Detect differentially expressed genes from bulk RNA-seq data

Name: Jingkai LAN; **Student number:** 23-107-832

Abstract:

Breast cancer is the most common cancer type among women. Utilizing tools such as DESeq2 and ClusterProfilers, we conducted differential expression analysis on RNA-seq data of three common subtypes of breast cancer and healthy controls. The results indicate that, Triple Negative Breast Cancer (TNBC) exhibits the largest gene expression differences compared to Normal samples. By constructing volcano plots, we identified RN7SL1 as the gene with the most significant difference in expression between HER2-positive breast cancer (HER2) and Normal. RN7SL4P, on the other hand, showed the most pronounced differential expression in TNBC and Non-Triple Negative Breast Cancer (NonTNBC) compared to Normal. Enrichment analysis revealed that the top confidently associated cellular component for all three breast cancer subtypes is the cytosolic ribosome. Further differential expression enrichment analysis identified Inflammation and Infection as the top two significantly associated biological pathways in all three tumor types compared to controls. The genes and pathways discovered in this study have scientific significance for the exploration of potential targets and biomarkers.

Key words: Differential Expression; Enrichment Analysis; Breast Cancer;

1 Introduction

Breast cancer is the most common cancer in women and the second leading cause of cancer-related death among women (Siegel et al., 2017). In China, breast cancer accounts for 12% of newly diagnosed cancers each year, and the associated mortality rate is on the rise, representing 9.6% of global breast cancer deaths (Fan et al., 2014). Meanwhile, breast cancer is the most prevalent malignancy in American women, with related deaths constituting approximately 15% of all cancer cases (Giuliano et al., 2018). Currently, bioinformatics tools for differential expression analysis play a crucial role in identifying potential therapeutic targets (Deng et al., 2019). Breast cancer can be broadly classified into three subtypes: HER2-positive breast cancer (HER2), Triple Negative Breast Cancer (TNBC), and Non-Triple Negative Breast Cancer (NonTNBC). This study aims to utilize bioinformatics tools for RNA-seq differential expression analysis of the three tumor types compared to healthy controls, aiming to obtain information on differentially expressed genes, cellular components where differentially expressed genes cluster, and associated biological processes.

2 Material and methods

Sample Acquisition and Quality Check:

The data used for differential expression analysis were obtained from the study by Eswaran et al. (GSE52194) (Eswaran et al., 2016). For each of the three tumor subtypes (HER2, TNBC, NonTNBC), the dataset's corresponding three replicates were used. Sequencing was performed in paired-end mode on Illumina HiSeq 2000. We utilized FastQC(v. 0.11.9) for sequencing quality checks and report generation. The reference genome (GCA_000001405.29) and annotation file (gtf) used in this study were sourced from the Ensembl website (https://asia.ensembl.org/Homo_sapiens/Info/Index?db=core).

Map Reads to the Reference Genome:

We used the Hisat2(v. 2.2.1) tool to index the reference genome and map reads to it. The Samtools(v. 1.10)

tool was employed for format conversion to obtain bam files, followed by sorting and indexing. FeatureCounts(Liao et al, 2013) was used to count the number of exons in each sample.

Data Analysis:

For inter-group differential expression analysis, the DESeq2 package based on the R language(v. 4.3.2) was employed (Love et al., 2014). To perform principal component analysis and generate a heatmap of expression differences, the VST function was used to transform the raw matrix, eliminating variance dependence on the mean. To identify genes that were significantly upregulated or downregulated compared to the Normal group, a volcano plot was generated using ggplot, with the threshold for upregulation set as padj < 0.5 and log2 FoldChange > 1 (downregulated genes: log2 FoldChange < 1). Pathway over-representation analysis was conducted using the ClusterProfiler package. The annotation for organisms utilized the human genome annotation org.Hs.eg.db recommended by ClusterProfiler (Carlson M, 2019). Differential expression analysis results between the three tumor types and normal samples were subjected to Gene Ontology (GO) over-representation analysis for cellular components (CC). Additionally, enrichment analysis was performed using DisGeNET, and the results were visualized using a dotplot.

3 Results

Quality Check and Mapping:

We conducted quality checks using FastQC, summarizing read counts and GC% for each sample (Table S1). Across all nine diseased samples, the average GC% was 54%, with minimal differences among different tumor types. The average GC% for the normal group was 50%. In the HER2 sample group, the range of reads was 52,010,599-68,888,018; TNBC1 ranged from 44,434,722-48,256,786; NonTNBC ranged from 51,565,654-64,355,558, and Normal ranged from 15,886,336-37,178,138. In all samples, the average base quality gradually decreased with the lengthening of reads, with mates2 exhibiting a larger decline compared to mates1. Adapter Content checks revealed the removal of adapters in all samples. Additionally, most samples showed slight base abnormalities in the first 10 bases of Per Base Sequence Content. If abnormalities were severe, trimming of the unstable first 10 bases was necessary. However, overall, the quality of all samples met the criteria for continued analysis, allowing for further data analysis.

We mapped the samples to the reference genome and obtained mapping summary data (Table S2). The overall alignment rates for all three diseased groups were lower than that of Normal. Among them, the TNBC group showed the greatest overall difference from the reference genome, with an average overall alignment rate of 87.48%, the lowest concordantly rate (75.28%), and the highest discordantly rate (1.05%).

Differential Expression Analysis:

For differential expression analysis, we used featureCounts to obtain exon counts for different genes across all samples. To assess gene expression clustering and overall differences, we conducted PCA analysis and generated a sample-to-sample distance heatmap (Figure 1). Except for one HER2 sample showing some differences, the clustering of the four groups was generally clear. Additionally, among all samples, the greatest difference was observed between Normal3 and TNBC1 samples.

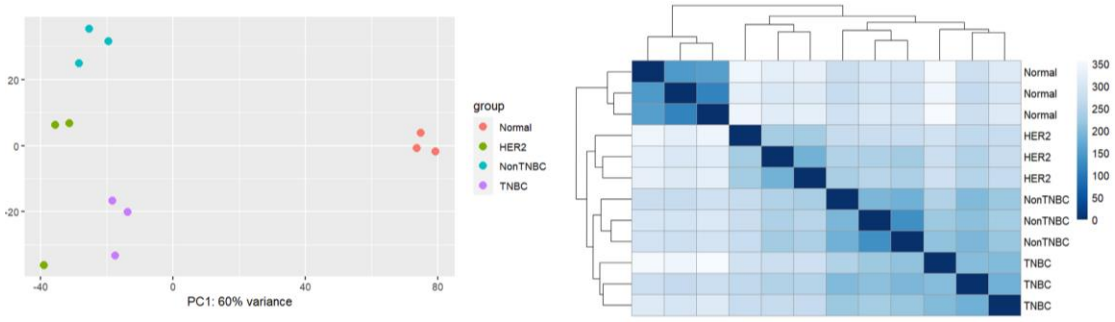

Figure 1. Principal Component Analysis plot and heatmap of expression distance between samples for the four groups.

We conducted statistical analysis of differential expression results between the three tumor types and the Normal group (Table S4). TNBC had the highest number of DE genes (16,060), followed by HER2 (14,694), and NonTNBC (11,876). In comparisons with Normal, the number of upregulated genes exceeded that of downregulated genes by more than twice in all three tumor types. For easier identification of differentially expressed genes associated with each tumor type, we created a volcano plot (Figure 2). In the HER2 group, RN7SL1 (ENSG00000276168) with a log2 FoldChange of 10.04 was identified, having a much higher standardized count mean (464,216) compared to other tumor types. RN7SL4P (ENSG00000263740) was found in both TNBC and NonTNBC, with the smallest padj values in both tumor types and log2 FoldChange values of 10.59 and 11.31, and standardized count means of 9,855 and 16,251, respectively.

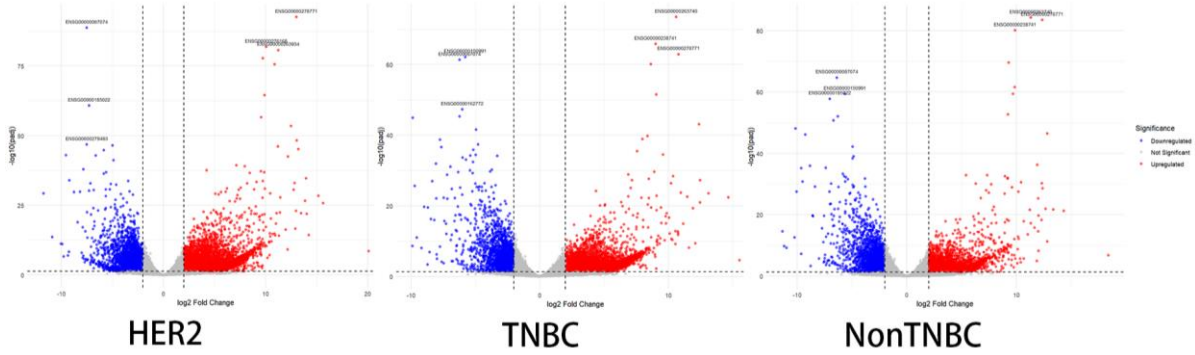

Figure 2. Volcano plot of differential expression analysis for the three tumor types compared to the Normal group. Blue represents downregulated genes, red represents upregulated genes. The top three genes with the smallest padj values are annotated. The log2 FoldChange threshold for color display was set to 2 for ease of visualization.

Enrichment Analysis:

By conducting Gene Ontology enrichment analysis on the differential expression results between the three tumor types and the normal group, we obtained the top 5 GO terms for each sample (Table 1). In all three tumor types, the top-ranked cellular component GO term based on confidence was cytosolic ribosome (GO:0022626), with GeneRatio ranging from 1.19% to 1.32%, and counts ranging from 80 to 85. Additionally, HER2 and TNBC shared the same top 3 GO terms. NonTNBC's enrichment in cellular component differed, focusing on cytosolic large ribosomal subunit and external side of plasma membrane, with the latter having the highest count among NonTNBC's top three GO terms.

Table 1. Top three GO terms for each tumor sample in Enrichment Analysis of Gene Ontology.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
HER2 VS Normal	GO:0022626	cytosolic ribosome	85/7113	116/19650	4.11E-16	3.10E-13	2.71E-13	RPL11/RPS8/RPS27/R	85
	GO:0005925	focal adhesion	232/7113	422/19650	1.62E-15	6.10E-13	5.34E-13	SLC9A1/EPHA2/DOCI	232
	GO:0030055	cell-substrate junction	234/7113	432/19650	1.13E-14	2.83E-12	2.48E-12	SLC9A1/EPHA2/DOCI	234
TNBC VS Normal	GO:0022626	cytosolic ribosome	80/6019	116/19650	1.87E-17	1.40E-14	1.23E-14	RPS8/RPS27/RPL5/RP	80
	GO:0005925	focal adhesion	208/6019	422/19650	4.61E-16	1.73E-13	1.52E-13	SLC9A1/EPHA2/BCAR	208
	GO:0030055	cell-substrate junction	211/6019	432/19650	9.80E-16	2.45E-13	2.15E-13	SLC9A1/EPHA2/BCAR	211
NonTNBC VS Normal	GO:0022626	cytosolic ribosome	85/7054	116/19650	2.33E-16	1.76E-13	1.52E-13	RPL11/RPS8/RPS27/R	85
	GO:0022625	cytosolic large ribosomal subunit	48/7054	58/19650	2.80E-13	1.06E-10	9.11E-11	RPL11/RPL5/RPL22/R	48
	GO:0009897	external side of plasma membrane	197/7054	388/19650	1.04E-09	2.61E-07	2.25E-07	S1PR1/F3/VCAM1/CD	197

We also conducted differential expression enrichment analysis (Figure 3). In the comparisons between the three tumor types and Normal, the most pronounced differences in expression were observed in the top two associated biological processes, Inflammation and Infection. In TNBC, the GeneRatios for these two biological processes were lower compared to the other two tumor types. In HER2, besides Inflammation and Infection, the highest-confidence differentially expressed biological pathways included B-cell Malignancy, LOW-GRAD, and Primary SjC6gren's syndrome. For TNBC and NonTNBC, the majority of differentially expressed biological pathways compared to Normal were similar. The main difference was that NonTNBC had more genes expressing differences associated with Neuroendocrine Tumors, while TNBC's differentially expressed genes were more related to Juvenile Arthritis.

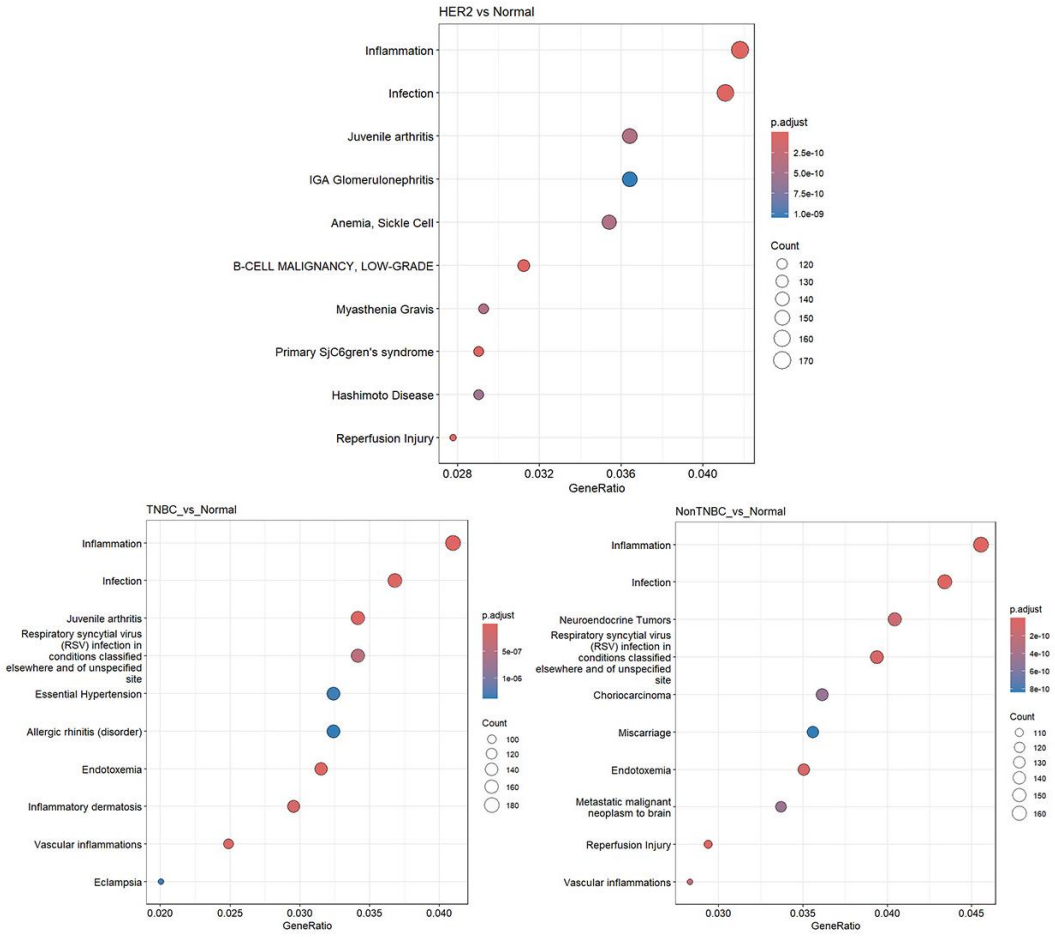

Figure 3. Enrichment Analysis of Differential Expression plot.

4 Discussion

During the inspection of sample sequence quality, an anomaly was observed in Per Base Sequence Content in the first 10 base pairs, despite the absence of adapters according to the report. This discrepancy may be attributed to

the sequencing instrument's unstable state during the initial sequencing stages.

For mapping, multimapped reads were observed in all samples. The presence of multimapped reads may lead to inaccurate counting in certain genes or regions, potentially impacting the reliability of gene expression analysis.

In PCA analysis, it was noted that one sample from the HER2 group clustered near the TNBC group. This suggests minimal genomic expression differences between this sample and TNBC, significantly affecting downstream analysis. As a result, the results of comparisons between HER2 and the control group, as well as TNBC and the control group, are likely to be closer, diminishing the confidence of the analysis.

In enrichment analysis, through Enrichment Analysis of Differential Expression, high levels of Inflammation and Infection were identified in all three tumor types, possibly due to the immune response against the tumor. Notably, in TNBC, the degree of immune response appeared slightly weaker compared to the other two types, hinting that TNBC patients might be more prone to recurrence or challenging to treat.

In summary, this study conducted differential expression analysis and enrichment analysis on data from public databases, revealing a series of crucial genes and enriched cellular components/biological processes associated with breast cancer.

Supplementary materials: Lanrem-LJK/rnaseq_course (github.com)

Reference:

Carlson, M. (2019). org.Hs.eg.db. [online] Bioconductor. Available at: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

Deng, J.-L., Xu, Y. and Wang, G. (2019). Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. Frontiers in Genetics, 10. doi:https://doi.org/10.3389/fgene.2019.00695.

Eswaran, J., Cyanam, D., Mudvari, P., Reddy, S.D.N., Pakala, S.B., Nair, S.S., Florea, L., Fuqua, S.A.W., Godbole, S. and Kumar, R. (2012). Transcriptomic landscape of breast cancers through mRNA sequencing. Scientific Reports, [online] 2(1), p.264. doi:https://doi.org/10.1038/srep00264.

Fan, L., Strasser-Weippl, K., Li, J.-J., St Louis, J., Finkelstein, D.M., Yu, K.-D., Chen, W.-Q., Shao, Z.-M. and Goss, P.E. (2014). Breast cancer in China. The Lancet Oncology, [online] 15(7), pp.e279–e289. doi:https://doi.org/10.1016/s1470-2045(13)70567-9.

Liao, Y., Smyth, G.K. and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), pp.923–930. doi:https://doi.org/10.1093/bioinformatics/btt656.

Giuliano, C.J., Lin, A., Smith, J.C., Palladino, A.C. and Sheltzer, J.M. (2018). MELK expression correlates with tumor mitotic activity but is not required for cancer growth. eLife, 7. doi:https://doi.org/10.7554/elife.32838.

Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). doi:https://doi.org/10.1186/s13059-014-0550-8.

Siegel, R.L., Miller, K.D. and Jemal, A. (2017). Cancer statistics, 2017. CA: A Cancer Journal for Clinicians, 67(1), pp.7–30. doi:https://doi.org/10.3322/caac.21387.