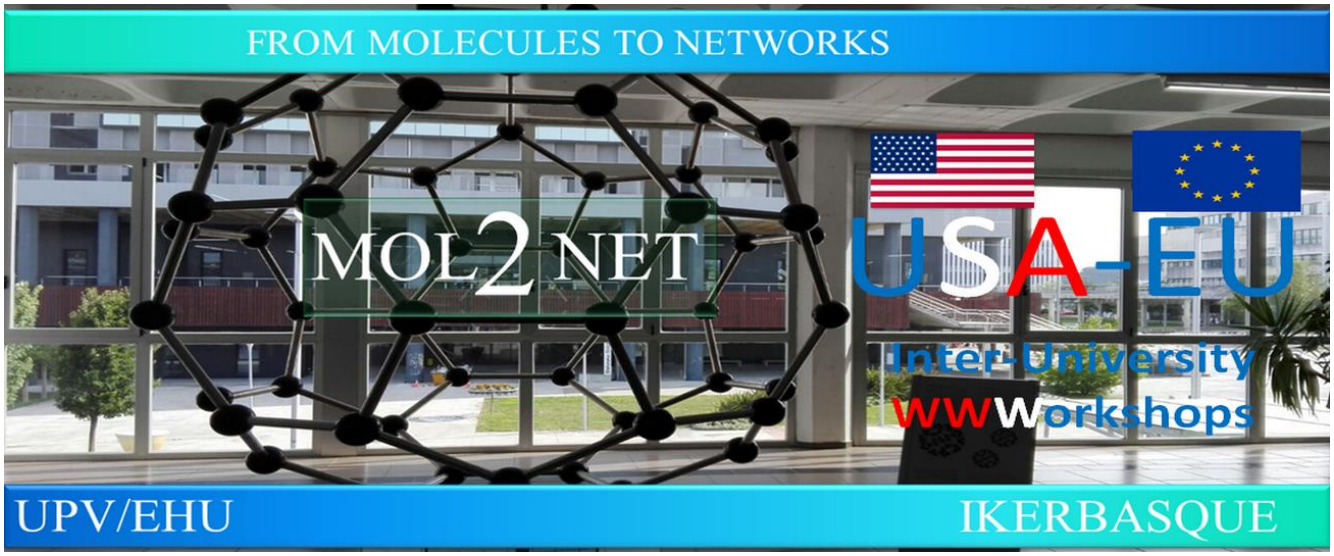
MDPI sciforum

MOL2NET, International Conference Series on Multidisciplinary Sciences



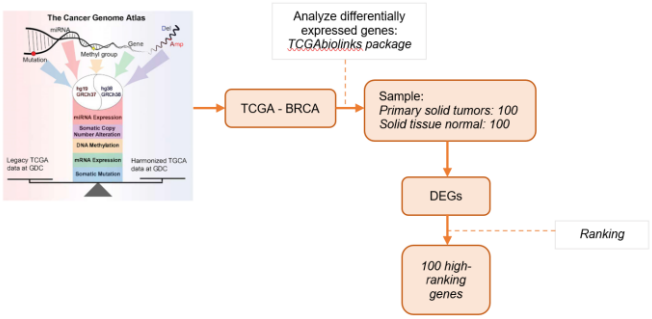

# Using *TCGAbiolinks* package in ranking breast cancer genes from The Cancer Genome Atlas (TCGA) to predict disease-associated genes

*NGUYEN Thi Ngoc Ha*[a], *BUI Tu Uyen*[b]

[a] *General and Inorganic Department, Hanoi University of Pharmacy, 13 – 15 Le Thanh Tong Street, Hoan Kiem District, Hanoi, Vietnam*

[b]*Hanoi Amsterdam High School for the Gifted*

**Graphical Abstract**




**Abstract.**

*Predicting genes which may associate with disease is one of the important goals of biomedical research. There have been many computational methods developed to rank genes involved in a particular disease. However, due to the complex relationship between genes and the diseases, many genes that cause genetic diseases have not yet been discovered. The problem of ranking genes to identify the disease-associated gene has drawn attention of many researchers. The Genomic Data Commons (GDC) Data Portal is a platform that contains different cancer*

| | |
|---|---|
| | *genomic studies. Such platforms have often been the primary focus on the data storage and they do not provide a comprehensive toolkit for analyses. In this study, we used the new functions of the R/Bioconductor TCGAbiolinks package to search and analyze differentially expressed genes between breast cancer samples with primary solid tumors (TP) and solid tissue normal (NT) samples to retrieve list of 100 high-ranking genes associated with breast cancer.* |

## Introduction

Cancer is one of the leading causes of mortality worldwide. It is a complex disease of which there are many causal mechanisms. Numerous large-scale studies have been conducted using state-of-the-art genome analysis technologies. One of the most important projects is The Cancer Genome Atlas (TCGA), which started in 2006 as a pilot project aiming to collect and analyze an unprecedented amount of clinical and molecular data including 33 tumor types spanning over 11,000 patients. This project has subsequently generated more than 2.5 petabytes of publicly available data over the past decade [2]. Publicly funded by The National Institute of Health (NIH), TCGA has made numerous discoveries regarding genomic and epigenomic alterations that are candidate drivers for cancer development. This was achieved through the creation of an "atlas" by applying large-scale genome-wide sequencing and multidimensional analyses. In 2016, TCGA was moved under the umbrella of the broader repository Genomic Data Commons (GDC) Data Portal together with other studies [3]. Many tools have been developed to interface with TCGA data and to help with the aggregation, pre- and post-processing of the datasets. Among them, *TCGAbiolinks* was developed as an R/Bioconductor package to address the challenges of comprehensive analyses of TCGA data. Software packages such as *TCGAbiolinks* regularly require enhancements and revisions in light of new biological or methodological evidence from the literature or new computational requirements, imposed by the platforms where the data are stored [6]. In this study, we use *TCGAbiolinks* package to search and analyze differentially expressed genes between primary solid tumors and solid tissue normal of breast cancer samples to retrieve a list of 100 high-ranking genes associated breast cancer.

## Materials and Methods

### *Samples*

Primary solid tumors: 100 samples

Solid tissue normal: 100 samples

### *Methods*

The data retrieval is handled by the three main *TCGAbiolinks* functions: *GDCquery*, *GDCdownload* and *GDCprepare*. This allows the user to interface with three main platforms: i) TCGA, ii) TARGET and, iii) The Cancer Genome Characterization Initiative (CGCI) (https://ocg.cancer.gov/programs/cgci). *TCGAbiolinks* also allows the user to interface with different -omics data including genomics and transcriptomics, clinical and pathological data, information on drug treatments, and subtypes.

*GDCprepare* allows the user to prepare gene expression data for downstream analyses. This step is done by restructuring data into a SummarizedExperiment (SE) object [5] that is easily managed and integrated

with other *R/Bioconductor* packages or as a dataframe for other forms of data manipulation, with which users can operate even decoupled from the *TCGAbiolinks* package.

The function *TCGAanalyze_Normalization* in *TCGAbiolinks* package adopts the EDASeq protocol [7] to apply between-lane normalization to adjust for distributional differences between samples or within-lane normalization.

According to the standard defined by the TCGA consortium, 60% tumor purity is the recommended threshold for analysis [8]. Thus, we apply a filtering step using the *TCGAtumor_purity* function of *TCGAbiolinks* whereby tumor samples that show a purity of less than 60% median CPE are excluded from the analysis.

*TCGAbiolinks*' function *TCGAanalyze_DEA* is used to analyze differentially expressed genes [4]. Using this function, we apply *edgeR* package in Bioconductor to analyze differentially expressed genes with |log(FoldChange)| > 1.

**Results and Discussion**

**Table 1.** 100 genes with highest |log(FC)|.

| Ranking | Gene name | \|log(FC)\| | Ranking | Gene name | \|log(FC)\| | Ranking | Gene name | \|log(FC)\| |
|---|---|---|---|---|---|---|---|---|
| 1 | LALBA | 12.10 | 34 | ADIPOQ | 6.82 | 67 | AQP7P1 | 5.79 |
| 2 | CPLX2 | 10.20 | 35 | SPHKAP | 6.81 | 68 | S100P | 5.79 |
| 3 | CDC20B | 10.06 | 36 | PCSK1 | 6.75 | 69 | CASP14 | 5.79 |
| 4 | CGA | 9.96 | 37 | SLC30A8 | 6.68 | 70 | STAC2 | 5.78 |
| 5 | CST4 | 9.89 | 38 | SYT13 | 6.68 | 71 | FABP4 | 5.72 |
| 6 | CSN2 | 9.80 | 39 | PCOLCE2 | 6.65 | 72 | GPD1 | 5.70 |
| 7 | CST5 | 9.00 | 40 | HEPACAM | 6.55 | 73 | SLC24A2 | 5.68 |
| 8 | CSN3 | 8.88 | 41 | GRM4 | 6.47 | 74 | KCNJ3 | 5.67 |
| 9 | MAGEA1 | 8.73 | 42 | SPDYC | 6.41 | 75 | FOXJ1 | 5.65 |
| 10 | MYOC | 8.71 | 43 | CBLN2 | 6.39 | 76 | SMYD1 | 5.60 |
| 11 | SULT1C3 | 8.66 | 44 | MMP1 | 6.33 | 77 | BMPR1B | 5.59 |
| 12 | CLEC3A | 8.64 | 45 | COL11A1 | 6.32 | 78 | CA4 | 5.57 |
| 13 | CA6 | 8.58 | 46 | MS4A15 | 6.30 | 79 | PLIN1 | 5.57 |
| 14 | CSN1S1 | 8.35 | 47 | C14orf180 | 6.26 | 80 | TNMD | 5.55 |
| 15 | NPY2R | 8.20 | 48 | MMP11 | 6.26 | 81 | TRPA1 | 5.53 |
| 16 | EPYC | 8.14 | 49 | PRAME | 6.23 | 82 | SLCO1A2 | 5.53 |
| 17 | PAX7 | 7.83 | 50 | ROPN1 | 6.15 | 83 | SLC5A8 | 5.52 |
| 18 | DLK1 | 7.80 | 51 | WIF1 | 6.13 | 84 | FABP7 | 5.47 |
| 19 | HS3ST4 | 7.77 | 52 | PLIN4 | 6.10 | 85 | TRDN | 5.47 |
| 20 | LRTM2 | 7.58 | 53 | ANGPTL7 | 6.03 | 86 | SFRP1 | 5.43 |
| 21 | COL10A1 | 7.48 | 54 | MRAP | 6.03 | 87 | ALDH1L1 | 5.39 |
| 22 | LEP | 7.39 | 55 | ACTL8 | 6.02 | 88 | TMEM145 | 5.36 |
| 23 | CIDEA | 7.29 | 56 | PENK | 6.02 | 89 | PYDC1 | 5.36 |
| 24 | CHRNA9 | 7.25 | 57 | BMP3 | 6.01 | 90 | SAA1 | 5.35 |
| 25 | CST1 | 7.24 | 58 | PI16 | 5.99 | 91 | AADAC | 5.35 |
| 26 | CSF3 | 7.13 | 59 | AQP7 | 5.93 | 92 | ASCL1 | 5.32 |
| 27 | SCARA5 | 7.08 | 60 | SLC7A10 | 5.92 | 93 | ADRA1A | 5.32 |
| 28 | S100A7A | 7.05 | 61 | ACADL | 5.92 | 94 | LGALS12 | 5.32 |
| 29 | GLYAT | 6.98 | 62 | CST2 | 5.91 | 95 | LYVE1 | 5.31 |
| 30 | S100A7 | 6.91 | 63 | CHRDL1 | 5.88 | 96 | C10orf90 | 5.31 |
| 31 | CIDEC | 6.88 | 64 | CD300LG | 5.86 | 97 | TF | 5.29 |
| 32 | ADH1B | 6.84 | 65 | HOXB13 | 5.81 | 98 | KCNJ16 | 5.29 |
| 33 | MMP13 | 6.83 | 66 | RBP4 | 5.80 | 99 | SLC19A3 | 5.28 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | 100 | APOB | 5.28 |

**Conclusions**

In conclusion, we have produced a list of 100 breast cancer genes with the highest differential expression between with primary solid tumors (TP) and solid tissue normal (NT) of breast cancer samples from The Cancer Genome Atlas.

**References**

1. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Comput Sci.* 2: e67 (2016).
2. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet. Nature Publishing Group*. 45, 1113–1120 (2013).
3. Grossman RL, Heath A, Murphy M. A Case for Data Commons: Toward Data Science as a Service. *Comput Sci Eng*. 18, 10–20 (2016).
4. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15: R29 (2014).
5. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 28, 882–883 (2012).
6. Zhang H. TSVdb: a web-tool for TCGA splicing variants analysis. *BMC Genomics*. 1–7 (2018).
7. Risso, D., Schwartz, K., Sherlock, G. et al. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics.* 12, 480 (2011).
8. Silva TC, Colaprico A, Olsen C, Bontempi G, Ceccarelli M, Berman BP, et al. TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*. 7, 439 (2018).