Appendix A Supplementary information for data collection and preprocessing

This section provides supplementary information for the data collection and preprocessing procedures. It includes detailed descriptions of dataset sources, sample composition, and tissue origins, together with the criteria and rationale for parameter selection in the quality-control process. Additional quality-control related visualizations are also provided to complement the results presented in the main text.

A.1 Dataset description and sample characteristics

To comprehensively investigate subtype-specific gene expression patterns of tumor-infiltrating $CD8^+$ T cells, we collected and integrated multiple publicly available single-cell RNA sequencing (scRNA-seq) datasets from diverse cancer types. As summarized in Table S1, a total of 11 cancer types were included, covering both solid tumors and hematological malignancies. All datasets were generated using the 10X Genomics platform to ensure technical consistency. The data were obtained from publicly accessible repositories, including the Gene Expression Omnibus (GEO) and Genome Sequence Archive (GSA), with corresponding PubMed IDs provided for reference.

Based on these datasets, $CD8^+$ T-cell populations were extracted and further processed for downstream analysis. The detailed sample composition is presented in Table S2. The number of cells varies substantially across cancer types, ranging from a few hundred cells (e.g., cholangiocarcinoma) to tens of thousands (e.g., thyroid cancer), reflecting differences in dataset scale and sequencing depth. In addition, the total number of detected genes and annotated lncRNAs is also reported. Notably, a considerable proportion of genes correspond to long non-coding RNAs, highlighting the complexity and richness of the transcriptomic landscape in tumor-infiltrating T cells.

Furthermore, to better characterize the biological context of the collected samples, we summarized the tissue origins of $CD8^+$ T cells in Table S3. The majority of datasets contain tumor-derived cells, while some also include cells from adjacent normal tissues or peripheral blood. This diversity in tissue origin provides a valuable basis for comparing tumor-associated and non-tumor immune states, and supports the identification of subtype-specific transcriptional signatures under different microenvironmental conditions.

Overall, these datasets provide a comprehensive and heterogeneous resource for studying immune cell heterogeneity in the tumor microenvironment, forming the foundation for the subsequent subtype-specific gene identification analysis.

A.2 Cell-level quality control metrics and filtering strategy

To ensure the reliability of downstream analyses, rigorous quality control (QC) procedures were applied to all single-cell RNA sequencing (scRNA-seq) datasets prior to subtype-specific gene identification[1, 2]. Given the intrinsic technical noise and sparsity of scRNA-seq data, particularly in tumor microenvironment studies[3, 4],

Table S1 Data source table

No.	Abbr.	Cancer type	Data source	PubMed ID
1	BC	Breast cancer	GSE114724;	29961579;
			GSE110686;	29942092;
			GSE156728	34914499
2	BCL	B-cell lymphoma	GSE156728	34914499
3	CHOL	Cholangiocarcinoma	GSE140228;	31675496;
			GSE125449	31588021
4	ESCA	Esophageal cancer	GSE156728	34914499
5	FTC	Follicular thyroid cancer	GSE156728	34914499
6	MM	Multiple myeloma	GSE156728	34914499
7	OV	Ovarian cancer	GSE156728	34914499
8	PACA	Pancreatic cancer	GSA:CRA001160;	31273297;
			GSE156728	34914499
9	RC	Rectal cancer	GSE156728	34914499
10	THCA	Thyroid cancer	GSE156728	34914499
11	UCEC	Uterine corpus endometrial carcinoma	GSE156728	34914499

Table S2 Sample information of $CD8^+$ T-cell subsets

No.	Abbr.	Cancer type	#Cell	#Gene	#lncRNA
1	BC	Breast cancer	4,291	21,999	5,911
2	BCL	B-cell lymphoma	3,482	26,293	8,957
3	CHOL	Cholangiocarcinoma	300	11,728	556
4	ESCA	Esophageal cancer	12,526	21,999	5,911
5	FTC	Follicular thyroid cancer	767	21,999	5,911
6	MM	Multiple myeloma	8,629	26,293	8,957
7	OV	Ovarian cancer	3,517	21,999	5,911
8	PACA	Pancreatic cancer	5,957	21,999	5,911
9	RC	Rectal cancer	16,544	21,999	5,911
10	THCA	Thyroid cancer	33,450	21,999	5,911
11	UCEC	Uterine corpus endometrial carcinoma	19,926	21,999	5,911

Table S3 Tissue origin information of $CD8^+$ T-cell subsets

No.	Abbr.	Cancer type	Tumor	Adjacent normal	Blood
1	BC	Breast cancer	✓	✓	×
2	BCL	B-cell lymphoma	✓	×	✓
3	CHOL	Cholangiocarcinoma	✓	✓	✓
4	ESCA	Esophageal cancer	✓	✓	×
5	FTC	Follicular thyroid cancer	✓	×	×
6	MM	Multiple myeloma	✓	×	✓
7	OV	Ovarian cancer	✓	✓	×
8	PACA	Pancreatic cancer	✓	✓	×
9	RC	Rectal cancer	✓	✓	×
10	THCA	Thyroid cancer	✓	×	×
11	UCEC	Uterine corpus endometrial carcinoma	✓	✓	×

we focused on two key cell-level QC metrics: the proportion of mitochondrial gene expression and the number of detected genes per cell[5, 6].

First, the proportion of mitochondrial gene expression (MT gene percentage) was used as an indicator of cell quality. Cells exhibiting abnormally high mitochondrial gene expression (typically exceeding 20%) are often associated with compromised cellular integrity[5, 7]. This threshold is widely adopted in scRNA-seq quality control practices, as elevated mitochondrial transcript fractions are indicative of compromised cellular integrity. In particular, during apoptosis or necrosis, disruption of mitochondrial membranes leads to the release of mitochondrial RNA into the cytoplasm, resulting in an abnormal increase in mitochondrial gene expression[8]. Therefore, a threshold around 20% provides a practical balance between retaining biologically meaningful cells and removing low-quality or dying cells. It should be noted that the exact cutoff may vary depending on tissue type and sequencing conditions; however, a threshold of 20% has been shown to be effective in a wide range of studies[1, 2, 5]. Biologically, when cells undergo apoptosis or necrosis, disruption of mitochondrial membranes leads to the leakage of mitochondrial RNA into the cytoplasm, resulting in an artificially elevated mitochondrial transcript fraction. Such cells no longer reflect normal physiological transcriptional states and are therefore considered low-quality and excluded during preprocessing.

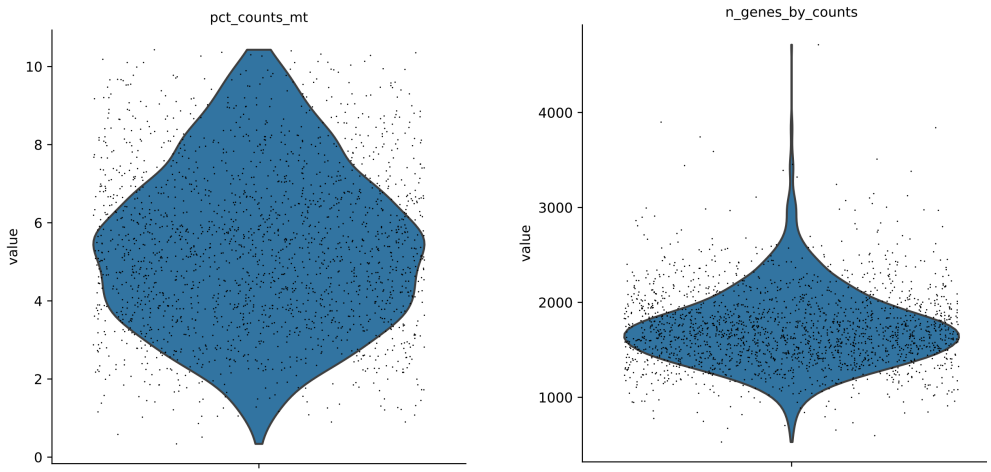
Second, the number of detected genes per cell was employed to identify potential outlier cells[1, 2]. In particular, we defined "outlier cells" as those whose detected gene counts significantly deviate from the population distribution, typically exceeding two standard deviations above the mean[6, 7]. These abnormal profiles are primarily attributed to technical artifacts introduced during droplet-based single-cell sequencing (e.g., 10X Genomics platform), where multiple cells may be encapsulated within the same droplet or insufficiently separated during cell dissociation, leading to so-called doublets or multiplets[9]. Such artifacts result in artificially inflated gene counts and can severely distort downstream analyses, including cell clustering and subtype identification.

Given that the primary objective of scRNA-seq analysis is to resolve cellular heterogeneity, the inclusion of low-quality cells or technical artifacts would obscure true biological signals. Therefore, strict filtering based on mitochondrial gene proportion and detected gene count is essential to ensure data quality and improve the robustness of subsequent subtype-specific gene identification.

A.3 Parameter selection rationale

The quality control thresholds for mitochondrial gene proportion and detected gene number were determined based on the empirical distributions of the dataset, rather than arbitrary fixed values. Specifically, these thresholds were selected according to the distributional characteristics of breast cancer $CD8^+$ T cells, as illustrated in Figure S1.

For mitochondrial gene proportion, the violin plot (Figure S1(a)) shows that the majority of cells exhibit mitochondrial content within the range of 0–10%. Only a small fraction of cells display higher mitochondrial proportions, which are typically associated with cellular stress, membrane damage, or apoptosis. Therefore, a threshold of 10% was adopted to effectively remove these abnormal cells while preserving the



(a) Distribution of mitochondrial gene proportion

(b) Distribution of detected gene numbers per cell

Fig. S1 Quality control metrics for CD8$^+$ T cells in breast cancer. (a) Distribution of mitochondrial gene proportion across CD8$^+$ T-cell subtypes. (b) Distribution of the number of detected genes per cell across CD8$^+$ T-cell subtypes.

majority of biologically meaningful cells. This choice follows the principle of retaining the dominant distribution while excluding outliers [2].

For the number of detected genes per cell, the distribution shown in Figure S1(b) indicates that most cells contain fewer than 3000 detected genes, whereas a small number of cells exhibit substantially higher gene counts. These high-count cells are likely to represent doublets or multiplets arising from droplet-based sequencing, where multiple cells are captured together and sequenced as a single observation. To avoid introducing artificial hybrid expression profiles, cells with more than 3000 detected genes were excluded. This threshold was chosen to align with the upper bound of the main distribution while filtering out potential outliers [9].

Importantly, the selection of these thresholds was data-driven and context-specific. As different tumor types and sample sources may exhibit distinct transcriptional characteristics, quality control parameters should be adapted accordingly rather than universally fixed[1, 2].

In addition, mitochondrial gene proportion and gene count were not considered independently. To ensure robust quality control, mitochondrial gene proportion and ribosomal protein gene expression (e.g., RPS/RPL family) were jointly evaluated[5, 10]. Cells exhibiting both high mitochondrial gene content ($> 30\%$) and low ribosomal gene expression ($< 5\%$) were regarded as low-quality cells, potentially reflecting severe technical noise, and were excluded from downstream analyses[7, 11].

Appendix B Rationale for subtype-level expression summary in τ calculation

The expression-specificity component of SDSGI was derived from the classical tissue-specificity index τ, which was originally proposed to quantify how selectively a gene is expressed across multiple tissues. In its original form, the index is defined as

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}, \tag{S1}$$

where N denotes the number of tissues and x_i represents the normalized expression component of a gene in tissue i. A larger τ value indicates stronger tissue specificity, whereas a smaller value suggests broader expression across tissues.

Since the objective of the present study is to measure gene specificity across immune cell subtypes rather than across tissues, the tissue dimension was replaced by the subtype dimension, yielding the subtype-specificity formulation in Eqs. (**??**) and (**??**).

A key issue in applying the τ index to scRNA-seq data is how to define the subtype-level expression summary x_i. Unlike bulk RNA-seq data, scRNA-seq data are highly sparse and are strongly affected by dropout events, resulting in an excessive number of zero values. Therefore, different choices of x_i may substantially affect the resulting specificity score distribution.

In this study, three candidate summaries were evaluated: the mean expression across all cells in a subtype, the mean expression across nonzero-expressing cells, and the median expression. As shown in Figure S2, using the mean across all cells led to distorted specificity-score distributions, because excessive zeros diluted subtype-preferential signals. By contrast, median-based summaries were overly compressed, especially for low-abundance genes such as lncRNAs, because the median expression was frequently zero in most subtypes. This often caused the normalized values to collapse toward zero and made subtype-specific variation difficult to resolve. In particular, for many genes, expression values were zero in more than 90% of cells, especially for lncRNAs, making median-based normalization unstable or even undefined in practice.

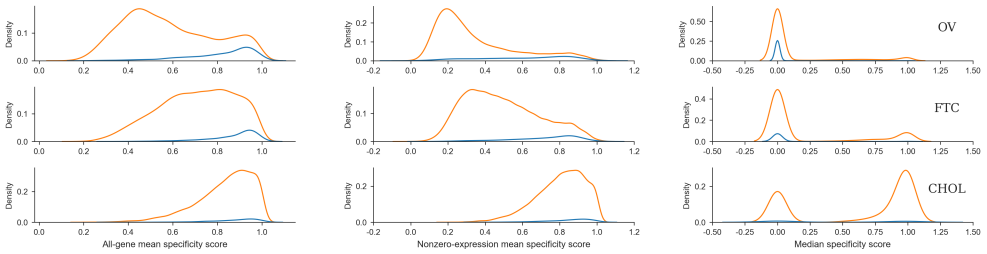


Fig. S2 Distribution of subtype-specificity scores under different expression summary strategies across three representative cancer types (OV, FTC, and CHOL)

Therefore, SDSGI defines x_i as the mean of nonzero expression values within each subtype. This choice mitigates the masking effect caused by dropout-dominated zeros while preserving subtype-level expression differences, and is particularly suitable for capturing low-abundance but biologically structured transcripts such as lncRNAs. The corresponding score distributions under the three alternative definitions are illustrated in Figure S2. Compared with the all-cell mean and median-based summaries, the nonzero-mean strategy yields more discriminative and interpretable subtype-specificity patterns across tumor types.

Appendix C Derivation and threshold selection of the surprise-information score

This section provides supplementary details for the surprise-information component of SDSGI. Specifically, we describe the statistical motivation of the score, the formal hypothesis-testing framework used for subtype-level comparison, the transformation from adjusted p-values to surprise-information scores, and the strategy used for threshold selection. These details are included here to keep the main text concise while preserving the full derivation of the method.

C.1 Wilcoxon rank-sum test for subtype-level comparison

The surprise-information component of SDSGI was introduced to improve sensitivity to low-abundance subtype-associated transcripts, especially lncRNAs, which are often weakly expressed and strongly affected by sparsity in scRNA-seq data. This component is motivated by Shannon's self-information theory, in which low-probability events carry larger amounts of information. Under this view, a gene showing an unexpectedly biased expression pattern in a specific subtype may still be biologically informative even if its absolute expression level is low.

For a given gene i and subtype j, let

$$G_{ij} = \{x_{1j}, x_{2j}, \ldots, x_{n_1 j}\}$$

denote the expression vector of gene i in subtype j, and let

$$\bar{G}_{ij} = \{x'_{1j}, x'_{2j}, \ldots, x'_{n_2 j}\}$$

denote the expression vector of the same gene in all remaining subtypes. To evaluate whether gene i is significantly enriched or depleted in subtype j, SDSGI applies the Wilcoxon rank-sum test to compare G_{ij} and $\bar{G}_{ij}$.

Let R_1 and R_2 denote the rank sums of the two groups. The corresponding Mann–Whitney statistics are

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \tag{S2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}, \tag{S3}$$

and the test statistic is defined as

$$U = \min(U_1, U_2). \tag{S4}$$

The raw p-value p_{ij} is adjusted for repeated comparisons as

$$\tilde{p}_{ij} = 1 - (1 - p_{ij})^t, \tag{S5}$$

where t denotes the number of repeated comparisons. When p_{ij} is extremely small, a Taylor-series approximation can be used to improve numerical stability.

The surprise-information score is then defined as

$$S_{ij} = -\log(\tilde{p}_{ij}). \tag{S6}$$

To distinguish subtype-specific enrichment from subtype-specific depletion, a sign is assigned according to the direction of the Wilcoxon statistic. The expected value of the Mann–Whitney statistic under the null hypothesis is

$$E(U) = \frac{n_1 n_2}{2}. \tag{S7}$$

If the expression of gene i tends to be higher in subtype j than in the remaining subtypes, S_{ij} is assigned a positive sign; otherwise, it is assigned a negative sign. Therefore, positive values indicate subtype-enriched expression, whereas negative values indicate subtype-depleted expression.

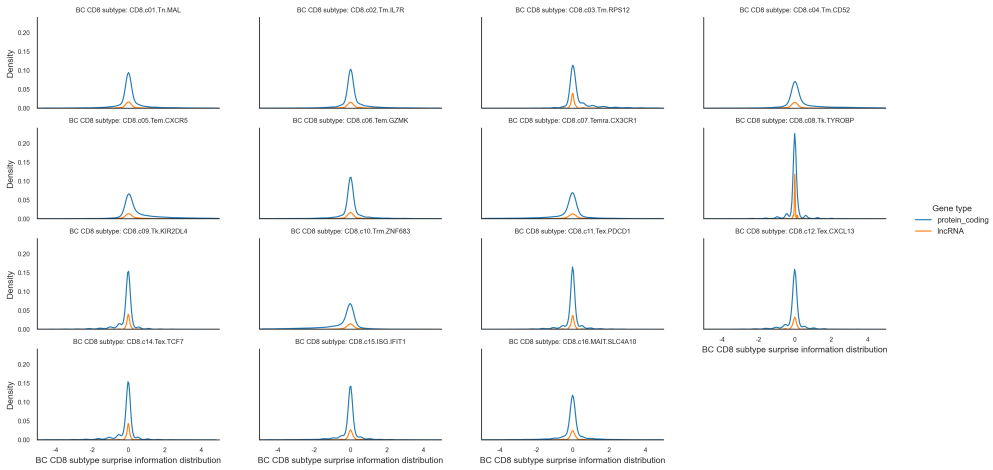


Fig. S3 Distribution of surprise-information-based subtype-specific scores in CD8$^+$ T-cell subtypes in breast cancer

As shown in Figure S3, the distributions of surprise-information scores in breast cancer CD8$^+$ T-cell subtypes are centered near zero and decay toward both sides,

with a visible elbow in the tails. Following the original analysis, Kneedle was used to identify the elbow point for threshold determination. The selected threshold was close to $|S_{ij}| = 3$, which is also consistent with statistical significance because $S_{ij} \approx 3$ when $p_{ij} = 0.05$ after transformation. This threshold therefore provides a practical criterion for selecting statistically informative subtype-associated genes.

C.2 Comparison and integration of the two candidate gene sets

SDSGI integrates two complementary scoring strategies for subtype-specific gene identification: the expression-based τ score and the surprise-information-based score. To evaluate the relationship between the candidate gene sets generated by these two components, we compared their overlap across breast cancer CD8$^+$ T-cell subtypes.

As shown in Figure S4, the overlap between the two candidate gene sets is generally limited across subtypes, whereas each strategy also identifies a substantial number of subtype-specific genes not recovered by the other. This pattern indicates that the two scoring strategies capture different but complementary aspects of subtype-associated expression.

Specifically, the expression-based τ score tends to prioritize genes with relatively clear subtype-restricted expression patterns, whereas the surprise-information-based score is more sensitive to genes with weaker absolute expression but statistically informative subtype bias. Therefore, neither strategy alone is sufficient to capture the full spectrum of biologically meaningful subtype-specific genes in sparse scRNA-seq data.

Based on this complementarity, SDSGI defines the final subtype-specific gene set as the union of the two candidate sets. This integration preserves both strongly subtype-enriched genes and low-abundance but informative genes that may be missed by either strategy alone, thereby improving the sensitivity and robustness of subtype-specific gene identification.

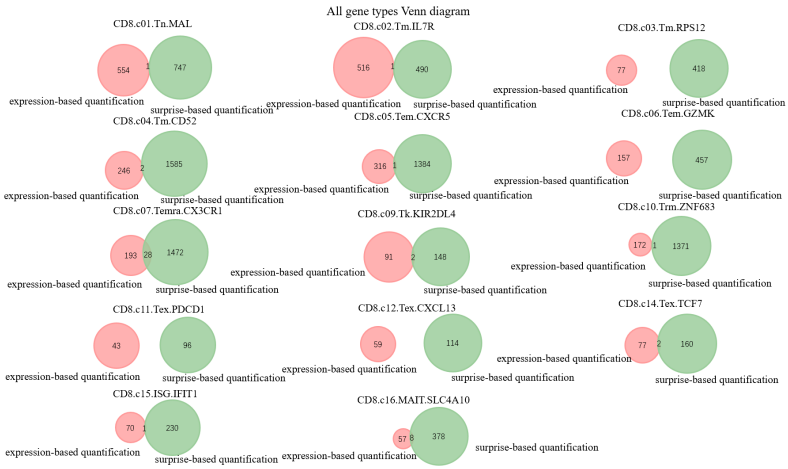


Fig. S4 Venn diagram of subtype-specific genes identified by the two features in breast cancer.

C.3 Correlation calculation and threshold selection for co-expression networks

To construct gene co-expression networks, pairwise correlations between genes were calculated based on their expression profiles across cells. Co-expression networks are built on the assumption that genes exhibiting coordinated expression patterns across samples are likely to be functionally related or co-regulated within shared biological processes.

C.3.1 Pearson correlation-based co-expression measurement

In this study, the Pearson correlation coefficient was used to quantify the linear association between gene-expression profiles. For two genes i and j, the Pearson correlation coefficient is defined as:

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}, \tag{S8}$$

where x_{ik} and x_{jk} denote the expression levels of genes i and j in sample k, respectively, and $\bar{x}_i$ and $\bar{x}_j$ are their mean expression values across all samples.

Pearson correlation was adopted due to its computational efficiency, interpretability, and its ability to capture coordinated linear expression patterns, which are commonly observed in gene regulatory systems. Positive correlations ($r > 0$) indicate coordinated up- or down-regulation, while negative correlations ($r < 0$) may reflect inhibitory or opposing regulatory relationships.

C.3.2 Matrix-based implementation

To improve computational efficiency for large-scale data, correlation coefficients were computed using matrix operations. Let $A \in \mathbb{R}^{m\times n}$ denote the gene-expression matrix with m genes and n samples. The computation proceeds as follows:

1. Compute the mean expression vector:

$$M = \frac{1}{n}A\mathbf{1}_n \tag{S9}$$

2. Center the expression matrix:

$$A' = A - M \tag{S10}$$

3. Compute the standard deviation vector:

$$S = \operatorname{diag}\left(\sqrt{\frac{1}{n-1}(A' \circ A')\mathbf{1}_n}\right) \tag{S11}$$

4. Compute the correlation matrix:

$$R = S^{-1}\left(\frac{1}{n-1}A'(A')^T\right)S^{-1} \tag{S12}$$

Here, $\mathbf{1}_n$ is a vector of ones, $\circ$ denotes element-wise multiplication, and S is the diagonal matrix of standard deviations.

C.3.3 Statistical significance of correlations

To assess the statistical significance of gene–gene correlations, the correlation coefficients were transformed into t-statistics:

$$T = R \times \sqrt{\frac{n-2}{1-R^2}}, \tag{S13}$$

and the corresponding two-sided p-values were computed as:

$$P = 2 \times (1 - F_t(|T|, \nu)), \tag{S14}$$

where $\nu = n - 2$ is the degree of freedom and $F_t(\cdot)$ denotes the cumulative distribution function of the t-distribution.

C.3.4 Threshold selection and network construction

Edges in the co-expression network were defined based on statistically significant correlations. Only gene pairs with correlation coefficients exceeding a predefined threshold and corresponding p-values below a significance level were retained.

To further ensure biological relevance and reduce noise, the threshold was selected by comparing the clustering coefficient of the real network with that of corresponding randomized networks. The optimal threshold was chosen as the point where the difference between real and random network clustering coefficients was maximized, thereby balancing network sparsity and functional modularity.

The resulting adjacency matrix was then used to construct subtype-specific gene co-expression networks, which served as the basis for downstream module detection and lncRNA function inference.

Appendix D Complete enrichment results for breast-cancer subtype-specific gene sets

This section provides the complete enrichment results corresponding to the representative GO and Reactome terms discussed in the main text. Breast cancer was selected as a representative case for downstream functional analysis. Here, we list the top 10 enriched GO terms for the subtype-specific gene set of CD8.c07.Temra.CX3CR1 and the top 10 enriched Reactome pathways for three exhausted T-cell-related subtypes, namely CD8.c11.Tex.PDCD1, CD8.c12.Tex.CXCL13, and CD8.c14.Tex.TCF7.

In addition to the tabular results, we provide visual summaries of the enrichment analysis for easier interpretation:

These supplementary tables and figures provide the full context for the enrichment patterns discussed in the main text, facilitating a comprehensive understanding of subtype-specific functional profiles.

Table S4 Complete GO enrichment results for the subtype-specific gene set of CD8.c07.Temra.CX3CR1 in breast cancer

GO term	Description	p value	Gene count	Odds ratio
GO:0001912	Positive regulation of leukocyte mediated cytotoxicity	1.09e-06	12	14.84
GO:0042269	Regulation of natural killer cell mediated cytotoxicity	1.31e-05	9	18.75
GO:0045953	Negative regulation of natural killer cell mediated cytotoxicity	1.32e-05	7	37.13
GO:0006968	Cellular defense response	2.79e-05	10	12.63
GO:1902533	Positive regulation of intracellular signal transduction	3.35e-05	32	3.21
GO:0045954	Positive regulation of natural killer cell mediated cytotoxicity	6.09e-05	7	23.86
GO:0046631	Alpha-beta T cell activation	1.19e-04	6	31.75
GO:0002717	Positive regulation of natural killer cell mediated immunity	1.28e-04	7	19.65
GO:0072678	T cell migration	2.22e-04	6	25.97
GO:0071222	Cellular response to lipopolysaccharide	4.47e-04	13	5.64

Table S5 Reactome enrichment results for CD8.c11.Tex.PDCD1 in breast cancer

Pathway description	p value	Odds ratio	Gene count
PD-1 Signaling	8.14e-16	135.42	9
Translocation of ZAP-70 to Immunological Synapse	2.80e-14	138.26	8
Phosphorylation of CD3 and TCR Zeta Chains	7.62e-14	118.49	8
Costimulation by the CD28 Family	1.97e-13	47.10	10
Generation of Second Messenger Molecules	1.64e-12	75.35	8
Downstream TCR Signaling	1.26e-09	29.88	8
TCR Signaling	7.15e-09	23.58	8
Interferon Gamma Signaling	2.49e-09	27.24	8
Cytokine Signaling in Immune System	8.73e-08	6.62	15
MHC Class II Antigen Presentation	3.49e-07	17.92	7

Table S6 Reactome enrichment results for CD8.c12.Tex.CXCL13 in breast cancer

Pathway description	p value	Odds ratio	Gene count
Immune System	5.58e-05	2.96	22
Cytokine Signaling in Immune System	9.29e-05	4.17	12
Translocation of ZAP-70 to Immunological Synapse	1.73e-04	32.03	3
MHC Class II Antigen Presentation	1.90e-04	10.35	5
Adaptive Immune System	2.26e-04	3.77	12
Phosphorylation of CD3 and TCR Zeta Chains	2.41e-04	28.33	3
NGF-stimulated Transcription	5.85e-04	20.45	3
Generation of Second Messenger Molecules	6.78e-04	19.37	3
Signal Transduction	9.13e-04	2.37	22
Disease	1.13e-03	2.46	19

Table S7 Reactome enrichment results for CD8.c14.Tex.TCF7 in breast cancer

Pathway description	p value	Odds ratio	Gene count
Peptide Chain Elongation	3.80e-26	55.19	20
Eukaryotic Translation Elongation	1.18e-25	51.69	20
Selenocysteine Synthesis	2.30e-22	45.01	18
Eukaryotic Translation Termination	2.30e-22	45.01	18
Nonsense Mediated Decay Independent of the Exon Junction Complex	3.40e-22	43.90	18
Viral mRNA Translation	5.00e-22	42.85	18
Response of EIF2AK4 (GCN2) to Amino Acid Deficiency	1.04e-21	40.90	18
Formation of a Pool of Free 40S Subunits	1.04e-21	40.90	18
L13a-mediated Translational Silencing of Ceruloplasmin Expression	5.79e-21	36.71	18

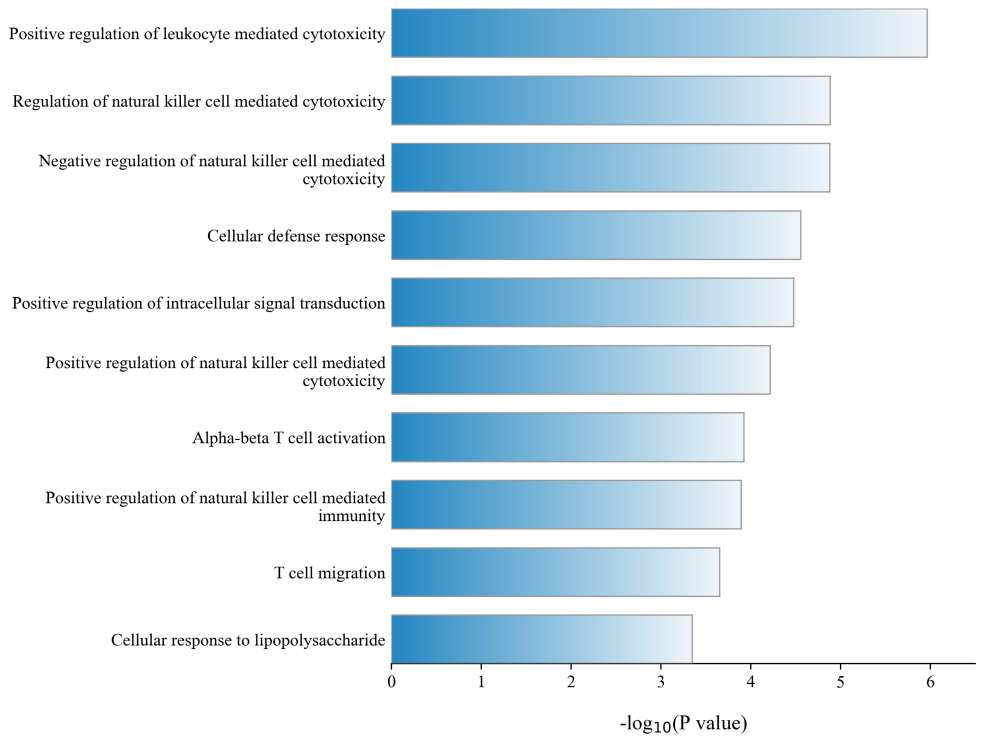


Fig. S5 Top 10 enriched GO terms for CD8.c07.Temra.CX3CR1 displayed as a horizontal bar chart. The length of each bar corresponds to $-\log_{10}$(P value), highlighting the statistical significance of each term.

D.1 Threshold selection for subtype-specific co-expression networks

To construct biologically meaningful subtype-specific co-expression networks, it was necessary to determine an appropriate correlation threshold for edge retention. If the threshold is set too low, the network remains overly dense and contains many weak or noisy correlations, making downstream module detection difficult to interpret. In contrast, if the threshold is set too high, a large number of informative edges may be removed, resulting in fragmented networks and loss of functional structure. Therefore, the correlation threshold should be selected in a data-driven manner rather than assigned arbitrarily.

In this study, gene co-expression was quantified using the Pearson correlation coefficient. For two genes i and j, the correlation coefficient was defined as

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}, \tag{S15}$$

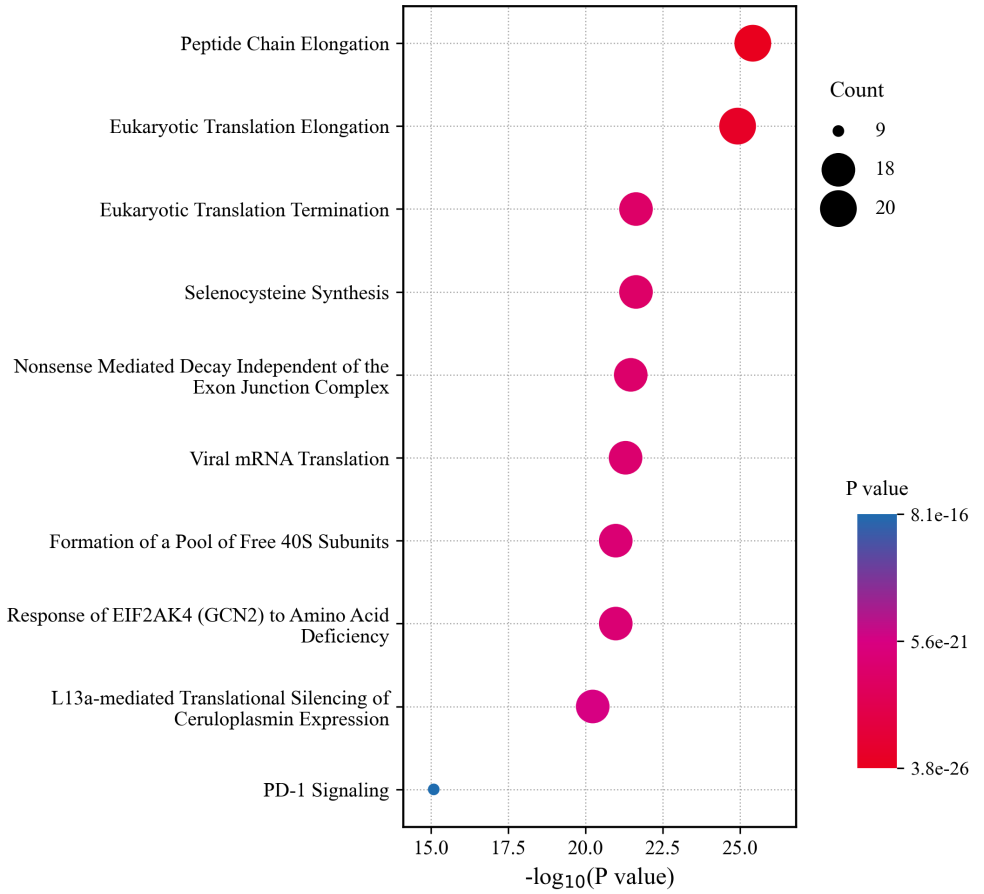


Fig. S6 Top 10 enriched Reactome pathways for CD8.c11.Tex.PDCD1, CD8.c12.Tex.CXCL13, and CD8.c14.Tex.TCF7 displayed as a bubble plot. The size of each bubble indicates the number of subtype-specific genes associated with the pathway, while the color intensity represents the statistical significance ($-\log_{10}$(P value)).

where x_{ik} and x_{jk} denote the expression values of genes i and j in cell k, and $\bar{x}_i$ and $\bar{x}_j$ are the corresponding mean expression values across all cells. Based on the resulting correlation matrix, candidate edges were retained under different correlation thresholds to generate subtype-specific co-expression networks.

Because different subtype-specific gene sets differed substantially in size and connectivity, a single fixed cutoff was not used for all subtypes. Instead, Pearson correlation thresholds were scanned within the range of 0.60–0.90. The distribution of pairwise gene correlations under this procedure is illustrated in Supplementary Figure S7.

To evaluate which threshold best preserves meaningful network structure, the clustering coefficient of the real network was compared with that of the corresponding

randomized network. Let $C(r)$ denote the clustering coefficient of the real network under threshold r, and let $C_0(r)$ denote the clustering coefficient of the randomized network under the same threshold. Their difference

$$\Delta C(r) = C(r) - C_0(r) \tag{S16}$$

was used to assess how strongly the observed network deviated from random structure. When $\Delta C(r)$ increases, the removed edges are mainly weak and noisy connections, while the local modular organization of the real network is still preserved. When $\Delta C(r)$ starts to decrease, it indicates that the threshold has become too stringent and that biologically meaningful edges are also being discarded.

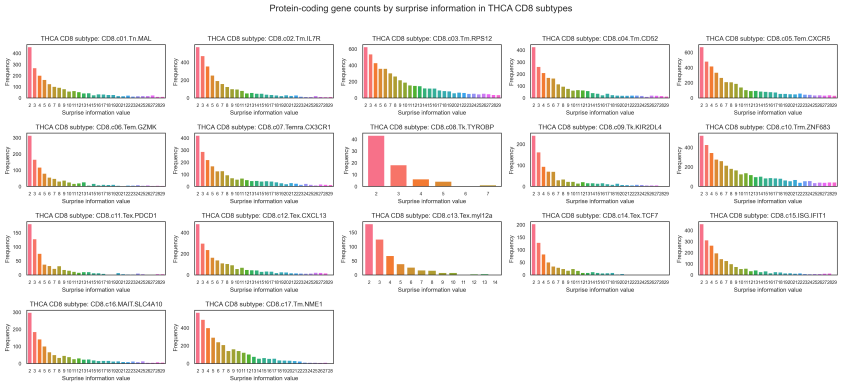


Fig. S7 Relationship between surprise-information thresholds and subtype-specific lncRNA counts in thyroid cancer

Accordingly, the optimal threshold was selected at the local maximum of the $\Delta C(r)$ curve. In other words, the chosen threshold corresponds to the point at which the difference between the real and randomized networks is maximized, representing the best balance between denoising and preservation of functional modularity. The mathematical criterion can be written as

$$C^* = \min_j \{r_j : C(r_j) - C_0(r_j) > C(r_{j+1}) - C_0(r_{j+1})\}. \tag{S17}$$

For the randomized network, the clustering coefficient was estimated as

$$C_0 = \frac{(\bar{k}^2 - \bar{k})^2}{\bar{k}^3 N}, \tag{S18}$$

where $\bar{k}$ is the average node degree, $\bar{k}^2$ is the second moment of the degree distribution, $\bar{k}^3$ is the average cubic degree, and N is the total number of nodes in the network.

Following this strategy, we calculated the clustering-coefficient differences between real and randomized networks across the threshold range of 0.60–0.90. A representative example is shown for the breast-cancer subtype CD8.c01.Tn.MAL in

Supplementary Table S9, and the corresponding $\Delta C(r)$ curve is shown in Supplementary Figure S8. The final correlation thresholds used for network construction, together with the numbers of retained nodes and retained edges after threshold filtering, are summarized in Table S8. These parameters provide the basis for subsequent module detection and functional analysis of lncRNA-containing co-expression modules.

Table S8 Subtype-specific co-expression network thresholds in breast cancer

Subtype	Correlation threshold	Retained nodes	Retained edges
CD8.c01.Tn.MAL	0.70	531	1,293
CD8.c02.Tm.IL7R	0.71	1,322	425
CD8.c03.Tm.RPS12	0.63	399	232
CD8.c04.Tm.CD52	0.70	2,163	1,064
CD8.c05.Tem.CXCR5	0.60	10,962	1,357
CD8.c06.Tem.GZMK	0.63	784	368
CD8.c07.Temra.CX3CR1	0.65	9,663	1,527
CD8.c09.Tk.KIR2DL4	0.62	164	123
CD8.c10.Trm.ZNF683	0.63	3,154	1,228
CD8.c11.Tex.PDCD1	0.63	67	49
CD8.c12.Tex.CXCL13	0.63	94	75
CD8.c14.Tex.TCF7	0.66	331	139
CD8.c15.ISG.IFIT1	0.65	154	135
CD8.c16.MAIT.SLC4A10	0.67	257	230

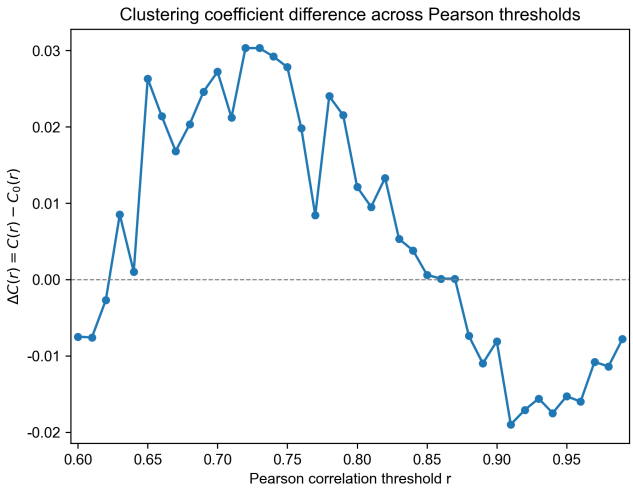
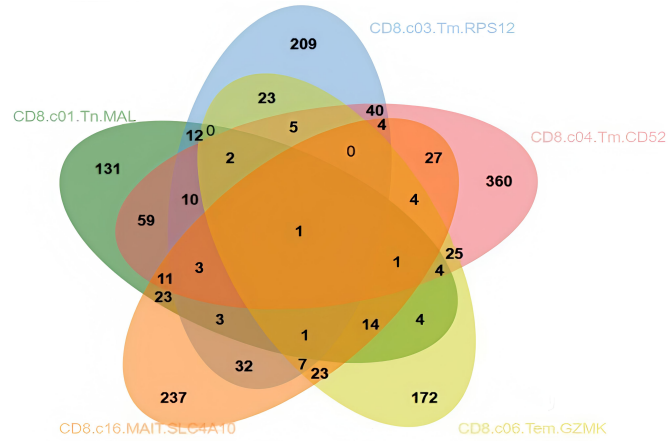


Fig. S8 Variation of the difference between the clustering coefficients of the real network and the random network

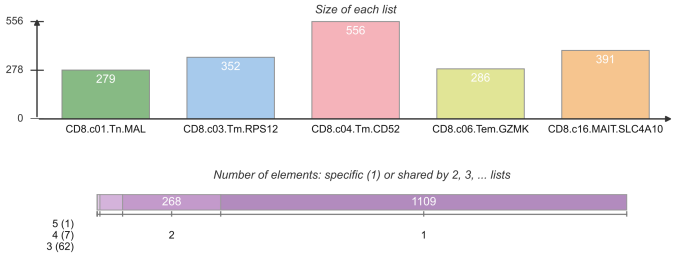
After selecting the optimal correlation thresholds for each subtype, subtype-specific co-expression networks were analyzed using the ClusterONE algorithm to identify densely connected modules. These results provide a comprehensive view of the modular organization across all breast-cancer T-cell subtypes and serve as the basis for subsequent functional analysis of lncRNA-containing modules.

To further support the functional interpretation of lncRNA-containing modules discussed in the main text, we visualized the overlap of enriched GO terms among representative CD8$^+$ T-cell subtypes. As shown in Supplementary Figure S9, most enriched GO terms were subtype-specific, whereas a small number of biological processes were shared across multiple subtypes. This supplementary visualization provides additional context for the GO-term overlap patterns and supports the module-based lncRNA functional inference.

To further illustrate the connectivity relationships among the three Tex cell subtypes, the PAGA graph is presented in Figure S10.



(a) Overlap of enriched GO terms among five CD8$^+$ T-cell subtypes



(b) Size of enriched GO term sets and distribution of shared GO terms

Fig. S9 Comprehensive visualization of enriched GO terms in CD8$^+$ T-cell subtypes. (a) Overlap relationships among subtypes. (b) Size of each GO term set and distribution of shared GO terms.

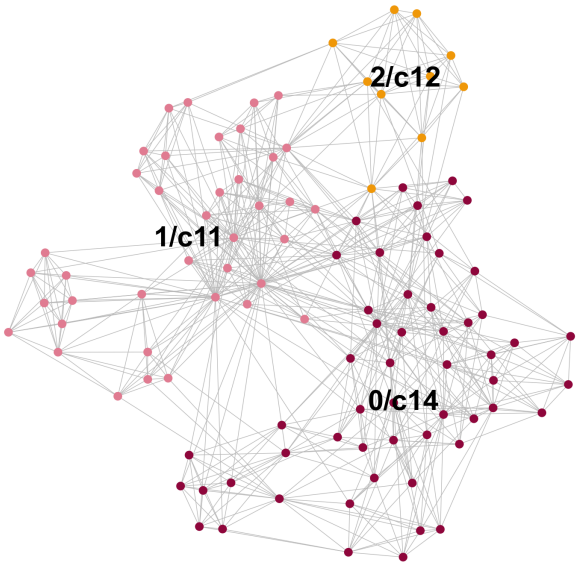


Fig. S10 PAGA graph of the three subtype Tex cells

Table S9 Changes in clustering-coefficient difference under different correlation thresholds for subtype CD8.c01.Tn.MAL

Threshold	C	C_0	$C - C_0$	Threshold	C	C_0	$C - C_0$
0.60	0.0977	0.1052	-0.0075	0.75	0.0574	0.0296	0.0278
0.61	0.0970	0.1046	-0.0076	0.76	0.0475	0.0277	0.0198
0.62	0.0939	0.0966	-0.0027	0.77	0.0400	0.0316	0.0084
0.63	0.0939	0.0854	0.0085	0.78	0.0389	0.0149	0.0240
0.64	0.0917	0.0907	0.0010	0.79	0.0313	0.0098	0.0215
0.65	0.0911	0.0648	0.0263	0.80	0.0289	0.0168	0.0121
0.66	0.0883	0.0669	0.0214	0.81	0.0276	0.0181	0.0095
0.67	0.0808	0.0640	0.0168	0.82	0.0255	0.0122	0.0133
0.68	0.0786	0.0583	0.0203	0.83	0.0196	0.0143	0.0053
0.69	0.0768	0.0522	0.0246	0.84	0.0186	0.0148	0.0038
0.70	0.0749	0.0477	0.0272	0.85	0.0121	0.0115	0.0006
0.71	0.0695	0.0483	0.0212	0.86	0.0091	0.0090	0.0001
0.72	0.0642	0.0339	0.0303	0.87	0.0052	0.0051	0.0001
0.73	0.0579	0.0276	0.0303	0.88	0.0049	0.0123	-0.0074
0.74	0.0579	0.0287	0.0292	0.89	0.0036	0.0146	-0.0110

References

[1] 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)

[2] Luecken, M.D., Theis, F.J.: Current best practices in single-cell rna-seq analysis: a tutorial. Molecular systems biology **15**(6), 188746 (2019)

[3] 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)

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

[5] 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)

[6] 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)

[7] Osorio, D., Cai, J.J.: Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell rna-sequencing data quality control. Bioinformatics **37**(7), 963–967 (2021)

[8] Glover, H.L., Schreiner, A., Dewson, G., Tait, S.W.: Mitochondria and cell death. Nature cell biology **26**(9), 1434–1446 (2024)

[9] Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.*: Massively parallel digital transcriptional profiling of single cells. Nature communications **8**(1), 14049 (2017)

[10] Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O.: Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. Nature biotechnology **33**(2), 155–160 (2015)

[11] Hicks, S.C., Townes, F.W., Teng, M., Irizarry, R.A.: Missing data and technical variability in single-cell rna-sequencing experiments. Biostatistics **19**(4), 562–578 (2018)