

Heterogeneity of vascular cells in different breast cancer subtypes revealed by single-cell RNA-seq analysis

Lei Xia *

Shenzhen college of international school, Shenzhen, China

* Corresponding author: S20084.Xia@stu.scie.com.cn

Abstract. Breast cancer is a common type of malignant tumor developing from breast tissue that happens in women, which can be triggered by both genetic and environmental factors. Nowadays, it is increasingly recognized as a global public health problem. Despite some treatments have been used, there are still many challenges in treating breast cancer due to the recurrence and metastasis. The development of single-cell RNA sequencing (scRNA-seq) technology has shown great potential in deeply discovering the molecular mechanism of breast cancer. In this study, we processed scRNA-seq data of patients from different types of breast cancer subtypes, extracted and integrated the vascular cells, and found a distinct tumor-specific endothelial cell. The results not only showed the power of meta-analysis of existing single-cell data, but also revealed the detailed characteristics of vascular cells in different conditions, yielding light to future medical treatments in breast cancer.

Keywords: Breast Cancer, ScRNA-seq Technology, Vasculature, Endothelial Cells.

1. Introduction

The Breast cancer is a common type of cancer that threatens the health of women from every age, especially aging from 45 to 55. There are various factors potentially affect the incidence of breast cancer, including genetics, reproductive history and age. Normal breast tissue is comprised of a sophisticated epithelial ductal system embedded in a stromal matrix that is composed of fibroblasts, adipocytes, endothelial, and immune cells. A complicated branched ductal system where the ducts end in a cluster of acini termed a terminal duct lobular unit (TDLU) is resulted in human breast puberty-induced branching [1].

With the emergency of ScRNA-seq technology, it plays a key role in the study of breast cancer. The first generation of sc-RNA seq is Sanger sequencing in 1977. The second generation is high-throughput sequencing in 2006. The third generation is single-molecule sequencing in 2010. The fourth and most popular generation is single-cell sequencing invented in 2015. The fifth and newest generation is spatial transcriptomics in 2016. Some distinct alternations in the gene expression and biology of breast cells during growth are disclosed further by reconstructing the mammary gland epithelial cell growth trajectory through scRNA-seq, disclosing the heterogeneity of breast epithelial cells [2, 3]. Meanwhile, scRNA-seq makes great contribution to revealing the pre-existing drug-resistant cells [1]. To further understand how treatment resistance is acquired by breast cancer cells at the single-cell level, Samuel et al. conducted extensive analyses of the genomic and phenotypic subclone evolution of four patients with metastatic ER+ breast cancer during lengthy following studies [4, 5]. Furthermore, these phenotypes show traits related to drug-resistant characteristics, such as mesenchymal signal transduction, consequently facilitating the drug resistance of tumor cells [6]. The downregulation of antigen presentation and TNF- α signaling make tumor cells to run away from the attack of the immune system more readily and easily [7]. The selection of drug-resistant phenotypes is demonstrated by these findings clearly. These results provide a feasible phenotype-targeted treatment strategy for enhancing chemotherapy resistance in breast cancer. Lastly, scRNA-seq can be used in the prognosis of breast cancer. The marker genes of BCSC identified by scRNA-seq is a viable way to predict the prognosis of breast cancer. Tong et al. analyzed the gene expression profile of breast cancer stem cells (BCSCs) at the single-cell level and found that 74 BCSC marker genes were enhanced during the transcription process [2]. Breast cancer patients with a higher risk of relapse expressed higher levels of BCSC markers, proving the medical significance of BCSC markers

in predicting the prognosis of breast cancer. These 74 BCSC markers are consequently in great likelihood to be new targets and prognostic markers for breast cancer therapy and medical treatment.

Studies have demonstrated that healthy and cancerous states in the human breast are exhibited. However, little is known about whether vascular cells (endothelial cells and mural cells) also adopt pro- or/ and anti-tumor adaptations, and if so, if the adaptations features are different among subtypes and when they start to emerge, and if it persists after treatment. Since the vascular genesis is an essential progress in the tumor formation and migration, this research studied breast cancer subtypes from the angel of vasculature, leveraged the single-cell data from Pal et al. which profiled several subtypes which would facilitate the reveal of subtype-specific heterogeneity. and found a distinct ER-specific endothelial cells, which will potentially be used for early diagnosis, prognosis, as well as drug target and drug delivery anchor [8].

2. Method

The data analyzed in this study comes from Pal et al. Single-cell data analysis follows the standard tutorial of scanpy. There are 13 normal individuals and 45 patients with various types of breast cancer. 4 with luminal B type, 8 with triple negative type, 6 with Her2 type, 3 with mER type and 24 with ER type.

3. Results and discussions

3.1. Data-processing and quality-control (ndata taken for instance)

To begin with, the data from patients of each breast cancer subtypes including ER subtype, Her2 subtype, luminal B subtypes, and healthy controls were processed. For each condition, the data was imported from each individual and combined into an overall data. In the healthy control for example, by checking the quality control matrix, namely percentage of mitochondrial gene expression, percentage of ribosomal gene expression, numbers of genes, and total counts, cells outside the normal range were removed (Fig 1A, B). The filtering criteria were that: $pct_counts_rb < 35$, $pct_counts_mt < 6$, $n_genes_by_counts \geq 200$, $total_counts \geq 500$, plus a reasonable cap of $n_genes_by_counts$ between 7000-9000, depending on the dataset. Following a quality control on the overall dataset of each condition, the matrix by donor was checked to remove any low-quality data input from specific donor due to technical factors (Fig 1C). Following a selection of highly variable genes, dimension reduction was conducted by PCA, which showed a preliminary separation of major cell types (Fig 1D). Data integration was carried out with $n_neighbors = 30$, and the integration homogeneity was checked by showing donor distribution on the UMAP. As it can be seen in Fig 1E, the integration was good for healthy controls. In certain cases of breast cancer patient data, epithelial cells exhibited patient-specific pattern and did not cluster together, and instead the stromal cells (endothelial cells, smooth muscle cells, and fibroblasts, etc) and immune cell (T cells, B cells, plasma cells, macrophages, mast cells, etc.) clusters are relatively homogenous. At the end, the quality control matrix on the integrated UMAP was checked to make sure there is no cluster formed due to artificial factors like a high-mitochondrial gene cluster (Fig 1F).

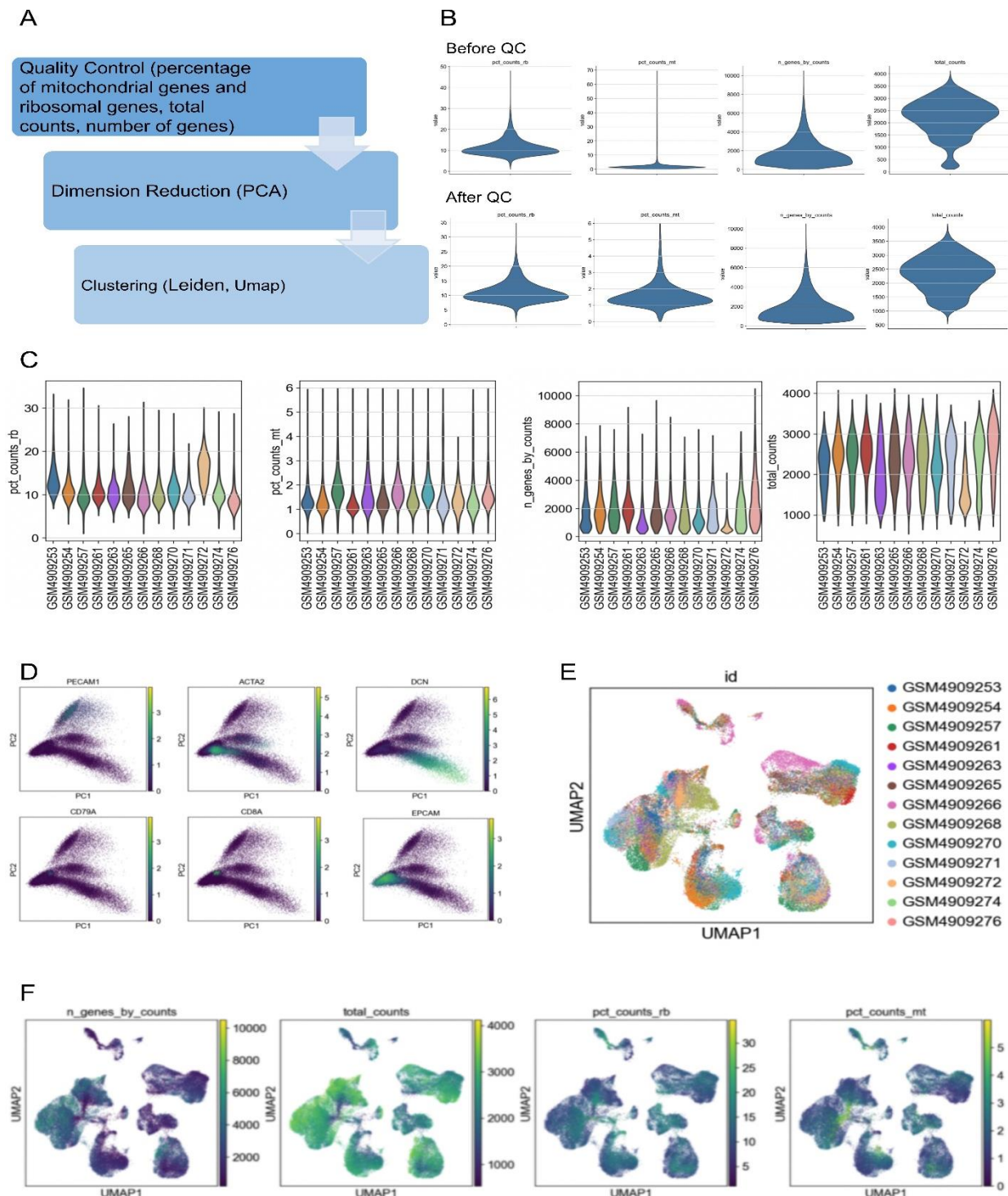


Figure 1. Data-processing & quality-control (ndata taken for instance).

(A) The workflow of data processing and quality control. (B) Violin diagrams showing percentage of mitochondrial gene expression, percentage of ribosomal gene expression, number of genes, and total counts before and after quality control. (C) The quality matrix by donor (sample id). (D) PCA diagrams showing marker genes for main cell types, including endothelial cells (PECAM1), smooth muscle cells (ACTA2), fibroblast (DCN), T cells (CD8A), B cells (CD79A), epithelial cells (EPCAM). (E) Diagram showing the data integration by donor (sample). (F) Umaps showing quality matrix after data integration.

3.2. The donor specificity mainly comed up in epithelial cells and fibroblasts

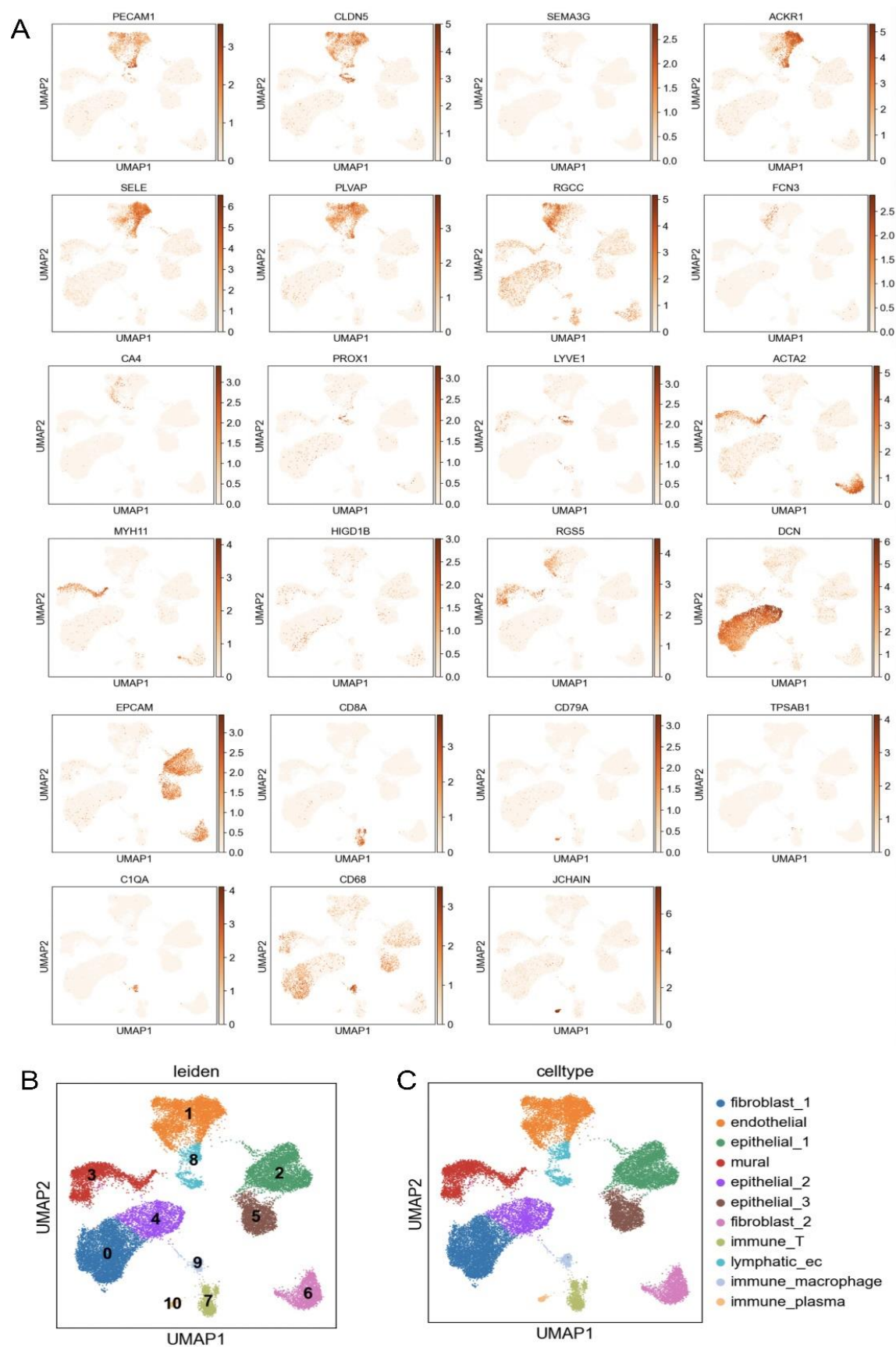


Figure 2. Major cell type annotation based on known marker genes.

(A) Umaps showing marker genes (like endothelial cells (PECAM1), smooth muscle cells (ACTA2).) (B) Unbiased leiden clustering. (C) Annotated clustering.

In order to correspond each cluster to major cell types, classical marker genes were plotted on the UMAP (Fig 2A). With a resolution = 0.35, the cells were separated by clear initial clustering (Fig 2B). Marker genes were used to identify as well as define cell types, such as endothelial cells, smooth muscle cells, pericytes, fibroblasts, epithelial cells, T cells, B cells, plasma cells, macrophages, mast cells (Table 1). For multi-cell type clusters that were not fully separated by the resolution of 0.35, they were further restricted clustering within that cluster, which was potentially likely to be the case for breast cancer subtype datasets. Fig 2C shows the final annotation of the healthy control, where it could be seen that the donor specificity mainly came up in epithelial cells and fibroblasts.

Table 1. Classical marker genes for major cell types

Gene name	Cell type (high resolution)	Cell type (low resolution)
PECAM1	Endothelial cells	Endothelial cells
SEMA3G	Artery endothelial cells	Endothelial cells
RGCC	Capillary endothelial cells	Endothelial cells
MYH11	Smooth muscle cells	Mural cells
RGS5	pericytes	Mural cells
EPCAM	Epithelial cells	Epithelial cells
CD8A	T cells	Immune cells
CLDN5	Endothelials	Endothelial cells
ACKR1	artery endothelial cells	endothelial cells
PLVAP	Vein endothelial cells & capillary endothelial cells	endothelial cells
PROX1	Lymphatic cells	immune cells
LYVE1	lymphatic cells	immune cells
ACTA2	mural cells	mural cells
C1QA	macrophage cells	immune cells
DCN	fibroblast cells	fibroblast cells
CD79A	B cells	Immune cells
TPSAB1	mast cells	immune cells
JCHAIN	plasma cells	blood cells
CA4	capillary endothelial cells	endothelial cells
FCN3	capillary endothelial cells	endothelial cells
HIGD1B	mural cells	mural cells

3.3. Analysis of various breast cancer subtypes

Epithelial cell diversity in breast cancer subtype TN and Her2 were very obvious and distinct when compared to the rest subtypes (Fig 3B). Subtype luminal B had a relatively higher fibroblast level than others and was divided into two clustering (Fig 3D). Immune cells were enriched in number in subtype TN, especially macrophages (Fig 3A). Macrophages in subtype luminal B seemed to be relatively limited (Fig 3D). Breast cancer subtype mER exhibited a simple cell type constitution and clear clustering, which may be a result of fewer donors and the small number of cells (Fig 3E). The largest heterogeneity among patients was exhibited in breast cancer subtype Her2 Fig 3B). Specifically, there were 6 epithelial cell cluster from 1 patient (Fig 3A).

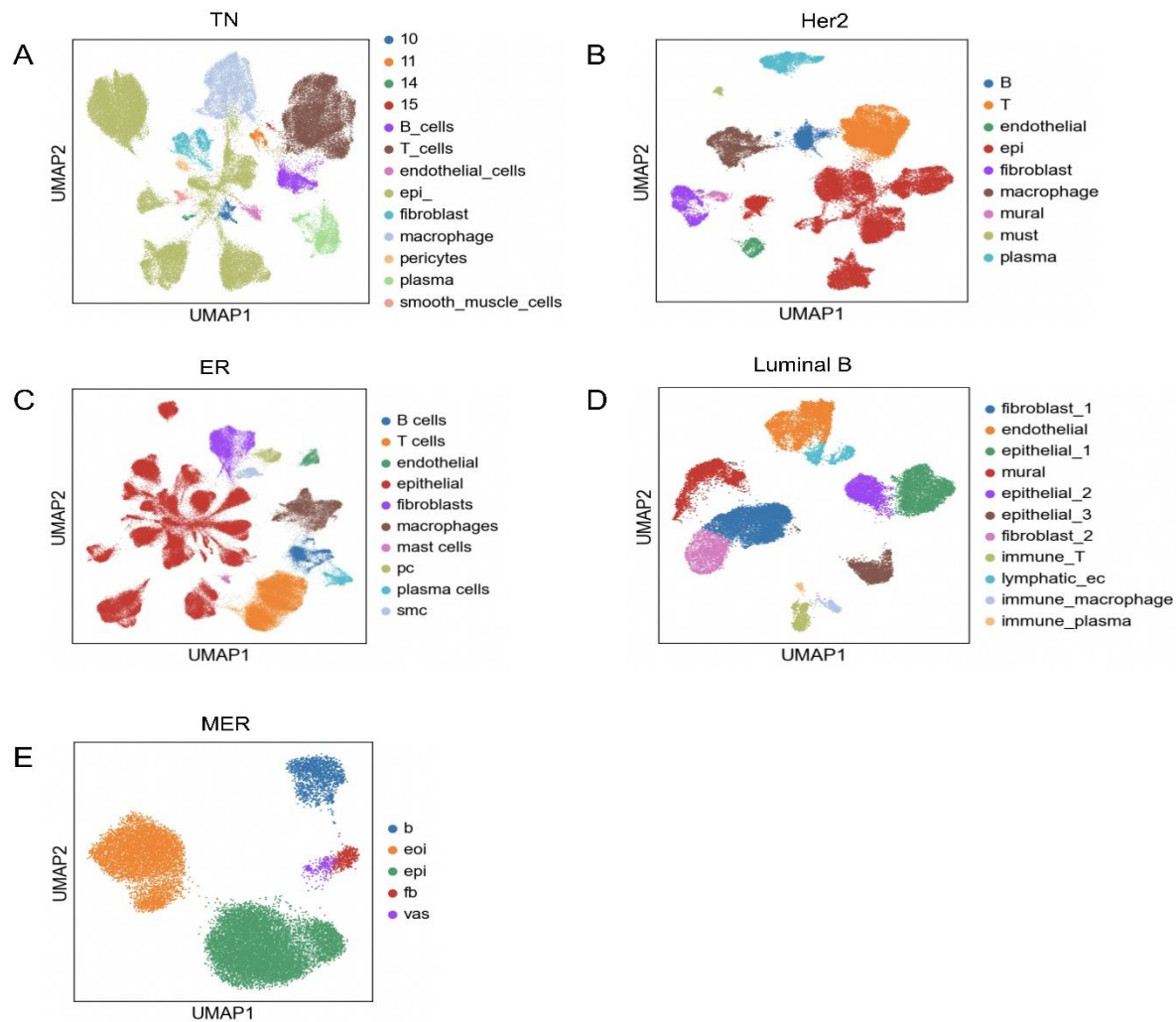


Figure 3. Major cell annotation of breast cancer subtypes.

(A-E) UMAPs showing cell type annotations for breast cancer TN, Her2, ER, mER, luminal B1. Having the annotations of major cell types, vascular cells were extracted including endothelial cells, smooth muscle cells, and pericytes (sometimes grouped as mural cells). The numbers of vascular cells from healthy controls and breast cancer subtypes patients are listed in Table 2.

Table 2. Numbers of vascular cells from healthy controls and breast cancer subtypes patients

	N	TN	B	ER	MER	Her2
Total cell number (post qc)	60660	64213	25910	140291	18617	46645
Total number of vascular cells	11648	1129	7084	3914	64	1129
Endothelial cells	6321	438	4429	1065	24	722
Mural cells	4217	581	2330	2794	40	383
Lymphatic EC	53	12	325	53	0	24

3.4. Removing artificial factors via a batch correction

From the initial integration, an obvious and distinct separation of conditions can be seen. Specifically, breast cancer subtypes ER, TN, her2, luminal B, were distantly separated from the healthy control on the UMAP (Fig 4A). By checking the markers of vascular sub cell types (ACKR1 for venous endothelial cells, RGCC for capillary endothelial cells), it is found that the same cell types were split into two clusters (Fig 4B), which was potentially dominated by the condition, either healthy or having breast cancer. Yet a homogeneous distribution of the quality control matrix excluded the possibility that the separation of the same type of vascular cells was caused by artificial factors like number of genes or counts, or mitochondrial and ribosomal genes (Fig 4C). Therefore, a batch correction using

bbknn is carried out to remove artificial factors brought in by each donor. Data from different conditions integrated better after batch correction, yet certain breast cancer-specific vascular clusters persisted, mainly dominated by subtype ER, TN, etc (Fig 4D, E).

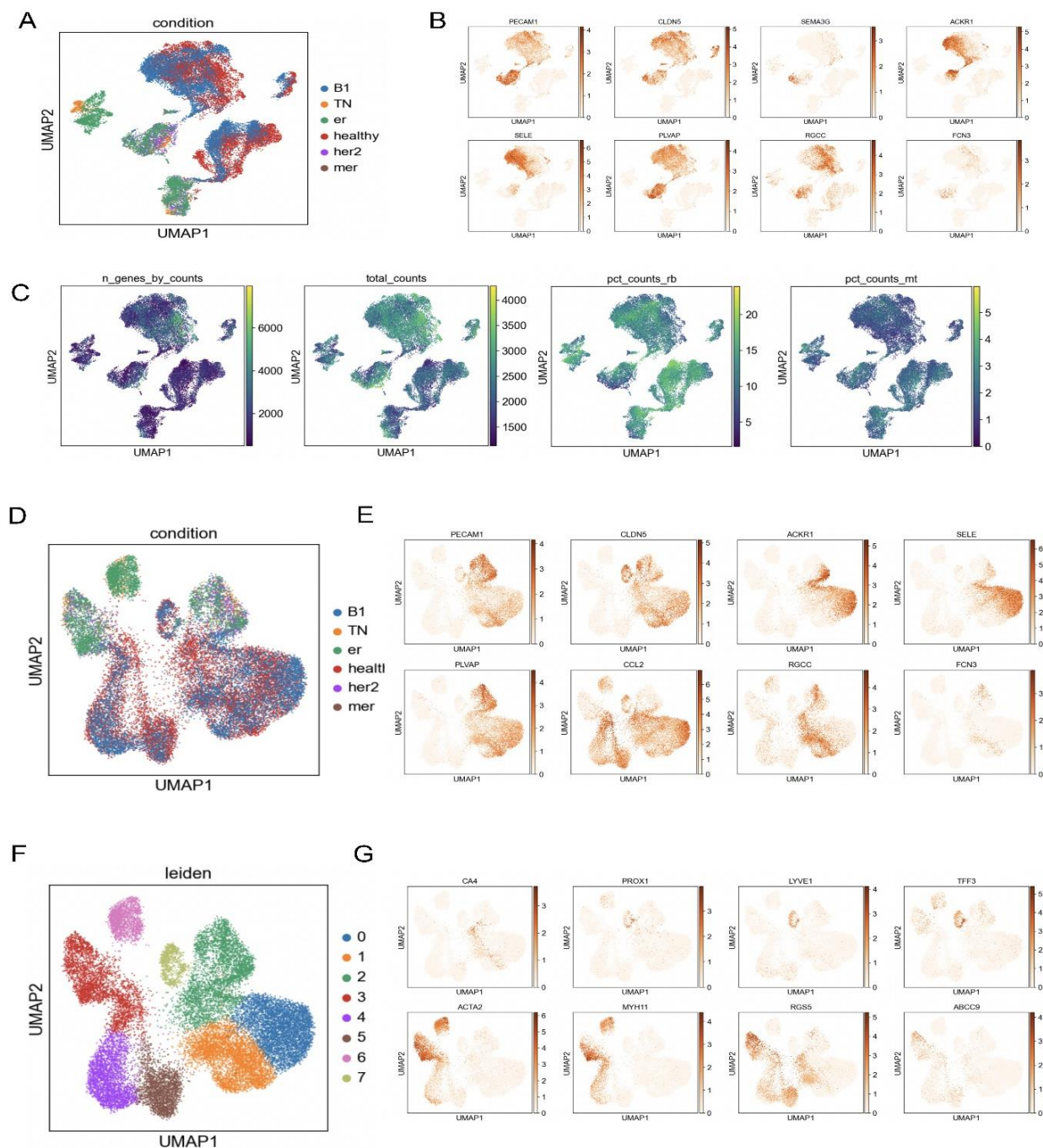


Figure 4. Integration of vascular cells from breast cancer datasets and healthy control.

(A) Umap showing subtype distribution before batch correction. (B) Umap showing marker genes expression before batch correction (ACKR1 for venous endothelial cells, RGCC for capillary endothelial cells). (C) Umap showing quality control factor distribution before batch correction. (D) Umap showing subtype distribution after batch correction. (E) Umap showing marker genes expression (ER, TN, Her2, luminal B) after batch correction. (F) Umap showing the leiden distribution after batch correction (G) Umap showing marker genes expression (like CA4, PROX1, LYVE1, TFF3, ACTA2 etc.)

3.5. Stromal cells like smooth muscle cells were able to facilitate tumour cell migration and coagulation

After batch correction and re-integration, the cells clusters were closer together, yet exhibiting distinction between healthy control and breast cancer patients (Fig 5A). No cluster was formed due to artificial factors (Fig 5B). Annotation was inferred by the distribution of classical vascular marker genes (Table 1), especially PECAM1 for endothelial cells, ACKR1 for venous endothelial cells, RGCC for capillary endothelial cells, MYH11 for smooth muscle cells, RGS5 for pericytes, and LYVE1 for lymphatic endothelial cells (Fig 5C). A total of 9 clusters were annotated, with 4 of them mainly consisting of breast cancer samples, 4 from healthy controls, and 1 lymphatic endothelial cell cluster shared by both groups (Fig 5D).

Intriguingly, a cluster dominated by ER condition, together with a few other breast cancer subtypes, stood out. It expressed high level of smooth muscle markers including ACTA2 (not shown), and MYH11 (Fig 5C), yet it was separated from the common smooth muscle cluster (“smc_bc”). A differential gene expression analysis showed it had a high level of KRT14 and KRT17. KRT14 encoded for Keratin 14, which was a type of keratin that had been found to be involved in cancer metastasis by regulating cell adhesion and evasion [9]. A study in breast cancer conducted both in vivo and in vitro experiments and demonstrated that Keratin 14 regulates the formation of polyclonal tumors deposit. Furthermore, our findings showed that tumor adaptation can not only be seen in epithelial cells, but also stromal cells like smooth muscle cells could also adopt features to facilitate tumor cell migration and coagulation.

Although the rest vascular cell types were shared between healthy control and breast cancer patients, they still possessed differentially expression gene features (Fig 5F-H). To be specific, the endothelial cells from breast cancer patients had a higher expression of genes including CRIP2, VWF, AQP1, CD34, and IGFBP4 (Fig 5F). IGFBP4 had been reported as an anti-angiogenic and anti-tumorigenic factor (Moreno et al) [10] indicating that the anti-tumorigenic program would be activated when tumor was presence. Yet the normal immune recruiting function of the venous endothelial cells (indicated by the expression of SELE) was inhibited in breast cancer pericytes. Smooth muscle normally expressed IL6 and IER3 as an immune defence and response under stress conditions, which went silent in breast cancer conditions (Fig 5G). The same trend was also seen in pericytes where the expression of anti-oxidation SOD2 and CXCL families were lowered in breast cancer pericytes, but more fibro was formed as indicated by COL1A2, COL3A1, and COL1A1, etc (Fig 5H).

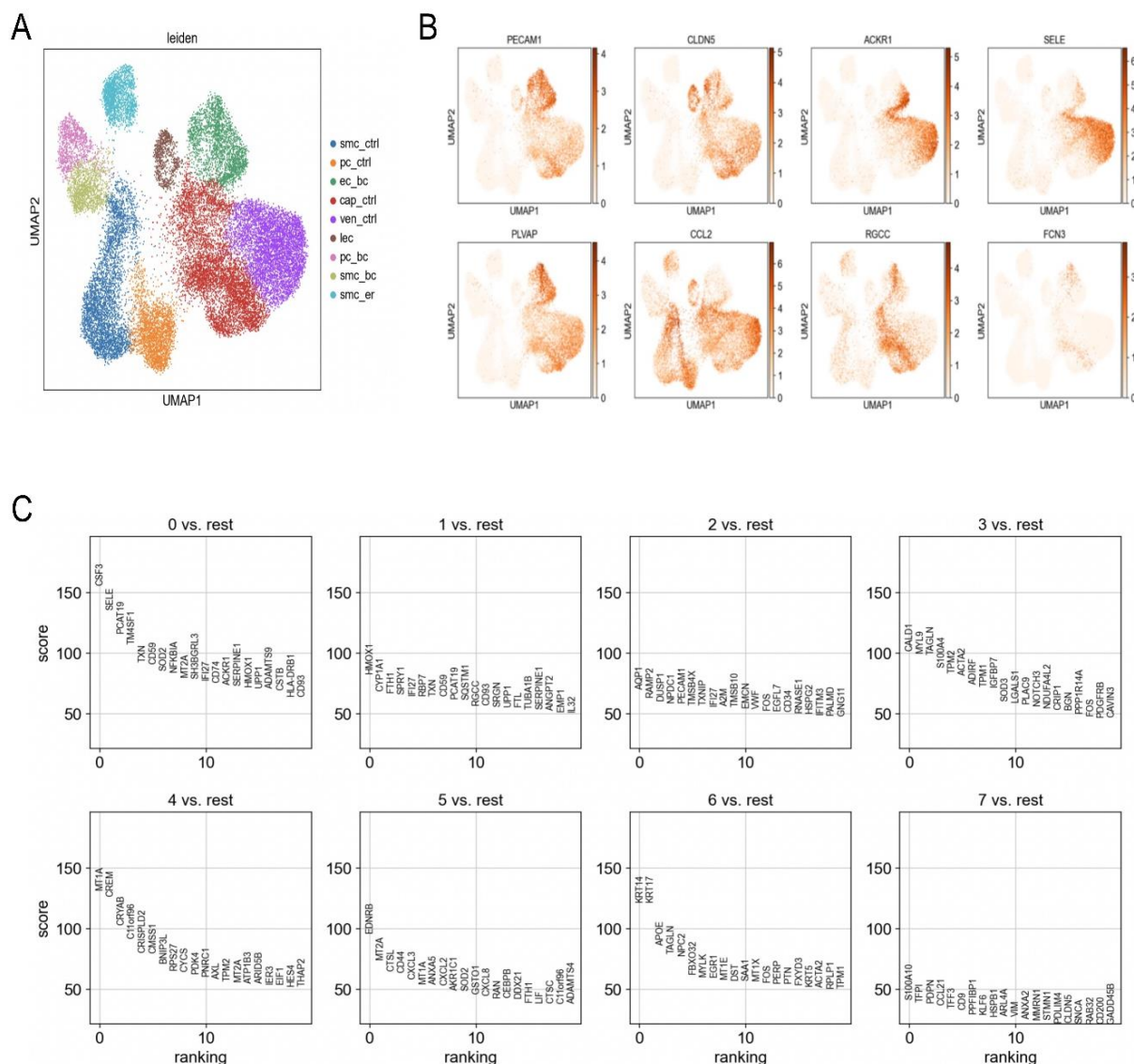


Figure 5. Subtype-specific vascular type or gene features

(A) Umap showing subtype distribution. (B) Expression of marker genes. (C) Differentially expressed genes by cluster.

4. Conclusions

After the quality-control and data-processing, it was clear to see that in epithelial cells and fibroblasts, the donor specificity was obvious. When analyzing and comparing all 5 subtypes, the epithelial diversity is obvious in TN and Her2 subtype. Moreover, in luminal B, the fibroblast level was very high. Immune cells were in large amount in TN subtype, especially for macrophages. The largest heterogeneity among patients was exhibited in Her2 subtype.

Going further, after checking the markers of vascular sub cell types for venous endothelial cells, the same cell sorts were split into two clusters as shown in the above diagrams, which was possibly dominated by the condition, either healthy or having breast cancer.

Interestingly, a cluster dominated in ER subtype, together with a few other breast cancer subtypes, stood out since it expresses a high level of smooth muscle markers including ACTA2 and MYH11. Besides, my finding showed that tumor adaptation cannot only be seen in epithelial cells, but also stromal cells like smooth muscle cells could also adopt features to facilitate tumor cell migration and coagulation.

References

- [1] Ren et al. Single cell RNA sequencing for breast cancer: present and future, *Cell Death Discovery*, 1, 5 (2021).
- [2] Nguyen, Q. H. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* 9, 2028 (2018).
- [3] Brady, S. W. et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* 8, 1231 (2017).
- [4] Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* 173, 879 – 893 (2018).
- [5] Mani, S. A. et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* 133, 704 – 715 (2008).
- [6] Vereecke, L., Beyaert, R. & van Loo, G. The ubiquitin-editing enzyme A20 (TNFAIP3) is a central regulator of immunopathology. *Trends Immunol.* 30, 383 – 391 (2009).
- [7] Tong, M. et al. Transcriptomic but not genomic variability confers phenotype of breast cancer stem cells. *Cancer Commun.* 38, 56 (2018).
- [8] Pal et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast, 1, (2021).
- [9] Cheung et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14 - expressing tumor cell clusters, *Cell Biology*, 2, 5 (2016).
- [10] Moreno et al. IGFBP-4 anti-angiogenic and anti-tumorigenic effects are associated with anti-cathepsin B activity, *Neoplasia*, 3, 7 (2013).