Gene Expression **2025** vol. 24(3) | 159–170
DOI: 10.14218/GE.2024.00071

Original Article

Unraveling Tumor Heterogeneity: Quantitative Insights from Single-cell RNA Sequencing Analysis in Breast Cancer Subtypes

Daniela Senra[1,2], Nara Guisoni[1,2,3] and Luis Diambra[1,2*]

[1]Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina; [2]CCT La Plata, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), La Plata, Buenos Aires, Argentina; [3]Instituto de Tecnología (INTEC), Universidad Argentina de la Empresa (UADE), Buenos Aires, Argentina

Received: December 08, 2024 | **Revised:** March 17, 2025 | **Accepted:** March 26, 2025 | **Published online:** April 25, 2025

Abstract

Background and objectives: Tumors are complex systems characterized by variations across genetic, transcriptomic, phenotypic, and microenvironmental levels. This study introduced a novel framework for quantifying cancer cell heterogeneity using single-cell RNA sequencing data. The framework comprised several scores aimed at uncovering the complexities of key cancer traits, such as metastasis, tumor progression, and recurrence.

Methods: This study leveraged publicly available single-cell transcriptomic data from three human breast cancer subtypes: estrogen receptor-positive, human epidermal growth factor receptor 2-positive, and triple-negative. We employed a quantitative approach, analyzing copy number alterations (CNAs), entropy, transcriptomic heterogeneity, and diverse protein-protein interaction networks (PPINs) to explore critical concepts in cancer biology.

Results: We found that entropy and PPIN activity related to the cell cycle could distinguish cell clusters with elevated mitotic activity, particularly in aggressive breast cancer subtypes. Additionally, CNA distributions varied across cancer subtypes. We also identified positive correlations between the CNA score, entropy, and the activities of PPINs associated with the cell cycle, as well as those linked to basal and mesenchymal cell lines.

Conclusions: This study addresses a gap in the current understanding of breast cancer heterogeneity by presenting a novel quantitative approach that offers deeper insights into tumor biology, surpassing traditional marker-based methods.

Introduction

Breast cancer is the most commonly diagnosed cancer and the fifth leading cause of cancer-related mortality worldwide.[1,2] It encompasses a diverse set of diseases, which has led to the development of various classification systems over time. Currently, immunohistochemistry for hormone receptors, including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), is one of the most extensively employed approaches for classification. This system identifies four main subtypes of breast cancer: luminal A, luminal B, HER2-amplified, and triple-negative.[3–5] Luminal A and B subtypes are characterized by the presence of estrogen receptors (ER+), and together, they constitute approximately 70% of all breast cancers, generally carrying a favorable prognosis. HER2-amplified (HER2+) breast cancer, which accounts for 15–20% of cases, is characterized by increased HER2 expression and the absence of ER. This subtype tends to behave more aggressively compared to the luminal subtypes.[4] Triple-negative (TN) breast cancer, which makes up about 15% of diagnoses, lacks ER, PR, and HER2 expression, exhibiting the most aggressive behavior with poor differentiation and high proliferation rates. The terms TN and basal breast cancer are often used interchangeably due to significant similarities in their signatures; however, not all TN breast cancers are basal.[6,7] Despite this broad classification based on receptor expression, histological categories

Keywords: Breast cancer; ER+; HER2+; Triple negative; Tumor heterogeneity; scRNA-seq; Copy number alteration; Protein networks.

[*]**Correspondence to:** Luis Diambra, Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, Boulevard 120 N° 1461 La Plata, Buenos Aires CP1900, Argentina. ORCID: https://orcid.org/0000-0001-8052-4880. Tel: +54-221-6086411, E-mail: ldiambra@gmail.com

How to cite this article: Senra D, Guisoni N, Diambra L. Unraveling Tumor Heterogeneity: Quantitative Insights from Single-cell RNA Sequencing Analysis in Breast Cancer Subtypes. *Gene Expr* 2025;24(3):159–170. doi: 10.14218/GE.2024.00071.

© 2025 The Author(s). This article has been published under the terms of Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0), which permits noncommercial unrestricted use, distribution, and reproduction in any medium, provided that the following statement is provided.
"This article has been published in *Gene Expression* at https://doi.org/10.14218/GE.2024.00071 and can also be viewed on the Journal's website at https://www.xiahepublishing.com/journal/ge".

also account for rare subtypes, such as medullary breast cancer or invasive micropapillary carcinomas.[8,9]

Tumors exhibit a wide range of phenotypic and molecular characteristics at both intertumor and intratumor levels. Heterogeneity, which has been extensively studied for its implications in diagnostics and treatment selection, presents significant challenges, as different tumors and cells respond differently to the same therapeutic approach.[10] Intertumor heterogeneity refers to variations between distinct tumors, whether from different or the same patient. Intratumor heterogeneity, on the other hand, describes the mixture of cancer cell subpopulations within a single tumor, each with diverse genetic, transcriptomic, and phenotypic profiles, along with complex interactions with the tumor microenvironment (TME). In recent years, several studies have explored intratumor heterogeneity at the transcriptomic level in breast cancer using single-cell RNA sequencing (scRNA-seq) data.[11–15] These investigations have focused on identifying distinct cell clusters and their corresponding gene expression signatures. However, while quantitative assessments of intratumor heterogeneity have been conducted in other cancer types,[11,16] the field of breast cancer research lacks a quantitative exploration of this phenomenon. scRNA-seq provides detailed insights into the molecular landscape of breast cancer subtypes, allowing for the characterization of variations in key biological processes, copy number alterations (CNAs), and pluripotency. This approach also offers the potential to explore correlations between these features and breast cancer subtypes, as well as their potential links to tumor aggressiveness and therapeutic resistance. Thus, scRNA-seq studies enhance our understanding of cancer dynamics at the single-cell level, paving the way for precision oncology strategies tailored to breast cancer subtypes.

In this study, we used quantitative measures on scRNA-seq data to analyze several hallmarks of cancer, focusing on cancer cells from the three primary breast cancer subtypes (ER+, HER2+, and TN). By quantifying these features and investigating their associations with the subtypes, our research aimed to deepen the understanding of the intricate relationships between these features and breast cancer subtypes. These measures include intratumor transcriptomic heterogeneity, CNAs, entropy, and key protein-protein interaction (PPIN) activity. We developed mathematical scores for each measure, as detailed in the Materials and Methods section.

Materials and methods

Dataset description

The dataset utilized in this study was obtained from scBrAtlas,[13] which is available in two forms: raw count matrices in the GEO database (series GSE161529) and preprocessed R objects on Figshare.[17] For our analysis, we utilized the preprocessed R objects from Figshare. The scBrAtlas comprises scRNA-seq samples derived from various human breast cancer states, including normal, preneoplastic, and cancerous conditions. This dataset includes approximately 430,000 individual cells from 69 surgical samples collected from 55 patients. Quality control metrics were applied to ensure the data's reliability. The samples were filtered based on criteria such as library size, number of genes per cell, and percentage of mitochondrial content per cell. Detailed filtering procedures are outlined in Table S1. Following this process, approximately 15% of the cells were excluded from each sample, leaving a total of 341,874 cells for further analysis. A comprehensive description of the preprocessing phase, along with the corresponding R code, is available in reference.[18] Since the data had already been preprocessed, we verified this step and directly used the available preprocessed data.

Our primary focus was on cancer cells, so samples from male patients, healthy individuals, precancerous tissues, and cancerous tissues associated with lymph nodes were excluded from the analysis. Table 1 provides a detailed description of the samples obtained from women with breast cancer, including patient age, cancer subtype, tumor size, grade, and number of cancer cells. Each sample contained a mixture of tumor cells, normal epithelial cells, and cells from the TME, such as fibroblasts, endothelial cells, and immune cells. The original data source labeled the cells, distinguishing cancer cells from normal cells using inferCNV.[19–21] For our downstream analysis, we excluded all microenvironment cells, retaining only epithelial cells. Furthermore, we subset the cancer cells by filtering out normal epithelial cells using the original labels. After cell filtering, we excluded samples containing fewer than 1,000 cancer cells. Based on these criteria, six ER+ samples (ER-0001, ER-0125, ER-0360, ER-0042, ER-0025, and ER-0163), five HER2+ samples (HER2-308, HER2-0337, HER2-0031, HER2-0161, and HER2-0176), and six TN samples (TN-0126, TN-0135, TN-B1-4031, TN-B1-0131, TN-B1-0554, and TN-B1-0177) were selected for further analysis. The sample labels provided by the data authors were preserved.

Data preprocessing and analysis were performed using R (version 4.3.1) and the Seurat package (version 4.4.0). Samples were integrated separately across subtypes using the Seurat 4 pipeline.[22,23] The ER+, HER2+, and TN breast cancer integrated datasets are shown in Figure 1a–c, with samples distinguished by the color of the cells.

Scores

To investigate potential links between tumor aggressiveness and specific biological features, we established several parameters to measure these features, including the level of CNAs, intratumor heterogeneity, entropy, and the activity of specific PPINs potentially associated with tumor aggressiveness, such as epithelial to mesenchymal transition (EMT) and cell cycle regulation. In this section, we provide precise definitions of these measures and the rationale for their application.

CNA score

CNAs are changes in the number of gene copies within tumor cells, which can provide a selective advantage, leading to increased expression of certain genes and reduced expression of others. This reflects the extent and type of genomic instability unique to each tumor.[24] CNAs have been linked to cancer progression and poor prognosis in breast cancer.[25,26] In the context of breast cancer, CNA inference has primarily been focused on discriminating cancer cells from non-tumor cells rather than quantifying the extent of CNAs.[11,13] One of our main interests involved exploring somatic CNAs. To deduce CNAs from scRNA-seq data, we employed the inferCNV R package (version 1.16.0),[19–21] which detects somatic chromosomal-scale CNAs by assessing the relative gene expression levels of contiguous genes along the genome and comparing them to a reference set of "normal" cells. We applied inferCNV to the samples based on cancer subtypes, using a breast normal epithelial sample from scBrAtlas (labeled as N0372) as our reference population, as done by the dataset creators.[13] A sliding window of 100 contiguous genes was used. The pre-existing classification of cells into cancer and non-cancer categories, provided with the preprocessed dataset, was validated using the inferCNV profiles

Table 1. Description of breast cancer samples

Sample ID	Patient age	Cancer subtype	Size (mm)	Grade	Cancer cells after filtering
TN-0126*	64	TN	64	3	1,235
TN-0135*	61	TN	22	3	1,433
TN-106	65	TN	25	3	54
TN-0114-T2	84	TN	17	3	177
TN-B1-4031*	25	TN (BRCA1)	20	3	5,129
TN-B1-0131*	84	TN (BRCA1)	25	3	5,513
TN-B1-0554*	29	TN (BRCA1)	37	3	2,337
TN-B1-0177*	30	TN (BRCA1)	13	3	1,575
HER2-0308*	32	HER2+	20	3	3,317
HER2-0337*	66	HER2+	67	3	3,924
HER2-0031*	47	HER2+	18	3	1,606
HER2-0069	71	HER2+	27	3	196
HER2-0161*	80	HER2+	45	3	4,124
HER2-0176*	60	HER2+	20	3	4,682
ER-0319	58	PR+	27	3	568
ER-0001*	58	ER+	32	3	4,559
ER-0125*	45	ER+	48	2	3,678
ER-0360*	70	ER+	50	2	1,934
ER-0032	55	ER+	90	3	417
ER-0042*	58	ER+	18	2	2,899
ER-0025*	52	ER+	23	2	4,499
ER-0151	49	ER+	35	2	749
ER-0163*	45	ER+	45	3	5,378

*Samples with more than 1,000 tumor cells after filtering were selected for analysis. BRCA1, breast cancer susceptibility gene 1 mutated; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; TN, triple-negative.

obtained in alignment with the labeled data.

Several computational tools, including inferCNV, copyKAT,[27] and CaSpER,[28] are commonly used to estimate CNAs in scRNA-seq data. These tools are primarily designed to distinguish tumor cells from non-tumor cells but can also be valuable for quantifying CNA extent within individual cells. To achieve this, we defined a CNA level for each cell i based on the residual expression matrix generated by infer CNV:

$$CNA_i = \frac{1}{m}\sum_{j=1}^{m}\left(X_{ij}-1\right)^2.$$

We denote the transposed residual expression matrix obtained from inferCNV by X, a matrix with n rows (cells) and m columns (genes). For simplicity, we avoid using the transposed symbol. This matrix serves as a surrogate for CNA in each gene j across every cell i. A value of 1 indicates neutrality, values exceeding 1

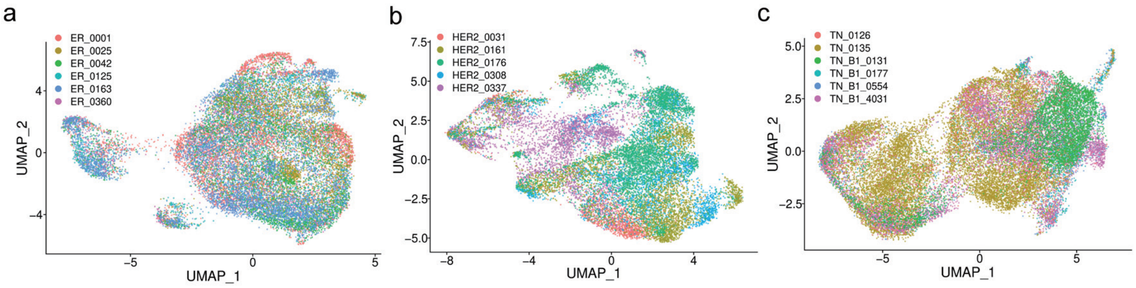

Fig. 1. UMAP visualization of cancer subtypes. Visualization in the UMAP space of samples corresponding to the subtypes ER+ (a), HER2+ (b), and TN (c). The samples were categorized by cancer subtype and integrated. Cells are color-coded according to the sample. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; TN, triple-negative; UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction.

Table 2. Details of gene sets used to compute the PPIN activity

Gene set	Activity symbol	Description	Reference
GO:0007049	$\langle ACT\rangle^{CC}$	Cell cycle: The progression of biochemical and morphological phases and events that occur in a cell during successive cell replication or nuclear replication events.	37
GO:0010718	$\langle ACT\rangle^{EMT}$	Positive regulation of epithelial to mesenchymal transition	37
CHARAFE: breast cancer luminal vs basal dn	$\langle ACT\rangle^{LB\text{-}dn}$	Characterization of breast cell lines: Luminal vs. Basal differentially expressed genes (downregulated)	38–40
CHARAFE: breast cancer luminal vs basal up	$\langle ACT\rangle^{LB\text{-}up}$	Characterization of breast cell lines: Luminal vs. Basal differentially expressed genes (upregulated)	38–40
CHARAFE: breast cancer luminal vs mesenchymal dn	$\langle ACT\rangle^{LM\text{-}dn}$	Characterization of breast cell lines: Luminal vs. Mesenchymal differentially expressed genes (downregulated).	38–40
CHARAFE: breast cancer luminal vs mesenchymal up	$\langle ACT\rangle^{LM\text{-}up}$	Characterization of breast cell lines: Luminal vs. Mesenchymal differentially expressed genes (upregulated).	38–40

ACT, activity; CC, cell cycle; CHARAFE, gene-set name; dn, downregulated; EMT, epithelial to mesenchymal transition; GO, Gene Ontology; LB, Luminal vs. basal; LM, Luminal vs. mesenchymal; PPIN, protein-protein interaction network; up, upregulated.

indicate gains, and values below 1 denote losses. Hence, all elements of the matrix are standardized by subtracting 1. It is important to highlight that since the terms are squared, the score reflects the magnitude of the CNA without distinguishing between gains and losses. This score indicates the mean squared dispersion of CNAs relative to the reference normal cells, and it is a variation of scores defined in previous works.[29,30] To quantify the degree of CNA in a given sample, we averaged the scores across all the cells within the sample. This sample score is denoted as $\langle CNA\rangle$.

Transcriptomic heterogeneity

The standard approach for evaluating intratumor heterogeneity in cancer using scRNA-seq data involves performing clustering to identify distinct cell types or states.[31,32] Subsequently, these clusters are characterized using differential expression analysis and enrichment analysis to identify gene markers for each cluster. While this method provides valuable insights into sample composition, it does not offer a score that measures the extent of heterogeneity at the transcriptomic level. To address this, we propose an alternative approach to assess transcriptomic heterogeneity, inspired by previous works.[23]

Our proposed approach involves assessing the variability of each gene from a scRNA-seq sample. Deriving variance from log-normalized data does not account for the inherent mean-variance relationship present in scRNA-seq data. To address this, variance-stabilizing methods are typically used.[33,34] In our study, we employed the variance-stabilizing method from the Seurat package, specifically employing the FindVariableFeatures and HVFinfo functions, which provide an estimator of variance adjusted for the feature mean.[22] This approach allows us to accurately assess gene variability while accounting for the underlying mean-variance relationship in scRNA-seq data. To quantify this variability across the entire transcriptome, we define a sample expression variability score as:

$$\langle VAR\rangle = \frac{1}{m}\sum_{j=1}^{m} VAR_j,$$

where VAR_j represents the variance stabilized for gene j. $\langle VAR\rangle$ denotes the mean variance across genes in a sample, serving as a metric to assess the degree of intratumor heterogeneity within a sample. A lower score indicates greater transcriptomic homogeneity, meaning that cells within a sample share more similar gene expression profiles. Conversely, higher values signify increased transcriptomic heterogeneity, suggesting greater diversity among the cells.

The activity of protein-protein interaction networks

In previous works, we developed a methodology to quantify the cell activity of a PPIN from scRNA-seq data.[35,36] A PPIN associated with a specific set of genes can be constructed by extracting the interactions from the full Human PPIN that involve proteins encoded by the genes within the defined gene set x. A score that reflects the activity of a PPIN can be defined for each cell i as follows:

$$ACT_i^x = \frac{1}{Ne}\sum_{j,k=1}^{m} A_{jk} Y_{ij} Y_{ik},$$

where A_{jk} is the upper triangular adjacency matrix, which characterizes the connectivity between genes j and k in the network. Y is the transposed normalized expression matrix of dimensions n x m, N is the number of edges in the PPIN, and e is the average expression of all genes in cell i. To enhance clarity, each row of Y represents the expression profile of the i-th cell. A normalization factor of $N\,e$ is introduced to account for differences in graph size and mean expression. This factor enables comparisons between samples and various PPINs. To quantify the overall activity levels of a PPIN associated with gene set x within a sample, we compute the average activity across all cells within the sample. This sample score will be denoted by $\langle ACT\rangle^x$.

We computed the activity of the PPINs associated with six *Homo sapiens* gene sets as detailed in Table 2. Genes associated with the biological process “cell cycle” (GO:0007049) were sourced from the QuickGO database,[37] and negative regulators of this process were excluded, as previously done.[36] The remaining gene sets were retrieved from the Molecular Signatures Database (MSigDB).[38,39] These include gene sets related to the positive regulation of EMT and four gene sets associated with a study that characterizes different breast cancer cell lines (basal, luminal, and mesenchymal).[40] These gene sets provide comprehensive insights into distinct cellular behaviors across different breast cancer subtypes. The activities of the PPINs associated with the cell cycle and EMT are denoted as ACT^{CC} and ACT^{EMT}, respectively. We refer to $ACT^{LB\text{-}up}$, $ACT^{LM\text{-}up}$, $ACT^{LB\text{-}dn}$, and $ACT^{LM\text{-}dn}$ as the PPIN activities corresponding to the upregulated and downregulated differentially expressed genes between luminal vs. basal and luminal vs. mesenchymal breast cancer cell lines.

Entropy

We compute the Shannon entropy, a well-established information theory metric used to measure the degree of uncertainty in a system configuration. Entropy is associated with a probability distribution p_j. In the context of scRNA-seq data, this probability can be obtained by dividing the expression of gene *j* by the total expression of cell *i*, as follows:

$$p_{ij} = \frac{Z_{ij}}{\Sigma_j Z_{ij}},$$

where *Z* represents the transposed count matrix.[41–43] Thus, we calculate the Shannon entropy for a specific cell *i* as follows:

$$H_i = -\sum_{j=1}^{m} p_{ij} \log(p_{ij}).$$

To quantify the degree of entropy for a sample, we simply average the entropy across all cells comprising the sample. This sample score will be denoted by $\langle H \rangle$.

Traditionally linked to heterogeneity, in this context, entropy measures the heterogeneity of individual cells, not the sample's intratumor heterogeneity. High *H* indicates a broad range of genes expressed simultaneously within a single cell, which is characteristic of non-specialized cells such as stem or progenitor cells.[41,44] Hence, samples with higher $\langle H \rangle$ scores exhibit more undifferentiated characteristics associated with cancer stemness. It has been reported that breast cancer aggressiveness and therapy resistance may be driven by breast cancer stem cells (CSCs).[45–47] Notably, CSCs have been found to be enriched in TN tumors compared to non-TN breast cancers.[48]

In summary, we have established several scores from scRNA-seq data that capture the levels of CNAs, entropy, and the activity of PPINs at the individual cell level. Overall cell score parameters at the sample level were derived by averaging across all cells within each sample. Hierarchical cluster analysis of the mean scores was performed with *hclust* from the stats R package (version 4.3.1).

Statistical analysis

To assess the statistical significance of the difference in the scores obtained from ER, HER2+, and TN subtypes, we used a custom script in Mathematica Wolfram software (version 13.0), which implements the Mann-Whitney test over samples belonging to cancer subtype pairs (https://reference.wolfram.com/language/ref/MannWhitneyTest). This non-parametric statistical test was used to compare the medians of two independent groups. In our case, the groups consist of samples from the same subtype. We compared the sample scores, indicated by <…>, corresponding to eight measures. Since three types of tumors were studied and eight scores were considered, 24 comparisons were possible. Thus, the Benjamini-Hochberg procedure was applied to compute adjusted *p*-values (*q*-values) for the multiple comparison test.[49] A *q*-value < 0.05 was considered statistically significant. The results of the statistical analysis are summarized in Table S2.

Results

We examined 17 individual samples obtained from breast cancer patients (refer to Table 1). These samples were categorized by cancer subtype and integrated (Fig. 1a–c). To explore CNAs, we conducted an inferCNV analysis for each sample. The resulting heatmaps of chromosomal copy gains or losses are shown in Figure 2a–c. Although the gain-deletion patterns do not exhibit strong synteny between the subtypes, some common characteristics can be highlighted. chr1 exhibited gains at chr1q, which contains several oncogenes such as *NRAS*, *JUN*, *MYCL*, *TAL1*, and *BLYM*, and losses at chr1p. Deletions in chr2 were observed in nearly all samples, irrespective of subtype. chr8q amplifications, encompassing the MYC proto-oncogene, were frequent across all subtypes. chr19 amplifications were seen in most samples but were more pronounced in ER+ and TN subtypes, aligning with previous findings.[50] Across each subtype, common characteristics among patients were noted. HER2+ samples exhibited consistent amplifications at chr17, particularly at the chr17q12 band harboring the HER2 gene, and frequent deletions at chr13. Note that samples HER2-0308 and TN-B1-0131 exhibit CNA profiles that differ from their respective subtype patterns. To explore the heterogeneity of tumors, we analyzed each cancer subtype using the scores defined in the Methods section. The cell distributions of eight derived scores for each sample are shown in the violin plots of Figure 3. Additionally, Figure 4 visualizes the distribution of a subset of scores for individual cells within the integrated Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) space, specific to each of the three cancer subtypes. The distributions of the *CNA* score are illustrated in Figure 3a, revealing significantly lower values in ER+ samples compared to those observed in HER2+ and TN samples (Mann-Whitney test, *q*-values = 0.015 in both cases). ER+ tumors exhibited lower dispersion of the *CNA* score compared to HER2+/TN tumors. Furthermore, no distinct cell clusters based on this score are observed within the UMAP embedding (see Fig. 4). In contrast, as shown in Figure 3b and c, the scores $\langle H \rangle$ and $\langle ACT \rangle^{CC}$ show a gradual increase across ER+, HER2+, and TN samples. TN samples demonstrated higher $\langle ACT \rangle^{CC}$ compared to the ER+ cancer subtype (Mann-Whitney test, *q*-values = 0.020). Furthermore, cell clusters exhibiting high ACT^{CC} were observed across all cancer subtypes, as depicted in Figure 4. This is consistent with clusters of cycling MKI67+ tumor cells identified across all cancer subtypes in a previous study.[13] On the other hand, the $\langle ACT \rangle^{EMT}$ score shows no significant differences among the cancer subtypes. Furthermore, the UMAP embedding does not reveal evidence of cell clusters exhibiting higher levels of activity in this PPIN (see Fig. S1). Additionally, we explored the activity of the PPINs associated with breast cancer cell lines (see Table 2). The distributions of $ACT^{LM\text{-}up}$ and $ACT^{LM\text{-}dn}$ in the UMAP space were similar to those of $ACT^{LB\text{-}up}$ and $ACT^{LB\text{-}dn}$, respectively (see Figs. 4 and S1). As expected, TN samples displayed significantly higher values of $\langle ACT \rangle^{LB\text{-}dn}$ and $\langle ACT \rangle^{LM\text{-}dn}$ compared to ER+ samples (Mann-Whitney test, *q*-values = 0.015 in both cases), as illustrated in Figure 3e, f, and Figure 4. Conversely, as seen in Figure 3g, h, and Figure 4, ER+ samples exhibited significantly higher $\langle ACT \rangle^{LB\text{-}up}$ and $\langle ACT \rangle^{LM\text{-}up}$ scores compared to TN samples (Mann-Whitney test, *p*-values = 0.015 in both cases). Similarly, the samples derived from the other luminal tumor (HER2+) also exhibited significantly higher $\langle ACT \rangle^{LB\text{-}up}$ and $\langle ACT \rangle^{LM\text{-}up}$ scores compared to TN samples (Mann-Whitney test, *p*-values = 0.020 and 0.015, respectively). These results indicate that many tumor cells preserve the transcriptional landscape of the original lineage. Furthermore, ER+ and HER2+ samples display cell clusters with high $ACT^{LB\text{-}up}$ scores. In contrast, TN samples exhibit clusters with high $ACT^{LB\text{-}dn}$ scores (Fig. 4). Interestingly, cells with strong original lineage features are co-localized in the UMAP space with cells exhibiting high entropy *H*. To assess the relationship between the scores, we computed the correlation matrix among sample averages using the Pearson correlation coefficient, depicted in Figure 5a.

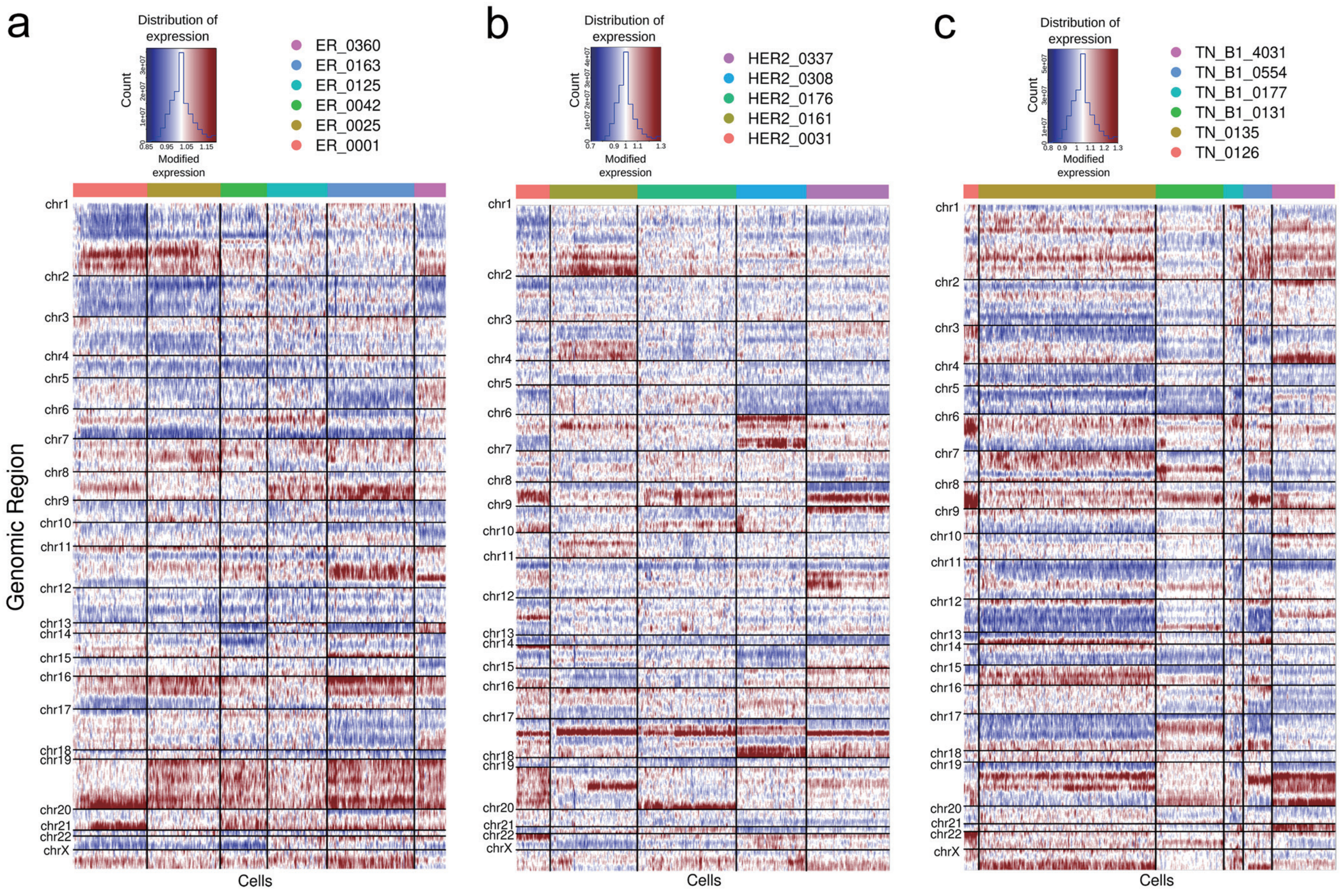

Fig. 2. Heatmap plots displaying the CNA. Average changes in the number of gene copies in tumor cells, obtained with inferCNV, for the three sample subtypes: ER+ (a), HER2+ (b), and TN (c). The genes are sorted by genome location (vertical axis) and grouped by chromosomes, while columns represent individual cells grouped by sample and color-coded according to the respective sample. Blue corresponds to regions with copy number loss, while red corresponds to regions with copy number gain. The histograms at the top of each panel depict the distribution of expression modification in tumor cells relative to reference cells. CNA, copy number alteration; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; TN, triple-negative.

At the sample level, the mean scores $\langle CNA\rangle$, $\langle ACT\rangle^{CC}$, $\langle H\rangle$, $\langle ACT\rangle^{LB\text{-}dn}$, $\langle ACT\rangle^{LM\text{-}dn}$, $\langle ACT\rangle^{EMT}$, and $\langle VAR\rangle$ revealed positive correlations. Notably, the first five scores exhibited stronger correlations among themselves, forming a cluster identified by performing hierarchical cluster analysis, as highlighted in Figure 5a. Specifically, $\langle CNA\rangle$, $\langle H\rangle$, and $\langle ACT\rangle^{CC}$ displayed the strongest Pearson correlation coefficients, ranging from 0.77 to 0.88. Moreover, both $\langle VAR\rangle$ and $\langle ACT\rangle^{EMT}$ exhibited positive correlations with these scores. Conversely, $\langle ACT\rangle^{LM\text{-}up}$ and $\langle ACT\rangle^{LB\text{-}u}$ demonstrated negative correlations with all other scores, except for $\langle VAR\rangle$. These findings suggest that samples with higher CNA burden tend to display increased cycling activity and entropy. Additionally, these samples exhibit a basal and mesenchymal-like phenotype, potentially indicative of the cancer cells' ability to undergo EMT in more aggressive tumors. Intratumor transcriptomic heterogeneity, $\langle VAR\rangle$, was also positively correlated with these scores. However, surprisingly, it also showed a positive correlation with $\langle ACT\rangle^{LM\text{-}dn}$ and $\langle ACT\rangle^{LB\text{-}dn}$, albeit with lower values.

Analysis of the nine sample score distributions across breast cancer subtypes revealed distinct patterns. $\langle CNA\rangle$ exhibited greater heterogeneity between the TN samples compared to the ER+ and HER2+ samples. Additionally, $\langle CNA\rangle$ levels were higher in HER2+ samples compared to ER+ samples, as shown in Figure 5b. Furthermore, the scores $\langle ACT\rangle^{CC}$, $\langle H\rangle$, $\langle ACT\rangle^{EMT}$, $\langle ACT\rangle^{LB\text{-}dn}$ and $\langle ACT\rangle^{LM\text{-}dn}$ were highest in TN samples, followed by intermediate levels in HER2+ samples, and lowest in ER+ samples, as visualized in Figure 5c–g. In Figure 5h and i, the opposite distribution order is observed. TN corresponds to a basal/mesenchymal phenotype, leading to lower activity related to luminal cancer types, while ER+ and HER2+ samples exhibit higher activity in these PPINs. However, it is important to note that these kernel density estimations are based on a limited number of samples, which may lead to estimated distributions with peaks that may not accurately represent the true underlying distributions. These peaks correspond to samples that deviate from the general behavior, as we will discuss later.

In terms of sample variability, ER+ samples showed the most left-skewed distribution, HER2+ samples showed the most right-skewed distribution, and the TN distribution fell between them (see Fig. 5j). One possible explanation for this finding is that HER2+ tumors present both luminal and basal features,[51] and therefore exhibit heterogeneous transcriptomic patterns. This observation suggests that even though variability correlates positively with other scores (albeit to a lesser extent), this score does not necessarily indicate more aggressive tumors.

For a detailed inspection, Figure 6 presents scatter plots of

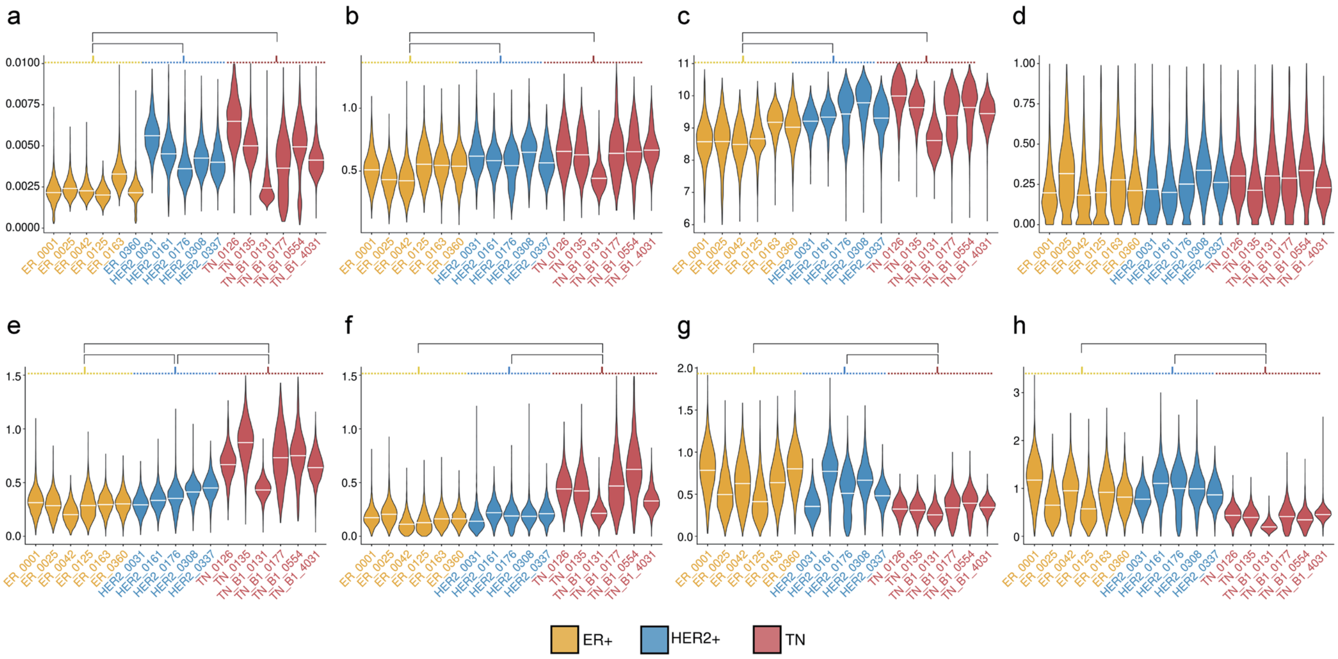

Fig. 3. Violin plots of the scores. Score distribution across the cell population of each sample for CNA (a), $\mathrm{ACT}^{\mathrm{CC}}$ (b), H (c), $\mathrm{ACT}^{\mathrm{EMT}}$ (d), $\mathrm{ACT}^{\mathrm{LB\text{-}dn}}$ (e), $\mathrm{ACT}^{\mathrm{LM\text{-}dn}}$ (f), $\mathrm{ACT}^{\mathrm{LB\text{-}up}}$ (g), and $\mathrm{ACT}^{\mathrm{LM\text{-}up}}$ (h). The white horizontal lines represent the mean value of the corresponding score for each sample. Yellow, blue, and red labels correspond to ER+, HER2+, and TN subtypes, respectively. The mean values obtained for samples of each subtype were compared between subtypes, with significant differences indicated by black lines. The resulting *p*- and *q*-values are listed in Table S2. ACT, activity; CNA, copy number alteration; dn, downregulated; EMT, epithelial to mesenchymal transition; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; LB, Luminal vs. basal; LM, Luminal vs. mesenchymal; TN, triple-negative; up, upregulated.

selected sample parameters to observe their relationships, distinguishing them by cancer subtype and labeling the samples. The positive correlation between the scores, as reported in the correlation matrix (Fig. 5a), can be verified by these scatter plots. Moreover, certain general trends are observable. ER+ samples tend to cluster in the lower regions of the scatter plots, indicating lower scores. In contrast, HER2+ and TN samples are more difficult to differentiate from each other, yet there is a tendency for TN samples to cluster at higher score levels compared to HER2+ samples.

Examining individual samples reveals certain exceptions. For instance, among the TN samples, TN-B1-0131 shows low scores across all parameters except for $\langle VAR \rangle$, resembling profiles found in ER+ samples. Further scrutiny of the metadata reveals that the TN-B1-0131 sample corresponds to an 84-year-old patient (see Table 1), making it 20 years older than the next oldest TN sample when sorted by age. Additionally, it exhibits an age gap of more than 30 years compared to the average age of the other TN samples, which is 42. A similar observation in the opposite direction can be made for HER2-0308. This particular sample shows notably higher scores compared to the others in the same cancer subtype, except for $\langle CNA \rangle$. The donor for this sample is 32 years old (Table 1), which is more than 30 years younger than the mean age of HER2+ samples.[52] The exceptional score values observed may be linked to age-dependent tumor aggressiveness. The residual expression matrices shown in Figure 2 provide complementary information regarding CNA. In contrast to other TN samples, TN-B1-0131 exhibited a low $\langle CNA \rangle$ value. This is reflected in its residual expression matrix shown in Figure 2c, which displays a pattern with fewer chromosomal gains and losses compared to the other TN samples. While HER2-0308 displayed intermediate levels of the $\langle CNA \rangle$ score within the HER2+ group (Fig. 2b), its residual expression matrix deviated from other HER2+ samples. Notably, it exhibited marked amplifications in chr6 and chr17, alongside deletions in chr14.

Discussion

This study presents a quantitative approach to assess key features related to cancer, specifically focusing on the three most prevalent subtypes of breast cancer: ER+, HER2+, and TN. Using scRNA-seq data obtained from human breast cancer samples, we conduct a comprehensive analysis of various cellular characteristics, including CNAs, entropy, and PPIN activity, which are linked to specific biological processes (e.g., EMT, cell cycle, luminal, mesenchymal, and basal breast cell lines). Additionally, we introduce a score that quantifies intratumoral transcriptomic heterogeneity. The novelty of this study lies in the quantitative assessment of these features—a comprehensive approach that has not been explored in the breast cancer single-cell transcriptomics field.

Our investigation at the single-cell level reveals intriguing signatures. The PPIN activity associated with the cell cycle and entropy demonstrates varying degrees of activity across the breast cancer subtypes: ER+, HER2+, and TN, in ascending order. Notably, clusters of cells displaying heightened mitotic activity are observed in all subtypes, with TN samples exhibiting a higher proportion of mitotic cells, consistent with previous studies.[13,18]

The study also highlights distinct CNA distribution patterns between ER+ and HER2+/TN tumors. $\langle CNA \rangle$ was significantly higher in HER2+ and TN samples compared to ER+ samples, but no significant difference was observed between HER2+ and TN tumors. Specifically, gains at chr1q have been reported in approxi-

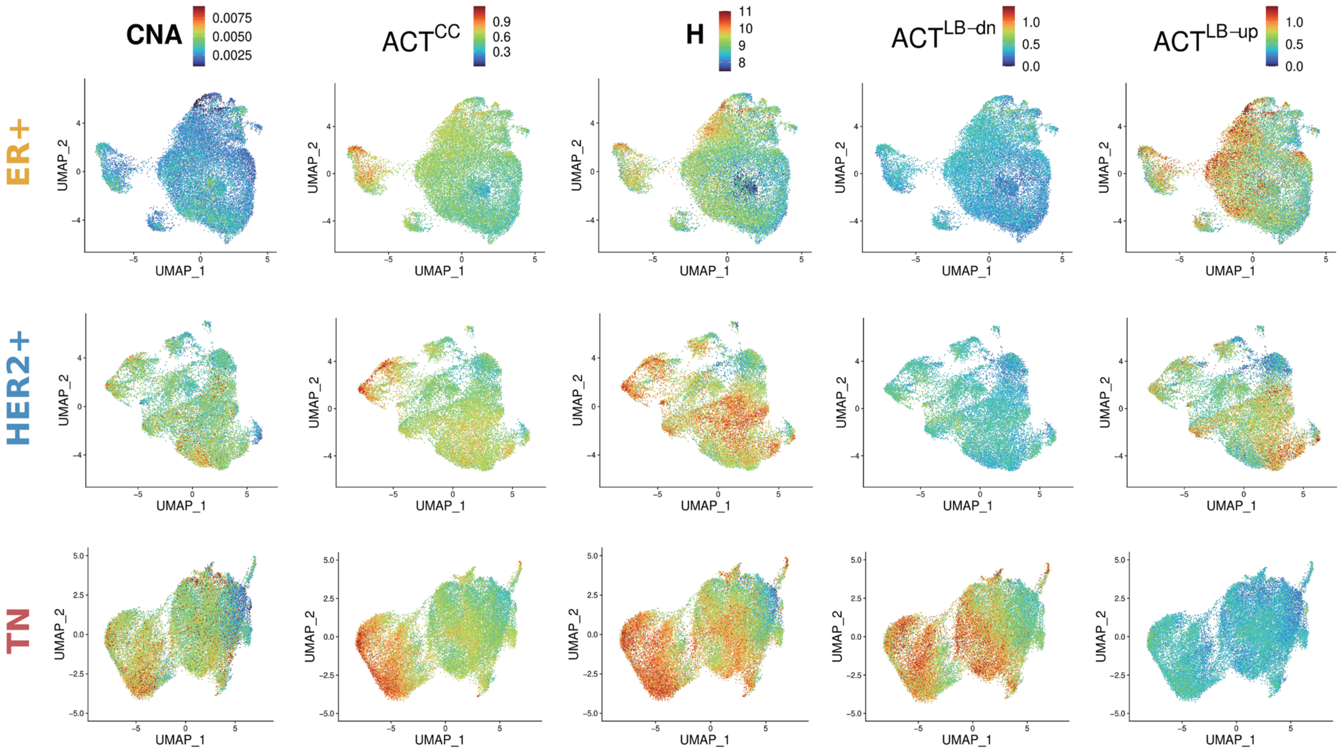

Fig. 4. Scores across cell populations of the three cancer subtypes. Visualization of the scores CNA, ACT^{CC}, H, ACT$^{LB\text{-}dn}$, and ACT$^{LB\text{-}up}$ in the UMAP space. Each point represents a cell, with color indicating the magnitude of the score in that cell. Samples were separated by cancer subtype: ER+ (top), HER2+ (middle), and TN (bottom) and then integrated for better visualization. Plots corresponding to the same score are organized in columns. ACT, activity; CNA, copy number alteration; dn, downregulated; ER, estrogen receptor; H, entropy; HER2, human epidermal growth factor receptor 2; LB, Luminal vs. basal; LM, Luminal vs. mesenchymal; TN, triple-negative; UAMP, Uniform Manifold Approximation and Projection for Dimension Reduction; up, upregulated.

mately 60% of ER+ patients,[53] while amplifications at chr8q and chr19 confirm their well-documented role in breast cancer.[54–56] Despite identifying these recurrent CNAs across subtypes, significant heterogeneity within each subtype was observed,[54] reflecting the complex and diverse nature of these malignancies. Furthermore, the activity profiles associated with basal and luminal cell

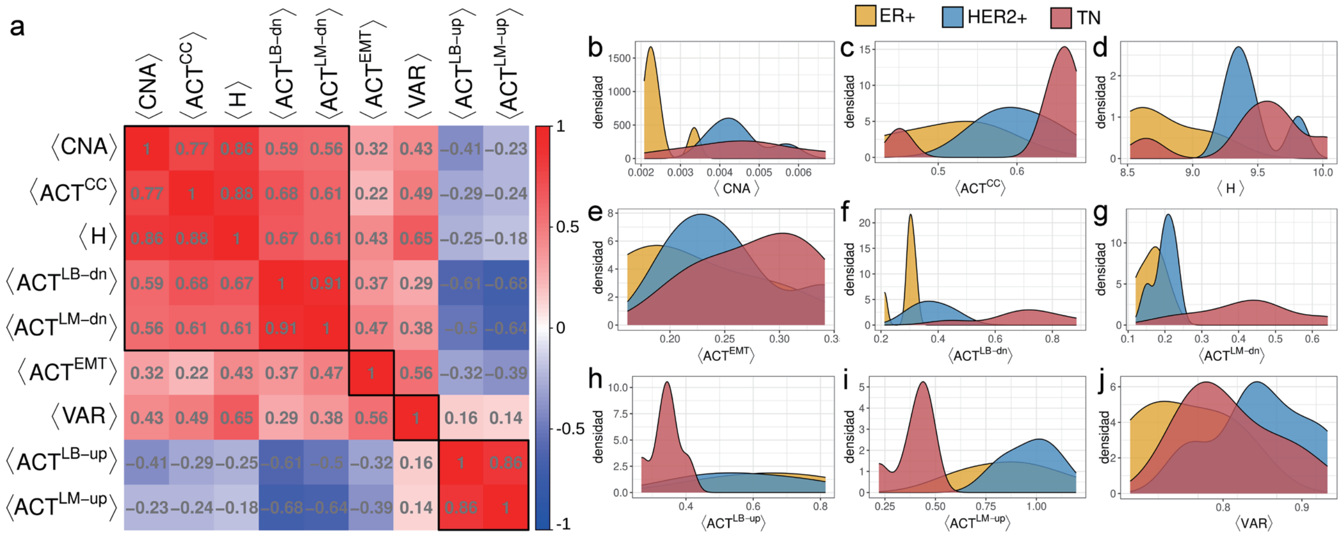

Fig. 5. Scores relationships. (a) Pearson correlation coefficient between the nine sample scores. Red indicates a positive correlation, blue indicates a negative correlation, and white indicates no correlation between the scores. The black squares identify clusters of five highly correlated scores. (b-j) Distribution estimation of sample scores, color-coded according to cancer subtypes: ER+, HER2+, and TN. ACT, activity; CNA, copy number alteration; ER, estrogen receptor; H, entropy; HER2, human epidermal growth factor receptor 2; LB, Luminal vs. basal; LM, Luminal vs. mesenchymal; TN, triple-negative; VAR, transcriptomic variability.

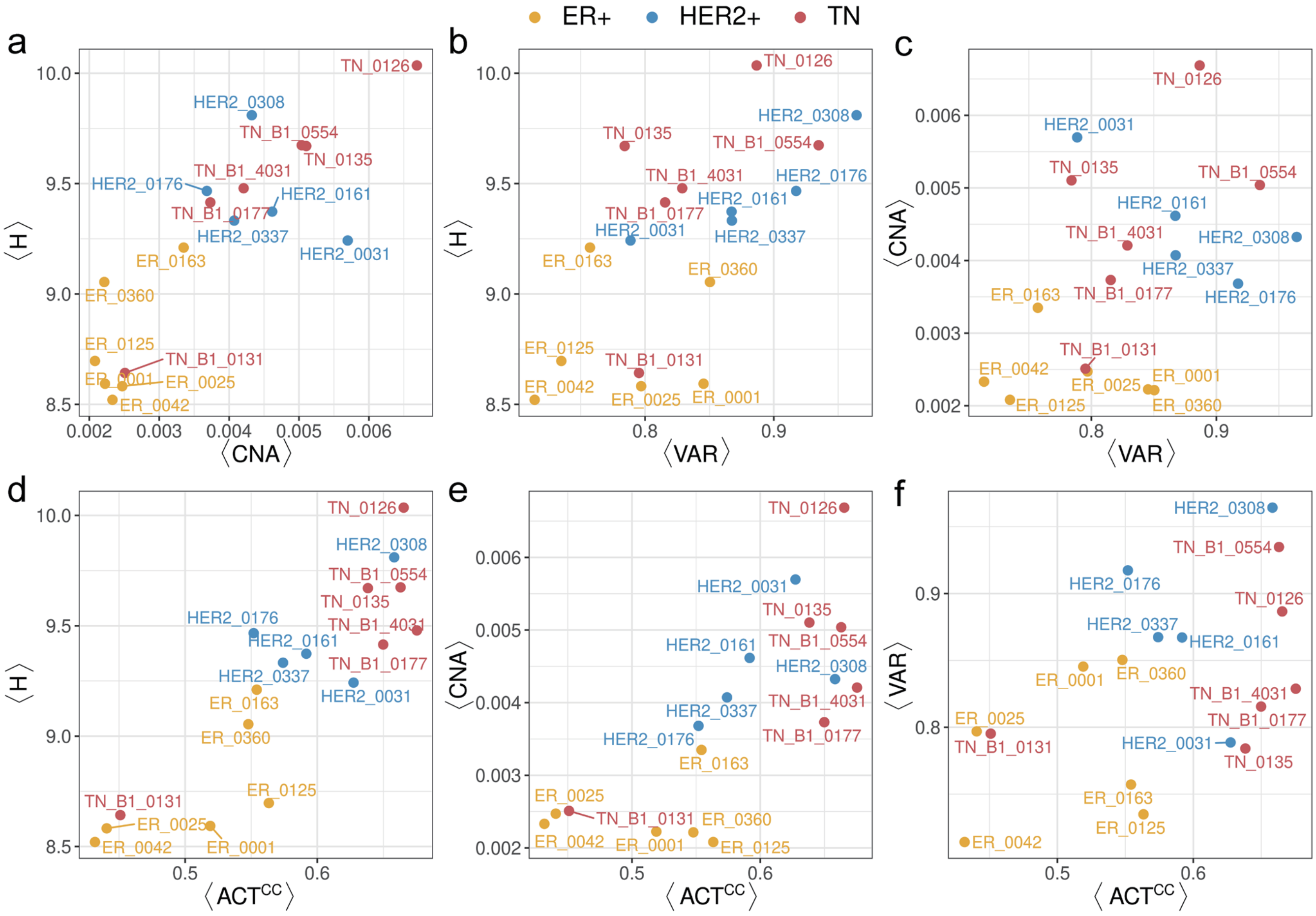

Fig. 6. Sample scores scatter plots. Scatter plots illustrating several relationships among sample scores that support the positive correlation. Each data point represents a sample, color-coded based on the cancer subtype, with labels included alongside the data points. ACT, activity; CNA, copy number alteration; ER, estrogen receptor; H, entropy; HER2, human epidermal growth factor receptor 2; TN, triple-negative; VAR, transcriptomic variability.

lines in our study differentiate basal and luminal tumors.

We did not observe cell groups with elevated ACT^{EMT} in any subtype, in agreement with prior studies.[13] This may be due to the reported scarcity of cells undergoing EMT, which could be masked within the large pool of cells analyzed.[57] Additionally, low expression levels of EMT-related genes (e.g., ZEB1, ZEB2, SNAIL) may suffice to trigger EMT, even without a substantial increase in $\mathrm{ACT}^{\mathrm{EMT}}$. Another factor could be the absence of metastasis in the studied samples, which would result in minimal EMT activity.

In analyzing the mean scores of the samples, we identified a positive correlation among $\langle CNA\rangle$, $\langle ACT\rangle^{CC}$, $\langle H\rangle$, $\langle ACT\rangle^{LB\text{-}dn}$, and $\langle ACT\rangle^{LM\text{-}dn}$ indicating that samples exhibiting basal characteristics present higher levels of these parameters. Moreover, these parameters show increasing levels in ER+, HER2+, and TN tumors (in ascending order), which aligns with the malignancy levels across cancer subtypes. Higher scores correspond to a more unfavorable prognosis. While various classifications of breast cancer subtypes exist, there is a consensus in this study that the order from better to worse prognosis is ER+, HER2+, and TN.[51,58–60] While $\langle ACT\rangle^{EMT}$ correlated positively with the previous scores, its correlation coefficient was lower than among them. An interesting exception emerged regarding the correlation with $\langle VAR\rangle$, a parameter quantifying transcriptomic heterogeneity within each sample, where HER2+ tumors showed the higher values, followed by TN and ER+ tumors. This is likely due to the fact that HER2+ tumors often exhibit both luminal and basal properties, resulting in a more diverse range of transcriptomic profiles. In various cancer types, including breast cancer, a distinct subset of cells known as CSCs has been identified. These cells comprise only a small fraction (approximately 0.1–1%) of the total tumor cell population and are associated with poor patient prognosis.[47,61,62] However, these cells may not significantly impact the transcriptomic variability score due to their low numerical abundance.

Our quantitative measures also uncovered distinct behavior in samples HER2-0308 and TN-B1-0131 compared to others within their respective subtypes. This difference may be attributed to patient age discrepancies relative to the age range of the other samples in each subtype. Numerous studies have reported more aggressive tumor biology, increased recurrence risk, treatment failure rates, and higher mortality in younger patients.[52,63–65]

This study has several limitations that could impact the accuracy of our conclusions. First, the relatively small sample size in the available database may limit the generalizability of our findings, especially given the high variability among samples. Although scRNA-seq is a powerful tool for studying tissue heterogeneity, larger breast cancer datasets will become available as this technology advances, enabling more comprehensive analyses. Second, PPIN activity assessment relies on existing interaction databases, which may be incomplete and fail to capture all relevant tumor biology interactions. This limitation could be addressed in the future by employing functional assays to validate the roles of specific genes in the biological processes of interest. Finally, the TME exerts selective pressures (e.g., hypoxia, nutrient deprivation, immune surveillance) that drive the evolution of tumor cell subpopulations.[66] As the TME is spatially heterogeneous, these selective pressures create distinct niches that can select tumor cells with specific adaptations, leading to regional heterogeneity.[67] Additionally, components of the TME, such as inflammatory cytokines and reactive oxygen species, can promote genomic instability in tumor cells, increasing mutation rates and contributing to the generation of diverse subclones.[68] Understanding the complex interplay between tumor cells and the TME is essential for a comprehensive view of tumor cell heterogeneity. A survey of the microenvironment (stromal/immune cells) in different subtypes from the same dataset was conducted in a previous study.[13] However, our focus is on cancer cells, and a detailed analysis of TME interactions lies beyond the scope of this study.

Conclusions

This study addresses a gap in the current understanding of breast cancer heterogeneity by presenting a novel quantitative approach that offers deeper insight into tumor biology, overcoming some limitations of traditional marker-based methods. Using single-cell RNA sequencing data, this work introduces a novel scoring framework that quantifies key cancer traits, such as CNAs, transcriptomic heterogeneity, entropy and activities of PPIN associated with biological processes relevant to cancer biology. The proposed methodology allows exploring these features at the individual cell level, revealing intra- and inter- tumor heterogeneity that may be relevant for tumor evolution and treatment response. We applied this methodology to human scRNA-seq datasets from ER+, HER2+, TN breast cancer subtypes. Our analysis revealed significant differences in several scores across the subtypes. Overall, this approach enables a better understanding of breast cancer heterogeneity, with the potential to identify novel therapeutic targets and strategies.

Acknowledgments

We thank Martín Tomé for his technical support and valuable insights. His expertise and assistance were essential in executing this project.

Funding

This work was supported by FONCyT, Agencia Nacional de Promoción Científica y Tecnológica, Argentina [PICT 2018-03713], and CONICET, Argentina [PIP 2021-1748].

Conflict of interest

The authors declare no conflicts of interest related to this publication.

Author contributions

Study concept and design (DS, NG, LD), analysis and interpretation of data (DS, NG, LD), drafting of the manuscript (DS), critical revision of the manuscript for important intellectual content (NG, LD), administrative, technical, or material support (NG, LD), and study supervision (NG, LD). All authors have made significant contributions to this study and have approved the final manuscript.

Ethical statement

The data used in this study were obtained from public databases and originated from previously published studies. These original studies were approved by the appropriate ethics committees and complied with all applicable regulations. As such, no additional ethical approval was required for this secondary analysis.

Data sharing statement

The dataset utilized in this study is publicly available in two forms: as raw count matrices in the GEO database (series GSE161529) and as preprocessed R objects on Figshare (DOI: 10.6084/m9.figshare.17058077). Additionally, all data generated during this study are included in the published article and listed in the methods section. All relevant codes for calculating the scores are also publicly available.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71(3):209–249. doi:10.3322/caac.21660, PMID:33538338.

[2] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, *et al*. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast 2022;66:15–23. doi:10.1016/j.breast.2022.08.010, PMID:36084384.

[3] Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA 2019;321(3):288–300. doi:10.1001/jama.2018.19323, PMID:30667505.

[4] Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of Breast Cancer. In: Mayrovitz HN (ed). Breast Cancer. Brisbane (AU): Exon Publications; 2022:31–42. doi:10.36255/exon-publications-breast-cancer-subtypes, PMID:36122153.

[5] Eliyatkın N, Yalçın E, Zengel B, Aktaş S, Vardar E. Molecular Classification of Breast Carcinoma: From Traditional, Old-Fashioned Way to A New Age, and A New Way. J Breast Health 2015;11(2):59–66. doi:10.5152/tjbh.2015.1669, PMID:28331693.

[6] Yin L, Duan JJ, Bian XW, Yu SC. Triple-negative breast cancer molecular subtyping and treatment progress. Breast Cancer Res 2020;22(1):61. doi:10.1186/s13058-020-01296-5, PMID:32517735.

[7] Alluri P, Newman LA. Basal-like and triple-negative breast cancers: searching for positives among many negatives. Surg Oncol Clin N Am 2014;23(3):567–577. doi:10.1016/j.soc.2014.03.003, PMID:24882351.

[8] Stelmach A, Patla A, Skotnicki P, Sas-Korczyńska B. Typical medullary breast carcinoma: Clinical outcomes and treatment results. Breast J 2017;23(6):770–771. doi:10.1111/tbj.12815, PMID:28421688.

[9] Verras GI, Mulita F, Tchabashvili L, Grypari IM, Sourouni S, Panagodimou E, *et al*. A rare case of invasive micropapillary carcinoma of the breast. Prz Menopauzalny 2022;21(1):73–80. doi:10.5114/pm.2022.113834, PMID:35388282.

[10] Liu J, Dang H, Wang XW. The significance of intertumor and intratumor heterogeneity in liver cancer. Exp Mol Med 2018;50(1):e416. doi:10.1038/emm.2017.165, PMID:29303512.

[11] Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, *et al*. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet 2021;53(9):1334–1347. doi:10.1038/s41588-021-00911-

1, PMID:34493872.

[12] Xu J, Qin S, Yi Y, Gao H, Liu X, Ma F, *et al*. Delving into the Heterogeneity of Different Breast Cancer Subtypes and the Prognostic Models Utilizing scRNA-Seq and Bulk RNA-Seq. Int J Mol Sci 2022;23(17):9936. doi:10.3390/ijms23179936, PMID:36077333.

[13] Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, *et al*. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. EMBO J 2021;40(11):e107333. doi:10.15252/embj.2020107333, PMID:33950524.

[14] Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, *et al*. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 2018;9(1):3588. doi:10.1038/s41467-018-06052-0, PMID:30181541.

[15] Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, *et al*. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017;8:15081. doi:10.1038/ncomms15081, PMID:28474673.

[16] Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, *et al*. Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. Cancer Cell 2019;36(4):418–430.e6. doi:10.1016/j.ccell.2019.08.007, PMID:31588021.

[17] Chen Y, Smyth G. Data, R code and output Seurat Objects for single cell RNA-seq analysis of human breast tissues. figshare. Dataset-doi:10.6084/m9.figshare.17058077.v1.

[18] Chen Y, Pal B, Lindeman GJ, Visvader JE, Smyth GK. R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue. Sci Data 2022;9(1):96. doi:10.1038/s41597-022-01236-2, PMID:35322042.

[19] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344(6190):1396–1401. doi:10.1126/science.1254257, PMID:24925914.

[20] Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352(6282):189–196. doi:10.1126/science.aad0501, PMID:27124452.

[21] Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, *et al*. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science 2017;355(6332):eaai8478. doi:10.1126/science.aai8478, PMID:28360267.

[22] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, *et al*. Comprehensive Integration of Single-Cell Data. Cell 2019;177(7):1888–1902.e21. doi:10.1016/j.cell.2019.05.031, PMID:31178118.

[23] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, *et al*. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–3587.e29. doi:10.1016/j.cell.2021.04.048, PMID:34062119.

[24] Harbers L, Agostini F, Nicos M, Poddighe D, Bienko M, Crosetto N. Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. Front Oncol 2021;11:700568. doi:10.3389/fonc.2021.700568, PMID:34395272.

[25] Zhang Y, Martens JW, Yu JX, Jiang J, Sieuwerts AM, Smid M, *et al*. Copy number alterations that predict metastatic capability of human breast cancer. Cancer Res 2009;69(9):3795–3801. doi:10.1158/0008-5472.CAN-08-4596, PMID:19336569.

[26] Han W, Jung EM, Cho J, Lee JW, Hwang KT, Yang SJ, *et al*. DNA copy number alterations and expression of relevant genes in triple-negative breast cancer. Genes Chromosomes Cancer 2008;47(6):490–499. doi:10.1002/gcc.20550, PMID:18314908.

[27] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, *et al*. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 2021;39(5):599–608. doi:10.1038/s41587-020-00795-2, PMID:33462507.

[28] Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun 2020;11(1):89. doi:10.1038/s41467-019-13779-x, PMID:31900397.

[29] Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, *et al*. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res 2019;29(9):725–738. doi:10.1038/s41422-019-0195-y, PMID:31273297.

[30] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, *et al*. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell 2019;178(4):835–849.e21. doi:10.1016/j.cell.2019.06.024, PMID:31327527.

[31] Choi YH, Kim JK. Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. Mol Cells 2019;42(3):189–199. doi:10.14348/molcells.2019.2446, PMID:30764602.

[32] Duan X, Wang W, Tang M, Gao F, Lin X. Dissecting Cellular Heterogeneity Based on Network Denoising of scRNA-seq Using Local Scaling Self-Diffusion. Front Genet 2021;12:811043. doi:10.3389/fgene.2021.811043, PMID:35082838.

[33] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 2019;20(1):296. doi:10.1186/s13059-019-1874-1, PMID:31870423.

[34] Mayer C, Hafemeister C, Bandler RC, Machold R, Batista Brito R, Jaglin X, *et al*. Developmental diversification of cortical inhibitory interneurons. Nature 2018;555(7697):457–462. doi:10.1038/nature25999, PMID:29513653.

[35] Senra D, Guisoni N, Diambra L. ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data. MethodsX 2022;9:101778. doi:10.1016/j.mex.2022.101778, PMID:35855951.

[36] Senra D, Guisoni N, Diambra L. Cell annotation using scRNA-seq data: A protein-protein interaction network approach. MethodsX 2023;10:102179. doi:10.1016/j.mex.2023.102179, PMID:37128282.

[37] Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 2009;25(22):3045–3046. doi:10.1093/bioinformatics/btp536, PMID:19744993.

[38] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102, PMID:16199517.

[39] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27(12):1739–1740. doi:10.1093/bioinformatics/btr260, PMID:21546393.

[40] Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adélaïde J, Cervera N, *et al*. Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene 2006;25(15):2273–2284. doi:10.1038/sj.onc.1209254, PMID:16288205.

[41] Gandrillon O, Gaillard M, Espinasse T, Garnier NB, Dussiau C, Kosmider O, *et al*. Entropy as a measure of variability and stemness in single-cell transcriptomics. Curr Opin Syst Biol 2021;7:100348. doi:10.1016/j.coisb.2021.05.009.

[42] Kannan S, Farid M, Lin BL, Miyamoto M, Kwon C. Transcriptomic entropy benchmarks stem cell-derived cardiomyocyte maturation against endogenous tissue at single cell level. PLoS Comput Biol 2021;17(9):e1009305. doi:10.1371/journal.pcbi.1009305, PMID:34534204.

[43] MacArthur BD, Lemischka IR. Statistical mechanics of pluripotency. Cell 2013;154(3):484–489. doi:10.1016/j.cell.2013.07.024, PMID:23911316.

[44] Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun 2017;8:15599. doi:10.1038/ncomms15599, PMID:28569836.

[45] Zeng X, Liu C, Yao J, Wan H, Wan G, Li Y, *et al*. Breast cancer stem cells, heterogeneity, targeting therapies and therapeutic implications. Pharmacol Res 2021;163:105320. doi:10.1016/j.phrs.2020.105320, PMID:33271295.

[46] Wanandi SI, Syahrani RA, Arumsari S, Wideani G, Hardiany NS. Profiling of Gene Expression Associated with Stemness and Aggressiveness of ALDH1A1-Expressing Human Breast Cancer Cells. Malays J Med Sci 2019;26(5):38–52. doi:10.21315/mjms2019.26.5.4, PMID:31728117.

[47] Luo M, Clouthier SG, Deol Y, Liu S, Nagrath S, Azizi E, *et al*. Breast cancer stem cells: current advances and clinical implications. Methods Mol Biol 2015;1293:1–49. doi:10.1007/978-1-4939-2519-3_1, PMID:26040679.

[48] Fultang N, Chakraborty M, Peethambaran B. Regulation of cancer stem cells in triple negative breast cancer. Cancer Drug Resist 2021;4(2):321–342. doi:10.20517/cdr.2020.106, PMID:35582030.

[49] Krzywinski M, Altman N. Comparing samples-part II. Nat Methods 2014;11(4):355–356. doi:10.1038/nmeth.2900.

[50] Rodrigues-Peres RM, de S Carvalho B, Anurag M, Lei JT, Conz L, Gonçalves R, *et al*. Copy number alterations associated with clinical features in an underrepresented population with breast cancer. Mol Genet Genomic Med 2019;7(7):e00750. doi:10.1002/mgg3.750, PMID:31099189.

[51] Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, *et al*. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res 2015;5(10):2929–2943. PMID:26693050.

[52] Klauber-DeMore N. Tumor biology of breast cancer in young women. Breast Dis 2005-2 2006;23:9–15. doi:10.3233/bd-2006-23103, PMID:16823162.

[53] Orsetti B, Nugoli M, Cervera N, Lasorsa L, Chuchana P, Rougé C, *et al*. Genetic profiling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identification of candidate genes on 1q. Br J Cancer 2006;95(10):1439–1447. doi:10.1038/sj.bjc.6603433, PMID:17060936.

[54] Baslan T, Kendall J, Volyanskyy K, McNamara K, Cox H, D'Italia S, *et al*. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. Elife 2020;9:e51480. doi:10.7554/eLife.51480, PMID:32401198.

[55] Zhang L, Feizi N, Chi C, Hu P. Association Analysis of Somatic Copy Number Alteration Burden With Breast Cancer Survival. Front Genet 2018;9:421. doi:10.3389/fgene.2018.00421, PMID:30337938.

[56] Ibragimova MK, Tsyganov MM, Pevzner AM, Litviakov NV. Transcriptome of Breast Tumors With Different Amplification Status of the Long Arm of Chromosome 8. Anticancer Res 2021;41(1):187–195. doi:10.21873/anticanres.14764, PMID:33419812.

[57] Wu Y, Sarkissyan M, Vadgama JV. Epithelial-Mesenchymal Transition and Breast Cancer. J Clin Med 2016;5(2):13. doi:10.3390/jcm5020013, PMID:26821054.

[58] Dai X, Cheng H, Bai Z, Li J. Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. J Cancer 2017;8(16):3131–3141. doi:10.7150/jca.18457, PMID:29158785.

[59] Herdiana Y, Wathoni N, Shamsuddin S, Muchtaridi M. α-Mangostin Nanoparticles Cytotoxicity and Cell Death Modalities in Breast Cancer Cell Lines. Molecules 2021;26(17):5119. doi:10.3390/molecules26175119, PMID:34500560.

[60] Schwarzenbach H, Gahan PB. Predictive value of exosomes and their cargo in drug response/resistance of breast cancer patients. Cancer Drug Resist 2020;3(1):63–82. doi:10.20517/cdr.2019.90, PMID:35582044.

[61] Crabtree JS, Miele L. Breast Cancer Stem Cells. Biomedicines 2018;6(3):77. doi:10.3390/biomedicines6030077, PMID:30018256.

[62] Zhang X, Powell K, Li L. Breast Cancer Stem Cells: Biomarkers, Identification and Isolation Methods, Regulating Mechanisms, Cellular Origin, and Beyond. Cancers (Basel) 2020;12(12):3765. doi:10.3390/cancers12123765, PMID:33327542.

[63] Narod SA. Breast cancer in young women. Nat Rev Clin Oncol 2012;9(8):460–470. doi:10.1038/nrclinonc.2012.102, PMID:22733233.

[64] Bharat A, Aft RL, Gao F, Margenthaler JA. Patient and tumor characteristics associated with increased mortality in young women (< or =40 years) with breast cancer. J Surg Oncol 2009;100(3):248–251. doi:10.1002/jso.21268, PMID:19330813.

[65] Voogd AC, Nielsen M, Peterse JL, Blichert-Toft M, Bartelink H, Overgaard M, *et al*. Breast Cancer Cooperative Group of the European Organization for Research and Treatment of Cancer. Differences in risk factors for local and distant recurrence after breast-conserving therapy or mastectomy for stage I and II breast cancer: pooled results of two large European randomized trials. J Clin Oncol 2001;19(6):1688–1697. doi:10.1200/JCO.2001.19.6.1688, PMID:11250998.

[66] Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer 2012;12(5):323–334. doi:10.1038/nrc3261, PMID:22513401.

[67] Emami Nejad A, Najafgholian S, Rostami A, Sistani A, Shojaeifar S, Esparvarinha M, *et al*. The role of hypoxia in the tumor microenvironment and development of cancer stem cell: a novel approach to developing treatment. Cancer Cell Int 2021;21(1):62. doi:10.1186/s12935-020-01719-5, PMID:33472628.

[68] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646–674. doi:10.1016/j.cell.2011.02.013, PMID:213 76230.