verileri işleyerek tahmin, sınıflandırma ve modelleme sağlayan araçlara dönüştürmek gelecek açısından oldukça önemlidir. Özellikle kanser alanında, hastalardan toplanan verilerin işlenmesi, yeni biyobelirteçlerin keşfedilmesi, kişiselleştirilmiş tedavi yöntemlerinin geliştirilmesi ve daha iyi prognostikler açısından önemli bir potansiyele sahiptir. Ancak, yüksek boyutlu verilerin incelenmesinde ve analiz edilmesinde önemli zorluklar vardır. Verileri bir araya getirmek ve farklı analiz yöntemleri uygulamak için iyi düzeyde kodlama becerisi gerekir. Bu çalışmada oluşturulan görsel arayüzlerle Kanser Genom Atlası programının halka açık binlerce kanser hastasına ait yüksek boyutlu verileri kodlama bilgisine ihtiyaç duymadan analiz edilebilir.

Geliştirdiğimiz bir biyoinformatik araç olan The Cancer Genome Explorer, kısaca TCGEx, bilim insanlarına kapsamlı bir araştırma fırsatı sunarak yüksek verimli kanser verisi analizlerini kolaylaştırırken aynı zamanda bu alandaki bilimsel çalışmların hızını arttırarak kansere dair olan bilgi birikimimizi arttırma potansiyeli taşır. Kanser alt tiplerini güçlü bir şekilde analiz edebilme ve çoklu kanser verilerini kullanarak karşılaştırmalı analizleri mümkün kılma gibi özellikleri ile TCGEx, özellikle hipotez odaklı araştırmalarda çalışmaları hızlandırarak literatüre katkıda bulunabilir. Bu çalışma ayrıca melanom için TCGExplorer kullanılarak hipoteze dayalı araştırmaların nasıl yapılabileceğini gösteren bir kullanım senaryosu sunmaktadır. Melanomda, tümör ve bağışıklık sistemi arasındaki etkileşimlerin miRNA düzeyinde aydınlatılması, yeni terapötiklerin geliştirilmesi için çok önemlidir. Bu çalışmada, makine öğrenmesi, boyut azaltma, hayatta kalma ve orantılı tehlike modelleme gibi çeşitli istatistiksel analiz yöntemlerini kullanarak belirlediğimiz tümör ve bağışıklık hücrelerine etki eden potansiyel terapötik hedeflerin özelliklerini TCGEx ile birlikte karakterize ediyoruz.

TABLE OF CONTENTS

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

LIST OF ABBREVIATIONS

TCGEx, TCGExplorer, The Cancer Genome Explorer

RNA-Seq, RNA sequencing

mRNA, Messenger RNA

lncRNA, Long Non-Coding RNA

miRNA, microRNA

NGS, Next Generation Sequencing

UV, Ultraviyole

NSCLC, Non-Small-Cell Lung Cancer

TCGA, The Cancer Genome Atlas

CNV, Copy Number Variation

ROC, Receiver Operating Characteristic Analysis

Cox-Ph, Cox Proportional Hazard Analysis

KM, Kaplan-Meier Curve Analysis

ML, Machine Learning

ncRNAs, Non-Coding RNAs

tRNA, Transfer RNA

rRNA, Ribosomal RNA

CRC, Colorectal Cancer

PC, Prostate Cancer

miR, miRNA

SKCM, Skin Cutaneous Melanoma

CTLs, Cytotoxic T Cells

TME, Tumor Microenvironment

Dedicated to all Cancer Patients…

CHAPTER 1

INTRODUCTION

1.1 Background of the Study and High-Throughput Data Analysis

Technology has advanced rapidly and steadily, making it feasible to process and store a large amount of data in many fields. With an increased ability of data, statistically significant deductions, consistent and accurate predictions can be obtained. Thus, processing and analyzing high-throughput data has gradually become more and more crucial. Especially in the context of biomedical sciences, the lowering costs of Next Generation Sequencing (NGS) led to an explosion in high-throughput data. It has now become possible to analyze the ever-changing transcriptomes of cells under different conditions (Z. Wang, Gerstein, and Snyder 2009; Lander et al. 2001) to understand the underlying biological mechanisms. With RNA-Seq, high-resolution sequencing and analysis of various RNAs such as Messenger RNA (mRNA) as well as non-coding Micro RNA (miRNA) and Long Non-Coding RNAs (lncRNA) has become possible. As next-generation sequencing methods have become increasingly common and usable, the data created by the results have begun to accumulate. With this accumulation, there is an increasing need to develop user-friendly analysis interfaces that can facilitate the analysis of high-dimensional data.

High throughput analysis approaches allow the study of biological systems at the epigenetic, transcriptomic, and genomic levels, which creates a remarkable ground for biomedical discoveries. The Cancer Genome Atlas (TCGA), is a sophisticated program initiated by the National Cancer Institute and the National Human Genome Research Institute in 2006 and today contains more than 20000 characterized primary cancer samples in 33 different cancer types. This ascendant cancer genomics program also includes normal samples that match primary cancer samples that are essential for comparison. TCGA made available to all users the omics data of thousands of patients in different cancer types, along with the clinical characteristics of the patients (Weinstein et al. 2013). Our knowledge about the molecular mechanisms of cancer has gradually

increased with the help of the TCGA data leading to the identification of therapeutic targets, the development of new treatment approaches and explore new potential biomarkers (Sanchez-Vega et al. 2018; Peng et al. 2018). The TCGA database provided the basis for many studies, one of which demonstrated the genomic classification of melanoma while simultaneously demonstrating gene signatures related to improved survival. Another study using the TCGA database provided a resource for exploring immunogenicity among cancer types when identifying immune subtypes including cancer tissue types and molecular subtypes (Thorsson et al. 2018). Accordingly, further interpretation and analysis of this data have been gaining great importance which accounts for the rapid expansion of the literature and thus, these explorations.

An effective analysis of large-scale genomic data is only achievable through programming because these high-dimensional data are highly challenging to handle and have too complicated of a structure to be evaluated in basic applications. Although there are pipelines ready to perform these analyses, it is challenging to perform these operations for scientists without basic coding, and statistics background. Web-based tools have been developed over time to accelerate cancer research and to perform these analyses quickly/effectively. Web-TCGA (Deng et al. 2016), is one of these bioinformatic tools and facilitates mutation, methylation, expression, and copy number variation (CNV) analysis. Another tool, UCSC Xena, can perform many functions, from discovering relationships between genomic and clinical cancer data to comparing normal and cancer tissue types (Goldman et al. 2020a). With GEPIA 2, it is possible to carry out research such as gene expression profiling and cancer survival analysis (Tang et al. 2019a). Many bioinformatics tools such as these have been used in numerous studies, contributing to the rapid progress of the cancer research field. These tools enabled petabytes of cancer data to be processed and analyzed, and the results of these analyses to be visualized and presented understandably. These bioinformatic tools that combine and present several different analysis methods have increased the usability of this TCGA data. The ability to effectively and rapidly analyze TCGA data has empowered scientists in their quest to discover novel treatments for various types of cancer. The development of visual interfaces for these tools has enabled scientists to obtain analysis results without requiring coding expertise. The bioinformatic approaches offered by these tools have played a crucial role in enabling scientists to take proactive measures against cancer. However, the existing bioinformatics tools are primarily designed for performing specialized scientific

tasks and offer limited analytical methods. There is a lack of an exploratory tool that enables a comprehensive examination of gene expression data.

1.2. Cancer

Cancer stands out as a major public health problem, the frequency of which is increasing every year. Although scientists have gained more knowledge about the onset of cancer and the processes that lead to the demise of patients compared to the past, many mechanisms underlying cancer still remain enigmatic. Cancer formation can occur through numerous diverse mechanisms. Genetic mutations and epigenetic changes in cell DNA play an important role in the formation of cancerous cells. Mutations can occur spontaneously during cell division or due to environmental factors. Environmental factors can include exposure to harmful mutagenic chemicals, an unhealthy diet, and overexposure to ultraviolet (UV) rays or radiation. Exposure to various mutagens and carcinogens can lead to the development of different types of cancer in the body. For instance, excessive exposure to UV radiation is a prominent risk factor for the skin cancer type known as Skin Cutaneous Melanoma (SKCM). This type of cancer originating from melanocytes is the most aggressive and deadly of skin cancers and nearly 50,000 people die each year because of this malignancy (Rastrelli et al. 2014). After the cancerous cells are formed by any agent, they are frequently eliminated by the immune system, which is very complex. Or it may remain in the equilibrium phase without being completely destroyed but suppressed by the immune system. Cancerous cells in this phase can increase their activation and escape with ease from the immune system at times when the immune system is weakened, such as aging, organ transplants, and so on. Or cancerous cells can actively suppress immunity without the need for any immune weakness. Cancer cells that manage to escape the immune system can form tumors over time (Dunn, Old, and Schreiber 2004).

Many features distinguish cancerous tissues from normal tissues. Some of these include unlimited proliferation, avoidance of growth suppressors, resisting cell death, achieving replicative immortality, inducing angiogenesis, and the ability to invade and metastasize (Hanahan and Weinberg 2011). Along with these distinctive features, genome instability and tumor microenvironment elements are also crucial for the course of cancer. To develop new treatment methods for cancer, it is of great importance to elucidate these

distinctive mechanisms. In particular, understanding the roles of the tumor microenvironment (TME) in tumor development and spread has been a turning point in the discovery of new immunotherapeutic targets (Sas et al. 2022). The TME encompasses the relationships within the framework of molecules released and produced by these cells in and around the tumor, including cancerous cells as well as non-cancerous cells (Xiao and Yu 2021). Gaining a more comprehensive understanding of the intricacies involved in the TME holds the potential to facilitate its regulation, consequently augmenting the effectiveness of cancer treatments (Bilotta, Antignani, and Fitzgerald 2022).

Immunotherapy has emerged as a promising treatment approach for cancer patients. Immunotherapy drugs can also be used in combination with other therapeutic approaches including chemotherapy, radiotherapy, and surgical resection (Herskind, Wenz, and Giordano 2017). Although traditional treatment approaches work in many patients and significantly increase survival, they do not work for all patients. It has become possible to offer personalized treatment options by dividing patients into cancer type, stage, mutation types, and many other subgroups, with immunotherapeutic approaches. For personalized treatment methods to be possible, it is necessary to have epigenetic, transcriptomic, and genomic levels of data about cancer patients and to be able to process them in the computer environment. Thus, patient-specific treatments can be offered by obtaining more information about the tumors and immune systems of the patients.

1.2.1. Cancer Research in Silico

With the diversification of immunotherapy methods and then the development of specific treatment methods for individuals and patient groups, cancer research in silico has gained great importance. Examination of patients' RNAseqs with various analysis methods has paved the way for specific treatments to be applied by finding the ones with the highest potential among thousands of therapeutic targets (Ding, Chen, and Shen 2020). In addition, the mechanisms discovered and verified by in silico methods became more easily targeted. A wide variety of methods have been developed to examine patient data in silico. For example, the Kaplan-Meier and Cox Proportional Hazard (Cox-Ph) methods used in survival analysis are robust tools for predicting how certain traits affect survival. Using these tools, genes that most affect prognosis can be identified and in vitro

or in vivo studies can be performed on them. Receiver Operating Characteristic (ROC) analysis can be used to measure the power of the detected potential targets and to understand their false positivity. Various analysis methods such as these have accelerated cancer studies and increased our knowledge of cancer exponentially.

In silico studies, which have gained a new vision with the use of artificial intelligence algorithms in cancer studies, which have been increasingly used recently, give hope in terms of a better understanding of cancer, predicting the response of patients to treatment, and increasing their survival (Rafique, Islam, and Kazi 2021). For example, in a study conducted on non-small-cell lung cancer (NSCLC) patients, it was shown that by applying machine learning methods based on RNA expression data from patients, response prediction to immunotherapy treatment could be performed (Wiesweg et al. 2019). In another study, artificial intelligence applications used to predict metastasis, the leading cause of cancer-related deaths, were compiled (Albaradei et al. 2021a). In addition to these, the results obtained by processing the DNA methylation data, pathological and radiological images of cancer tissues, personal and psychological data of the patients, as well as the collected RNAseq data, are used to obtain information about the mental and physiological prognosis as well as the progression of cancer through machine learning (Afshar et al. 2020; Kourou et al. 2021; Yousefi et al. 2022). To summarize, in silico cancer studies are of great importance for discovering potential powerful targets for patients and converting them to immunotherapeutic methods, predicting the prognosis of cancer, and predicting the response of patients to treatment.

1.3. Non-Coding RNAs

Noncoding RNAs (ncRNAs) are RNA molecules that are not translated into proteins. They are found in various regions of the genome, including intergenic and intronic regions. These gene sequences whose functions remained poorly understood for many years and were often regarded as non-functional were even referred to as "junk RNA". However, today it has been understood that ncRNAs play critical roles in important processes such as regulating cellular functions, regulating gene expression, and transferring genetic information (Mattick and Makunin 2006). Although the functions of many ncRNAs remain to be elucidated, they have been proven to be involved in many important cellular processes including gene regulation, post-transcriptional modulation,

antitumor immunity and tumor immunoevasion (Xu, Wang, and Huang 2021; Xing et al. 2021; Ekiz HA et al. 2019).

ncRNAs also have diversity within themselves. Transfer RNAs (tRNAs), which are involved in protein synthesis, Ribosomal RNAs (rRNAs), which are one of the structural components of ribosomes, and Long noncoding RNAs (lncRNAs), which play a role in the regulation of cellular functions, are just a few of them. miRNAs, on the other hand, usually target messenger RNAs (mRNAs) and act as regulators in regulating gene expression and in various cellular processes. miRNAs are short, single-stranded RNA molecules approximately 20-22 nucleotides in length. They play a crucial role in the post-transcriptional regulation of gene expression by selectively targeting the 3'-untranslated region of specific mRNAs (S. Zhu, Pan, and Qian 2013). This interaction leads to either the degradation of the targeted mRNA or the inhibition of its translation, thereby influencing gene expression levels. The primary transcripts, called pri-miRNA, are processed, first into short stem-loop structures called pre-miRNA, and then into functional miRNA. miRNAs play important roles in host-pathogen relationships. In addition, it has been shown that miRNAs are involved in critical steps in developmental timing, cell differentiation, apoptosis, proliferation, and formation of cancerous tissues (Jiao et al. 2021; Di Leva, Garofalo, and Croce 2014). While miRNAs, which act as regulators in many intracellular processes, perform basic processes such as miRNA regulation, epigenetic changes such as methylation, and the circadian clock, they are in coordination with various effectors such as lncRNAs (Cai et al. 2009).

1.3.1. miRNAs Controlling the Cancer Biology

miRNAs, whose overexpression has been shown to be associated with many human diseases, are regulators of many cellular processes including carcinogenesis (Yoshida, Yamamoto, and Ochiya 2021). miRNAs are capable of targeting multiple mRNAs, but multiple miRNAs can act as co-regulators for a single mRNA. The situation is not much different for miRNAs in tumor suppression and immune responses. Therefore, it is essential to elucidate the miRNAs and miRNA interactions involved in all these processes to better understand the formation, immune escape and spread of cancer cells (Dragomir et al. 2018). A better understanding of these mechanisms can lead to the discovery of therapeutic targets and promising biomarkers. However, while it is not

simple to elucidate the role of a single miRNA involved in these processes, predicting miRNA linkages and networks is a challenge. Nevertheless, successful studies have been carried out to elucidate the mechanisms between miRNA, tumors, and the immune system. For example, these studies showed that miRNAs play a critical role in the Wnt signaling pathway dysregulation (Balacescu et al. 2018), which has an important role in the emergence of colorectal cancer (CRC). It is now accepted that miRNAs cause pathogenesis in CRC by triggering and inhibiting the Wnt signaling pathway (Jafarzadeh and Soltani 2021). In another similar study, the role of specific miRNAs in the pathogenesis of prostate cancer (PC) was shed light on. miRNAs-145,148, and 185 have been shown to be involved in regulating and directing PC stem cell behavior (Coradduzza et al. 2022). In other research, the role of miRNA-155 on cancer cells was investigated (Mattiske et al. 2012) and the regulatory importance of miRNA-155 on different immune cell populations was tried to be determined. As a result, miR-155 has been shown to have different roles depending on the cell type (Thompson et al. 2023).

Taken together, a more complete understanding of the regulatory roles of miRNAs in immune and cancer cells is important for combating diseases better. Furthermore, examination of high throughput data will continue to help miRNAs that can serve as biomarkers and potential therapy targets.

1.4. Aim of the Study

The aim of this study was to develop a bioinformatic tool we named The Cancer Genome Explorer shortly TCGExplorer or TCGEx that will enable rapid, effective, and comprehensive analysis of the high-dimensional cancer data. Thus we aim to facilitate cancer research and enable scientists without coding expertise to benefit from the TCGA data sets. Thus, we establish a bridge between cancer immunology, statistics, and software sciences. Our application also aims to contribute to the research of everyone working in the field of cancer by providing a user-friendly interface. In this way, scientists can quickly analyze data and direct their studies with a few clicks. Some bioinformatic tools that enable the analysis of cancer data are already available. These tools, which mostly offer several analysis methods at the same time, have filled important gaps in the literature and inspired our work. TCGEx aims to be a resource for comprehensive in silico

cancer research, designed to make a difference with its original aspects, by putting these tools on top of the gains.

The study aimed to develop TCGEx software, which encompasses a multitude of powerful features setting it apart from other tools in its category. These distinctive features include interactive interfaces, multiple modules for conducting sophisticated analyses, customizable graphics, subset-specific analysis capabilities, and the implementation of linear regression-based machine-learning algorithms. It is also completely free and easy to use by all users. With the following modules, TCGEx becomes a unique tool that scientists can use in exceptionally comprehensive/in-depth cancer studies: Principal Component Analysis (PCA), Kaplan-Meier Analysis Module, Receiver Operating Characteristic (ROC) Curve Analysis, Heatmap Analysis, Cox Proportional Hazards (Cox-Ph) Model Analysis Module, Gene Set Enrichment Analysis (GSEA), feature correlation analysis, correlated gene table analysis, metadata analysis module, and machine learning algorithms analysis. The high throughput cancer patient data can be rapidly, efficiently, and thoroughly analyzed with a variety of approaches using this user-friendly interface, which was created using the R programming language. The outcomes are then shown in a fashion that is suitable for publication in the paper.

CHAPTER 2

IMPLEMENTATION

2.1. Summary of the Pipeline

The TCGEx tool was created using the R programming language and the Shiny framework (Chang W et al. 2023). The TCGEx application was supported by online servers and opened to the access of the entire scientific community. The open-source code of the application is available on GitHub (https://github.com/atakanekiz/TCGEx).

The transcriptional data obtained from The Cancer Genome Atlas (TCGA) encompasses gene expression levels of both coding and non-coding genes. Analyzing this vast dataset from thousands of patients can unveil previously unknown mechanisms underlying cancer. We gathered RNA sequencing (RNAseq), and miRNA sequencing (miRNAseq) expression datasets obtained from cancerous and normal tissues of patients at different cancer stages by downloading via TCGAbiolinks package and normalizing them. The normalization procedure enables us to make meaningful comparisons between samples that have undergone sequencing with varying depths, by aligning them on a similar scale. This essential step ensures that the gene expression measurements are adjusted appropriately, allowing for accurate and reliable analyses across different samples. While it is possible to use raw data directly in differential expression analysis and some other analysis methods, we scaled the raw expression counts by considering the sequencing depth of different samples to perform the analyzes in TCGEx in the most accurate way. Converting raw data to scaled harmonized data and using it for analysis is a common approach and there are some methods for this such as counts per million (CPM) and log2-counts per million (log-CPM). It allows for the comparison of gene expression levels between samples by scaling the raw gene counts to a common denominator. By dividing the raw counts by the library size (total counts) of each sample and then scaling it to a million, CPM normalization provides a relative measure of gene expression that is independent of the sequencing depth. This normalization method ensures that differences in library size do not introduce bias when analyzing gene expression levels across samples. In the data preparation part, we transformed our counting matrices using log-

CPM without taking into account gene lengths. In the data preparation part, we transformed our counting matrices using log-CPM without taking into account gene lengths. Since we are not comparing absolute expression levels of genes among each other and were examined in correlation analyses and related analyzes, this normalization process is the best fit for data to be used in TCGEx. For the normalization process, we used the cpm function from the edgeR package in the R software. This function basically applies the formula log(CountsPerMillion(data)+1) to each expression value. Normalization of gene expression data in TCGA facilitates its utilization in TCGEx.

Then, we created large data frames containing hundreds of rows of samples and thousands of columns containing genes or features to be used in analyses for each cancer type by downloading clinical data containing detailed information about cancer patients, from gender to mutation patterns, via TCGAbiolinks package, and combining them with normalized RNAseq and miRNAseq data. As a final step, the immune scores from the seminal article "The Immune Landscape of Cancer", which describes general immune patterns in tumors, have been added to each cancer dataset to enhance its utility for analysis (Thorsson et al. 2018). As a result of all these strategies, the processed data sets that will be used in the modular analyses were prepared.

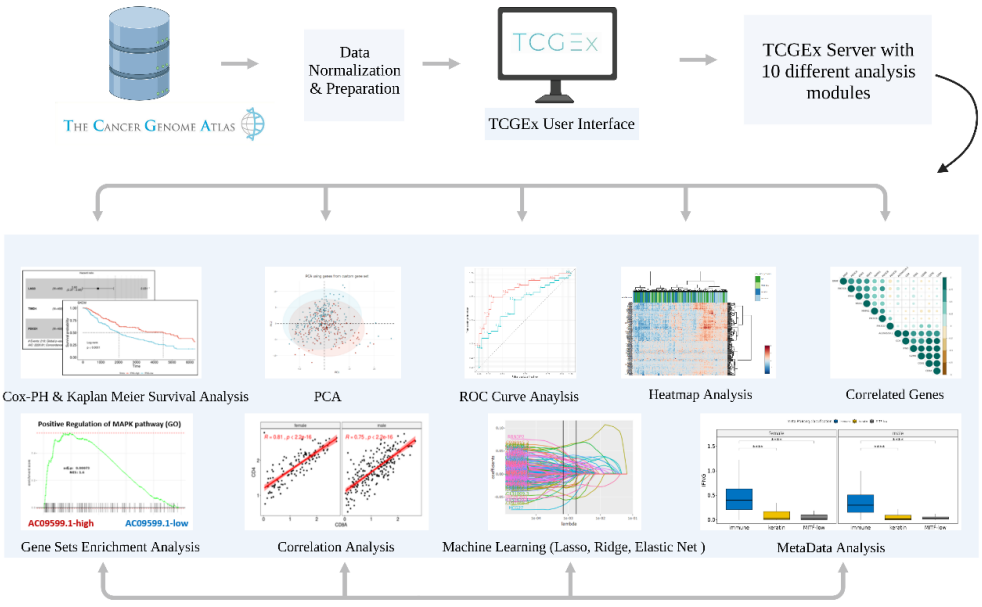


Figure 1.1. The General Operating Mechanism of TCGExplorer.

2.2. The TCGEx User Interface and Server

Shiny apps are composed of two main components: the user interface (UI) where inputs are received by interacting with the user and the server side for the computational processes. One of the prominent features of TCGEx provides a wide range of motion by allowing the user to enter many inputs to customize the analysis thus an input selection screen makes it possible to develop tailored analyses. In addition, different tabs were created for each module, allowing the user to easily navigate between analyses. Correspondingly, the code of each analysis tab was written in a modular way. This approach facilitates troubleshooting and new feature development without affecting the other parts of the app. In addition to troubleshooting, modularized writing of the code in the app increases readability and provides code cross-usability.

2.3. Data Selection Module

Data Selection is the first module in the application and through this module, the user can select the type of cancer to investigate, load the relevant data and proceed to the analysis part. Moreover, this module visualizes various descriptive statistics such as gender, age at diagnosis, and cancer stage of these specific cancer patients. Hence, the user gains more knowledge about the selected cancer data and gains insight into which subsets the data can be analyzed. While the evaluation of selected cancer types alone reveals mechanisms specific to that cancer, being able to examine different cancer types at the same time may offer an advantage for discovering common patterns in different cancer types. Thus, we also added the ability to select multiple cancer projects and obtain aggregated data.

GEPIA 2, Stanford TCGA Clinical Explorer, and UCSC Xena are tools that can offer advanced analysis in their fields (Lee et al. 2015; Goldman et al. 2020b; Tang et al. 2019b). While these tools make it possible to analyze only one cancer data at a time, TCGEx provides the user the right to select multiple cancer data simultaneously. In this way, the user can aggregate and examine more than one cancer data in aggregate and perform comparative analyses between these cancer data. It allows users to use cancer type as a variable in ensuing analyses including survival modeling and graphing. Offering

a personalized analysis experience, TCGEx presents prospects for users to perform the analysis they want with its state-of-the-art and expansive variety of analysis methods.

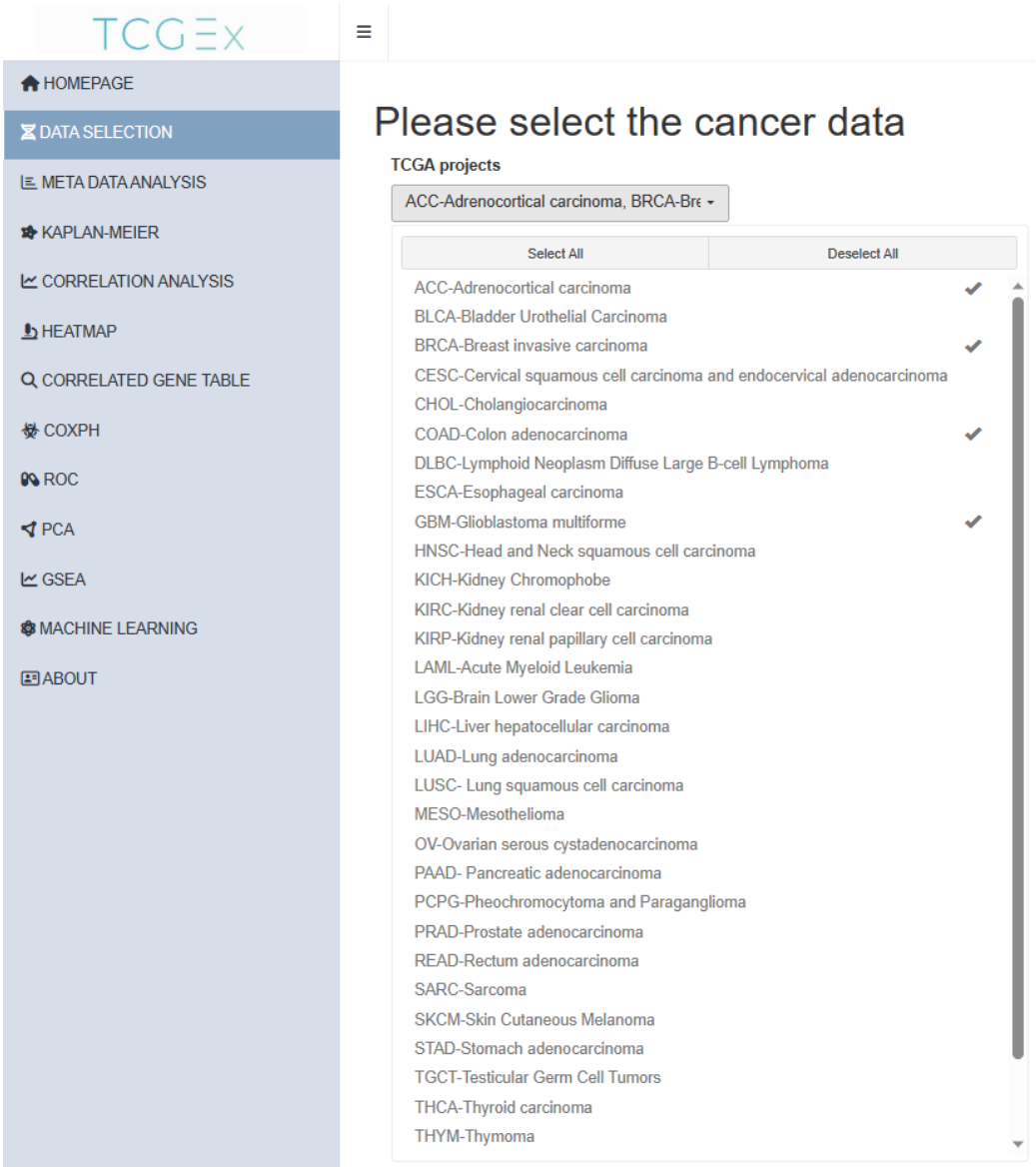


Figure 2.1. Data Selection Module User Interface. With the user interface in the Data Selection Module, the user can start to perform analysis by selecting one or more cancer types among 33 harmonized cancer types downloaded from TCGA.

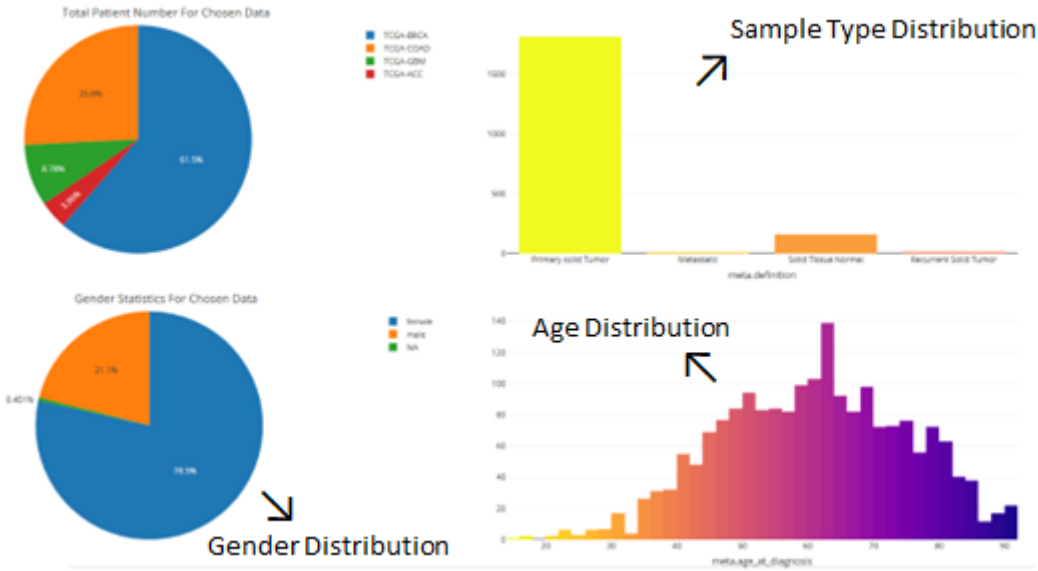


Figure 2.2. Representative visualizations from the Data Selection Module. After the data selection is made by the user, various information is presented to create a general impression about the selected cancer types. Age, gender distributions, and sample definition are some descriptive stats offered. After this stage, the user can start researching by choosing the analysis method.

2.4. Principal Component Analysis Module

PCA analysis tool provides visualization of the large tumor sample datasets via dimensionality reduction. This tool also offers users studying on the entire RNAseq/ miRNAseq genes or a subset of genes. The gene subsets can be selected from previously defined annotated pathways from The Molecular Signatures Database (MSigDB) or users' custom gene sets. By using this feature, users can eliminate the masking effect and noisy information (Yeung and Ruzzo 2001). While the PCA analysis is performed on the gene subsets as described above, users can annotate data points using clinical metadata or categorize gene expression as high and low. This feature allows studying whether there is that feature-related separation on the gene subset. For instance, to see whether high expression of a specific miRNA divides the patients on the principal component space calculated by relevant gene sets such as immune-related genes (Atakan Ekiz et al. 2019). In another case, it can be observed that clinical metadata such as PAM50

subtypes of breast cancer is separated well on a PCA plot using gene subsets (Tsai et al. 2015). This tool also allows for deciding on the colors of the plot. Users can select color palettes for specific journals including Nature, Science, and Journals of Clinical Oncology for conveniently generating publication-ready plots. In addition to customizing the plotting parameters, users can also modify the input data by changing the centering and scaling features. Centering involves subtracting the mean from each observation, which centers the graph at the origin and makes it easier to interpret relationships between variables. Meanwhile, scaling involves dividing each variable by its standard deviation, which avoids the clustering effect that binary variables can have in principal component analysis (Jollife and Cadima 2016). Using centering and scaling helps researchers interpret their data accurately and avoid common pitfalls, providing more reliable analysis and insights. The "keep highly variable genes" feature in the principal component analysis includes only genes with specified levels of variance, limiting the number of genes in analysis, which may lead to more distinctive patterns in the PCA analysis.

The interface designed for users to use the PCA Module is shown in Figure 2.3. In the first input in the PCA Module, the user can select sample types such as "Primary Solid Tumor", "Metastatic" and "Solid Tissue Normal" to include in the analysis here. Then, in the second input, the user can select the genes that will be used in PCA. Users can choose 5 options: i)All genes (RNAseq and miRNAseq data), ii)miRNAs (mature miRNAs), iii)RNAseq(includes only coding genes), iv)genes annotated in MSigDB gene sets, v)a custom list of genes by uploading the file. The Center Data radio button comes by default and gene expression values are centered by subtracting the mean expression value. A variable that is on a different scale from the others may dominate the variance direction. Scaling (default) gene expression values by the scale variable radio button prevent this effect. Then MSigDB Collection and desired gene sets can be selected. With the Apply variance filtering radio button, the user can apply variance filtering to keep the most highly variable genes in the analysis. Setting this value to 10, for instance, will select the genes having the top 10% highest variation in the dataset. In the feature annotation input, Users can color code the data points on the graph using gene expression values or clinical metadata. If the user selects a gene name here, gene expression will be categorized at the median value per sample and points will be annotated. Users can also select clinical metadata (eg. meta.gender) to color points accordingly. Finally, with the color palette

input, the user can choose to create the graphic in a format that can be published in certain scientific journals such as "Nature" or the desired color palette.

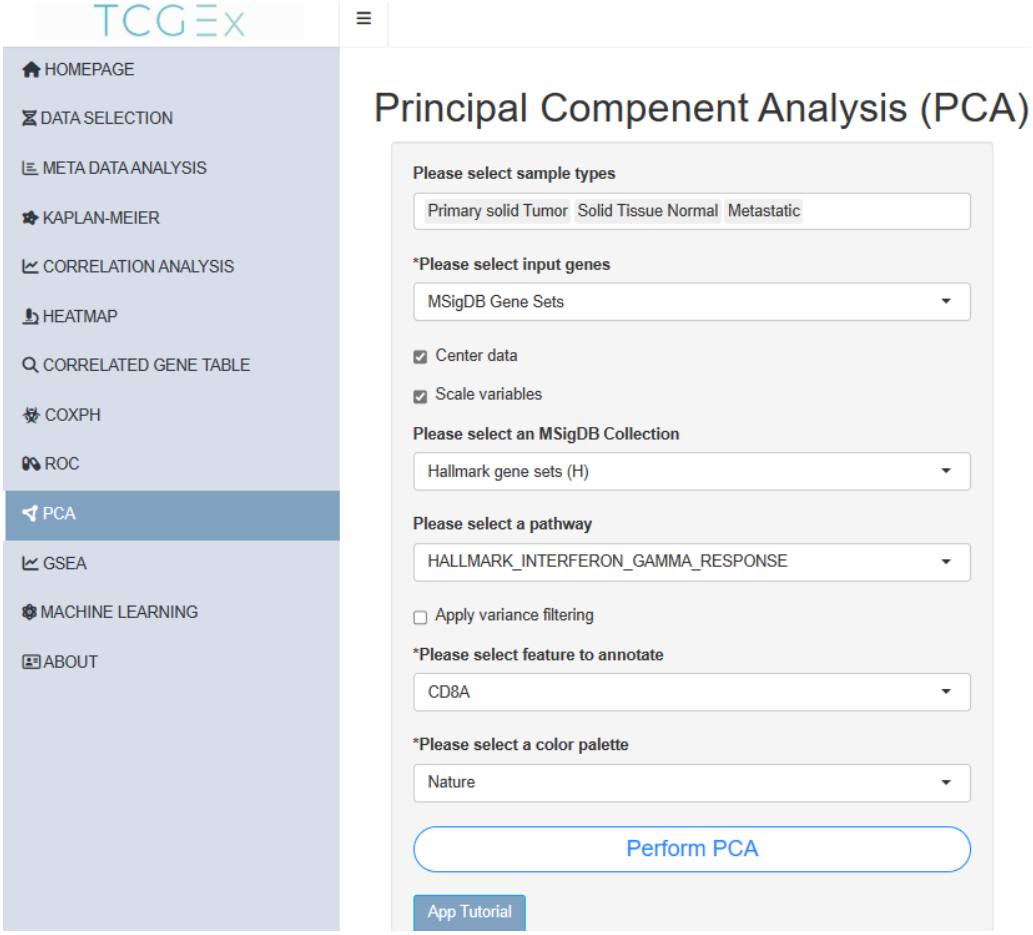


Figure 2.3. PCA Module User Interface

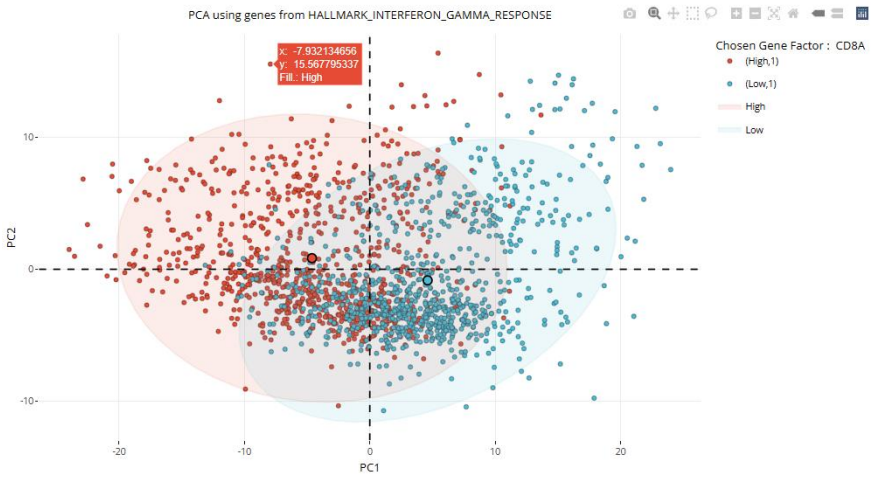


Figure 2.4. PCA Module Graphical Output. The graph was formed by annotating the "CD8A" gene a cytotoxic T lymphocyte marker, and PCA analysis was

performed using the “Hallmark Interferon Gamma Response” gene set selected from MSigDB on the “Primary Solid Tumor”, “Metastatic” and “Solid Tissue Normal” samples selected by the user. With the panel in the upper right, a detailed analysis can be made of the analysis result and the graphic can be downloaded if desired.

2.5. Receiver Operating Characteristic Curve Analysis Module

Receiver operating characteristic (ROC) analysis is used to find the efficiency of a created model. Points on the ROC plot portray a curve that contains a sensitivity/specificity match for each data point. The sensitivity of a signal can be identified by the percentage of true positives and the specificity of a signal can be observed by the percentage of true negatives(Nahm et al. 2022). The best result is obtained when there are no false positives and false negatives exist(Sø et al. 2009). The area under the curve (AUC) on the ROC plot is a widely used parameter to compare the performances of different ROC analyses. AUC> ~0.5 means the classifier is working better than random chance (Sonego, Kocsor, and Pongor 2008). The bigger AUC corresponds the better the classifier prediction results. ROC curve analysis is very useful in terms of testing new potential biomarkers. Biomarkers can be used to determine whether a disease is present, or absent, or whether the treatment is effective or not (Hsu, Chang, and Hsueh 2014). Finding new biomarkers and testing them is a challenging process. Multiple biomarkers may be combined to reduce the false positive rate. Still, new biomarkers are found, and different combinations are created to boost performance. By using ROC curve analysis, the sensitivity and specificity of potential biomarkers can easily be seen, combined, and measured.

The ROC module in this website will help users to use the already cleaned and prepared TCGA cancer database. This allows new biomarkers to be assessed, and different biomarkers to be combined and comparable within seconds. In this module, ROC curves are drawn according to the gene expressions in the data. The option to add MSigDB gene sets and their subcategories will make the newly created/improved biomarkers to be more comparable by taking the already existing gene sets as reference. Users can customize their analysis by choosing whether they would like to examine numerical gene expressions or categorical metadata. This flexibility increases the analysis

perspective of the module as not only diseased and non-diseased tissues will be examined, but anything that a user wants that is in the data set. Users who are interested in more specific areas of gene expression such as the top and bottom 20% of expression can easily filter the data and narrow the expression range. A point that should be carefully considered is, this procedure lowers the sample size which might result in low accuracy of the curve according to the situation. In a simple scenario where the diseased and non-diseased tissue is required to be analyzed by a potential biomarker, users can choose the normal and tumor tissue on the data and display the AUC values for each curve. There is no consensus but in some papers in the literature, AUCs higher than 0.80 are notable biomarkers (English et al. 2016). AUC values can also be used to compare it with other plots to choose the optimum curve. With this module, users who are interested in using ROC plots to review TCGA data and use it in aims such as improving biomarkers or more will be able to do it in seconds with the flexibility to run it repeatedly.

The interface created for users to use the ROC Module is shown in Figure 2.5. In the first sample type selection input, users can subset data. ROC analysis is performed between two classes in the response variable as '1' (desired outcome group) and '0' (undesired outcome group). Users can specify these classes for both categorical and numeric variables in the TCGA data. In the second input, users need to select what kind of response variable they are interested in. Users can select the specific numerical or categorical variable they want to binarize according to the previous selection in the third input. High cutoff input selection allows the users to assign samples to class-1 if their values are more than the specified quantile here. For instance, setting this value to 25 would mean binarizing the top 25% of the data as '1'. If the user wants to binarize at the median value this value should be 50 (default). Low cutoff input allows the user to assign samples to class-0 if their values are less than the specified quantile here. For instance, setting this value to 25 would mean binarizing the bottom 25% of the data as '0'. If the user wants to binarize at the median value this value should be 50 (default). In the next input selection, users select the predictor variable. If the user enters more than one variable, their values will be averaged. Subsequently, if users desired the average expression value of a specific MSigDB gene set can be added to the graph. If the user would like to include a curve for a specific MSigDB gene set, this can be accomplished at the last input selection.

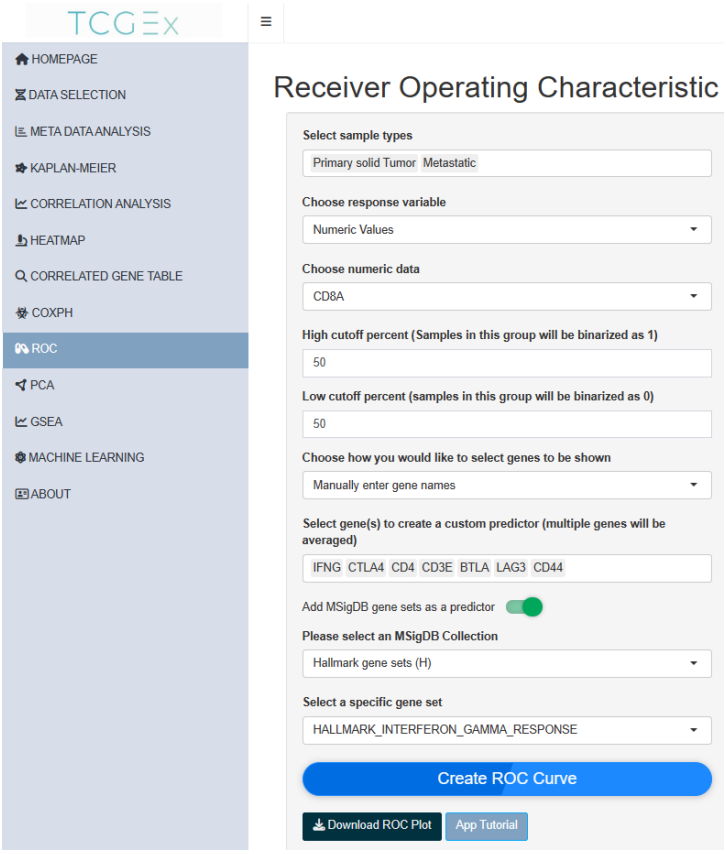


Figure 2.5. ROC Module User Interface

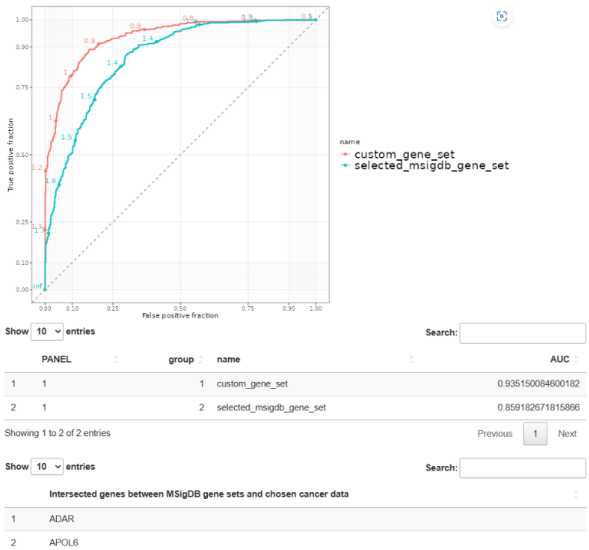


Figure 2.6. ROC Module Result. The graph shows the ROC analysis result and the area under the curve scores of the predictor custom and MSigDB gene sets selected just below. The table at the bottom of the result contains intersected genes between MSigDB gene sets and chosen cancer data.

2.6. Heatmap Analysis Module

In biology, heatmaps can be used to review the expression values of genes (Gehlenborg and Wong 2012). The ability to display a high number of gene expression data in one heatmap simplifies the high-throughput data analysis. By using TCGA cancer data, RNA and miRNA gene expression patterns can be detected and reviewed within the heatmap module. The heatmaply package in R is used to create the heatmap. Just like on the other modules on the website, user can either create their own gene set and look them up on the heatmap or choose from the Human MSigDB gene sets. This gives the user the flexibility of adding the already curated MSigDB gene sets within seconds. At the same time, users can specific scaling per gene or sample to create different visualizations. It also has the flexibility to only see the genes which have the top expression variability. This feature will hide the genes that do not have distinct expression values on the heatmap. One of the most important parts of heatmaps is annotations. The samples can have multiple annotation bars on top of the heatmap and, different categorical metadata or categorized high and low expression values of genes can be shown. By choosing the desired features on the interface, the heatmap which was created in a few seconds will give us the information of hundreds of patients.

The interface developed for users to use the Heatmap Module is shown in Figure 2.7. Users can select the genes to be used in the heatmap analysis in the second input, after substituting their data by selecting the sample type. The user can complete this process by manually selecting the gene, using the MSigDB gene sets, or uploading their own gene set. After selecting the genes for the plot, with the variation filter the user can apply a variance filter to keep only highly variable genes in the plot. 100 (default) means no filter is applied. If the user like to see the top 10% variable genes only, set this value to 10. Such filtering can help see more informative genes. Then users can select categorical clinical metadata features to show as annotations on top of the heatmap. Users can also create an annotation bar by categorizing the patients based on their gene expression levels. Users can specify one or more genes in this input. When multiple genes are entered, their average is calculated. Patients are categorized as 'high' and 'low' according to the median gene expression value. In the last 4 inputs, the user can select distance calculation and clustering methods for samples and genes.

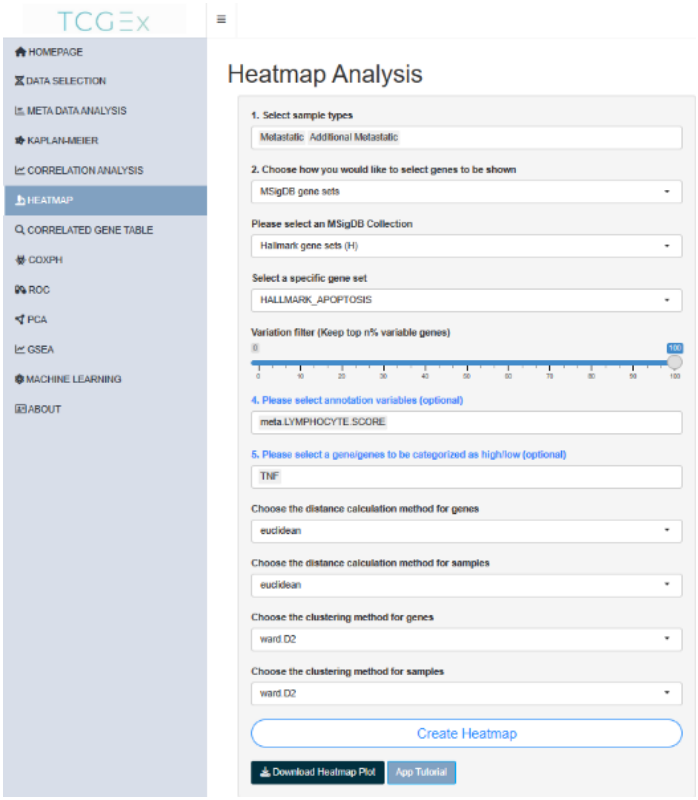


Figure 2.7. Heatmap Module User Interface

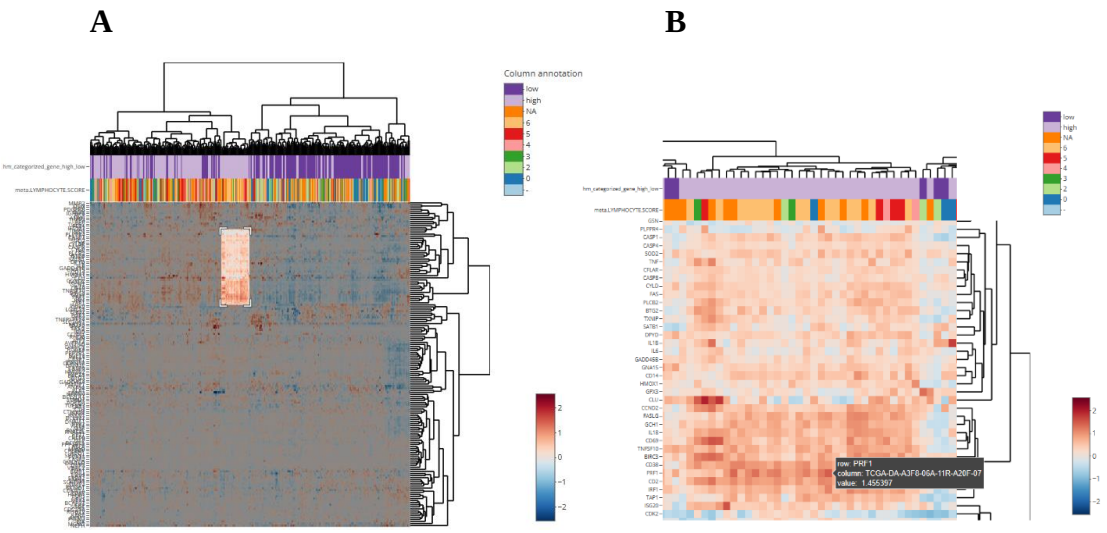


Figure 2.8. Heatmap Module Results. Heatmap analysis graph created with user inputs where expression patterns on selected genes can be examined. Zoom-in and zoom-out features are included to focus on specific genes. In this way, the user has the opportunity to examine the parts of interest more closely and analyze them in depth. In the B panel, it is possible to closely see a selected region of the plot that emerged as a result of the first analysis.

2.7. Gene Sets Enrichment Analysis Module

Although a significant part of the human genome has been elucidated, the functions and processes of many genes are unknown. Characterizing these unknown or currently supposed unimportant genes has the potential to illuminate unexplored mechanisms in cancer. The gene set enrichment analysis (GSEA) module is used to determine whether a gene or clinical feature can define subsets of samples that are enriched or depleted in certain pathways or biological processes. In this way, the relationship of the investigated feature with various cancer phenotypes can be investigated.

With this TCGEx module, GSEA can be performed quickly for the selected feature by filtering the samples in the selected cancer data. With the inputs entered by the user, gene sets categories such as "Hallmark" in the Human MSigDB are decided for the selected feature, and the most enriched or depleted gene sets are defined in this category. Moreover, many gene sets are analyzed simultaneously, enriched and depleted ones are presented to the user with their statistical results. Furthermore, the customized interface allows the user to manually determine the High and Low cutoff percent for the selected feature. Thus, the user can categorize samples based on gene expression at desired cutoffs and perform GSEA between two subsets of data. That way the features of samples expressing different levels of a gene of interest can be studied. The user can find significant enrichments by looking at all gene sets, as well as can obtain detailed results about the selected feature by performing the analysis on the specific gene set that is specified. The fgsea package was used to generate the gene set enrichment analysis results (Korotkevich G et al. 2019).

The interface generated for users to use the GSEA Module is shown in Figure 2.9. The user can start the analysis by choosing which sample types to include in the analysis. GSEA is performed between two groups of data. In the second input, if the user would like to perform GSEA for a categorical clinical feature, the user is expected to select two data subsets and define one of them as the 'sample' for the analysis (the other one will become a reference). If the user would like to perform GSEA for a numerical feature such as gene expression, then the user can categorize samples based on gene expression values as 'high' and 'low' through user-defined quantiles. Setting high and low cutoffs to 50 will categorize gene expression at the median value. Users can set these numbers to 25 to

compare the top 25% of expressors to the bottom 25% of expressors. Then in the third input, the user can increase the number of permutations for preliminary estimation of P-values. Afterward, the user can continue the analysis by pulling the gene sets to be used in the analysis from MSigDB or uploading them locally. Finally, if the user is continuing the analysis on MSigDB gene sets, the user can select the "Top Pathways" option and choose an MSigDB gene set collection and determine which gene sets in that collection are enriched or depleted the feature the user has determined. If "Specific Pathway" is selected, a specific gene set in the selected MSigDB Collection can be examined and a specific GSEA graph can be drawn. With the download buttons at the bottom, the user can download the analysis result they want to their device.

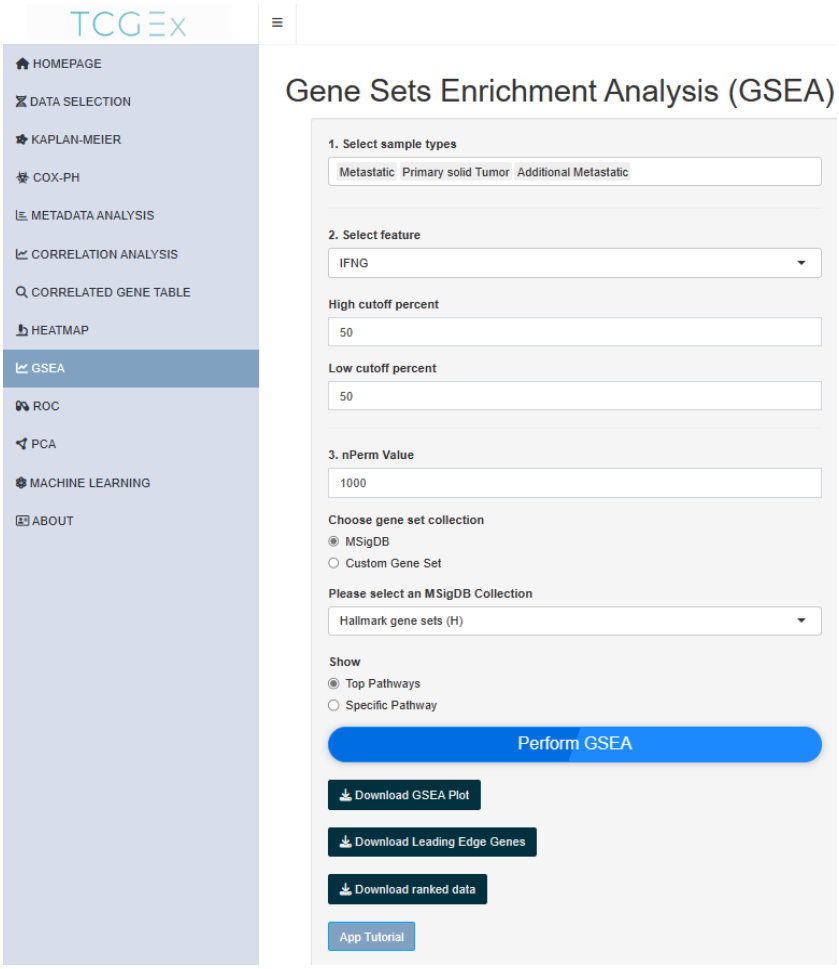
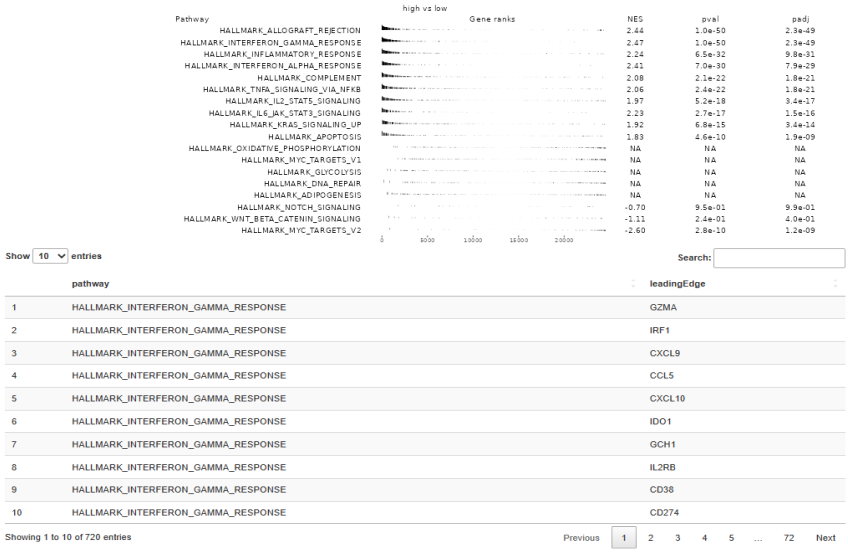


Figure 2.9. GSEA Module User Interface

A



high vs low

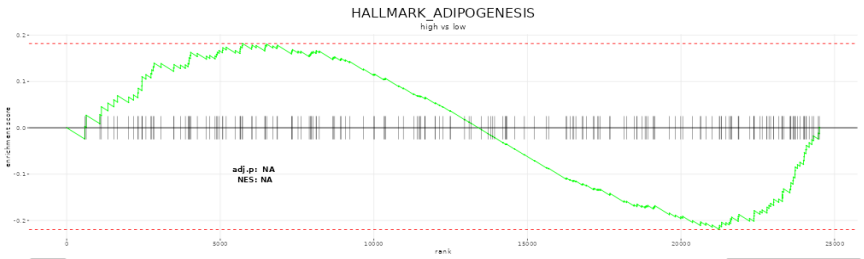
Pathway	Gene ranks	NES	pval	padj
HALLMARK_ALLOGRAFT_REJECTION		2.44	1.0e-50	2.3e-49
HALLMARK_INTERFERON_GAMMA_RESPONSE		2.47	1.0e-50	2.3e-49
HALLMARK_INFLAMMATORY_RESPONSE		2.24	6.5e-32	9.8e-31
HALLMARK_INTERFERON_ALPHA_RESPONSE		2.41	7.0e-30	7.9e-29
HALLMARK_COMPLEMENT		2.08	2.1e-22	1.8e-21
HALLMARK_TNFA_SIGNALING_VIA_NFKB		2.06	2.4e-22	1.8e-21
HALLMARK_IL2_STAT5_SIGNALING		1.97	5.2e-18	3.4e-17
HALLMARK_IL6_JAK_STAT3_SIGNALING		2.23	2.7e-17	1.5e-16
HALLMARK_KRAS_SIGNALING_UP		1.92	6.8e-15	3.4e-14
HALLMARK_APOPTOSIS		1.83	4.6e-10	1.9e-09
HALLMARK_OXIDATIVE_PHOSPHORYLATION		NA	NA	NA
HALLMARK_MYC_TARGETS_V1		NA	NA	NA
HALLMARK_GLYCOLYSIS		NA	NA	NA
HALLMARK_DNA_REPAIR		NA	NA	NA
HALLMARK_ADIPOGENESIS		NA	NA	NA
HALLMARK_NOTCH_SIGNALING		-0.70	9.5e-01	9.9e-01
HALLMARK_WNT_BETA_CATENIN_SIGNALING		-1.11	2.4e-01	4.0e-01
HALLMARK_MYC_TARGETS_V2		-2.60	2.8e-10	1.2e-09

Show 10 entries Search:

	pathway	leadingEdge
1	HALLMARK_INTERFERON_GAMMA_RESPONSE	GZMA
2	HALLMARK_INTERFERON_GAMMA_RESPONSE	IRF1
3	HALLMARK_INTERFERON_GAMMA_RESPONSE	CXCL9
4	HALLMARK_INTERFERON_GAMMA_RESPONSE	CCL5
5	HALLMARK_INTERFERON_GAMMA_RESPONSE	CXCL10
6	HALLMARK_INTERFERON_GAMMA_RESPONSE	IDO1
7	HALLMARK_INTERFERON_GAMMA_RESPONSE	GCH1
8	HALLMARK_INTERFERON_GAMMA_RESPONSE	IL2RB
9	HALLMARK_INTERFERON_GAMMA_RESPONSE	CD38
10	HALLMARK_INTERFERON_GAMMA_RESPONSE	CD274

Showing 1 to 10 of 720 entries Previous 1 2 3 4 5 ... 72 Next

B



Show 10 entries Search:

	pathway	leadingEdge
1	HALLMARK_ADIPOGENESIS	MDH2
2	HALLMARK_ADIPOGENESIS	SLC19A1
3	HALLMARK_ADIPOGENESIS	PEMT
4	HALLMARK_ADIPOGENESIS	SLC25A1
5	HALLMARK_ADIPOGENESIS	REEP6
6	HALLMARK_ADIPOGENESIS	CYC1
7	HALLMARK_ADIPOGENESIS	VEGFB
8	HALLMARK_ADIPOGENESIS	RETSAT
9	HALLMARK_ADIPOGENESIS	ESRRA
10	HALLMARK_ADIPOGENESIS	ADCY6

Showing 1 to 10 of 44 entries Previous 1 2 3 4 5 Next

Figure 2.10. GSEA Module Results. Graph showing the result of enrichment analysis of gene sets in the identified MSigDB Collection if "Top Pathways" is selected in the A panel. The table at the bottom of the graph was created to identify genes that were leading that most strongly contribute to the enrichment score for different comparisons. In the B part of this graph, the GSEA chart is created by selecting the "Specific Pathway" option. Just below the parts of the figure, it is possible to see the leading edge genes for that particular pathway.

2.8. Machine Learning Module

Machine learning (ML), a developing arm of artificial intelligence, refers to a range of applications and algorithms that aims to extract relevant and useful information from the obtained data. In the context of bioinformatics, ML is utilized to perform classification, prediction and feature selection from biological data (Shastry and Sanjay 2020). ML is a powerful tool for the establishment of the relationship between independent variables. Thus, apart from its usefulness in the areas of evolutionary biology or genomics, it offers a strong approach to modeling biological networks that can illuminate gene expression regulation in a cell at various metabolic states and disease conditions (C. Xu and Jackson 2019).

Regularized regression, a type of supervised machine learning technique, is a derivative of linear regression that allows one to simultaneously create a model and perform feature selection in the high dimensional data (Witten and Tibshirani 2009). It is optimized to minimize the sum of squared residuals by penalizing the generated model coefficient estimates. The penalty term is applied to the model equation to reduce model complexity and make the prediction with the limited mean squared error. By reducing model complexity and providing parameter shrinkage, regularized regression allows high prediction performance and prevents overfitting on large data sets (Ahrens, Hansen, and Schaffer 2020). Ridge, Lasso, and Elastic Net regression are all types of regularized regression methods and each has varying strengths depending on the structure and behavior of the data. Ridge regression shrinks the estimated coefficients without making them zero and penalizes correlated parameters in a similar fashion (Friedman, Hastie, and Tibshirani 2010). Lasso regression shrinks model coefficients down to zero and performs feature selection. Elastic Net regression can refer to a middle ground in between Ridge and Lasso regression and it is ideal for data sets in which the number of predictor variables significantly exceeds the number of samples and/or there are a high number of correlating variables. The degree of penalization for each approach can be optimized with cross-validation which is a widely accepted methodology for tuning parameter selection. TCGEx performs regularized regression using the algorithms described above on the transcriptomic profiling data obtained from the TCGA data portal. The machine learning module integrated into the application runs by the functions of the Glmnet package in R Programming language, which allows users to conduct and interpret various regularized

regression tasks. Glmnet provides a useful workflow that can perform generalized linear model fitting with Ridge, Lasso, and Elastic Net regression and also for penalization tuning, i.e. regularization parameter lambda optimization, via functions that perform cross-validation on a given data set for specified parameters. In addition, informative plots that describe model efficacy and demonstrate the amount of parameter shrinkage for varying sets of lambda values are also integrated into the ML module by utilizing the plotting functions of packages Glmnet and ggplot2. Accordingly, TCGEx compiles these tools in an efficient, user-friendly user interface that allows non-specialist researchers to utilize ML algorithms on the transcriptomic profiling data.

The machine learning module integrated into the TCGEx application requires a data matrix, the type of the regularized regression method to be applied to this data, and finally, response and predictor variables that are all determined by the user. Transcriptomic profiling data available on the ML module have the same structure that is described earlier in the introductory sections and therefore, the cancer subtype of interest should be specified before the analysis. Regularized regression method is controlled by the alpha parameter (α) as the mathematical equation for the algorithm requires. Accordingly, the choice of method is set by the input controller ranging from 0 to 1, where 1 corresponds to the lasso; 0 corresponds to the ridge; and the numbers in between determine elastic mixing. Response and predictor variables can be entered manually, uploaded via Excel files, or selected from the pre-formed sets obtained by the MSigDB database. The chosen response set can include a list of protein-coding and/or non-coding transcript names. TCGEx utilizes gene expression data to generate a response variable (also known as a dependent variable) whose linear relationship with specified independent variables is examined. This expression profile is computed by taking the row-wise mean of the normalized expression counts for a given set. For the predictor variables (independent variables) the application provides a diverse and extensive selection repertoire. The regularization path is computed with cross-validation and thus, the behavior of the model coefficients as the penalization parameter varies can be viewed. In addition, the TCGEx ML module presents users with an option to test model accuracy by allowing the splitting of the input into training and test subsets.

The interface generated for users to use the ML Module is shown in Figure 2.11. Slider input, which appears after the sample type selection, allows the user to eliminate genes that are expressed at low levels. The selection here specifies the maximum allowed

percentage of zero expression in a given gene. For instance, if this number is set to 50, genes that are not expressed in 50% or more of the samples in the analysis. The default value of 100 indicates that there is no filtering applied. We anticipate that, by removing lowly expressed genes, the model can perform more robustly because, in the presence of many non-expressors, samples with expression can appear spuriously associated with the response outcome. In the "Response Variables" panel, users can specify response variables by i) entering gene names manually, ii)using genes from MSigDB gene sets, or iii) selecting one of the previously calculated immune cell signatures. When multiple genes are entered or gene sets are selected, a single response variable is calculated by averaging the expression values. Users can type gene names in the box or upload an xlsx/xls file for manual gene selection. In the "Predictor Variable" panel, users can enter predictor variables either by entering them manually (users can type or upload a file), or using genes from MSigDB gene sets. Regularized regression models will examine the relationship between these predictor variables and the previously specified response variable. The user can switch to the regression tab after making their selections in this tab. Finally, the users can see the response and predictor variables that are determined with the panels opened on the right after the variable selections users have made.

The ML Module analysis results are shown in Figure 2.12. After specifying the response and predictor variables in the previous tab, users can determine the necessary parameters for regularized regression analysis in this regression tab. Users can perform regression on the whole data or split the data set into "training" and "test" subsets. Splitting allows examining the model accuracy through the mean-squared error. Users determine how much of the data will be used as a train set and how much will be reserved for the test set by using the data splitting toggle. Users can also choose the lambda value for coefficients at which the test set will be predicted. With the regression input, users can choose the method of regularized regression using this slider. alpha = 1 corresponds to LASSO regression where some coefficients will be shrunk (ie, penalized) to zero. alpha = 0 corresponds to Ridge regression where some coefficients will converge to (but not reach) zero. 0 < alpha < 1 corresponds to Elastic-Net regression where the penalty is a mixture of both. Finally, users can choose the lambda value at which the variable coefficients of the model are displayed. Lambda is the regularization parameter in the model. Minimum lambda is the value that gives the minimum cross-validation error in the regression. Lambda + 1 se is the value of lambda that gives the most regularized (ie.

more penalized and simpler) model where the cross-validation error is within the one standard error of the minimum. After all inputs are entered, the user can perform the analysis by clicking the "Train the model" button and downloading the results with the download button. Subsequently, a graph that will appear in the right bottom panel will show how the increasing levels of model penalization effect predictor coefficient shrinkage and the overall mean-squared error. The regression graph in the upper panel shows the changing coefficients of the predictors according to the increasing lambda values. In the table next to it, these predictors are ordered to be presented to the user according to their coefficient magnitude.

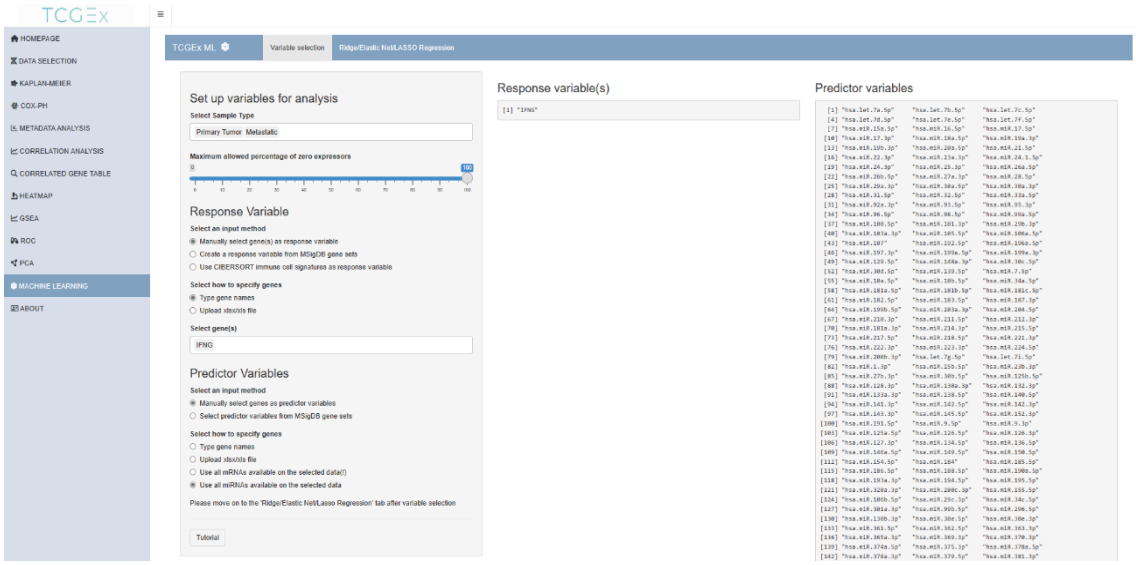


Figure 2.11. ML Module User Interface showing the input parameters to the algorithm

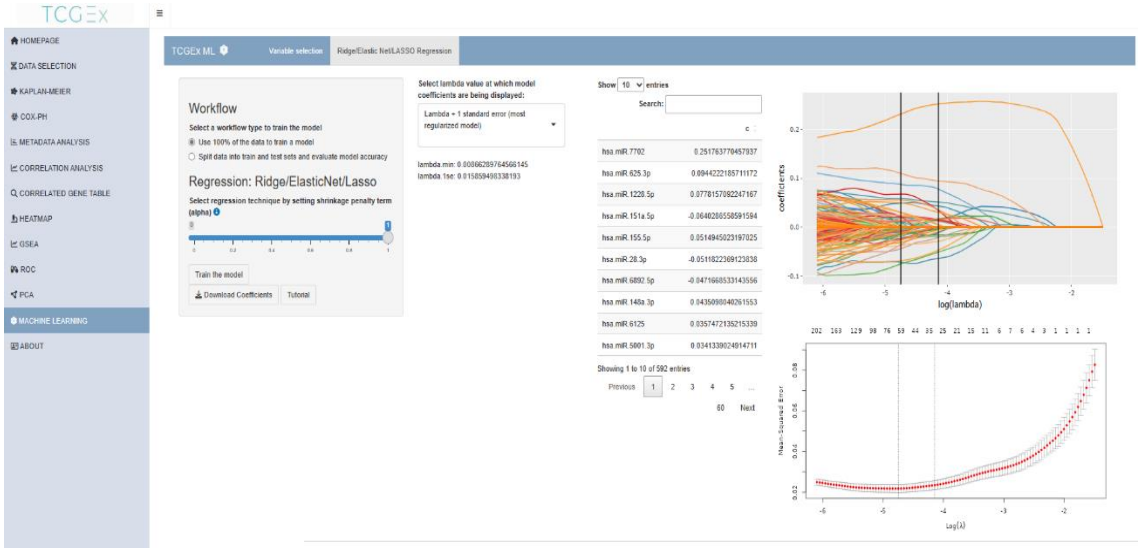


Figure 2.12. ML Module results tab showing penalized coefficients in different graphs

2.9. Kaplan-Meier Analysis Module

Survival analysis is used to determine how long it will likely be until a particular event occurs such as death or recurrence of cancer (Baek et al. 2021). The Kaplan-Meier (KM) survival curve functions as a time-varying estimator of the probability of surviving over a given period of time while accounting for several brief intervals of time. (Kishore, Goel, and Khanna 2010). It assumes that the event occurs at a certain time and the null hypothesis is that, the probability of survival is equal for data subsets at a given point in time. The KM survival curve is frequently used in cancer studies to illustrate how various variables including treatment, mutation type, and gene expression may affect survival (Huang et al. 2018; Wu et al. 2020). As a tool for comparing variables, this approach can facilitate the analysis of the large-scale data sets offered by TCGA with thousands of genes, miRNAs, and clinical metadata parameters (gender, BMI, ethnicity or genes, miRNAs categorized as low, high, etc.).

To facilitate the use of survival analysis in cancer studies, we have developed a user-friendly interface with a simple design and an orderly layout. The KM module allows users to target the desired subset of data by allowing the selection of sample types, features (thousands of genes and miRNAs), and covariates (genes, miRNAs, or clinical metadata). Both numerical and categorical features can be selected and users can define which data subsets are to be included in the analysis, enabling customization of research. Additionally, users can select all or separate cancer types for which survival graphs can be generated, enhancing the versatility of the module.

After the selection, the KM survival curve will be presented to the user. This tool allows customization of the figures with the addition of statistical hypothesis testing and color options for the user. This way, our tool creates publication-ready figures with a high level of customization, allowing researchers to present their findings accurately and effectively. Overall, the tool we developed offers numerous gene, miRNA, and clinical data selection options, advanced data selection, and a range of figure customization options.

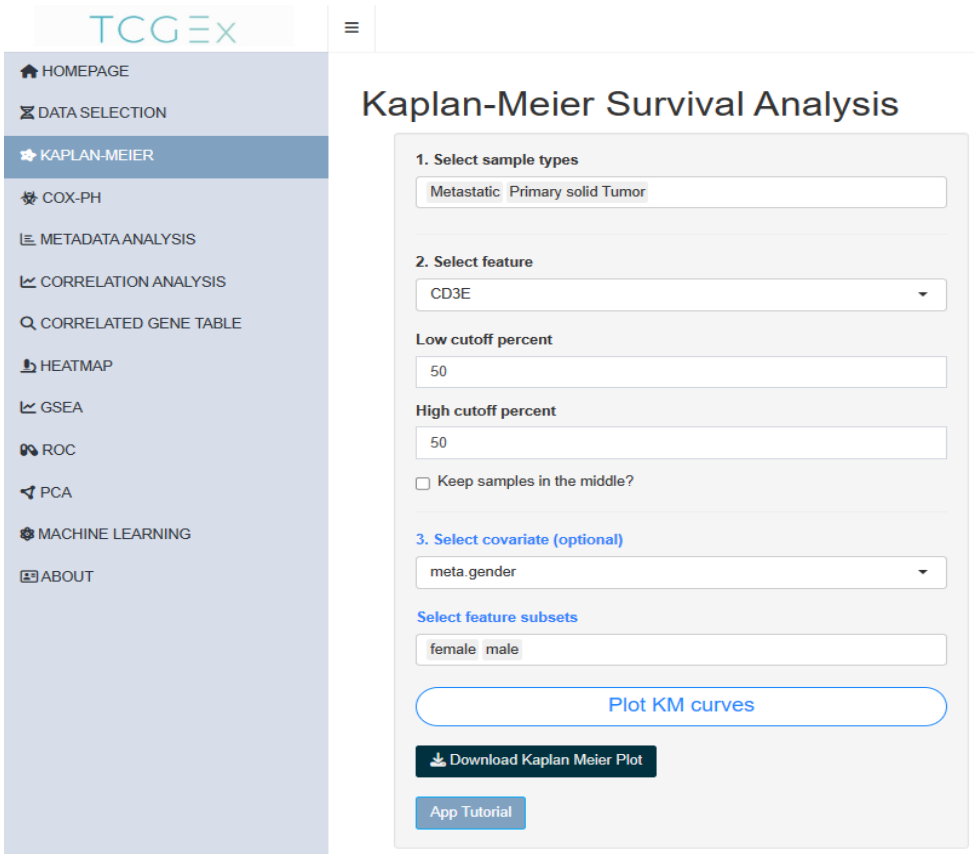


Figure 2.13. KM Module User Interface. In the first input, users can select the sample types (eg. primary and/or metastatic) to tailor the analysis to their needs. KM analysis is performed between groups of data. Subsequently, the user can select genes, miRNAs, or clinical metadata features. If the user selection is a categorical data type (eg. patient gender, tumor subtype), the user will be asked to select which subsets to be included in the analysis. If the user selection is a numerical data type (eg. gene expression), the user will be asked to define quantile cutoffs to categorize gene expression as "high" and "low". After defining the cutoff percentages, If the user's numeric categorization results in three groups (ie low, middle, high), the user can hide (default) or show the middle group in the graph by clicking the radio button. The user can add a covariate to the analysis and perform the survival analysis on the data subsets. Finally, the user can download the graph resulting from the analysis by pressing the download button.

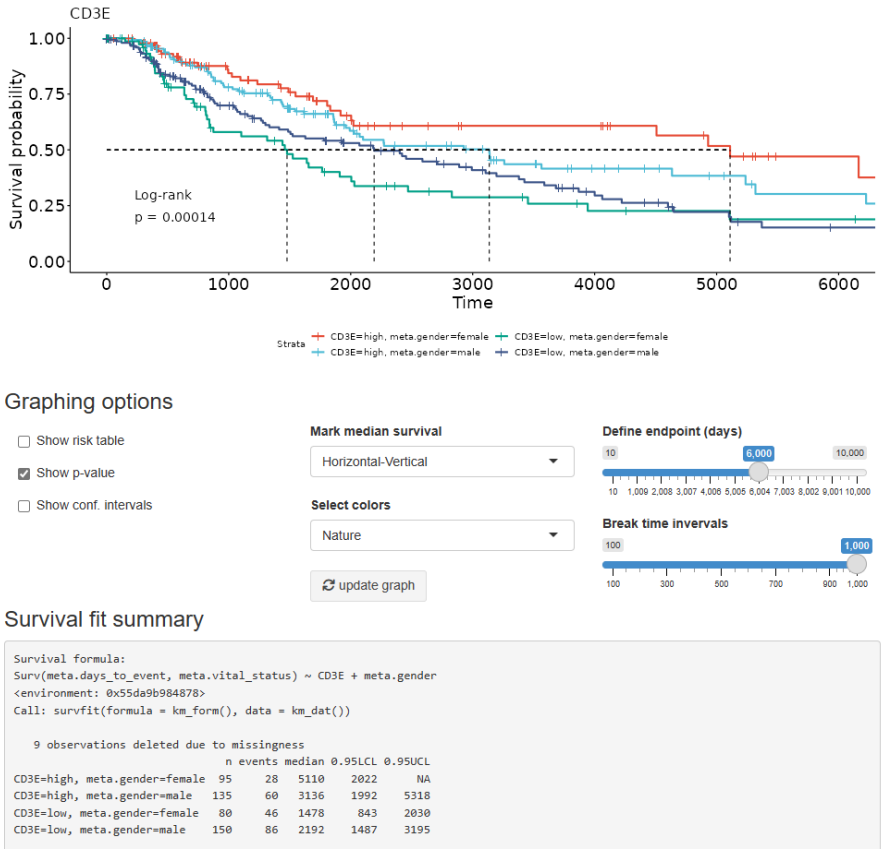


Figure 2.14. KM Module Results. KM plot generated for a user-selected gene across subset groups using clinical data. Users also can show the risk table on the graph. A table will be added below the KM curves showing the number of surviving patients at different time points. With the "Show p-value" option, users can show the log-rank p-value on the graph. If there are more than two groups in the analysis, the p-value is calculated by testing the null hypothesis that all the samples come from populations with identical survival. Users can select two specific data subsets to show pair-wise p-values. Confidence interval bands can be added to the graph. In addition to these, the user can change the plotted time interval and the breaks on the x-axis. This does not affect the results of survival analysis. Again, with the options below the graph, the user can plot dashed lines to highlight median survival in data subsets and change the color palette of the graph. Finally, there is the survival fit summary at the bottom of the chart. With this summary, the user can find the formula for the survival analysis and detailed statistical information about the research results.

2.10. Cox Proportional Hazards Model Survival Analysis Module

The Cox proportional hazard regression model is the most often employed survival model in the medical field where survival probabilities are modeled by one or several covariates (Kamphorst et al. 2022; Deo, Deo, and Sundaram 2021; Baek et al. 2021). In this model, no assumption is made on the baseline hazard function form because it is a semi-parametric model. Thus, the proportional risk of death is calculated as a factor of time. Patients' expected survival can be predicted using the coefficients of covariates. According to the proportionality independent of time, which is one of the assumptions of the model, the effect of the factors does not change over time. That is, the hazard ratio of a factor remains constant over time. For example, this assumption is valid if the effect of a drug does not change over time.

As mentioned before Cox module also allows users to focus on data subsets. The user can select features (thousands of genes and miRNAs) and covariates (genes, miRNAs, or clinical data). The Cox module presents an option where the impact of multiple variables can be observed together. This module allows for the simultaneous adjustment of multiple risk factors. In addition, the user can see the Log-Rank (p-value) for both individual predictors and global models to determine their statistical significance. With the confidence intervals given, how the risk changes across the strata can be seen. Survival formula, coefficients, number of events, concordance, likelihood test, Wald test, and Chi-squared test values are visualized to describe details of the Cox proportional hazards analysis. This module allows Cox proportional hazards to be carried out with several different options, so the research can be tailored to meet the interests of the researcher and present reliable results in a short period.

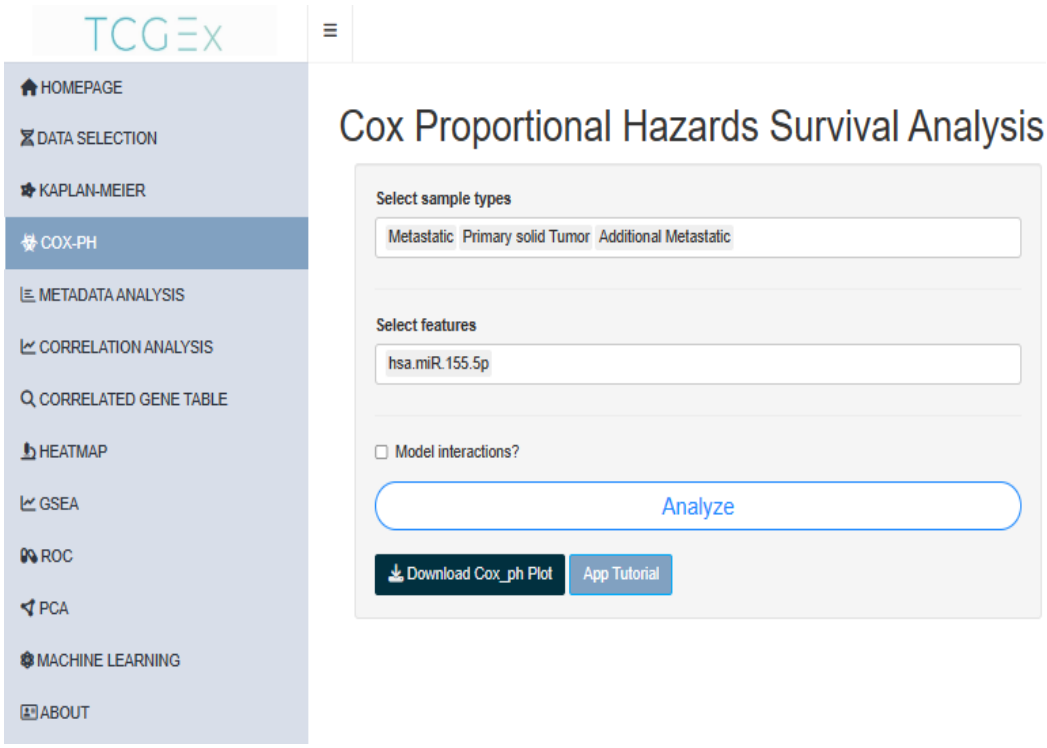
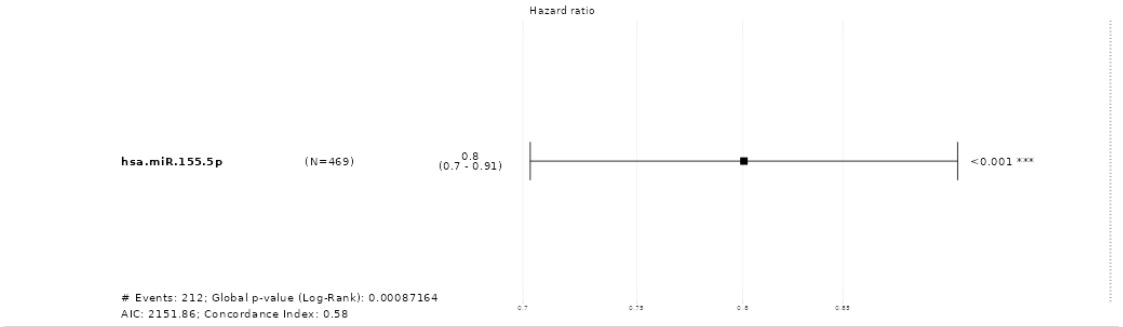


Figure 2.15. Cox-Ph Module User Interface. As in every module, the user can select the sample types (eg. primary and/or metastatic samples) to target specific data subsets in the analysis. Then, the user can select one or more features desired to analyze (eg. genes or clinical metadata). If a single feature is selected, univariate Cox-Ph analysis is performed. When two or more features are selected, multivariate Cox-Ph analysis is performed where the effects of individual features are reported along with the overall effects. Users can also perform model interactions between features (optional). This more complex modeling allows examining whether covariates have an impact on each other's effect. The user can, for instance, investigate whether “Gene_A” has a different survival impact in males and females by specifying “Gene_A*meta.gender”.




```
######################################################################
Survival formula
Surv(meta.days_to_event, meta.vital_status) ~ hsa.miR.155.5p
<environment: 0x55da51989068>
######################################################################


Call:
coxph(formula = cox_form(), data = cox_dat())

  n= 438, number of events= 212
   (31 observations deleted due to missingness)

                 coef exp(coef) se(coef)      z Pr(>|z|)
hsa.miR.155.5p -0.22253   0.80049  0.06617 -3.363 0.000771 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
hsa.miR.155.5p    0.8005      1.249    0.7031    0.9113

Concordance= 0.584  (se = 0.023 )
Likelihood ratio test= 11.08  on 1 df,   p=9e-04
Wald test            = 11.31  on 1 df,   p=8e-04
Score (logrank) test = 11.35  on 1 df,   p=8e-04


######################################################################
Testing proportional hazards assumption.
p-values less than 0.05 indicates the proportionality assumption is violated.
Consider using time-dependent modeling
              chisq df    p
hsa.miR.155.5p  1.47  1 0.23
GLOBAL          1.47  1 0.23
######################################################################
```

Figure 2.16. Cox-Ph Module Results. The graph obtained as a result of the Cox Proportional Hazards analysis shows the hazard ratio for the features specified. In the script just below, users can learn to assess the statistical significance of the model.

2.11. Correlation Analysis Module

The Correlation Analysis Module offers the opportunity to compare the expression levels of two genes from selected cancer data. The correlation analysis is visualized via scatter plots. Users can customize the features of the scatter plot via their choices. This module provides information about each patient when users hover their mouse over each point if they prefer. Thus, users can observe the outlier patients in the plot. Users also can see the regression line on the gene comparison plot optionally. With these features, researchers can detect tendencies via the regression line and patients who deviate from the tendency, and find out their gender, race, and patient ID. In addition, this

module makes it possible to identify the differences between the classes of clinical features such as tumor type and mutation type by dividing the samples into groups through an optional facet parameter. The Correlation Analysis module also displays the p-value value and Pearson's correlation coefficient on the graph depending on the user's request (Garcia-Diaz et al. 2017). The P-value shows whether the linear regression slope is statistically non-zero. Researchers can identify the strong or weak correlations between expressions of two genes which especially play important roles in certain immune processes such as immune checkpoint blockade via correlation plot analysis. The main advantage of this module is that ggplot2 is enhanced with the ggiraph package for flexible user input and informative data points.

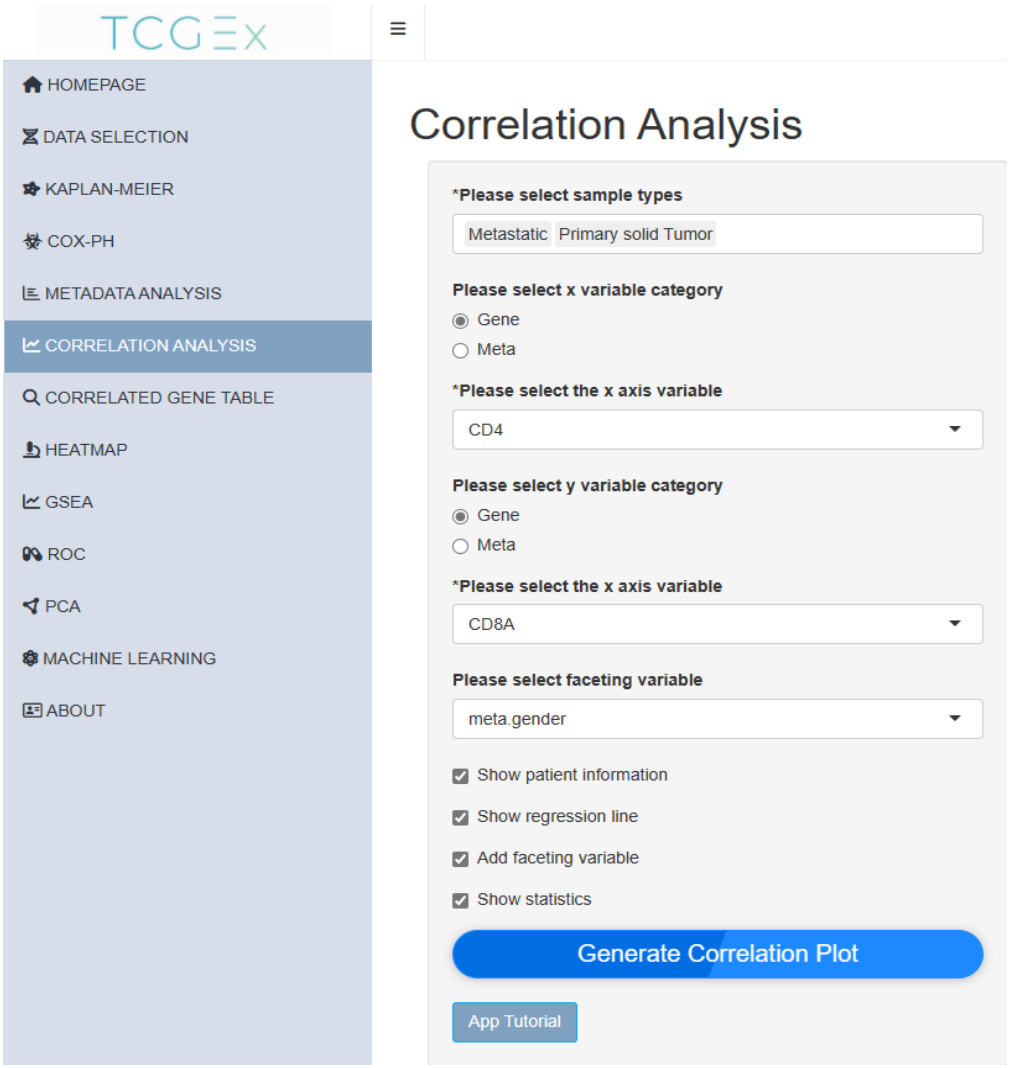


Figure 2.17. Correlation Analysis Module User Interface. After selecting the sample type, the user determines the variables to place in the x and y axes. These variables

can be gene or clinical data. Then, users can also add a faceting variable and regression line to the graphic to be created. Besides, the user can view the correlation analysis statistically. If "Show patient information" is selected, hovering over each point shown in the graph will show which patient the data came from.

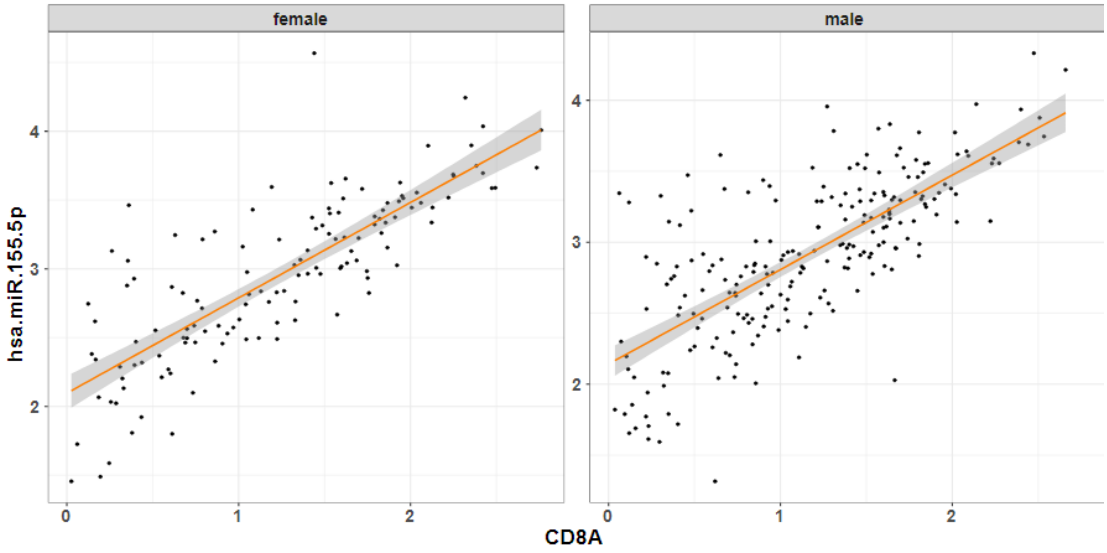


Figure 2.18. Correlation Analysis Module Results. Graph showing the correlation of cytotoxic T cell marker CD8A and antitumor-immunity regulator miRNA-155 in melanoma patients. If the users wish, can download it by hovering over the graphic and use the results obtained from it in their studies.

2.12. Metadata Analysis Module

In the Metadata Analysis Module, users have the opportunity to explore the relationships between a gene and a categorical clinical parameter. While performing this analysis, the user can choose to include or omit specific data subsets of the selected clinical feature. Optionally, the generated plot can be examined in more depth by using an extra faceting variable. For instance, the change of the "CD8A" gene in metastatic and primary cancer tissue samples in melanoma cancer can be visualized per mutation subtypes. Users have flexible options for customizing the appearance of the plots. Jitter, mean, and standard error can be added to the graphic.

Another feature that comes with this module is to determine and show how statistically significant the results are. For comparisons the user, can specify a parametric T-test and a nonparametric Wilcoxon test. Then, the process continues by choosing one of the single reference or pairwise comparison methods. Thus, the user can determine how significant the result is by seeing the adjusted or normal p values as numerical or symbols on the resulting graph. The ggpubr package was used to generate the metadata correlation plots.

The interface designed for users to use the Metadata Analysis Module is shown in Figure 2.19. The user is expected to select the categorical variables to be placed on the x-axis. After selecting the categorical variables, the feature to be placed on the y-axis is entered by the user. In addition to all these, the users can specify the scientific journal color palettes of their graphics on the graph creation panel and download them in a manner suitable for publication in a journal.

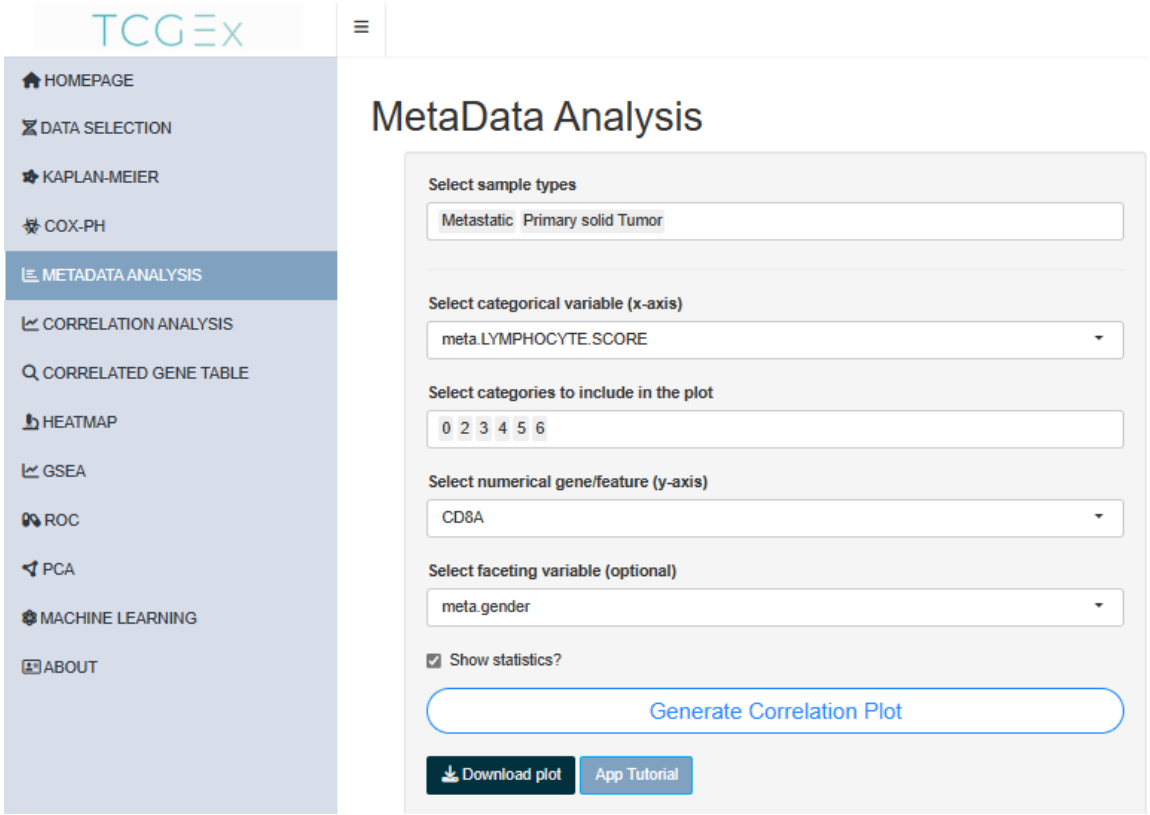


Figure 2.19. Metadata Analysis Module User Interface.

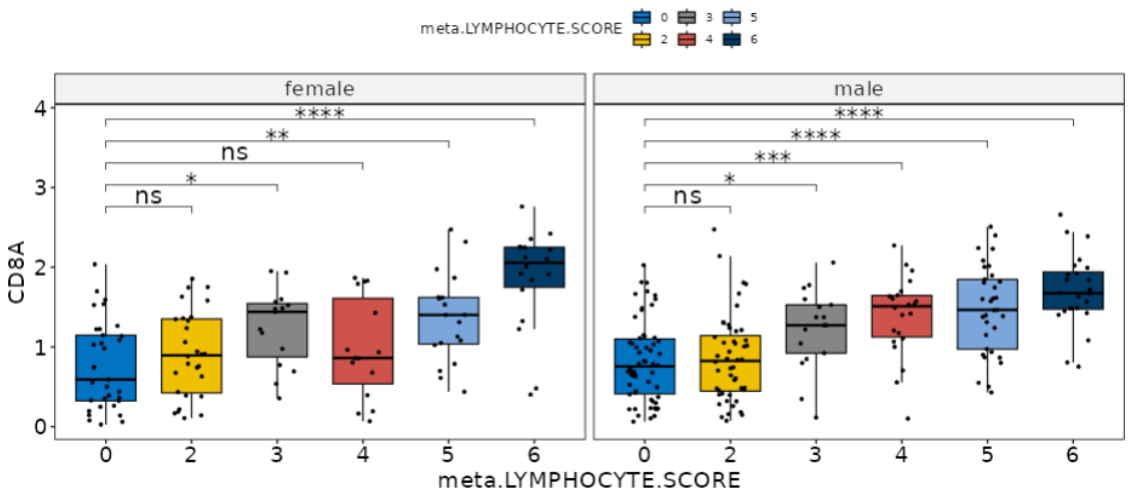


Figure 2.20. Metadata Analysis Module Results. Metadata analysis enables a selected gene to be examined through different clinical parameters, and while doing so, ensure graphics are ready to be published in journals by presenting how statistically significant it is. Here is a box-plot graph showing the change of the CD8A gene selected by the user depending on the lymphocyte score which includes a combination of different parameters, such as the number or density of areas of lymphocyte in the tumor tissue in male and female patients in melanoma.

2.13. Correlated Gene Analysis Module

Correlated Genes Analysis (CGA) is a useful tool to detect correlations between a selected gene and thousands of genes in the selected cancer project. Users can focus the analysis on specific sample types as mentioned before. The highest positively and negatively correlated genes are shown in a table in the numbers specified by the user. This module also provides the number of samples lacking expression of selected genes. The correlated genes plot is beneficial for users to highlight the most and least correlated genes. In this plot, correlation coefficients are shown with colored circles. Thus, users can visualize the correlations among desired genes conveniently. This tool also provides method options, “Pearson” and “Spearman”, for correlation coefficient calculation. Since normalized datasets are used, significant differences between the two methods are not

expected, but in some cases rank-based Sperman's correlation may be preferred (Bland and Altman 1999). This module can be conveniently used to visualize miRNAs and their targets to reveal possible regulatory interactions. For instance, one can visualize miR-145 against its supposed targets such as SOX2 controlling the regulation of colon cancer stem cells (Syeda et al. 2020; Yu et al. 2015). This module also can be used to enhance the results of our other modules. For example, upregulated methylation-related genes can be detected from heatmap analysis and correlations between these genes can be studied via correlation analysis (Liu et al. 2020).

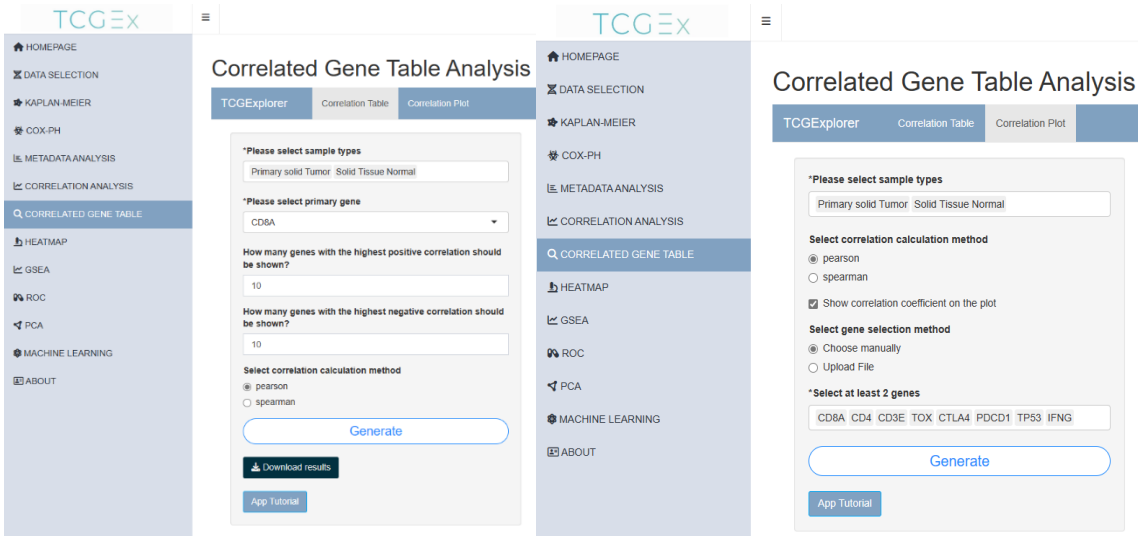


Figure 2.21. Correlated Gene Analysis Module User Interface. In this "Correlation Table" tab, users can select a gene and tabulate its top positively and negatively correlated genes. Users can also visualize correlations in the "Correlation Plot" tab. In the first tab, users select their genes of interest. Users can then change the number of best positively and negatively correlated genes they want to appear in the table. Users can specify how the correlation should be calculated and download the obtained table. In the second tab, the user continues the analysis by choosing which method to use while calculating the correlation. Then the user determines whether the correlation coefficients will be displayed on the plot or not. Finally, users determine the genes they want to include in the correlation plot manually by selecting them or uploading them to the system and getting the result.

A

Number of samples where the chosen gene is not expressed

0

Show 10 entries Search:

	Genes	p_value	correlation_coefficient	zero_patient
1	CD2	1.61e-195	0.909	0
2	SLA2	1.16e-194	0.908	0
3	NKG7	2.1e-194	0.908	0
4	CD3D	4.1e-190	0.904	0
5	CXCR6	1.64e-188	0.902	0
6	GZMH	1.12e-182	0.897	0
7	CD3E	2.57e-175	0.889	0
8	TRBC2	3.74e-166	0.879	0
9	SH2D1A	4.11e-165	0.878	0
10	CD3G	6.46e-162	0.874	0

Showing 1 to 10 of 20 entries Previous 1 2 Next

B

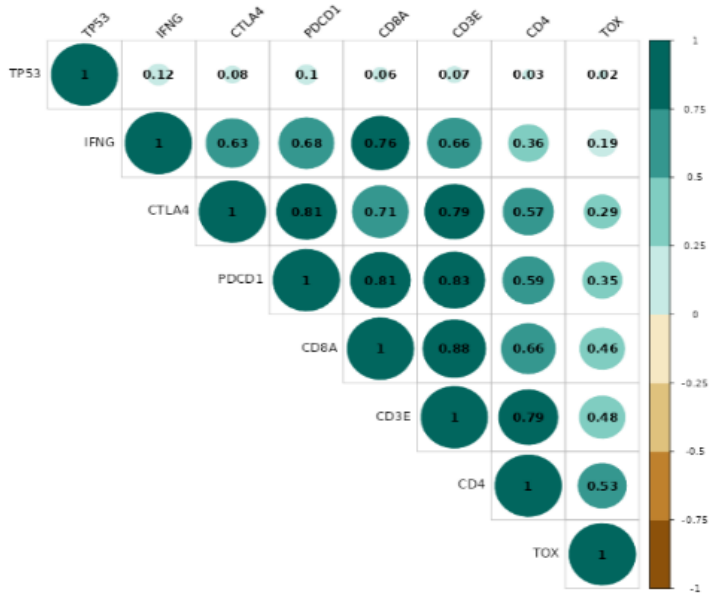


Figure 2.22. Correlated Gene Analysis Module Results. The table in the A part shows the analysis result, which includes the genes with the highest and least correlations of a gene selected from the user-subset data, and which was created by the user-determined correlation calculation method. The table includes the correlation coefficients of the genes as well as the p-values showing how statistically significant they are. In the correlation plot in the B part, the correlation of user-specified genes with each other is visualized. It is important to detect positive and negative correlations between dozens of genes at a single glance.

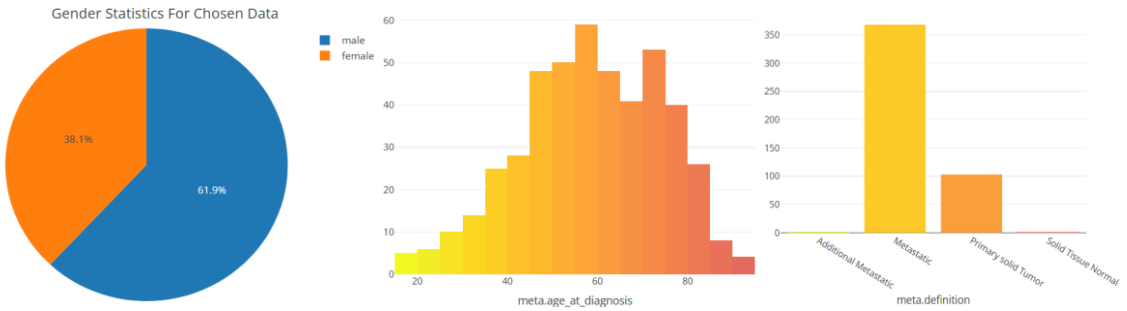
CHAPTER 3

USE CASE SCENARIO

Skin Cutaneous Melanoma (SKCM) is a highly aggressive form of cancer that originates from melanocytes. It is presently the most fatal type of skin cancer, and its incidence has been on the rise in recent times (Leonardi et al. 2018). Protracted exposure to the sun is often associated with the development of this cancer, and if left untreated, it can rapidly metastasize throughout the body, posing a significant threat to the patient's life (Hartman and Lin 2019). Therefore, early diagnosis and treatment are paramount in the effective management of SKCM.

While therapeutic approaches have been applied recently to improve survival rates in melanoma patients significantly, a considerable number of patients do not respond to this type of treatment (Y. Chen et al. 2022). It has become increasingly crucial to develop personalized treatment methods alongside studies aimed at identifying the subtypes of the disease. It is essential to understand why some patients do not respond to treatment and to determine in advance which patients will respond to specific treatments. Through the analysis of data from 473 melanoma patients in TCGA, studies have been conducted to identify mutation types in melanoma (Akbani et al. 2015). These studies have made it possible to classify the disease into subtypes. At this point, TCGEx offers scientists the opportunity to conduct comprehensive analyses of the desired cancer type.

After selecting the cancer project, TCGEx provides descriptive statistics about the data sets, allowing the user to have a general idea of the data. Many data such as age, gender, demographic distribution, mutation types and subgroups of patients in the selected cancer type can be examined by the user and then analyses can be performed based on these. Of note, after selecting melanoma as the cancer type, “Metastatic”, “Additional Metastatic” and “Primary Solid Tumor” were used as sample types in all analyses performed in the remainder of the chapter.



Gender Distribution **Age Distribution** **Sample Type Distribution**

Figure 3.1. Informative Statistics about SKCM Patients. Using the TCGEx data selection module, the user can perform cancer selection and obtain descriptive information about that cancer. For instance, Melanoma patients are mostly diagnosed at an advanced stage. The sample type distribution in the figure reveals this.

In the article published by The Cancer Genome Atlas Network in 2015 (Weinstein et al. 2013), mutations in important genes observed in melanoma patients were revealed and disease subtypes were created according to mutation types. These mutation subtypes differ significantly in survival and other prognostic indicators. With TCGEx, it is possible to perform detailed analyzes and better understand these subgroups.

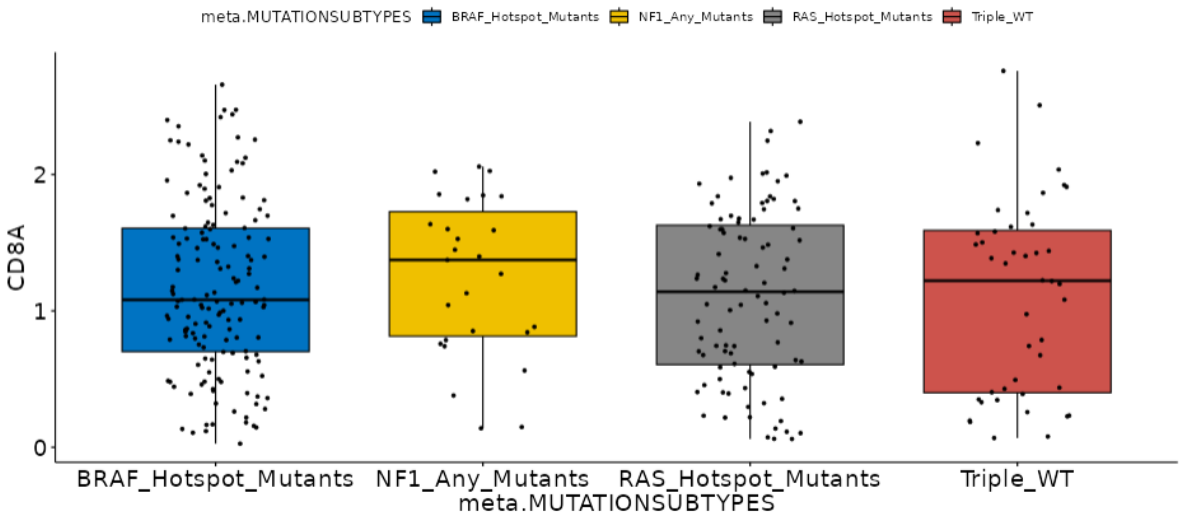


Figure 3.2. Profiling of subgroups with mutation types in terms of CD8A. In this graphic, there is profiling of subgroups of patients with different mutation types in cancerous tissues of melanoma in terms of CD8A expression, one of the important markers of the cytotoxic T cells which are the key regulators for anti-tumor immunity, using TCGEx.

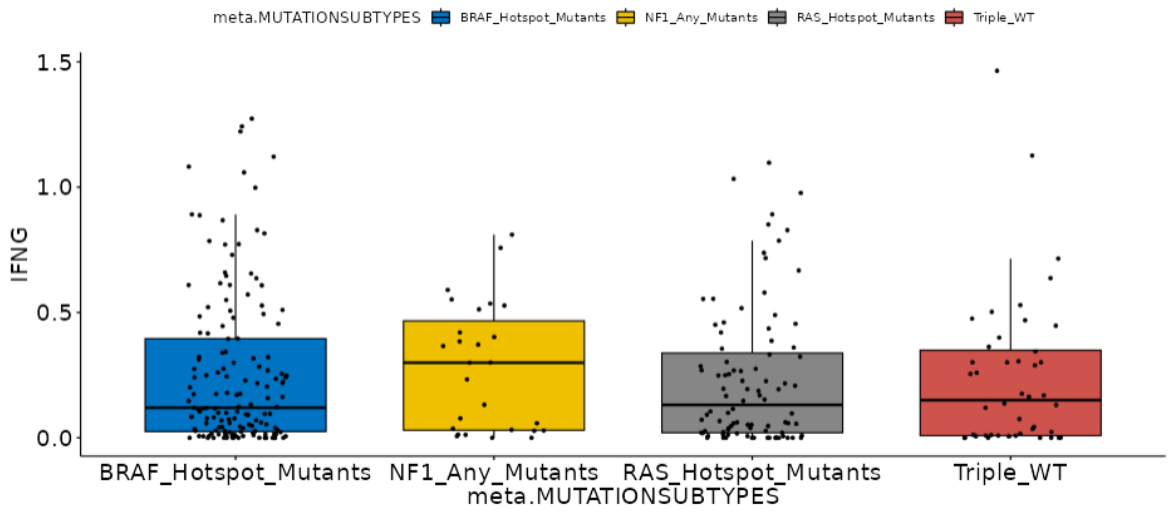


Figure 3.3. Profiling of subgroups with mutation types in terms of IFNG. While this graph shows the interferon-gamma expression levels of melanoma mutants, it is also striking that the data here are in parallel with the CD8 gene expression levels of the mutants in the previous graph. Similar expression patterns suggest that T cell presence is commonly associated with more IFNg and its effector function.

Melanoma tumors display considerable heterogeneity, exhibiting a wide range of biological traits, metastatic capabilities, risks of survival, and responsiveness to various treatments. Hence, the categorization of melanoma tumors into different clinically distinguishable and prognostic subtypes becomes essential to ensure precise diagnosis, guide appropriate treatments, and facilitate the development of subtype-specific drugs (Netanely et al. 2021). In addition to BRAF, NF1, RAS, and Triple Wild Type mutants, 3 more subtypes have been defined in the Genomic Classification of Cutaneous Melanoma article (Akbani et al. 2015), and the interferon-gamma response of these subtypes is a distinguishing feature for each other. These approaches are extremely important as they enable personalized solutions by classifying melanoma into subtypes, distinguishing patients from each other, and determining the treatment method to be applied. By using TCGEx modules, it is possible to examine many features that distinguish these subtypes from each other. For example, the effect of these subtypes on patient survival was also demonstrated by Kaplan-Meier analysis in Figure 3.4.

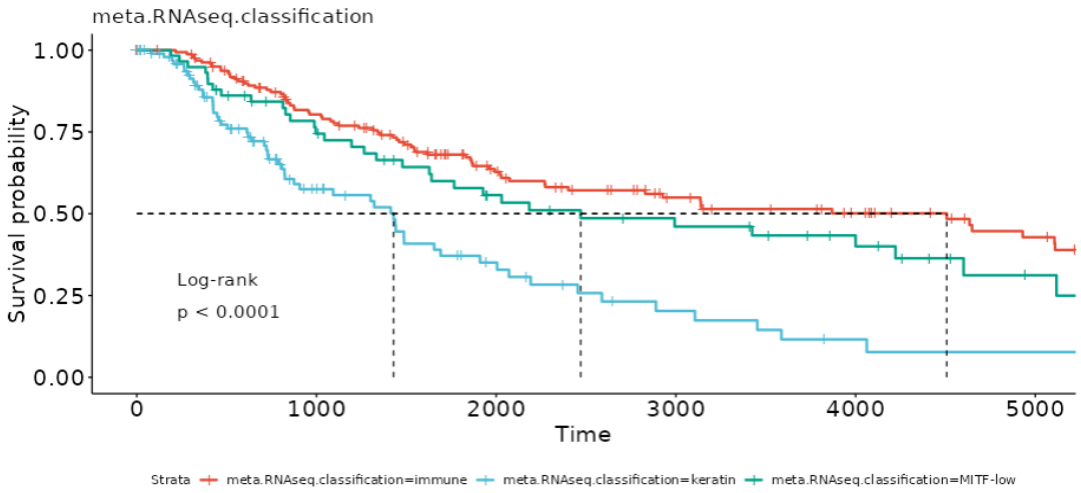


Figure 3.4. The Kaplan-Meier Curve of subgroups in Melanoma. The KM Curve constructed using TCGEx shows that patients belonging to the "immune" group from the "immune", "keratin" and "MITF-low" subgroups in melanoma have significantly higher survival than the other groups.

Increasing evidence suggests that expression levels of miRNAs may have a crucial role in human tumors and could even be key players in the response to therapy (B. Q. Chen et al. 2022). miRNAs can exhibit specificity for particular tumors or be linked to immune cells that infiltrate the tumor microenvironment. Different disease subsets may have varying miRNA expressions, which can originate from tumor cells or infiltrating immune cells, such as T cells. Through heatmap analysis using TCGEx, it is possible to identify specific miRNAs that are upregulated in the "immune" subgroup, which has better survival (Akbani et al. 2015). In addition, expression levels of genes that play an important role in the interferon-gamma response can be compared in cases where miRNAs hypothesized to be immune-related are high or low subsets. It has been shown in Figure 3.5. that genes involved in the interferon-gamma response, which is important for inflammation and cell-mediated immunity, are upregulated at the intersection of high subsets of specific miRNAs such as miR-142, miR-150, and miR-155 and immune subtype. Also, these miRNAs are good prognosticators along with the genes involved in immune processes in the selected cancer type. This analysis can be performed simultaneously with the "immune", "keratin", and "MITF-low" subgroups, providing valuable insights into the molecular mechanisms underlying cancer and potentially leading to new therapeutic strategies.

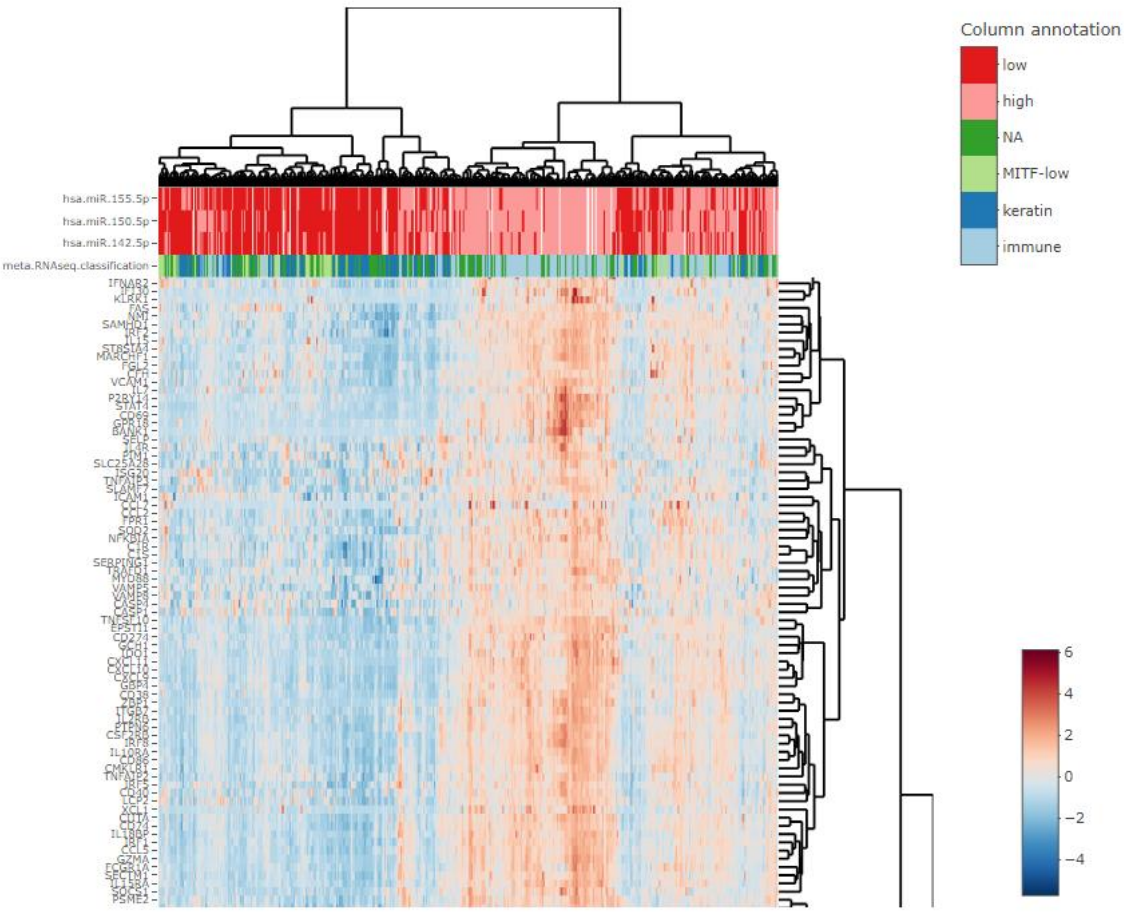


Figure 3.5. Heatmap analysis results of SKCM patients data. Heatmap analysis results were created by using data from SKCM patients and an interferon-gamma response gene set. In the graph, it is seen that the genes involved in the interferon-gamma response are more upregulated in areas where the immune subgroup, which shows better survival, and the regions with high levels of certain miRNAs, which are thought to regulate the immune response. With the Heatmap Module of TCGEx, users have also options to perform column-wise or row-wise scaling as used in this figure.

When comparing miRNAs that are upregulated in the "immune" subgroup using heatmap analysis to the Correlated Genes Analysis module provided by TCGEx, it was observed that miR-155 and a few other miRNAs that could be potential predictors of anti-tumor immunity are highly correlated with the CD8A gene, a marker for cytotoxic T cells (CTLs). This finding suggests a potential regulatory role for miRNAs in the immune response and highlights the importance of studying their interactions with immune cell

markers in cancer. When reviewing the literature, the effect of miR-155 on malignant melanoma cell migration and invasion remains largely unclear. The origin of these miRNA signals remains uncertain, as they could emanate from either the tumor cells themselves or the immune cells present in the tumor microenvironment. To explore and resolve these potential sources, a more detailed investigation of the tumor microenvironment using scRNAseq with higher resolution is warranted. However, it is worth noting that these specific miRNAs have been linked to a heightened immune response within the tumor microenvironment, suggesting their potential role in modulating the immune activity in that context (Jayawardana et al. 2016).

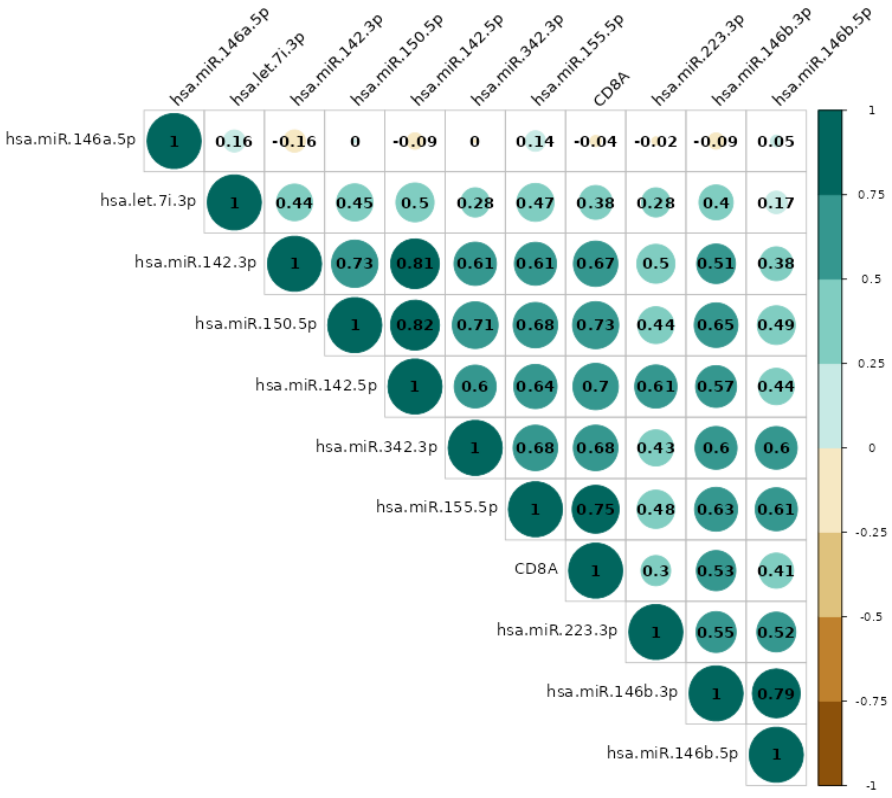
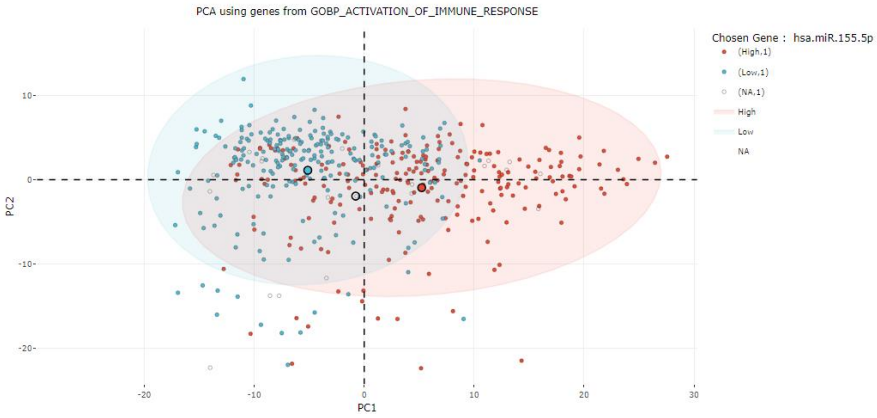


Figure 3.6. Relationship between miRNAs that are upregulated within the "immune" Subclass in melanoma**.** The graph is illustrating the relationship between miRNAs that are upregulated within the "immune" subclass in melanoma, as identified through the TCGEx Heatmap module, and the CD8A gene, which serves as a marker for CTLs.

It can be hypothesized that miR-155 has a regulatory effect on the immune response in melanoma data. The results of PCA analysis using the activation of the

immune response gene set showed a significant differentiation in patient subsets identified as high and low based on miR-155 expression level. These results provided insight into the role of miR-155 in the immune response. Then, the data of patients in 33 cancer types in TCGA are combined, together with the ability of TCGEx to examine different cancer types together to investigate whether miR-155 will have a similar role in only melanoma or all cancer data. When this combined cancer data is examined specifically for activation of immune response genes using PCA, another analysis method in this bioinformatics tool, a clear separation is observed in patients with high and low miR-155 as shown in Figure 3.7. Indicating its potential role as a biomarker in immune response regulation. These findings suggest that bioinformatics tools such as TCGEx may facilitate the identification of potential therapeutic targets and biomarkers for cancer diagnosis and treatment.

A



B

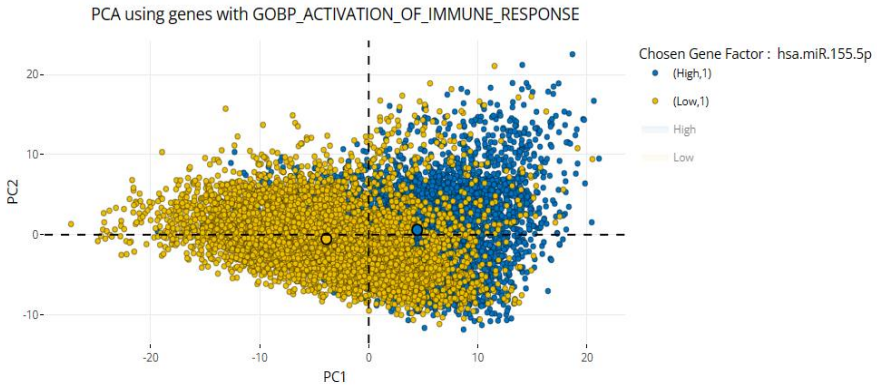


Figure 3.7. The distinction between patients overexpressing and underexpressing miR-155. In the A part, the PCA result demonstrates the potential role of miR-155 in immune activation in melanoma. The graph on B part, created using the

PCA module from TCGEx, examines all cancer types for activation of the immune response, showing the clear distinction between patients overexpressing and less than miR-155.

Continuing the analysis of melanoma using the Correlated Genes Analysis module, it is possible to generate a list of genes that are most highly correlated with miR-155. Upon further investigation of this list, it becomes evident that the genes with which this miRNA shows high correlation are key genes including CD8A, CCL5 and NKG7 that are crucial for the immune system (Korbecki et al. 2020; Aldinucci, Borghese, and Casagrande 2020). This finding highlights the potential role of miR-155 in regulating the immune response in melanoma and suggests that uncovering these genes and/or miRNAs could be a promising therapeutic approach in the treatment of this cancer type. These results demonstrate the potential of bioinformatics tools such as the Correlated Genes Analysis in identifying novel targets for cancer therapy.

	Genes	p_value	correlation_coefficient
1	CCL5	6.59e-92	0.778
2	NKG7	2.24e-86	0.763
3	CD2	1.68e-85	0.76
4	TRAC	2.89e-84	0.756
5	TRBC2	1.87e-83	0.754
6	CD8A	7.18e-83	0.752
7	MIR155HG	9.23e-83	0.752
8	CD3D	1.22e-82	0.752
9	CD8B	1.35e-82	0.752
10	TIGIT	3.33e-82	0.75

Figure 3.8. Correlated Genes Analysis results show that the genes most correlated with miR-155 in SKCM patients are those involved in the immune response.

Subsequently, using Gene-to-Gene Correlation Analysis, the absolute correlation between miR-155 and CD3E and TNF genes, which play a regulatory role in immune cells, was observed in the generated graphs. These findings suggest a potential regulatory

role of miR-155 in immune cell function, specifically in the regulation of immune regulatory gene expression such as CD3E and TNF. The results of this study highlight the importance of miR-155 as a potential therapeutic target for immune-related diseases, including cancer.

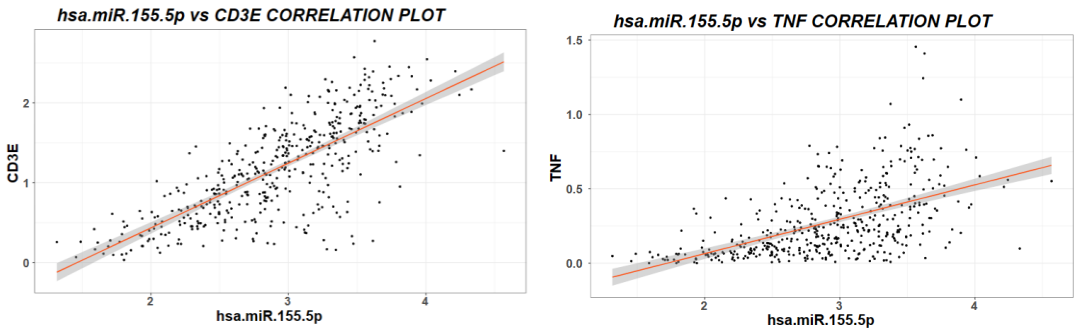


Figure 3.9. Correlation of miR-155 with the immunomodulatory genes CD3E and TNF. Graphs created using the Correlation Analysis module in TCGEx, showing the correlation of miR-155 with the immunomodulatory genes CD3E and TNF.

In order to obtain more detailed insights about mir-155, examining the pathways where miR-155 is enriched on TCGEx's GSEA module may be a meaningful next step. The TCGEx GSEA module scores and prioritizes gene sets based on gene expression data. It then compares the gene sets and associates them with a particular phenotype so that the relationship of the gene sets to the phenotype can be evaluated statistically. It then calculates the enrichment score for each gene set, taking into account the sequence of genes and the overall gene expression profile. This enrichment score gives information about which gene sets the feature searched by the user is associated with. The results of this analysis show that mir-155 is highly enriched in pathways such as activation of immune response, adaptive immune response, cellular response to cytokine stimulus, cell activation and inflammatory response. These results can increase our knowledge of mir-155's functions and its potential clinical applications.

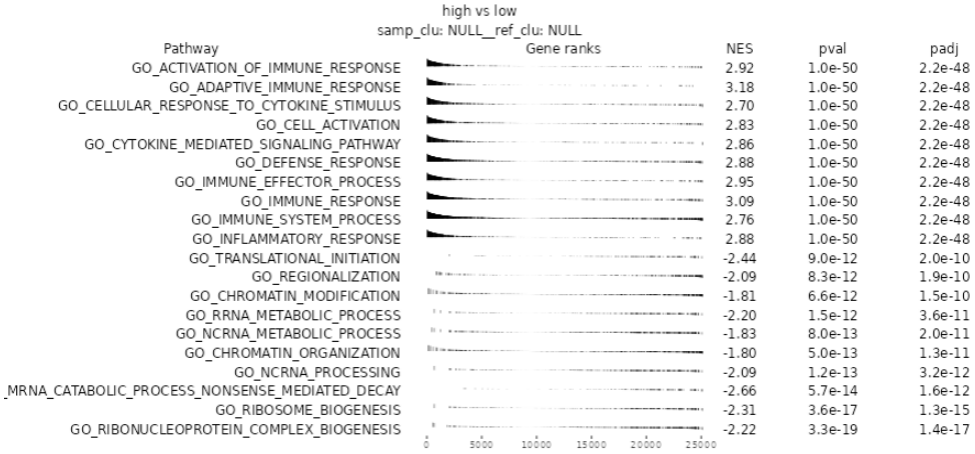


Figure 3.10. GSEA analysis results that includes information about pathways where miR-155 is enriched and depleted in melanoma using the TCGEx Gene Sets Enrichment Analysis module by using the MSigDB ontology gene sets.

In particular, Interferon Gamma Response, which is one of the remarkable pathways in which miR-155 is enriched, can be used to examine False Positive and True Positive Fractions with ROC Curve Analysis against CD8A and IFNG, another module offered by TCGEX. Gene set that contained genes up-regulated in response to IFNG from MSigDB were used to perform this analysis. It is obvious from these analysis results that miR-155 performs close to markers such as CD8A and IFNG used to determine IFNG Response. With the expansive range of analyses offered by TCGEx, scientists have the option to test their hypotheses in many different ways.

A **B**

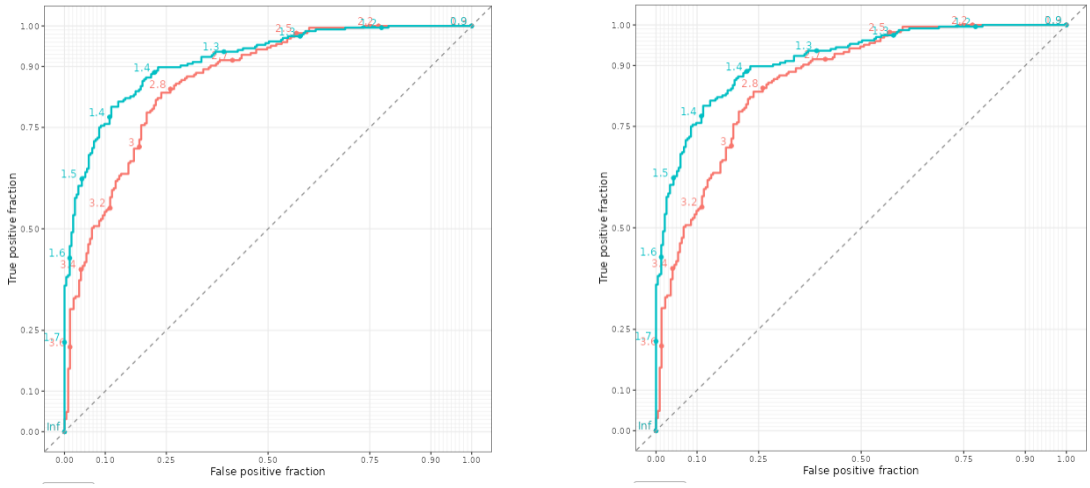


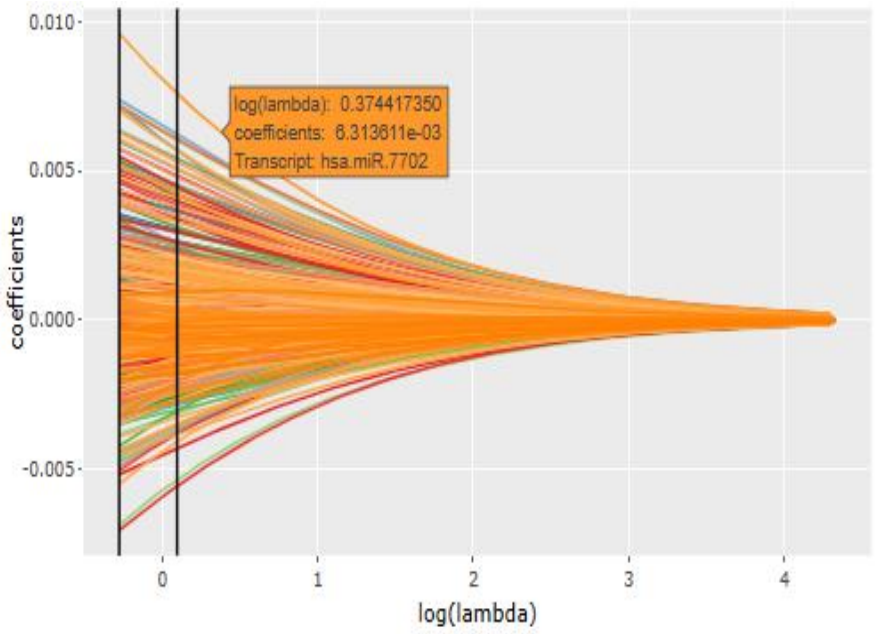
Figure 3.11. True and False Positive Fractions of the IFNG Response Geneset and hypothetically potential biomarker miR-155 in melanoma. The A panel shows

the True and False Positive Fractions defined by the IFNG Response gene set (blue) and mir-155 (red) used against the CD8A. The B panel shows the True and False Positive Fractions of the IFNG Response gene set (blue) and miR-155 (red), which are similarly used to predict IFNG, a biomarker of innate and adaptive immunity. When the results are taken together, miR-155 gives promising results in predicting important genes that play a role in immunity and is a potent biomarker candidate.

Identifying miRNAs that are correlated with differential survival and immune involvement in cancer can lead to the discovery of novel therapeutic targets. Machine learning algorithms, such as linear regression-based models, have been used in several studies to identify the strongest predictors of various clinical parameters, including protein-coding genes and noncoding RNAs (Albaradei et al. 2021).

In this study, we present a streamlined interface that enables researchers to use Lasso, Ridge, and Elastic Net machine learning algorithms with custom response and predictor variables. This interface can be used to identify miRNAs that are strongly associated with immune signatures within the tumor.

A



(figure 3.12. continued on the next page)

B

hsa.miR.7702	0.00951038323482098
hsa.miR.1228.5p	0.00737136535955316
hsa.miR.6803.3p	0.00720989105420552
hsa.miR.155.5p	0.00716959466362634
hsa.miR.1976	0.00711324587089296
hsa.miR.5586.5p	0.00701351081314841
hsa.miR.342.5p	0.00633144251377122
hsa.miR.625.5p	0.00624241570610634
hsa.miR.6842.3p	0.00601373108169037
hsa.miR.1228.3p	0.00591066212291546

Figure 3.12. Potential miRNAs to explain and predict the CD8 T Cell Score. Graph A and Table B were obtained with the TCGEx Machine Learning module. The graph on the A part of this figure is the Ridge Regression model. Each line represents a predictor. The graph allows us to determine the predictors with the highest score among the coefficients shrinking to zero. With this model, the potential miRNAs to explain and predict the CD8+ T cell score were determined. In the table on the B part, there is an ordered list of miRNAs with the highest potential in predicting the CD8+ T cell score from Thorsson's cybersort data, which is a selected response variable. The score next to the best predictor miRNAs shows their coefficients in the B part of this figure.

Utilizing the ML module offered by TCGEx to identify miRNAs that best predict the CD8 T cell score reveals that miR-155 holds strong potential to predict T cells. In addition to miR-155, miRNAs such as miR-7702, with high coefficient scores, may deserve further investigation. Subsequently, potential candidates that best predict lymphocyte infiltration signature score and interferon-gamma response can be identified using machine learning. As seen in the graphs and tables, miR-7702 and a few other miRNAs are potential predictors here. In this example, TCGEx is used to conveniently perform sophisticated analyses suggesting that it can facilitate cancer research.

Table 3.1. The best predictor miRNAs for Lymphocyte Infiltration Score and Interferon-Gamma Response, respectively in melanoma patients. The numbers next to the best predictors exhibit their coefficients.

hsa.miR.5586.5p	0.11453274905518
hsa.miR.342.5p	0.114457660501537
hsa.miR.342.3p	0.101425872357191
hsa.miR.150.3p	0.0971754512666782
hsa.miR.1976	0.0967520132846892
hsa.miR.7702	0.0853980545866301
hsa.miR.142.5p	0.0845230748793213
hsa.miR.6842.3p	0.084227377202264
hsa.miR.146b.3p	0.0821334519439558
hsa.miR.29c.3p	0.079848410976221

hsa.miR.3614.3p	0.155297188378112
hsa.miR.3614.5p	0.10413515628511
hsa.miR.5586.5p	0.0618105781364399
hsa.miR.4658	0.0593328092937445
hsa.miR.1228.3p	0.0581247091551678
hsa.miR.16.1.3p	0.0554059693015674
hsa.miR.4728.3p	0.0499327237093806
hsa.miR.4999.5p	0.0487362318570367
hsa.miR.7702	0.0470331247150946
hsa.miR.342.5p	0.0455915884795359

As mentioned before, more than 150 immune expression signatures were scored in the study of Thorsson et al. These scores have facilitated our understanding of the immune system and the mechanisms involved. In Table 3.1. shows the best potential predictor miRNAs associated with Lymphocyte Infiltrated Signature and Interferon-Gamma Response scores, respectively, were analyzed with the TCGEx Machine Learning Module.

Many studies have been carried out in the literature on miRNAs that are thought to have important regulatory functions such as miR-155, miR-142, and miR-150 in these tables (Mahesh and Biswas 2019; Sur et al. 2020; Shabani et al. 2019). However, very little research has been done on some of the candidates in the list of potential predictor miRNAs that we obtained as a result of these ML algorithms. We highlight the importance of further study and characterizing the other strong candidates found here. The projects to be realized with the potential candidates obtained will play an important role in closing the gaps in the literature.

CHAPTER 4

DISCUSSION AND CONCLUSION

Compared to mentioned existing tools that analyze and visualize cancer data, TCGEx offers a wide range of analysis methods, integrating machine learning algorithms into cancer data. This tool makes it possible to perform comparative analyses using multiple cancer data simultaneously, the ability to produce images ready for publication in the article stands out in a fast, robust, extended functionality, and presents a user-friendly and comprehensive approach to cancer research. In the table below, where the features of TCGEx and other current solutions are compared, it is possible to see the features that this new application offers scientists in the field of research.

Table 4.1. Cross-property analysis of TCGEx and other applications.

Features/Tools	TCGExplorer	UCSC Xena	GEPIA 2	Standford TCGA-CE	Regulome Explorer	Onc-DB	Web-TCGA	OncoLnc	CBioPortal
Kaplan Meier Survival Analysis	✓	✓	✓	×	×	✓	×	✓	✓
Cox Proportional Hazards	✓	×	×	×	×	×	×	✓	×
Metadata Analysis	✓	✓	×	✓	✓	✓	✓	×	×
Correlation Analysis	✓	✓	✓	×	✓	✓	✓	×	✓
Hierarchical Clustering (Heatmap)	✓	×	×	×	×	×	×	×	✓
Gene Set Enrichment Analysis	✓	×	×	×	×	×	×	×	×
Receiver Operating Characteristic Analysis	✓	×	×	×	×	×	×	×	×
Dimensionality Reduction (PCA)	✓	×	✓	×	×	×	×	×	×
Machine Learning Algorithms	✓ (Ridge-Lasso-Elastic Net)	×	×	✓ (Elastic Net)	×	×	×	×	×
Subset-Specific Analysis	✓	×	×	×	×	×	×	×	✓
Integration of Multiple Datasets	✓	×	✓	×	×	×	×	×	×

In addition, TCGEx enables comprehensive analysis with the wide-range analyses modules as shown in Figure 4.1. and without any confusion with its user-friendly, simple interface. It provides precise results with its customized input options and step-by-step clear instructions that scientists will carry out their research. With the tutorials presented to the user, it is possible to carry out the desired analyses, and it can enable the users to explore new hypotheses and ideas on-the-fly during the analysis.

With this open-source and publicly available application, it is possible to analyze high-dimensional cancer data from many different perspectives in seconds, while at the same time, the user is allowed to explore possible common and distinctive aspects of different cancer types. This tool, which complements existing applications such as cBioPortal and TCGA-Assembler (Y. Zhu, Qiu, and Ji 2014; Cerami et al. 2012) and can be a guide for future studies in this field, offers the opportunity to be used equally for all large or small-scale studies. In order to better understand how TCGEx works and to validate its superior capabilities, users can visit the website in the supplementary material or install it on their computer and continue to analyze without the need for the internet. (https://tcgex.iyte.edu.tr)

TCGEx can play an essential role in the writing and execution of many research projects. Potential biomarkers and immunotherapeutic targets identified using TCGEx modules can generate preliminary data for scientists. Studies can be carried out to obtain new scientific insights as a result of supporting these data with the literature. In all these processes, this bioinformatics tool will be a resource researchers can easily access and save time. Being able to aggregate and analyze cancer types can help users understand common patterns in different cancers. Thus, the common mechanisms underlying processes such as metastasis, which is one of the most important causes of cancer becoming fatal, can be illuminated. Furthermore, examining patient groups in subsets can easily be used in in-silico studies, which are necessary for the development of personalized treatment methods.

With its compact structure, TCGEx is suitable software for development and updates. The data in the TARGET database (Y. Wang et al. 2020), which contains data on childhood cancers, can be integrated into the bioinformatics application, which currently includes only the cancer data in the TCGA database. Thus, with the in-depth analysis of childhood cancers, it will be possible to compare them with other cancer types. In addition, the inclusion of mouse tumor data in The Mouse Tumor Biology Database

(Krupke et al. 2008) into the TCGEx environment may allow users for cross-species analysis. In this way, scientists can have foresight about the future of the study before starting their research. In the long term, it may be possible for the user to analyze the data in the GEO Database or the user's data with more than 10 methods in TCGEx. Thus, increasingly accumulating high-throughput cancer data can be easily examined and the emergence of unknowns about cancer can be accelerated.

TCGEx bioinformatics tool offers subset-specific analysis as well as the ability to group different RNAs such as miRNAs from different RNAseq data and analyze them according to the user's request. Modules such as PCA and ML allow users to conduct more extensive research on miRNAs. In this way, the place of miRNAs in tumor immunology can be more clearly defined. In the close future, TCGEx will have the ability to perform other ncRNAs group-specific analyses such as lincRNA. In this way, our knowledge about non-coding ncRNAs will increase rapidly.

TCGEx, which offers many users from all over the world the opportunity to connect at the same time and perform bioinformatic analysis, also stands out with its powerful server and infrastructure. TCGEx server, established at Izmir Institute of Technology, offers users a fast, effective, and robust analysis opportunity with its multiple virtual machines. In addition, the application, which has effective security certificates, offers users a quite safe and free research opportunity. This powerful infrastructure can also be upgraded depending on the increase in usage and load in the future, so users will have an uninterrupted cancer research experience.

It can be accessed from TCGEx (https://tcgex.iyte.edu.tr) address and also has the feature of working in the docker environment. With this feature, users can download the TCGEx application image to their computers and run it locally with the Docker application. Thus, users can easily continue their analysis locally in case of any server issues.

TCGEx, which makes it possible to examine dozens of different cancer types by combining many analysis methods, can become an extremely valuable resource for cancer research with new modules to be added. One of the new features to be added to the application can be a module that can perform differential gene expression analysis using raw cancer data. In this way, genes that show the most different expressions in different samples can be detected. This feature will also provide scope to support other modules. Another feature that could be added would be to develop a scoring function that would

describe the correlation of a selected gene with a gene set in correlation plots. This way, the relationship between individual genes and gene sets associated with specific phenotypes can be explored.

Eventually, TCGEx arises as a unique research tool with features such as subset-specific analysis, integration of multiple datasets, machine learning algorithms, and 10 robust bioinformatics analysis modules, while at the same time, promising for in-silico cancer research with its modularized structure that is open to upgrades.

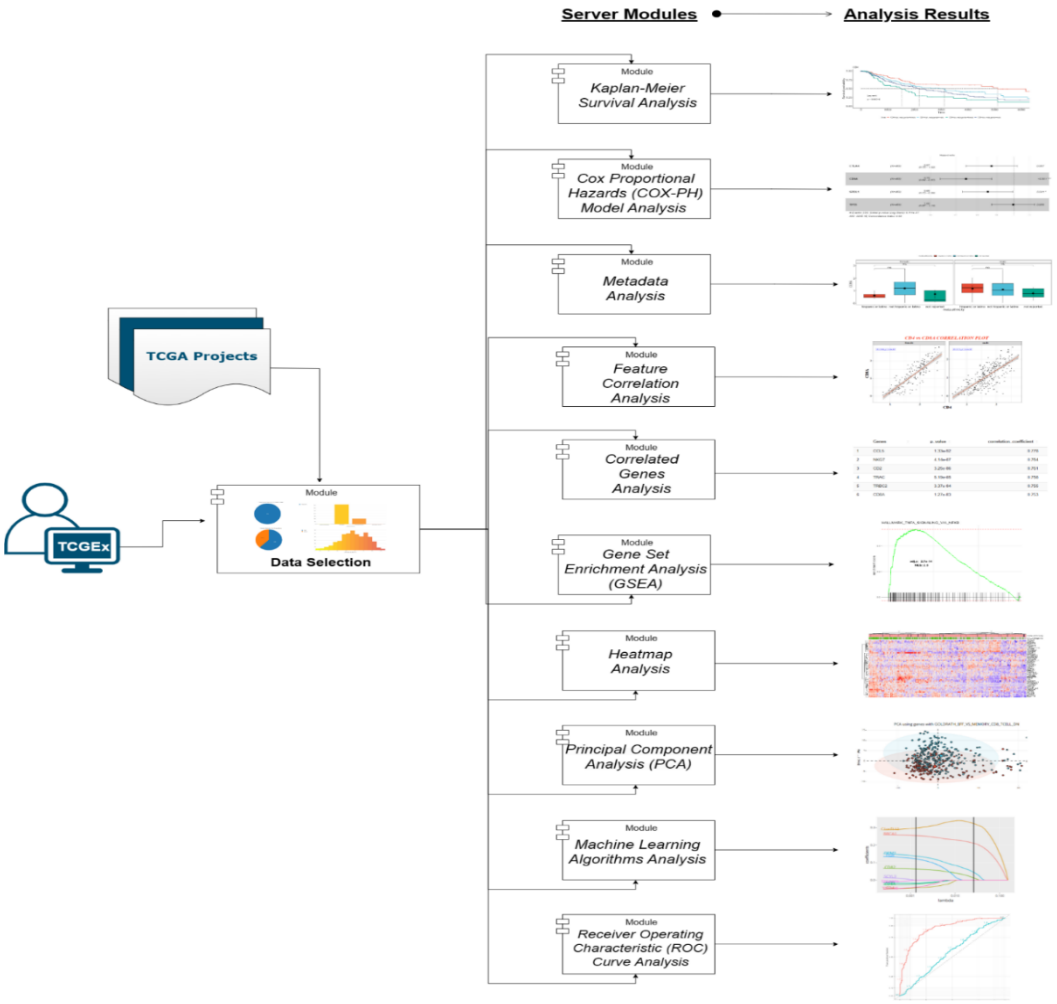


Figure 4.1. TCGExplorer's wide analysis range and general workflow chart. In the TCGEx application, one or more cancer data to be analyzed by the user is selected first. And then it is possible to perform the following analyses: Principal Component Analysis, Kaplan-Meier Analysis Module, Receiver Operating Characteristic Curve Analysis, Heatmap Analysis, Cox Proportional Hazards Model Analysis Module, Gene Set Enrichment Analysis, Correlation Analysis, Correlated Gene Table Analysis, Metadata Analysis Module, and Machine Learning Algorithms Analysis.

REFERENCES

Afshar, Parnian, Arash Mohammadi, Pascal N. Tyrrell, Patrick Cheung, Ahmed Sigiuk, Konstantinos N. Plataniotis, Elsie T. Nguyen, and Anastasia Oikonomou. 2020. ”Deep Learning-Based Radiomics for the Time-to-Event Outcome Prediction in Lung Cancer.” *Scientific Reports 2020 10:1* 10 (1): 1–15. https://doi.org/10.1038/s41598-020-69106-8.

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer. 2020. “Lassopack: Model Selection and Prediction with Regularized Regression in Stata.” 20 (1): 176–235. https://doi.org/10.1177/1536867X20909697.

Akbani, Rehan, Kadir C. Akdemir, B. Arman Aksoy, Monique Albert, Adrian Ally, Samirkumar B. Amin, Harindra Arachchi, et al. 2015. “Genomic Classification of Cutaneous Melanoma.” *Cell* 161 (7): 1681–96. https://doi.org/10.1016/J.CELL.2015.05.044.

Albaradei, Somayah, Maha Thafar, Asim Alsaedi, Christophe Van Neste, Takashi Gojobori, Magbubah Essack, and Xin Gao. 2021a. “Machine Learning and Deep Learning Methods That Use Omics Data for Metastasis Prediction.” *Computational and Structural Biotechnology Journal* 19 (January): 5008–18. https://doi.org/10.1016/J.CSBJ.2021.09.001.

Aldinucci, Donatella, Cinzia Borghese, and Naike Casagrande. 2020. “The CCL5/CCR5 Axis in Cancer Progression.” *Cancers* 12 (7): 1–30. https://doi.org/10.3390/CANCERS12071765.

Ekiz, Hüseyin Atakan., Thomas B. Huffaker, Allie H. Grossmann, W. Zac Stephens, Matthew A. Williams, June L. Round, and Ryan M. O’Connell. 2019. “MicroRNA-155 Coordinates the Immunological Landscape within Murine Melanoma and Correlates with Immunity in Human Cancers.” *JCI Insight* 4 (6). https://doi.org/10.1172/jci.insight.126543.

Baek, Eu Tteum, Hyung Jeong Yang, Soo Hyung Kim, Guee Sang Lee, In Jae Oh, Sae Ryung Kang, and Jung Joon Min. 2021. “Survival Time Prediction by Integrating Cox Proportional Hazards Network and Distribution Function Network.” *BMC Bioinformatics* 22 (1). https://doi.org/10.1186/S12859-021-04103-W.

Balacescu, Ovidiu, Daniel Sur, Calin Cainap, Simona Visan, Daniel Cruceriu, Roberta Manzat-Saplacan, Mihai Stefan Muresan, Loredana Balacescu, Cosmin Lisencu, and Alexandru Irimie. 2018. “The Impact of MiRNA in Colorectal Cancer Progression and Its Liver Metastases.” *International Journal of Molecular Sciences 2018, Vol. 19, Page 3711* 19 (12): 3711. https://doi.org/10.3390/IJMS19123711.

Bilotta, Maria Teresa, Antonella Antignani, and David J. Fitzgerald. 2022. “Managing the TME to Improve the Efficacy of Cancer Therapy.” *Frontiers in Immunology* 13 (October): 954992. https://doi.org/10.3389/FIMMU.2022.954992/BIBTEX.

Bland, J Martin, and Douglas G Altman. 1999. “Measuring Agreement in Method Comparison Studies.” *Statistical Methods in Medical Research* 8 (2): 135–60. https://doi.org/10.1177/096228029900800204.

Cai, Yimei, Xiaomin Yu, Songnian Hu, and Jun Yu. 2009. “A Brief Review on the Mechanisms of MiRNA Regulation.” *Genomics, Proteomics & Bioinformatics* 7 (4): 147–54. https://doi.org/10.1016/S1672-0229(08)60044-3.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. “The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” *Cancer Discovery* 2 (5): 401. https://doi.org/10.1158/2159-8290.CD-12-0095.

Chen, Bao Qing, Mihnea P. Dragomir, Chen Yang, Qiaoqiao Li, David Horst, and George A. Calin. 2022. “Targeting Non-Coding RNAs to Overcome Cancer Therapy Resistance.” *Signal Transduction and Targeted Therapy* 7 (1). https://doi.org/10.1038/S41392-022-00975-3.

Chen, Yong, Jingqin Zhong, Wei Sun, Wangjun Yan, Chunmeng Wang, Wanlin Liu, Xinyi Lin, and Zijian Zou. 2022. “BRAF Inhibitor Resistance in Melanoma: Mechanisms and Alternative Therapeutic Strategies.” *Current Treatment Options in Oncology* 23 (11): 1503–21. https://doi.org/10.1007/S11864-022-01006-7/TABLES/2.

Coradduzza, Donatella, Sara Cruciani, Caterina Arru, Giuseppe Garroni, Aleksei Pashchenko, Mosab Jedea, Silvia Zappavigna, et al. 2022. “Role of MiRNA-145, 148, and 185 and Stem Cells in Prostate Cancer.” *International Journal of Molecular Sciences 23, Page 1626* 23 (3): 1626. https://doi.org/10.3390/IJMS23031626.

Deng, Mario, Johannes Brägelmann, Joachim L. Schultze, and Sven Perner. 2016. “Web-TCGA: An Online Platform for Integrated Analysis of Molecular Cancer Data Sets.” *BMC Bioinformatics* 17 (1): 1–7. https://doi.org/10.1186/S12859-016-0917-9/FIGURES/4.

Deo, Salil Vasudeo, Vaishali Deo, and Varun Sundaram. 2021. “Survival Analysis—Part 2: Cox Proportional Hazards Model.” *Indian Journal of Thoracic and Cardiovascular Surgery* 37 (2): 229. https://doi.org/10.1007/S12055-020-01108-7.

Ding, Shuning, Xiaosong Chen, and Kunwei Shen. 2020. “Single-Cell RNA Sequencing in Breast Cancer: Understanding Tumor Heterogeneity and Paving Roads to Individualized Therapy.” *Cancer Communications* 40 (8): 329–44. https://doi.org/10.1002/CAC2.12078.

Dragomir, Mihnea, Ana Carolina P. Mafra, Sandra M.G. Dias, Catalin Vasilescu, and George A. Calin. 2018. “Using MicroRNA Networks to Understand Cancer.” *International Journal of Molecular Sciences 2018, Vol. 19, Page 1871* 19 (7): 1871. https://doi.org/10.3390/IJMS19071871.

Dunn, Gavin P., Lloyd J. Old, and Robert D. Schreiber. 2004. “The Three Es of Cancer Immunoediting.”*104803* 22 (March): 329–60. https://doi.org/10.1146/ANNUREV.IMMUNOL.22.012703.104803.

English, Patricia A., J. Andrew Williams, Jean François Martini, Robert J. Motzer, Olga Valota, and Richard E. Buller. 2016. “A Case for the Use of Receiver Operating Characteristic Analysis of Potential Clinical Efficacy Biomarkers in Advanced Renal Cell Carcinoma.” *Future Oncology (London, England)* 12 (2): 175–82. https://doi.org/10.2217/FON.15.290.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/JSS.V033.I01.

Garcia-Diaz, Angel, Daniel Sanghoon Shin, Blanca Homet Moreno, Justin Saco, Helena Escuin-Ordinas, Gabriel Abril Rodriguez, Jesse M. Zaretsky, et al. 2017. “Interferon Receptor Signaling Pathways Regulating PD-L1 and PD-L2 Expression.” *Cell Reports* 19 (6): 1189–1201. https://doi.org/10.1016/J.CELREP.2017.04.031.

Gehlenborg, Nils, and Bang Wong. 2012. “Points of View: Heat Maps.” *Nature Methods* 9 (3): 213. https://doi.org/10.1038/NMETH.1902.

Goldman, Mary J., Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, et al. 2020a. “Visualizing and Interpreting Cancer Genomics Data via the Xena Platform.” *Nature Biotechnology 2020 38:6* 38 (6): 675–78. https://doi.org/10.1038/s41587-020-0546-8.

Hanahan, Douglas, and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The Next Generation.” *Cell* 144 (5): 646–74. https://doi.org/10.1016/J.CELL.2011.02.013.

Hartman, Rebecca I., and Jennifer Y. Lin. 2019. “Cutaneous Melanoma—A Review in Detection, Staging, and Management.” *Hematology/Oncology Clinics of North America* 33 (1): 25–38. https://doi.org/10.1016/J.HOC.2018.09.005.

Herskind, Carsten, Frederik Wenz, and Frank A. Giordano. 2017. “Immunotherapy Combined with Large Fractions of Radiotherapy: Stereotactic Radiosurgery for Brain Metastases-Implications for Intraoperative Radiotherapy after Resection.” *Frontiers in Oncology* 7 (JUL): 269688. https://doi.org/10.3389/FONC.2017.00147/BIBTEX.

Hsu, Man Jen, Yuan Chin Ivan Chang, and Huey Miin Hsueh. 2014. “Biomarker Selection for Medical Diagnosis Using the Partial Area under the ROC Curve.” *BMC Research Notes* 7 (1). https://doi.org/10.1186/1756-0500-7-25.

Huang, Bo, Lu Tian, Enayet Talukder, Mace Rothenberg, Dae Hyun Kim, and Lee Jen Wei. 2018. “Evaluating Treatment Effect Based on Duration of Response for a Comparative Oncology Study.” *JAMA Oncology* 4 (6): 876–79. https://doi.org/10.1001/JAMAONCOL.2018.0275.

Jafarzadeh, Meisam, and Bahram M. Soltani. 2021. “MiRNA-Wnt Signaling Regulatory Network in Colorectal Cancer.” *Journal of Biochemical and Molecular Toxicology* 35 (10): e22883. https://doi.org/10.1002/JBT.22883.

Jayawardana, Kaushala, Sarah Jane Schramm, Varsha Tembe, Samuel Mueller, John F. Thompson, Richard A. Scolyer, Graham J. Mann, and Jean Yang. 2016. “Identification, Review, and Systematic Cross-Validation of MicroRNA Prognostic Signatures in Metastatic Melanoma.” *The Journal of Investigative Dermatology* 136 (1): 245–54. https://doi.org/10.1038/JID.2015.355.

Jiao, Peng, Xing Ping Wang, Zhuo Ma Luoreng, Jian Yang, Li Jia, Yun Ma, and Da Wei Wei. 2021. “MiR-223: An Effective Regulator of Immune Cell Differentiation and Inflammation.” *International Journal of Biological Sciences* 17 (9): 2308–22. https://doi.org/10.7150/IJBS.59876.

Jollife, Ian T., and Jorge Cadima. 2016. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065). https://doi.org/10.1098/RSTA.2015.0202.

Kamphorst, Bart, Thomas Rooijakkers, Thijs Veugen, Matteo Cellamare, and Daan Knoors. 2022. “Accurate Training of the Cox Proportional Hazards Model on Vertically-Partitioned Data While Preserving Privacy.” *BMC Medical Informatics and Decision Making* 22 (1): 1–18. https://doi.org/10.1186/S12911-022-01771-3/FIGURES/3.

Kishore, Jugal, ManishKumar Goel, and Pardeep Khanna. 2010. “Understanding Survival Analysis: Kaplan-Meier Estimate.” *International Journal of Ayurveda Research* 1 (4): 274. https://doi.org/10.4103/0974-7788.76794.

Korbecki, Jan, Szymon Grochans, Izabela Gutowska, Katarzyna Barczak, and Irena Baranowska-Bosiacka. 2020. “CC Chemokines in a Tumor: A Review of Pro-Cancer and Anti-Cancer Properties of Receptors CCR5, CCR6, CCR7, CCR8, CCR9, and CCR10 Ligands.” *International Journal of Molecular Sciences* 21 (20): 1–34. https://doi.org/10.3390/IJMS21207619.

Kourou, Konstantina, Georgios Manikis, Paula Poikonen-Saksela, Ketti Mazzocco, Ruth Pat-Horenczyk, Berta Sousa, Albino J. Oliveira-Maia, et al. 2021. “A Machine Learning-Based Pipeline for Modeling Medical, Socio-Demographic, Lifestyle and Self-Reported Psychological Traits as Predictors of Mental Health Outcomes after Breast Cancer Diagnosis: An Initial Effort to Define Resilience Effects.” *Computers in Biology and Medicine* 131 (April). https://doi.org/10.1016/J.COMPBIOMED.2021.104266.

Krupke, Debra M., Dale A. Begley, John P. Sundberg, Carol J. Bult, and Janan T. Eppig. 2008. “The Mouse Tumor Biology Database.” *Nature Reviews. Cancer* 8 (6): 459–65. https://doi.org/10.1038/NRC2390.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature 2001 409:6822* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Lee, Ho Joon, Jennifer Palm, Susan M. Grimes, and Hanlee P. Ji. 2015. “The Cancer Genome Atlas Clinical Explorer: A Web and Mobile Interface for Identifying Clinical–Genomic Driver Associations.” *Genome Medicine* 7 (1). https://doi.org/10.1186/S13073-015-0226-3.

Leonardi, Giulia C., Luca Falzone, Rossella Salemi, Antonino Zanghì, Demetrios A. Spandidos, James A. Mccubrey, Saverio Candido, and Massimo Libra. 2018. “Cutaneous Melanoma: From Pathogenesis to Therapy (Review).” *International Journal of Oncology* 52 (4): 1071–80. https://doi.org/10.3892/IJO.2018.4287/HTML.

Leva, Gianpiero Di, Michela Garofalo, and Carlo M. Croce. 2014. “MicroRNAs in Cancer.” 9 (January): 287–314. https://doi.org/10.1146/ANNUREV-PATHOL-012513-104715.

Liu, Jun, Guili Sun, Shangling Pan, Mengbin Qin, Rong Ouyang, Zhongzhuan Li, and Jiean Huang. 2020. “The Cancer Genome Atlas (TCGA) Based M6A Methylation-Related Genes Predict Prognosis in Hepatocellular Carcinoma.” *Bioengineered* 11 (1): 759. https://doi.org/10.1080/21655979.2020.1787764.

Mahesh, Guruswamy, and Roopa Biswas. 2019. “MicroRNA-155: A Master Regulator of Inflammation.” *Journal of Interferon & Cytokine Research : The Official Journal of the International Society for Interferon and Cytokine Research* 39 (6): 321–30. https://doi.org/10.1089/JIR.2018.0155.

Mattick, John S., and Igor V. Makunin. 2006. “Non-Coding RNA.” *Human Molecular Genetics* 15 Spec No 1. https://doi.org/10.1093/HMG/DDL046.

Mattiske, Sam, Rachel J. Suetani, Paul M. Neilsen, and David F. Callen. 2012. “The Oncogenic Role of MiR-155 in Breast Cancer.” *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 21 (8): 1236–43. https://doi.org/10.1158/1055-9965.EPI-12-0173.

Netanely, Dvir, Stav Leibou, Roma Parikh, Neta Stern, Hananya Vaknine, Ronen Brenner, Sarah Amar, et al. 2021. “Classification of Node-Positive Melanomas into Prognostic Subgroups Using Keratin, Immune, and Melanogenesis Expression Patterns.” *Oncogene 2021* 40 (10): 1792–1805. https://doi.org/10.1038/s41388-021-01665-0.

Peng, Xinxin, Zhongyuan Chen, Farshad Farshidfar, Xiaoyan Xu, Philip L. Lorenzi, Yumeng Wang, Feixiong Cheng, et al. 2018. “Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers.” *Cell Reports* 23 (1): 255. https://doi.org/10.1016/J.CELREP.2018.03.077.

Rafique, Raihan, S. M.Riazul Islam, and Julhash U. Kazi. 2021. “Machine Learning in the Prediction of Cancer Therapy.” *Computational and Structural Biotechnology Journal* 19 (January): 4003–17. https://doi.org/10.1016/J.CSBJ.2021.07.003.

Rastrelli, Marco, Saveria Tropea, Carlo Riccardo Rossi, and Mauro Alaibac. 2014. “Melanoma: Epidemiology, Risk Factors, Pathogenesis, Diagnosis and Classification.” *In Vivo (Athens, Greece)* 28 (6): 1005–12. https://pubmed.ncbi.nlm.nih.gov/25398793/.

Sanchez-Vega, Francisco, Marco Mina, Joshua Armenia, Walid K. Chatila, Augustin Luna, Konnor C. La, Sofia Dimitriadoy, et al. 2018. “Oncogenic Signaling Pathways in The Cancer Genome Atlas.” *Cell* 173 (2): 321-337.e10. https://doi.org/10.1016/J.CELL.2018.03.035.

Sas, Zuzanna, Ewa Cendrowicz, Isabel Weinhäuser, and Tomasz P. Rygiel. 2022. “Tumor Microenvironment of Hepatocellular Carcinoma: Challenges and Opportunities for New Treatment Options.” *International Journal of Molecular Sciences 2022*, 23 (7): 3778. https://doi.org/10.3390/IJMS23073778.

Shabani, Parastoo, Sama Izadpanah, Ali Aghebati-Maleki, Elham Baghbani, Amir Baghbanzadeh, Ali Fotouhi, Babak Bakhshinejad, Leili Aghebati-Maleki, and Behzad Baradaran. 2019. “Role of MiR-142 in the Pathogenesis of Osteosarcoma and Its Potential as Therapeutic Approach.” *Journal of Cellular Biochemistry* 120 (4): 4783–93. https://doi.org/10.1002/JCB.27857.

Shastry, K. Aditya, and H. A. Sanjay. 2020. “Machine Learning for Bioinformatics,” *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications* 25–39. https://doi.org/10.1007/978-981-15-2445-5_3.

Sonego, Paolo, András Kocsor, and Sándor Pongor. 2008. “ROC Analysis: Applications to the Classification of Biological Sequences and 3D Structures.” *Briefings in Bioinformatics* 9 (3): 198–209. https://doi.org/10.1093/BIB/BBM064.

Sur, Daniel, Claudia Burz, Shanthi Sabarimurugan, and Alexandru Irimie. 2020. “Diagnostic and Prognostic Significance of MiR-150 in Colorectal Cancer: A Systematic Review and Meta-Analysis.” *Journal of Personalized Medicine* 10 (3): 1–12. https://doi.org/10.3390/JPM10030099.

Syeda, Zainab Ali, Siu Semar Saratu’ Langden, Choijamts Munkhzul, Mihye Lee, and Su Jung Song. 2020. “Regulatory Mechanism of MicroRNA Expression in Cancer.” *International Journal of Molecular Sciences* 21 (5). https://doi.org/10.3390/IJMS21051723.

Tang, Zefang, Boxi Kang, Chenwei Li, Tianxiang Chen, and Zemin Zhang. 2019a. “GEPIA2: An Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis.” *Nucleic Acids Research* 47 (W1): W556–60. https://doi.org/10.1093/NAR/GKZ430.

Thompson, Jacob W., Ruozhen Hu, Thomas B. Huffaker, Andrew G. Ramstead, H. Atakan Ekiz, Kaylyn M. Bauer, William W. Tang, et al. 2023. "MicroRNA-155 Plays Selective Cell-Intrinsic Roles in Brain-Infiltrating Immune Cell Populations during Neuroinflammation." *The Journal of Immunology* 210 (7): 926–34. https://doi.org/10.4049/JIMMUNOL.2200478.

Thorsson, Vésteinn, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai Hsien Ou Yang, Eduard Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4): 812-830.e14. https://doi.org/10.1016/J.IMMUNI.2018.03.023.

Tsai, Yihsuan S., Daniel Dominguez, Shawn M. Gomez, and Zefeng Wang. 2015. "Transcriptome-Wide Identification and Study of Cancer-Specific Splicing Events across Multiple Tumors." *Oncotarget* 6 (9): 6825–39. https://doi.org/10.18632/ONCOTARGET.3145.

Wang, Yunxia, Song Zhang, Fengcheng Li, Ying Zhou, Ying Zhang, Zhengwen Wang, Runyuan Zhang, et al. 2020. "Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics." *Nucleic Acids Research* 48 (D1): D1031. https://doi.org/10.1093/NAR/GKZ981.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63. https://doi.org/10.1038/NRG2484.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* (10): 1113–20. https://doi.org/10.1038/ng.2764.

Wiesweg, M., F. Mairinger, H. Reis, M. Goetz, R. F.H. Walter, T. Hager, M. Metzenmacher, et al. 2019. "Machine Learning-Based Predictors for Immune Checkpoint Inhibitor Therapy of Non-Small-Cell Lung Cancer." *Annals of Oncology : Official Journal of the European Society for Medical Oncology* 30 (4): 655–57. https://doi.org/10.1093/ANNONC/MDZ049.

Witten, Daniela M., and Robert Tibshirani. 2009. "Covariance-Regularized Regression and Classification for High Dimensional Problems." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71 (3): 615–36. https://doi.org/10.1111/J.1467-9868.2009.00699.X.

Wu, Fengfeng, Xuesheng Jiang, Qun Wang, Qian Lu, Fengxiang He, Jianyou Li, Xiongfeng Li, Mingchao Jin, and Juntao Xu. 2020. "The Impact of MiR-9 in Osteosarcoma: A Study Based on Meta-Analysis, TCGA Data, and Bioinformatics Analysis." *Medicine* 99 (35): e21902. https://doi.org/10.1097/MD.0000000000021902.

Xiao, Yi, and Dihua Yu. 2021. "Tumor Microenvironment as a Therapeutic Target in Cancer." *Pharmacology & Therapeutics* 221 (May). https://doi.org/10.1016/J.PHARMTHERA.2020.107753.

Xing, Yun, Zhiqiang Wang, Zhou Lu, Jie Xia, Zhangjuan Xie, Mengxia Jiao, Ronghua Liu, and Yiwei Chu. 2021. “MicroRNAs: Immune Modulators in Cancer Immunotherapy.” *Immunotherapy Advances* 1 (1). https://doi.org/10.1093/IMMADV/LTAB006.

Xu, Chunming, and Scott A. Jackson. 2019. “Machine Learning and Complex Biological Data.” *Genome Biology* 20 (1): 1–4. https://doi.org/10.1186/S13059-019-1689-0/FIGURES/2.

Xu, Qianhui, Yuxin Wang, and Wen Huang. 2021. “Identification of Immune-Related LncRNA Signature for Predicting Immune Checkpoint Blockade and Prognosis in Hepatocellular Carcinoma.” *International Immunopharmacology* 92 (March): 107333. https://doi.org/10.1016/J.INTIMP.2020.107333.

Yeung, K. Y., and W. L. Ruzzo. 2001. “Principal Component Analysis for Clustering Gene Expression Data.” *Bioinformatics (Oxford, England)* 17 (9): 763–74. https://doi.org/10.1093/BIOINFORMATICS/17.9.763.

Yoshida, Kosuke, Yusuke Yamamoto, and Takahiro Ochiya. 2021. “MiRNA Signaling Networks in Cancer Stem Cells.” *Regenerative Therapy* 17 (June): 1. https://doi.org/10.1016/J.RETH.2021.01.004.

Yousefi, Paul D., Matthew Suderman, Ryan Langdon, Oliver Whitehurst, George Davey Smith, and Caroline L. Relton. 2022. “DNA Methylation-Based Predictors of Health: Applications and Statistical Considerations.” *Nature Reviews in Genetics* 23 (6): 369–83. https://doi.org/10.1038/S41576-022-00465-W.

Yu, Yingjie, Pratima Nangia-Makker, Lulu Farhana, Sindhu G. Rajendra, Edi Levi, and Adhip P.N. Majumdar. 2015. “MiR-21 and MiR-145 Cooperation in Regulation of Colon Cancer Stem Cells.” *Molecular Cancer* 14 (1): 98. https://doi.org/10.1186/S12943-015-0372-7.

Zhu, Shu, Wen Pan, and Youcun Qian. 2013. “MicroRNA in Immunity and Autoimmunity.” *Journal of Molecular Medicine (Berlin, Germany)* 91 (9): 1039–50. https://doi.org/10.1007/S00109-013-1043-Z.

Zhu, Yitan, Peng Qiu, and Yuan Ji. 2014. “TCGA-Assembler: An Open-Source Pipeline for TCGA Data Downloading, Assembling, and Processing.” *Nature Methods* 11 (6): 599. https://doi.org/10.1038/NMETH.2956.