Investigating Candidate Marker Genes Across Breast Cancer Subtypes

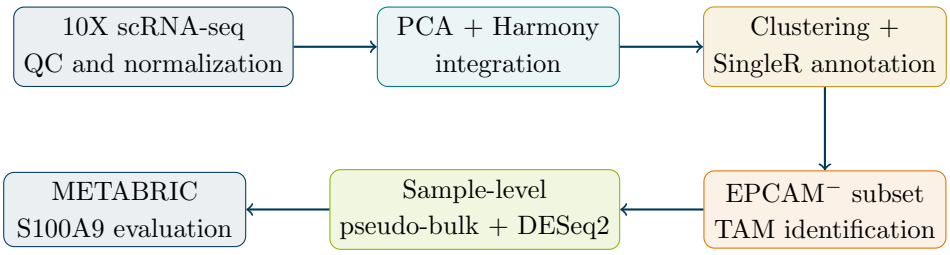
Single-cell transcriptomic analysis of tumour-associated macrophages and independent METABRIC evaluation

Turky AlSulaiman

KAUST / University of Cambridge
Department of Genetics, University of Cambridge
Mentor: Chandra Chilamakuri

Research Project Report

24 August 2026



Contents

1 Abstract **2**

2 Introduction **2**
2.1 Breast cancer heterogeneity . . . 2
2.2 The tumour microenvironment and tumour-associated macrophages . . . 3
2.3 Single-cell RNA sequencing as an analytical framework . . . 3
2.4 Rationale for pseudo-bulk analysis . . . 4
2.5 Candidate gene evaluation and independent validation . . . 4
2.6 Objectives . . . 4

3 Related Work **5**
3.1 Single-cell analysis of heterogeneous tissues . . . 5
3.2 Harmony integration . . . 5
3.3 Reference-based cell annotation . . . 5
3.4 Differential expression and pseudo-bulk analysis . . . 6
3.5 METABRIC and independent cohort analysis . . . 6

4 Methodology **6**
4.1 Study design . . . 6
4.2 Single-cell dataset . . . 7
4.3 Quality control . . . 7
4.4 Normalization and highly variable genes . . . 8
4.5 Principal component analysis . . . 8
4.6 UMAP and clustering . . . 8
4.7 Harmony batch correction . . . 8
4.8 SingleR cell-type annotation . . . 9
4.9 EPCAM-negative analysis . . . 9
4.10 TAM identification . . . 9
4.11 TAM sample filtering . . . 10
4.12 Pseudo-bulk aggregation . . . 10
4.13 Subtype definition for pseudo-bulk analysis . . . 10
4.14 DESeq2 differential expression . . . 10
4.15 Candidate gene selection . . . 11
4.16 METABRIC data . . . 11
4.17 METABRIC subtype classification . . . 11
4.18 METABRIC subtype statistics . . . 11
4.19 Lymph-node analysis . . . 12
4.20 Software . . . 12

5 Results **12**
5.1 Single-cell dataset and preprocessing . . . 12
5.2 Integrated single-cell representation . . . 13
5.3 Cell-type annotation . . . 13

5.4 Identification of the EPCAM-negative population . . . 13
5.5 Identification of TAMs . . . 13
5.6 Pseudo-bulk differential expression . . . 13
5.7 S100A9 expression across METABRIC subtypes . . . 14
5.8 S100A9 and lymph-node involvement . . . 14
5.9 Summary of principal findings . . . 15

6 Discussion **15**
6.1 Integration of single-cell and bulk-cohort evidence . . . 15
6.2 Biological interpretation of the subtype pattern . . . 15
6.3 Interpretation of the lymph-node finding . . . 16
6.4 Why pseudo-bulk was appropriate . . . 16
6.5 Role of Harmony and clustering evaluation . . . 16
6.6 Importance of independent validation . . . 17
6.7 Potential relevance to TNBC . . . 17
6.8 Interpretation of the candidate gene . . . 17

7 Limitations and Future Work **17**
7.1 Limitations . . . 17
7.1.1 Limited single-cell sample size . . . 18
7.1.2 TAM identification is computational rather than experimentally confirmed 18
7.1.3 Bulk validation does not establish TAM-specific expression . . . 18
7.1.4 Association does not establish causation . . . 18
7.1.5 Clinical relevance requires additional endpoints . . . 18
7.2 Future work . . . 18

8 Conclusion **19**

9 Reproducibility and Computational Organization **19**

10 References **20**

A Appendix: Key Analysis Parameters **21**

B Appendix: Suggested Figure Placement **21**

C Appendix: Interpretation Guide **22**

1 Abstract

Breast cancer is a biologically heterogeneous disease in which clinically important subtypes differ in receptor status, molecular programs, treatment options, and disease behaviour. Tumour-associated macrophages (TAMs) are an important component of the tumour microenvironment and can display transcriptional states that vary with the surrounding tumour context. This project investigated candidate marker genes associated with breast cancer subtypes, with particular emphasis on transcriptional differences within TAMs.

Single-cell RNA sequencing (scRNA-seq) data from 30 breast cancer samples were analyzed using an R-based workflow. The dataset comprised 16 oestrogen receptor-positive (ER+) samples, 6 human epidermal growth factor receptor 2-positive (HER2+) samples, and 8 triple-negative breast cancer (TNBC) samples. Individual 10X Genomics datasets were merged and subjected to quality control, normalization, highly variable gene selection, principal component analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and clustering using Seurat. Harmony was used to reduce sample-associated technical variation. Clustering parameters were evaluated using silhouette analysis, and cells were annotated using SingleR with a reference-based cell-type annotation strategy. An EPCAM-negative population was subsequently analyzed to focus on non-epithelial cells, and a macrophage/TAM population was identified using clustering together with expression of macrophage-associated markers, particularly CD14 and CD68.

Because cells originating from the same biological sample are not independent biological replicates, TAM counts were aggregated at the sample level to create pseudo-bulk expression profiles. Differential expression was then performed using DESeq2 for HER2+ versus ER+, TNBC versus ER+, and TNBC versus HER2+ comparisons. S100A9 emerged as a candidate gene for further evaluation. An independent analysis of the METABRIC cohort was performed using microarray expression and clinical metadata. S100A9 expression was compared across ER+ ($n = 1,398$), HER2+ ($n = 134$), and triple-negative ($n = 320$) tumours and was also examined in relation to lymph-node involvement.

The METABRIC analysis supported the subtype-associated pattern identified in the single-cell pseudo-bulk analysis, with higher S100A9 expression in HER2+ and triple-negative tumours compared with ER+ tumours. S100A9 expression also showed a statistically significant but modest increase among tumours with lymph-node involvement ($p = 0.0002092$), indicating an association that warrants further investigation rather than establishing a strong clinical effect. Overall, the project demonstrates a workflow that connects cell-level transcriptomic analysis with sample-level differential expression and independent cohort evaluation. S100A9 was identified as a subtype-associated candidate gene whose biological and clinical relevance should be investigated further, particularly in TNBC and lymph-node disease.

Keywords: breast cancer; single-cell RNA sequencing; tumour-associated macrophages; pseudo-bulk; DESeq2; Harmony; SingleR; S100A9; METABRIC; TNBC.

2 Introduction

2.1 Breast cancer heterogeneity

Breast cancer is the most common cancer affecting women worldwide and represents a major global health burden. Its clinical behaviour is highly heterogeneous, reflecting substantial

variation at the histological, molecular, cellular, and genomic levels. Modern breast cancer classification relies heavily on the expression of hormone receptors and HER2, among other molecular characteristics. Oestrogen receptor (ER), progesterone receptor (PR), and HER2 status are particularly important because they influence disease classification and treatment selection. TNBC is defined by the absence of ER, PR, and HER2 expression and therefore represents a clinically challenging group in which endocrine and HER2-directed therapies are not applicable.

The three broad groups examined in this project were ER+, HER2+, and TNBC. These groups were selected because they represent biologically distinct breast cancer contexts and because comparison among them provides an opportunity to identify genes associated with subtype-specific biology. Importantly, receptor-based classification describes the tumour compartment, whereas the tumour microenvironment contains additional populations that can influence tumour behaviour.

2.2 The tumour microenvironment and tumour-associated macrophages

The tumour microenvironment is composed of malignant cells together with immune cells, stromal cells, endothelial cells, extracellular matrix components, and other non-malignant populations. Among infiltrating immune populations, macrophages are frequently abundant in solid tumours. Tumour-associated macrophages are not a single uniform population; rather, they can adopt diverse transcriptional and functional states in response to local signals.

This cellular heterogeneity creates an important analytical problem. Bulk RNA sequencing averages expression across many cell types and can therefore obscure cell-type-specific transcriptional programs. scRNA-seq provides an alternative by measuring transcriptomic profiles at the level of individual cells, allowing cellular populations to be identified and studied separately.

The present project therefore focused on TAMs rather than analyzing the entire tumour transcriptome as one mixture. The central question was whether TAM-associated gene expression differed among ER+, HER2+, and TNBC tumours and whether candidate genes identified in this setting could be evaluated independently in a large breast cancer cohort.

2.3 Single-cell RNA sequencing as an analytical framework

scRNA-seq enables transcriptomic measurements from individual cells and can reveal cellular populations that would otherwise be masked in bulk measurements. However, the increased resolution comes with analytical challenges, including sparse count matrices, variation in sequencing depth, sample-to-sample technical variation, high dimensionality, and the non-independence of cells collected from the same biological sample.

A typical scRNA-seq workflow includes quality control, normalization, feature selection, dimensionality reduction, clustering, cell-type annotation, and downstream differential expression. In this project, Seurat was used as the primary framework for these operations. Integration was performed using Harmony, which is designed to reduce unwanted dataset-associated variation while retaining biologically meaningful structure (Butler et al., 2018; Korsunsky et al., 2019).

2.4 Rationale for pseudo-bulk analysis

A major consideration in comparing biological groups using scRNA-seq is the experimental unit. Thousands of cells from a single patient do not represent thousands of independent patients. Treating cells as independent observations can therefore lead to pseudoreplication and inflated statistical significance.

To address this issue, the project used a pseudo-bulk strategy for the TAM differential expression analysis. Raw counts from TAM cells were summed within each sample, producing a single gene-count profile for each biological sample. Differential expression was then conducted using DESeq2, a negative-binomial model designed for count-based expression data (Love et al., 2014).

This approach preserves the sample as the biological replicate and allows subtype comparisons to be performed at the appropriate level.

2.5 Candidate gene evaluation and independent validation

Differential expression analysis was used to identify genes whose expression differed between breast cancer subtypes within the TAM population. S100A9 was selected for further investigation as a candidate marker gene.

A second major objective was to determine whether a pattern observed in the single-cell analysis could also be observed in an independent cohort. METABRIC provides a large breast cancer dataset combining molecular measurements with clinical information and has been widely used to investigate breast cancer molecular heterogeneity (Curtis et al., 2012).

The METABRIC analysis therefore served as an independent evaluation of S100A9 expression across breast cancer subtypes and its relationship with lymph-node involvement. This distinction is important: the METABRIC analysis does not demonstrate that S100A9 expression originates specifically from TAMs, because the expression measurements represent bulk tumour samples containing multiple cellular populations. Instead, it asks whether the candidate gene identified through the TAM-focused single-cell analysis shows a consistent subtype-associated pattern in an independent breast cancer cohort.

2.6 Objectives

The overall objective was to identify and characterize candidate marker genes associated with breast cancer subtype differences within TAMs.

The specific objectives were:

1. To process and quality-control scRNA-seq data from ER+, HER2+, and TNBC breast cancer samples.
2. To integrate samples while reducing sample-associated technical variation using Harmony.
3. To identify cellular populations using unsupervised clustering and characterize them using SingleR annotation and marker-gene expression.
4. To isolate the EPCAM-negative population and identify a macrophage/TAM population using clustering and macrophage marker expression.

5. To generate sample-level pseudo-bulk profiles from TAM cells and perform subtype-associated differential expression using DESeq2.
6. To identify candidate genes from the TAM differential expression results and investigate S100A9 in greater detail.
7. To independently evaluate S100A9 expression in METABRIC across ER+, HER2+, and triple-negative tumours.
8. To explore the relationship between S100A9 expression and lymph-node involvement.

3 Related Work

3.1 Single-cell analysis of heterogeneous tissues

The development of scRNA-seq has made it possible to characterize complex tissues at single-cell resolution. Integration methods allow cells from multiple samples and experimental conditions to be analyzed jointly while attempting to distinguish biological structure from technical variation. Butler et al. demonstrated an integration strategy for combining multiple scRNA-seq datasets and identifying shared cellular populations across datasets (Butler et al., 2018). Stuart and Satija subsequently described broader principles for integrative single-cell analysis, emphasizing the importance of identifying shared cellular states while accounting for dataset-specific variation (Stuart and Satija, 2019).

These principles are directly relevant to breast cancer, where multiple patients and tumour subtypes can contain overlapping cellular populations but also differ in their biological composition. In this project, integration was performed before clustering so that cellular populations could be compared across the different breast cancer groups.

3.2 Harmony integration

Harmony is an iterative integration method that operates on a low-dimensional representation such as PCA and attempts to produce a shared embedding in which cells are organized primarily according to biological state rather than dataset-specific effects (Korsunsky et al., 2019). The method was selected for this project because multiple patient samples were analyzed jointly and sample-associated variation could otherwise influence clustering.

Harmony was not treated as evidence that all technical variation had been eliminated. Rather, it was used as a computational correction step whose output was evaluated through dimensionality-reduction plots and clustering quality measures.

3.3 Reference-based cell annotation

Cell-type annotation is a critical step in scRNA-seq analysis. Unsupervised clusters do not inherently have biological labels, and marker expression can sometimes be ambiguous. SingleR provides a reference-based approach that compares single-cell expression profiles with reference transcriptomes to assign cell-type labels (Aran et al., 2019).

In this project, SingleR was used together with marker-gene inspection. The annotation was therefore not treated as the sole evidence for macrophage identity. The macrophage population was examined using additional markers, including CD14, CD68, LST1, C1QC, and CX3CR1.

3.4 Differential expression and pseudo-bulk analysis

DESeq2 models RNA-seq count data using negative-binomial distributions and incorporates library size normalization and dispersion estimation for differential expression testing (Love et al., 2014). The method is designed around biological replicates, which makes it appropriate for sample-level pseudo-bulk analysis.

The pseudo-bulk strategy used here differs from performing a conventional Wilcoxon test across individual cells. Instead of considering every TAM cell as an independent observation, counts were aggregated within each sample. Consequently, subtype comparisons were driven by between-sample variation, which better corresponds to the experimental design.

3.5 METABRIC and independent cohort analysis

METABRIC integrated molecular and clinical characterization across a large breast cancer cohort and helped demonstrate the molecular diversity of breast cancer (Curtis et al., 2012). The dataset is useful for independent evaluation because it contains a much larger number of tumours than the single-cell dataset used for this project.

However, an important methodological distinction remains. METABRIC expression measurements are bulk measurements. Therefore, higher S100A9 expression in a METABRIC subgroup cannot by itself establish that the difference is caused by TAMs. The role of the METABRIC analysis in this project is to evaluate whether the candidate gene's overall tumour expression pattern is consistent with the subtype-associated signal identified in the TAM-focused single-cell analysis.

4 Methodology

4.1 Study design

The project consisted of two connected analytical components.

The first component was a single-cell discovery workflow. It began with raw 10X Genomics data, proceeded through quality control and integration, identified cellular populations, isolated TAMs, and used sample-level pseudo-bulk differential expression to identify candidate genes.

The second component was an independent bulk-expression evaluation. S100A9, selected from the single-cell differential expression results, was evaluated using METABRIC expression and clinical data.

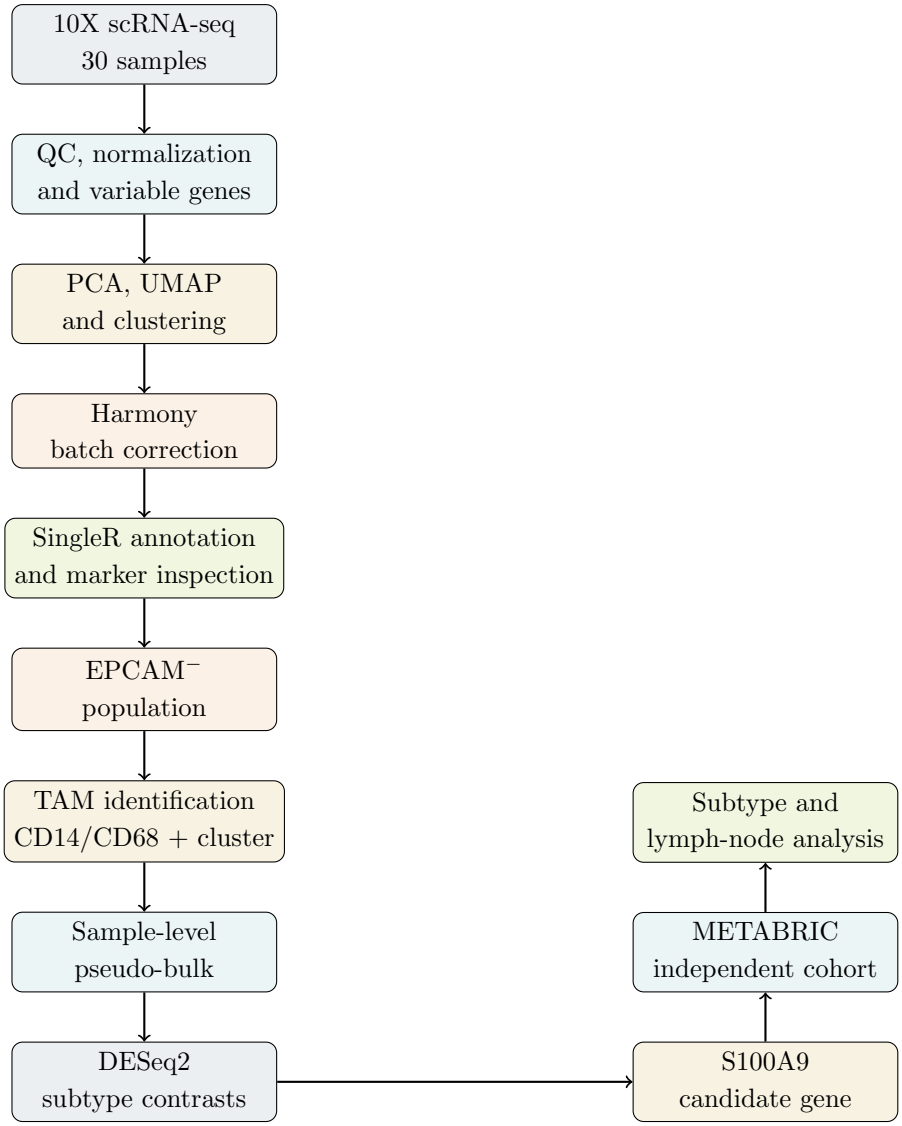


Figure 1: Overview of the analytical workflow. The single-cell component progresses from 10X Genomics data through quality control, integration, annotation, TAM identification, sample-level pseudo-bulk aggregation, and DESeq2 differential expression. S100A9 was subsequently evaluated in the independent METABRIC cohort.

4.2 Single-cell dataset

The single-cell dataset consisted of 30 breast cancer samples: 16 ER+, 6 HER2+, and 8 TNBC samples. The TNBC group included BRCA1-associated TNBC samples, which were incorporated into the TNBC category for the downstream TAM pseudo-bulk comparisons.

The data were processed as individual 10X Genomics datasets and subsequently merged. Sample identity and experimental group were retained as metadata to allow sample-level aggregation and subtype comparisons.

4.3 Quality control

Quality control was performed at the sample level before downstream integration. The number of detected genes per cell and mitochondrial transcript proportion were used as primary QC

variables.

Cells were retained using the following criteria:

- $nFeature_RNA \geq 500$;
- mitochondrial percentage $\leq 20\%$.

Mitochondrial genes were removed from the expression matrix after mitochondrial content had been evaluated. These filters were intended to reduce the influence of low-quality cells and cells exhibiting unusually high mitochondrial transcript abundance.

4.4 Normalization and highly variable genes

Each sample was normalized using Seurat's normalization workflow. Highly variable genes were then identified, with 2,000 variable features selected for downstream dimensionality reduction.

Variable feature selection reduces the dimensionality of the expression matrix while retaining genes that contribute strongly to cell-to-cell variation. This is particularly useful in scRNA-seq because the complete gene-by-cell matrix is high-dimensional and sparse.

4.5 Principal component analysis

PCA was performed on the normalized expression data to generate a lower-dimensional representation of transcriptional variation. The first 30 principal components were calculated, and the first 20 PCs were used for the principal neighborhood and UMAP analyses in the main workflow.

An elbow plot was used to examine the variance captured by the principal components and to support selection of an appropriate number of dimensions.

4.6 UMAP and clustering

UMAP was used to visualize the cellular landscape in two dimensions. UMAP does not directly classify cells; instead, it provides a low-dimensional representation in which cells with similar high-dimensional expression profiles tend to be positioned near one another.

Graph-based clustering was then performed using Seurat. The analysis explored multiple values of the nearest-neighbour parameter and clustering resolution. Rather than selecting these parameters arbitrarily, silhouette analysis was used to assess separation between clusters.

The parameter search included:

$$k \in \{10, 15, 20, 25, 30, 40\}$$

with multiple clustering resolutions evaluated. The final analysis used $k = 15$ and a clustering resolution of 0.15.

The resulting cluster identities were stored as:
`Louvain_k15_res0.15`.

4.7 Harmony batch correction

Because multiple patient samples were combined, sample-specific technical variation could influence the low-dimensional structure. Harmony was therefore applied to the PCA representa-

tion. Harmony attempts to align cells across datasets while preserving biologically meaningful differences (Korsunsky et al., 2019).

The corrected embedding was used for UMAP, nearest-neighbour construction, and clustering. Harmony correction was evaluated visually by examining UMAPs coloured by sample and biological group.

4.8 SingleR cell-type annotation

The Human Primary Cell Atlas reference was used with SingleR to obtain reference-based cell-type annotations. The resulting labels provided a broad characterization of the cellular populations.

Annotation was supplemented by examination of marker genes. In particular, immune and stromal populations were investigated using markers such as PTPRC, LST1, CD3D, COL1A1, COL1A2, DCN, LUM, PECAM1, VWF, and epithelial keratins.

4.9 EPCAM-negative analysis

Because EPCAM is a marker associated with epithelial tumour cells, an EPCAM-negative population was isolated for downstream analysis:

```
epcam_neg <- subset(combined, subset = EPCAM == 0)
```

The EPCAM-negative population was not simply carried forward without further analysis. It underwent a separate PCA, Harmony integration, UMAP, clustering, and annotation workflow. This allowed non-epithelial populations to be examined at greater resolution.

4.10 TAM identification

A macrophage/TAM population was identified within the EPCAM-negative dataset using both cluster identity and macrophage-associated gene expression. Cluster 2 was selected based on the clustering structure and then evaluated using CD14 and CD68 expression.

Additional macrophage-associated genes examined included:

- CD14;
- CD68;
- LST1;
- C1QC;
- CX3CR1.

The combined use of clustering, reference-based annotation, and marker expression was intended to reduce reliance on a single criterion for defining the TAM population.

4.11 TAM sample filtering

The downstream pseudo-bulk analysis required sufficient TAM representation within individual samples. Samples containing fewer than 20 TAM cells were excluded from the pseudo-bulk analysis.

This filtering step was applied at the sample level rather than removing individual TAM cells based on expression values. Its purpose was to reduce the instability that could result from generating pseudo-bulk profiles from very small numbers of cells.

4.12 Pseudo-bulk aggregation

For each sample, raw TAM counts were summed across cells:

$$Y_{gs} = \sum_{c \in s} Y_{gc},$$

where Y_{gs} represents the aggregated count for gene g in sample s, and Y_{gc} represents the raw count for gene g in cell c.

This generated a gene-by-sample matrix suitable for DESeq2.

Importantly, the sample was treated as the biological replicate. This avoids treating thousands of cells originating from the same patient as independent observations.

4.13 Subtype definition for pseudo-bulk analysis

The primary pseudo-bulk comparisons were performed among ER+, HER2+, and TNBC samples. BRCA1-associated TNBC samples were incorporated into the TNBC category for these comparisons.

The resulting pairwise contrasts were:

1. HER2+ versus ER+;
2. TNBC versus ER+;
3. TNBC versus HER2+.

4.14 DESeq2 differential expression

DESeq2 was used to model pseudo-bulk count data. Genes were retained when at least two samples had counts of 10 or greater:

$$\sum_{s} I(Y_{gs} \geq 10) \geq 2.$$

The filtered count matrix was then analyzed using DESeq2. Differential expression results were ordered according to adjusted p-value.

The three comparisons were obtained using the following contrast structure:

```
results(dds, contrast = c("group", "HER2", "ER"))
results(dds, contrast = c("group", "TNBC", "ER"))
results(dds, contrast = c("group", "TNBC", "HER2"))
```

The purpose of these analyses was discovery: to identify genes whose expression patterns differed among the subtype groups within TAMs.

4.15 Candidate gene selection

S100A9 was selected as a candidate gene from the TAM differential expression results. The gene was then carried forward into the independent cohort analysis.

The candidate-gene strategy was deliberately followed by an independent evaluation rather than interpreting the single-cell result alone as evidence of clinical relevance.

4.16 METABRIC data

The METABRIC analysis used three data sources:

- clinical sample metadata;
- clinical patient metadata;
- Illumina microarray gene-expression data.

Sample-level and patient-level clinical data were merged using patient identifiers, after which S100A9 expression was extracted and joined with the clinical metadata.

4.17 METABRIC subtype classification

Tumours were classified into ER+, HER2+, and triple-negative categories using receptor-status information. For the main subtype analysis, the comparison was restricted to these three groups.

The sample sizes used for the S100A9 subtype analysis were:

Table 1: METABRIC samples included in the S100A9 subtype analysis.

Subtype	Number of tumours
ER+	1,398
HER2+	134
Triple Negative	320
Total	1,852

4.18 METABRIC subtype statistics

Because expression distributions were not assumed to be normally distributed, a Kruskal–Wallis test was used to evaluate the overall difference in S100A9 expression across the three subtype groups.

Where an overall difference was observed, Dunn's post-hoc test with Benjamini–Hochberg adjustment was used for pairwise comparisons.

4.19 Lymph-node analysis

Lymph-node involvement was defined from the number of positive lymph nodes examined. Tumours with zero positive nodes were classified as having no lymph-node involvement, whereas tumours with one or more positive nodes were classified as having lymph-node involvement.

S100A9 expression was compared between these groups using the Wilcoxon rank-sum test.

The purpose of this analysis was to examine whether the candidate gene showed an association with a clinically relevant feature of disease spread. It was not interpreted as proof of metastatic causation.

4.20 Software

The principal software and packages used in the analysis are summarized in Table 2.

Table 2: Major computational tools used in the project.

Tool / package	Role in the analysis
R	Main statistical programming environment
Seurat	scRNA-seq data processing, normalization, PCA, UMAP, neighbours, clustering, and visualization
Harmony	Batch/sample integration in low-dimensional space
SingleR	Reference-based cell-type annotation
celldex	Reference datasets used for cell annotation
DESeq2	Sample-level pseudo-bulk differential expression
ggplot2	Statistical visualization and publication-style plots
patchwork	Combination of multiple plots into composite figures
ComplexHeatmap / circlize	Heatmap and expression visualization utilities used during exploratory analysis
cluster	Silhouette-based clustering evaluation
FSA	Dunn's post-hoc test for METABRIC subtype comparisons
TCGAbiolinks	TCGA data-access and clinical-data exploration during downstream analysis

5 Results

5.1 Single-cell dataset and preprocessing

The study analyzed 30 breast cancer samples distributed across three primary subtype groups: 16 ER+, 6 HER2+, and 8 TNBC samples. The dataset was processed as individual 10X Genomics samples before being merged for downstream analysis.

Quality control removed cells that did not meet the predefined gene-detection and mitochondrial-content thresholds. Following QC, the data were normalized and highly variable genes were selected for dimensionality reduction.

The preprocessing stage produced a common Seurat object containing the selected samples and their associated metadata. This object formed the basis for the integrated clustering analysis.

5.2 Integrated single-cell representation

PCA was used to summarize major transcriptional axes before Harmony integration. Harmony was subsequently applied to reduce sample-associated variation in the low-dimensional representation.

The corrected UMAP provided a visual representation of the cellular landscape. Cells positioned near each other in the UMAP generally had more similar transcriptional profiles, whereas separated regions represented populations with more distinct transcriptional states.

Clustering parameter evaluation using silhouette analysis supported the use of $k = 15$ and a resolution of 0.15 for the final clustering structure.

5.3 Cell-type annotation

SingleR produced reference-based cell-type annotations for the integrated dataset. These annotations were used to characterize the major cellular populations and to identify candidate immune populations for further analysis.

Marker expression was examined alongside SingleR results. In particular, immune markers such as PTPRC and LST1, epithelial markers such as EPCAM and keratins, endothelial markers such as PECAM1 and VWF, and stromal markers including COL1A1, COL1A2, DCN, and LUM were used to interpret the major populations.

5.4 Identification of the EPCAM-negative population

The EPCAM-negative subset provided a focused representation of non-epithelial cells. A separate dimensionality-reduction and clustering workflow was performed rather than assuming that the original clusters would remain optimal after subsetting.

This additional analysis improved the ability to inspect immune and stromal populations and supported downstream identification of macrophage-associated cells.

5.5 Identification of TAMs

Cluster 2 within the EPCAM-negative analysis was identified as a macrophage population using SingleR annotation and further supported by macrophage-associated marker expression.

CD14 and CD68 were used as key markers for selecting TAM cells. LST1, C1QC, and CX3CR1 were also examined as additional markers associated with macrophage identity.

This step produced the TAM population used for pseudo-bulk analysis.

5.6 Pseudo-bulk differential expression

TAM expression counts were aggregated by sample. Samples with fewer than 20 TAM cells were excluded to avoid generating unstable sample-level profiles.

The resulting pseudo-bulk matrix allowed the analysis to be performed using the biological sample as the replicate. DESeq2 was used to evaluate the following comparisons:

- HER2+ versus ER+;
- TNBC versus ER+;
- TNBC versus HER2+.

The differential expression results provided the discovery set from which S100A9 was selected for further investigation.

5.7 S100A9 expression across METABRIC subtypes

The independent METABRIC analysis included 1,398 ER+ tumours, 134 HER2+ tumours, and 320 triple-negative tumours.

S100A9 expression showed a subtype-associated pattern, with higher expression observed in HER2+ and triple-negative tumours compared with ER+ tumours. The overall subtype comparison was evaluated using the Kruskal–Wallis test, followed by Dunn's multiple-comparison procedure.

The independent cohort therefore reproduced the principal direction of the subtype-associated pattern identified through the TAM pseudo-bulk analysis.

5.8 S100A9 and lymph-node involvement

S100A9 expression was also evaluated according to lymph-node involvement in the METABRIC cohort. The analysis showed a statistically significant difference between tumours without lymph-node involvement and those with lymph-node involvement:

$$p = 0.0002092.$$

Although statistically significant, the magnitude of the observed difference was modest. This distinction is important because statistical significance in a large cohort does not necessarily indicate a large biological effect. Therefore, the result is best interpreted as evidence of an association that warrants additional investigation rather than evidence of a strong predictive or prognostic effect.

5.9 Summary of principal findings

Table 3: Summary of the principal findings.

Analysis	Principal finding
Single-cell clustering	Distinct cellular populations were resolved after QC, dimensionality reduction, Harmony correction, and graph-based clustering.
Cell annotation	SingleR and marker expression identified major cellular populations, including macrophage-associated cells.
EPCAM-negative analysis	A focused non-epithelial population was analyzed to facilitate immune-cell characterization.
TAM identification	Cluster 2 was identified as a macrophage population and further evaluated using CD14, CD68, LST1, C1QC, and CX3CR1.
Pseudo-bulk DE	Sample-level aggregation enabled subtype comparisons using DESeq2 while retaining biological samples as replicates.
Candidate gene	S100A9 was selected for independent investigation.
METABRIC subtype analysis	S100A9 expression was higher in HER2+ and triple-negative tumours than in ER+ tumours.
Lymph-node analysis	S100A9 expression showed a statistically significant but modest association with lymph-node involvement ($p = 0.0002092$).

6 Discussion

6.1 Integration of single-cell and bulk-cohort evidence

The central strength of this project is the connection between two complementary analytical scales. The single-cell analysis was used for discovery and cellular localization, whereas the METABRIC analysis was used to evaluate whether a candidate gene showed a similar expression pattern in a large independent cohort.

The single-cell analysis was specifically focused on TAMs. This is important because S100A9 was not selected solely because it was highly expressed in breast tumours generally. It emerged from a workflow designed to compare transcriptional profiles within a defined macrophage-associated population across breast cancer subtypes.

The subsequent METABRIC analysis addressed a different question. Because METABRIC expression data are bulk measurements, they cannot independently establish that S100A9 expression is derived from TAMs. Nevertheless, observing a similar subtype-associated expression pattern in a large independent cohort provides supportive evidence that the gene is associated with the biological differences between breast cancer subtypes.

6.2 Biological interpretation of the subtype pattern

The observed higher S100A9 expression in HER2+ and triple-negative tumours relative to ER+ tumours is consistent with the broader concept that inflammatory and immune-associated transcriptional programs can differ across breast cancer subtypes.

The result is particularly relevant to TNBC because TNBC lacks ER, PR, and HER2 expression and consequently does not benefit from endocrine or HER2-targeted therapies. The identification of genes associated with TNBC biology can therefore provide hypotheses for future mechanistic investigation.

However, the present study does not establish that S100A9 drives TNBC biology. Expression differences alone cannot distinguish whether a gene is a causal driver, a consequence of the tumour microenvironment, a marker of a particular cellular state, or a correlate of another biological process.

6.3 Interpretation of the lymph-node finding

The association between S100A9 expression and lymph-node involvement provides an additional clinical context for the candidate gene. The result was statistically significant, with $p = 0.0002092$, but the observed effect was modest.

This is an important distinction. A large cohort can provide substantial statistical power to detect relatively small expression differences. Consequently, a low p-value should not be interpreted as evidence of a large or clinically important effect without considering the effect size and distribution of the measurements.

The lymph-node result therefore supports the hypothesis that S100A9 may be related to features of tumour dissemination or an associated tumour microenvironmental state, but it does not establish S100A9 as a marker of metastasis.

6.4 Why pseudo-bulk was appropriate

The pseudo-bulk design is a particularly important methodological feature of the project. If every TAM cell had been treated as an independent replicate, the analysis could have been dominated by the number of cells rather than the number of biological samples.

By aggregating TAM counts within each sample, the analysis preserved the patient/sample as the unit of replication. DESeq2 could then estimate dispersion across biological samples and test subtype-associated expression differences using a framework appropriate for count data.

This approach also makes the biological interpretation more straightforward: a gene is considered subtype-associated when its sample-level expression differs systematically between the groups, rather than simply because many cells from one sample express the gene.

6.5 Role of Harmony and clustering evaluation

The use of Harmony was intended to reduce unwanted sample-associated variation before clustering. This was particularly relevant because the study combined multiple patients and breast cancer subtypes.

However, integration methods can potentially remove biological variation if they are applied too aggressively. The use of silhouette analysis to examine clustering structure provided an additional quantitative criterion for choosing clustering parameters. The final clustering was therefore not selected solely based on visual appearance.

Nevertheless, clustering quality metrics should not be interpreted as proof that a cluster corresponds to a biologically discrete cell type. Biological annotation still requires reference

labels and marker-gene evidence, which is why SingleR and macrophage markers were used together.

6.6 Importance of independent validation

Independent cohort evaluation is important when identifying candidate marker genes. A gene that appears differentially expressed in a relatively small single-cell dataset may reflect sample composition, technical variation, or cohort-specific characteristics.

The METABRIC analysis addressed this issue by testing S100A9 in a substantially larger independent breast cancer cohort. The reproduction of the subtype-associated direction strengthens the hypothesis that the observed pattern is not limited to the original single-cell dataset.

At the same time, independent validation should not be overstated. METABRIC is based on bulk expression data and therefore cannot independently validate the TAM-specific cellular origin of S100A9. The result is best described as independent support for the subtype-associated expression pattern.

6.7 Potential relevance to TNBC

The high S100A9 expression observed in TNBC is potentially interesting because TNBC has fewer established receptor-directed treatment options than ER+ and HER2+ disease. This creates a rationale for investigating alternative biological pathways and candidate targets.

S100A9 should therefore be regarded as a candidate for further study rather than an established therapeutic target. Functional experiments would be required to determine whether altering S100A9 activity changes tumour-cell behaviour, macrophage states, tumour–immune interactions, or treatment response.

6.8 Interpretation of the candidate gene

The project supports three related observations:

1. S100A9 was identified through a TAM-focused single-cell pseudo-bulk discovery analysis.
2. S100A9 showed a consistent subtype-associated expression pattern in METABRIC, with higher expression in HER2+ and triple-negative tumours than in ER+ tumours.
3. S100A9 expression was associated with lymph-node involvement in METABRIC, although the magnitude of this association was modest.

Taken together, these findings make S100A9 a reasonable candidate for further biological investigation. They do not, however, demonstrate that S100A9 is a therapeutic target, a prognostic biomarker, or a direct mediator of metastasis.

7 Limitations and Future Work

7.1 Limitations

7.1.1 Limited single-cell sample size

The single-cell discovery dataset contained 30 samples distributed across three main subtype groups. Although the number of cells can be large, the number of independent biological samples remains the limiting factor for subtype-level inference.

Future work should include larger numbers of patients per subtype to improve statistical power and the stability of candidate-gene ranking.

7.1.2 TAM identification is computational rather than experimentally confirmed

The TAM population was identified using clustering, SingleR annotation, and marker-gene expression. These approaches provide strong computational evidence for cell identity, but they do not constitute experimental validation.

Flow cytometry, immunohistochemistry, spatial transcriptomics, or other orthogonal approaches could provide additional confirmation of the cellular population and its localization.

7.1.3 Bulk validation does not establish TAM-specific expression

The METABRIC expression measurements represent bulk tumour tissue. Therefore, the observed S100A9 expression pattern cannot be attributed exclusively to TAMs.

Future studies should examine S100A9 expression directly within macrophages in tissue, ideally using spatial transcriptomics, multiplex immunofluorescence, or single-cell/spatial datasets containing matched clinical information.

7.1.4 Association does not establish causation

The observed relationships between S100A9 expression, subtype, and lymph-node involvement are associative. They do not establish a causal role for S100A9 in tumour progression or metastasis.

Functional experiments are required to determine whether S100A9 contributes directly to tumour progression, immune-cell recruitment, macrophage polarization, or metastatic behaviour.

7.1.5 Clinical relevance requires additional endpoints

The current analysis explored lymph-node involvement but did not establish prognostic value. A clinically useful biomarker would require evaluation in relation to outcomes such as overall survival, disease-free survival, distant metastasis, treatment response, or recurrence.

7.2 Future work

The following studies would strengthen the biological and clinical interpretation of the findings:

1. **Spatial validation:** Investigate S100A9 expression directly within lymph-node tissue and tumour tissue to determine its cellular localization and relationship with macrophages.
2. **Larger independent cohorts:** Evaluate S100A9 in additional breast cancer cohorts to determine whether the subtype-associated pattern is reproducible.
3. **Survival analysis:** Investigate the relationship between S100A9 expression and overall survival, disease-free survival, and other clinically relevant outcomes.

4. **Functional studies:** Determine whether S100A9 perturbation alters tumour-cell behaviour or macrophage-associated phenotypes.
5. **Mechanistic studies:** Investigate pathways through which S100A9 may influence tumour–immune interactions and metastatic processes.
6. **Therapeutic evaluation:** Assess whether S100A9 represents a therapeutically actionable pathway, particularly in TNBC, using appropriate experimental models.
7. **Cellular deconvolution or spatial methods:** Determine whether bulk-cohort S100A9 expression is driven primarily by macrophages or by additional tumour or stromal populations.

8 Conclusion

This project developed an integrated computational workflow to investigate candidate marker genes associated with breast cancer subtypes, with a specific focus on tumour-associated macrophages. Single-cell RNA-seq data from ER+, HER2+, and TNBC samples were processed using Seurat, Harmony, SingleR, and graph-based clustering, followed by EPCAM-negative subsetting and macrophage identification using cluster structure and macrophage-associated markers. Sample-level pseudo-bulk aggregation and DESeq2 differential expression were then used to compare TAM transcriptional profiles across breast cancer subtypes while retaining biological samples as the units of replication.

S100A9 emerged as a candidate subtype-associated gene and was subsequently evaluated in the independent METABRIC cohort. The METABRIC analysis reproduced the overall subtype-associated pattern, with higher S100A9 expression in HER2+ and triple-negative tumours compared with ER+ tumours. S100A9 expression was also statistically associated with lymph-node involvement, although the magnitude of the difference was modest. These findings provide a rationale for further investigation of S100A9, particularly in TNBC and in the context of tumour–immune interactions, while avoiding the stronger claim that the current data establish S100A9 as a therapeutic or prognostic marker.

The main contribution of the project is therefore not the identification of a definitive biomarker, but the development of a reproducible analytical framework linking single-cell cellular resolution, sample-level differential expression, and independent cohort evaluation. Further spatial, functional, and clinical studies are required to determine the biological significance and potential translational relevance of S100A9.

9 Reproducibility and Computational Organization

The analysis was organized into five main Quarto scripts. The workflow is sequential for the single-cell component, while the METABRIC analysis forms an independent downstream validation component.

Table 4: Analysis scripts and their primary roles.

Script	Main purpose	Principal output
01_Data_Loading_and_QC	Import 10X data, merge samples, QC, normalization, highly variable genes	QC-normalized Seurat object
02_Clust_Harm_Ann_HER2vsERvsTN	PCA, UMAP, Harmony, clustering, silhouette analysis, SingleR annotation	Integrated annotated Seurat object
03_EPCAMNeg_BRCA1_HER2vsERvsTN	EPCAM-negative subsetting, reintegration, clustering, annotation	EPCAM-negative Seurat object
04_TAMs_HER2vsERvsTN	TAM identification, sample filtering, pseudo-bulk aggregation, DESeq2	Pairwise differential expression results
05_METABRIC	S100A9 subtype and lymph-node analysis in an independent cohort	METABRIC statistics and figures

Intermediate Seurat objects were saved as RDS files to allow downstream scripts to begin from processed objects without repeating earlier computationally expensive steps.

The analysis used a fixed random seed where stochastic procedures were involved:

```
set.seed(123)
```

Absolute file paths were used in the original analysis environment. For reproducibility on another computer, these paths should be converted to project-relative paths or updated to the corresponding local directory structure.

10 References

References

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. 2018;36:411–420. doi:10.1038/nbt.4096.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*. 2019;16:1289–1296. doi:10.1038/s41592-019-0619-0.

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*. 2019;20:163–172. doi:10.1038/s41590-018-0276-y.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15:550. doi:10.1186/s13059-014-0550-8.

Stuart T, Satija R. Integrative single-cell analysis. *Nature Reviews Genetics*. 2019;20:257–272. doi:10.1038/s41576-019-0093-7.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486:346–352. doi:10.1038/nature10983.

A Appendix: Key Analysis Parameters

Table 5: Key parameters used throughout the single-cell workflow.

Parameter	Setting
Minimum detected genes	500
Maximum mitochondrial percentage	20%
Highly variable genes	2,000
PCA dimensions calculated	30
Dimensions used for main analysis	1–20
Harmony correction	Yes
Harmony theta	2
Nearest-neighbour values evaluated	10, 15, 20, 25, 30, 40
Final k	15
Final clustering resolution	0.15
TAM cluster	Cluster 2
Key TAM markers	CD14, CD68
Additional TAM markers	LST1, C1QC, CX3CR1
Minimum TAM cells per sample	20
Pseudo-bulk unit	Biological sample
DESeq2 contrasts	HER2/ER, TNBC/ER, TNBC/HER2

B Appendix: Suggested Figure Placement

The report can be strengthened by adding the actual project figures in the following locations:

1. **QC figure:** violin plots showing $nFeature_RNA$ and mitochondrial percentage before and after filtering.
2. **Integrated UMAP:** UMAP coloured by sample and by subtype.

3. **Cluster annotation:** UMAP showing Seurat clusters and SingleR labels.
4. **EPCAM-negative UMAP:** focused UMAP showing the non-epithelial cellular landscape.
5. **TAM identification:** UMAP or feature plots showing CD14, CD68, LST1, C1QC, and CX3CR1.
6. **Pseudo-bulk differential expression:** volcano plots or ranked gene plots for the three DESeq2 contrasts.
7. **S100A9 validation:** METABRIC S100A9 expression across ER+, HER2+, and triple-negative tumours.
8. **Lymph-node analysis:** S100A9 expression according to lymph-node involvement.

C Appendix: Interpretation Guide

For clarity, the principal claims supported by the current analysis can be separated from claims requiring additional evidence.

Table 6: Interpretation of findings and appropriate strength of claims.

Finding	Appropriate interpretation
S100A9 identified in TAM pseudo-bulk analysis	Candidate subtype-associated gene
Higher S100A9 in HER2+ and TNBC in METABRIC	Independent support for a subtype-associated expression pattern
S100A9 associated with lymph-node involvement	Statistical association; magnitude should be considered
Higher expression in TNBC	Rationale for further biological investigation
Potential therapeutic target	Hypothesis requiring functional and therapeutic validation
Prognostic biomarker	Not established by the current analysis; requires outcome-based validation
TAM-specific expression in METABRIC	Cannot be established from bulk expression alone
Causal role in metastasis	Not established by expression association