Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Identifying the specifically expressed lncRNAs in the Tumor Microenvironment

**Xingli Guo**

xlguo@mail.xidian.edu.cn

Xidian University

**Zijia Zhang**

Xidian University

**Yujie Ren**

Xidian University

**YaXuan Hu**

Xidian University

**Bingbo Wang**

Xidian University

**Research Article**

**Keywords:** tumor microenvironment, single-cell RNA sequencing, lncRNA, T cells, subtype-specific genes

**Posted Date:** June 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9835966/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at BMC Bioinformatics on October 13th, 2026. See the published version at https://doi.org/10.1186/s12859-026-06659-x.

# Identifying the specifically expressed lncRNAs in the Tumor Microenvironment

Xingli Guo[1*], Zijia Zhang[1], Yujie Ren[1], YaXuan Hu[1], Bingbo Wang[1]

[1*]School of Computer Science and Technology, Xidian University, 266 Xinglong Section of Xifeng Road, Xi'an, 710126, Shaanxi, China.

*Corresponding author(s). E-mail(s): xlguo@mail.xidian.edu.cn;
Contributing authors: 23009200586@stu.xidian.edu.cn;
renran520240@gmail.com; huyx1221@163.com;
bingbowang@xidian.edu.cn;

**Abstract**

**Background:** Understanding immune cell heterogeneity in the tumor microenvironment (TME) is crucial for studying tumor immunity and improving immunotherapy strategies. Diverse immune cell populations, particularly T cells, exhibit functional heterogeneity and play distinct roles in the TME, and even within the same immune cell lineage, multiple subtypes or states with distinct transcriptional and functional characteristics can coexist across different cancer types and individual patients. Recent single-cell studies have further revealed remarkable heterogeneity among tumor-infiltrating T cells, particularly within the $CD8^+$ T-cell compartment. Since gene expression programs reflect the functional properties of these cellular subtypes, identifying subtype-specific regulatory molecules is essential for understanding their roles in antitumor immunity. As important regulatory molecules in immune-related molecular mechanisms, long non-coding RNAs (lncRNAs), which often exhibit strong cell-type and context specificity, may serve as important regulators of subtype-specific immune programs. Therefore, identifying lncRNAs that are specifically associated with different subtypes of tumor-infiltrating $CD8^+$ T cells may provide important insights into the regulatory mechanisms of immune responses in cancer. However, detecting subtype-associated lncRNAs from single-cell RNA sequencing (scRNA-seq) data remains challenging due to sparse expression patterns and the low abundance of many lncRNAs.

**Results:** In this study, we developed SDSGI (Surprise-Driven Subtype-Specific Gene Identification), a computational method for identifying subtype-specific

genes and lncRNAs from scRNA-seq data. This framework integrates two complementary scoring strategies: an expression-based specificity score derived from the $\tau$ index and a surprise-information-based score that captures statistically informative subtype-enrichment signals. By combining expression specificity with statistical evidence, the method improves the detection of low-abundance but subtype-associated genes, including lncRNAs that may be missed by conventional differential expression analysis. We applied SDSGI to tumor-infiltrating $CD8^+$ T-cell datasets across multiple cancer types. The results show that the method identifies clear subtype-specific gene signatures and biologically meaningful lncRNAs associated with immune cell states. Specifically, 6,899 subtype-specific genes in breast cancer including 497 lncRNAs; 5,035 genes in B-cell lymphoma including 361 lncRNAs; and 11,092 genes in esophageal cancer including 944 lncRNAs have been identified by our method. Functional enrichment, co-expression network analysis, and survival validation further suggested that the identified lncRNAs are associated with immune regulation, subtype-specific functional modules, and cancer prognosis.
**Conclusions:** These results indicate that SDSGI provides a useful framework for discovering subtype-specific lncRNAs in the tumor immune microenvironment.

**Keywords:** tumor microenvironment, single-cell RNA sequencing, lncRNA, T cells, subtype-specific genes

# 1 Introduction

The tumor microenvironment (TME) is a highly complex ecosystem composed of diverse immune and stromal cell populations whose interactions play critical roles in tumor progression and therapeutic response[1, 2]. Among these immune components, T cells exhibit remarkable functional heterogeneity, including naïve, effector, memory, and exhausted states, each characterized by distinct transcriptional programs [3–5]. Understanding the molecular mechanisms underlying this heterogeneity is essential for elucidating tumor–immune interactions and improving immunotherapy strategies[6].

Long non-coding RNAs (lncRNAs), defined as non-coding transcripts longer than 500 nucleotides, have emerged as important regulators in diverse biological processes and diseases [7, 8]. Increasing evidences indicate that lncRNAs participate in immune regulation, T-cell differentiation, and tumor-associated immune responses[7–9]. In particular, lncRNAs have been implicated in modulating immune cell activation, cytokine signaling pathways, and T-cell exhaustion within the tumor microenvironment [9]. For example, the lncRNA NKILA has been reported to promote tumor immune evasion by increasing the sensitivity of tumor-specific cytotoxic T cells and type 1 helper T cells to activation-induced cell death through inhibition of NF-$\kappa$B activity [10]. In addition, the immunogenic lncRNA LIMIT has been shown to regulate MHC-I antigen-presentation machinery and antitumor immunity, suggesting that lncRNAs can affect $CD8^+$ T-cell-mediated immune responses and immunotherapy efficacy [11]. Despite their functional importance, the landscape of lncRNAs in tumor-infiltrating immune cells remains largely unexplored.

Recent advances in single-cell RNA sequencing (scRNA-seq) have enabled the high-resolution characterization of cellular heterogeneity in complex tissues such as tumors [3, 4, 12, 13]. By capturing transcriptomic profiles at the single-cell level, scRNA-seq provides unprecedented opportunities to identify subtype-specific regulatory genes across immune cell populations. However, functionally elucidating lncRNAs based on scRNA-seq data remains challenging. Compared with protein-coding genes, lncRNAs are typically expressed at lower levels and exhibit sparse expression patterns, especially in single-cell RNA-seq data[14, 15]. Most existing approaches rely on differential expression (DE) analysis to identify relevant genes in the specific context, but these methods may overlook low-abundance transcripts because of limited statistical power and strong dependence on expression magnitude. Alternatively, specificity-based metrics such as the $\tau$ index have been used to quantify the degree of cell-type specificity of gene expression [16]. As a result, important specific signals—especially those involving low-abundance lncRNAs may remain undetected.

To address these limitations, we propose SDSGI, a computational framework designed to identify specifically expressed genes from single-cell transcriptomic data. SDSGI integrates the $\tau$ specificity index with a statistical surprise-information metric derived from significance testing, thereby combining expression specificity with statistical confidence. This strategy enables the detection of specifically expressed genes, including low-abundance lncRNAs that are often missed by conventional approaches.

Our method is applied to single-cell RNA-seq data from tumor-infiltrating T cells, to identify specifically expressed gene in the subtype of T cells, which are called subtype-specific genes. These genes showed clear subtype-biased expression patterns, including low-abundance transcripts that were less evident using conventional differential expression analysis. Functional analyses further reveal that the genes detected by SDSGI are strongly associated with immune-related biological processes and pathways. In addition, trajectory and network analyses highlight potential regulatory roles of several candidate lncRNAs in T-cell functional states within the tumor microenvironment. These findings indicate that SDSGI provides a useful framework for uncovering context-specific regulatory programs in complex cellular systems.

# 2 Materials and Methods

## 2.1 Overview of the SDSGI framework

We developed a computational method, named Surprise-Driven Subtype-Specific Gene Identification (SDSGI), to identify immune cell subtype-specific genes from single-cell RNA sequencing (scRNA-seq) data in the tumor microenvironment (TME). SDSGI is motivated by tissue-specificity analysis and self-information theory[17–19], and is designed to address two major challenges of scRNA-seq data: the high sparsity caused by extensive zero inflation and dropout events[12, 14], and the low abundance of many transcripts, particularly long non-coding RNAs (lncRNAs), whose biologically meaningful subtype-biased signals can be masked by highly expressed genes and are therefore often missed by conventional analyses[15, 20].

The overall workflow of SDSGI is illustrated in Figure 1. Briefly, scRNA-seq data from tumor-infiltrating T cells are first collected and integrated in the work of Zheng

et al. [13]. Especially, the subtype annotation for T cells are given in that work. Here some preprocessings are performed including standard quality-control and normalization procedures. SDSGI then quantifies subtype specificity using two complementary scoring strategies. One is a $\tau$-based specificity score, which evaluates how selectively a gene is expressed across cell subtypes. The other is a surprise-information-based score, which transforms the statistical significance of subtype-biased expression into an information-theoretic measure and is particularly effective for detecting genes with weak but subtype-enriched expression signals.

Based on the two scoring schemes, a set of subtype-specific genes can be identified. The low expressed lncRNAs with specific expression patterns are included in these gene sets. Functional enrichment analysis is performed to investigate potential biological mechanisms underlying tumor-infiltrating T-cell subtypes. A co-expression network is also constructed using these genes to characterize subtype-associated functional modules. And then potential functions of lncRNAs can be inferred.

## 2.2 Data collection and preprocessing

The single-cell RNA sequencing (scRNA-seq) data of tumor-infiltrating T cells (GSE156728) downloaded from the GEO database[21]. These datasets were all generated using the 10X Genomics platform and spanning 11 cancer types and 17 T-cell subtypes, thus providing a pan-cancer landscape of T-cell heterogeneity within the tumor microenvironment. Detailed sample information and tissue sources are provided in Additional file 1: Tables S1–S3.

In our method, several preprocessings are performed as follows,

1. Cells with fewer than 200 detected genes were removed.
2. Genes expressed in fewer than three cells were discarded.
3. Only protein-coding genes and lncRNAs were retained, the gene annotation are based on GENCODE(V44) and LNCipedia(V5.2).

Because scRNA-seq data are highly sparse and affected by technical noise, cell quality was further evaluated using biologically meaningful indicators, particularly the proportion of expressed mitochondrial genes is considered in each cell. Quality-control thresholds were determined in a data-dependent manner according to the distribution of these metrics and were therefore not fixed across cancer types. For breast cancer $CD8^+$ T cells, the mitochondrial proportion threshold was set to 10%, and cells with more than 3000 detected genes which refer to all retained genes detected in each cell after gene annotation and filtering were excluded as potential outliers [22, 23]. Detailed descriptions of the quality-control metrics and the rationale for parameter selection are provided in Additional file 1: Figure S1 and Supplementary Methods A.2–A.3. The post-preprocessing data statistics are summarized in Table 1.

## 2.3 Identification of subtype-specific genes

### 2.3.1 Expression specificity measurement using $\tau$ score

To quantify subtype-restricted expression patterns, SDSGI adopts a subtype-specificity score motivated by the classical tissue-specificity index ($\tau$). Here, the tissue

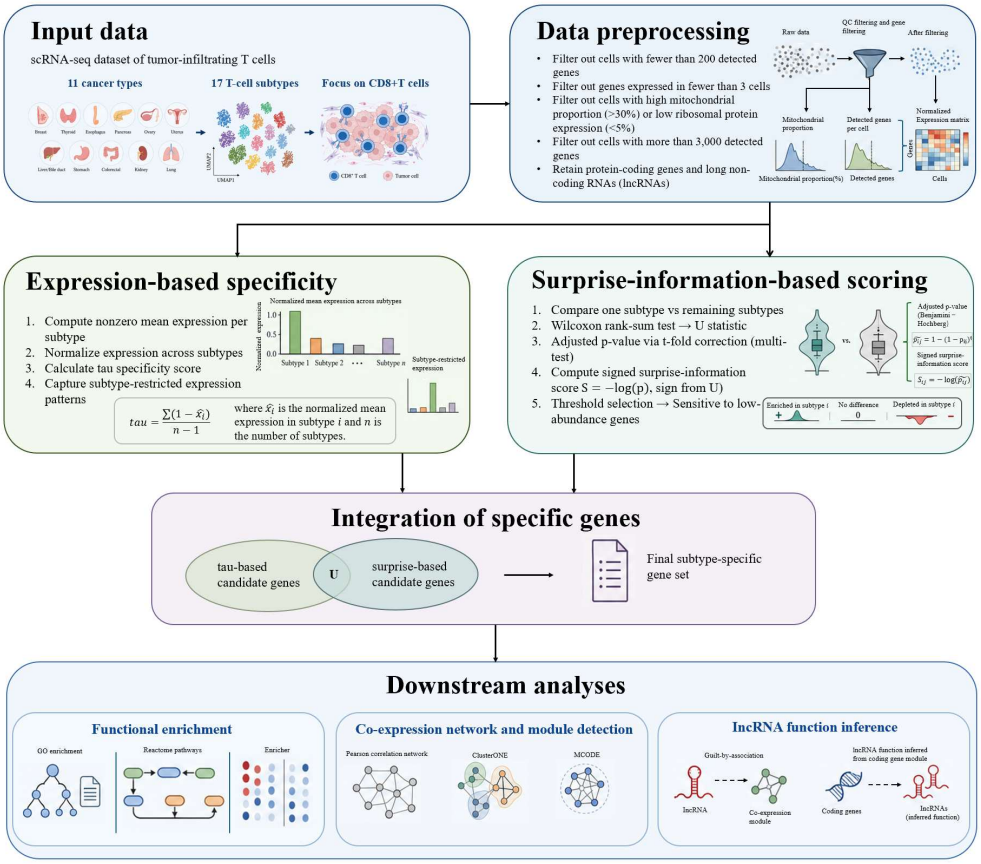



**Fig. 1** Overview of the SDSGI framework for identifying subtype-specific lncRNAs from scRNA-seq data. The workflow includes data preprocessing, subtype-specific gene identification using the $\tau$ specificity index and surprise-information analysis, followed by downstream functional analysis and validation.

dimension in the original $\tau$ index was replaced by the immune-cell-subtype dimension, yielding

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1}, \tag{1}$$

where $n$ denotes the number of immune cell subtypes and $\hat{x}_i$ is the normalized expression level of a given gene in subtype $i$, computed as

$$\hat{x}_i = \frac{x_i}{\max_{1\le i\le n}(x_i)}. \tag{2}$$

Here, $x_i$ denotes the subtype-level expression summary of the given gene in subtype $i$. To improve robustness to sparsity and dropout in scRNA-seq data, $x_i$ was calculated as the mean of nonzero expression values within subtype $i$. Under this formulation, genes with concentrated expression in one or a few subtypes receive larger

**Table 1** Summary of $CD8^+$ T-cell transcriptome data after quality control

| ID | Tumor Abbreviation | Tumor Type | #Cell | #Gene | #lncRNA |
|---|---|---|---|---|---|
| 1 | BC | Breast Cancer | 4,291 | 14,068 | 2,018 |
| 2 | BCL | B-cell Lymphoma | 3,482 | 13,826 | 1,890 |
| 3 | CHOL | Cholangiocarcinoma | 300 | 11,431 | 531 |
| 4 | ESCA | Esophageal Cancer | 12,526 | 16,040 | 2,854 |
| 5 | FTC | Follicular Thyroid Cancer | 767 | 10,986 | 787 |
| 6 | MM | Multiple Myeloma | 8,629 | 15,403 | 2,515 |
| 7 | OV | Ovarian Cancer | 3,517 | 13,636 | 1,742 |
| 8 | PACA | Pancreatic Cancer | 5,957 | 14,734 | 2,280 |
| 9 | RC | Rectal Cancer | 16,544 | 16,090 | 2,819 |
| 10 | THCA | Thyroid Cancer | 33,450 | 17,518 | 3,647 |
| 11 | UCEC | Uterine Corpus Endometrial Carcinoma | 19,926 | 16,522 | 3,119 |

$\tau$ values, whereas broadly expressed genes receive smaller values [24]. More details on the comparison of different expression summaries are provided in Additional file 1: Supplementary Methods B and Figure S2.

### 2.3.2 Surprise-information-based statistical scoring

Although the expression-based $\tau$ score is effective for identifying highly subtype-specific genes, it remains less sensitive to low-abundance transcripts, especially lncRNAs, in sparse scRNA-seq data. To better capture subtype-associated signals that may be statistically informative despite weak absolute expression, SDSGI introduces a surprise-information-based statistical score.

For each gene $i$ and subtype $j$, let $G_{ij}$ denote the expression vector of gene $i$ in subtype $j$, and let $\bar{G}_{ij}$ denote the expression vector of the same gene in all remaining subtypes. To evaluate whether gene $i$ shows subtype-biased expression in subtype $j$, SDSGI compares $G_{ij}$ with $\bar{G}_{ij}$ using the Wilcoxon rank-sum test. This non-parametric test does not require normally distributed expression values and is therefore suitable for sparse and zero-inflated scRNA-seq data. The computational procedure for the surprise-information score is summarized in Algorithm 1.

In this way, the surprise-information score quantifies subtype-associated statistical deviations while also preserving the direction of change, thereby complementing the expression-based $\tau$ score in the identification of low-abundance subtype-associated genes. Details on the statistical derivation, adjusted p-value transformation, sign assignment of the surprise-information score, and threshold selection are provided in Additional file 1: Supplementary Methods C.1 and Figure S3.

### 2.3.3 Integration of subtype-specific genes

The final subtype-specific gene set in SDSGI was defined as the union of the candidate genes identified by the expression-based $\tau$ score and the surprise-information-based score. This integration strategy allows SDSGI to retain both genes with clear subtype-restricted expression patterns and genes with weaker absolute expression but

**Algorithm 1** Surprise-information-based statistical scoring in SDSGI

**Require:** Gene expression vectors $G_{ij}$ and $\bar{G}_{ij}$ for gene $i$ in subtype $j$
**Ensure:** Signed surprise-information score $S_{ij}$

1: Perform Wilcoxon rank-sum test on $G_{ij}$ and $\bar{G}_{ij}$
2: Obtain test statistic $U$ and raw $p$-value $p_{ij}$
3: Compute adjusted $p$-value:

$$\tilde{p}_{ij} = 1 - (1 - p_{ij})^t$$

4: Compute surprise-information score:

$$S_{ij} = -\log(\tilde{p}_{ij})$$

5: Compute the expected Wilcoxon statistic under the null hypothesis:

$$E(U) = \frac{n_1 n_2}{2}$$

6: **if** $U < E(U)$ **then**
7: $\quad S_{ij} \leftarrow -S_{ij}$
8: **end if**
9: **return** $S_{ij}$

statistically informative subtype bias, thereby improving the sensitivity of subtype-specific gene identification in sparse scRNA-seq data. Detailed comparisons between the two candidate sets across subtypes are provided in Additional file 1: Supplementary Methods C.2 and Figure S4.

## 2.4 Gene co-expression network construction and module detection

Subtype-specific gene sets identified by SDSGI were used to construct gene co-expression networks for downstream lncRNA function inference. For each subtype, pairwise gene-expression correlations were calculated across all the cells in each subtype, and significantly co-expressed gene pairs were retained to form subtype-specific co-expression networks. Both ClusterONE[25] and MCODE[26] were employed to detect functional modules in the co-expression network for each subtype. These densely connected gene modules are used to infer potential lncRNA functions based on the guilt-by-association principle[27], and the functional annotation of their co-expressed protein-coding genes[28]. Detailed descriptions of the correlation metric, threshold selection strategy, matrix-based implementation, and module-detection procedure are provided in Additional file 1: Supplementary Methods C.3.

## 2.5 Function analysis

Gene Ontology (GO) enrichment analysis and Reactome pathway analysis[29] were performed using the Enrichr platform[30, 31]. These results were used to analyze the functions of subtype-specific coding genes and to support lncRNA functional inference based on co-expression relationships.

# 3 Results

## 3.1 Selection of the surprise-information threshold for subtype-specific gene identification

The surprise-information threshold $|S|$ was set to 3, based on the elbow point identified by the Kneedle algorithm[32] and its approximate correspondence to the statistical significance level $-\log(0.05) \approx 3$. Details are provided in Additional file 1: Supplementary Methods C.1 and Figure S3.

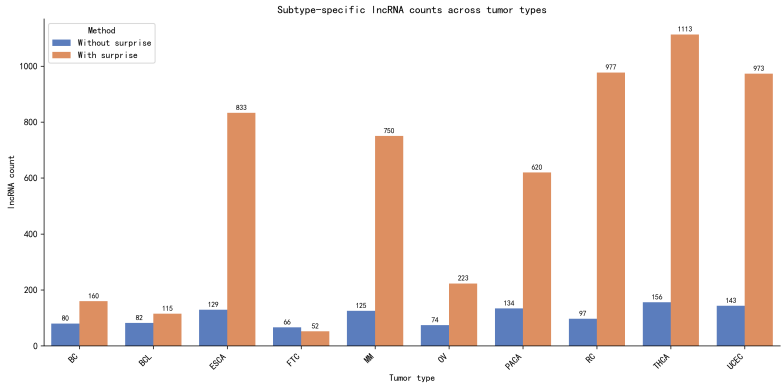



**Fig. 2** Subtype-specific lncRNAs identified by SDSGI

Under this parameter setting, we then examined whether the incorporation of surprise-information scoring altered the identified subtype-specific lncRNA sets. As shown in Figure 2, compared with the strategy without surprise-information scoring, the SDSGI framework identified substantially more subtype-specific lncRNAs across tumors under the same significance level, indicating that surprise-information scoring improved the sensitivity for low-abundance transcript detection.

Further analyses showed that the surprise-information threshold has a direct influence on the size of the resulting subtype-specific gene sets. As expected, as the threshold increased, the number of identified subtype-specific lncRNAs decreased monotonically and gradually approached zero. This pattern indicates that increasing the threshold imposed a stricter subtype-specific screening criterion, such that only genes with stronger subtype-restricted expression patterns were retained. However, if the threshold is set too high, too few genes would remain for meaningful downstream

analyses. Therefore, the final threshold should be selected as a compromise between screening stringency and the feasibility of downstream analyses.

## 3.2 Visual validation of subtype-specific gene sets

We first summarized the subtype-specific gene sets identified across all cancer types, including the numbers of protein-coding genes and lncRNAs, to provide an overall view of the composition of the identified gene sets (Table 2).

**Table 2** Summary of subtype-specific genes identified across different cancers

| Cancer Type | Protein-coding Genes | lncRNAs | Total |
|---|---|---|---|
| BC | 6,402 | 497 | 6,899 |
| BCL | 4,674 | 361 | 5,035 |
| ESCA | 10,148 | 944 | 11,092 |
| FTC | 4,212 | 300 | 4,512 |
| MM | 9,965 | 825 | 10,790 |
| OV | 7,398 | 515 | 7,913 |
| PACA | 10,115 | 833 | 10,948 |
| RC | 10,151 | 909 | 11,060 |
| THCA | 10,720 | 1,224 | 11,944 |
| UCEC | 10,864 | 1,153 | 12,017 |

Next, to evaluate whether the genes identified by SDSGI indeed exhibited subtype-specific expression patterns, $CD8^+$ T cells for breast cancer are illustrated for analysis. Random selected 3 genes for each subtype were displayed in Figure 3 and Figure 4. The size of each bubble corresponds to the proportion of cells in which the corresponding gene is expressed, and the color intensity indicates the mean expression level of the gene among those cells, with the darker color indicating the higher average expression. Compared with the genes identified by conventional differential expression analysis, the genes selected by SDSGI showed clearer subtype-associated expression patterns.

Importantly, subtype specificity in scRNA-seq data is not limited to highly expressed genes. Some genes appeared as small size or weakly colored bubbles in the global visualization because of their relatively low abundance, yet still showed clear subtype-biased expression patterns. This observation indicates that low expression does not imply a lack of subtype specificity. This important property of our method will facilitate the analysis of lncRNAs.

Representative examples are shown in Figure 5. Both PCSK5 and MBNL1-AS1 displayed distinct subtype-restricted expression in $CD8^+$ T cells, illustrating the ability of SDSGI to retain biologically meaningful low-abundance genes, including lncRNAs, that may be overlooked by conventional differential expression analysis.

## 3.3 Functional analysis

To further investigate the biological relevance of the subtype-specific gene sets identified by SDSGI, we selected breast cancer as a representative case for downstream

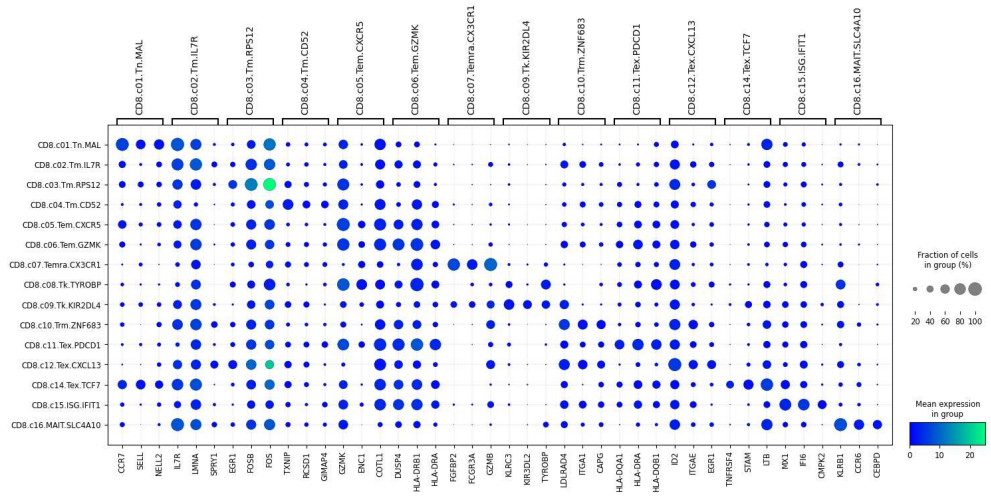



**Fig. 3** Top three subtype-specific gene expression patterns in each CD8$^+$ T-cell subtype identified by SDSGI

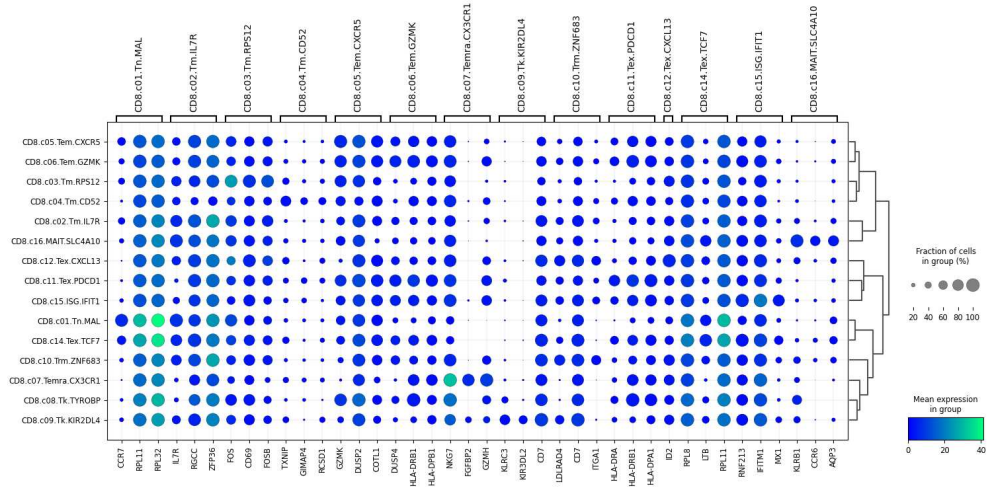



**Fig. 4** Gene expression patterns of CD8$^+$ T cells derived from differential expression analysis

enrichment analysis. Based on the subtype-specific gene sets identified for breast cancer CD8$^+$ T-cell subtypes, we performed both Gene Ontology (GO) enrichment analysis and Reactome pathway analysis[33].

For GO enrichment analysis, we selected CD8.c07.Temra.CX3CR1 as a representative subtype for interpretation. Several enriched GO terms were closely related to

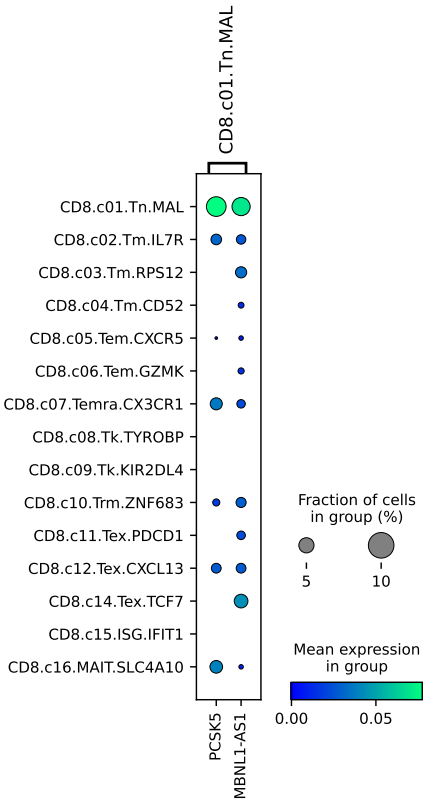



**Fig. 5** Subtype-specific expression of PCSK5 and MBNL1-AS1 in $CD8^+$ T cells

immune effector activity, including GO:0001912 (positive regulation of leukocyte mediated cytotoxicity), GO:0045953 (negative regulation of natural killer cell mediated cytotoxicity), and GO:0046631 (alpha-beta T cell activation). These results suggest that the subtype CD8.c07.Temra.CX3CR1 is functionally associated with activated cytotoxic immune responses. In the study of Zheng et al.[3], tumor-infiltrating T cells exhibited substantial functional heterogeneity, and distinct T-cell subsets were associated with different effector states. Similarly, the work of Zhang et al.[4] showed that different intratumoral T-cell states were linked to distinct functional programs, including activation and effector differentiation. Therefore, the representative GO terms enriched in this subtype may be a suggest of the functional properties of Temra-like T cells.

For Reactome pathway analysis, we focused on three exhausted T-cell-related subtypes in breast cancer: CD8.c11.Tex.PDCD1, CD8.c12.Tex.CXCL13, and CD8.c14.Tex.TCF7, which all exhibit interesting functional heterogeneity within the tumor microenvironment. These three Tex subtypes are characterized by distinct

marker genes, namely PDCD1, CXCL13, and TCF7. To elucidate functional differences among them, we performed Reactome pathway enrichment analysis on the subtype-specific gene sets of each Tex subtype.

From the enrichment results, CD8.c11.Tex.PDCD1 was significantly enriched in PD-1 signaling and interferon-gamma signaling pathways, suggesting its association with T-cell exhaustion and chronic inflammatory conditions. In accordance with Arasanz et al. [34], PD-1 serves as a central axis in immune evasion and T-cell inhibitory regulation, and PD-1/PD-L1 blockade may help restore TCR surface expression and downstream signaling. The subtype-specific genes identified in our study are significantly enriched in these pathways, providing potential insights for PD-1/PD-L1 blockade strategies. Additional downstream pathways, including CD28 family costimulation and TCR signaling, indicate that this subtype may be under persistent antigen stimulation but regulated by inhibitory signals, highlighting functional limitations while retaining partial cytokine secretion capabilities.

CD8.c12.Tex.CXCL13 exhibited unique transcriptional features, with enrichment in adaptive immune-related pathways and neurotrophin signaling, suggesting enhanced tissue residency and immune-regulatory functions. Notably, neurotrophin expression in immune cells is dynamically regulated [35], correlating with differentiation stages and functional activity. Compared to CD8.c11.Tex.PDCD1, this subtype lacks PD-1 pathway enrichment but retains signaling and MHC-II antigen presentation capacity, potentially contributing to tertiary lymphoid structure formation and the recruitment of B cells and follicular helper T cells, which may indicate stronger immune surveillance.

In contrast, CD8.c14.Tex.TCF7 was enriched in pathways related to translation regulation and cellular stress responses, including peptide chain elongation, translation termination, ribosome assembly, amino acid deficiency response, viral mRNA translation, and selenocysteine synthesis. This enrichment pattern suggests a resting but proliferation- and differentiation-competent state, providing a metabolic basis for long-term survival and functional maintenance, and functionally complementing CD8.c11.Tex.PDCD1. This subtype may serve as a key target for T-cell reactivation in immunotherapy.

These findings indicate that CD8.c11.Tex.PDCD1 and CD8.c12.Tex.CXCL13 may cooperate in chronic antigen-stimulated immune responses, with the former exhibiting a more inhibitory phenotype and the latter showing tissue-adaptive features. The translation-related pathways in CD8.c14.Tex.TCF7 suggest a "standby" state, potentially critical for reversing T-cell exhaustion. Collectively, these three subtypes represent the functional heterogeneity of Tex cells and imply a potential dynamic differentiation trajectory from resting (c14) $\rightarrow$ activated/regulatory (c12) $\rightarrow$ exhausted (c11), motivating further analysis of their dynamic differentiation trajectories.

Furthermore, to quantify the role of lncRNAs in these subtypes, we analyzed all enriched pathways (p¡0.05) and counted lncRNAs directly associated with these pathways. In some subtypes, lncRNAs are widely associated with multiple pathway-enriched genes, suggesting their involvement in key regulatory processes and providing insights into subtype-specific lncRNA functions within the tumor microenvironment. Detailed GO enrichment results are provided in Additional file 1: Table S4 and

Figure S5. Detailed top 10 Reactome enrichment results for CD8.c11.Tex.PDCD1, CD8.c12.Tex.CXCL13, and CD8.c14.Tex.TCF7 are provided in Additional file 1: Tables S5–S7 and Figure S6.

## 3.4 Tex trajectory

Here, CD8.c11.Tex.PDCD1, CD8.c12.Tex.CXCL13, and CD8.c14.Tex.TCF7 cell subtypes were selected for further trajectory analysis (hereafter referred to as c11, c12 and c14 for convenience), because the preceding pathway enrichment analysis suggested functional relationships among these subtypes.

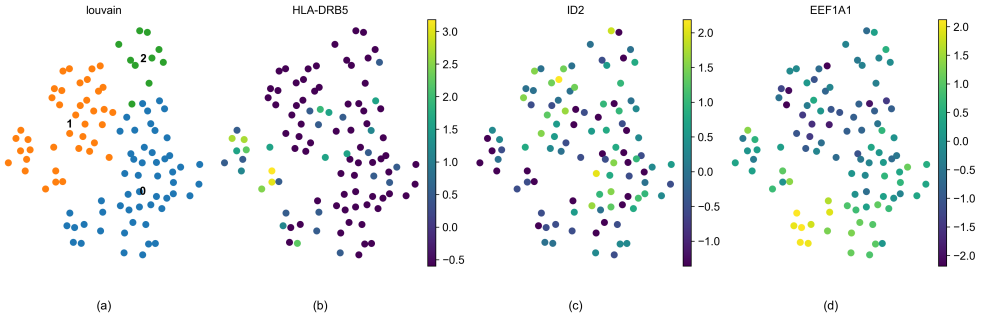



**Fig. 6** Louvain clustering and subtype-specific marker gene expression of the three Tex cell subtypes. (a) Louvain clustering results of Tex cells. Each point represents one cell, and the three colors indicate three clusters. Cluster 0 is colored as blue, the cluster 1 is colored as orange and the cluster 2 is colored as green. (b–d) Expression patterns of three representative subtype-specific marker genes: HLA-DRB5 for c11, ID2 for c12, and EEF1A1 for c14. Darker colors indicate lower gene expression, whereas brighter colors indicate higher gene expression.

First, Louvain clustering[36] was used to identify three cell clusters, as shown in Figure 6(a). Also the subtype specific genes are labeled in the cell clusters. We can find the subtype specific gene HLA-DRB for c11 is specifically high expressed in the orange cluster 1, similarly the ID2 for c12 with a high expression in green cluster 2, and EEF1A1 with a high expression in blue cluster 0. Partition-based Graph Abstraction (PAGA) [37] was applied to reconstruct the coarse-grained connectivity among the three subtypes. The resulting connectivity structure is shown in Additional file 1: Figure S10. Based on the results, trajectory analysis was then performed among these cells. The preceding functional enrichment analysis of subtype-specific genes gave clues about the functional relationships among the three Tex subtypes. In accordance with these functional clues, the results shown in Figure 7 indicate a potential developmental trajectory from a resting state, through an intermediate activation/regulatory state, and ultimately to a terminally exhausted state.

Furthermore, pseudotime analysis was performed to assign each cell a relative progression value along the inferred trajectory [5]. Cells at the early pseudotime points predominantly belonged to the resting Tex subtype, whereas cells at the terminal pseudotime points corresponded to the fully exhausted subtype. Analysis

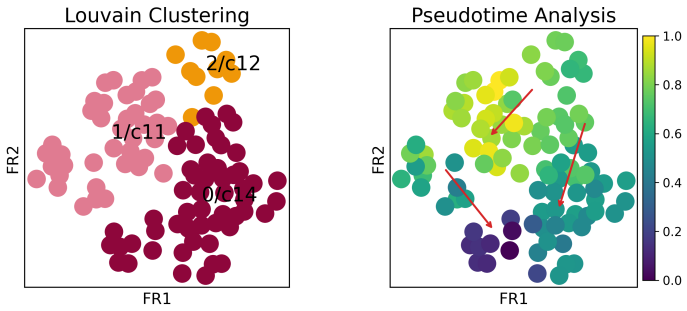



**Fig. 7** The trajectory analysis of the three Tex subtypes

of key exhaustion-associated genes, including PDCD1 and other inhibitory receptors, revealed a gradual increase in expression along the pseudotime axis, validating the inferred differentiation direction. Overall, these results demonstrate a continuous, dynamic differentiation process among Tex subtypes in the breast cancer tumor microenvironment. The integration of subtype-specific genes, including lncRNAs, into the trajectory reconstruction not only enhances the resolution of Tex developmental dynamics but also provides candidate molecular targets for future studies aimed at preventing or reversing T-cell exhaustion.

## 3.5 Co-expression network construction

To further explore the biological associations among the subtype-specific genes, we constructed subtype-specific co-expression networks for breast cancer $CD8^+$ T-cell subtypes. For each subtype, the correlation threshold was determined separately, and the selected thresholds ranged from 0.60 to 0.71 (Additional file 1: Table S8). Detailed rationale for threshold selection is provided in Additional file 1: Supplementary Methods D.1 and Figures S7–S8. Under these thresholds, the retained numbers of nodes and edges varied substantially across subtypes, reflecting differences in network size and transcriptional coordination.

We then applied the ClusterONE algorithm to the resulting co-expression networks to identify densely connected functional modules. The final number of modules obtained is presented in Table 3. Details are presented in Additional file 1: Supplementary Methods D.1 and Table S9.

## 3.6 lncRNA function prediction

We summarize the overlap of enriched GO terms of each module. Details is provided in Additional file 1: Figure S9. Overall, the overlap is minor, indicating that the module may play the specific roles in the subtype of T cells. At the same time, a small number of GO terms were shared by multiple cell subtypes. An example of GO:0006090 indicated metabolic regulation may act as a shared functional background across several $CD8^+$ T-cell states. However, the associated lncRNAs and protein-coding gene partners differed across subtypes, indicating that this shared biological process may be mediated by subtype-dependent lncRNA–mRNA coordination patterns.

**Table 3** Complete module detection results for subtype-specific co-expression networks in breast cancer

| Subtype | #Modules | #Modules containing lncRNAs |
|---|---|---|
| CD8.c01.Tn.MAL | 44 | 6 |
| CD8.c02.Tm.IL7R | 9 | 2 |
| CD8.c03.Tm.RPS12 | 25 | 2 |
| CD8.c04.Tm.CD52 | 61 | 6 |
| CD8.c05.Tem.CXCR5 | 6 | 2 |
| CD8.c06.Tem.GZMK | 30 | 8 |
| CD8.c07.Temra.CX3CR1 | 17 | 4 |
| CD8.c09.Tk.KIR2DL4 | 4 | 1 |
| CD8.c10.Trm.ZNF683 | 32 | 6 |
| CD8.c11.Tex.PDCD1 | 6 | 0 |
| CD8.c12.Tex.CXCL13 | 5 | 1 |
| CD8.c14.Tex.TCF7 | 3 | 1 |
| CD8.c15.ISG.IFIT1 | 16 | 0 |
| CD8.c16.MAIT.SLC4A10 | 35 | 8 |

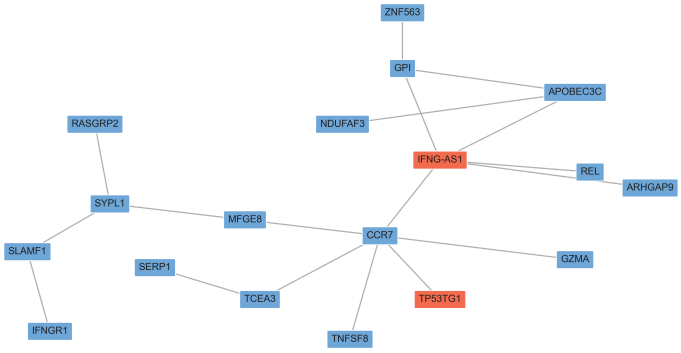



**Fig. 8** The GO:0006090-associated module in CD8.c01.Tn.MAL subtype

We therefore selected two representative modules associated with GO:0006090 for analysis, as shown in Figure 8 and Figure 9. In the CD8.c01.Tn.MAL-associated module (Figure 8), the lncRNA IFNG-AS1 was directly connected with protein-coding genes involved in pyruvate-related metabolic processes. In the CD8.c04.Tm.CD52-associated module (Figure 9), the lncRNA MMP25-AS1 was linked to a distinct metabolic module containing coding genes related to glycolytic regulation. These topological patterns suggest that different T-cell subtypes may recruit different lncRNAs into partially convergent metabolic programs, even when the enriched biological process is nominally shared. Compared with the module in CD8.c01.Tn.MAL, the CD8.c04.Tm.CD52 module showed a denser local structure and a more explicit lncRNA-centered association pattern, making it a particularly informative example of subtype-specific lncRNA–mRNA coordination.

Overall, module-based network analysis identified 105 lncRNAs associated with subtype-related functional modules in breast cancer, among which 14 ones have

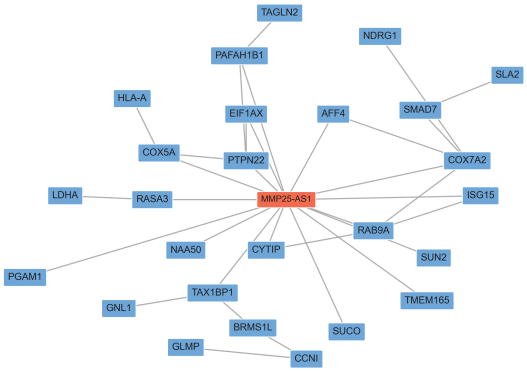



**Fig. 9** The GO:0006090-associated module in the CD8.c04.Tm.CD52 subtype

independent literature support (Table 4). Rather than listing all candidates in the main text, we highlight that this literature-supported subset provides orthogonal evidence for the biological relevance of the network-derived predictions. Representative examples include IFNG-AS1, MMP25-AS1, LINC00152, MIR22HG, and AC092580.4, several of which have previously been implicated in cancer progression, immune regulation, or patient prognosis.

**Table 4** Literature-supported breast-cancer-related lncRNAs identified from subtype-specific co-expression networks

| Subtype | Validated lncRNA | PubMed ID |
|---|---|---|
| CD8.c01.Tn.MAL | IFNG-AS1 | 30336781 |
| CD8.c01.Tn.MAL | TP53TG1 | 35871431 |
| CD8.c01.Tn.MAL | PVT1 | 34922601 |
| CD8.c02.Tm.IL7R | LINC00116 | 36736170 |
| CD8.c02.Tm.IL7R | LINC-PINT | 37660994 |
| CD8.c03.Tm.RPS12 | SNHG12 | 38833118 |
| CD8.c06.Tem.GZMK | SNHG8 | 33275231 |
| CD8.c06.Tem.GZMK | LINC00299 | 37706018 |
| CD8.c09.Tk.KIR2DL4 | HCG18 | 34976998 |
| CD8.c14.Tex.TCF7 | DLEU2 | 38296962 |
| CD8.c16.MAIT.SLC4A10 | AC092580.4 | 32754579 |
| CD8.c16.MAIT.SLC4A10 | LINC00152 | 33842324 |
| CD8.c16.MAIT.SLC4A10 | MIR22HG | 34446703 |
| CD8.c16.MAIT.SLC4A10 | MIR4435-2HG | 36105009 |

Taken together, these results show that subtype-specific co-expression networks not only preserve distinct functional modules across breast cancer T-cell subtypes, but also provide a useful framework for linking lncRNAs to biologically interpretable module functions.

# 4 Discussion and Conclusion

In this study, we proposed a computational method for identifying subtype-specific genes in tumor-infiltrating T cells using single-cell RNA-seq data. By integrating an expression-based subtype-specificity score derived from the $\tau$ index with a surprise-information-based statistical score, the framework captures subtype-restricted expression patterns from complementary perspectives. Compared with conventional differential expression analysis, this integrative strategy is better suited to sparse scRNA-seq data and enables the retention of genes that show subtype-preferential enrichment despite low expression levels.

Although the present analysis focused mainly on $CD8^+$ T cells, the underlying design is not restricted to this cell population. Because the method relies on subtype-level expression specificity and statistically informative enrichment signals rather than predefined cell-type-specific markers, it may be extended to other immune cell populations in the TME, such as $CD4^+$ T cells, regulatory T cells, natural killer cells, B cells, and macrophages. The same strategy may also be useful for identifying subtype- or state-specific regulatory genes in non-tumor inflammatory conditions, autoimmune diseases, or other complex tissue microenvironments. With the increasing availability of spatial transcriptomic and multi-omics data, future extensions could further incorporate spatial localization or epigenetic information to improve the interpretation of context-dependent lncRNA functions[38, 39].

The results in breast cancer $CD8^+$ T cells demonstrate that the identified subtype-specific gene sets include both protein-coding genes and lncRNAs with potential relevance to distinct immune cell states. In particular, the analyses of exhausted $CD8^+$ T cells suggest structured relationships among Tex subtypes, and the PAGA connectivity analysis reveals potential transitions between exhaustion-related cell states [36, 37]. In addition, several lncRNAs identified from subtype-specific co-expression modules showed independent literature support. Representative examples, such as *LINC00152*, are consistent with previously reported tumor-promoting roles in breast cancer [40]. Moreover, it may be a inspiring object in the immune therapeutic strategy for breast cancer. These observations support the biological relevance of the identified gene sets and suggest that low-abundance lncRNAs may contribute to subtype-specific immune regulation in the tumor microenvironment.

Several limitations should also be noted. First, although the use of nonzero expression summaries and surprise-information-based scoring helps reduce the influence of sparsity, scRNA-seq data are still affected by dropout events, zero inflation, technical noise, and potential batch effects[12, 41]. These factors may influence the detection of low-abundance transcripts, especially lncRNAs. Second, the identification of subtype-specific genes and the construction of co-expression networks still depend on several parameter choices, including the $\tau$-score threshold, the surprise-information threshold, and the correlation cutoff used for network construction. Although these parameters were selected based on score distributions, statistical significance, and network-topology considerations, their robustness should be further evaluated across additional datasets. Third, the functional interpretation of lncRNAs in this study is mainly based on co-expression relationships, and the guilt-by-association principle [42, 43]. Therefore, the prioritized lncRNAs should be regarded as computationally

inferred candidates rather than experimentally validated therapeutic targets. Further experimental studies are needed to validate their regulatory functions and mechanistic roles in tumor-infiltrating immune cells. Finally, the downstream functional and trajectory analyses were mainly performed in breast cancer $CD8^+$ T cells. Future studies should extend this framework to additional cancer types, other immune cell populations, and independent validation cohorts.

Overall, these findings highlight the value of combining expression specificity, statistical enrichment, and network-based functional inference for discovering low-abundance but functionally informative lncRNAs in the tumor immune microenvironment. By providing a computational strategy for prioritizing subtype-associated lncRNAs, this work may facilitate further investigation of immune-cell heterogeneity and lncRNA-mediated regulation in cancer.

# Abbreviations

| | |
|---|---|
| BC | Breast cancer |
| BCL | B-cell lymphoma |
| CHOL | Cholangiocarcinoma |
| DE | Differential expression |
| ESCA | Esophageal cancer |
| FTC | Follicular thyroid cancer |
| GEO | Gene Expression Omnibus |
| GO | Gene Ontology |
| lncRNA | Long non-coding RNA |
| MCODE | Molecular Complex Detection |
| MHC | Major histocompatibility complex |
| MM | Multiple myeloma |
| OV | Ovarian cancer |
| PACA | Pancreatic cancer |
| PAGA | Partition-based Graph Abstraction |
| PD-1 | Programmed cell death protein 1 |
| PD-L1 | Programmed death-ligand 1 |
| QC | Quality control |
| RC | Rectal cancer |
| scRNA-seq | Single-cell RNA sequencing |
| SDSGI | Surprise-Driven Subtype-Specific Gene Identification |
| TCR | T-cell receptor |
| Tex | Exhausted T cells |
| THCA | Thyroid cancer |
| TME | Tumor microenvironment |
| UCEC | Uterine corpus endometrial carcinoma |

# Supplementary information

**Additional file 1:** In this file we provide the supplementary materials supporting this study, including extended methodological details, Supplementary Tables

S1–S9, and Supplementary Figures S1–S10. These materials cover data collection and preprocessing, quality-control parameter selection, $\tau$-based specificity scoring, surprise-information score derivation and threshold selection, candidate-gene integration, enrichment analysis, trajectory analysis, co-expression network construction, module detection, and lncRNA functional interpretation.

# Declarations

- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Availability of data and materials:** All single-cell RNA-seq datasets analyzed in this study are publicly available in the GEO database under accession number GSE156728. All other datasets and related information supporting the conclusions of this article are included within the article and its supplementary materials. The code is currently maintained in a Github repository (https://github.com/guo-lab-xidian/GGXD_Bioinformatics.git).
- **Competing interests:** The authors declare that they have no competing interests.
- **Funding:** This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61672407, 62572372, and 62172318.
- **Authors' contributions:** GX designed the overall methodology, conceptual framework of the study and writing. ZJ developed and implemented the programming and writing. RY contributed to both the design of the method and programming aspects. HY carried out data preprocessing and contributed to programming. WB participated in discussions on study design and provided critical insights. All authors read and approved the final manuscript.
- **Acknowledgements:** Not applicable.

# References


[1] De Visser, K.E., Joyce, J.A.: The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. Cancer cell **41**(3), 374–403 (2023)

[2] Liu, Y., Zhang, Q., Xing, B., Luo, N., Gao, R., Yu, K., Hu, X., Bu, Z., Peng, J., Ren, X., *et al.*: Immune phenotypic linkage between colorectal cancer and liver metastasis. Cancer cell **40**(4), 424–437 (2022)

[3] Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q., *et al.*: Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing. Cell **169**(7), 1342–1356 (2017)

[4] Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J.Y., *et al.*: Lineage tracking reveals dynamic relationships of t cells in colorectal cancer. Nature **564**(7735), 268–272 (2018)

[5] Chen, Z., Ji, Z., Ngiow, S.F., Manne, S., Cai, Z., Huang, A.C., Johnson, J., Staupe, R.P., Bengsch, B., Xu, C., *et al.*: Tcf-1-centered transcriptional network drives an effector versus exhausted cd8 t cell-fate decision. Immunity **51**(5), 840–855 (2019)

[6] Baldominos, P., Barbera-Mourelle, A., Barreiro, O., Huang, Y., Wight, A., Cho, J.-W., Zhao, X., Estivill, G., Adam, I., Sanchez, X., *et al.*: Quiescent cancer cells resist t cell attack by forming an immunosuppressive niche. Cell **185**(10), 1694–1708 (2022)

[7] Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A., *et al.*: Long non-coding rnas: definitions, functions, challenges and recommendations. Nature reviews Molecular cell biology **24**(6), 430–447 (2023)

[8] Zhang, Y.: Lncrna-encoded peptides in cancer. Journal of hematology & oncology **17**(1), 66 (2024)

[9] Guo, Y., Xie, Y., Luo, Y.: The role of long non-coding rnas in the tumor immune microenvironment. Frontiers in Immunology **13**, 851004 (2022)

[10] Huang, D., Chen, J., Yang, L., Ouyang, Q., Li, J., Lao, L., Zhao, J., Liu, J., Lu, Y., Xing, Y., *et al.*: Nkila lncrna promotes tumor immune evasion by sensitizing t cells to activation-induced cell death. Nature immunology **19**(10), 1112–1125 (2018)

[11] Li, G., Kryczek, I., Nam, J., Li, X., Li, S., Li, J., Wei, S., Grove, S., Vatan, L., Zhou, J., *et al.*: Limit is an immunogenic lncrna in cancer immunity and immunotherapy. Nature cell biology **23**(5), 526–537 (2021)

[12] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., *et al.*: Eleven grand challenges in single-cell data science. Genome biology **21**(1), 31 (2020)

[13] Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., *et al.*: Pan-cancer single-cell landscape of tumor-infiltrating t cells. Science **374**(6574), 6474 (2021)

[14] Kharchenko, P.V., Silberstein, L., Scadden, D.T.: Bayesian approach to single-cell differential expression analysis. Nature methods **11**(7), 740–742 (2014)

[15] Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A., *et al.*: Single-cell analysis of long non-coding rnas in the developing human neocortex. Genome biology **17**(1), 67 (2016)

[16] Kryuchkova-Mostacci, N., Robinson-Rechavi, M.: A benchmark of gene expression tissue-specificity metrics. Briefings in bioinformatics **18**(2), 205–214 (2017)

[17] DeMeo, B., Berger, B.: Sca: recovering single-cell heterogeneity through information-based dimensionality reduction. Genome Biology **24**(1), 195 (2023)

[18] Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., *et al.*: Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21**(5), 650–659 (2005)

[19] Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

[20] Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., *et al.*: The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. Genome research **22**(9), 1775–1789 (2012)

[21] Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)

[22] Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., Teichmann, S.A.: Classification of low quality cells from single-cell rna-seq data. Genome biology **17**(1), 29 (2016)

[23] McGinnis, C.S., Murrow, L.M., Gartner, Z.J.: Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. Cell systems **8**(4), 329–337 (2019)

[24] Joshi, C.J., Ke, W., Drangowska-Way, A., O'Rourke, E.J., Lewis, N.E.: What are housekeeping genes? PLoS computational biology **18**(7), 1010295 (2022)

[25] Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods **9**(5), 471–472 (2012)

[26] Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics **4**(1), 2 (2003)

[27] Oliver, S.: Guilt-by-association goes global. Nature **403**(6770), 601–602 (2000)

[28] Van Dam, S., Vosa, U., Graaf, A., Franke, L., Magalhaes, J.P.: Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in bioinformatics **19**(4), 575–592 (2018)

[29] Jupe, S., Akkerman, J.W., Soranzo, N., Ouwehand, W.H.: Reactome-a curated knowledgebase of biological pathways: megakaryocytes and platelets. Journal of thrombosis and haemostasis: JTH **10**(11), 2399 (2012)

[30] Xie, Z., Bailey, A., Kuleshov, M., Clarke, D., Evangelista, J., Jenkins, S., et al.: Gene set knowledge discovery with enrichr, curr. Protoc. 132021e90 (2021)

[31] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., *et al.*: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research **44**(W1), 90–97 (2016)

[32] Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171 (2011). IEEE

[33] Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., Mi, H.: Panther: Making genome-scale phylogenetics accessible to all. Protein Science **31**(1), 8–22 (2022)

[34] Arasanz, H., Gato-Cañas, M., Zuazo, M., Ibañez-Vea, M., Breckpot, K., Kochan, G., Escors, D.: Pd1 signal transduction pathways in t cells. Oncotarget **8**(31), 51936 (2017)

[35] Minnone, G., De Benedetti, F., Bracci-Laudiero, L.: Ngf and its receptors in the regulation of inflammatory response. International journal of molecular sciences **18**(5), 1028 (2017)

[36] Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.-a.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., *et al.*: Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis.

Cell **162**(1), 184–197 (2015)

[37] Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., Theis, F.J.: Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome biology **20**(1), 59 (2019)

[38] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. cell **177**(7), 1888–1902 (2019)

[39] Vandereyken, K., Sifrim, A., Thienpont, B., Voet, T.: Methods and applications for single-cell and spatial multi-omics. Nature Reviews Genetics **24**(8), 494–515 (2023)

[40] Wu, J., Shuang, Z., Zhao, J., Tang, H., Liu, P., Zhang, L., Xie, X., Xiao, X.: Linc00152 promotes tumorigenesis by regulating dnmts in triple-negative breast cancer. Biomedicine & Pharmacotherapy **97**, 1275–1281 (2018)

[41] Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., Chen, J.: A benchmark of batch-effect correction methods for single-cell rna sequencing data. Genome biology **21**(1), 12 (2020)

[42] Rinn, J.L., Chang, H.Y.: Genome regulation by long noncoding rnas. Annual review of biochemistry **81**(1), 145–166 (2012)

[43] Ehsani, R., Drabløs, F.: Measures of co-expression for improved function prediction of long non-coding rnas. BMC bioinformatics **19**(1), 533 (2018)

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf