**Conceptual Design Report-Paula Vazquez**

**Searching for novel breast and prostate cancer RNA biomarkers and identifying common genes disturb in these cancers by applying machine learning algorithms on transcriptome data from the Cancer Genome Atlas (TACG) database.**

**Abstract:**

The search for novel RNA biomarkers in cancer can significantly advance the development of RNA-based diagnostic and therapeutic strategies, leading to more effective and personalized approaches for cancer treatment and management. This project's primary objective is to find novel RNA biomarkers for breast and prostate cancer, both of which rank among the most prevalent cancers in the human population. Additionally, the project seeks to investigate the feasibility of predicting or diagnosing these cancers by applying machine learning algorithms to the gene expression profiles of tumor and normal tissue. Through an examination of the gene expression patterns disturb in both cancer types, the project also aims to enhance our comprehension of the shared pathways disrupted or contributing to disease progression. This conceptual project report describes the data, methods, data flow, and models that will be used for this project.

**Table of Contents:**

Abstract ……………………………………………………………………………………………………………………………….1

Introduction………………………………………………………………………………………………………………...........3

Project Objectives.……………………………………………………………………………………………………………….4

Methods.……………………………………………………………………………………………………………………………..5

Data..………………………………………………………………………………………………………………………………….11

Metadata..………………………………………………………………………………………………………………………….13

Data Quality……………………………………………………………………………………………………………………….14

Data flow……………………………………………………………………………………………………………………………15

Risks……………………………………………………………………………………………………………………………………16

Conclusions and Outlook…………………………………………………………………………………………………….17

References and Bibliography……………………………………………………………………………………………...18

**Introduction:**

Female breast cancer has now surpassed lung cancer as the most frequently diagnosed cancer in women, while prostate cancer stands as the second leading cause of cancer-related death for men globally, second only to lung cancer (1,2). A family history of breast cancer was reported to be associated with an increased risk of prostate cancer suggesting that certain genetic mutations that increase the risk of breast cancer may result in an increased risk of prostate cancer in men (3). This suggests that there may be a link between prostate cancer and breast cancer disease pathways. Early diagnosis plays a pivotal role in the prevention and successful treatment of patients with breast and prostate tumors. Therefore, the identification of novel biomarkers and targets linked to these diseases are crucial for enhancing diagnostic methods, developing personalized therapies, and predicting treatment outcomes.

The Cancer Genome Atlas (TCGA) database houses comprehensive information on the molecular attributes of over 20,000 primary cancer and matched normal samples, encompassing 33 types of cancer (4). One of the key components of this database is the transcriptome profiles, representing genes actively transcribed into RNA within the samples. Perturbations in RNA expression levels are frequently observed in various diseases, including cancer. Consequently, RNAs exhibiting differential expression in cancer cells compared to normal cells can serve as valuable diagnostic biomarkers or therapeutic targets.

This project aims to shed light on the common genes and molecular pathways activated or inactivated in breast and prostate cancers through the identification of differentially expressed genes (DE) in tumors and corresponding normal tissue samples. Additionally, machine learning algorithms will be employed to assess whether RNA expression levels can predict patients' cancer outcomes, as well as to identify cancer-type-specific biomarkers for breast and prostate cancers using transcriptome profile data.

**Project Objectives:**

1. **Identification of Common and Specific Gene Disruptions in Breast and Prostate Cancer:**

Considering the association between a family history of breast cancer and the increased risk of prostate cancer, our project seeks to elucidate the shared genes and molecular pathways activated or deactivated by both cancer types. The primary focus will be on identifying the genes with disrupted expression in both breast and prostate cancers.

1. **Predicting patient outcomes based on gene expression and identifying the best RNA biomarkers.**

RNA biomarkers are specific RNA molecules that can serve as indicators or predictors of various biological processes or disease states, including cancer. They can be crucial in understanding disease mechanisms, diagnosing diseases, predicting disease progression, and monitoring treatment responses. They can also provide valuable insights into the molecular and cellular changes associated with tumorigenesis and tumor progression. Our approach involves the application of machine learning algorithms on transcriptome profiles of breast cancer and prostate cancer samples alongside their corresponding normal tissue to discover novel biomarkers for these types of cancer. Moreover, our project involves exploring whether machine learning algorithms applied to gene expression profiles can effectively predict the presence of the disease within the samples.

**Methods:**

**Infrastructure and packages:**

Local Python installation (5,6) within Anaconda (7) will serve as the project's primary Infrastructure. The project will utilize the following modules/packages:

1. **OS Package (8) and TCGA Downloader Python Script (9):** These tools will facilitate the download of transcriptome profile raw count data from the TCGA database (4). Initially, a 'manifest' file will be acquired directly from the web interface of the database.
2. **PyDESeq2 package (10):** Acting as a wrapper around the R-based DESeq2 package, PyDESeq2 provides a more convenient means of implementing DESeq2 within Python, eliminating the need to handle the R code explicitly. This package will be used to normalize the row count data and to perform the differential expression (DE) gene analysis.

1. **Pandas (11) and NumPy (12):** These libraries will aid in the creation of the merged data frames. They will also facilite data filtering and preparation of the data for subsequent analysis.
2. **Data Visualization Libraries:** Matplotlib (13), matplotlib\_venn (14), Seaborn (15), and Plotly (16) will be used for diverse data visualization and exploratory data analysis purposes.
3. **Machine Learning Libraries:** Scikit-learn (17), **XGBoost (18),**  TensorFlow (19) will be employed for the implementation of various machine learning methods.

**Data downloading:**

Transcriptome profiles (RNAseq data) of breast, and prostate cancer and their corresponding normal tissue samples were downloaded from The Cancer Genome Atlas (TCGA) data set, accessible via the National Cancer Institute’s (NCI’s) Genomic Data Commons (GDC) platform (4). The retrieval process involved the generation of a manifest file directly on the GDC Data portal, following the specified portal instructions. The following steps were undertaken to produce the manifest files:

1. **Selection of required samples:**

a) Under the ‘Repository’ section, the criteria for downloading transcriptome data were set as follows:

* Data category: Transcriptome Profiling
* Data type: Gene Expression Quantification
* Experimental Strategy: RNA-Seq
* Workflow Type: Star-Counts

b) Within the ‘Cases’ section, the data selected included:

* TCGA-BRCA for breast cancer project samples.
* TCGA-PRAD for prostate gland cancer project samples.

1. **Manifest File Generation:**

The selected samples were added to the browser's 'shopping cart', leading to the download of two distinct manifest files as text files—one for retrieving breast cancer samples and the other for prostate cancer samples.

Subsequently, the tcga\_downloader.py Python script, developed by Vincent Appiah (5), was employed to facilitate the download of data utilizing the generated manifest files. The execution of the script, guided by the developer's tutorial instructions, resulted in the retrieval of individual patient sample data organized into various folders segregated by clinical outcomes, including 'Primary Tumor Samples’, 'Metastatic', and 'Solid Tissue Normal'. Each of these folders contained specific TSV files, with each file corresponding to an individual sample (a sample file example is presented in Table 1). Moreover, the script enabled the acquisition of the associated metadata file, which provided comprehensive information on sample preparation, analysis type, and pertinent clinical details.

**Data Pre-processing:**

The use of fpkm (**fragments per kilobase of transcript per million mapped reads**) and tpm (transcripts per million) gene counts for sample-to-sample comparisons and performing differential gene expression analysis is discouraged as they are only suitable for gene expression comparison within the same sample (20-21). To compare gene expression between samples, normalized raw counts are recommended (20-21). Therefore, the columns labeled ‘unstranded’ that represent the raw sequencing counts for each gene will be selected for further analysis. Since the raw sequencing data for each sample exists in individual files, a merged data frame will be created by appending the ‘unstranded’ raw counts column for each sample. A merged table will be generated for each project, one for breast cancer and one for prostate cancer using pandas. A Python script will be devised to iterate through the files within each folder, extracting the 'unstranded' column from each TSV file, using the file name (sample identifier) as the column name, and merging all the data into a single frame. This process will be complemented by incorporating the folder name or information from the metadata files to add the clinical data outcome for each sample. We will also ensure that the common columns (gene ID, gene\_type, gene name) are preserved in the merged data and that the extracted columns correspond to the correct values. The combined data will be written into a new CSV file. This procedure will be replicated for folders containing both normal and tumor samples, eventually consolidating the final files into a single CSV file for each project.

There are many undetected genes, where the raw counts are 0 across all samples. These genes will be removed since genes that do not express at all cannot be differentially expressed.  We will also consider removing genes with very low counts. To facilitate accurate comparisons of gene expression between samples, normalization of the raw counts is essential. The PyDESeq2 package employs the median of ratios method to normalize the raw counts of each gene in the samples, thereby accounting for variations in sequencing depth and RNA composition. As a prerequisite for utilizing PyDESeq2, the count matrices and metadata matrices need to be prepared, aligning with the requirements outlined in the PyDESeq2 package documentation (10). The count matrix, essentially a data frame with genes in the columns and sample IDs in the rows, will be transposed from the original merged data frame. Likewise, the metadata data frame will be structured accordingly, with the sample ID in the rows and the outcome/condition in the columns.

**Differential expression gene analysis and exploratory analysis of common upregulated and downregulated genes in breast cancer and prostate cancer:**

Differential gene expression (DE) analysis for breast and prostate cancer will be conducted using the PyDESeq2 package (10). This analysis will focus on comparing gene expression levels between tumor and normal tissue samples for each cancer type separately. The resulting output will yield a comprehensive table comprising gene identifiers, expression level ratios between cancer and control samples as fold changes and log2-fold changes, and pertinent information concerning the statistical significance of these changes. This information will include p-values, adjusted p-values (p-adj), test statistics, or confidence intervals. In the output table, genes exhibiting log2-fold values greater than 0 will be categorized as upregulated genes, whereas those with values less than 0 will be classified as downregulated genes. Genes meeting the criteria of a log2-fold value greater than or equal to 2 or less than or equal to -2, along with a padj value of less than 0.05, will be shortlisted for further comprehensive analysis and data exploration. Nevertheless, varying log2-fold values will be tested to determine the most appropriate cut-off for subsequent analysis.

The differential expression (DE) data will be examined and visualized through the creation of volcano plots and interactive volcano plots, utilizing the Plotly and Matplotlib libraries. Additionally, heat maps of the most differentially expressed (DE) genes will be generated using the Sns package, enabling the identification of prominent patterns in gene expression across the conditions. To gain deeper insights into the shared and distinct gene expression patterns between breast and prostate cancers, Venn diagrams will be constructed using the matplotlib\_venn library. This analysis will facilitate a comprehensive exploration of the most crucial genes exhibiting differential upregulation or downregulation in tumor samples compared to normal tissue samples. This process will aid in the identification and selection of the most pertinent DE genes, paving the way for a more comprehensive and targeted downstream analysis.

**Machine learning algorithms for detecting the best biomarkers and predicting the outcome of breast and prostate cancer by the pattern of gene expression:**

**1. Feature dimensionality reduction:**

For applying machine learning algorithms, a log2-fold (> 2 or < -2, as starting point), and p-adj value < 0.05 cut-off, will be selected to reduce the feature dimensionality (number of genes) used within the algorithms. Additionally, a further cut-off will be applied to the overall gene expression levels, primarily by excluding genes with considerably low row counts in both tumor and normal samples. This step is crucial, as biomarkers exhibiting weak expression or detection via RNAseq might not serve as effective markers for diagnosis. Furthermore, genes that do not demonstrate significant differential expression between tumor and normal samples will not contribute to the identification of valuable biomarkers.

PCA analysis can also help in reducing the dimensionality of the data by transforming the original features into a new set of uncorrelated variables, known as principal components. This reduction can simplify the analysis and improve the computational efficiency of the machine learning algorithms. We will also consider applying SNE, or t-distributed Stochastic Neighbor Embedding clustering for dimensionality reduction. Both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) will be applied using the scikit-learn library (17). PCA analysis can also help in identifying the most important genes or gene combinations that contribute significantly to the prediction of disease progression. Patients with cancers can have different patterns of gene expression and the goal of precision medicine is to look for specific/ individual biomarkers to use as target therapies. Given the potential variations in gene expression patterns among cancer patients, the implementation of PCA analysis aims to discern distinct clusters or groups within breast cancer and prostate cancer patients. For example, breast cancer can be classified into four molecular subtypes based on the expression of the human epidermal growth factor receptor 2 (HER2) and hormone receptor (HR). PCA analysis can help to distinguish these patterns and find additional patterns if present in the samples. Moreover, performing PCA analysis across all samples, including prostate cancer and breast cancer, could shed light on potential shared gene expression patterns among patients with both cancer types.

1. **Supervised Machine learning:**

The next step involves the application of supervised learning approaches to uncover potential biomarkers and predicting disease outcome. We will start by the implementation of the simplest linear regression model (from the Scikit-learn package (17) aiming to determine the most effective model for identifying biomarkers and predicting disease outcomes. Subsequently, we will employ random forest (from the Scikit-learn package) and boosted classification trees to forecast patient outcomes (from the **XGBoost**  package (18). Considering the complexity and high dimensionality of RNA-seq data, we will also consider to apply neuronal network from the tensor flow packages (19). To comprehensively evaluate the performance of the machine learning model, metrics including Accuracy, Precision, Recall, and F1 scores will be calculated. Since these cancers are sex-dependent and it does not make sense to distinguish/predict outcomes between them, analysis of breast cancer vs no cancer and prostate cancer vs no cancer will be done separately. To facilitate the loading of data into these models, the categorical outcome variable will be converted into numerical values, with '0' denoting normal and '1' representing tumor samples.

Two distinct data frames will be created, one for breast cancer data and the other for prostate cancer data. These data frames will feature the outcome variable (y) in the first column, with subsequent columns representing the gene or features (x) and their corresponding expression levels in normalized counts for each gene in the sample.

Both the features (x) and the outcome (y) variables will be converted to Numpy-arrays. The data sets will be randomly split into a training set and a test set using the train\_test\_split function from the Scikit\_learn module (17). Adhering to convention, approximately 80% of the data will be allocated to the training set, with the remaining 20% assigned to the validation set. The overarching goal is to achieve an accuracy rate exceeding 80%. This comprehensive approach aims to facilitate the identification of robust predictive models and subsequently, the potential development of efficient diagnostic and prognostic tools for breast and prostate cancers.

**Data:**

The RNA seq data that will be used in this project is available on the Cancer Genome Atlas (TCGA) web page (4). The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program that has molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. This is a joint effort between NCI and the National Human Genome Research Institute that began in 2006, bringing together researchers from diverse disciplines and multiple institutions. The TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data is publicly available for anyone in the research community to use.

For this project, RNA sequencing data from breast cancer and prostate cancer tissue samples, along with their respective control samples, has been retrieved from the TCGA database and is as follows:

* **Breast Cancer:** transcriptome profiles of 1109 Primary Tumor samples, 113 samples from Solid Normal Tissue, and 7 samples of metastatic tissue (this data will not be used in this project).
* **Prostate Cancer:** transcriptome profiles of 306 Primary Tumors and 52 solid normal tissue. 1 metastatic sample that will not be used in this study.

The data retrieval script categorizes the data into distinct folders based on clinical outcome classifications. Each folder contains individual TSV files, representing data from each specific sample (an exemplary file is shown in Table 1).

**Table 1.** 

**Table 1.** Example of the single sequencing file data from a breast tumor sample. The first rows showing the expression profile of the first 14 genes are shown from a total of 60.000 genes detected in the RNA seq data.

File name: 0a9e33db-2527-4cc3-8669-a7c10fed7a7f.rna\_seq.augmented\_star\_gene\_counts.tsv

The initial rows of the data represent the counts for mapped, multi-mapped, and unmapped sequencing reads in the sample with reference to the genome. Following this, the data provides details in the following format:

* **Gene\_Id:** Ensembl ID gene identifier.
* **Gene\_name**: Annotated name of the gene**.**
* **Gene\_type**: Type of gene categorized based on from the Ensembl biotype annotation, accessible at: http://www.ensembl.org/info/genome/genebuild/biotypes.html.
* **Unstranded**: Total raw reads mapped to each gene using an Unstranted library preparation method. This approach does not retain information about the originating RNA strand, making it valuable for quantifying overall gene expression levels.
* **Stranded\_first**: Number of raw reads mapped to the first cDNA strand generated during library preparation using a Stranded RNA-seq method.
* **Stranded\_second:** Number of raw reads mapped to the second cDNA strand generated in a Stranded RNA-seq method.
* **Tpm-unstranded** **(Transcripts per million):** Tpm is a measure of gene expression normalized by the total number of transcripts in the sample and scaled to a million.
* **Fpkm- unstranded (Fragments per kilobase of transcript per million mapped reads):** Fpkm is another metric used to quantify gene expression levels. It takes into account the length of the gene and the total number of fragments (reads) that map to that gene.

### **Fpkm-uq-unstranded (Upper quartile fpkm):** Fpkm-uq is a modified version of fpkm that is normalized by the upper quartile of the fpkm values across all genes in a given sample.

It is essential to note that for differential gene expression (DE) analysis (sample-to-sample comparisons), the use of fpkm and tpm is not recommended; instead, the 'unstranded' raw count column will be extracted for subsequent analysis (see methods).

**Metadata:**

In the context of RNA sequencing data or transcriptomics, metadata serve to describe the sample from which the RNA sequence was derived, including crucial details such as the organism, cell line, and the specific library-preparation method employed. For the dataset used within this project, a metadata TSV (Tab-Separated Values) file is downloaded simultaneously with the data, following the protocol outlined in the method section. This metadata file encompasses pertinent information related to each downloaded file. The most pertinent information includes:

* **Sample ID:** A unique identifier assigned to each specific sample/patient.
* **Disease Information:** Details regarding the disease associated with the patient from whom the sample was obtained for analysis.
* **Sample Type:** Designation specifying whether the sample corresponds to a tumor sample or solid normal tissue.
* **Data Category and Data Type:** Information outlining the specific data category and type corresponding to each sample.
* **File name:** the name of the CSV file for each sample, is included within the metadata. This particular detail can be used to extract sample-specific information from the metadata file, thereby facilitating comprehensive data analysis and interpretation.

Top of Form

Bottom of Form

**Data Quality:**

Before proceeding with a formal analysis, it is imperative to conduct an initial assessment of data quality for the RNA-seq data. The preliminary examination reveals numerous undetected genes across different samples, where the raw count values are observed to be 0 in both the control and tumor samples. It is essential to remove these genes from the analysis, as the lack of expression or detection through the sequencing method precludes the possibility of identifying differential expression.

Another critical aspect to evaluate is the library size, denoting the total number of mapped reads in each sample (the aggregate of mapped reads for all genes). Any substantial deviations in the total read counts among samples, particularly if a specific sample exhibits significantly fewer mapped reads compared to others, may suggest potential technical issues during sequencing. These particular samples warrant further scrutiny, especially if they do not align with the overall data clustering patterns.

Additionally, examining the percentage of unmapped reads in each sample, as indicated in the initial rows of the data, is crucial. Generally, a consistent percentage of unmapped reads is anticipated across all samples. Significant deviations in these percentages could imply sample degradation or contamination issues, prompting the need for a careful reconsideration of their inclusion in the analysis.

**Data flow:**

****

**Figure 1.** Flow chart of the research process.

**Risks:**

While the utilization of machine learning in predicting disease outcomes and identifying biomarkers based on RNA-seq data presents promising opportunities, it is vital to carefully consider the associated risks and challenges. RNA-seq data, being inherently noisy, can be susceptible to batch effects and various technical artifacts that may impact the reliability and robustness of the results. Furthermore, the presence of biological variability introduces complexities that might not be entirely captured within the data. In the context of the Cancer Genome Atlas (TCGA) project database, specimens of histologically normal tissue adjacent to surgically removed tumors are considered normal. However, these tissues may not be completely normal due to the numerous effects tumors may have on the neighboring cells, including biased growth factors and cytokine balances, pathological inflammation, and altered vascularization (22). Consequently, the control samples derived from the TCGA database might not be optimal control samples for detecting differentially expressed genes. Alternative databases containing normal sample data, such as the ones deposited in the Oncobox Atlas of Normal Tissue Expression (ANTE) platform*,* could be used as a source to either analyze or validate our findings (22). Furthermore, the number of control samples is much lower than the number of tumor samples and it is at the ‘low limit’ for performing machine learning experiments. Therefore, the inclusion of RNAseq data form other sources may improve the predictions.

The TCGA includes Age and ethnicity data. Given that the age of the patients can significantly influence gene expression patterns, it could be also advisable to analyze the data in groups comprising age-aggregated samples. Although we have information on the age of the patients from where the samples were taken since age-related gene expression can affect the results it might be useful to analyze the data in groups of age-aggregated samples. Additionally, the ethnicity of the patients from whom the samples were obtained introduces another potential bias that must be carefully considered, particularly when generalizing the results for application in the healthcare sector across diverse populations. Therefore it will be crucial to assess whether the demographics represented within the dataset used in this project correspond to the ethnic composition observed within the general population.

**Conclusions and Outlook:**

Nowadays, the gold standard for prostate cancer diagnosis involves a prostate biopsy performed after a previous clinical suspicion based on prostate-specific antigen (PSA) levels and digital rectal examination (DRE). On the other side, mammogram screening has been widely used for breast cancer screening. However, concerns persist regarding the reduced sensitivity of DRE, the low specificity of PSA, the elevated rates of false positives and false negatives, and the potential radiation exposure associated with mammography.

Over the past two decades, an increasing amount of ‘omics’ data has become publicly available. In this project, the main objective is to identify RNA biomarkers of prostate cancer and breast cancer by applying machine learning algorithms to transcriptome data. The final aim will be to use these RNA markers for early and non-invasive diagnosis of cancer. Notably, circulating cell-free or extracellular vesicle-packaged tumor-related nucleic acids, encompassing mRNAs, microRNAs, and lncRNAs in bodily fluids such as peripheral blood, nipple aspirate fluid, sweat, urine, and tears, have emerged as potential non-invasive diagnostic biomarkers, supplementing existing clinical methods for early cancer detection (23-24). Following the identification of biomarker combinations capable of predicting breast and prostate cancer, subsequent wet lab experiments are imperative to ascertain the detectability of these RNA biomarkers in fluids using techniques like quantitative PCR (qPCR). The goal is to achieve a prediction of cancer outcomes with enhanced or equivalent accuracy, sensitivity, and specificity compared to conventional and diagnostic methods currently in use or under development (23,24) and that are easily detected through a non -invasive technique.

This project also aim to to explore whether breast cancer and prostate cancer trigger similar genetic responses, especially considering the elevated risk of prostate cancer associated with a family history of breast cancer. While the absence of specific genes affected in both cancer types is plausible, complementary gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses can also be conducted to investigate potential disruptions in common biological pathways leading to disease progression.

**References:**

1. Hyuna Sung et.al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. **CA: A *Cancer Journal for Clinicians***, 71(3), 209-249.
2. Le Wang et al. (2022). Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019. ***Front Public Health,***10:811044.
3. Zheng-Ju Ren et al. (2020). First-degree family history of breast cancer is associated with prostate cancer risk: a systematic review and meta-analysis. ***The Prostate*,** 80(6), 422-432.
4. TCGA Research Network Data: https://www.cancer.gov/tcga.
5. Python Software Foundation. Python Language Reference, version 2.7. Available at:http://www.python.org
6. G. van Rossum. (1995). Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
7. Anaconda Software Distribution. (2020). Anaconda Inc. Available at: https://docs.anaconda.com/.
8. Van Rosuum. (2022). os-Miscellaneous operating system interfaces. Available at: https://docs.python.org/3/library/os.html
9. Vicent Vappiah (2021). tcga downloader python script*.* Github repository: https://github.com/vappiah/DataMiner. Tutorial: https://www.youtube.com/watch?v=YJxcsm4aJXI&t=150s.
10. Muzellec et.al. (2022). Pydeseq2: a python package for bulk rna-seq differential expression analysis. bioRxiv. doi:10.1101/2022.12.14.520412.
11. The Pandas Development Team. (2020). pandas-dev/pandas: Pandas. v 2.1.2. Zenodo. Available at: https://doi.org/10.5281/zenodo.3509134
12. Harris Charles R. et al. (2020). Array programming with Numpy. *Nature*, 585:7825, 357-362.
13. J. D. Hunter (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering.**IEEE COMPUTER SOC*. 9 (3), 90-95.
14. Konstantin Tretyakov. (2012). Matplotlib-venn 0.11.9. Available at: https://pypi.org/project/matplotlib-venn/
15. Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021, https://doi.org/10.21105/joss.03021
16. Plotly Technologies Inc. (2015). Collaborative data science Publisher: Plotly Technologies Inc. Montréal, QC. Available at: https://plot.ly
17. Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825–2830.
18. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, August 13-17, 2016. ACM, New York, NY, USA, 785-794. [https://doi.org/10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785" \t "_new)
19. TensorFlow Developers. (2022). TensorFlow (v2.8.2). Zenodo. https://doi.org/10.5281/ZENODO.4724125.
20. Yingdong Z et al. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine,* 19:269.
21. Harvard Chan Bioinformatics Core (HBC) team. Introduction to DGE-Archived. Available at:https://hbctraining.github.io/DGE\_workshop/lessons/02\_DGE\_count\_normalization.html.
22. Suntsova M. et al. (2019). Atlas of RNA sequencing profiles for normal human tissues. *Scientific Data* 6: 36.
23. Li J. et al. (2020). Non-Invasive Biomarkers for Early Detection of Breast Cancer.*Cancers(Basel),* 12(10): 2767.
24. McGrath S. et. al. (2016). Prostate cancer biomarkers: Are we hitting the mark? *Prostate International,* 4 (4):130-135.