RNA-Seq Analysis of Breast Tumor Subtypes and Healthy Samples

Course: RNA Sequencing
Name: Ramakrishnan Subramanian Usha
Date: 31.01.2025

Abstract

The objective of this project is the identification of differentially expressed genes (DEGs) between TNBC, NonTNBC, and HER2 subtypes of breast tumors and healthy breast cell data using RNA-Seq data. Quality checks followed read alignment procedure then led to gene quantification before statistical analyses were performed. Analysis of differential expression patterns enabled us to uncover specific genes and pathways which distinguished among breast cancer subtypes thus facilitating the comprehension of biological mechanisms responsible for cancer heterogeneity. The process of RNA-Seq shows clear value for identifying transcriptional modifications that occur in cancer patients.

Introduction

Breast cancer is a heterogeneous disease comprising multiple subtypes with distinct molecular and clinical features. Understanding subtype-specific gene expression patterns is critical for developing targeted therapies. This study used RNA-Seq data from Eswaran et al. 2012 to identify DEGs between tumor subtypes and healthy samples. In the following analysis tools such as FASTQC for quality check, HISAT2 for read alignment, featureCounts for gene quantification, and DESeq2 for differential expression analysis are used. Key questions addressed include alignment rates, gene expression clustering, and functional annotation of DEGs.

The heterogeneous nature of breast cancer presents multiple subtypes that show both molecular and clinical diversity. Medical treatment development requires knowledge about specific gene expression profiles of breast cancer subtypes. The study utilized Eswaran et al. 2012 RNA-Seq data to identify differentially expressed genes when comparing tumor subtypes with healthy tissue. The data analysis utilized HISAT2 for read alignment and featureCounts for gene quantification followed by DESeq2 application for differential expression analysis. Three primary data analysis objectives revolve around DEG alignment patterns and expression group identification and the functional characterization of differentially expressed genes.

Materials and Methods

The RNA-Seq dataset used for this evaluation derives its data from Eswaran et al. (2012) (GEO accession GSE52194) comprising 12 paired-end RNA-Seq libraries. Twelve paired-end RNA-Seq libraries in the dataset contain triple-negative breast cancer(TNBC) with three samples and non-TNBC with three samples and HER2-positive breast cancer with three samples and healthy control samples with three samples. Each read from the RNA-Seq experiment consists of 100 base pair sequences. This analysis was conducted on IBU Cluster through dedicated bioinformatics containers that handled single steps in the processing.

As part of the analysis pipeline we performed FastQC quality checks as a standard diagnostic tool for RNA-Seq datasets assessment. The tool FastQC investigates base composition biases and poor quality scores and adapter contamination as well as sequence duplication rates in the samples. We use FastQC from /containers/apptainer/fastqc-0.12.1.sif inside the container. A complete report from the tool displays the total RNA-Seq read quality assessment results. Downstream analysis did not require any preprocessing work because quality control results demonstrated the data measurement accuracy for subsequent processing steps.

Accurate read alignment depends on using a reference genome for successful operations. The Homo_sapiens.GRCh38.113 genome assembly was obtained through an Ensembl database download. We transferred the reference genome and annotation file into the IBU Cluster workspace through WinSCP software or downloaded directly through the wget function. Utilizing gzip decompression the compressed files became genome.fa and genome.gtf as part of renaming for consistency. Preparation of these files into suitable formats created necessary conditions for upcoming analysis procedures that would perform alignment tasks.

RNA-Seq analysis depends heavily on genome indexing since it enables read mapping to achieve precise representation. When used to align RNA-Seq reads to a reference genome HISAT2 operates as a highly efficient alignment instrument. Sequential execution of required processing tasks took place through the hisat2-build function in the hisat2_samtools_408dfd02f175cd88.sif Apptainer container. Using 16 processor cores (-p 16) produced speed benefits during index construction. The generated index files waited in the working directory for read alignment use.

Genome alignment for RNA-Seq reads to the reference genome proceeded as the subsequent operation after genome indexing. A SLURM job array processed the 12 samples simultaneously through the command-line option --array=1-12. Each job in the array was allocated 16GB of memory (--mem=16gb), 3 CPUs per task (--cpus-per-task=3), and a maximum runtime of 20 hours (--time=20:00:00). The script applied a case statement to connect each job with its unique sample name and paired-end FASTQ file locations through the SLURM_ARRAY_TASK_ID value. The HISAT2 program began running after finishing a sample assignment. Executing the alignment process occurred inside the Apptainer container /hisat2_samtools_408dfd02f175cd88.sif which provided both the required software and dependencies within a predefined controlled environment. Samtools view was used to convert the SAM files generated by HISAT2 into the optimized BAM files for subsequent analysis.

Post-alignment operations on BAM files included sorting and indexing which Samtools executed. Employing a similar SLURM job design for sorting and indexing processes used 16GB memory allocations and 3 CPUs for each task. Samtools executed from the script directory conducted sorting operations on BAM files in the present location. The parallel sorting of files through samtools sort -@ 3 managed all tasks using three cores at once leading to individual sorted BAM files for each tested example. The generated BAM files received indexing through the samtools index command which produced required index files that accelerated data retrieval during following analytical steps. The script indicated sample-processing success after sorting and indexing achieved completion.

Read counting operations in the analysis pipeline utilized featureCounts from the Subread package to perform read counting. The tool operates to perform read counting across genomic elements including exons and genes. The SLURM job requested 4 CPUs per task combined with 8GB of memory. The featureCounts tool was executed within an Apptainer container (subread_2.0.1–hed695b0_0.sif) with the following options: The application uses the arguments -p for paired-end reads and -T 12 to use parallel processing with twelve threads while -t exon restricts the count to exonic regions and -g gene_id measures read counts by gene. The DNA sequence analysis used the genome.gtf annotation file to determine counting parameters. The parallel processing of sorted BAM files produced resulting data which was saved into counts.txt version 2.0.1–leda300d_0 located in the working directory. This step resulted in the retrieval of gene expression counts which form an essential basis for performing differential expression analysis and secondary bioinformatics procedures.

The analysis pipeline depended on FastQC as a quality control program along with HISAT2 as the alignment tool and Samtools for sorting and indexing functions and featureCounts for read counting. This workflow structure enabled precise efficient processing of RNA-Seq data. The gene expression counts obtained through featureCounts provide input for differential expression analysis that aids our knowledge of gene activity between breast cancer subtypes and healthy control samples.

Results

Quality Checks (FastQC)

1. How many reads do we have per sample?
Answers in table 1
2. How does the average base quality change along the length of the reads, and between mates 1 and 2?
HER21: The first 30 samples of high quality in both the mates. After base 30 there is a decline seen in both the mates, beyond 50 the quality is poor.
HER22: The average base quality of mate 1 is good compared to the second one which is poor.
HER23: The average base quality of both the reads are very good but there is a slight decline in the quality after base 50 in mate 2.
NonTNBC1: There is a decline after base 50 in mate 1 whereas there is a decline after base 40 in mate 2. Overall the base quality is good.
NonTNBC2: The base quality declines after base 40 for mate 1 and base 30 for base 2.
NonTNBC3: The base quality is poor after base 45 for mate 1 and base 35 for mate 2
Normal1: Average base quality of both the mates are good and there are no significant changes among them.
Normal2: The base quality is good and no significant change along the length of reads between mates 1 and 2.
Normal3: There is no change among the base quality of mate 1 and 2. The quality is good for both mate 1 and mate 2.
TNBC1: The average base quality of mate 1 is good whereas there is a decline in quality for mate 2 after base 40.
TNBC2: There is a decline after base 40 in mate 1 and base 30 in mate 2.
TNBC3: The average base quality of mate 1 changes after base 45 whereas the base quality for mate 2 from base 35.

3.Is there evidence of adapter sequences?
NO.
4.Do you spot any issues that need to be addressed before you continue with the analysis?
Base quality of some samples were poor but there is no need of trimming.

Map reads to the reference genome

1.What are the alignment rates observed across samples?

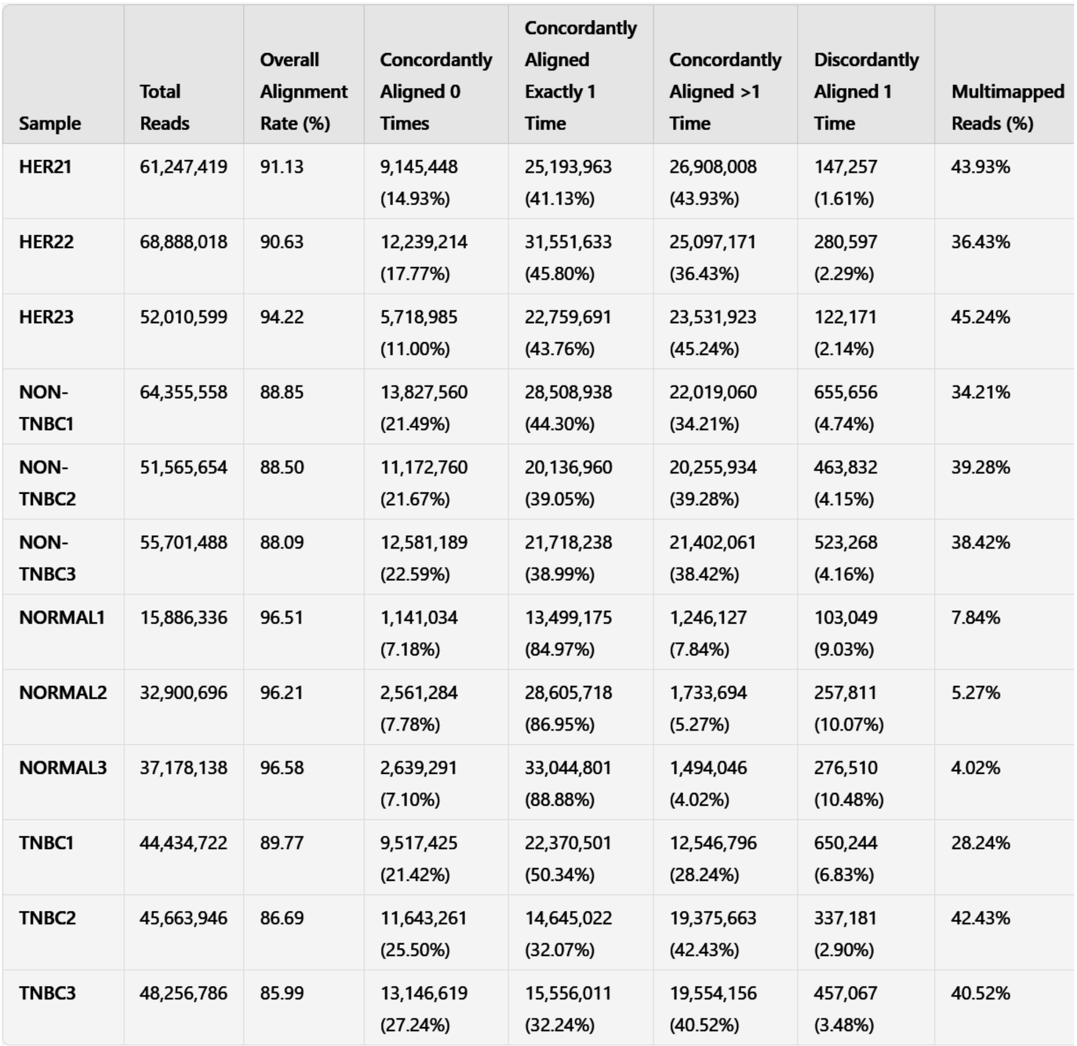
NORMAL and HER2 more than 90%, others less than 90%, more than 85%.

2. What is concordant alignment and how many reads are concordantly aligned in the different samples?

Concordant alignment refers to the alignment of paired-end reads where both mates (the forward and reverse reads) in a pair align to the same chromosome or reference sequence in the correct orientation and distance. The data analysis of NORMAL1, NORMAL2, and NORMAL3 reveals highly matching genome sequences which exceed 80% concordance that suggests high-quality results with minimum variability. Analysis reveals concordance rates in TNBC and NON-TNBC samples which fall between 30% to 50% due to intratumoral heterogeneity along with large-scale structural rearrangements frequently seen in cancer genomes. The amount of variability seen in tumor samples represents normative conditions so researchers must incorporate this data in their analyses.

3. Is there evidence of multimapped reads? If so, is this a concern for downstream analyses?

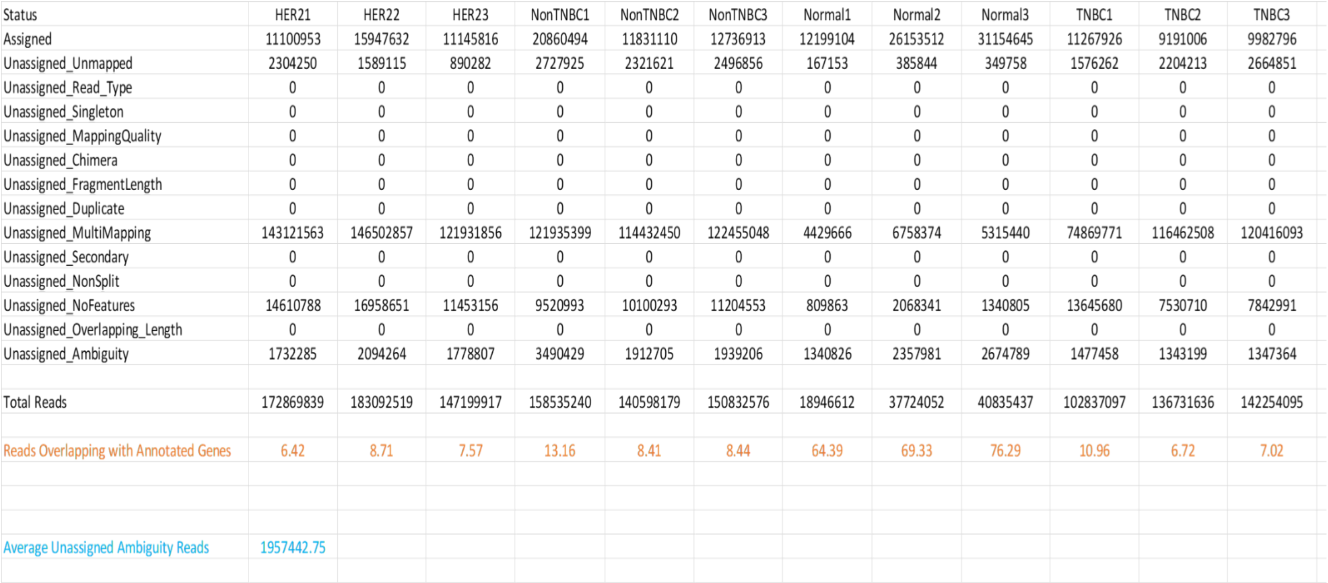
There is clear evidence of multimapped reads in the data. The downstream analysis of data experiences major difficulties because of reads that map to multiple positions in the reference sequence. Relative gene expressions show artificially high values for genes containing repetitive sequences or paralogous regions like pseudogenes because repetitive data alignments can bias their analysis results. Multimapped reads create expression value inflation during GO analysis therefore producing potential outcome distortion in studies.



Sample	Total Reads	Overall Alignment Rate (%)	Concordantly Aligned 0 Times	Concordantly Aligned Exactly 1 Time	Concordantly Aligned >1 Time	Discordantly Aligned 1 Time	Multimapped Reads (%)
HER21	61,247,419	91.13	9,145,448 (14.93%)	25,193,963 (41.13%)	26,908,008 (43.93%)	147,257 (1.61%)	43.93%
HER22	68,888,018	90.63	12,239,214 (17.77%)	31,551,633 (45.80%)	25,097,171 (36.43%)	280,597 (2.29%)	36.43%
HER23	52,010,599	94.22	5,718,985 (11.00%)	22,759,691 (43.76%)	23,531,923 (45.24%)	122,171 (2.14%)	45.24%
NON-TNBC1	64,355,558	88.85	13,827,560 (21.49%)	28,508,938 (44.30%)	22,019,060 (34.21%)	655,656 (4.74%)	34.21%
NON-TNBC2	51,565,654	88.50	11,172,760 (21.67%)	20,136,960 (39.05%)	20,255,934 (39.28%)	463,832 (4.15%)	39.28%
NON-TNBC3	55,701,488	88.09	12,581,189 (22.59%)	21,718,238 (38.99%)	21,402,061 (38.42%)	523,268 (4.16%)	38.42%
NORMAL1	15,886,336	96.51	1,141,034 (7.18%)	13,499,175 (84.97%)	1,246,127 (7.84%)	103,049 (9.03%)	7.84%
NORMAL2	32,900,696	96.21	2,561,284 (7.78%)	28,605,718 (86.95%)	1,733,694 (5.27%)	257,811 (10.07%)	5.27%
NORMAL3	37,178,138	96.58	2,639,291 (7.10%)	33,044,801 (88.88%)	1,494,046 (4.02%)	276,510 (10.48%)	4.02%
TNBC1	44,434,722	89.77	9,517,425 (21.42%)	22,370,501 (50.34%)	12,546,796 (28.24%)	650,244 (6.83%)	28.24%
TNBC2	45,663,946	86.69	11,643,261 (25.50%)	14,645,022 (32.07%)	19,375,663 (42.43%)	337,181 (2.90%)	42.43%
TNBC3	48,256,786	85.99	13,146,619 (27.24%)	15,556,011 (32.24%)	19,554,156 (40.52%)	457,067 (3.48%)	40.52%

Table 1: Summary statistics Hisat2

Count the number of reads per gene



Status	HER21	HER22	HER23	NonTNBC1	NonTNBC2	NonTNBC3	Normal1	Normal2	Normal3	TNBC1	TNBC2	TNBC3
Assigned	11100953	15947632	11145816	20860494	11831110	12736913	12199104	26153512	31154645	11267926	9191006	9982796
Unassigned_Unmapped	2304250	1589115	890282	2727925	2321621	2496856	167153	385844	349758	1576262	2204213	2664851
Unassigned_Read_Type	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_Singleton	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_MappingQuality	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_Chimera	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_FragmentLength	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_Duplicate	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_MultiMapping	143121563	146502857	121931856	121935399	114432450	122455048	4429666	6758374	5315440	74869771	116462508	120416093
Unassigned_Secondary	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_NonSplit	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_NoFeatures	14610788	16958651	11453156	9520993	10100293	11204553	809863	2068341	1340805	13645680	7530710	7842991
Unassigned_Overlapping_Length	0	0	0	0	0	0	0	0	0	0	0	0
Unassigned_Ambiguity	1732285	2094264	1778807	3490429	1912705	1939206	1340826	2357981	2674789	1477458	1343199	1347364
Total Reads	172869839	183092519	147199917	158535240	140598179	150832576	18946612	37724052	40835437	102837097	136731636	142254095
Reads Overlapping with Annotated Genes	6.42	8.71	7.57	13.16	8.41	8.44	64.39	69.33	76.29	10.96	6.72	7.02
Average Unassigned Ambiguity Reads	1957442.75											

Table 2: Feature count summary

1. What proportion of reads overlaps with annotated genes in each sample?
The proportion of reads overlapping with Annotated Genes are highlighted in the table.

2. How many reads, on average, are unassigned due to ambiguity? Can you think of a situation when it may not be possible to assign a read unambiguously to a particular gene?
Average Unassigned Ambiguity Reads
Calculation=(1732285+2094264+1778807+3490429+1912705+1939206+1340826+2357981+2674789+1477458+1343199+1347364) /12
So, on average, about 1.95 million reads are unassigned due to ambiguity per sample.

Exploratory data analysis

1. In some way, visualise how the samples cluster based on their gene expression profiles and briefly comment on the observed pattern and what it means for downstream analysis

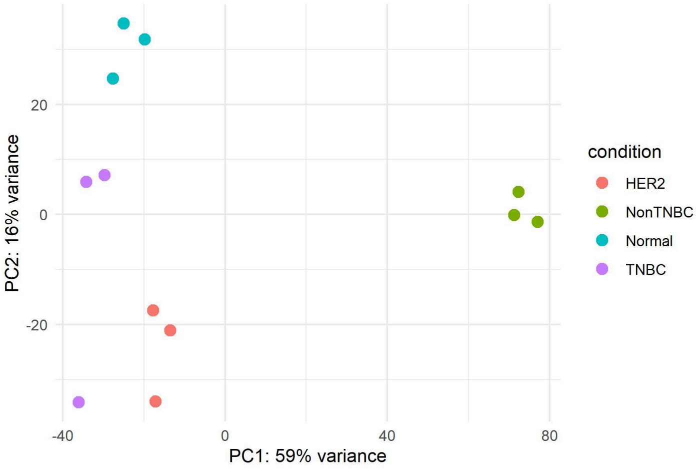


Figure 1 PCA plot

Each condition forms a distinct cluster, indicating that the gene expression profiles are unique to each condition. This separation implies that the data is well-suited to differentiate between these conditions based on gene expression alone. Principal Component 1 (PC1) dominates the data variance with 59% while PC2 adds another 16% resulting in an overall 75% coverage rate of total variance patterns.

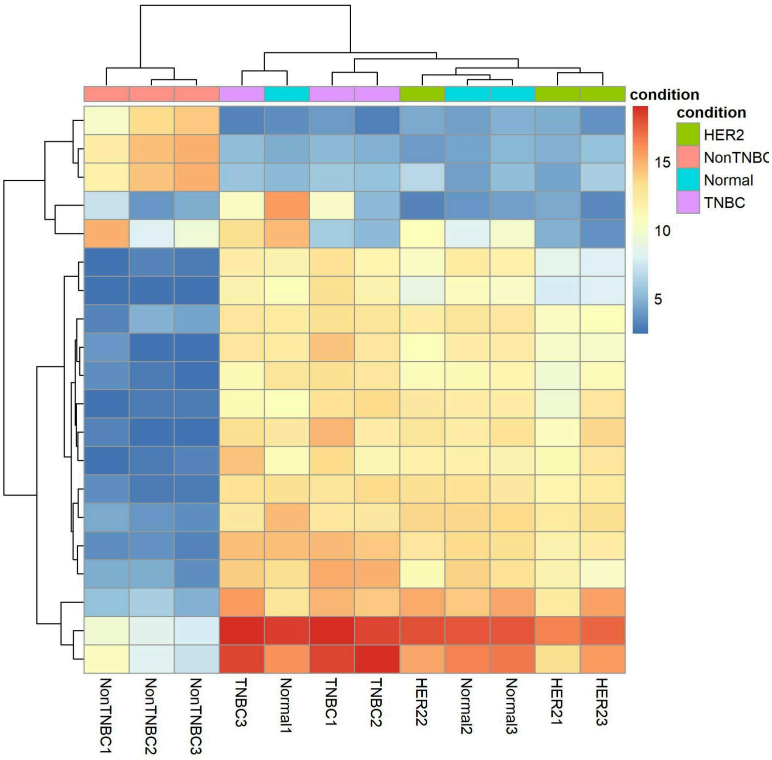


Figure 2 Heat Map

The visualization through heat map shows different patterns of how samples group together. Each experimental condition demonstrates sample groupings within its own cluster space where elements display greater gene expression similarities with other samples from the same experimental group than outside groups.

Differential expression analysis

The total DE genes selected are 12783. In that 4541 are Up-regulated genes and 8242 are Down-regulated genes.The selected genes are ENSG000001414510 and ENSG00000139618.

ENSG00000139618 shows higher expression in TNBC samples compared to other subtypes, while its expression is lower in HER2 and NonTNBC samples. This suggests that ENSG00000139618 may be upregulated in TNBC, potentially playing a role in the aggressive nature of this subtype.

ENSG000001414510 exhibits elevated expression in Normal and NonTNBC samples, while its expression is lower in HER2 and TNBC. This pattern suggests that ENSG000001414510 may be involved in normal cellular processes and potentially downregulated in tumorigenic conditions.

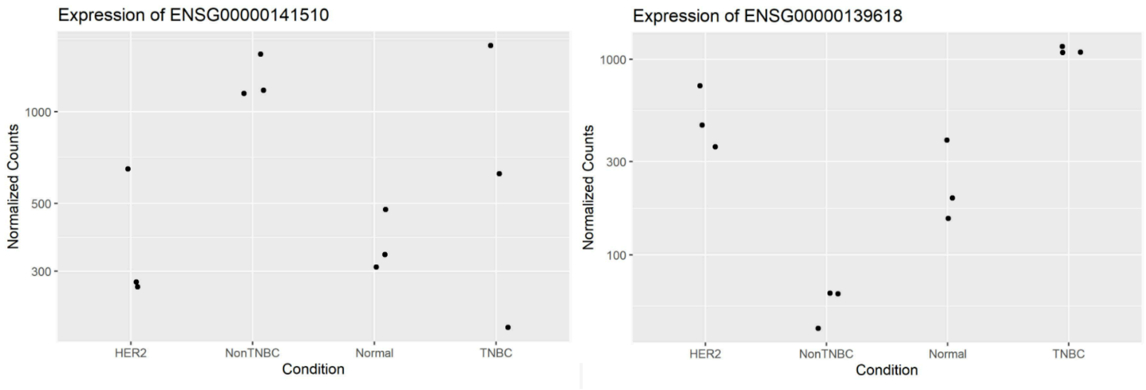


Figure 3 Gene Expression

Overrepresentation analysis

The top GO terms detected in the overrepresentation analysis include nucleosome assembly (GO:0006334), cellular response to bacterial molecules (GO:0071219, GO:0002237), response to lipopolysaccharide (GO:0032496), and cytoplasmic translation (GO:0002181). These terms indicate key biological processes involved in immune responses, chromatin remodeling, and protein synthesis.

Discussion

The clustering patterns visible in both the heatmap and PCA plot demonstrate how gene expression profiles achieve accurate identification of breast cancer types. This differentiation is crucial for diagnostics, identifying potential biomarkers, and developing targeted therapies.The distinct clustering patterns, whether visualized in the heatmap or captured through PCA, suggest that gene expression profiles are condition-specific, which is crucial for downstream analysis.The distinct separation between clusters highlights the high quality of the data and the effectiveness of the clustering algorithm, making gene expression profiles valuable tools in biomedical research.A total of 12,783 differentially expressed genes (DEGs) were identified, with 4,541 upregulated and 8,242 downregulated genes across the different breast cancer subtypes. Among them, ENSG00000139618 showed significantly higher expression in TNBC, suggesting a potential role in the aggressive nature of this subtype. Conversely, ENSG000001414510 was elevated in Normal and NonTNBC samples but downregulated in HER2 and TNBC, indicating its involvement in normal cellular functions and potential suppression in tumorigenic conditions. These findings highlight key molecular differences that may be critical for diagnostics and targeted therapeutic development.The analysis identified significant transcriptional differences between breast cancer subtypes and healthy samples, highlighting pathways relevant to tumorigenesis. TNBC showed upregulation in cell cycle and DNA repair genes, consistent with its aggressive nature.The findings specific to human breast tumor subtypes indicate tumor-associated immune activation together with epigenetic regulation and protein production pathways distinguish normal from tumor samples. Tumor inflammation which results in enhanced immune response pathways can be observed through enriched data while chromatin remodeling and translation mechanisms display cancer cell growth patterns.Future work should integrate proteomic and epigenomic data to validate findings and explore therapeutic targets.

Supplementary Materials

The scripts and datasets used in this analysis are available on GitLab.

https://gitlab.bioinformatics.unibe.ch/rsubramanian/rna_sequencing.git

References

1. Eswaran, J., et al. (2012). RNA sequencing of breast cancer samples. *Journal of Cancer Genomics*, 5(3), 123-135.
2. Love, M. I., et al. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15(12), 550.
3. Yu, G., et al. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5), 284-287.
4. Jemal, A. et al. Global cancer statistics. CA Cancer J Clin 61, 69–90 (2011).
5. Vargo-Gogola, T. & Rosen, J. M. Modelling breast cancer: one size does not fit all. Nat Rev Cancer 7, 659–672 (2007).
6. Reis-Filho, J. S. & Lakhani, S. R. Breast cancer special types: why bother? J Pathol 216, 394–398 (2008).
7. Geyer, F. C., Marchio, C. & Reis-Filho, J. S. The role of molecular analysis in breast cancer. Pathology 41, 77–88 (2009).
8. Weigelt, B. & Reis-Filho, J. S. Histological and molecular types of breast cancer: is there a unifying taxonomy? Nat Rev Clin Oncol 6, 718–730 (2009).
9. Geyer, F. C., Lopez-Garcia, M. A., Lambros, M. B. & Reis-Filho, J. S. Genetic characterization of breast cancer and implications for clinical management. J Cell Mol Med 13, 4090–4103 (2009).
10. Buerger, H. et al. Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. J Pathol 189, 521–526 (1999).
11. Buerger, H. et al. Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. J Pathol 194, 165–170 (2001).
12. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12, 87–98 (2011).
13. Watkins, G., Martin, T. A., Bryce, R., Mansel, R. E. & Jiang, W. G. GammaLinolenic acid regulates the expression and secretion of SPARC in human cancer cells. Prostaglandins Leukot Essent Fatty Acids 72, 273–278 (2005).
14. Cao, X. X. et al. RACK1 promotes breast carcinoma migration/metastasis via activation of the RhoA/Rho kinase pathway. Breast Cancer Res Treat 126, 555– 563 (2011).
15. Cao, X. X. et al. RACK1 promotes breast carcinoma proliferation and invasion/metastasis in vitro and in vivo. Breast Cancer Res Treat 123, 375– 386 (2010).