The Universal Breast cancer Subtyping 93 finds that claudin-low breast cancer may originate from basal breast cancer.

Abstract

Background

Breast cancer is a complex disease with diverse molecular characteristics, significantly impacting patient prognosis, outcomes, and treatment decisions. Previous studies have introduced PAM50 classifiers and claudin-low classifiers based on bulk RNA-seq samples. However, single-cell analysis has revealed the existence of distinct subtypes within the same tumor, indicating that classifiers relying on gene signatures derived from bulk samples may not accurately capture the true molecular features of breast cancer.

Method

To address this limitation, we utilized single-cell data from breast cancer patients to define the E-M ratio parameter. We identified 93 epithelial-specific genes and developed a Universal Breast cancer Subtyping 93 (UBS93). To validate the efficacy of UBS93, we conducted separate analyses using bulk RNA-seq and single-cell RNA-seq datasets of human breast cancer cell lines, as well as bulk RNA-seq data from mice. Additionally, we compared the performance of UBS93 with that of the genefu package to highlight its advantages.

Results

UBS93 demonstrated excellent performance in human and mouse datasets, including bulk RNA-seq and single-cell RNA-seq data. It exhibited higher epithelial

specificity and accuracy compared to PAM50 genes. When predicting bulk RNA-seq data from breast cancer cell lines and mouse models, UBS93 outperformed the genefu package. Single-cell validation revealed the coexistence of basal and claudin-low subtypes in the HDQP1 cell line and two TNBC patients, suggesting a shared origin. Differential gene expression analysis identified ELF3 loss as a potential driver for basal-to-claudin-low differentiation. Experimental validation confirmed that the downregulation of ELF3 resulted in the downregulation of CLDN3, CLDN4, and CLDN7, facilitating the transition from basal to claudin-low cells.

Conclusion

Our study constructed a comprehensive breast cancer classification, UBS93, based on 93 epithelial-specific genes identified using single-cell data. By applying UBS93, we unveiled the coexistence of basal and claudin-low subtypes and illuminated the molecular mechanism underlying basal-to-claudin-low differentiation, with ELF3 loss playing a significant role in this process.

Background

Breast cancer is a heterogeneous disease in terms of molecular alterations, cellular composition, and clinical outcomes. However, this heterogeneity poses challenges regarding clinically relevant tumor classification for prognosis and prediction [36931265] [1]. Fortunately, researchers have utilized microarray technology to develop an intrinsic breast cancer classifier called PAM50, which categorizes cancer into five subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like [19204204] [2]. This classification system significantly enhances the prognostic

and predictive value over traditional approaches, including pathological staging, histological grading, and standard clinical biomarkers.

Through in-depth investigation of gene expression profiles in breast cancer, researchers have discovered a novel subtype characterized by low expression levels of cell adhesion components such as CLDN3, CLDN4, CLDN7, and CDH1, which is associated with mesenchymal features. This new subtype exhibits increased proliferative capacity and poorer prognosis [17493263] [3]. In response, Alexi et al. developed a classification method called the nine-cell line claudin-low predictor, categorizing breast cancer into two groups: Claudin-low and Others [20813035] [4]. However, the origin of this new subtype has been a subject of ongoing debate. Some researchers suggest that the occurrence and progression of triple-negative breast cancer from luminal epithelium are driven by carcinogenic RAS signal transduction., although this conclusion has yet to be validated in human data [34145248] [5]. Another study, based on genetic, epigenetic, and gene expression analyses, found that claudin-low breast cancer originates from three subgroups, with two subgroups associated with luminal and basal-like subtypes, and the third subgroup closely related to normal human breast stem cells [32647202] [6]. In conclusion, the origin of claudin-low breast cancer is a complex process that requires further research to explore the underlying mechanisms and contributing factors.

Single-cell transcriptomic analysis has provided us with deeper insights into the heterogeneity among different subtypes of breast cancer [35352511] [7]. By analyzing the intrinsic subtypes within individual malignant cells, it has been discovered that there

exist cells of different subtypes within a single tumor. This indicates that the tumor subtypes defined by gene signatures obtained through bulk RNA sequencing may not always accurately reflect the true molecular phenotype of the tumor [36931265][1]. To gain a more comprehensive understanding of tumor cells, single-cell analysis is necessary. Additionally, there is a lack of standardized prediction for breast cancer subtypes. To address these issues, we have developed a novel epithelial cell-specific prediction factor called Universal Breast Cancer Subtyping 93 (UBS93). UBS93 classifies breast cancer into four subtypes: Basal, Claudin-low, HER2-amp, and Luminal. UBS93 focuses on epithelial-specific markers and aims to provide a more consistent and comprehensive approach for breast cancer subtype classification. This advancement holds the potential to improve the accuracy of classification, facilitate personalized treatment strategies, and guide clinical decision-making.

Results

Development of UBS93 gene panel

Breast cancer is a solid tissue composed of multiple cell types, including cancer cells and tumor microenvironment cells. The tumor microenvironment cells encompass endothelial cells, cancer-associated fibroblasts, T cells, B cells, macrophages, and others[34493872] [8]. When examining the expression values of genes in transcriptome sequencing of this tissue, the value for each gene is a combination of its expression in cancer cells and tumor microenvironment cells. We can denote the expression value of a gene in cancer cells as E and in tumor microenvironment cells as M. The tumor purity of a bulk sample, represented by α, indicates the proportion of cancer cells in the sample.

Thus, the expression value of a gene in a bulk tumor sample can be calculated as αE + (1-α) M. In the context of breast cancer subtyping, the previously widely used method called PAM50 utilized the Spearman correlation technique, which proved to be effective[19204204][2]. Consequently, we aim to continue using this method and prioritize the rank values of gene expression. Genes with epithelial specificity typically exhibit negligible expression in tumor microenvironment cells, making their expression in a bulk sample approximately equal to the product of the sample's tumor purity and the gene's expression value in cancer cells. Consequently, when determining the rank value of a gene with epithelial specificity, it tends to align with the rank value observed in pure tumors. However, for randomly selected genes, their rank values do not tend to align with those observed in pure tumors (Figure 1a). Therefore, our objective is to develop a gene panel consisting of genes with epithelial specificity to guide the molecular subtyping of breast cancer.

To identify genes with epithelial specificity, we utilized published single-cell data from human breast cancer patients and defined a parameter called "E-M-ratio". [34493872][8]. Based on this, we select the top 100 genes with the highest E-M-ratio (refer to the method section for detailed information). To confirm the epithelial specificity of the top 100 gene panel, we examine the average expression values of these genes across different cell types in four patients (Figure 1b, Figure S2). Whether in non-triple negative breast cancer patients (CID3963, CID4066, CID45171) or triple-negative breast cancer(TNBC) patients, we consistently observe significantly higher average expression values of the top 100 gene panel in cancer cells compared to other

tumor microenvironment cells.

To determine the subtypes of breast cancer, we employ Uniform Manifold Approximation and Projection (UMAP) (Figure 1c) on the transcriptome data of breast cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE). We use the lineage molecular subtype labels of the cell lines as a reference for our subtype classification. We observe that the cell lines predominantly formed three clusters, and within the "basal_A" cluster, there were numerous well-established basal cell lines[17157791] [9]. Therefore, we conclude that this cluster should be classified as the basal subtype. Many of the cell lines in the "basal_B" cluster overlap with the claudin-low cell lines described in published articles, suggesting that this cluster should be classified as the claudin-low subtype[20813035][4]. Although the luminal and HER2-amp clusters form a single cluster, there were distinct boundaries between them, indicating the need to separate them. For the two cell lines labeled as luminal-HER2-amp, they do not cluster together as a distinct group, suggesting that the annotation of this subtype may not be accurate, and we do not intend to use it in our classification. Therefore, we determine that breast cancer can be divided into four subtypes based on their molecular characteristics: Basal, Claudin-low, Luminal, and HER2-amp. In the CCLE breast cancer transcriptome data, genes are represented in the form of ensemble IDs. Out of the TOP100 genes, 93 genes have ensemble IDs. We further use the data from these 93 genes to perform T-distributed Stochastic Neighbor Embedding (t-SNE) analysis on the transcriptome data of breast cancer cell lines in CCLE (Figure S3). Remarkably, these 93 genes achieve satisfactory classification results, enabling effective subtype

classification.

Finally, we develop a molecular classifier called Universal Breast cancer Subtyping 93 (UBS93) based on these 93 genes. The schematic process is illustrated in Figure 1d. When a vector containing all gene expression values of a sample is inputted into UBS93, it intersects the genes present in the sample with the genes in the subtype centroid matrix (refer to the method for detailed information). Then, UBS93 calculates the correlation coefficients between the sample vector and the subtype centroid matrix using Spearman correlation. The sample is assigned to the subtype with the highest correlation coefficient. However, due to the clustering of luminal and HER2-amp subtypes in close proximity, the correlation coefficient calculated using the centroid is insufficient to separate them. Therefore, we employ the ssGSEA method to distinguish these two subtypes. If the sample's subtype is not HL, it is classified as either Basal or Claudin-low. In the case of a sample with the HL subtype, we calculate the ssGSEA score of HER2-amp signature genes (refer to the method for detailed information). If the ssGSEA score exceeds the specified cutoff value (refer to the method for detailed information), the sample is classified as HER2-amp subtype. If the ssGSEA score is below the cutoff value, the sample is classified as a Luminal subtype.

Validation of UBS93

To evaluate the effectiveness of UBS93, we utilize it to predict the CCLE data and bulk RNA-sequence dataset of three breast cancer cell lines from published studies[36388465, 24176112, 26771497] [10] [11] [12]. The prediction accuracies are 95.8%, 93.3%, 85.0%, and 84.9%, respectively. The subsequent cell line data were

predicted using UBS93 after excluding cell lines with molecular subtypes other than the four subtypes defined by us in CCLE. We group the data sets according to subtypes, calculate the accuracy, and find that most of the data sets achieved an accuracy rate higher than 70%. (Table 1). Additionally, we perform UBS93 predictions on TNBC samples from The Cancer Genome Atlas Program (TCGA). 87.8% of the samples are classified as the Basal subtype, which is consistent with the reported findings[21076464, 18398844] [13] [14].

To validate the effectiveness of UBS93 in analyzing single-cell transcriptome data, we utilize two published datasets from human breast cancer cell lines (GSE173634, GSE202771) for UBS93 prediction[35361816, 38100350] [15] [16]. In GSE173634, 21 out of 28 cell lines exhibit a proportion of cells correctly predicted subtypes higher than 70% (Figure 2a). Similarly, in GSE202771, 11 out of 15 cell lines show a proportion of cells with accurate subtype prediction higher than 76% (Figure S5). We observe a luminal cell line, ZR751, in GSE173634, where only 21.1% of cells are predicted as Luminal subtype, while 78.9% are predicted as HER2-amp. In GSE202771, we also notice cases where a significant proportion of cells in two luminal cell lines are predicted as HER2-amp. As a result, we examine the distribution of ERBB2 expression values in breast cancer cell lines from the CCLE dataset (Figure S6). We find that the ERBB2 expression value of ZR751 is similar to that of other HER2-amp cell lines, and the ERBB2 expression value of EFM223 is close to the median of all breast cancers. These observations suggest that the molecular subtype labels in the CCLE dataset may be controversial and require further analysis and discussion. Interestingly, we observe

that only 56.4% of cells in the HDQP1 cell lines are predicted as Basal subtype. We conduct further analysis of this finding, and the specific results are presented in the fourth section of our study. Furthermore, we discover a nearly one-to-one mixture of Claudin-low and HER2-amp cells in both batches of data from the CAL51 cell line. However, their UAMP plot chart does not reveal any evident division into two distinct clusters, so we do not pursue further analysis in this regard.

To verify the effectiveness of UBS93 in mice, we utilize bulk RNA-sequence data from 10 published mouse models with KRAS mutations, including 5 Basal and 5 Claudin-low subtypes[34145248][5]. Given the unclear labeling of Basal and Claudin-low in the article, we perform Hierarchical clustering on the 10 mouse datasets using five genes, including Cldn3, Cldn13, Cldn7, Vim, and Epcam, to distinguish between Basal and Claudin-low subtypes (Figure 2b). Additionally, we include the other 17 Claudin-low reference models mentioned in the article for clustering[31730857] [17]. Subsequently, we apply UBS93 to predict the subtypes of 27 mouse samples. Remarkably, the predicted subtypes align perfectly with the true subtypes(Supplementary Table 1). The aforementioned findings demonstrate the strong performance of UBS93 across human and mouse datasets, as well as in analyzing both bulk RNA-sequence and single-cell sequence data.

Comparison to genefu

To demonstrate the advantages of UBS93, we compare it with the current widely used method for molecular subtyping of breast cancer. Typically, the molecular typing method involves classifying breast cancer as either the "Claudin-low" or "Others"

subtype using a nine-cell line Claudin-low predictor [20813035][4]. The "Others" samples are further classified into Basal-like, Normal-like, Luminal A, Luminal B, and HER2-enriched subtypes using the Prediction Analysis of Microarray 50 (PAM50) [19204204][2]. Both of these methods are included in the genefu package, so we will compare the results of UBS93 with those of the genefu package [26607490] [18]. To begin, we identify the intersection between the UBS93 gene set and the PAM50 gene set and find that there are six genes common to both sets (Figure S7). To determine whether UBS93 genes are more epithelial-specific than PAM50 genes, we compare the gene scores of the two panels and their average expression values in different cell types of four patients[34493872][8]. We observe that the E-M-ratio values of UBS93 genes are significantly higher than those of PAM50 genes. Additionally, the average expression values of UBS93 genes in epithelial cells are significantly higher than those of PAM50 genes, as observed in both non-TNBC and TNBC (Figure 3a and Figure S8). Notably, in patient CID3963, we observe that the expression level of PAM50 genes in cancer-associated fibroblasts is equal to that in epithelial cells, while the expression level of UBS93 genes in cancer-associated fibroblasts is significantly higher than that in epithelial cells.

Next, we compare the prediction accuracy of UBS93 and the genefu package using bulk RNA-sequence data from human breast cancer cell lines and mouse models. In the case of human breast cancer cell lines, UBS93 demonstrates higher overall prediction accuracy compared to genefu (Figure 3b), as well as improved prediction accuracy for different subtypes (Figure S9). Moving on to the mouse data, the genefu package

incorrectly identifies all claudin-low samples as "Others," whereas UBS93 accurately predicts them as "Claudin-low" subtypes (Table 2).

Finally, due to the sparsity of single-cell data[23267174] [19], there may be situations where calculating the correlation coefficients between cells and subtypes is not possible. Considering this limitation, we compare UBS93 and PAM50 and find that UBS93 has a significantly higher proportion of cells where the correlation coefficients can be calculated compared to PAM50. These results consistently indicate that UBS93 has distinct advantages over PAM50 in subtype prediction.

UBS93 enables the discovery of TNBC cell lines and patients of hybrid subtype

In the second section of our analysis, We observe the presence of a cell line called HDQP1 in the GSE173634 dataset, which is a recognized basal cell line. However, we make an interesting observation: only 56.4% of cells in HDQP1 are predicted as basal, while the remaining 43.6% of cells are predicted as claudin-low. To further investigate this, we perform t-SNE visualization on the single-cell data and find that basal and claudin-low cells form two distinct clusters (Figure 4a). To gain more insights, we conduct hierarchical clustering using the expression levels of CLDN3, CLDN4, CLDN7, EPCAM, and VIM in the HDQP1 single-cell data. This analysis reveals that the cells segregate into two clusters, with claudin-low cells predominantly clustering together. In comparison to the basal cells in the other cluster, claudin-low cells exhibit lower expression of cell adhesion molecules such as CLDN3, CLDN4, and CLDN7. Additionally, they show lower expression of the breast epithelial cell marker EPCAM and higher expression of markers associated with epithelial-mesenchymal transition

processes, such as VIM (Figure S10). We also notice a similar phenomenon in the HDQP1 cell line from the GSE202771 dataset, although in this case, claudin-low cells account for only 3.4% of the total population (Figure S11, and Figure S12).

Does this phenomenon also occur in patients? We analyze published single-cell data from TNBC patients and discover the presence of basal and claudin-low cells (91.1% and 7.9%, respectively) within the malignant epithelial cells of a TNBC patient in the GSE161529 dataset. On the t-SNE plot, claudin-low cells cluster together in a distinct area (Figure 4b). When compared to basal cells, claudin-low cells also exhibit lower expression of CLDN3, CLDN4, CLDN7, and EPCAM, as well as higher expression of VIM (Figure S13, and Figure S14). We group the cancer cells of this patient based on subtypes and perform inferCNV analysis. Surprisingly, we find that the copy number variations (CNVs) in basal cells and claudin-low cells are highly similar (Figure S15). We calculate the means of chromosomal CNVs in cells according to their cell types and discover that the CNVs in basal and claudin-low cells are similar, causing these two cell types to cluster together (Figure 4c). Additionally, we identify a patient with similar characteristics in the GSE118389 dataset. However, in this case, claudin-low cells split into two clusters on the t-SNE plot based on VIM expression values, with one cluster having a VIM expression equal to zero and the other having a higher VIM expression. We designate the latter as the true claudin-low cell group (Figure S16, and Figure S17). However, when we perform inferCNV analysis, we find that the high VIM group does not cluster with basal cells and does not exhibit the same CNVs (Figure S17, and Figure S18). This discrepancy may be due to the small number of cells in this particular sample.

Basal cells and Claudin-low cells exhibit genomic similarities, indicating a potential common origin. However, they show significant differences in their transcriptomes. This suggests that there may be specific molecular mechanisms driving the trans-differentiation process from Basal to Claudin-low cells.

Loss of ELF3 expression drives basal to claudin-low trans-differentiation

To explore the molecular mechanisms underlying the trans-differentiation of Basal cells into Claudin-low cells, we perform gene differential expression analysis (DEG) on Basal and Claudin-low cells from both the patient 0554 data and the HDQP1cell lines datasets. However, due to the limited number of Claudin-low cells in the PT126 patient, the gene DEG is not conducted for this group. We find 23 up-regulated genes and 20 down-regulated genes in claudin-low cells than that in basal cells that are consistently present in the DEGs of both patient 0554 and two HDQP1 cell lines. Among these genes, it has been reported that ELF3 inhibited the transcription of miR-485-5p, leading to an enhanced expression of CLDN4 in ovarian cancer [34354393] [20]. Notably, the TCGA breast cancer project has observed that the basal-like subtype of breast cancer shares many genetic alterations with high-grade serous ovarian cancer [23000897; 23329768] [21] [22]. The overexpression of ELF3 in highly invasive TNBC cell lines MDA-MB-231 and BT549 inhibits EMT by reducing the expression of several EMT-related proteins[23]. Given these findings, it raises the question of whether the loss of ELF3 in basal cells of breast cancer could also contribute to their trans-differentiation into Claudin-low cells through the loss of CLDN4 expression. We want to explore it.

Discussion

When applying UBS93 to single-cell data of breast cancer, we observe that the prediction accuracies of Luminal and HER2-amp subtypes are lower compared to Basal and Claudin-low subtypes. This discrepancy arises because we initially distinguish Luminal and HER2-amp from Basal and Claudin-low subtypes, and subsequently employ ssGSEA to differentiate between Luminal and HER2-amp subtypes. To mitigate the influence of tumor microenvironment cells, we select UBS93 genes as background genes. However, single-cell data often exhibits sparsity, resulting in many gene expression values being zero, which leads to suboptimal outcomes. To address this issue, we explore an alternative scoring method suitable for single-cell data called AUCell. However, when we use UBS93 genes as background genes, only four genes are selected, constituting a mere 5% proportion. This selection is highly unreasonable for calculating the Area Under the Curve (AUC) metric. Consequently, we decide to abandon the AUCell approach. Therefore, when applying UBS93 to single-cell data, it may be more appropriate to consider them as a single subtype without distinguishing between Luminal and HER2-amp.

We specifically focus on the trans-differentiation between basal and claudin-low subtypes in both cell lines and patients. Is there any evidence of trans-differentiation from luminal to claudin-low subtype? Previous reports indicate that when there is an exogenous or endogenous KRAS mutation expressed in an epithelial cell lineage-independent manner, luminal epithelial cells can give rise to basal-like and claudin-low breast cancer [34145248][5]. However, these findings have only been verified in mouse

models and have not yet been validated with human data. In our observations, we notice the coexistence of basal and claudin-low cells in the HDQP1 cell line and patients. It is worth noting that the HDQP1 cell line does not possess a KRAS mutation, suggesting that there may be other molecular mechanisms driving the development of claudin-low breast cancer. To investigate this further, we conduct differential gene expression analysis between basal and claudin-low cells. We find that ELF3 is down-regulated in claudin-low cells compared to basal cells. Interestingly, a study in ovarian cancer reports that ELF3 can enhance the expression of CLDN4 by inhibiting the transcription of miR-485-5p [34354393][20]. Additionally, other articles suggest that basal-like breast cancer shares several common genetic alterations with high-grade serous ovarian cancer [23000897, 23319768][21,22]. Recent studies have revealed that overexpression of ELF3 in highly invasive TNBC cell lines, namely MDA-MB-231 and BT549, exerts its inhibitory effects on EMT by downregulating the expression of various EMT-associated proteins. [1] Based on these findings, we hypothesize that the loss of ELF3 in basal cells of breast cancer may lead to the down-regulation of CLDN4, resulting in the trans-differentiation of these cells into the claudin-low subtype. To explore this further, we perform a knock-down experiment targeting ELF3 in the HCC70 breast cancer cell line, which is known to be relatively representative [31092827] [24].

We examine data from over 50 patients with triple-negative breast cancer (TNBC) and identify only two cases with mixed subtypes of basal and claudin-low, suggesting a relatively low incidence rate. However, further data are required to accurately determine the incidence rate. Since we lack survival data for these two patients, the

impact of transitioning from the basal to claudin-low subtype on survival time remains unknown. Previous studies demonstrate that claudin-low breast cancer exhibits better overall survival (OS) and disease-free survival (DFS) rates at 3 years, 5 years, and 10 years compared to the basal-like subtype [28045912] [25]. The favorable outcomes observed in claudin-low breast cancer may be attributed to extensive lymphocyte infiltration observed in many of these tumors. Numerous researchers show that increased numbers of infiltrating lymphocytes are associated with improved outcomes in breast cancer [25534375, 24608200, 19917869, 25071121] [26] [27] [28] [29]. Therefore, subtype transition may lead to better prognosis and longer survival times for patients. Fougner et al. highlight that claudin-low breast cancer exhibits high lymphocyte infiltration, along with increased expression of PD-L1 and T lymphocytes, which may present opportunities for immunotherapy [32286297] [30]. Consequently, subtype transition could impact the adjustment of patient treatment plans. Further research and clinical studies are necessary to validate these findings and fully understand the implications of subtype transitions on patient prognosis and treatment strategies.

The TCGA breast cancer project has identified similarities in the copy number landscapes and molecular alterations between basal-like breast cancer and high-grade serous ovarian cancer. These findings suggest a shared etiology and potential for similar treatment approaches [23000897, 23319768][21, 22]. Based on this information, it is worth exploring whether mixed subtypes, similar to those observed in TNBC patients, may also occur in ovarian cancer. This direction of research presents an opportunity for

further investigation and exploration.

In summary, we utilize single-cell data from breast cancer to identify a set of epithelial-specific genes and develop a molecular subtyping tool called UBS93 for breast cancer. This tool aims to provide a unified and improved classification system for breast cancer. Using UBS93, we identify TNBC cell lines and patients with a mixed subtype of basal and claudin-low. Through genome analysis, we observe significant similarities between basal and claudin-low cells, indicating a shared origin. However, transcriptomic differences suggest the presence of molecular mechanisms driving the transition between these subtypes. To further investigate this transition, we conduct a differential gene expression analysis between basal and claudin-low cells. Our analysis highlights the ELF3 gene as a potential key transcription factor involved in the transformation from basal to claudin-low subtype. Experimental verification reveals that the loss of ELF3 leads to the downregulation of CLDN3, CLDN4, CLDN7, and EPCAM, along with the upregulation of VIM. These findings shed light on the molecular mechanisms underlying subtype transitions in breast cancer and provide insights into potential regulatory factors and genes involved in this process.

Materials and Methods

Establishment of UBS93 gene panel

We obtained a breast cancer scRNA-seq dataset consisting of 8 samples from Sunny Z. Wu's research. The IDs of the samples are respectively CID3586, CID3921, CID4495, CID3948, CID4066, CID45171, CID44041 CID3963. For a specific gene G, we calculated its E-M-ratio value in sample S using the following formula:

E-M-ratio = (Epi(G) + 0.1) / (Mic(G) + 0.1)

Here, Epi(G) represents the average expression of gene G in cancer cells, and Mic(G) represents its average expression in non-cancer cells of sample S.

Next, we divided all the samples into two groups: the TNBC group (2 samples) and the Non-TNBC group (6 samples). We computed the group-wise E-M-ratio of gene G by taking the mean E-M-ratio within each group. Consequently, we obtained two group-wise E-M-ratio values for each gene G.

Finally, we ranked all the genes based on their average group-wise E-M-ratio. We performed a quantile regression analysis between average group-wise E-M-ratio (Y) and rank (X) and noticed that there is a significant change in the slope of the score after the 25000th. Subsequently, we extracted the top 2720 genes with higher scores and repeated the above steps. We found a significant change in the slope of the score after ranking 2220. We then extracted the top 500 genes with higher scores and repeated the above steps. We found that there was a significant change in the slope of the score after ranking 400th. Therefore, we believe that the selection of the top 100 genes seems to be a reasonable choice. (see Figure S1). After excluding 7 genes that lacked associated ENSEMBL gene IDs, we established the UBS93 gene panel.

Construction of the subtype centroid

The 50 CCLE breast cell lines are categorized into three groups: basal, claudin-low, and HL (including HER2_amp, luminal, and luminal_HER2_amp). Subsequently, the median of each gene in the UBS93 gene panel is calculated to construct a 93x3 centroid matrix.

Differentiating Luminal and HER2-amp subtypes

We conducted a differential gene expression analysis between Luminal and HER2-amp cell lines in a bulk RNAseq dataset of breast cancer cell lines (CCLE). From this analysis, we identified 860 common up-regulated genes specific to the HER2-amp

subtype. Among these genes, we selected the ones present in the UBS93 gene panel to define the HER2-amp signature.

In the UBS93 classifier, when a sample is classified as the HL subtype, we further classify it into either Luminal or HER2-amp subtype through an additional ssGSEA analysis (UBS93 genes used as the background and HER2-amp signature as a geneset). We plotted the ROC curve by pooling two datasets(GSE212143 and GSE48213) together and obtained an area under the curve of 0.964. (refer to Figure S4). For the final UBS93 classifier, we determined the cutoff value to be 1.662, based on Youden's index calculated in the merged dataset.

Breast cancer subtyping with genefu

In the genefu package, the gene sets used for Claudin-low subtyping and PAM50 subtyping are distinct. When analyzing a breast cancer sample, we initially employed the function claudinLow( ) to determine whether its subtype is Claudin-low. If not, we subsequently conducted PAM50 subtyping on it.

Processing of scRNA-seq data

The Seurat package was used to perform. In the t-SNE analysis, the distance between two cells C1, C2 was defined as 1- cor(C1, C2) where cor(C1, C2) is the Spearman rank correlation value across UBS93 genes. The copy number was computed with inferCNV.

Processing of bulk RNA-seq data

We mapped raw RNA-seq reads to transcriptome index with STAR and then utilized RSEM to quantify each gene's TPM (Transcripts Per Million) value. The genome versions used for humans and mice were hg38 and mm10, respectively. The genome annotation files were downloaded from GENCODE.

Software tools and statistical methods

All of the analysis was conducted in R. The ggplot2 package was used for data visualization and DESeq2 was used for identifying differentially expressed genes (FDR < 0.05 and abs(log2FC) > 1). If not specified, the two-sided Wilcoxon rank-sum test was used to compute the p-value in hypothesis testing. All heat maps were drawn using the heatmap package.

Reference

1. Nolan, E., G.J. Lindeman, and J.E. Visvader, *Deciphering breast cancer: from biology to the clinic.* Cell. **186**(8): p. 1708-1728.
2. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* Journal of clinical oncology: official journal of the American Society of Clinical Oncology. **27**(8): p. 1160-7.
3. Herschkowitz, J.I., et al., *Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.* Genome biology. **8**(5): p. R76.
4. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast cancer research: BCR. **12**(5): p. R68.
5. Rädler, P.D., et al., *Highly metastatic claudin-low mammary cancers can originate from*

luminal epithelial cells. Nature communications. **12**(1): p. 3742.

6. Pommier, R.M., et al., *Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution.* Nature communications. **11**(1): p. 3431.

7. Jovic, D., et al., *Single-cell RNA sequencing technologies and applications: A brief overview.* Clinical and translational medicine. **12**(3): p. e694.

8. Wu, S.Z., et al., *A single-cell and spatially resolved atlas of human breast cancers.* Nature genetics. **53**(9): p. 1334-1347.

9. Neve, R.M., et al., *A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.* Cancer cell. **10**(6): p. 515-27.

10. Greer, Y.E., et al., *Mitochondrial Matrix Protease ClpP Agonists Inhibit Cancer Stem Cell F unction in Breast Cancer Cells by Disrupting Mitochondrial Homeostasis.* Cancer research communications. **2**(10): p. 1144-1161.

11. Daemen, A., et al., *Modeling precision treatment of breast cancer.* Genome biology. **14**(10): p. R110.

12. Marcotte, R., et al., *Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance.* Cell. **164**(1-2): p. 293-309.

13. Badve, S., et al., *Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists.* Modern Pathology: an official journal of the United States and Canadian Academy of Pathology, Inc. **24**(2): p. 157-67.

14. Bertucci, F., et al., *How basal are triple-negative breast cancers?* International journal of

cancer. **123**(1): p. 236-40.

15. Gambardella, G., et al., *A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response.* Nature communications. **13**(1): p. 1714.

16. Jovanović, B., et al., *Heterogeneity and transcriptional drivers of triple-negative breast cancer.* Cell reports. **42**(12): p. 113564.

17. Hollern, D.P., et al., *B Cells and T Follicular Helper Cells Mediate Response to Checkpoint Inhibitors in High Mutation Burden Mouse Models of Breast Cancer.* Cell. **179**(5): p. 1191-1206.e21.

18. Gendoo, D.M.A., et al., *Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer.* Bioinformatics (Oxford, England). **32**(7): p. 1097-9.

19. McDavid, A., et al., *Data exploration, quality control and testing in single-cell qPCR-base d gene expression experiments.* Bioinformatics (Oxford, England). **29**(4): p. 461-7.

20. Kuang, L. and L.a. Li, *E74-like factor 3 suppresses microRNA-485-5p transcription to trigger growth and metastasis of ovarian cancer cells with the involvement of CLDN4/Wnt/β-catenin axis.* Saudi journal of biological sciences. **28**(8): p. 4137-4146.

21. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature. **490**(7418): p. 61-70.

22. Chentli, F., et al., *Congenital craniopharyngioma: a case report and literature review.* Journal of pediatric endocrinology & metabolism: JPEM. **25**(11-12): p. 1181-3.

23. Park, S. and Y. Kwon, *117P - The role of ELF3 in metastasis of triple negative breast cancer.* Annals of Oncology, 2017. **28**: p. x34.

24. Liu, K., et al., *Evaluating cell lines as models for metastatic breast cancer through i ntegrative analysis of genomic data.* Nature communications. **10**(1): p. 2138.

25. Dias, K., et al., *Claudin-Low Breast Cancer; Clinical & Pathological Characteristics.* PloS one. **12**(1): p. e0168669.

26. Denkert, C., et al., *Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers.* Journal of clinical oncology: official journal of the American Society of Clinical Oncology. **33**(9): p. 983-91.

27. Loi, S., et al., *Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial.* Annals of Oncology: official journal of the European Society for Medical Oncology. **25**(8): p. 1544-50.

28. Denkert, C., et al., *Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer.* Journal of clinical oncology: official journal of the American Society of Clinical Oncology. **28**(1): p. 105-13.

29. Adams, S., et al., *Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199.* Journal of clinical oncology: official journal of the American Society of Clinical Oncology. **32**(27): p. 2959-66.

30. Fougner, C., et al., *Re-definition of claudin-low as a breast cancer phenotype.* Nature communications. **11**(1): p. 1787.