Gene Expression based classification for identifying the significant genes in Non-small Cell Lung Cancer samples

Neelambika B Hiremath (neelambika@ieee.org)
J S S Academy of Technical Education Bengaluru https://orcid.org/0000-0001-8833-1290
Dayananda P
J S S Academy of Technical Education Bengaluru https://orcid.org/0000-0001-8445-3469
Mrityunjaya V Latte
J S S Academy of Technical Education Bengaluru https://orcid.org/0000-0002-9624-0556

Research Article

Keywords: differential gene expression, next-generation Sequencing, RNA sequence, transcripts, non-small cell lung cancer, classification, data enrichment, infomap algorithm

Posted Date: July 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1843296/v1

Abstract

Lung cancer is the most common and fatal type of cancer. NSCLC refers to any kind of epithelial lung cancer that isn't small cell lung cancer (SCLC), which results for 85 percent of lung cancer cases. Differential gene expression is a type of gene analysis in which the RNA sequence data from next-generation sequencing is shown for any quantitative changes in the experimental data set's levels. Transcriptome analysis focuses on obtaining transcript statistics from a gene transcript file with a fold change of genes on a normalised scale in order to find quantitative differences in gene expression levels between the reference genome and NSCLC samples. The data has a significant clinical influence in terms of identifying and characterising candidate genes in order to validate them. The resultant data set and the plot display depicts the significant candidate genes in the respective location which are significant in expressing their changes in samples of NSCLC. The samples are differentiated with prominent gene labels of NSCLC disease samples. The significant values of this quantized analysis on read count data of expression, data tables prompt the candidate genes data set of NSCLC samples also the results explain the differential expression of particular samples across samples from genders namely male and female. The current research experiment focuses on the computational difficulty of read, search, match, and data enrichment of unstructured data with the goal of classifying biomarkers based on differential expression results and pathways found by classification algorithms.

Introduction

Lung cancer is the most common cancer, claiming the lives of 1.6 million people each year [1]. Only around 15% of cancer patients are detected at an early stage, with the rest of patients being diagnosed later [2]. Because of differences in intrinsic oncogenic signalling pathways, NSCLC (non-small cell lung cancer) is a common kind of lung cancer with a lot of molecular heterogeneity.

Nearly 80–90% of them develop (NSCLC), with lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) these are the most prevalent kinds. Adenocarcinoma involves women, non-smokers, and most eastern groups, whereas small cell lung cancer affects Caucasian males [3]. Tobacco use might be a significant factor. Patients with small cell lung cancer are found not just among smokers. Non-smokers are more exposed to second-hand smoke, pollution, occupational carcinogens, and genetics than smokers are. In comparison to histological categorization, genes have a major influence on NSCLC diagnosis, treatment, and prognosis. Treatment for lung cancer based on the epidermal growth factor receptor (EGFR) gene genotype reduces NSCLC progression and death. Based on molecular aspects of the tumour, including as activation of epidermal growth factor (EGF) receptor pathways, anti-EGF receptor treatment with tyrosine kinase inhibitors is utilised to treat NSCLC [4]. MYC deregulation, MET/HGF activation, and RB1/TP16/CCND1 pathway inactivation have all been associated to the development of NSCLC. Despite the fact that the mechanism of NSCLC remains unknown [5].

Gene expression at different phases of carcinogenesis and the molecular mechanism of anticancer medicines are studied using differentially expressed genes. Multiple gene expression investigations are

carried out on NSCLC, however the sample size, tumour stages, grading, and other variables all have an impact on the results[6]. Because it offers biological information about the cells, RNA sequencing would be excellent for discovering the key gene. There are also a few bioinformatics detailed tools for RNA sequencing, as well as significant stumbling blocks in cancer applications. The procedure includes raw data quality control, transcription assembly and read alignment, expression qualification, and differential expression analysis. In this case, a variety of bioinformatic methods are used [7]. As a result, next-generation sequencing is important in RNA sequencing, because of research into cancer biomarkers and differential gene expression, evolution, cancer heterogeneity, cancer medication resistance, the microenvironment, and immunotherapy, to name a few., all require next-generation sequencing[8] .

Materials And Methods

Datasets

This work is based on real biological samples data, the dataset is available on Sequence read Archive database maintained by the National Centre for Biotechnology Information. The accession reads from the project SRP117020 were selected for analysing representative samples on both genders (Male & Female) with age criteria aged more than 50 years. The data consisted of RNA sequences with distribution of poor to well differentiated adenocarcinomas and squamous cell cancers which were sequenced using Illumina Hiseq2500 [9].

Data Analysis

The investigation began with raw sequencing data in fastq format. NCBI's assembly page (https://www.ncbi.nlm.nih.gov/assembly/GCF 000001405.26/) provided the reference genome.

Read Alignment and Assembly of Transcripts

HISAT2, a rapid and sensitive alignment software, was used to align RNA-seq data to the human reference genome.[10]. The Visual exploration and differences in the expressions were obtained using Omics box (with Edge R in the backend). StringTie was used to obtain the transcripts from the alignment files. As a result, from StrintTie the gtf files are obtained and merged together. Count files were obtained from the Omics box (Using HTseq count in the backend) [11],[12],[13],[14].

Differential gene expression and Pathway analysis

The transcripts and expression levels obtained from StringTie were subjected to get the differentially expressed genes which was performed using Omics box. The package uses statistical methods to get the differentially expressed genes.

Pathway and Gene ontology analyses were performed on the list of differentially expressed genes collected. This was carried out using KOBAS 3.0 annotation module [15]. The annotate module accepts

gene sequences in FASTA format or Gene ID as input and presently covers 5944 different species to run the annotations. Additionally, Pathways were obtained from EnrichR[16].

The following parameters are followed for Differential expression analysis

- CPM filter: Normalised expression units are required to overcome technological limits in sequenced data, the gene length and depth of sequencing and to make gene expressions directly comparable within and between samples. This is because a larger read count for a gene expressed at the same level arises from increased sequencing depth, whereas gene length differences result in uneven read counts for genes expressed at the same level (longer the gene more the read count). Counts per million mapped reads are the number of sequenced fragments multiplied by one million.

CPM Filter: It's standard practice to filter and eliminate genes with low counts across libraries; the filtering is done on a count per million bases to account for changes in library size (transcript bp) between samples. In a sample with 6 million reads, a CPM of 1 corresponds to a count of 6. In order to pass the filter, the gene's CPM must be higher than the filter level in at least one sample.

TMM: Trimmed Mean of M-values (TMM) is the normalization method used in edgeR.

All other samples are referred to as test samples, while reference samples have the average expressions that are the closest to the mean of all samples. The weighted mean (weighted by expected asymptotic variance) of log ratios between the test and reference genes from a gene set that includes the most and least expressed genes, as well as genes with the highest and lowest log ratios, is used to calculate the scaling factor for each test sample.

The main experimental variable is: is based on the parameters set out in this project's scenario. Male and female conditions exist. For Male/Female, the log2fold change is determined.

Male is the primary condition, while Female is the secondary condition.

The majority of the p-values we generate are predicated on the assumption that our test statistic fits into a certain distribution. If the conditions are satisfied, these distributions are a suitable technique to derive p-values. However, this isn't the only technique to compute a p-value. Exact tests compute a p-value experimentally rather than generating a theoretical probability based on a distribution. To determine the likelihood of gene expression, the p value is determined for the conditions given (male as reference and female as contrast).

Enrichment analysis

Based on the pathway analysis results, functional enrichment analysis was performed on genes implicated exclusively in the NSCLC disease pathway. Enrichment analysis was carried out with the help of the KOBAS 3.0 enrichment module and EnrichR. Enrichment identifies statistically significant pathways, illnesses, and gene ontology.

Obtaining the Candidate genes

The mutations responsible for NSCLC were retrieved using the ClinVar database[17], and then genes implicated in mutations were considered, with their matching expression levels derived using Omics box differential expression findings.

Results And Discussion

Read Alignment and Assembly of transcripts

Reads alignment to the reference genome is a very important step. When the alignment does not occur properly transcriptome reconstruction becomes difficult for genes especially expressed at lower levels. Alignment of the reads to the human reference genome gave SAM files whose alignment rates were above 85% which was later converted to BAM which is the compressed binary version of SAM files. Later when alignment files were subjected to transcriptome reconstruction using StringTie, the annotation file was obtained with all the expression levels of all genes and transcripts. Figure 1 depict the flow of processes.

Differential Expression and pathway analysis

The relevant packages under Omics box were loaded, and the phenotype data was loaded in .csv format which contained sample ID and Gender of the sample. Omics box gives two tables with Genes and its counts and differentially expressed genes. The differentially expressed genes and gene counts is shown in **supplementary table S1** and **supplementary table S2**. The table gives three types of values. One, fold change value referring to the ratio between expression levels in male and female. Two, the log of the fold change value. The values that are <1 indicate that it is expressed at a lower level and values >1 indicate that it is expressed at a higher level. Lastly, P-value indicating the probability of the expression value. Figure 2 displays the graph.

A typical differentiation of interest between 2 treatment conditions is adjusted log fold-change vs P-value. A volcano plot is displayed in Figure 4, it looks like explosion of volcano with data source points, with data points concentrated at the origin and a fanning effect distant from it. Volcano plots demonstrate the statistical significance of the difference in relation to the amount of the difference for each individual gene in the comparison, typically using the negative base-10 log and base-2 log fold-change methods. Because P-values have a negative transformation, the lower the P-value, the higher the data point on the y-axis. In general, Threshold indicators for adjusted P-values are included in volcano plots to identify which genes are statistically differentially expressed based on the adjusted P-value of their difference between treatments. The data points closer to 0 denote genes with equal or identical mean expression levels, while the extreme values of the log fold-change along the x-axis reflect more substantial changes. Volcanic charts anticipate a high degree of dispersion, as the name implies. A higher dispersion indicates that gene expression between two treatment groups differs significantly. The majority of data points in a volcanic image are grouped near the origin, which is unusual.

MA plots as depicted in figure 3, are a popular approach to compare two treatments' log fold-change vs mean expression. This is visualised using a scatter plot with base-2 log fold-change on the y-axis and normalised mean expression on the x-axis. Data points with extreme values along the y-axis represent genes with considerably differing expression levels (although, not necessarily differentially expressed). The log fold-change variability of lower mean expression values is generally larger than that of higher expression values. The data points spread out as the graph moves from right to left, generating a fanning effect. Because log fold-changes have standard cut-offs, MA plots will frequently exhibit these cut-offs.

Because it lacks any measure of statistical significance, this image does not readily identify which data points are statistically different expressed. To account for this, data points with P-values below the threshold adjusted will be coloured in some MA plots. A well-constructed MA plot can provide some useful information, with each data point representing a particular gene. A base-2 log fold-change criterion of 1 indicates whether genes in the related comparison have doubled or halved. A MA plot with a lot of data points above the one threshold on the y-axis suggests that a lot of genes are upregulated, whereas a plot with a lot of data points below 1 shows that a lot of genes are downregulated. MA plots often have a fairly equal dispersion relative to the y-axis that tightens when the x-axis is raised. When studying dormant and non-dormant plants, biological significance may indicate a y-axis dispersion that is larger or smaller than usual. In the case where all or most data points on the y-axis are close to 0, the expression patterns of the two treatment groups will be very similar.

When all of the differentially expressed genes were put through pathway analysis with Homo sapiens as the organism and Gene symbol ID mapping as the technique, a total of 5071 genes were found to be involved in various pathways. Different parts of output for a certain gene were shown in the results. One was the gene's involvement in many pathways. Second, the illness in which the genes were implicated. Third, gene ontology was used to identify the biological roles of the genes. A companion document contains the whole route analysis output. To gain a better understanding of the pathways developed, the genes from the Differentially expressed analysis were submitted to pathway analysis using EnrichR, which yielded substantial hits of pathways; the details are shown in **Supplementary table S3**.

PATHWAY CLASSIFICATION

The pathways were obtained from KOBAS and EnrichR and the genes causing NSCLC were classified. Details of it is given in **supplementary Table S3**. The genes involved in NSCLC were subjected to Biomarker analysis which was carried out using GEPIA. From the results obtained from ClinVar database only for obtaining the Single nucleotide polymorphisms and the corresponding expression analysis from Omics Box it is seen that the genes, BRAF, NRAS, KRAS, EGFR and MAP2K1 have been involved in multiple mutations as **shown in supplementary table S4** which gives vast idea about the SNPs. The table consists of detailed information about the nucleotide change and the corresponding amino acid change, chromosome number and the position in the human genome (GRCh38). The expression values of genes

involved in mutations are highlighted in Table 1. In addition, Supplementary table S5 lists all SNPs linked to NSCLC. The table 1 displays the single nucleotide polymorphism with genes.

Table 1 Biomarker genes with expression and Polymorphism

GENE NAME	Log2fold	SNP
BRAF	-0.2327	G464V, G427V, G376V, G442V, G412V, G430V, G467V, G504V
NRAS	0.38354	Q61K
KRAS	0.17279	G12C
EGFR	-0.5007	V834L, V567L, V781L, V789L

The figure 5 shows the Biomarkers which are Significant for NSCLC. The graph depicts the expression of each gene as well as its various cancer connections. All the genes show the significant expression in Lung cancer (LUAD).

DISCUSSION

In NSCLC, the epidermal growth factor receptor was revealed to be the initial oncogenic target (EGFR). In Asian patients, the frequency of EGFR mutations ranges from 40 percent to 11-17 percent, while in Caucasian patients, the frequency ranges from 11 to 17 percent. Nearly all EGFR mutations include exons 18 to 21. Exon 19 (del 19) minor in-frame deletions contribute for 40-50 percent of EGFR mutations, while exon 21 p.Leu858Arg amino acid changes account for 30-40 percent[18]. BRAF mutations can be identified in 2 to 8% of NSCLC patients. 50% of all BRAF mutations are caused by the BRAF exon 15 p.Val600Glu activating mutation. Exons 11 and 15 have additional mutations that are either triggering (p.Gly469X, p.Leu597Arg, or p.Lys601Glu) or defective (p.Gly466Val, p.Asp594X, p.Gly596Cys) mutants. Single BRAF inhibitors (e.g., vemurafenib or dabrafenib) cause cell cycle arrest and death in p.Val600Glu mutated-NSCLC, as predicted[19]. KRAS-activating mutations are discovered in about 30% of cancer patients and are used as a biomarker for exclusion. Smokers are more likely to have KRAS-mutated tumours, which seldom have other drug-related drivers. Drugs targeting the most common KRAS mutation in lung cancer are being developed (p.Gly12Cys). The kind of KRAS mutation can potentially

offer details about the severity of the disease or medication sensitivity[20]. The G12D mutation, for example, has been associated to a better prognosis in NSCLC than the G12V or G12R mutations.

PIK3. It affects a protein that aids in the growth and survival of lung cancer cells. According to Trusted Source, PIK3 mutations were found in up to 4% of those with NSCLC.

Conclusion

The most dependable approaches are used in the current protocol. After aligning all reads in the initial data sets to the whole genome and then removing only those reads that aligned to the standard human reference genome and their mates, we expect a mapping rate of close to 100% for the reads in our reduced data set. When the annotation file is used in the merge stage, all transcripts in the annotation file, including those with zero expression levels, as well as any novel transcripts, are included. When the annotation file is available, we encourage utilising it since it aids StringTie in accurately assembling transcripts, especially those expressed at low levels. The present classification technique includes machine learning techniques that aid in the classification of biomarkers linked to NSCLC. These categorization algorithms aid in reliable biomarker assessment, as well as rich visualisation and enrichment. In this study, seven new biomarkers were discovered, and their visualisations and interactions with other cancer kinds were studied. It was discovered that the biomarkers listed had the highest percentage of transcripts mapped for lung cancer. The discovered biomarkers are novel for both male and female samples and seen to be significant in both the genders. This experiment has proposed the following genes: BRAF, NRAS, KRAS, EGFR, MAP2K1, PIK3CA, MET; which are mutated and expressed more for the selected data of patients with Lung adenoma cancer. Clinically, the findings may be used as information, and medicines can be focused. By looking at the gene mapping for the specific action, these medicines can be classified as tumour suppressors or chemotherapy resistant.

Declarations

Acknowledgements

Authors acknowledge JSS Academy of Technical Education Bengaluru research centre for providing the facility to carry out the research.

Author contributions

Conceptualization, Investigation, Methodology by Neelambika B Hiremath. Supervision by Dayananda P, review by Mrithyunjay V Latte, all authors read and approved the final manuscript.

Funding

No funding is received for this research work by any organization.

Competing Interests

The authors declare no competing interests.

Consent for Publication

Not Applicable.

Declarations

Ethics approval not required, no wet lab experiment or clinical trails conducted

References

1. P. Goldstraw *et al.*, “Non-small-cell lung cancer,” Lancet, vol. 378, no. 9804, pp. 1727–1740, 2011.
2. R. S. Herbst, D. Morgensztern, and C. Boshoff, “The biology and management of non-small cell lung cancer,” Nature, vol. 553, no. 7689, pp. 446–454, 2018.
3. Y. Chen and P. Chi, “Basket trial of TRK inhibitors demonstrates efficacy in TRK fusion-positive cancers,” J. Hematol. Oncol., vol. 11, no. 1, pp. 1–5, 2018.
4. E. Kettunen *et al.*, “Differentially expressed genes in nonsmall cell lung cancer: Expression profiling of cancer-related genes in squamous cell lung cancer,” Cancer Genet. Cytogenet., vol. 149, no. 2, pp. 98–106, 2004.
5. Q. Ma *et al.*, “Identification and validation of key genes associated with non-small-cell lung cancer,” J. Cell. Physiol., vol. 234, no. 12, pp. 22742–22752, 2019.
6. J. Wang *et al.*, “Analysis of gene expression profiles of non-small cell lung cancer at different stages reveals significantly altered biological functions and candidate genes,” Oncol. Rep., vol. 37, no. 3, pp. 1736–1746, 2017.
7. S. Wang *et al.*, “Alternatively Expressed Transcripts Analysis of Non-Small Cell Lung Cancer Cells under Different Hypoxic Microenvironment,” *J. Oncol.*, vol. 2021, 2021.
8. M. Zhou *et al.*, “Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications,” Radiology, vol. 286, no. 1, pp. 307–315, 2018.
9. Illumina Inc., “HiSeq 2500 Sequencing System Specification Sheet,” no. Figure 2, pp. 1–4, 2015.
10. Y. Zhang, C. Park, C. Bennett, M. Thornton, and D. Kim, “Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N,” Genome Res., vol. 31, no. 7, pp. 1290–1295, 2021.
11. “Biobam.” [Online]. Available: https://www.biobam.com/omicsbox/. accessed on 20/feb/2022
12. M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: A Bioconductor package for differential expression analysis of digital gene expression data,” Bioinformatics, vol. 26, no. 1, pp. 139–140, 2009.
13. S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea, “Transcriptome assembly from long-read RNA-seq alignments with StringTie2,” bioRxiv, pp. 1–13, 2019.

14. S. Anders, P. T. Pyl, and W. Huber, “HTSeq-A Python framework to work with high-throughput sequencing data,” Bioinformatics, vol. 31, no. 2, pp. 166–169, 2015.
15. D. Bu *et al.*, “KOBAS-i: Intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis,” Nucleic Acids Res., vol. 49, no. W1, pp. W317–W325, 2021.
16. G. V. M. Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang and N. R. C. and A. Ma’ayan, “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,” *BMC Bioinforma.* 2013, vol. 14/128, pp. 1471–2105, 2013.
17. M. J. Landrum *et al.*, “ClinVar: Improvements to accessing data,” Nucleic Acids Res., vol. 48, no. D1, pp. D835–D844, 2020.
18. A. Midha, S. Dearden, and R. McCormack, “EGFR mutation incidence in non-Small-cell lung cancer of adenocarcinoma histology: A systematic review and global map by ethnicity (mutMapII),” Am. J. Cancer Res., vol. 5, no. 9, pp. 2892–2911, 2015.
19. J. W. P. Bracht *et al.*, “BRAF mutations classes I, II, and III in NSCLC patients included in the SLLIP trial: The need for a new pre-clinical treatment rationale,” Cancers (Basel)., vol. 11, no. 9, 2019.
20. F. Barlesi *et al.*, “Routine molecular profiling of patients with advanced non-small-cell lung cancer: Results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT),” Lancet, vol. 387, no. 10026, pp. 1415–1426, 2016.

Supplementary Information

Supplementary Tables 1-5 not available with this version.

Figures

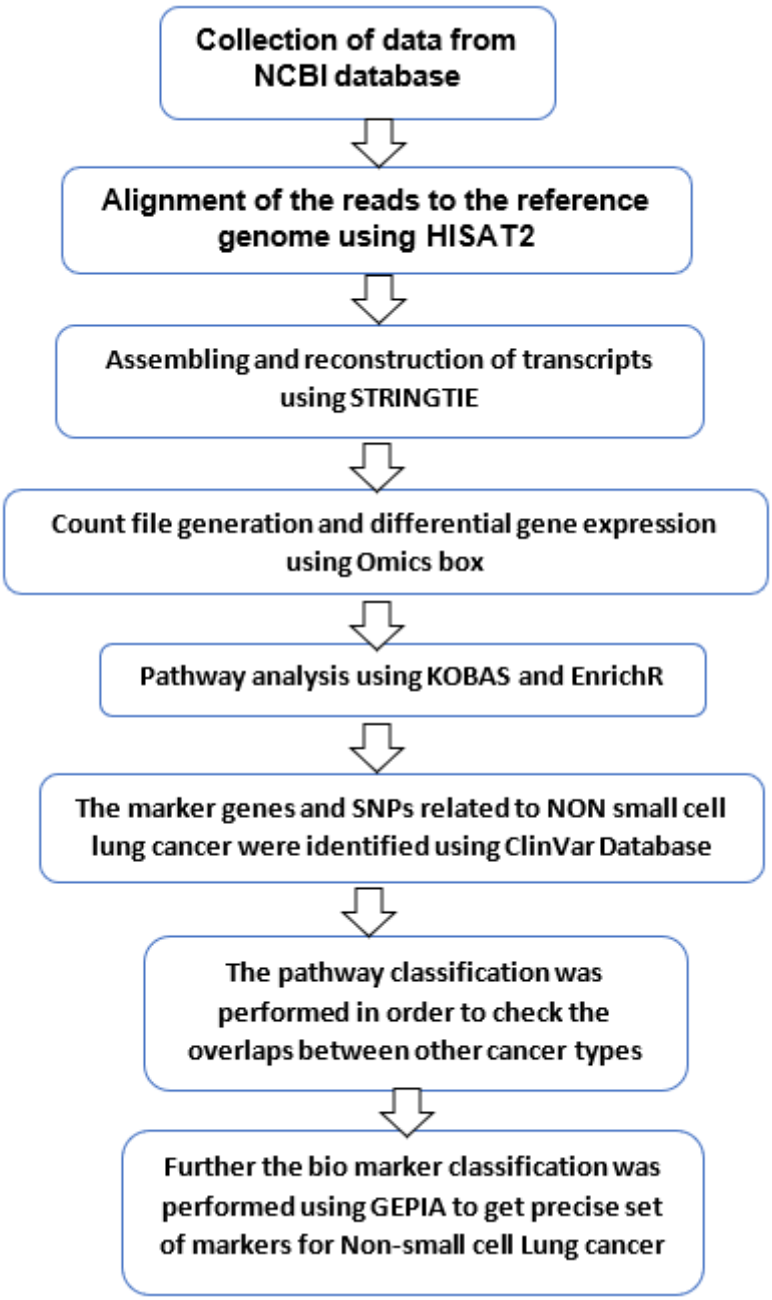

Figure 1

Workflow for Transcriptome analysis

Figure 2

Bar graph indicating the total number of genes that are;

a) Total number of genes b) Number of genes that are considered after filtration and normalization c) Number of genes differentially expressed d) Number of genes upregulated and downregulated

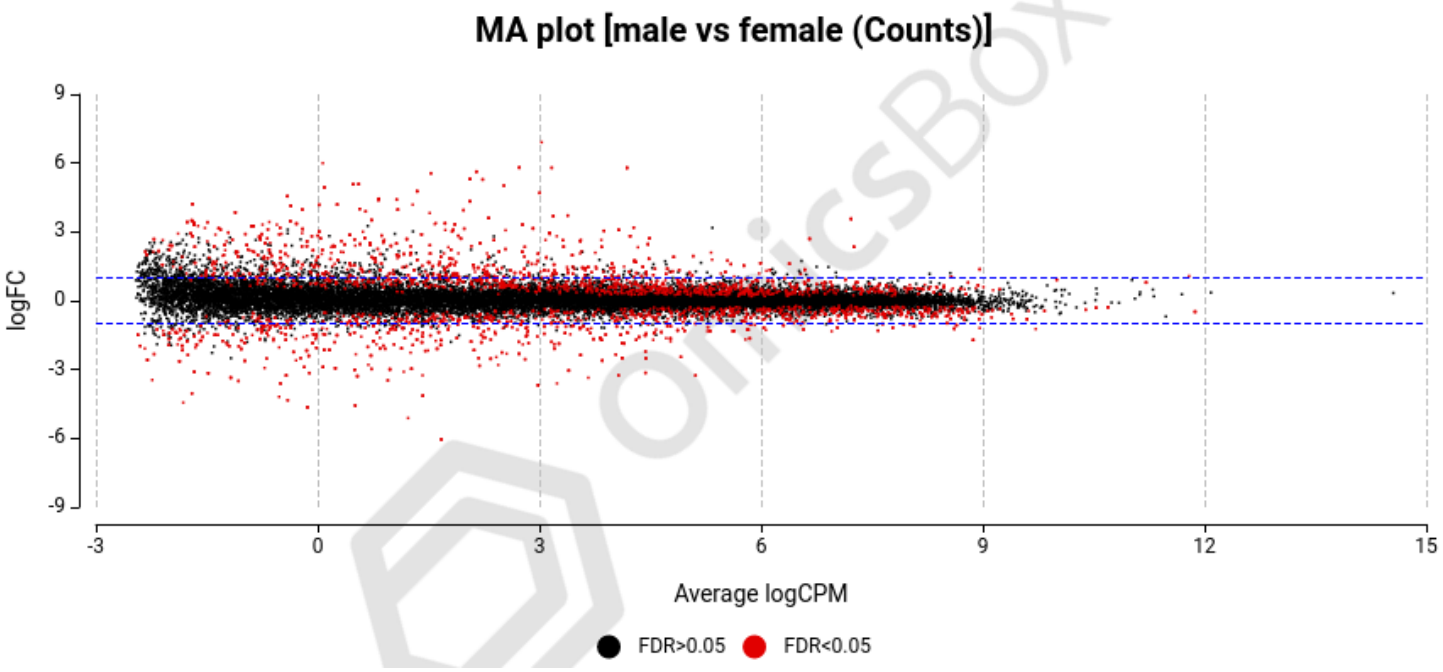

Figure 3

MA Plot

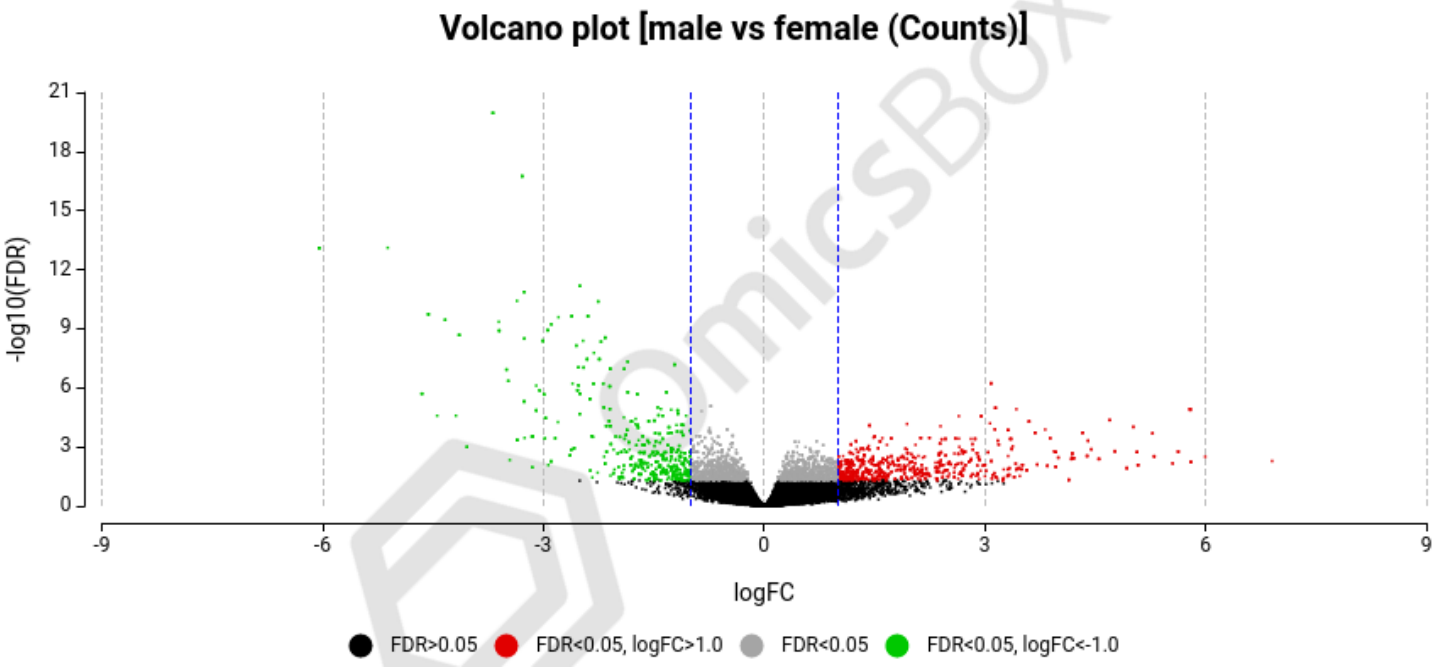

Figure 4

Volcano plot

Figure 5

Bar graphs showing the gene expression interaction with Adenoid Cystic Carcinoma, Breast cancer, cholangiocarcinoma, Diffuse b cell carcinoma, glioblastoma, renal cell carcinoma, brain tumors, ovarian cancer, paraganglioma, cervical cancer, skin cutaneous melanoma, tenosynovial giant cell tumor, thymus cancer, uterine cancer and lung cancer.

X- axis – Gene Y-axis - Expression