UBS93 enables delineating intratumor subtype heterogeneity in breast cancer

Jing Li[1, #], Yuan Gao[2, #], Shouhui Guo[1], Shisong Zhang[3, 4, *], Weixing Yu[5, *], Ke Liu[1,*]

Affiliations:

1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China
2. Precision Medicine Laboratory for Chronic Non-communicable Diseases of Shandong Province, Institute of Precision Medicine, Jining Medical University, Jining, Shandong Province, China.
3. Department of Thoracic and Oncological Surgery, Children's Hospital Affiliated to Shandong University, Jinan, China
4. Department of Thoracic and Oncological Surgery, Jinan Children's Hospital, Jinan, China
5. Department of Biochemistry and Molecular Biology, School of Basic Medicine, Jining Medical University, Jining, China

# These authors contributed equally

* To whom correspondence should be addressed

zsli323@163.com

weixingy@mail.jnmc.edu.cn

keliu.iluke@email.sdu.edu.cn

Abstract

Breast cancer is a heterogeneous disease that comprises multiple molecular subtypes. Although several methods have been developed to subtype bulk tumor samples, their capacity to delineate intratumor heterogeneity at the single-cell level remains limited. In this study, we present UBS93, a novel classifier specifically designed for breast cancer subtyping in both bulk tumor samples and individual cancer cells. We rigorously validate UBS93 and demonstrate its superior performance compared to existing methods, particularly in the identification of Claudin-low samples. Application of UBS93 to single-cell RNA sequencing (scRNA-seq) data from triple-negative breast cancer patients uncovered a previously unrecognized "Basal-like/Claudin-low mixed" subtype, suggesting a Basal-like origin for Claudin-low cancer cells. Subsequent computational analyses and experimental validations reveal that downregulation of transcription factor *ELF3* induces a partial Claudin-low phenotype in Basal-like breast cancer cells. Collectively, our findings provide a powerful tool for elucidating intratumor subtype heterogeneity in breast cancer, with potential applicability to other cancer types.

Introduction

Breast cancer is a heterogeneous disease characterized by diverse genomic profiles, transcriptomic alterations, and clinical outcomes. The heterogeneity presents challenges in clinically relevant tumor classification[1].

Parker *et al.* utilized microarray technology to develop an intrinsic breast cancer classifier PAM50, which stratifies the disease into five subtypes: LuminalA, LuminalB, HER2-enriched, Basal-like, and Normal-like[2]. This classification system provides superior prognostic and predictive capabilities compared to traditional methods, improving upon pathological staging, histological grading, and standard clinical biomarkers[2]. In addition to PAM50, Herschkowitz *et al.* identified another subtype known as Claudin-low which is characterized by low expression levels of cell adhesion molecules (such as *CLDN3*, *CLDN4*, *CLDN7*, and *CDH1*) and marked by high stemness and mesenchymal traits[3, 4, 5, 6]. Fougner *et al.* conducted integrative multi-omics analyses on breast cancer samples from multiple cohorts and further confirmed Claudin-low as a distinct subtype[6].

The use of single-cell omics technology has profoundly enhanced our understanding of the complexity of breast cancer. Several recent studies have uncovered previously overlooked intratumor subtype heterogeneity, challenging the traditional classification strategy[7, 8, 9, 10, 11,12, 13, 14, 15, 16]. For instance, Wu *et al*. computationally identified Basal-like cancer cells within ER+ breast cancer patients and further confirmed the finding experimentally[16]. Similarly, the emergence of such “mixed subtype” phenomenon has been observed in mouse models through the activation of endogenous mutant *KRAS* or

72 exogenous *KRAS* expression[5]. These discoveries underscore the limitations of
73 traditional bulk omics-based subtyping and highlight the importance of achieving
74 single-cell resolution for more precise classification.
75 Despite advancements in computational methods for breast cancer subtyping, two
76 significant challenges remain. First, most of the existing methods are designed for
77 bulk omics data, which restricts their utility in single-cell analyses. Second, these
78 approaches often focus on PAM50 subtypes while overlooking the Claudin-low
79 subtype, a critical oversight that can lead to misclassification. Addressing these issues
80 is essential for refining breast cancer subtyping and improving diagnostic and
81 therapeutic strategies.
82 In this study, we developed a novel gene panel specifically tailored for breast cancer
83 subtyping and designed a classifier based on this panel. We rigorously evaluated its
84 performance using both bulk and single-cell RNA sequencing (scRNA-seq) data in
85 multiple species. Notably, we identified the presence of Claudin-low cancer cells
86 within Basal-like breast cancer cell lines and patient samples (a phenomenon we
87 define as the "Basal-like/Claudin-low mixed" subtype), demonstrating the utility of
88 our method in delignating intratumor subtype heterogeneity. Further computational
89 analysis and experimental assays revealed that *ELF3* knock-down induced a partial
90 Claudin-low phenotype in Basal-like breast cancer, suggesting a potential Basal-
91 like/Claudin-low trans-differentiation. Our work provides a powerful tool for
92 delineating intratumor subtype heterogeneity in breast cancer and should be
93 applicable to other cancers.

Results

Development of a uniform gene panel for breast cancer subtyping

Solid tumor tissue is a heterogeneous mixture of various cell types[16]. When quantifying the transcriptome of a tumor sample with bulk RNA sequencing technology, the expression of a gene could be considered as the combination of its expression in cancer cells and non-cancer cells. Given a gene g, suppose its average expression in cancer cells and non-cancer cells are C_g and M_g, respectively. If the tumor purity is α, then the expression value of g can be modeled as $\alpha * C_g + (1-\alpha) * M_g$. Given several random genes, their rankings (according to the measured expression in a bulk tumor sample) should significantly differ from those derived from pure cancer cells due to the contribution of microenvironment (**Figure 1a, left panel**); however, if these genes exhibit high cancer cell specificity (i.e., the value M for each gene is approximately zero), then the two ranking lists should be highly correlated (**Figure 1a, right panel**). Therefore, the rankings of cancer cell-specific genes might be informative for molecular subtyping.

To identify these genes, we analyzed the scRNA-seq dataset from Wu *et al.*'s study and calculated a value, termed the "C-M ratio", for each gene (see Materials and Methods). Among the top 100 genes with the highest C-M ratios, seven lacked ENSEMBL IDs and were subsequently removed, resulting in a uniform gene panel for breast cancer subtyping (UBS93) (**Supplementary Table 1**). As expected, the expression of UBS93 genes was significantly higher in cancer cells when compared to non-cancer cells (**Figure 1b**).

116 The Cancer Cell Line Encyclopedia (CCLE) project classifies breast cancer into five
117 subtypes: basal_A, basal_B, luminal, HER2-amp, and luminal-HER2-amp[17]. We
118 applied t-distributed Stochastic Neighbor Embedding (t-SNE) to the CCLE breast
119 cancer bulk RNA-seq data and observed that cell lines clustered according to their
120 subtypes, thereby validating the CCLE subtyping criteria (**Figure 1c**). As a result, we
121 primarily used CCLE subtypes in our study. However, to align with existing studies,
122 we made the following adjustments to subtype names. First, we renamed basal_A and
123 basal_B to Basal-like and Claudin-low, respectively. Second, we excluded the
124 luminal-HER2-amp subtype (see Materials and Methods). Ultimately, we classified
125 breast cancer into four molecular subtypes: Basal-like, Claudin-low, Luminal, and
126 HER2-amp.

127 We repeated the t-SNE analysis on the CCLE breast cancer bulk RNA-seq data with
128 only UBS93 genes (**Supplementary Figure 1**) and obtained results similar to
129 previous t-SNE analysis, confirming the effectiveness of the UBS93 panel in breast
130 cancer subtyping. Building on this, we developed our subtype classifier **(Figure 1d**).
131 We first constructed a 93 × 3 subtype centroid matrix, where each column
132 represented the average expression vector of UBS93 genes across the CCLE cell lines
133 corresponding to Basal-like, Claudin-low, Luminal or HER2-amp, respectively. Given
134 the gene expression profile of a patient sample (or a single cancer cell), we computed
135 its Spearman rank correlation with each column of the subtype centroid matrix and
136 assigned it to the subtype with the highest correlation coefficient. If the sample is
137 subtyped as Luminal or HER2-amp, we further calculated an enrichment score of

138 HER2-amp signature genes and utilized this score to determine the sample's final

139 subtype (see Materials and Methods).

140

141 **Evaluation of UBS93 classifier in human and mice**

142 We first evaluated the performance of the UBS93 classifier using an assembled

143 dataset which contains the bulk RNA-seq profile of 47 breast cancer cell lines

144 extracted from the Gene Expression Omnibus (GEO) database[18, 19, 20]**(Supplementary**

145 **Table 2)**. We found 93.6% of the samples were distributed on the diagonal of the

146 confusion matrix, suggesting the high overall accuracy of our method (**Figure 2a**).

147 Notably, our classifier also demonstrated optimal subtype-specific performance and

148 the F1 score for the four subtypes were 0.97 (Basal-like), 0.80 (Claudin-low), 1.00

149 (HER2-amp), and 0.92 (Luminal), respectively. In addition to cell lines, we also

150 evaluated our classifier on the breast cancer samples from The Cancer Genome Atlas

151 (TCGA) and derived ideal overall accuracy (81.9%) and subtype-specific

152 performance[21] (**Figure 2b**).

153 Next, we validated our classifier using an assembled scRNA-seq dataset of human

154 breast cancer cell lines (**Supplementary Table 3**)[22, 23]. For each cell line, we

155 computed the accuracy as the proportion of cells that were correctly subtyped. Our

156 classifier achieved an accuracy higher than 80% in 19 (63.3%) of the 30 cell lines and

157 the median accuracy was 85.7%. The subtype-specific median accuracy values were

158 99.0% (Basal-like), 98.3% (Claudin-low), 78.0% (HER2-amp), and 75.7% (Luminal)

159 (**Figure 2c**).

160 We assembled a dataset of 27 bulk RNA-seq samples to validate the performance of
161 our classifier in murine-derived datasets. In the dataset, 17 samples were from
162 Hollern's study[24] and experimentally confirmed as Claudin-low; the other 10 samples
163 were from Rädler's study[5] but lacked subtype labels. According to the results reported
164 in the literature, five of them were Basal-like and the other five were Claudin-low. We
165 performed hierarchical clustering based on the expression of five "Claudin-low
166 indicative" genes (*Cldn3, Cldn13, Cldn7, Vim, and Epcam*) to impute the missing
167 subtype labels (**Figure 2d**) and then applied our classifier to the 27 samples. We
168 found the subtype assigned by our classifier aligned with the true subtype in all
169 samples (**Supplementary Table 4**).
170
171 **Comparison to existing methods**
172 We systematically compared our classifier with genefu, a widely utilized R package
173 for performing PAM50 subtyping of breast cancer[25]. Notably, only six genes are
174 presented in both panels (**Supplementary Figure 2a**). In addition, UBS93 genes
175 exhibited significantly higher C-M-ratio values and cancer cell expression than
176 PAM50 genes (**Supplementary Figure 2b**). We compared the performance of the
177 two classifiers using the bulk RNA-seq data of TCGA breast cancer samples. The
178 genefu achieved slightly better overall accuracy (83.7%); however, the UBS93
179 classifier achieves better performance in Basal-like, Claudin-low, and HER2-amp
180 subtypes (**Figure 3a**). In murine datasets, the genefu package misclassified all Basal-
181 like samples as Claudin-low while our classifier accurately identified them as Basal-
182 like (**Supplementary Table 5**).

183 In addition to genefu, we also compared our classifier with SCSubtype (a method
184 developed for subtyping a single breast cancer cell) [16] on the assembled scRNA-seq
185 dataset that is used to evaluate our classifier. Since the Claudin-low subtype is not
186 included in SCSubtype, we therefore only made the comparisons on the three other
187 subtypes and found UBS93 classifier exhibited significantly higher accuracy than
188 SCSubtype in the Basal-like subtype (**Figure 3b**).

189 **Discovery of "Basal-like/Claudin-low mixed" subtype**

190 We observed that although our classifier generally performed well on scRNA-seq data,
191 the cell line HDQP1, which is believed to be a Basal-like cell line, appeared to be an
192 exception. Surprisingly, only 56.4% of HDQP1 single cells were subtyped as Basal-
193 like with the remainder subtyped as Claudin-low (**Figure 2c)**. Interestingly, when
194 performing t-SNE visualization of the HDQP1 scRNA-seq data, the cells formed two
195 distinct clusters (**Figure 4a**). One cluster consisted of only Basal-like cells and 84.2%
196 of the other cluster were Claudin-low cells. We performed differential gene expression
197 analysis between UBS93-identified Claudin-low and Basal-like cells and found the
198 cells of Claudin-low group exhibited established characteristics of Claudin-low tumors[4,
199 5, 6, 26, 27]. First, there was significantly lower expression of cell-cell adhesion genes such
200 as *CLDN4*, *CLDN7*, *KRT19,* and *EPCAM* (**Figure 4b, Supplementary Figure 3,**
201 **Supplementary Table 6**). Second, there were higher levels of epithelial-to-
202 mesenchymal transition (EMT)-associated genes, as illustrated by up-regulation of *VIM*
203 and the enrichment of EMT signature genes in the up-regulated differentially expressed
204 (DE) genes (**Figure 4c**). Third, higher stemness was identified as indicated by an

205 elevated *CD44* (a marker of cancer stem cell) level and *CD44*/*CD24* ratio[4, 27, 28](**Figure**
206 **4d**). We next assembled a pseudo-bulk sample using the cells subtyped as Claudin-low
207 and computed its transcriptomic correlation with CCLE breast cancer cell lines[17]. The
208 cell line with the highest correlation coefficient was CAL120, a Claudin-low cell line
209 **(Figure 4e)**. Moreover, four of the five cell lines with the highest correlation were
210 Claudin-low. In contrast, when we performed the same analysis using cells subtyped as
211 Basal-like, HDQP1 had the highest correlation coefficient (**Figure 4f**), which is
212 expected. Our computational analysis suggested that HDQP1 cell line was a mixture of
213 Basal-like and Claudin-low cells, therefore it is a cell line of "Basal-like/Claudin-low
214 mixed" subtype. To double confirm our finding, we replicated our analysis on the
215 HDQP1 scRNA-seq data from another dataset (GSE202771) and derived the same
216 conclusion (**Supplementary Figure 4, 5 and Supplementary Table 7**).
217 We next investigated whether the "Basal-like/Claudin-low mixed" subtype
218 phenomenon occurs in patients. We applied the UBS93 classifier to the scRNA-seq
219 data of 50 triple-negative breast cancer (TNBC) samples and ranked them according to
220 the median correlation coefficient (with the Claudin-low centroid) of UBS93-identified
221 Claudin-low cancer cells (**Figure 4g, Supplementary Table 8**). The metaplastic breast
222 cancer patient CID4513 (from dataset GSE176078) had the highest ranking, which may
223 be not surprising since previous studies have confirmed that metaplastic breast cancer
224 cells exhibit characteristics of Claudin-low subtype[29, 30]. Interestingly, the patient 0554
225 (from dataset GSE161529), who was reported as a Basal-like breast cancer patient in
226 the literature, had the second highest ranking[15]. Upon further exploration, we found

227 most (91.1%) of the cancer cells were subtyped as Basal-like while 7.9% were subtyped
228 as Claudin-low. Similar to HDQP1, the Claudin-low cells tended to form a distinct
229 cluster on the t-SNE plot and meanwhile exhibited the three established characteristics
230 of Claudin-low tumor (**Figure 5a-5d, Supplementary Figure 6, and Supplementary**
231 **Table 9)**. In addition, the transcriptomic correlation analysis using pseudo-bulk
232 samples validated the results (**Figure 5e, 5f**). Therefore, we concluded that the Claudin-
233 low cancer cells identified in patient 0554 were highly convincible and this patient was
234 an instance of "Basal-like/Claudin-low mixed" subtype.
235 Together, these results highlight the existence of a Basal-like/Claudin-low mixed
236 subtype in both TNBC cell lines and patient samples, underscoring the complexity
237 within TNBC subtypes and the utility of UBS93 in delineating intratumor subtype
238 heterogeneity.
239
240 ***ELF3* knock-down induces a partial Claudin-low phenotype in Basal-like cancer**
241 To date, the cellular origin of Claudin-low cancer cells remains unclear. While
242 Creighton *et al*. proposed that these cells originate from breast stem cells[31], evidence
243 from mouse models suggests they may instead arise from normal luminal epithelial
244 cells[5]. Pommier *et al.* performed large scale analysis on bulk cancer genomics data
245 and proposed the three potential origins of Claudin-low cancer cells: malignant
246 conversion from normal human mammary stem cells, transformation from luminal (or
247 basal-like) cancer cells[32]. However, no robust single cell-level evidence from patient
248 samples was provided in their study. The case of patient 0554, who exhibits the
249 "Basal-like/Claudin-low mixed" subtype, offers a unique opportunity to investigate

250 the cellular origin of Claudin-low cancer cells in greater depth. Hierarchical clustering
251 of various cell types based on their CNV (copy number variation) profiles revealed
252 that Claudin-low cancer cells share the highest similarity with Basal-like cancer cells
253 (**Figure 6a**), supporting the Basal-like origin of Claudin-low cancer cells in human
254 TNBC patients.
255 To investigate the genes which may play a pivotal role in the Basal-like/Claudin-low
256 transformation, we intersected the DE results performed in HDQP1 cell line and patient
257 0554, identifying 18 commonly upregulated and 37 commonly downregulated genes
258 (**Supplementary Table 10**). Among these, the transcription factor *ELF3* emerged as a
259 key player for two reasons **(Supplementary Figure 7).** First**,** overexpression of *ELF3*
260 in highly invasive Claudin-low cell lines (MDA-MB-231 and BT549) inhibits the EMT
261 process[33]; second, *ELF3* enhances the expression of *CLDN4* in ovarian cancer and non-
262 small cell lung cancer[34, 35]. Therefore, we questioned whether depletion of this gene in
263 Basal-like breast cancer would trigger the Basal-like/Claudin-low transformation.
264 Since HCC70 was previously identified as the cell line most closely resembling the
265 transcriptome of Basal-like breast cancer, we selected it for further studies[36]. First, we
266 performed scRNA-seq on HCC70 cells and applied the UBS93 classifier to the data.
267 As expected, 98.7% of the cells were classified as Basal-like, with no cells classified as
268 Claudin-low, confirming the high quality of the cell line. We then conducted scRNA-
269 seq on *ELF3* knockdown HCC70 cells to determine whether *ELF3* depletion could
270 convert Basal-like cells into Claudin-low cells, however UBS93 analysis indicated that
271 the majority of the cells (98.1%) remained Basal-like.

272 Despite failing to induce complete Basal-like/Claudin-low transformation, we explored
273 whether *ELF3* knockdown could partially induce Claudin-low characteristics in
274 HCC70 cells. To account for variable knockdown efficiency, we selected the 100 cells
275 with the lowest *ELF3* expression (sh-*ELF3* group) and the 100 cells with the highest
276 *ELF3* expression (control group). Interestingly, the sh-*ELF3* group exhibited a
277 significantly lower correlation with the Basal-like centroid compared to the control
278 group (**Supplementary Figure 8a**). Differential gene expression analysis identified
279 210 upregulated and 79 downregulated genes (**Figure 6b, Supplementary Table 11**).
280 As expected, *CLDN4* was significantly down-regulated (p-value = 3.57e-14); in
281 addition, we also observed the down-regulation of the epithelial cell marker *KRT19*.
282 Interestingly, we found several of the down-regulated DE genes (such as *CLDN4*)
283 showed Claudin-low specific depletion (**Figure 6c**). We curated a set of such genes
284 (which we termed as the CLSD genes) and GSEA analysis revealed significant
285 enrichment of CLSD genes in the down-regulation part of DE gene list
286 (**Supplementary Figure 8b**).
287 We next examined how *ELF3* knock-down influenced the expression of *CLDN3*,
288 *CLDN4*, *CLDN7*, *EPCAM*, and *VIM* with RT-PCR (**Figure 6d**). Consistent with
289 scRNA-seq DE analysis, *CLDN4* was significantly down-regulated upon *ELF3* knock-
290 down (p-value = 0.002254). Interestingly, we also observed significant down-
291 regulation of *CLDN7* although it was not in the DE gene list (p-value = 0.018557); no
292 significant changes were detected for *CLDN3* and *EPCAM*. Unexpectedly, *VIM* was
293 significantly down-regulated (p-value = 0.003883), which is inconsistent with other

294 multiple studies reporting the suppression role of *ELF3* in EMT process. We conclude
295 there may be unknown mechanisms underlying the regulation of EMT by *ELF3* and
296 further investigation is needed. To double confirm our findings, we performed
297 replicated experiments and got same conclusion (**Supplementary Figure 9**). In
298 summary, our results suggest that *ELF3* knock-down induces a partial Claudin-low
299 phenotype in Basal-like cancer cell line, as evidenced by the down-regulation of CLSD
300 genes such as *CLDN4* and *CLDN7*.
301

302 **Discussion**

303 In this study, we introduced a novel classifier, UBS93, for the molecular subtyping of
304 breast cancer. We validated its performance using both bulk and single-cell RNA
305 sequencing data from human and mouse models. Notably, our classifier demonstrates
306 superior performance compared to classical methods, particularly in the identification
307 of Claudin-low subtypes. Unlike genefu, the most widely used method for breast
308 cancer subtyping, our classifier utilizes a single, uniform gene signature.
309 A key advantage of this approach is the simplification of the probability assignment
310 process. In UBS93, all correlation values (with subtype centroids) are calculated using
311 UBS93 genes, ensuring that the values are directly comparable. The probability of a
312 sample being classified as a specific subtype S_i is given by the formula:
313

314 $$Pr(Si) = cor(S_i) \,/\, \sum cor(S_j)$$

315 Where cor(Si) represents the correlation with the i-th centroid (values below zero are
316 assigned a value of zero). In contrast, when correlation values are computed across
317 different gene signatures (e.g., in genefu, the correlation with the Claudin-low
318 centroid is computed using non-PAM50 genes), additional adjustments are required to
319 make the correlations comparable.
320 Beyond computational analysis, our study lays the groundwork for a novel breast
321 cancer subtyping assay. We demonstrate that quantifying the expression of UBS93
322 genes is sufficient for molecular subtyping and can be achieved using cost-effective
323 technologies such as qRT-PCR. Future experiments will be conducted to validate our
324 concept.
325 Using UBS93, we identified human breast cancer cell lines and patients with the
326 "Basal-like/Claudin-low mixed" subtype. While Wu *et al*. previously reported "mixed
327 subtype" phenomenon in ER+ breast cancer patients[16], our finding was overlooked in
328 their study due to the exclusion of the Claudin-low subtype in the computational tool
329 SC Subtype. Our research complements and extends Wu *et al*.'s work, contributing to
330 a deeper understanding of intratumor heterogeneity especially in triple-negative breast
331 cancer.
332 Prior to our study, Pommier *et al*., through integrative analysis of large-scale cancer
333 genomics data, proposed three potential origins for Claudin-low cancer cells:
334 malignant conversion from normal human mammary stem cells, transformation from
335 luminal (or basal-like) cancer cells[32]. The discovery of the "Basal-like/Claudin-low
336 mixed" subtype, alongside our experimental findings, provides compelling evidence

337 supporting the Basal-like origin of Claudin-low cancer cells. In the future, we plan to
338 utilize UBS93 to obtain evidence for the other two hypotheses.
339 Our research also raises several interesting scientific questions for further
340 investigation. First, besides *ELF3* knock-down, which additional genetic perturbations
341 are needed to fully covert Basal-like cancer cells to Claudin-low cancer cells? High
342 EMT level is one of the three hallmark characteristics of Claudin-low cancer cells;
343 however, we did not detect significant up-regulation of mesenchymal marker genes
344 (such as *VIM* and *FN1*) upon *ELF3* knock-down. We noticed that *ZEB1,* a key EMT
345 regulator, was consistently up-regulated in the DE analysis between Claudin-low and
346 Basal-like cancer cells. Therefore, we intend to explore whether double genetic
347 perturbations (*ELF3*-knockdown and *ZEB1*-overexpression) would induce complete
348 Claudin-low phenotype.
349 Second, is *KRAS* mutation necessary for generation of Claudin-low cancer cells?
350 Rädler *et al.* reported that in mouse models activation of endogenous *KRAS* mutation
351 (or expression of exogenous *KRAS*) drove normal luminal epithelial cells to Claudin-
352 low and Basal-like cancer cells[5]; however, these findings have not been thoroughly
353 validated in human. In fact, *KRAS* mutation is rare in human breast cancer. According
354 to cBioPortal, we only identified 6 TCGA breast cancer patients carrying *KRAS*
355 mutation; in addition, the cell line HDQP1 also does not carry *KRAS* mutation.
356 Therefore, there might be *KRAS*-mutation-independent mechanisms which may
357 promote the generation of Claudin-low cancer cells.

358 Third, how does the presence of "Basal-like/Claudin-low mixed" subtype influence
359 the clinical outcomes? Distant metastasis remains the leading cause of cancer-related
360 deaths[37], and Claudin-low breast cancer has been shown to be highly aggressive
361 compared to other subtypes[27]. Therefore, it is reasonable to hypothesize that the
362 "Basal-like/Claudin-low mixed" subtype is an unfavorable prognostic marker in
363 TNBC. We plan to integrate scRNA-seq data with survival information to explore this
364 hypothesis further.
365 Last but not least, Fougner *et al.* reported that Claudin-low breast cancer exhibited high
366 lymphocyte infiltration and increased expression of PD-L1[6]. Therefore, we wonder
367 whether the presence of "Basal-like/Claudin-low mixed" subtype re-shapes the tumor
368 microenvironment. Briefly, we plan to apply the UBS93 classifier to spatial
369 transcriptomics data from patients with this "Basal-like/Claudin-low mixed" subtype,
370 identifying areas of Claudin-low cancer cells and comparing their surrounding
371 microenvironments to those of Basal-like cancer cells. This analysis could provide
372 valuable insights into potential immunotherapy opportunities.
373 In conclusion, our work presents a powerful tool for delineating intratumor subtype
374 heterogeneity in breast cancer. As high-quality tumor scRNA-seq data become
375 increasingly available, we anticipate that our strategy will be applicable to other
376 cancers and will help uncover novel biological mechanisms underlying tumor
377 progression.
378
379

380 # Materials and Methods

381 **Development of UBS93 gene panel**

382 We obtained a breast cancer scRNA-seq dataset consisting of 8 samples from Wu *et*
383 *al.*'s research[16]. Given a gene g, its C-M-ratio value in sample S is calculated with the
384 following formula:

385 $$\text{C-M-ratio} = (C_g + 0.1) / (M_g + 0.1)$$

386 Here, C_g and M_g represent the average expression of G in cancer cells and non-cancer
387 cells (of sample S), respectively.

388 Next, we divided all the samples into two groups: the TNBC group (2 samples) and
389 the non-TNBC group (6 samples). We computed the group-wise C-M-ratio of gene g
390 by taking the mean C-M-ratio within each group. Consequently, we obtained two
391 group-wise C-M-ratio values for each gene g.

392 Finally, we ranked all genes according to their average group-wise C-M-ratio. We
393 performed a quantile regression analysis between average group-wise C-M-ratio (Y)
394 and rank (X) and noticed that there appears to be a change point of the slope. We
395 picked out the top 400 genes to re-perform the analysis and found 300 appeared to be
396 a reasonable estimate of the change point. Therefore, we finally kept the top 100
397 genes (**Supplementary Figure 10**). After excluding 7 genes that are lacked
398 ENSEMBL gene IDs, we established the UBS93 gene panel.

399

400 **Determination of breast cancer subtypes**

401 The CCLE categorized breast cancer into five subtypes: basal_A, basal_B, Luminal,
402 HER2-amp, and luminal-HER2-amp. In the t-SNE plot (**Figure 1c**), basal_A and
403 basal_B form their own distinct cluster while the other three subtypes form the third
404 cluster. The basal_A cluster is significantly enriched of cell lines whose PAM50

subtype is Basal-like; therefore, it is reasonable to rename it as Basal-like; similarly, basal_B was renamed as Claudin-low. The luminal-HER2-amp subtype is not used in our classifier since the two cell lines of this subtype do not appear to form a distinct cluster.

Differentiating Luminal and HER2-amp subtypes

We conducted differential gene expression analysis between CCLE HER2-amp and Luminal cell lines and identified 860 significantly up-regulated genes. We picked out the 20 genes which are present in the UBS93 panel and derived the HER2-amp signature.

In our classifier, ssGSEA analysis (UBS93 genes as background, and HER2-amp signature as a geneset) is utilized to differentiate Luminal and HER2-amp subtypes. We validated such strategy on a pooled bulk RNA-seq dataset (GSE212143 and GSE48213) and obtained an AUC (area under the curve) value of 0.964 (**Supplementary Figure 11**). In the final classifier, we determined the cutoff value to be 1.662 based on Youden's index calculated in the validation dataset.

Breast cancer subtyping with genefu

In genefu package, the gene sets used for Claudin-low and PAM50 subtyping are different. When analyzing a breast cancer sample, we first employ the function claudinLow to determine whether its subtype is Claudin-low; if not, we subsequently conduct PAM50 subtyping. In PAM50 subtyping, we removed the Normal-like subtype and merged LuminalA and LuminalB as Luminal.

Gene set enrichment analysis

430 The EMT signature was downloaded from MSigDB[38]. To derive the Claudin-low-
431 specific depletion (CLSD) gene set, z-score normalization was applied to gene
432 expression values (across all CCLE breast cancer cell lines) and the median
433 expression value for each gene was then calculated specifically for Claudin-low cell
434 lines. Genes with median expression values below -1 were selected. In total, resulting
435 in a total of 960 CLSD genes. The clusterProfiler R package was used to compute
436 enrichment score and p-value[39].
437

438 **Processing of bulk RNA-seq data**

439 We mapped raw bulk RNA-seq reads to transcriptome index with STAR[40] and then
440 utilized RSEM[41] to quantify each gene's TPM (Transcripts Per Million) value. The
441 genome versions used are hg38 and mm10, respectively. The genome annotation files
442 were downloaded from GENCODE[42].
443

444 **The t-SNE analysis**

445 Given two single cells C1 and C2, their distance is defined as 1- cor(C1, C2) where
446 cor(C1, C2) is the Spearman rank correlation value across used genes. After the
447 computation of the distance matrix, Rtsne function is used to generate the final
448 embedding.
449

450 **Cell lines and cell culture conditions**

451 HCC70 cells and HEK293T cells were obtained from Pricella Biotechnology Co., Ltd
452 and American Type Culture Collection. HCC70 cells maintained in RPMI 1640
453 (catalog no.22400089, Gibco). HEK293T cells maintained in DMEM (catalog no.

454 SH30285.01, Cytiva Life Sciences). All medium supplemented with 10% FBS (catalog
455 no. 164210-50, Pricella Biotechnology).
456
457 **Lentivirus packaging and infection**
458 HEK293T cells (2 × 106) were seeded in a 100-mm cell culture dish and transfected
459 with plasmids using PolyJet in vitro DNA transfection reagent PEI (C0537-20mg,
460 Beyotime Biotechnology). The transfection experiments were operated according to the
461 manufacturer's instructions. For knock-down target proteins, we co-transfected the
462 cells with the plasmids of pMDL, pVSVG, pREV and pLKO.1 (sh-*ELF3*, scrambled)
463 to generate lentivirus. For lentivirus infection, 500 μl of virus-containing medium was
464 added to the 5 × 104 cells with 1% polybrene.
465
466 **Real-time PCR**
467 Total RNA was extracted from cells using an RNAsimple Total RNA Kit
468 (TIANGEN, DP419). Complementary DNA was prepared using the High-Capacity
469 cDNA Reverse Transcription Kit purchased from Vazyme Biotech (catalog no. R333–
470 01). Real-time PCR analysis was performed using 2 × Universal SYBR Green Fast
471 qPCR Mix (catalog no. RK21203, Abclonal Technology) with the following
472 conditions: 3 min at 95 °C, and then 40 cycles at 95 °C for 5 s, 60 °C for 25 s, using a
473 Real-Time System (Roche LightCycler® 480). Data were normalized according to
474 expression of a control gene (ACTB) in each experiment. The data are presented as
475 the mean ± s.d. from three independent experiments. The p-values were computed
476 with unpaired t-test in GraphPad Prism 10 software.
477
478 **Single-cell RNA-sequencing of HCC70**

479 The single-cell RNA-sequencing of HCC70 cell line was performed by a commercial
480 company Novogene. Briefly, RNA from the barcoded cells was subsequently reverse-
481 transcribed and sequencing libraries were constructed with reagents from a Chromium
482 Single Cell 3' v2 reagent kit (10X Genomics) according to the manufacturer's
483 instructions. Sequencing was performed with Illumina sequencing machines
484 according to the manufacturer's instructions. After filtering low-quality read with
485 fastp[43], raw reads were demultiplexed and mapped to the human reference genome by
486 cellranger 7.2.0[44].
487

488 **Software tools and statistical methods**

489 All the analysis was conducted in R. Seurat[45, 46, 47, 48] was used for processing scRNA-
490 seq data. In DE analysis, genes with adjusted p-value < 0.05 and abs(log2FC) > 0.5)
491 were considered as differentially expressed genes. The ggplot2[49] package was used
492 for data visualization, pheatmap and ComplexHeatmap packages were used to
493 generate heat maps[50, 51], GSVA was used for ssGSEA analysis[52, 53]. If not specified,
494 the two-sided Wilcoxon rank-sum test was used to compute the p-value in hypothesis
495 testing.
496

497 **Data Availability**

498 The authors declare that all data supporting the findings of this study are available
499 within the article and its supplementary information files or from the corresponding
500 author upon reasonable request.
501

502 **Code Availability**

503 The code is available at github (https://github.com/bioklab/UBS93).

Acknowledgments

We thank Dr. Xiaohui Lin for helping us revise the manuscript. The research is supported by National Natural Science Foundation of China (Fund 32370715), Science Fund for Distinguished Oversea Young Scholars of Shandong Province (2023HWYQ-015), Taishan Young Scholar Program of Shandong Province (tsqn202312020), and Cheeloo Young Scholar Program of Shandong University. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

Authors' contribution

J.L. and K.L. conceived the study. J.L. performed the majority of computational analysis, Y.G. and W.X.Y. performed all the experimental assays, and K.L. supervised the study. All authors contributed to writing, reviewing, and editing the manuscript and approved the manuscript.

Competing Interests

The researchers declare no competing interests.

Reference

1. Nolan E, Lindeman GJ, Visvader JE. Deciphering breast cancer: from biology to the clinic. *Cell* **186**, 1708-1728.

2. Parker JS*, et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167.

3. Herschkowitz JI, *et al.* Identification of conserved gene expression features between murine ma mmary carcinoma models and human breast tumors. *Genome Biol* **8**, R76.

4. Prat A, *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68.

5. Rädler PD, *et al.* Highly metastatic claudin-low mammary cancers can originate from lumin al epithelial cells. *Nat Commun* **12**, 3742.

6. Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. *Nat Commun* **11**, 1787.

7. Jordan NV, *et al.* HER2 expression identifies dynamic functional states within circulatin g breast cancer cells. *Nature* **537**, 102-106.

8. Chung W, *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profi ling in primary breast cancer. *Nat Commun* **8**, 15081.

9. Azizi E, *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Micro environment. *Cell* **174**, 1293-1308.e1236.

10. Karaayvaz M, *et al.* Unravelling subclonal heterogeneity and aggressive disease states in T NBC through single-cell RNA-seq. *Nat Commun* **9**, 3588.

11. Kim C, *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e813.

12. Savas P, *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-reside nt memory subset associated with improved prognosis. *Nat Med* **24**, 986-993.

13. Gulati GS, *et al.* Single-cell transcriptional diversity is a hallmark of developmental p otential. *Science* **367**, 405-411.

14. Bassez A, *et al.* A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nat Med* **27**, 820-832.

15. Pal B, *et al.* A single-cell RNA expression atlas of normal, preneoplastic and tumori genic states in the human breast. *EMBO J* **40**, e107333.

16. Wu SZ, *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347.

17. Barretina J, *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anti

cancer drug sensitivity. *Nature* **483**, 603-607.

18. Greer YE*, et al.* Mitochondrial Matrix Protease ClpP Agonists Inhibit Cancer Stem Cell F unction in Breast Cancer Cells by Disrupting Mitochondrial Homeostasis. *Cancer Res Commun* **2**, 1144-1161.

19. Daemen A*, et al.* Modeling precision treatment of breast cancer. *Genome Biol* **14**, R110.

20. Marcotte R*, et al.* Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabi lities, and Resistance. *Cell* **164**, 293-309.

21. Cancer Genome Atlas Research N*, et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120.

22. Gambardella G, Viscido G, Tumaini B, Isacchi A, Bosotti R, di Bernardo D. A single-cell analysis of breast cancer cell lines to study tumour het erogeneity and drug response. *Nat Commun* **13**, 1714.

23. Jovanović B*, et al.* Heterogeneity and transcriptional drivers of triple-negative breast ca ncer. *Cell Rep* **42**, 113564.

24. Hollern DP*, et al.* B Cells and T Follicular Helper Cells Mediate Response to Checkpoint I nhibitors in High Mutation Burden Mouse Models of Breast Cancer. *Cell* **179**, 1191-1206.e1121.

25. Gendoo DMA*, et al.* Genefu: an R/Bioconductor package for computation of gene expression-b ased signatures in breast cancer. *Bioinformatics* **32**, 1097-1099.

26. Pan C*, et al.* Research progress of Claudin-low breast cancer. *Front Oncol* **13**, 1226118.

27. Dias K*, et al.* Claudin-Low Breast Cancer; Clinical & Pathological Characteristics. *PLoS One* **12**, e0168669.

28. Li W, Ma H, Zhang J, Zhu L, Wang C, Yang Y. Unraveling the roles of CD44/CD24 and ALDH1 as cancer stem cell marker s in tumorigenesis and metastasis. *Scientific reports* **7**, 13856.

29. Reddy TP, Rosato RR, Li X, Moulder S, Piwnica-Worms H, Chang JC. A comprehensive overview of metaplastic breast cancer: clinical featur es and molecular aberrations. *Breast Cancer Res* **22**, 121.

30. Weigelt B*, et al.* Metaplastic breast carcinomas display genomic and transcriptomic heter ogeneity [corrected]. *Mod Pathol* **28**, 340-351.

31. Creighton CJ*, et al.* Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc Natl Acad Sci U S A* **106**, 13820-13825.

32. Pommier RM, *et al.* Comprehensive characterization of claudin-low breast tumors reflects t he impact of the cell-of-origin on cancer evolution. *Nat Commun* **11**, 3431.

33. Park S, Kwon Y. The role of ELF3 in metastasis of triple negative breast cancer. *Annals of Oncology* **28**, x34 (2017).

34. Kuang L, Li La. 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 J Biol Sci* **28**, 4137-4146.

35. Zhang X, Wang X, Wen Y, Chen S, Zhou C, Wu F. Single-cell transcriptomics reveal metastatic CLDN4+ cancer cells unde rlying the recurrence of malignant pleural effusion in patients with a dvanced non-small-cell lung cancer. *Clin Transl Med* **14**, e1649.

36. Liu K, *et al.* Evaluating cell lines as models for metastatic breast cancer through i ntegrative analysis of genomic data. *Nat Commun* **10**, 2138.

37. Park M, Kim D, Ko S, Kim A, Mo K, Yoon H. Breast Cancer Metastasis: Mechanisms and Therapeutic Implications. *Int J Mol Sci* **23**, 6806.

38. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740.

39. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among ge ne clusters. *OMICS* **16**, 284-287.

40. Dobin A, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.

41. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or wit hout a reference genome. *BMC Bioinformatics* **12**, 323.

42. Frankish A, *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942-D949.

43. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890.

44. Zheng GXY, *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049.

45. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502.

46. Stuart T, *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821.

47. Hao Y, *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529.

48. Hao Y, *et al.* Dictionary learning for integrative, multimodal and scalable single-ce ll analysis. *Nat Biotechnol* **42**, 293-304.

49. The ggplot2 package.https://cran.r-project.org/web/packages/ggplot2/index.html.).

50. The pheatmap package.https://github.com/raivokolde/pheatmap.).

51. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849.

52. Subramanian A, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpret ing genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550.

53. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7.