Differently expressed genes in breast cancer subtypes

Anna Boss

24-208-928

467713 RNA Seq – Fall Semester 2024

Abstract

This study investigates the transcriptomic differences between three breast cancer subtypes: Triple-Negative Breast Cancer (TNBC), Non-TNBC, and HER2-positive, using RNA-seq data. The goal was to identify differentially expressed genes (DEGs) that could provide insights into the molecular mechanisms underlying each subtype and inform personalized treatment strategies. The data were analyzed using a bulk RNA-seq approach, including quality control, alignment, and differential gene expression revealing 9165 DEGs across the subtypes. Functional enrichment analysis highlighted key pathways, such as sensory perception in TNBC, and immune responses in HER2-positive breast cancer. The analysis highlighted significant heterogeneity within each subtype, reflecting the diverse tumor microenvironments, genomic instability, and epigenetic regulation that shape breast cancer's biological landscape. These differences suggest that even within established subtypes, there is considerable variability that may influence clinical behavior and response to treatment. The findings suggest that identified pathways and involved genes could serve as potential therapeutic targets for more tailored treatments. However, limitations such as the lack of unique molecular identifiers and missing metadata of the samples indicate the need for further studies. Future research should validate these DEGs in clinical settings and explore their therapeutic potential through functional studies and larger cohort analyses.

Introduction

Breast cancer is the most common malignancy in women, with 2.3 million new cases diagnosed in 2022 and approximately 670 000 deaths worldwide[1]. The incidence of breast cancer continues to rise, but advancements in early detection and treatment have led to improved survival rates. However, prognosis varies significantly depending on the molecular subtype of the disease[2].

Breast cancer can be classified based on various criteria, with immunohistochemical surrogate classification being one of the most widely used methods[3]. This classification relies on the expression of hormone receptors (estrogen receptor (ER) and progesterone receptor (PR)), human epidermal growth factor receptor 2 (HER2), and the proliferation marker Ki67. Based on these markers, breast cancer is categorized into five subtypes:

- Luminal A (ER^+ and/or PR^+, $HER2^-$, Ki67 < 30%)
- Luminal B $HER2^-$ (ER^+ and/or PR^+, $HER2^-$, Ki67 ≥ 30%)
- Luminal B $HER2^+$ (ER^+ and/or PR^+, $HER2^+$), referred to in this study as Non-triple negative breast cancer (Non-TNBC)
- HER2-enriched (non-luminal) (ER^-, PR^-, $HER2^+$), also called HER2-positive
- Basal-like (ER^-, PR^-, $HER2^-$), also known as triple negative breast cancer (TNBC)

Survival outcomes differ across these subtypes, with luminal A typically having the most favorable prognosis and basal-like breast cancer being associated with poor survival rates[2]. TNBC poses a unique challenge due to its aggressive nature and high molecular heterogeneity[2]. Lacking ER, PR, and HER2 expression, it is unresponsive to targeted therapies, leaving chemotherapy as the primary option. However, response rates vary, and resistance frequently leads to recurrence.

Transcriptomic analyses have further classified TNBC into six molecular subtypes with distinct biological features, complicating treatment strategies[2]. Its dynamic tumor microenvironment, genomic instability, and high metastatic potential further hinder therapeutic success.

Understanding the molecular distinctions between breast cancer subtypes is crucial for developing more effective, personalized treatment strategies. Transcriptomic analysis offers a powerful approach to investigate gene expression profiles, revealing subtype-specific molecular signatures that could serve as potential therapeutic targets. In this study, I analyzed a publicly available RNA-seq dataset to identify differentially expressed genes (DEGs) across three breast cancer subtypes, aiming to uncover molecular differences that could inform personalized treatment approaches.

Methods

Source of RNA-seq reads

This study used RNA-seq data from Eswaran et al.[4]. The corresponding FASTQ files for selected samples were obtained from the Gene Expression Omnibus under accession number GSE52194 and were provided by Dr. Alexander Nater via the IBU cluster. The number of reads for the provided samples matched the original study for all samples except "Normal 2", for which a subset of 15 886 336 reads (out of 32 854 596) was used. Additionally, the samples "TNBC2" and "TNBC3" in this analysis correspond to "TNBC5" and "TNBC6" in the original study, respectively.

The dataset included human breast cancer tissue RNA samples from three tumor types and normal tissue (control). Classification was performed using immunohistochemistry. Tumor types were classified as:

- Triple-negative breast cancer (TNBC): Negative ER, PR, and HER2.
- Non-TNBC: Positive for ER, PR, and HER2.
- HER2-positive: Negative for ER and PR but positive for HER2.

For library preparation, ribosomal RNA was depleted to enrich for transcriptomic RNA. Sequencing was performed on an Illumina HiSeq 2000 in paired-end sequencing mode.

Quality assessment of reads

The quality of raw sequencing reads was evaluated using FastQC[5] (v0.12.12). Reads were then processed with Fastp[6] (v0.23.03) to enhance the dataset quality. This included filtering out reads with low complexity (threshold < 30%), identifying and removing adapter sequences, and discarding reads where more than 40% of bases had a Phred quality score below 15. Additionally, bases with a Phred score below 20 were trimmed from the read tails, duplicates and reads containing more than one ambiguous base were removed, and base correction was applied to improve overall sequence quality. After these steps, only reads with a minimum length of 50 bp were retained to minimize mapping ambiguity and increase confidence in alignment. The effectiveness of these quality control measures was validated using MultiQC[7] (v1.19), which generated comparative metrics before and after filtering.

Alignment of reads

The cleaned reads were aligned to the indexed reference genome GRCh38 build 113[8] (downloaded on 12 November 2024) using HISAT2[9] (v2.2.1). The mapped reads were subsequently processed with SAMtools[10] (v1.20) for sorting, indexing, and conversion to a binary format (BAM). Unmapped reads and non-chromosomal contigs (e.g. KI* and GL*) were excluded to focus on the target genome. After alignment, mapped reads were assessed using SAMtools flagstat, and a combined report was generated with MultiQC[7] (v1.194).

Counting reads

Gene expression quantification was performed using featureCounts of the subread[11] package (v2.0.6). Preprocessed and aligned BAM files served as input, with gene features annotated using the indexed annotation gtf file associated with GRCh38 build 113[8] (downloaded on 12 November 2024). Expression counts were quantified at the exon level, with a minimum overlap of one base required for inclusion. Only paired-end reads that were properly paired and had a minimum mapping quality score of 10 were considered. This reduced the impact of ambiguous or incorrect alignments, while still retained a sufficient number of reads for robust statistical analysis.

Differential gene expression analysis

Differential gene expression analysis was performed in R[12] (v4.4.1) using DESeq2[13] package (v1.46.0). The featureCounts output table was preprocessed to retain only gene IDs and raw counts. A metadata table describing the sample groups (HER2-positive, Non-TNBC, Normal, and TNBC) was prepared for downstream analysis. A DESeqDataSet object was created using the DESeqDataSetFromMatrix function, with the four groups specified as the design. To improve statistical power and focus on genes with sufficient expression, only genes with a total count greater than 10 across all samples were retained.

Subsequent steps included data normalization, dispersion estimation, and fitting statistical models to identify DEGs. Principal Component Analysis (PCA) was conducted using the PCAplot function on the top 500 most variable genes to evaluate sample clustering and identify potential data abnormalities. Variance-stabilizing transformation (VST) was applied to normalize data to account for sequencing depth differences and stabilize variance across expression values. The PCA plot was visualized using ggplot2[14] (v3.5.1) and gridExtra[15] (v2.3).

To identify DEGs within breast cancer subtypes, pairwise comparisons were performed with DESeq2: TNBC vs. HER2-positive, Non-TNBC vs. HER2-positive, and TNBC vs. Non-TNBC. Results were extracted using the results function and classified into "Upregulated," "Downregulated," or "Not Significant" categories based on an adjusted p-value < 0.05 and a log2 fold change threshold. Genes with a positive log2 fold change and adjusted p-value < 0.05 were considered upregulated, while those with a negative log2 fold change were classified as downregulated. Adjusted p-values were calculated using the Benjamini-Hochberg procedure.

Visualization of the pairwise comparisons was done with volcano plots using the EnhancedVolcano[16] (v1.24.0), ggrepel[17] (v0.9.6), and gridExtra[15] (v2.3) packages, with a log2 fold change cutoff of 2 and p-value 0.05. Gene symbols were annotated using org.Hs.eg.db[18] package (v3.20.0).

To visualize the expression levels of significant DEGs, two heatmaps using the pheatmap[19] package (v1.0.12) were generated: one including all genes with an absolute log2 fold change > 2 and another displaying the top 50 DEGs according to absolute log2 fold change. Gene counts were normalized using the counts function from the DESeq2[13] package (v1.46.0) with the median of ratios method and subsequently scaled using the Z-score.

Gene ontology enrichment analysis

To identify enriched Gene Ontology (GO) terms in the pairwise comparisons, the clusterProfiler[20] package (v4.14.4) was used. Overrepresentation analysis (ORA) was applied to identify biologically relevant processes associated with cancer subtypes, focusing on the Biological Processes category. Genes with an absolute log2 fold change ≥ 2 were selected, and multiple testing correction was performed using the Benjamini-Hochberg method ($p < 0.05$). Gene annotation was conducted with the org.Hs.eg.db[18] package (v3.20.0), and all genes within the experiment served as the reference set.

Enrichment results were visualized using gene-concept network plots created with the DOSE[21] package (v4.0.0), highlighting the relationships between enriched GO terms and input genes.

All scripts for data preprocessing, alignment, differential expression analysis, and visualization are available at: https://github.com/ann4boss/rna_course.

Results

Quality control and preprocessing

Quality assessment of the raw reads showed substantial variability across samples, with read counts ranging from 15.9 million to 68.9 million (Table 1). FastQC analysis indicated declining Phred scores towards the end of reads, particularly in reverse reads. Sample HER22, for instance, had a Phred score of 24 at the end of the reverse read, indicating low quality. Duplication rates were high across all samples, ranging from 46% to 71%. However, the absence of unique molecular identifiers made it impossible to distinguish between biological duplicates and PCR artifacts.

Additionally, GC content abnormalities were observed in 12 read pairs, likely due to duplication or adapter contamination. Adapter sequences were detected at low levels (≤ 0.9%) in all samples, while elevated poly-G sequences were observed in the reverse reads of TNBC and Non-TNBC samples, with a maximum of 1.85%. Overrepresented sequences, including Illumina Paired End PCR Primer 2, were detected in the forward reads of TNBC and Non-TNBC samples.

Cleaning the reads with Fastp improved data quality, with ≥ 95% of bases achieving a Phred score ≥30. However, duplication rates remained elevated (31.6% to 52.0%), potentially impacting downstream analyses. Reads were removed primarily because they were shorter than the 50 bp threshold after trimming low-quality bases from the ends.

Read alignment and mapping statistics

Alignment rates varied significantly across sample (Table 1). Control samples exhibited the highest percentages of uniquely mapped reads (> 88%), while HER2-positive and Non-TNBC samples showed moderate rates (67% to 79% and 63% to 65%, respectively). TNBC samples displayed the lowest and most variable alignment rates (53% to 67%).

Unmapped reads, where neither mate aligned to the reference genome, accounted for less than 10% of reads across all samples. Control samples had the lowest unmapped read rate (average 1.8%), followed by HER2-positive samples (2.1% to 3.0%). Non-TNBC and TNBC samples exhibited higher unmapped read rates (7.5% to 8.2% and 6.4% to 9.6%, respectively). A small percentage of reads (≤ 2%) were mapped discordantly, indicating improper orientation or insert size.

Table 1: Overview of read counts and retention rates across processing steps

Sample	Raw reads	Reads retained after cleaning (%)	Uniquely mapped reads (%)	Reads assigned to exons (%)
Normal 1	15 886 336	13 854 402 (87.2)	12 273 072 (88.6)	10 850 007 (88.4)
Normal 2	32 900 696	28 454 435 (86.5)	25 872 306 (90.9)	23 007 195 (88.9)
Normal 3	37 178 138	32 678 292 (87.9)	30 103 910 (92.1)	27 698 214 (92.0)
HER21	61 247 419	23 721 620 (38.7)	16 441 614 (69.3)	7 743 728 (47.1)
HER22	68 888 018	26 336 306 (38.2)	20 820 413 (79.1)	10 002 203 (48.0)
HER23	52 010 599	25 122 599 (48.3)	16 711 352 (66.5)	8 721 450 (52.2)
NonTNBC1	64 355 558	29 013 043 (45.1)	18 725 762 (64.5)	14 836 758 (79.2)
NonTNBC2	51 565 654	22 857 342 (44.3)	14 387 038 (62.9)	8 660 851 (60.2)
NonTNBC3	55 701 488	23 469 184 (42.1)	14 765 344 (62.9)	8 957 886 (60.7)
TNBC1	44 434 722	26 968 226 (60.7)	18 122 515 (67.2)	8 901 508 (49.1)
TNBC2	45 663 946	19 176 605 (42.0)	10 232 982 (53.4)	7 212 993 (70.5)
TNBC3	48 256 786	20 162 570 (41.8)	10 648 313 (52.8)	7 401 233 (69.5)

The percentage refers to the number of reads retained from the previous step. Example for sample Normal 1: 87.2% of the raw reads passed the cleaning step. Of the cleaned reads, 88.6% were uniquely mapped and subsequently 88.4% assigned to an exon.

Multimapped reads, where both mates aligned to multiple locations, were more prevalent in breast cancer samples (15.4% to 26.3%) compared to control samples (≤ 7%). This elevated multimapping rate may reflect genomic instability or repetitive regions in cancer genomes, which are less prevalent in normal tissues. These multimapped reads limit the number of assigned features in downstream analysis.

Feature count and read assignment

Reads assigned to exons ranged from 7.2 to 27.7 million (Table 1). A large fraction of reads remained unassigned due to multimapping, especially in breast cancer samples (37.7% to 59.3%) compared to controls (10.0% to 18.5%).

Additionally, a portion of reads could not be assigned to any exon. This was more pronounced in breast cancer samples, ranging from 11.9% to 30.2%, compared to normal samples (3.6% to 6.2%). A smaller fraction of reads (2.7% to 7.4%) was excluded due to ambiguity, where alignments overlapped two or more exons. This may result from repetitive regions or the shorter read lengths (median of 60 bp for breast cancer, 75 bp for controls).

Differential gene expression analysis

Of 78 932 detected genes, 26 562 were excluded due to low expression, leaving 52 370 genes for pairwise analysis. PCA showed that PC1 (57%) separated healthy from cancerous tissues (Figure 1A).

Most samples clustered within subtypes, except HER22, which aligned closer to TNBC, even after removing normal samples (Figure 1B). Despite this outlier behavior, no issues were identified after sample cleaning to suggest that HER22 was compromised. The sample was retained in subsequent analyses to maintain statistical power and ensure sufficient replicates for robust conclusions. This anomaly could also indicate that sample HER22 exhibits transcriptomic characteristics more aligned with TNBC, potentially due to underlying genomic alterations such as activation of basal-like gene expression programs.

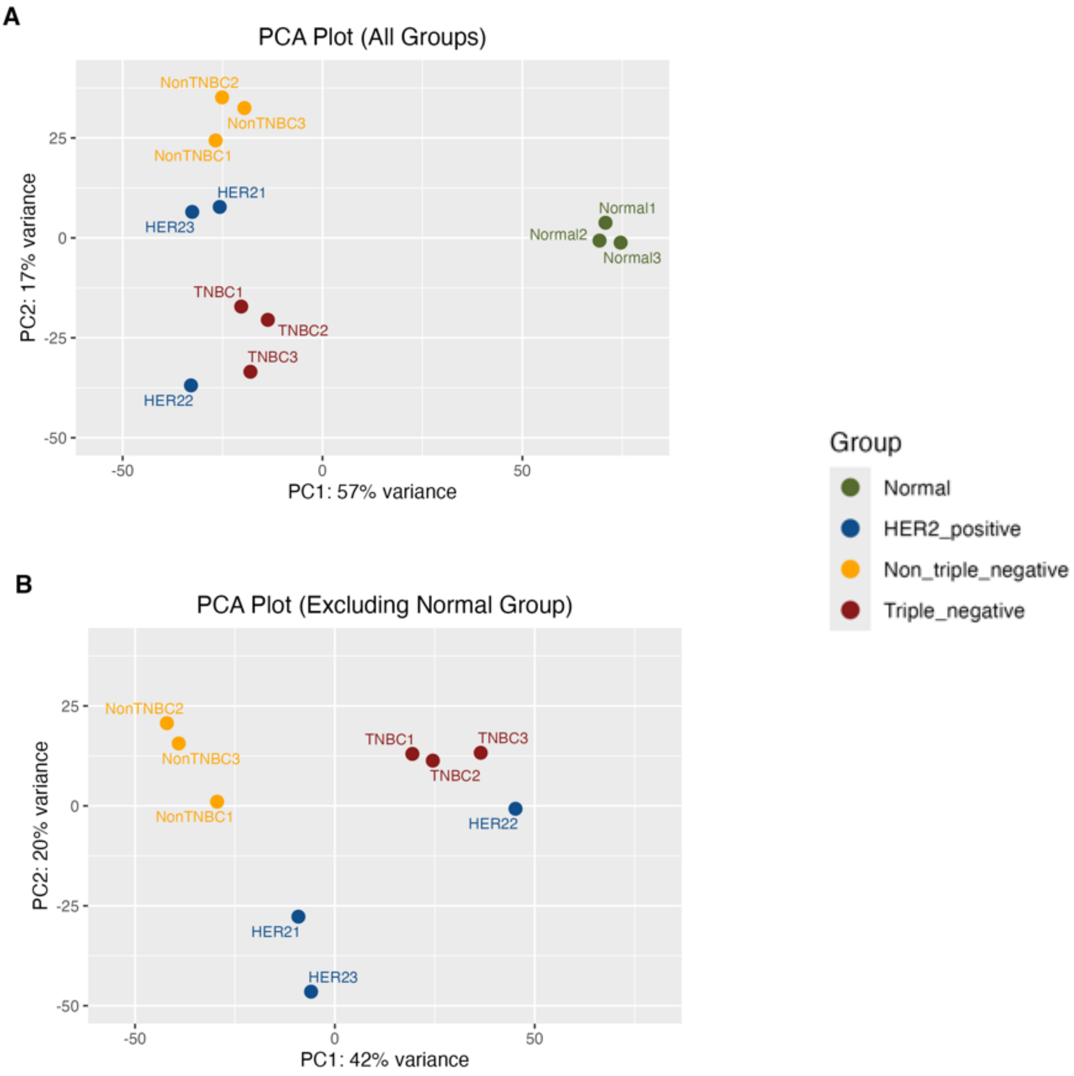

***Figure 1**: **Principal component analysis (PCA) plots of transcriptomic data for breast cancer subtypes.** (A) PCA plot including all groups: HER2-positive, TNBC, Non-TNBC, and Normal. (B) PCA plot excluding the Normal group to focus on differences between cancer subtypes.*

Pairwise comparisons between breast cancer subtypes showed significant differences in gene expression. Comparing TNBC to the HER2-positive group revealed 6338 significant DEGs, with 4137 upregulated and 2201 downregulated in TNBC (Figure 2A). In the Non-TNBC vs. HER2-positive comparison, 5507 genes were differentially expressed, including 2791 upregulated and 2716 downregulated in Non-TNBC (Figure 2B). The Non-TNBC vs. TNBC comparison showed the fewest DEGs (1620), with 757 upregulated and 863 downregulated in TNBC (Figure 2C).

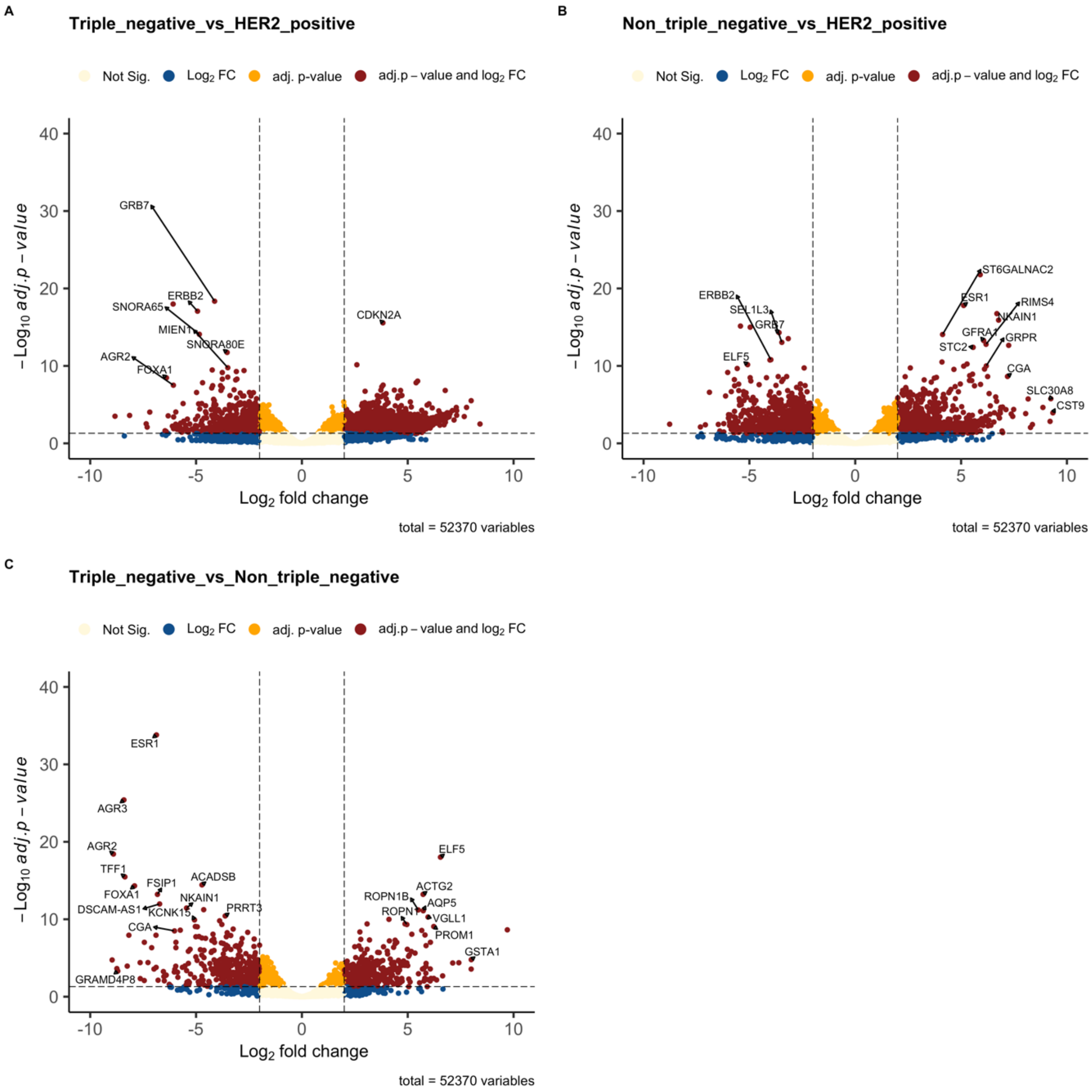

Figure 2: Volcano plots showing differentially expressed genes (DEGs) in pairwise comparisons between breast cancer subtypes. *Each plot displays the log2 fold change (x-axis) versus the -log10 adjusted p-value (y-axis) for genes in the comparison. Horizontal dashed lines indicate the adjusted p-value cutoff of 0.05, and vertical dashed lines represent the log2 fold change cutoff of ±2. Genes meeting both cutoffs (adjusted p-value < 0.05 and |log2 fold change| ≥ 2) are highlighted in red and considered significantly differentially expressed. Comparisons include (A) TNBC vs. HER2-positive, (B) Non-TNBC vs. HER2-positive, and (C) Non-TNBC vs. TNBC.*

Normalized expression levels of all significant DEGs (n = 9165) revealed clustering within cancer subtypes, though expression patterns varied (Figure 3A). This highlights the intricate genetic makeup of breast cancer, shaped by diverse microenvironments, epigenetic modifications, and regulatory mechanisms. Examining the top 50 DEGs according to absolute log2 fold change underscores this heterogeneity (Figure 3B). Interestingly, most of these genes were highly expressed only in one of the three replicates of a subtype. This variability not only reflects the complexity of gene regulation in breast cancer but also suggests that tumor heterogeneity extends beyond subtype classification.

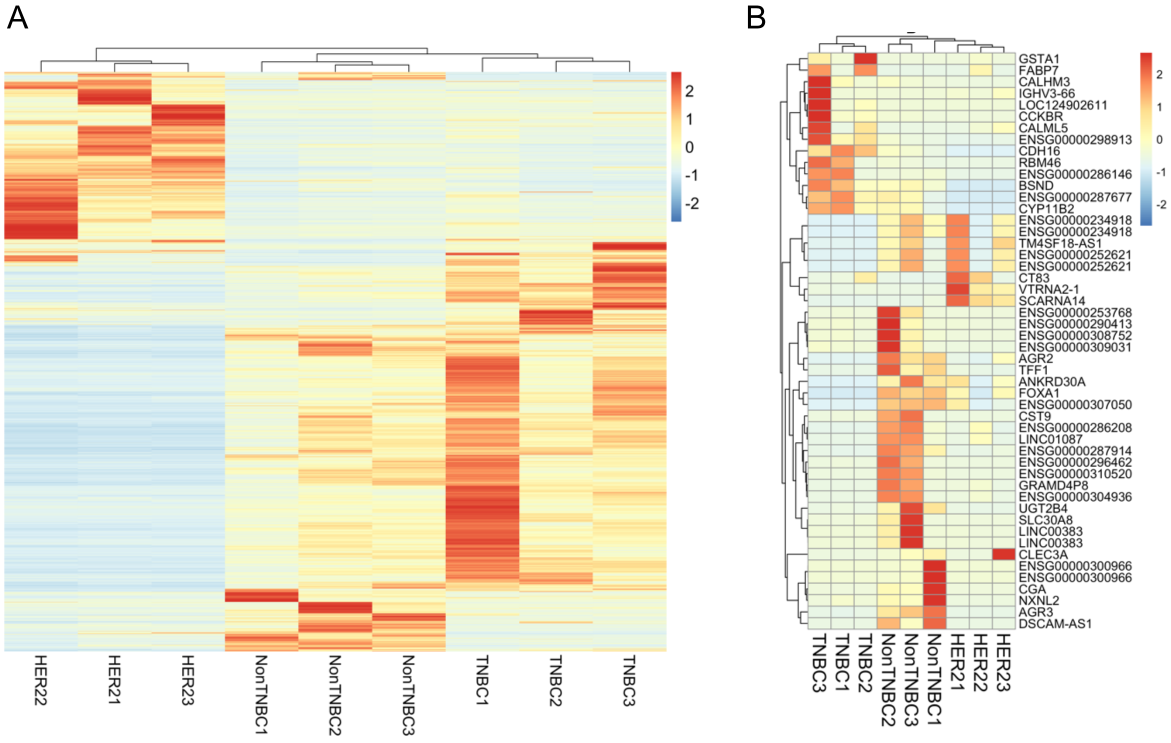

***Figure 3: Heatmap of expression patterns of differentially expressed genes in breast cancer subtypes** (A) Heatmap of normalized expression levels for all significantly differentially expressed genes (DEGs) (n = 9165), showing clustering within breast cancer subtypes despite variability in expression patterns. (B) Heatmap of the top 50 DEGs ranked by absolute log2 fold change, highlighting expression heterogeneity.*

Overrepresentation analysis

To identify biological processes associated with the differences between breast cancer subtypes, ORA was performed on sets of DEGs using the entire dataset of genes in the experiment as the background. Of the 52 370 background genes, 17 241 were assigned and included in the analysis.

For the TNBC vs. HER2-positive comparison, ORA identified 97 significantly enriched GO terms, including "detection of stimulus involved in sensory perception" and "ion transport" (Figure 4A). The majority of genes involved in these processes were upregulated in TNBC, suggesting that TNBC may have distinct sensory and ion regulatory mechanisms compared to HER2-positive breast cancer. These pathways could be important for the aggressiveness and metastatic potential observed in TNBC. It was previously seen that olfactory receptors are associated with metastases[22].

In the Non-TNBC vs. HER2-positive comparison, ORA revealed 69 significantly enriched GO terms, including "import across plasma membrane" and "lymphocyte-mediated immunity" (Figure 4B). The lower amount of immune-related genes in Non-TNBC compared to HER2-positive could be due to increase tumor infiltrating lymphocytes seen in HER2-positive breast cancer[23].

Finally, in the TNBC vs. Non-TNBC comparison, ORA identified 8 significantly enriched GO terms, with "morphogenesis of an epithelium" and "tissue regeneration" standing out (Figure 4C). These terms point to potential differences in the capacity for tissue remodeling and epithelial organization between TNBC and Non-TNBC, which could influence clinical behaviors and responses to therapy.

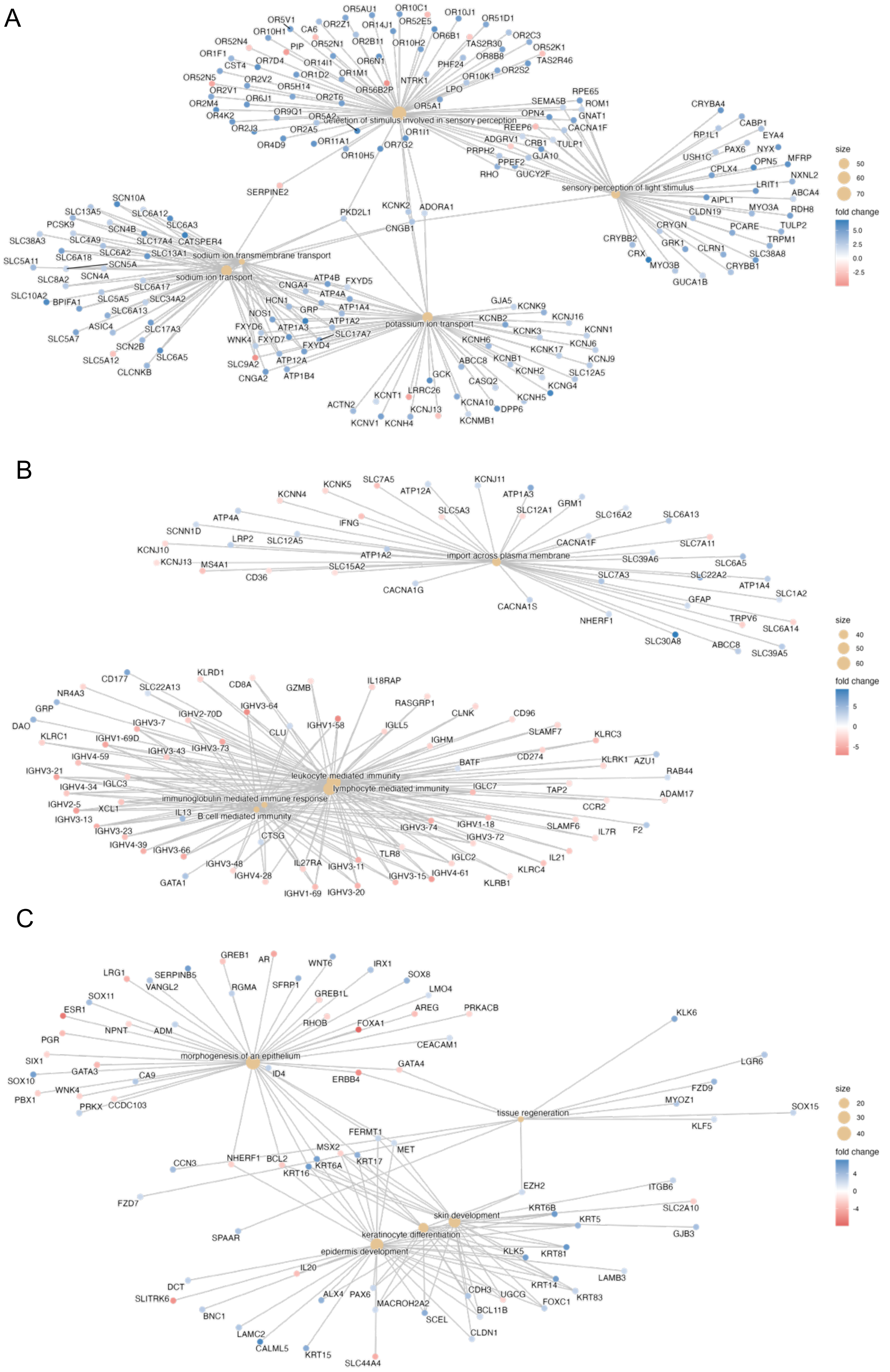
A
detection of stimulus involved in sensory perception
sensory perception of light stimulus
sodium ion transmembrane transport
sodium ion transport
potassium ion transport
OR5AU1 OR10C1 OR10J1 OR51D1 OR5V1 OR10H1 CA6 OR2Z1 OR14J1 OR52E5 OR6B1 OR2C3 OR52N4 PIP OR52N1 OR2B11 OR10H2 TAS2R30 OR52K1 OR1F1 OR7D4 OR14I1 OR6N1 OR8B8 TAS2R46 CST4 OR1D2 OR1M1 PHF24 OR2S2 OR52N5 OR2V2 OR5H14 OR10K1 OR2V1 OR56B2P NTRK1 LPO OR6J1 OR2T6 RPE65 OR2M4 OR9Q1 SEMA5B ROM1 OR4K2 OR5A2 OR5A1 OPN4 GNAT1 OR2J3 OR2A5 REEP6 CACNA1F OR4D9 OR11A1 OR1I1 ADGRV1 CRB1 TULP1 OR10H5 OR7G2 PRPH2 GJA10 PPEF2 RHO GUCY2F
CRYBA4 CABP1 RP1L1 EYA4 USH1C PAX6 NYX CPLX4 OPN5 MFRP NXNL2 LRIT1 ABCA4 AIPL1 CLDN19 MYO3A RDH8 CRYGN PCARE TULP2 CRYBB2 GRK1 TRPM1 CLRN1 SLC38A8 CRX MYO3B CRYBB1 GUCA1B
SERPINE2 PKD2L1 KCNK2 ADORA1 CNGB1
SCN10A SLC13A5 SLC6A12 PCSK9 SLC6A3 SCN4B SLC38A3 SLC4A9 SLC17A4 CATSPER4 SLC6A18 SLC6A2 SLC13A1 SLC5A11 SCN5A SLC8A2 SCN4A SLC6A17 SLC10A2 BPIFA1 SLC5A5 SLC34A2 SLC6A13 ASIC4 SLC17A3 SLC5A7 SCN2B SLC6A5 SLC5A12 CLCNKB
ATP4B CNGA4 FXYD5 ATP4A HCN1 ATP1A4 NOS1 GRP ATP1A2 FXYD6 ATP1A3 SLC17A7 WNK4 FXYD7 FXYD4 SLC9A2 ATP12A CNGA2 ATP1B4
GJA5 KCNK9 KCNB2 KCNJ16 KCNK3 KCNN1 KCNH6 KCNJ6 KCNK17 ABCC8 KCNB1 KCNJ9 KCNH2 SLC12A5 GCK CASQ2 KCNG4 LRRC26 KCNT1 KCNH5 KCNA10 ACTN2 KCNJ13 DPP6 KCNV1 KCNH4 KCNMB1
size 50 60 70
fold change 5.0 2.5 0.0 -2.5
B
import across plasma membrane
KCNJ11 SLC7A5 KCNK5 ATP12A ATP1A3 KCNN4 GRM1 SLC5A3 SLC12A1 SLC16A2 SLC6A13 IFNG ATP4A SCNN1D LRP2 SLC12A5 CACNA1F SLC7A11 KCNJ10 ATP1A2 SLC39A6 KCNJ13 MS4A1 SLC15A2 SLC6A5 CD36 SLC7A3 SLC22A2 ATP1A4 CACNA1G SLC1A2 GFAP CACNA1S TRPV6 NHERF1 SLC6A14 SLC30A8 ABCC8 SLC39A5
leukocyte mediated immunity
lymphocyte mediated immunity
immunoglobulin mediated immune response
B cell mediated immunity
KLRD1 CD177 IL18RAP SLC22A13 CD8A GZMB NR4A3 IGHV3-64 RASGRP1 GRP IGHV2-70D CLNK CD96 IGHV3-7 IGHV1-58 DAO CLU IGLL5 SLAMF7 KLRC1 IGHV1-69D IGHM KLRC3 IGHV3-43 IGHV3-73 CD274 IGHV4-59 BATF KLRK1 AZU1 IGHV3-21 RAB44 IGHV4-34 IGLC3 IGLC7 TAP2 ADAM17 IGHV2-5 CCR2 XCL1 IL13 IGHV3-13 IGHV1-18 SLAMF6 IGHV3-23 IGHV3-74 IL7R F2 CTSG IGHV3-72 IGHV4-39 IGHV3-66 TLR8 IL21 IGHV3-48 IL27RA IGHV3-11 IGLC2 KLRC4 IGHV4-28 IGHV3-15 IGHV4-61 GATA1 KLRB1 IGHV1-69 IGHV3-20
size 40 50 60
fold change 5 0 -5
C
morphogenesis of an epithelium
tissue regeneration
skin development
keratinocyte differentiation
epidermis development
GREB1 AR WNT6 IRX1 LRG1 SERPINB5 SOX8 VANGL2 RGMA SFRP1 LMO4 SOX11 GREB1L ESR1 AREG PRKACB RHOB FOXA1 NPNT ADM CEACAM1 PGR SIX1 GATA4 GATA3 ID4 CA9 ERBB4 SOX10 WNK4 PBX1 PRKX CCDC103
KLK6 LGR6 FZD9 MYOZ1 SOX15 KLF5
FERMT1 MSX2 MET CCN3 NHERF1 BCL2 KRT17 KRT6A KRT16 FZD7 EZH2 ITGB6 KRT6B SLC2A10 KRT5 SPAAR GJB3 KLK5 KRT81 IL20 LAMB3 DCT KRT14 SLITRK6 ALX4 PAX6 CDH3 UGCG FOXC1 KRT83 MACROH2A2 BCL11B BNC1 SCEL LAMC2 CLDN1 CALML5 KRT15 SLC44A4
size 20 30 40
fold change 4 0 -4

***Figure 4: Gene-concept network plots for enriched biological processe**. (A) Overrepresentation analysis (ORA) of differentially expressed genes (DEGs) in TNBC vs. HER2-positive breast cancer. (B) ORA for Non-TNBC vs. HER2-positive breast cancer. (C) ORA for TNBC vs. Non-TNBC.*

Discussion

This study analyzed the transcriptomic profiles of HER2-positive, TNBC, and Non-TNBC breast cancer subtypes to identify DEGs with therapeutic potential. A total of 9165 DEGs were identified, with functional enrichment analysis highlighting key pathways, such as increased olfactory receptor activation in TNBC and elevated tumor-infiltrating lymphocytes in HER2-positive tumors. These pathways may represent promising therapeutic targets.

Comparison with the original findings of Eswaran et al.[4] confirmed differential expression of key genes (e.g., *ESR1, AGR3, FOXA1*) while also identifying new DEGs (such as *SNORA53, CYP2B7P, TMEM178B*). A key observation was the substantial variability in gene expression within subtypes, underscoring the heterogeneity of breast cancer tissue. This variability likely reflects differences in the tumor microenvironment, epigenetic regulation, and sample-specific factors. Spatiotemporal analyses showed unique transcriptomic features at tumor borders and dynamic expression changes over time, with initial differences in primary tumors diminishing at later stages[24].

Despite the robustness of this analysis, certain limitations must be acknowledged. The absence of unique molecular identifiers limited the ability to distinguish between PCR duplicates and true biological variation, potentially introducing bias into expression profiles. Additionally, longer sequencing read length could enhance mapping accuracy. Moreover, GO term analysis was restricted by incomplete gene annotations, limiting pathway insights as only a third of all genes were annotated.

Future research should validate these DEGs as therapeutic targets through functional studies (e.g., knockdown/overexpression experiments) and single-cell RNA sequencing to resolve intra-subtype heterogeneity. Larger, spatiotemporal cohorts with detailed clinical metadata (such as tumor stage and treatment status) would enhance the generalizability of findings and clarify the role of immune infiltration in shaping gene expression patterns.

While this study advances understanding of transcriptomic differences among breast cancer subtypes, translating these findings into clinical applications requires integrative approaches combining genomic, proteomic, and clinical data. The identified pathways suggest distinct therapeutic vulnerabilities, but further validation is essential to determine their potential for targeted therapies.

In conclusion, this study provides valuable insights into the molecular landscape of breast cancer subtypes, highlighting potential therapeutic targets and biological pathways. However, further validation and large-scale studies will be essential to fully determine whether the identified DEGs can be effectively leveraged for targeted therapies.

References

1. WHO. Breast cancer. March 13, 2024. Accessed January 29, 2025. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

2. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primer*. 2019;5(1):66. doi:10.1038/s41572-019-0111-2

3. Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol*. 2015;26(8):1533-1546. doi:10.1093/annonc/mdv221

4. Eswaran J, Cyanam D, Mudvari P, et al. Transcriptomic landscape of breast cancers through mRNA sequencing. *Sci Rep*. 2012;2(1):264. doi:10.1038/srep00264

5. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. Published online January 8, 2019. Accessed November 9, 2024. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

6. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

7. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354

8. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Res*. 2024;52(D1):D891-D899. doi:10.1093/nar/gkad1049

9. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37(8):907-915. doi:10.1038/s41587-019-0201-4

10. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008

11. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma Oxf Engl*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

12. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2024. https://www.R-project.org/

13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

14. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org

15. Auguie B. gridExtra: Miscellaneous Functions for “Grid” Graphics. Published online 2017. https://CRAN.R-project.org/package=gridExtra

16. Blighe K, Rana S, Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. Published online 2024. doi:10.18129/B9.bioc.EnhancedVolcano

17. Slowikowski K. ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2.” Published online 2024. https://CRAN.R-project.org/package=ggrepel

18. Carlson M, Falcon S, Pages H, Li N. org. Hs. eg. db: Genome wide annotation for Human. *R Package Version*. 2019;3(2):3.

19. Kolde R. *Pheatmap: Pretty Heatmaps*.; 2019. https://CRAN.R-project.org/package=pheatmap

20. Xu S, Hu E, Cai Y, et al. Using clusterProfiler to characterize multiomics data. *Nat Protoc*. 2024;19(11):3292-3320. doi:10.1038/s41596-024-01020-z

21. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. 2015;31(4):608-609. doi:10.1093/bioinformatics/btu684

22. Li M, Schweiger MW, Ryan DJ, Nakano I, Carvalho LA, Tannous BA. Olfactory receptor 5B21 drives breast cancer metastasis. *iScience*. 2021;24(12):103519. doi:10.1016/j.isci.2021.103519

23. Onkar SS, Carleton NM, Lucas PC, et al. The Great Immune Escape: Understanding the Divergent Immune Response in Breast Cancer Subtypes. *Cancer Discov*. 2023;13(1):23-40. doi:10.1158/2159-8290.CD-22-0475

24. Greenwald NF, Nederlof I, Sowers C, et al. Temporal and spatial composition of the tumor microenvironment predicts response to immune checkpoint inhibition. Published online January 28, 2025:2025.01.26.634557. doi:10.1101/2025.01.26.634557