Cancer Biomarker Discovery, CaBiD: A Tool for Investigating Differential Gene Expression in Cancer Data sets

Tony Okeke [1], Cooper Molloy[1], Ali Youssef [1]

[1] School of Biomedical Engineering, Drexel University, USA

Course : BMES 550-Advanced Biocomputational Languages
Instructor: Ahmet Sacan
Date : 2022-12-07

Abstract

Identifying genes with higher expression in tumor or metastatic tissues can help in better understanding tumor formation and can serve as biomarkers of progression or as potential therapy targets. In this project, the main goal is to discover potential cancer biomarkers by assessing genomic data from cancer patients with different cancer types by performing differential gene expression. A graphical user interface (GUI) is created in this project as a tool to highlight the difference between normal and tumoral patients for a given cancer type and preprocessed GSE data set selected ondemand by the user. The GUI is created in Python and data utilized in this project is from gene expression omnibus (GEO) and CuMiDa. The key features of this tool revolve around performing differential gene expression analysis between normal and tumoral samples for specific cancer types and visualizing results in volcano plots and heatmaps.

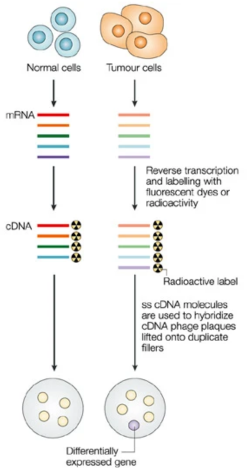
1 Introduction

With the development of research on cancer genomics and microenvironment, a new era of oncology focusing on the complicated gene regulation of cells and cancer immunotherapy is emerging. Especially the comprehensive analysis of differential gene expression in cancer cells. This project is aimed to identify the common gene expression characteristics of multiple cancers – lung cancer, liver cancer, kidney cancer, cervical cancer, and breast cancer – and the potential therapeutic targets in public databases [1]. Investigating the differences between diseased and healthy states helps us understand the pathology of diseases and, eventually, treat them. One particular focus of investigation is differentially-expressed genes (DEGs), which involves the identification of genes that are differentially expressed in disease. In pharmaceutical and clinical research, DEGs can be valuable to pinpoint candidate biomarkers, therapeutic targets and gene signatures for diagnostics [2]. Cancer is a highly complex, heterogeneous, and robust disease. It arises due to the failure at multiple levels in multicellular organisms. The complexity of genomic profiles, expression patterns, and cellular interactions within the tumor microenvironment are the major challenges in understanding the disease mechanism. This complexity results from intratumoral heterogeneity (the substantial genetic diversity within tumors), where cancer cells have distinct molecular and phenotypic features established by different genetic alterations and environmental factors. The goal of differential expression testing is to determine which genes are expressed more in different conditions. These genes can offer biological insight into the processes affected by the condition(s) of interest and become potential discovery tools for biomarkers for drug discovery or therapeutics. Figure 1 shows the differences between normal and tumor cells when differentially expressed [3].

This application strives to become an easy-access early stage discovery tool for researchers and investigators trying to highlight main biomarkers that are associated with certain types of cancer based on differential gene expression.

The main end users for this application are drug discovery scientists at pharmaceutical companies, bioinformatics researchers, and oncologists. End-users can implement this tool as an early stage discovery method to screen for potential biomarkers and genes that can indicate abnormal expression in cancer cases.

If this application becomes successful, it will empower the research field to understand the differences in gene expression of healthy and cancer patients. The broad potential applications of differential gene expression to generate biological insights, by predicting DE between tissues, differential transcript-usage, and drivers of aging throughout the human lifespan, of gene coexpression relationships on a genome-wide scale, and of frequently DE genes across diverse conditions.


Normal cells
Tumour cells
mRNA
Reverse transcription and labelling with fluorescent dyes or radioactivity
cDNA
Radioactive label
ss cDNA molecules are used to hybridize cDNA phage plaques lifted onto duplicate fillers
Differentially expressed gene

Nature Reviews | Cancer

Figure 1. Cancer vs. non-cancer cell. The figure illustrates the characteristic differences in gene expression between cancer and non-cancer cells.

Many researchers thrive to create applications that facilitate the understanding of different biological mechanisms related to gene expression. An application created by Bartha & Györffy, in the form of a web tool for the comparison of gene expression in normal, tumor and metastatic tissues. The pan-cancer analysis page of the tool displays the expression range for a selected gene across all tissues in all available normal and tumor RNA Seq data [4]. Our application is similar in the case of trying to assess DE in cancer gene data, however, our application uses data from different repositories and it’s more customizable based on the type of cancer the user is interested in. Another application by Zang et. al., called GEPIA which is a web server for cancer and normal gene expression profiling and interactive analyses. GEPIA provides key interactive and customizable functions including differential expression analysis, profiling plotting, correlation analysis, patient survival analysis, similar gene detection and dimensionality reduction analysis [5]. Our application is more specific than GEPIA in terms of using the DE gene data to evaluate and assess the potential biomarkers for cancer.

2 DATA SET

The data analyzed by our software was sourced from CuMiDa, a database featuring data sets of gene expression in humans with various types of cancer. This database was created in 2019 with the intention of training and validating machine learning algorithms to be used in cancer research (Feltes et al., 2019) [6]. CuMiDa consists of 78 gene microarray data sets that represent the cleanest and most discernible data from the GEO database, a comprehensive database of gene expression studies used to assess gene expression data within the framework of existing data. These 78 data sets were selected based on stringent exclusion criteria, namely that no treatments were applied to subjects and that appropriate sample sizes were available for each classification group within the study.

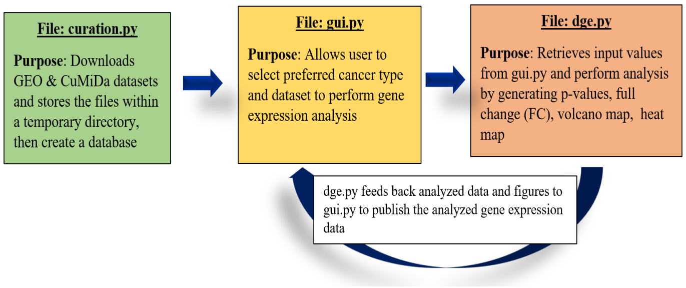
To align with the scope of this project, we further limited the CuMiDa data sets into those with binary classifications: tumor or non-tumor. We also chose to restrict our selection to data sets generated on the *Affymetrix GeneChip Human Genome U133 Plus 2.0* Array (GPL570). After filtering by classification, 21 data sets met the criteria to be used in this analysis. The types of cancer, number of samples, and GEO Accession number in the filtered data sets are enumerated in Table 1.

Cancer Type	Samples	GEO Accession
Bladder	85	GSE31189
Breast	116	GSE42568
Breast	12	GSE26910
Colorectal	63	GSE8671
Colorectal	33	GSE32323
Colorectal	18	GSE41328
Gastric	24	GSE19826
Gastric	20	GSE79973
Leukemia	46	GSE71935
Liver	91	GSE62232
Lung	114	GSE19804
Lung	90	GSE18842
Lung	48	GSE27262
Pancreatic	51	GSE16515
Prostate	49	GSE46602
Prostate	17	GSE55945
Prostate	12	GSE26910
Renal	143	GSE53757
Renal	28	GSE66270
Throat	103	GSE42743
Throat	40	GSE12452

Table 1. Selected CuMiDa data set information. The above table enumerates the cancer types, sample sizes, and GEO Accession numbers of the data sets that utilized binary classification in the CuMiDa database. These data sets were analyzed individually in the GUI.

3 Methods and Implementation

A graphical user interface (GUI) was developed to allow users to perform differential gene expression analysis on any of the 21 selected data sets. For the user-selected data set, differentially expressed genes were identified by conducting a Welch's t-test across all 54, 675 genes in the data set. The resulting p-values were corrected for false discovery rate via the Benjamini-Hochberg method. The fold-change value for each gene was calculated by subtracting the mean *expression* of the tumor group from the *normal* group (since data from CuMiDa are already log-transformed). Default thresholds of 0.05 for the adjusted p-values and 2.0 for fold-changes are provided in the GUI, though the user is able to alter these thresholds.
The *CaBiD* GUI was implemented in Python 3.10 using the *wxPython* library which is compatible with both Windows and MacOS. The figures - volcano plots and heatmaps - are generated using the *Seaborn* and *Matplotlib* packages. To improve the responsiveness of the GUI, we decided to exclude low variance genes (keep only the 99th percentile) when generating the heatmap and the corresponding dendrograms. The *GEOparse* package was used to download data from GEO. The *pandas, scipy* and *statsmodels* packages were used to perform the differential expression. Figure 2 shows a flowchart depicting the function of each script in our GUI.



File: curation.py

Purpose: Downloads GEO & CuMiDa datasets and stores the files within a temporary directory, then create a database

File: gui.py

Purpose: Allows user to select preferred cancer type and dataset to perform gene expression analysis

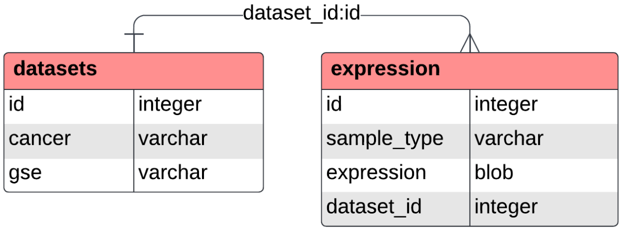
File: dge.py

Purpose: Retrieves input values from gui.py and perform analysis by generating p-values, full change (FC), volcano map, heat map

dge.py feeds back analyzed data and figures to gui.py to publish the analyzed gene expression data

Figure 2. GUI design path. The above figure illustrates the purpose and function of each GUI component as well as the flow of data.

The *curation* module was developed to download the selected expression matrices from CuMiDa and replace the probe IDs with corresponding GenBank identifiers from the GPL file. Once the files are downloaded, a SQLite database is created to store the gene expression information. The *data sets* table stores details about the GSE numbers and cancer types for each data set, and the *expression* table contains the expression arrays for each patient across all data sets. Due to the size of the expression arrays (1 x 54675), we chose to serialize the expression values using the *pickle* library and stored the binary strings as *BLOBs* in the database. The *curation.CuMiDa.retrieve_data set* function was written to deserialize the results of select queries to the database and return a series matrix as a *pandas* data frame. Figure 3 shows the ER diagram for our database.



datasets	
id	integer
cancer	varchar
gse	varchar

expression	
id	integer
sample_type	varchar
expression	blob
dataset_id	integer

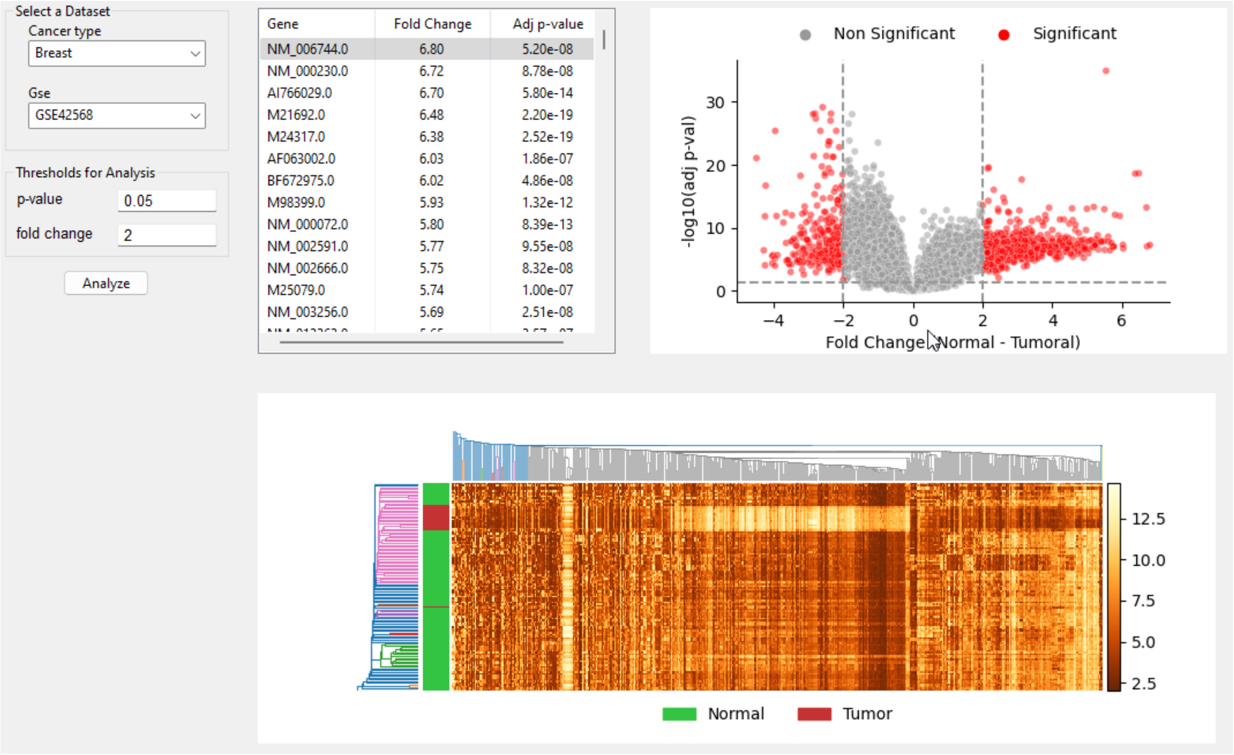
Figure 3. Database ERD. The ERD shows the schema for our SQLite database, including the *data sets* and *expression*.

4 Experiments and Results

Our tool can be used to evaluate each of the 21 data sets present in the filtered CuMiDa database. The GUI allows the user to select the cancer type to be analyzed and any of the corresponding GSE identifiers associated with that cancer type. Additionally, the user may enter their preferred p-value and fold-change thresholds for identifying significantly differentially expressed genes. The GUI generates a volcano plot of gene expression between normal and tumoral samples with regard to the FC and -log(p-value). The input values determine the boundary lines shown on the Graph in Figure 4.

Figure 4 shows a screenshot of our GUI illustrating the analysis results for the GSE42568 breast cancer data set. This data set contains samples which exhibit numerous significantly differentially expressed genes and demonstrates that our GUI enables easy visual identification of said genes. Data points in red mark genes that are expressed significantly differently between normal and tumoral data. It is clear that tumor cells both over- and under-express certain genes relative to normal cells considering the significance in both directions of fold change. The heatmap in Figure 4 features dendrograms grouping cells with similar expression and genes with similar expression intensities. The leftmost solid bar identifies either normal or tumor type cells for the expression row. The heatmap in Figure 5 illustrates noticeably higher gene expression of middle genes and lower expression in the rightmost genes in tumor cells as compared to normal cells. The heatmap is a useful illustration because it confirms differential gene expression and provides rationale for further investigation into which genes are differentially expressed.

Table 2 represents the five most differentially expressed genes between normal and tumor cells in this data set. It provides the FC and adjusted p-value of each gene to demonstrate the overall significance. These genes had the highest y-values on the graph in Figure 4 and likely had the largest intensity discrepancy between normal and tumor types on the heatmap in Figure 5.



Select a Dataset
Cancer type: Breast
Gse: GSE42568

Thresholds for Analysis
p-value: 0.05
fold change: 2

Gene	Fold Change	Adj p-value
NM_006744.0	6.80	5.20e-08
NM_000230.0	6.72	8.78e-08
AI766029.0	6.70	5.80e-14
M21692.0	6.48	2.20e-19
M24317.0	6.38	2.52e-19
AF063002.0	6.03	1.86e-07
BF672975.0	6.02	4.86e-08
M98399.0	5.93	1.32e-12
NM_000072.0	5.80	8.39e-13
NM_002591.0	5.77	9.55e-08
NM_002666.0	5.75	8.32e-08
M25079.0	5.74	1.00e-07
NM_003256.0	5.69	2.51e-08

Figure 4. GUI window with analysis results. This figure shows the GUI populated with the results of differential expression analysis for a breast cancer data set (GSE42568). The table shows a list of all differentially expressed genes in the data set.The volcano plot shows the identified genes in red. The heatmap depicts the expression for a subset of genes; lighter colors correspond to higher gene expression.

Gene	Fold Change	Adj P-Value
FABP4	5.54	1.26E-35
CEP55	-2.59	7.37E-30
GINS1	-2.81	7.69E-29
NUSAP1	-2.35	7.69E-29
COL110A	-2.87	9.47E-29

Table 2. Top 5 differentially expressed genes. The table shows the top 5 significantly differentially expressed genes identified in the breast cancer data set (GSE42568) with p-value and fold-change thresholds of 0.05 and 2 respectively.

5 Discussion

The results of our GUI show that it is a viable way to establish preliminary differential gene expression in binary normal/tumor data sets. The figures can be visually inspected to assess the presence of differential gene expression and evaluate which genes had the most dissimilar expression between normal and tumor groups. The most differentially expressed genes, enumerated in Table 2, are known to be biologically associated with breast cancer. CEP55, GINS1, NUSAP1, and COL10A1 have all been found to be upregulated in breast cancer tissue [7-10]. However, FABP4 has been loosely shown to be a minor player in obesity-related breast cancer, so its high significance in our analysis disagrees with the literature [11]. However, FABP4 has been shown to prevent the proliferation of endometrial cancer cells so perhaps it could play a similar role in breast cancer [12]. That mechanism would justify lack of expression of FABP4 in tumor cells in our analysis.

The primary limitation of this project is that we used a very specific data set with binary classification types and "clean" data. This GUI needs to analyze numerous data sets that cannot be generalized to a larger population. Additionally, we only analyzed a small subset of cancer types, each with previously known differential gene expression. More analysis is needed to determine if the GUI works as a general analysis tool.

More research needs to occur with respect to the roles specific genes play in breast (and other types of) cancer. The literature on some genes is sparse, which poses a challenge when evaluating the validity of our output. Our GUI can further be improved by extending the number of datasets included - including additional datasets generated on other microarray platforms would be an ideal place to start. We could also implement the ability to retrieve gene information from the *Ensembl Rest API* which would aid biologists in deriving greater insights from the results in the GUI. Finally, the approach used for the differential expression analysis is not very robust; one way of improving the GUI would be to implement the use of a more widely implemented model such as *limma* (which is implemented in R) [13].

6 References

[1] Xue, J. M., Liu, Y., Wan, L. H., & Zhu, Y. X. (2020). Comprehensive analysis of differential gene expression to identify common gene signatures in multiple cancers. Medical science monitor: international medical journal of experimental and clinical research, 26, e919953-1.

[2] Rodriguez-Esteban, R., & Jiang, X. (2017). Differential gene expression in disease: a comparison between high-throughput studies and the literature. BMC medical genomics, 10(1), 1-10.

[3] Liang, P., & Pardee, A. B. (2003). Analysing differential gene expression in cancer. Nature Reviews Cancer, 3(11), 869-876.

[4] Bartha, Á., & Győrffy, B. (2021). TNMplot. com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues. International journal of molecular sciences, 22(5), 2622.

[5] Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research, 45(W1), W98-W102.

[6] Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. Journal of Computational Biology, 26(4), 376-386.

[7] Kalimutho, M., Sinha, D., Jeffery, J., Nones, K., Srihari, S., Fernando, W. C., ... & Khanna, K. K. (2018). CEP 55 is a determinant of cell fate during perturbed mitosis in breast cancer. EMBO molecular medicine, 10(9), e8566.

[8] Nakahara, I., Miyamoto, M., Shibata, T., Akashi‐Tanaka, S., Kinoshita, T., Mogushi, K., ... & Ohta, T. (2010). Up‐regulation of PSF1 promotes the growth of breast cancer cells. Genes to Cells, 15(10), 1015-1024.

[9] Qiu, J., Xu, L., Zeng, X., Wu, Z., Wang, Y., Wang, Y., ... & Du, Z. (2021). NUSAP1 promotes the metastasis of breast cancer cells via the AMPK/PPARγ signaling pathway. Annals of Translational Medicine, 9(22).

[10] Zhang, M., Chen, H., Wang, M., Bai, F., & Wu, K. (2020). Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. Bioscience reports, 40(2).

[11] Zeng, J., Sauter, E. R., & Li, B. (2020). FABP4: a new player in obesity-associated breast cancer. Trends in molecular medicine, 26(5), 437-440.

[12] Wu, Z., Jeong, J. H., Ren, C., Yang, L., Ding, L., Li, F., ... & Lu, J. (2021). Fatty Acid–Binding Protein 4 (FABP4) Suppresses Proliferation and Migration of Endometrial Cancer Cells via PI3K/Akt Pathway. OncoTargets and therapy, 14, 3929.

[13] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47–e47.