Single-cell molecular subtyping reveals novel intratumor heterogeneity in human Basal-like breast cancer

Jing Li[1, #], Yuan Gao[2, #], Shouhui Guo[1], Qingzhen Hou[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 molecularly heterogeneous disease composed of multiple intrinsic subtypes. Recent studies have highlighted the substantial intratumor heterogeneity of breast cancer, wherein malignant cells of distinct intrinsic subtypes co-exist within the same tumor. However, most existing subtyping methods are designed for bulk transcriptomic data and are therefore limited in their ability to resolve such intratumor heterogeneity at single-cell resolution. To address this gap, we develop UBS93, a computational framework that enables robust molecular subtyping of both bulk tumor samples and individual breast cancer cells. We rigorously validate UBS93 and demonstrate its superior performance relative to existing approaches, particularly in identifying the highly aggressive Claudin-low subtype. Applying UBS93 to single-cell RNA sequencing data from human Basal-like breast cancers, we identify the co-existence of Basal-like and Claudin-low cancer cell populations within the same tumor—a form of intratumor heterogeneity previously observed only in mouse models with genetically engineered RAS pathway alterations. Further analyses suggest that Claudin-low cancer cells originate from Basal-like population, with down-regulation of transcription factor *ELF3* playing a pivotal role in the Basal-like/Claudin-low transition. Together, our findings establish UBS93 as a powerful tool for breast cancer subtyping and uncover a previously unrecognized layer of intratumor heterogeneity in human Basal-like breast cancer.

45 **Introduction**

46 Breast cancer is a heterogeneous disease characterized by diverse genomic profiles,
47 transcriptomic alterations, and clinical outcomes. This heterogeneity presents
48 significant challenges for achieving clinically meaningful tumor classification[1]. Parker
49 *et al.* utilized microarray technology to develop an intrinsic breast cancer classifier
50 PAM50, which stratifies the disease into five subtypes: LuminalA, LuminalB, HER2-
51 enriched, Basal-like, and Normal-like[2]. This classification system provides superior
52 prognostic and predictive capabilities compared to traditional methods, such as
53 pathological staging, histological grading, standard clinical biomarkers, and so on[2]. In
54 addition to PAM50, Herschkowitz *et al.* identified the highly aggressive Claudin-low
55 subtype which is characterized by low expression levels of cell adhesion molecules
56 (e.g., *CLDN3*, *CLDN4*, *CLDN7*, and *CDH1*) and enriched for high stemness and
57 mesenchymal features[3, 4, 5]. Fougner *et al.* further validated Claudin-low as a distinct
58 subtype through integrative multi-omics analyses across multiple cohorts, although its
59 cellular origin remains unclear[6].

60

61 The use of single-cell omics technology has profoundly enhanced our understanding of
62 intratumor heterogeneity in breast cancer[7-16]. Especially, several recent studies have
63 reported the co-existence of cancer cells of multiple molecular subtypes within the same
64 tumor. For instance, Wu *et al*. identified confident Basal-like cancer cells within a
65 human estrogen receptor-positive (ER^+) breast tumor (which are usually subtyped as
66 LuminalA or LuminalB)[16]; in addition, Patrick *et al.* demonstrated that modulation of

67 the RAS signaling pathway in mouse models can induce tumors containing both Basal-
68 like and Claudin-low cancer cell populations[5]. These discoveries underscore the
69 limitations of traditional bulk omics-based subtyping strategy and highlight the
70 importance of achieving single-cell resolution for more precise classification.
71
72 Despite advances in computational subtyping, significant challenges remain in
73 delineating the intratumoral heterogeneity described above. Most existing tools are
74 optimized for bulk omics data, limiting their applicability in single-cell analyses. To
75 our knowledge, SCSubtype is the only available tool which enables molecular
76 subtyping of breast cancer at single-cell resolution; however, it overlooks the highly
77 aggressive Claudin-low subtype. Addressing these limitations is critical for refining
78 breast cancer subtyping and guiding precise diagnostic and therapeutic interventions.
79
80 In this study, we developed a computational framework (UBS93) which enables
81 molecular subtyping of both bulk tumor samples and single cancer cells. Applying
82 UBS93 to single-cell RNA sequencing (scRNA-seq) data, we identified a previously
83 unrecognized form of intratumor heterogeneity in human Basal-like breast cancer,
84 characterized by co-existence of Basal-like and Claudin-low cancer cell populations
85 within the same tumor. Further analysis suggested that Claudin-low cancer cells arise
86 from Basal-like populations, with downregulation of the transcription factor *ELF3* can
87 induce a partial Claudin-low phenotype.
88

Results

Development of the UBS93 classifier for breast cancer subtyping

Solid tumor tissue is a heterogeneous mixture of various cell types[16]. In bulk RNA sequencing, the measured expression of a gene reflects a weighted average of its expression across all constituent cells. Specifically, for a given gene g, if its average expression in cancer cells and non-cancer cells is denoted by C_g and M_g, respectively, and the tumor purity is α, the observed expression in a bulk sample can be approximated as $\alpha * C_g+(1-\alpha) * M_g$. As a result, gene expression rankings in a bulk tumor sample may deviate substantially from those in pure cancer cells due to the influence of the tumor microenvironment **(Figure 1a, left panel)**; however, for genes highly specific to cancer cells (i.e., $M_g \approx 0$), the rankings in a bulk tumor sample and pure cancer cells are expected to remain highly concordant **(Figure 1a, right panel)**. This observation suggests that the relative expression ranks of cancer cell-specific genes may serve as robust features for molecular subtyping.

To identify such genes, we analyzed the single-cell RNA sequencing (scRNA-seq) dataset from Wu *et al.*'s study and computed "C-M ratio" value for each gene, representing its expression specificity to cancer cells. Among the top 100 genes with the highest C-M ratios, seven lacked ENSEMBL IDs and were excluded. The remaining 93 genes formed a uniform gene panel for breast cancer subtyping, termed UBS93 (**Supplementary Table 1**). As expected, UBS93 genes exhibited significantly higher expression in cancer cells compared to non-cancer cells (**Figure 1b**).

112 To validate this gene panel, we utilized breast cancer cell line data from the Cancer Cell
113 Line Encyclopedia (CCLE), which classifies cell lines into five subtypes: basal_A,
114 basal_B, luminal, HER2-amplified (HER2-amp), and luminal-HER2-amp[17]. A t-
115 distributed stochastic neighbor embedding (t-SNE) analysis of CCLE bulk RNA-seq
116 data revealed clustering patterns consistent with these subtypes (**Figure 1c**), supporting
117 the validity of the CCLE classification. For consistency with widely used breast cancer
118 nomenclature, we renamed basal_A and basal_B as Basal-like and Claudin-low,
119 respectively, and excluded the luminal-HER2-amp subtype from further analysis. This
120 yielded four refined subtypes: Basal-like, Claudin-low, Luminal, and HER2-amp. We
121 then performed t-SNE using only the UBS93 gene set on the same CCLE dataset
122 (**Supplementary Figure 1**). The resulting subtype-specific clusters mirrored those
123 from the full transcriptome analysis, confirming the utility of UBS93 in distinguishing
124 breast cancer subtypes.
125
126 Building on these observations, we developed our UBS93 subtype classifier **(Figure**
127 **1d**). We first created a 93 x 3 subtype centroid matrix, where each column
128 represented the average expression profile of UBS93 genes across the CCLE cell lines
129 corresponding to Basal-like, Claudin-low, Luminal or HER2-amp, respectively. For a
130 given bulk RNA-seq or scRNA-seq gene expression profile, we computed its Spearman
131 rank correlation with each subtype centroid and assigned it to the subtype with the
132 highest correlation. If the sample is subtyped as Luminal or HER2-amp, we further

133 calculated an enrichment score based on HER2-amp signature genes to determine the
134 final subtype assignment.
135
136 **Evaluation of UBS93 classifier in human and mice**
137 We first evaluated the performance of the UBS93 classifier using an assembled dataset
138 which contains the bulk RNA-seq profile of 47 breast cancer cell lines extracted from
139 the Gene Expression Omnibus (GEO) database[18, 19, 20]**(Supplementary Table 2)**. We
140 found 93.6% of the samples were distributed on the diagonal of the confusion matrix,
141 suggesting the high overall accuracy of our method (**Figure 2a**). Notably, our classifier
142 also demonstrated optimal subtype-specific performance and the F1 scores for the four
143 subtypes were 0.97 (Basal-like), 0.80 (Claudin-low), 1.00 (HER2-amp), and 0.92
144 (Luminal), respectively. In addition to cell lines, we also evaluated our classifier on the
145 breast cancer samples from The Cancer Genome Atlas (TCGA) and derived ideal
146 overall accuracy (81.9%) and subtype-specific performance[21] (**Figure 2b**).
147
148 Next, we validated our classifier using an assembled scRNA-seq dataset of human
149 breast cancer cell lines (**Supplementary Table 3**)[22, 23]. For each cell line, we computed
150 the accuracy as the proportion of cells that were correctly subtyped. Our classifier
151 achieved an accuracy higher than 80% in 19 of the 30 cell lines (63.3%), with a median
152 accuracy of 85.7%. The subtype-specific median accuracy values were 99.0% for
153 Basal-like, 98.3% for Claudin-low, 78.0% for HER2-amp, and 75.7% for Luminal
154 (**Figure 2c**).
155

156 We also validated the classifier on a dataset of 27 bulk RNA-seq samples from murine-
157 derived studies. This dataset included 17 samples from Hollern's study[24], which were
158 experimentally confirmed as Claudin-low, and 10 samples from Rädler's study[5], with
159 unknown subtype labels. Based on hierarchical clustering of five "Claudin-low
160 indicative" genes (*Cldn3, Cldn13, Cldn7, Vim*, and *Epcam*), we imputed the missing
161 subtype labels (**Figure 2d**). The UBS93 classifier accurately assigned subtypes to all
162 27 samples, with results aligning with the known subtypes (**Supplementary Table 4**).
163
164 **Comparison to existing methods**
165 We systematically compared UBS93 classifier with genefu, a widely used R package
166 for molecular subtyping of breast cancer[25]. Notably, only six genes were shared
167 between UBS93 and PAM50 panels (**Supplementary Figure 2a**); moreover, UBS93
168 genes exhibited significantly higher C-M-ratio values and cancer cell specificity
169 compared to PAM50 genes (**Supplementary Figure 2b**).
170
171 When comparing classifier performance using bulk RNA-seq data from TCGA breast
172 cancer samples, genefu achieved a slightly higher overall accuracy (83.7%). However,
173 UBS93 outperformed genefu in classifying Basal-like, Claudin-low, and HER2-amp
174 subtypes (**Figure 3a**). In murine datasets, genefu misclassified all Basal-like samples
175 as Claudin-low, whereas UBS93 accurately identified them as Basal-like
176 (**Supplementary Table 5**).
177

178 In addition to genefu, we also compared UBS93 with SCSubtype, a method designed
179 for subtyping individual breast cancer cells[16], using the same scRNA-seq dataset used
180 to evaluate our classifier. Since SCSubtype does not include the Claudin-low subtype,
181 the comparison focused on the other three subtypes. UBS93 demonstrated significantly
182 higher accuracy than SCSubtype for the Basal-like subtype (**Figure 3b**).
183
184 **UBS93 uncovers novel intratumor heterogeneity in Basal-like breast cancer**
185 Prior to our study, Wu *et al.* identified Basal-like cancer cells within human ER^+ breast
186 tumors, highlighting the presence of intratumor heterogeneity in ER^+ breast cancer[16].
187 However, the intratumoral heterogeneity of triple-negative breast cancer (TNBC)
188 remains largely underexplored. To investigate this, we re-examined UBS93 subtyping
189 results derived from the scRNA-seq datasets of 12 established Basal-like breast cancer
190 cell lines. In 10 of these lines, over 80% of single cells were classified as Basal-like,
191 consistent with subtyping outcomes derived from bulk RNA-seq data. Strikingly, in the
192 canonical Basal-like cell line HDQP1, only 56.4% of cells were classified as Basal-like,
193 while the remaining cells were subtyped as Claudin-low (**Figure 2c**). These findings
194 led us to hypothesize that, despite being classified as Basal-like based on bulk
195 transcriptomic profile, HDQP1 in fact consists of a mixture of Basal-like and Claudin-
196 low cancer cell populations.
197
198 To test this hypothesis, we performed t-SNE visualization of the HDQP1 scRNA-seq
199 data and observed two distinct clusters (**Figure 4a**). One cluster consisted exclusively
200 of Basal-like cells, while 84.2% of the cells in the other cluster were subtyped as

201 Claudin-low. Differential gene expression (DE) analysis suggested that the UBS93-
202 identified Claudin-low cells exhibited canonical features of Claudin-low breast cancer[4,
203 5, 6, 26, 27]. These included: (1) reduced expression of cell adhesion genes (e.g., *CLDN4*,
204 *CLDN7*, *KRT19,* and *EPCAM*) (**Figure 4b, Supplementary Figure 3, and**
205 **Supplementary Table 6**); (2) elevated epithelial-to-mesenchymal transition (EMT)
206 level, evidenced by up-regulation of *VIM* and the enrichment of EMT signature genes
207 among up-regulated DE genes (**Figure 4c**); and (3) heightened stemness, as indicated
208 by increased *CD44* (marker of cancer stem cell) expression and higher *CD44*/*CD24*
209 ratio[4, 27, 28](**Figure 4d**).
210
211 To further validate these findings, we generated a pseudo-bulk transcriptomic profile
212 using Claudin-low cells and computed its transcriptomic correlation with breast cancer
213 cell lines from the CCLE project[17]. CAL120—a well-established Claudin-low cell
214 line—exhibited the highest correlation, with four of the top five most correlated lines
215 being Claudin-low **(Figure 4e)**. Conversely, pseudo-bulk analysis of Basal-like cells
216 yielded HDQP1 as the top correlated cell line, with all five top correlates being Basal-
217 like (**Figure 4f**). To double confirm our finding, we replicated our analysis on an
218 independent HDQP1 scRNA-seq data (GSE202771) and derived the same conclusion
219 (**Supplementary Figure 4, 5 and Supplementary Table 7**). Therefore, our
220 computational analysis suggested that the Basal-like cell line HDQP1 was an admixture
221 of Basal-like and Claudin-low cell populations.
222

223 To our knowledge, such a "Basal-like/Claudin-low admixture" phenotype has been
224 previously reported only in mouse models following genetic modulation of the RAS
225 signaling pathway[5], with no prior evidence in human Basal-like breast cancer cell lines
226 and patient samples. To investigate whether this phenotype also exists in TNBC patient
227 tumors, we applied the UBS93 classifier to the scRNA-seq data from 50 TNBC samples.
228 These samples were then ranked based on the median transcriptomic correlation of
229 Claudin-low cells with the Claudin-low reference centroid (**Figure 4g, Supplementary**
230 **Table 8**). As expected, the metaplastic breast cancer patient CID4513 (from dataset
231 GSE176078) ranked highest, which may not be surprising since previous studies have
232 confirmed that metaplastic breast cancer exhibits characteristics of Claudin-low
233 subtype[29, 30]. Intriguingly, the patient 0554 (from dataset GSE161529), who was
234 classified as Basal-like in the literature, ranked second[15]. Of the cancer cells in this
235 patient, 91.1% were subtyped as Basal-like, while 7.9% were Claudin-low. Similar to
236 HDQP1, Claudin-low cells tended to form a distinct cluster on the t-SNE plot (**Figure**
237 **5a**), and meanwhile exhibited canonical features of Claudin-low breast cancer: down-
238 regulation of adhesion genes, enrichment of EMT signatures, and elevated stemness
239 markers (**Figure 5b, 5c, 5d, Supplementary Figure 6, and Supplementary Table 9)**.
240 In addition, pseudo-bulk transcriptomic correlation analysis confirms their resemblance
241 to known Claudin-low breast cancer cell lines (**Figure 5e, 5f**).
242
243 In summary, our integrative analyses reveal a previously unrecognized form of
244 intratumoral heterogeneity in human Basal-like breast cancer, characterized by the co-

245 existence of Basal-like and Claudin-low cancer cells. These findings highlight the
246 capability of UBS93 to detect subtle but biologically meaningful substructures in breast
247 cancer and underscore its potential utility for studying tumor evolution and plasticity.
248

249 ***ELF3* knockdown induces a partial Claudin-low phenotype in Basal-like breast**
250 **cancer**

251 To investigate the underlying cause of the observed heterogeneity, we first examined
252 the cellular origin of Claudin-low cancer cells. Pommier *et al.* conducted large-scale
253 computational analyses of bulk cancer genomics data and proposed three potential
254 origins: direct malignant transformation from normal mammary stem cells, or trans-
255 differentiation from either Luminal or Basal-like cancer cells[32]. However, robust single-
256 cell level evidence from patient samples has been lacking. The identification of Basal-
257 like/Claudin-low admixture in our study provides a unique opportunity to revisit this
258 question. Using copy number variation (CNV)-based hierarchical clustering in patient
259 0554, we observed that Claudin-low cancer cells were most closely related to Basal-
260 like cancer cells (**Figure 6a**), supporting the hypothesis that Claudin-low cancer cells
261 may arise from a Basal-like lineage.
262

263 To identify genes which may drive the Basal-like/Claudin-low transition, we
264 intersected DE gene lists from both the HDQP1 cell line and patient 0554, revealing 18
265 commonly up-regulated and 37 commonly down-regulated genes (**Supplementary**
266 **Table 10**). Among these, the transcription factor *ELF3* emerged as a gene of interest
267 for two reasons **(Supplementary Figure 7).** First**,** overexpression of *ELF3* in Claudin-

268 low cancer cell lines (MDA-MB-231 and BT549) inhibits the EMT process[33]; second,
269 *ELF3* promotes *CLDN4* expression in ovarian cancer and non-small cell lung cancer[34,
270 35]. These findings prompted us to test whether depletion of *ELF3* in Basal-like cancer
271 cells might induce a Claudin-low phenotype.
272
273 We selected HCC70, a cell line with a transcriptomic profile closely resembling Basal-
274 like breast cancer[36], as our experimental model. scRNA-seq followed by UBS93
275 subtyping confirmed that 98.7% of untreated HCC70 cells were classified as Basal-like.
276 We then performed *ELF3* knockdown in HCC70 and repeated scRNA-seq to determine
277 whether this perturbation could promote Claudin-low features. UBS93 analysis showed
278 that 98.1% of cells remained Basal-like, suggesting that *ELF3* knockdown alone is
279 insufficient to trigger a complete subtype switch.
280
281 Despite failing to induce complete Basal-like/Claudin-low transition, we asked whether
282 *ELF3* knockdown could induce partial Claudin-low features. To account for variable
283 knockdown efficiency, we selected the 100 cells with the lowest *ELF3* expression (sh-
284 *ELF3* group) and the 100 cells with the highest *ELF3* expression (control group) for
285 further analysis. Interestingly, the sh-*ELF3* group exhibited a significantly lower
286 correlation with the Basal-like centroid compared to the control group (**Supplementary**
287 **Figure 8a**) and DE analysis between these two groups identified 210 up-regulated and
288 79 down-regulated genes, respectively (**Figure 6b, Supplementary Table 11**). As
289 anticipated, *CLDN4* was significantly down-regulated (p-value = 3.57e-14), as was the

290 epithelial cell marker *KRT19* (p-value = 3.59e-13). Interestingly, when examining the
291 expression pattern of down-regulated DE genes across CCLE breast cancer cell lines,
292 several exhibited Claudin-low specific depletion (**Figure 6c**). We curated a set of such
293 genes (which we termed as CLSD genes) and gene set enrichment analysis revealed
294 significant enrichment of CLSD genes among the down-regulated DE genes
295 (**Supplementary Figure 8b**).
296
297 In addition to scRNA-seq, we also used qRT-PCR to assess how *ELF3* knockdown
298 influenced the expression of *CLDN3*, *CLDN4*, *CLDN7*, *EPCAM*, and *VIM* (**Figure 6d**).
299 Consistent with scRNA-seq results, *CLDN4* was significantly down-regulated upon
300 *ELF3* knockdown (p-value = 0.002254), and *CLDN7* also showed reduced expression,
301 despite not meeting DE gene thresholds (p-value = 0.018557). No significant changes
302 were observed in *CLDN3* and *EPCAM*. Unexpectedly, *VIM* expression was down-
303 regulated (p-value = 0.003883), which contrasts with the reported role of *ELF3* as a
304 suppressor of EMT and warrants further investigation. Replicate experiments
305 confirmed these findings (**Supplementary Figure 9**).
306
307 Together, these results suggest that *ELF3* knockdown induces a partial Claudin-low
308 phenotype in Basal-like cancer cells, as evidenced by the down-regulation of CLSD
309 genes such as *CLDN4* and *CLDN7*.
310
311

Discussion

In this study, we introduced a novel classifier UBS93 for the molecular subtyping of breast cancer. We validated its performance using both bulk and single-cell RNA sequencing data from human and mouse models. Notably, UBS93 outperformed classical methods, particularly in the identification of Claudin-low subtypes.

Using UBS93, we uncovered a previously unrecognized form of intratumoral heterogeneity in human Basal-like breast cancer, characterized by the co-existence of Basal-like and Claudin-low cancer cell populations within the same tumor. In our opinion, this finding carries three major implications. First, it advances our understanding of Basal-like breast cancer biology. Previously, such heterogeneity had only been reported in mouse models following genetic modulation of the RAS signaling pathway[5]. Whether this phenotype also occurs in humans remained unclear, given the substantial molecular differences between human and murine models. Our study provides direct evidence that this heterogeneity exists in human Basal-like breast tumors. Second, our findings offer compelling support for the hypothesis that Claudin-low cancer cells may originate from a Basal-like lineage, an idea previously proposed by Pommier et al. but lacking conclusive evidence. Third, our findings highlight the necessity of incorporating all intrinsic subtypes—including the Claudin-low subtype—into molecular subtyping frameworks, particularly in single-cell analyses. Notably, although Wu et al. introduced the SCSubtype classifier, its omission of the Claudin-low

333 subtype led to the failure to detect this heterogeneity in human Basal-like scRNA-seq
334 datasets.
335 Beyond the advancements in computational subtyping, our research has raised several
336 intriguing scientific questions that warrant further investigation.
337
338 First, besides *ELF3* knockdown, what additional genetic perturbations are needed to
339 fully convert Basal-like cancer cells into Claudin-low cancer cells? While high
340 epithelial-to-mesenchymal transition (EMT) activity is a hallmark of Claudin-low cells,
341 *ELF3* knockdown alone did not significantly up-regulate mesenchymal markers such
342 as *VIM* or *FN1*. Interestingly, *ZEB1*—a key EMT regulator—was consistently up-
343 regulated in the DE analyses comparing Claudin-low and Basal-like populations. This
344 prompts the hypothesis that combined *ELF3* knockdown and *ZEB1* overexpression
345 may be necessary to induce a complete Claudin-low phenotype.
346
347 Second, is *KRAS* mutation necessary for generation of Claudin-low cancer cells? Rädler
348 *et al.* induced the tumor of "Basal-like/Claudin-low admixture" through activation of
349 endogenous *KRAS* mutations (or exogenous *KRAS* expression). However, *KRAS*
350 mutation is rare in human breast cancer. Using cBioPortal, we only identified six TCGA
351 breast cancer patients carrying *KRAS* mutation; in addition, the cell line HDQP1 does
352 not carry KRAS mutation either. These observations suggest that there might be *KRAS*
353 independent mechanisms which may also give rise of Claudin-low cancer cells.
354

355 Third, what are the clinical implications of the “Basal-like/Claudin-low admixture”?
356 Given that distant metastasis is the primary cause of cancer-related mortality and that
357 Claudin-low tumors are associated with greater aggressiveness, it is plausible that the
358 presence of such a phenotype may correlate with poor prognosis in Basal-like breast
359 cancer. In future, we will integrate single-cell RNA-seq data with clinical outcome
360 information to test this hypothesis.
361
362 Finally, does the presence of “Basal-like/Claudin-low admixture” alter the tumor
363 microenvironment? Fougner *et al.* reported that Claudin-low tumors are characterized
364 by high lymphocyte infiltration and elevated PD-L1 expression. We therefore
365 hypothesize that Claudin-low cells within the tumors may shape distinct immune niches.
366 To explore this, we will apply the UBS93 classifier to spatial transcriptomics datasets
367 to identify regions enriched with Claudin-low cells and compare their surrounding
368 microenvironments to those of Basal-like regions. This analysis may offer insights into
369 immunotherapeutic vulnerabilities in this heterogeneous Basal-like breast cancer.
370
371 In conclusion, our study introduces a robust and versatile tool for delineating
372 intratumoral heterogeneity in breast cancer. As high-quality single-cell transcriptomic
373 data continue to emerge across diverse tumor types, we anticipate that this strategy will
374 be broadly applicable to other cancers. By enabling precise dissection of tumor
375 subpopulations, our approach holds promise for uncovering novel biological
376 mechanisms that drive tumor progression and therapeutic resistance.

377 # Materials and Methods

378 **Development of UBS93 gene panel**

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

382 $$\text{C-M-ratio} = (C_g + 0.1) / (M_g + 0.1)$$

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

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

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

396

397 **Determination of breast cancer subtypes**

398 The CCLE categorized breast cancer into five subtypes: basal_A, basal_B, Luminal,
399 HER2-amp, and luminal-HER2-amp. In the t-SNE plot (**Figure 1c**), basal_A and
400 basal_B form their own distinct cluster while the other three subtypes form the third
401 cluster. The basal_A cluster is significantly enriched with 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

427 The EMT signature was downloaded from MSigDB[38]. To derive the Claudin-low-
428 specific depletion (CLSD) gene set, z-score normalization was applied to gene
429 expression values of CCLE breast cancer cell lines. For each gene, the median
430 expression value across Claudin-low cell lines was computed. If this value is below -1,
431 then the gene was considered as a CLSD gene. In total, this approach identified 960
432 CLSD genes. The clusterProfiler R package was used to compute enrichment score and
433 p-value[39].
434

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

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

441 **The t-SNE analysis**

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

447 **Cell lines and cell culture conditions**

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

451 SH30285.01, Cytiva Life Sciences). All medium supplemented with 10% FBS (catalog
452 no. 164210-50, Pricella Biotechnology).
453
454 **Lentivirus packaging and infection**
455 HEK293T cells (2×10^6) were seeded in a 100-mm cell culture dish and transfected
456 with plasmids using PolyJet in vitro DNA transfection reagent PEI (C0537-20mg,
457 Beyotime Biotechnology). The transfection experiments were operated according to the
458 manufacturer's instructions. For knockdown target proteins, we co-transfected the cells
459 with the plasmids of pMDL, pVSVG, pREV and pLKO.1 (sh-*ELF3*, scrambled) to
460 generate lentivirus. For lentivirus infection, 500 μl of virus-containing medium was
461 added to the 5×10^4 cells with 1% polybrene.
462
463 **Quantitative Real-time PCR**
464 Total RNA was extracted from cells using an RNAsimple Total RNA Kit (TIANGEN,
465 DP419). Complementary DNA was prepared using the High-Capacity cDNA Reverse
466 Transcription Kit purchased from Vazyme Biotech (catalog no. R333–01). Real-time
467 PCR analysis was performed using 2 × Universal SYBR Green Fast qPCR Mix (catalog
468 no. RK21203, Abclonal Technology) with the following conditions: 3 min at 95 °C,
469 and then 40 cycles at 95 °C for 5 s, 60 °C for 25 s, using a Real-Time System (Roche
470 LightCycler® 480). Data were normalized according to expression of a control gene
471 (*ACTB*) in each experiment. The data are presented as the mean ± s.d. from three
472 independent experiments. The p-values were computed with unpaired t-test in
473 GraphPad Prism 10 software.
474
475 **Single-cell RNA-sequencing of HCC70**

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

485 **Software tools and statistical methods**

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

493 **Data Availability**

494 The authors declare that all data supporting the findings of this study are available
495 within the article and its supplementary information files or from the corresponding
496 author upon reasonable request.
497

498 **Code Availability**

499 The code is available at github (https://github.com/bioklab/UBS93).
500

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

mammary 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 luminal 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 circulating breast cancer cells. *Nature* **537**, 102-106.

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

9. Azizi E, *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *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-resident 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 potential. *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 tumorigenic 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 Function 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, Vulnerabilities, 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 heterogeneity and drug response. *Nat Commun* **13**, 1714.

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

24. Hollern DP, *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**, 1191-1206.e1121.

25. Gendoo DMA, *et al.* Genefu: an R/Bioconductor package for computation of gene expression-based 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 features and molecular aberrations. *Breast Cancer Res* **22**, 121.

30. Weigelt B, *et al.* Metaplastic breast carcinomas display genomic and transcriptomic heterogeneity [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 the 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 integrative 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 gene 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 without 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-cell 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 interpreting 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.

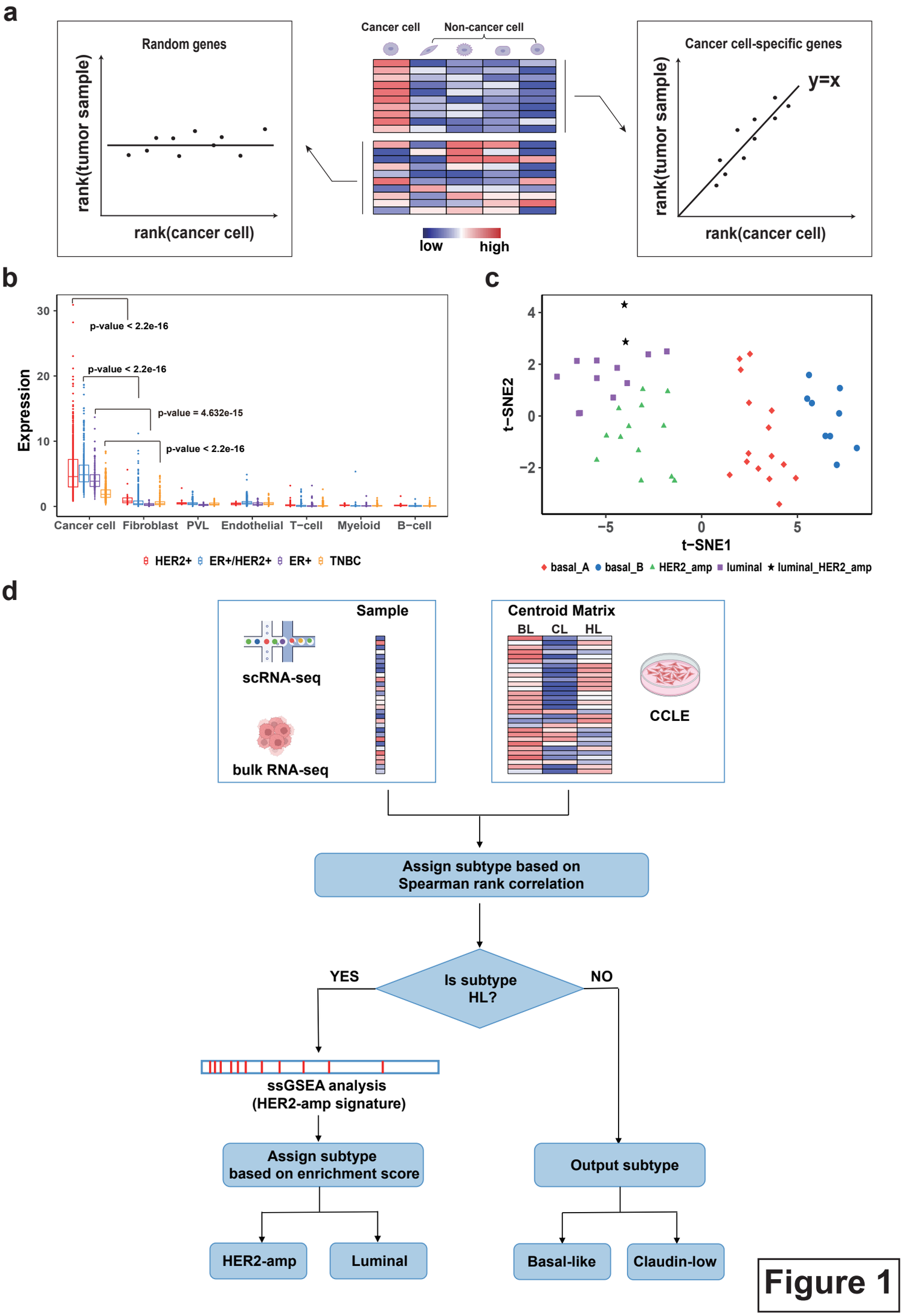
a
Random genes
rank(tumor sample)
rank(cancer cell)
Cancer cell
Non-cancer cell
low
high
Cancer cell-specific genes
y=x
rank(tumor sample)
rank(cancer cell)
b
Expression
p-value < 2.2e-16
p-value < 2.2e-16
p-value = 4.632e-15
p-value < 2.2e-16
Cancer cell
Fibroblast
PVL
Endothelial
T−cell
Myeloid
B−cell
HER2+
ER+/HER2+
ER+
TNBC
c
t−SNE2
t−SNE1
basal_A
basal_B
HER2_amp
luminal
luminal_HER2_amp
d
Sample
scRNA-seq
bulk RNA-seq
Centroid Matrix
BL
CL
HL
CCLE
Assign subtype based on Spearman rank correlation
Is subtype HL?
YES
NO
ssGSEA analysis (HER2-amp signature)
Assign subtype based on enrichment score
Output subtype
HER2-amp
Luminal
Basal-like
Claudin-low

Figure 1

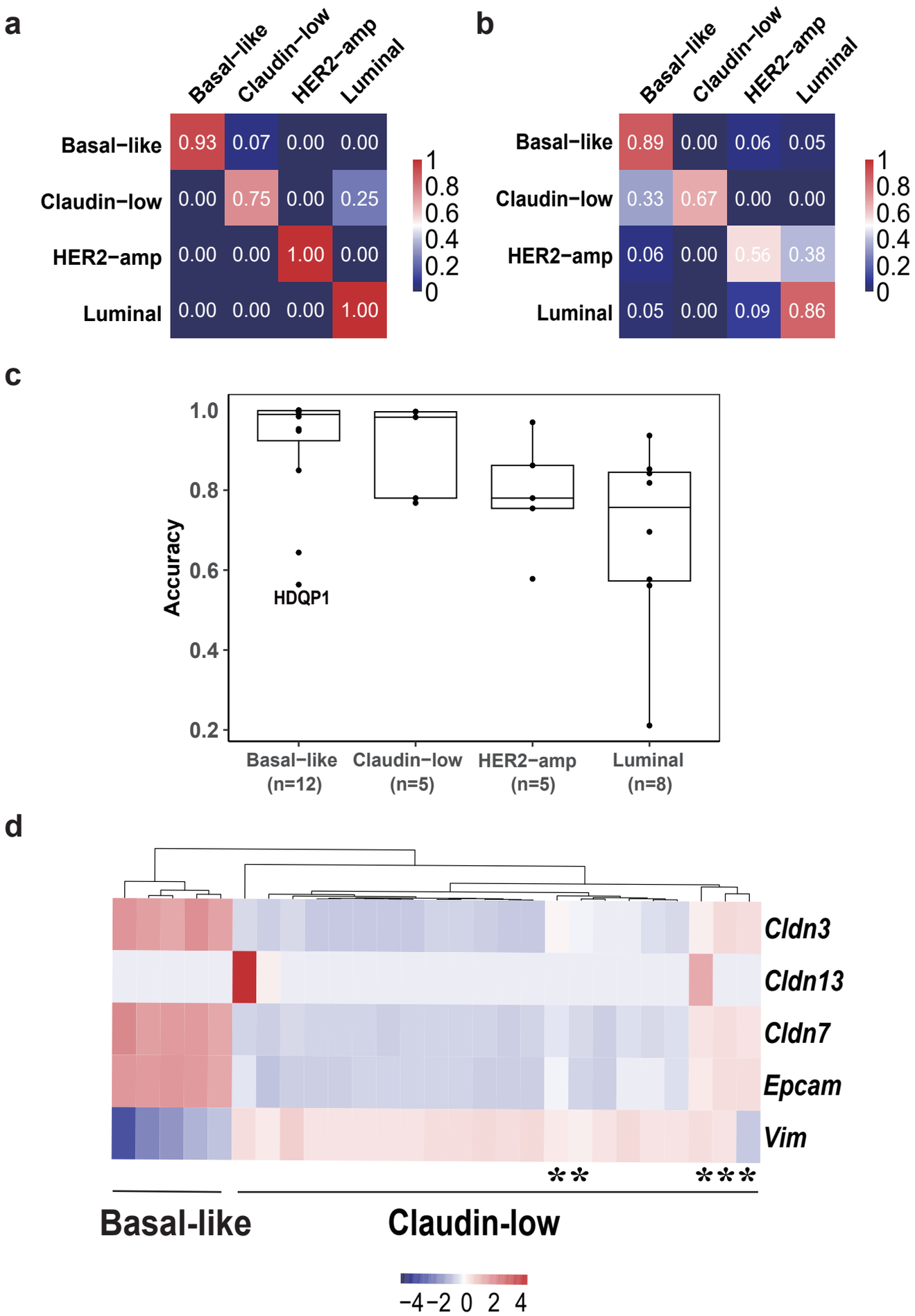
a
Basal-like
Claudin-low
HER2-amp
Luminal
0.93 0.07 0.00 0.00
0.00 0.75 0.00 0.25
0.00 0.00 1.00 0.00
0.00 0.00 0.00 1.00
1 0.8 0.6 0.4 0.2 0
b
Basal-like
Claudin-low
HER2-amp
Luminal
0.89 0.00 0.06 0.05
0.33 0.67 0.00 0.00
0.06 0.00 0.56 0.38
0.05 0.00 0.09 0.86
1 0.8 0.6 0.4 0.2 0
c
Accuracy
1.0
0.8
0.6
0.4
0.2
HDQP1
Basal-like (n=12)
Claudin-low (n=5)
HER2-amp (n=5)
Luminal (n=8)
d
Cldn3
Cldn13
Cldn7
Epcam
Vim
**
***
Basal-like
Claudin-low
−4 −2 0 2 4

Figure 2

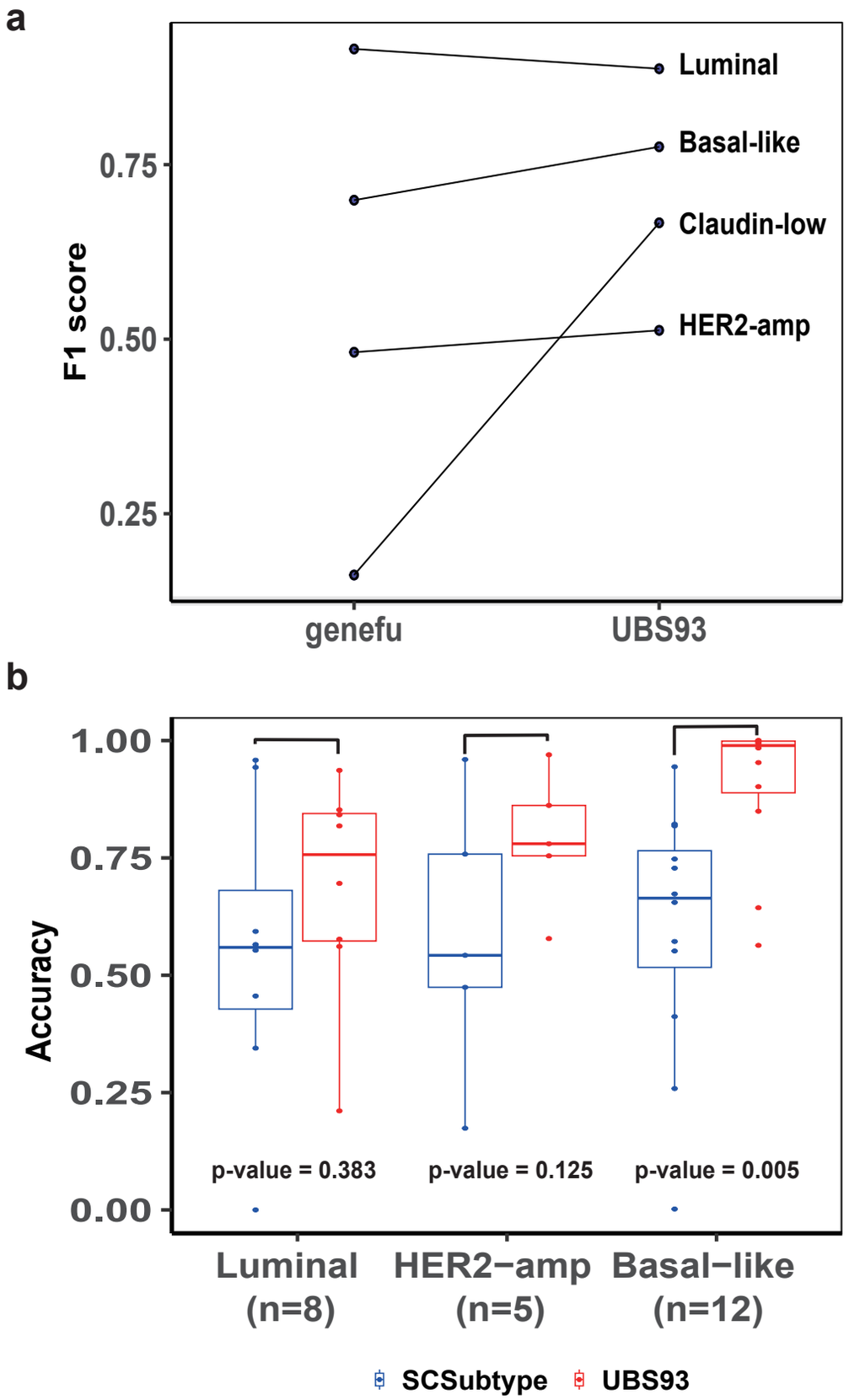
a
F1 score
0.75
0.50
0.25
Luminal
Basal-like
Claudin-low
HER2-amp
genefu
UBS93
b
Accuracy
1.00
0.75
0.50
0.25
0.00
p-value = 0.383
p-value = 0.125
p-value = 0.005
Luminal (n=8)
HER2−amp (n=5)
Basal−like (n=12)
SCSubtype
UBS93

Figure 3

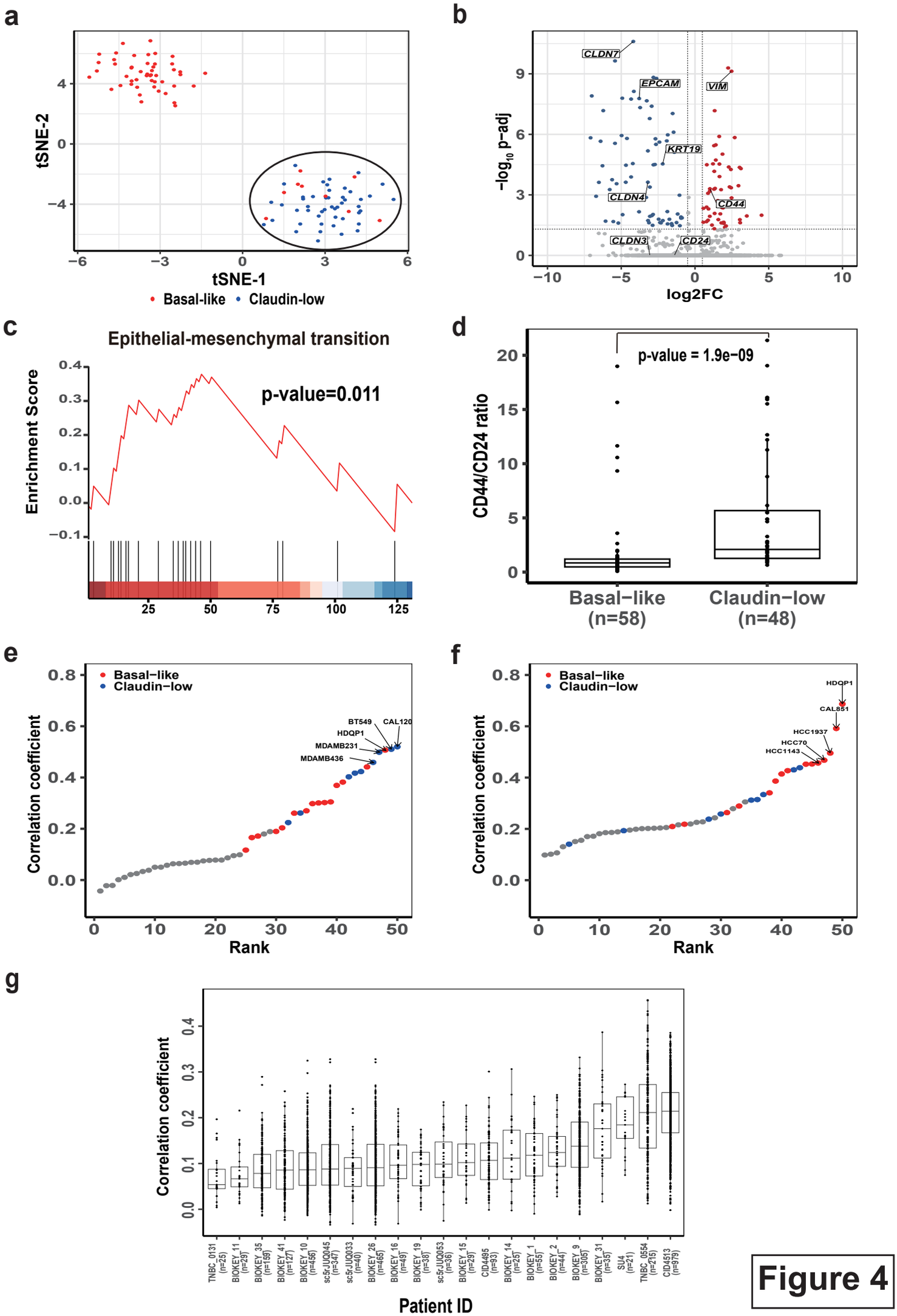
a
tSNE-2
tSNE-1
Basal-like
Claudin-low
b
CLDN7
EPCAM
VIM
KRT19
CLDN4
CD44
CLDN3
CD24
−log10 p−adj
log2FC
c
Epithelial-mesenchymal transition
p-value=0.011
Enrichment Score
d
p-value = 1.9e−09
CD44/CD24 ratio
Basal−like
(n=58)
Claudin−low
(n=48)
e
Basal−like
Claudin−low
BT549
CAL120
HDQP1
MDAMB231
MDAMB436
Correlation coefficient
Rank
f
Basal−like
Claudin−low
HDQP1
CAL851
HCC1937
HCC70
HCC1143
Correlation coefficient
Rank
g
Correlation coefficient
TNBC_0131 (n=25)
BIOKEY_11 (n=29)
BIOKEY_35 (n=159)
BIOKEY_41 (n=127)
BIOKEY_10 (n=456)
sc5rJUQ045 (n=347)
sc5rJUQ033 (n=40)
BIOKEY_26 (n=485)
BIOKEY_16 (n=49)
BIOKEY_19 (n=38)
sc5rJUQ053 (n=36)
BIOKEY_15 (n=29)
CID4495 (n=93)
BIOKEY_14 (n=25)
BIOKEY_1 (n=55)
BIOKEY_2 (n=44)
BIOKEY_9 (n=305)
BIOKEY_31 (n=35)
SU4 (n=21)
TNBC_0554 (n=215)
CID4513 (n=979)
Patient ID

Figure 4

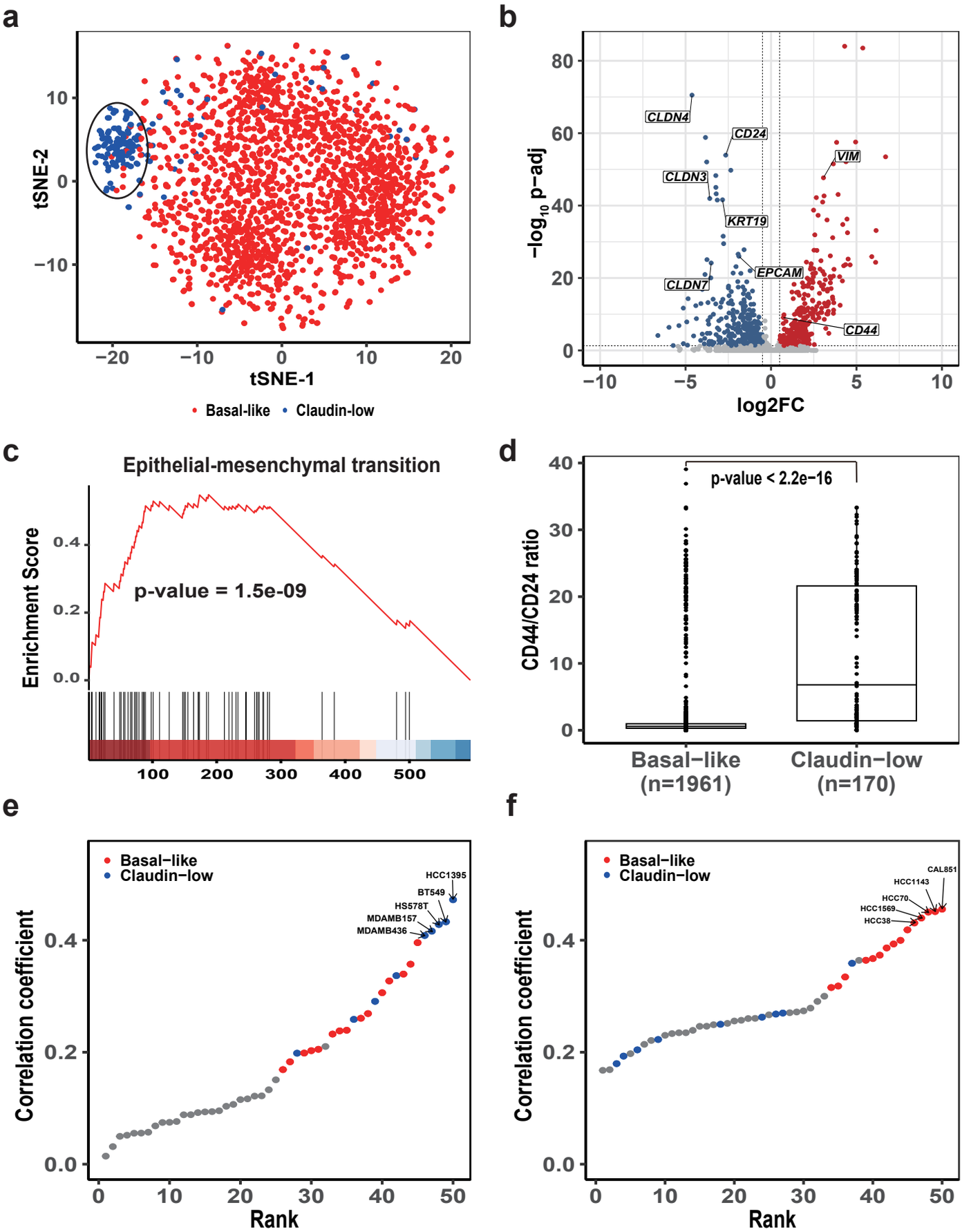
a
tSNE-2
tSNE-1
Basal-like
Claudin-low
b
CLDN4
CD24
CLDN3
KRT19
CLDN7
EPCAM
VIM
CD44
−log10 p−adj
log2FC
c
Epithelial-mesenchymal transition
p-value = 1.5e-09
Enrichment Score
d
p-value < 2.2e−16
CD44/CD24 ratio
Basal−like
(n=1961)
Claudin−low
(n=170)
e
Basal−like
Claudin−low
HCC1395
BT549
HS578T
MDAMB157
MDAMB436
Correlation coefficient
Rank
f
Basal−like
Claudin−low
CAL851
HCC1143
HCC70
HCC1569
HCC38
Correlation coefficient
Rank

Figure 5

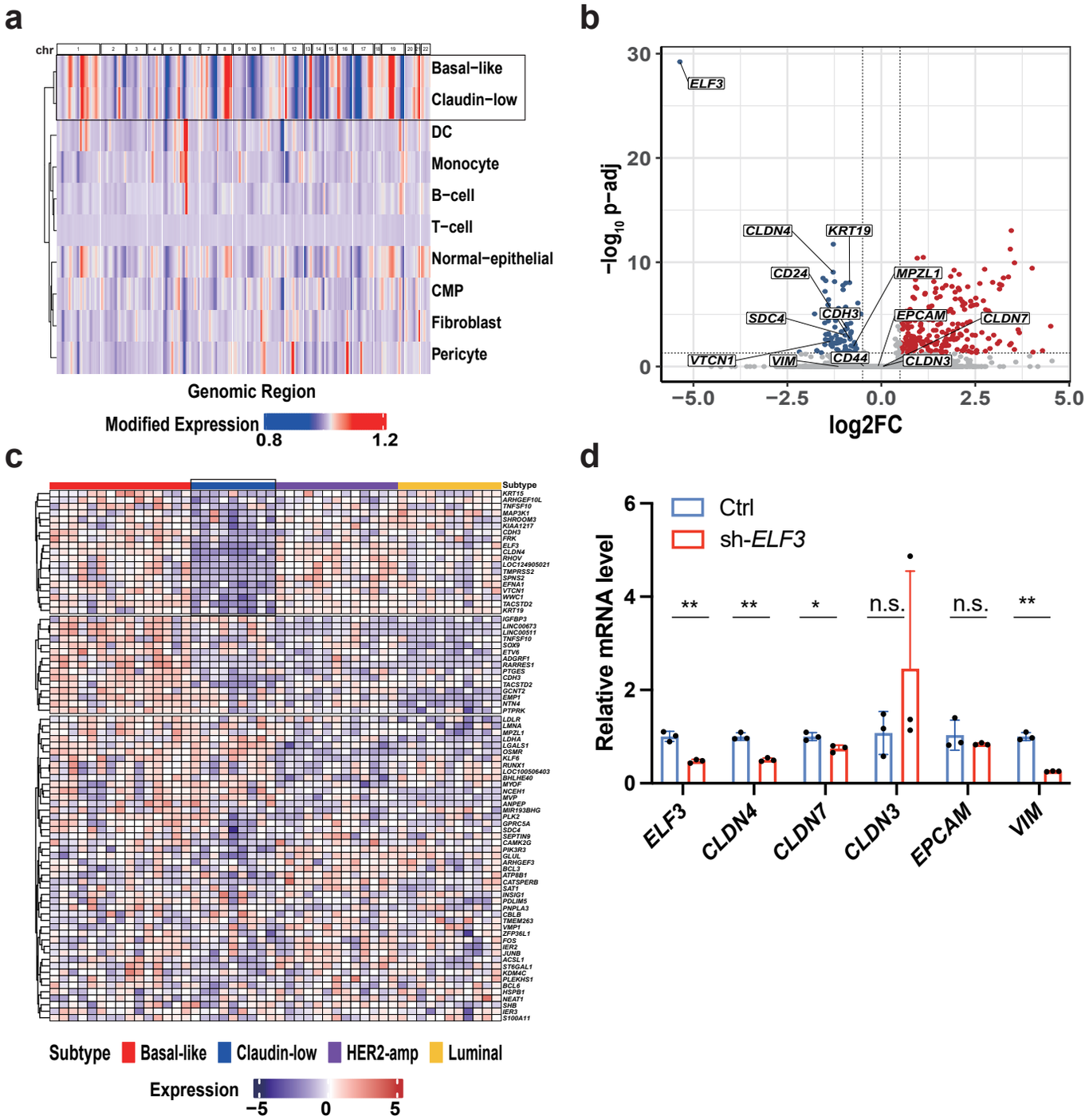
a
chr
Basal-like
Claudin-low
DC
Monocyte
B-cell
T-cell
Normal-epithelial
CMP
Fibroblast
Pericyte
Genomic Region
Modified Expression
0.8
1.2
b
ELF3
CLDN4
KRT19
CD24
MPZL1
SDC4
CDH3
EPCAM
CLDN7
VTCN1
VIM
CD44
CLDN3
−log10 p−adj
log2FC
c
Subtype
Subtype
Basal-like
Claudin-low
HER2-amp
Luminal
Expression
−5
0
5
d
Ctrl
sh-ELF3
Relative mRNA level
**
**
*
n.s.
n.s.
**
ELF3
CLDN4
CLDN7
CLDN3
EPCAM
VIM

Figure 6

Figure 1: Development of the UBS93 classifier

(a) Schematic representation illustrating the rationale for utilizing cancer cell-specific genes in molecular subtyping.

(b) Average expression values of UBS93 genes across various cell types in breast cancer patients. In the boxplots, each dot represents a single cell, the central line indicates the median, and the bounds represent the 25th and 75th percentiles (interquartile range). Whiskers extend to 1.5 times the interquartile range. P-values were calculated using the two-sided Wilcoxon rank-sum test.

(c) The t-SNE visualization of CCLE breast cancer cell lines, colored by molecular subtypes.

(d) Workflow diagram for the UBS93 classifier. BL: Basal-like, CL: Claudin-low, HL: HER2-amp or Luminal.

Figure 2: Evaluation of UBS93 classifier in human and mouse

(a) Evaluation of UBS93 using bulk RNA-seq data of breast cancer cell lines. In the confusion matrix, rows represent the actual subtypes, while columns represent the predicted subtypes. In each entry, the value indicates the relative proportion of samples (with the actual subtype corresponding to the row) predicted as the subtype indicated by the column.

(b) Evaluation of UBS93 using bulk RNA-seq data from TCGA breast cancer samples. The confusion matrix follows the same format as in (a).

(c) Evaluation of UBS93 using scRNA-seq data of breast cancer cell lines. In the boxplots, each dot represents a cell line, the central line indicates the median, and the bounds represent the 25th and 75th percentiles (interquartile range). Whiskers extend to 1.5 times the interquartile range.

(d) Scaled expression values of five Claudin-low indicative genes in the assembled murine dataset (27 samples). Asterisks highlight Claudin-low samples whose subtypes were not reported in Rädlers' study.

Figure 3: Comparison to existing methods

(a) Comparison of UBS93 with genefu using TCGA breast cancer samples.

(b) Comparison of UBS93 with SCSubtype using scRNA-seq data. In the boxplots, each dot represents a cell line, the central line indicates the median, and the bounds represent the 25th and 75th percentiles (interquartile range). Whiskers extend to 1.5 times the interquartile range. P-values were calculated using the two-sided Wilcoxon signed-rank test.

Figure 4: HDQP1 is a cell line with "Basal-like/Claudin-low admixture" phenotype

(a) The t-SNE visualization of HDQP1 scRNA-seq data (GSE17364), with cells colored according to their UBS93 subtypes.

(b) Differential gene expression analysis comparing Claudin-low and Basal-like cells of the HDQP1 cell line. Red and blue dots represent genes significantly up-regulated and down-regulated in Claudin-low cells, respectively.

(c) Gene set enrichment analysis (GSEA) of epithelial-mesenchymal transition (EMT) signature in the differentially expressed genes. The barcode beneath the heatmap represents the ranked position of EMT genes, from red (high log2FC) to blue (low log2FC).

(d) Comparison of stemness between Basal-like and Claudin-low cells. Each dot represents a cell, the central line indicates the median value and the bounds represent the 25th and 75th percentiles (interquartile range). Whiskers extend to 1.5 times the interquartile range. P-values were calculated using the two-sided Wilcoxon rank-sum test.

(e-f) Ranking of CCLE breast cancer cell lines based on transcriptomic correlation with Claudin-low (or Basal-like) pseudo-bulk sample. The Basal-like and Claudin-low cell lines are colored red and blue, respectively.

(g) Ranking of TNBC patients based on median correlation coefficients with the Claudin-low centroid. Each dot represents a single cell subtyped as Claudin-low by UBS93, the central line indicates the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers extend to 1.5 times the interquartile range.

Figure 5: Identification of TNBC patient with "Basal-like/Claudin-low admixture" phenotype

(a) The t-SNE visualization of scRNA-seq data from TNBC patient 0554, with cancer cells colored according to their UBS93 subtypes.

(b) Differential gene expression analysis comparing Claudin-low and Basal-like cells in TNBC patient 0554. Red and blue dots represent genes significantly up-regulated and down-regulated in Claudin-low cells, respectively.

(c) Gene set enrichment analysis (GSEA) of epithelial-mesenchymal transition (EMT) signature in the differentially expressed genes. The barcode beneath the heatmap represents the ranked position of EMT genes, from red (high log2FC) to blue (low log2FC).

(d) Comparison of stemness between Basal-like and Claudin-low cells. Each dot represents a cell, the central line indicates the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers extend to 1.5 times the interquartile range. P-values were calculated using the two-sided Wilcoxon rank-sum test.

(e-f) Ranking of CCLE breast cancer cell lines based on transcriptomic correlation with the Claudin-low (or Basal-like) pseudo-bulk sample. The Basal-like and Claudin-low cell lines are colored red and blue, respectively.

Figure 6: *ELF3* knock-down induces a partial Claudin-low phenotype in Basal-like cancer

(a) Hierarchical clustering of different cell types (from TNBC patient 0554) based on average copy number variation profile.

(b) Differential expression analysis comparing sh-*ELF3* and control groups. Red and blue dots represent genes significantly up-regulated and down-regulated in the sh-*ELF3* group, respectively.

(c) Scaled expression of down-regulated genes across CCLE breast cancer cell lines of different subtypes.

(d) Relative mRNA levels (quantified by RT-PCR) of *ELF3*, *CLDN4*, *CLDN7*, *CLDN3*, *EPCAM*, and *VIM* in control (blue) and sh-*ELF3* (red) groups. Each dot represents a parallel sample. P-values were calculated using an unpaired two-sided t-test. n.s. ($p > 0.05$), *($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).