IET Systems Biology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

CFMF: A Clustering-Free Cell Marker Finder for Single-Cell Transcriptomic Data

Shumin Yin | Ke Liu

Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China

Correspondence: Ke Liu (keliu.iluke@email.sdu.edu.cn)

Received: 30 October 2025 | **Revised:** 8 January 2026 | **Accepted:** 13 January 2026

Keywords: marker gene identification | rare cell detection | single-cell RNA sequencing | tumor heterogeneity

ABSTRACT

The discovery of novel cell types and their marker genes is a fundamental goal of single-cell RNA sequencing (scRNA-seq). Nevertheless, most conventional methods rely on cell clustering, a process in which parameter selection is often non-trivial and subjective. To address such limitations, we developed Clustering-Free Cell Marker Finder (CFMF), a novel computational framework that enables the discovery of marker genes without clustering. We first validated CFMF using the PBMC3K dataset and found that it not only recovered canonical marker genes, but also uncovered novel marker genes. When validating with glioblastoma datasets, CFMF successfully recapitulated established cellular state signatures. Using human normal lung scRNA-seq datasets, we demonstrated its superior sensitivity in rare cell type detection even at very low prevalence. We utilised CFMF to perform integrative analysis on colorectal cancer scRNA-seq datasets and defined seven distinct transcriptional states including an $LGR5^{+}$ state that is strongly associated with metastasis. Finally, we systematically benchmarked CFMF against widely used gene selection methods across multiple scRNA-seq datasets and revealed that CFMF appeared to more effectively identify biologically meaningful genes. In summary, we provide a powerful framework for the discovery of marker genes and the analysis of tumour heterogeneity, which may facilitate our understanding of complex biological systems.

1 | Introduction

Recent advances in single-cell transcriptomic technologies have enabled the construction of comprehensive cellular atlases across diverse species and organs [1–5], providing novel insights into tissue architecture and cellular function. A critical step in scRNA-seq data analysis is cell type annotation, which involves assigning individual cells to established cell types or functional states [6]. Annotation strategies can be broadly categorised into manual and automated approaches: manual annotation defines cell identity primarily based on the expression of known marker genes within cell clusters, whereas automated annotation typically leverages machine learning or statistical modelling methods to infer cell identity by comparing single-cell transcriptomic profiles with reference datasets representing established cell types [7].

Despite its widespread adoption, manual annotation exhibits several limitations. First, it inherently depends on the availability of highly specific marker genes [8] and requires extensive domain knowledge, thereby introducing subjectivity into the annotation process. This limitation is particularly evident in tumour subtyping, where substantial intratumoral heterogeneity [9, 10] obscures the delineation of distinct subclonal populations and hinders the identification of reliable subtype-specific marker genes. Second, manual annotation typically begins with clustering, yet the clustering outcome is strongly influenced by parameter choices (e.g., the number of clusters). The use of inappropriate parameters may artificially split a single cell type into multiple clusters or fail to capture rare cell types [11]. In particular, rare cell types [12, 13] are often masked by more abundant populations, leading to their misclassification and the missed opportunity to discover novel cell types. Therefore, there is

an urgent need to develop clustering-independent methods that can robustly identify highly specific marker genes and thereby improve the reliability of cell type annotation.

In this study, we present CFMF, an innovative framework for marker gene identification that operates independently of both clustering and prior domain knowledge. CFMF is founded on the hypothesis that cells expressing the same marker gene exhibit spatial proximity in the transcriptomic landscape, thereby forming localized aggregations. We evaluated CFMF across multiple normal and malignant datasets and demonstrated its ability to recover both canonical marker genes of established cell types and novel marker genes, including those specific to rare cell types. Moreover, when using CFMF to delineate intratumoral heterogeneity of colorectal cancer (CRC), we uncovered seven distinct transcriptional states. Notably, transcriptional state 4 was enriched in metastatic CRC, and exhibited an inverse correlation with $CD8^+$ T cell infiltration, implicating a potential role in immune evasion and metastatic progression. Finally, we systematically compared CFMF with other established gene selection methods and found that CFMF identified more biologically meaningful genes.

2 | Methods

2.1 | Data Acquisition and Preprocessing

A total of 13 scRNA-seq datasets [14–25] and one CRC spatial transcriptomics dataset [26] were collected (Table S1). Raw gene expression matrices were processed using Seurat (v4.3.0) [27]. Dimensionality reduction was performed using principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) was computed on the first 30 principal components using Seurat default parameters for two-dimensional visualisation.

2.2 | Cell Type Annotation

SingleR was used to perform automatic cell type annotation with the BlueprintEncodeData as the reference [28]. Copy number variation (CNV) profiles were inferred by inferCNV (v1.12.0) (using immune cells as the reference population) [29], and were further used to identify malignant cells.

2.3 | The Full CFMF Pipeline

The CFMF pipeline comprised the following steps:

1. Gene Expression Filtering. Genes detected in at least 3 cells but not expressed in all cells were retained.
2. Dimensionality reduction. PCA was performed on HVGs and the first 50 PCs were used for subsequent neighbour calculations.
3. Local neighbourhood enrichment. For each gene, positive cells were defined as those with nonzero expression. For each positive cell, the 50 nearest neighbours were identified in PCA space using get.knn in the FNN package with Euclidean distance. Fisher's exact test was applied per positive cell to assess enrichment of positive neighbours. The resulting per-cell p-values were combined via meta-analysis and subsequently adjusted using the Benjamini-Hochberg procedure to yield a final meta p.
4. Global compactness. Silhouette [30] values were computed (positive vs. non-positive) in PCA space for each positive cell and the median was recorded.
5. Threshold selection and validation. To determine optimal thresholds, we established a reference set of high-confidence marker genes using the PBMC3K dataset. Specifically, we first identified 'marker genes' using the FindAllMarkers function of Seurat (avg_log2FC $>$ 0.75, adjusted p-value $<$ 0.01) and then selected those with high cell-type specificity using the scoreMarkers function (AUC $>$ 0.8) [31]. For the significance threshold, we observed that genes with a -log10(meta p) $>$ 200 were enriched for high-specificity markers (AUC $>$ 0.95, Table S2). For the silhouette score, a cutoff of $>$ 0 was chosen as it inherently indicates that cells expressing the gene are more transcriptionally similar to each other than to non-expressing cells. Consequently, operational cutoffs of -log10(meta p) $>$ 200 and silhouette $>$ 0 were selected.

2.4 | Pathway Activity Inference

Single-cell pathway activity was computed with AUCell (v1.20.2) [32] on raw RNA counts using MSigDB hallmark [33] and Reactome [34] gene sets. The active region was defined as the top 5% of expressed genes per cell (aucMaxRank = 5% of all genes). Hallmark and Reactome gene sets were obtained from the Enrichr database [35].

2.5 | Immune Infiltration Analysis

For each CRC scRNA-seq sample, the proportion of immune cells within the non-malignant compartment, and the relative abundance of each CRC transcriptional state among malignant cells were calculated. Associations between the two values were assessed using the Spearman rank correlation.

2.6 | Site Distribution Preference of CRC Transcriptional States

$R_{o/e}$ (Ratio of observed to expected frequency) was calculated with the Startrac (v0.1.0) package [36]. The $R_{o/e}$ value was defined as:

$$R_{o/e} = \frac{\text{Observed}_{i,j}}{\text{Expected}_{i,j}},$$

$$\text{Expected}_{i,j} = \text{Total}_i \times \frac{\text{Total}_j}{\text{Total}_{\text{all}}},$$

where $\text{Observed}_{i,j}$ is the number of cells from state i in tissue j, Total_i is the total number of cells in state i, Total_j is the total

number of cells in tissue j, and $\text{Total}_{\text{all}}$ is the total number of all cells across both tissues.

2.7 | Spatial Transcriptomic Data Analysis

A total of 26 CRC metastasis tumour sections were curated for analysis. ESTIMATE scores were computed per spatial spot using the ESTIMATE R package (v1.0.13), and converted to tumour purity via the established formula [37]:

$$\text{Tumor purity} = \cos(0.6049872018 + 0.0001467884 \times \text{ESTIMATE score}).$$

For each section, k-means clustering ($k = 5$) was applied to normalised expression profiles and the two clusters with highest ESTIMATE scores were designated as neoplastic.

The spatial deconvolution was performed with spacexr package (v2.2.1) [38]. A reference matrix was constructed by randomly sampling 5000 malignant CRC cells; transcriptional state proportions per spatial spot were estimated using the RCTD algorithm implemented in spacexr.

2.8 | Comparison With Established Feature-Selection Methods

Cell-type specificity was quantified using the scoreMarkers function in the scran package [31]. Given a gene, the area under the curve (AUC) statistic was first computed for every cell type to evaluate its discriminative power in distinguishing the target cell type from others. Then, the maximum AUC value across all cell types was defined as the final AUC score, serving as the global indicator of cell-type specificity.

3 | Results

3.1 | Overview of the CFMF Method

To address the limitations of conventional cell type annotation (which heavily relies on clustering algorithms and predefined marker genes), we developed a method named Clustering-Free cell Marker Finder (termed CFMF). CFMF is founded on the hypothesis that cells expressing the same marker gene should exhibit inherent transcriptional similarity and thus localise adjacently in transcriptomic space. Guided by this principle, CFMF first applies PCA to reduce the dimensionality of the expression matrix, and subsequently designates a gene as a candidate marker if it demonstrates specific expression across spatially adjacent cells (Figure 1A).

Given a gene g, we first identified all cells expressing g (termed g^+ cells). Specifically, for each g^+ cell, we then quantified the local enrichment of g^+ cells by applying Fisher's exact test to its 50 nearest neighbours in PCA space (Figure 1A). The resulting p-values derived for each g^+ cell were subsequently integrated via meta-analysis, yielding a single meta p-value that reflects the overall degree of local enrichment for gene g. Recognising that local metrics alone may fail to capture global distribution patterns, we further computed silhouette coefficients of g^+ cells to assess their transcriptomic cohesiveness (Figure 1A). Accordingly, an ideal marker gene is expected to exhibit both a low p-value and a high silhouette coefficient. Thresholds for both the p-value and silhouette score were established through pilot experiments (Section 2: Methods), with genes that satisfy the criteria of $\log_{10}(\text{meta } p\text{-value}) > 200$ and silhouette score > 0 being classified as potential marker genes.

3.2 | Validating CFMF With Public scRNA-Seq Datasets

To validate the performance of CFMF, we first applied it to the PBMC3K dataset, which comprises 2700 peripheral blood mononuclear cells from 10x Genomics [14]. As expected, CFMF accurately recovered canonical lineage marker genes of mononuclear cells, including *S100A8* for myeloid cells, *CD3E* for $CD4^+$ T cells, and *CD14* for classical monocytes (Figure 2A, Figure S1A). In addition, CFMF identified subtype-specific marker genes, such as *CCR7* and *LEF1*, which were specifically expressed in $CD4^+$ naïve T cells (Figure 2A, Figure S1A). Notably, CFMF also uncovered several novel marker genes for known cell types. For example, *PRSS23* and *AKR1C3* were broadly expressed in NK cells, *FOLR3* and *RBP7* in $CD14^+$ monocytes, and *HMOX1* in $CD16^+$ monocytes (Figure 2A,B, Figure S1B), thereby demonstrating the ability of CFMF to identify cell type-specific marker genes.

To assess the generalisability of CFMF, we next applied it to a glioblastoma (GBM) scRNA-seq dataset [39] comprising five cellular states. As anticipated, the CFMF-selected genes recapitulated previously reported GBM cellular state signature genes [39], including *CST3* and *S100B* for the astrocytic (AC) signature, as well as *BCAN* and *GPR17* for the oligodendrocyte progenitor cell (OPC) signature (Figure 2C). Moreover, CFMF revealed additional genes not included in previous signature lists but displaying pronounced cellular state-specific expression, including *LGALS3* and *TNFRSF1A* for the mesenchymal-like (MES) signature, *MLC1* and *AQP4* for the AC signature, *TNR* and *COL20A1* for the OPC signature, and *DCX* and *ELAVL4* for the neural progenitor-like (NPC) signature (Figure 2C,D, Figure S1C). These results demonstrate that CFMF effectively reveals the transcriptional heterogeneity of GBM.

Furthermore, we assessed the ability of CFMF-selected genes to recover cell-type specific upregulated genes in both the PBMC3K and GBM datasets. As expected, CFMF-selected genes were significantly enriched for upregulated DEGs across multiple cell types in PBMC3K, as well as for established GBM cellular state signature genes when compared with non-selected genes (OR = 73.7, $p = 1.65 \times 10^{-65}$, and OR = 50.4, $p = 1.18 \times 10^{-166}$, respectively; Fisher's exact test; Figure 2E,F). Collectively, CFMF provides a robust framework for identifying cellular heterogeneity across both normal and malignant tissues.

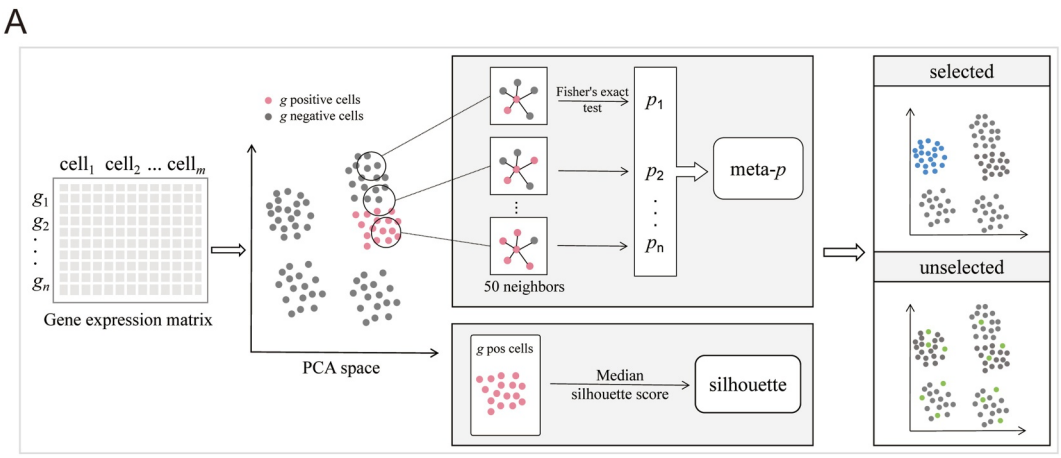

FIGURE 1 | Overview of the CFMF method. (A) Schematic illustrating the CFMF pipeline. Individual points in PCA space represent single cells. Grey points indicate cells with no detectable expression of the indicated gene (gene-negative), whereas pink points indicate cells with detectable expression (gene-positive). Right panels show the expression patterns of an ideal marker (top, cells expressing this gene are coloured blue) and a non-marker (bottom, cells expressing this gene are coloured green). The term 'selected' denotes genes satisfying the thresholds for both meta-*p* and silhouette metrics in CFMF, whereas the term 'unselected' comprises genes falling below these thresholds.

3.3 | CFMF Fuels the Identification of Rare Cell Types

Growing evidence suggests that rare cell types are pivotal in development, immune responses, and disease progression [40, 41]. Yet, their scarcity across tissues renders them difficult to resolve by conventional clustering-based scRNA-seq methods, particularly when confined to individual samples. Given the demonstrated ability of CFMF to identify biologically informative marker genes, we investigated whether it could also be applied to detect rare cell types. To this end, we first curated several normal lung scRNA-seq datasets previously reported to contain rare cell types [3], such as ionocytes and pulmonary neuroendocrine cells, and applied CFMF at each individual sample level to identify potential marker genes. Among the 42 samples subjected to individual clustering analysis, ionocytes did not form a distinct Seurat cluster in 28 cases, prompting us to focus subsequent analyses on these samples. When examining genes expressed in fewer than 5% of cells, we found that CFMF successfully identified marker genes of rare cell types, such as *SCG3* for pulmonary neuroendocrine cells (Figure 3A, Figure S2A–D), and *ASCL3* and *BSND* for ionocytes (Figure 3B, Figure S2A–D). CFMF identified at least one marker gene in 88% of the samples containing more than three rare cells. Remarkably, CFMF recovered ionocyte marker genes even when their prevalence within a sample declined to as low as 0.02%. Collectively, these findings highlight the superior sensitivity of CFMF in detecting transcriptional signatures of rare cell types.

3.4 | CFMF as a Powerful Tool for Delineating Tumour Heterogeneity

Tumour heterogeneity constitutes a critical barrier to effective therapeutic intervention, underscoring the necessity of delineating its molecular determinants [9, 10, 42]. Building on our previous findings that CFMF can effectively identify cellular state-associated marker genes in GBM, we applied CFMF to the scRNA-seq datasets of 109 CRC samples to identify CRC transcriptional states. As the identified genes might include marker genes of non-malignant cells (such as *NKG7* for NK cells), we implemented additional filtering steps to refine the CFMF-selected candidates (Figure 4A). For each sample, genes were processed as follows: (i) genes meeting the CFMF criterion were identified; (ii) genes for which more than half of the expressing cells were malignant were retained; (iii) genes expressed in more than half of all malignant cells were excluded; (iv) the top 1000 genes ranked by silhouette were selected. The candidate gene sets from all samples were then aggregated, and the 2000 most recurrent genes were designated as the final marker gene set. Notably, substituting these CFMF-selected genes for HVGs during batch correction enabled effective integration of malignant cells across samples, successfully mitigating batch effects that conventional HVG-based approaches failed to resolve (Figure 4B). These findings indicate that the recurrent CFMF-selected genes capture core biological features that are shared across diverse cancer samples.

Clustering analysis of malignant cells was performed using the top 2000 recurrent genes. After excluding 6 low-quality clusters characterised by reduced transcript counts and elevated ribosomal gene expression (Figure S3A), the remaining cells were reclustered, yielding 8 distinct transcriptional states (Figure 4C). To characterise the biological features of these states, we quantified MSigDB hallmark gene set activities [33] in each malignant cell using AUCell [32] and compared the activity profiles across groups (Figure 4D). TS1 (transcriptional state 1), marked by *S100A11* and *PHLDA2*, exhibited predominant upregulation of epithelial-mesenchymal transition and inflammatory response pathways, while TS2, defined by *ADH1C* and *OLFM4*, exhibited dysregulation of metabolic pathways. TS3 demonstrated a hybrid profile, combining features of both TS1 and TS2, thereby suggesting an intermediate state. TS4 displayed upregulation of the WNT signalling pathway and TS5 was distinguished by the activation of the MYC pathway, while TS6 lacked dominant

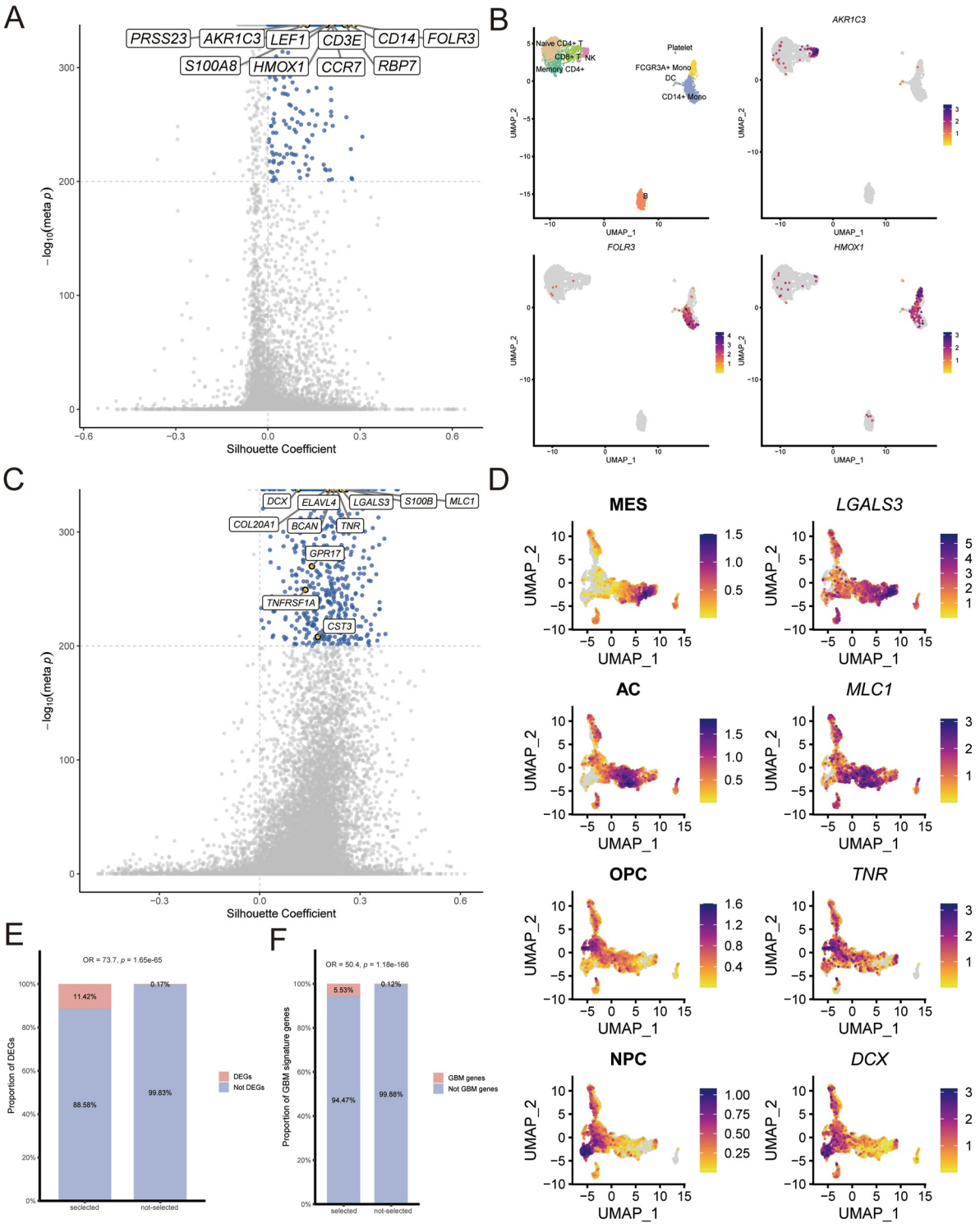

FIGURE 2 | Validation of CFMF using public single-cell RNA-seq datasets. (A) Scatterplot of silhouette coefficient versus meta-p value for all genes in the PBMC3K dataset. Points coloured in blue denote candidate marker genes ($-\log_{10}$(meta-p) > 200 and silhouette coefficient > 0); selected genes are labelled. (B) UMAP plots of the PBMC3K dataset showing annotated cell types and novel markers identified by CFMF. (C) Scatterplot of silhouette coefficient versus meta-p value for all genes in a representative GBM sample. Blue points denote candidate markers ($-\log_{10}$(meta-p) > 200 and silhouette coefficient > 0); selected genes are labelled. (D) UMAP plots of the GBM dataset showing four GBM cellular state signature scores (left) and the expression of the corresponding genes identified by CFMF (right). (E) Bar plot comparing the distribution of the top 10 upregulated DEGs per cell type in the PBMC3K dataset between CFMF-selected and non-CFMF genes. p-values were calculated by two-tailed Fisher's exact test. The term 'selected' denotes genes satisfying the thresholds for both meta-p and silhouette metrics in CFMF, whereas the term 'not-selected' comprises genes falling below these thresholds. (F) Bar plot comparing the distribution of cellular state-specific signature genes in the GBM dataset between CFMF-derived and non-CFMF genes. p-values were calculated by two-tailed Fisher's exact test. The term 'selected' denotes genes satisfying the thresholds for both meta-p and silhouette metrics in CFMF, whereas the 'not-selected' comprises genes falling below these thresholds.

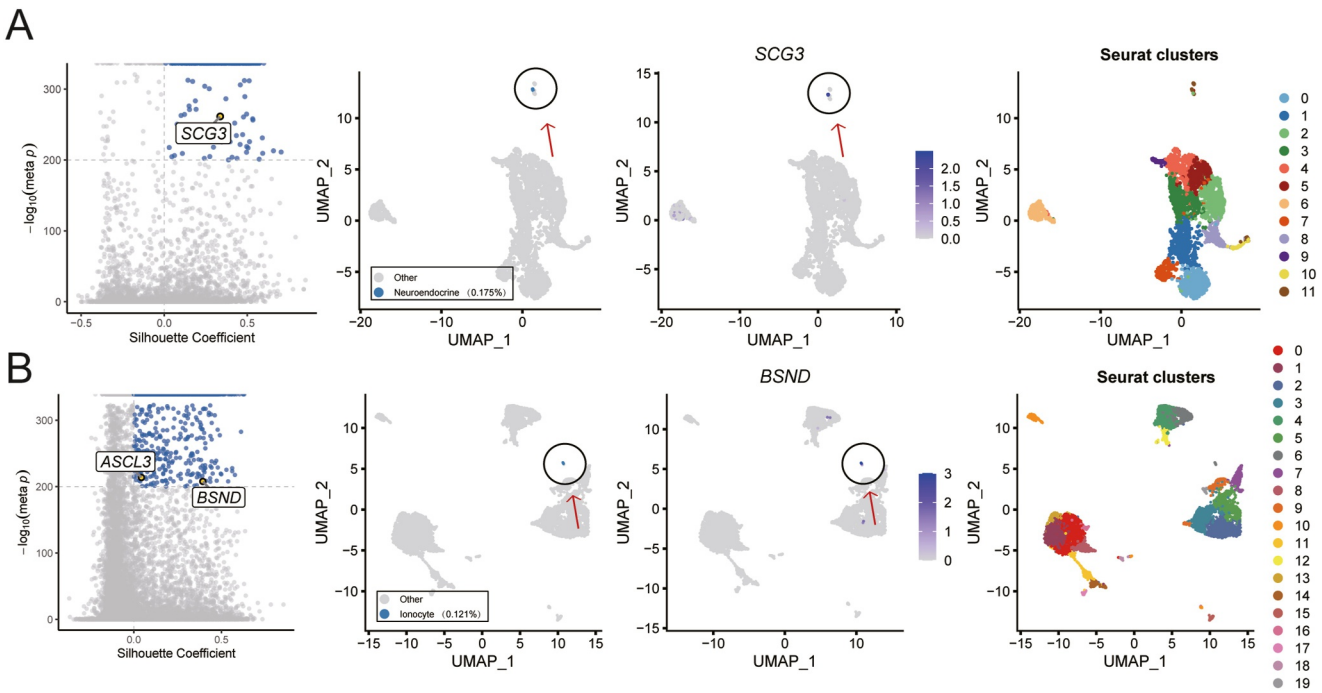

FIGURE 3 | CFMF facilitates identification of rare cell types. (A) Scatterplot of silhouette coefficient versus meta-p value for genes expressed in $< 5\%$ of cells in a representative human lung sample. Blue points denote candidate markers ($-\log_{10}$(meta-p) > 200 and silhouette coefficient > 0); selected genes are labelled. UMAP plots showing annotation and proportion of neuroendocrine cells, expression of the CFMF-identified marker, and Seurat cluster assignments obtained with default resolution, respectively. (B) Scatterplot of silhouette coefficient versus meta-p value for genes expressed in $< 5\%$ of cells in a representative human lung sample. Blue points denote candidate markers ($-\log_{10}$(meta-p) > 200 and silhouette coefficient > 0); selected genes are labelled. UMAP plots showing annotation and proportion of ionocytes, expression of the CFMF-identified marker, and Seurat cluster assignments obtained with default resolution, respectively.

hallmark pathway activities. TS7 and TS8 shared highly similar MYC pathway upregulation and proliferative characteristics, prompting their consolidation into a single transcriptional state. The robustness of these transcriptional state-specific characteristics was further validated by Reactome pathway analysis [34] (Figure S3B), which corroborated the hallmark-derived profiles. Consequently, we identified a total of 7 transcriptional states with distinct biological characteristics (Figure S4A,B).

When comparing transcriptional state distributions between primary tumours and metastatic sites, we observed that primary lesions were predominantly composed of TS2, TS3, TS6 and TS7, whereas metastatic sites were dominated by TS1, TS4 and TS5 (Figure 4E). Notably, TS4 showed the most pronounced enrichment at metastatic sites (Figure 4E) and was therefore prioritised for further analysis. This transcriptional state was characterised by elevated expression of *LGR5* (Figure S4B), a gene previously implicated in metastatic progression [43, 44]. Across samples, the proportion of TS4 was significantly negatively correlated with $CD8^+$ T cell abundance (Figure 4F), suggesting an immune-evasive or immune-poor state. Together, these findings underscore a potential role for TS4 in tumour metastasis. In addition, we compared the proportion of these transcriptional states in malignant spots using CRC spatial transcriptomic data [26], including 26 metastatic slices (Section 2: Methods). The results revealed that TS4 predominated within metastatic regions (Figure 4G), further highlighting its association with metastasis.

3.5 | Comparison of CFMF With Established Methods

To further assess the ability of CFMF to identify biologically informative genes, we benchmarked it against several established gene selection methods, including Fano [45], HVG [27], M3Drop [46], RaceID3 [47], SCTransform [48], and Entropy (SE) [49]. We investigated whether CFMF-selected genes could more effectively capture biological distinctions among cell types by exhibiting cell type-specific expression patterns. For each method, we calculated the AUC for the identified marker genes to quantify their expression specificity across cell types in the PBMC3K dataset and four GBM datasets. Notably, CFMF consistently outperformed all other methods in most cases, achieving higher AUC scores across a broad range of gene selection thresholds (top 30 to 2000 genes) (Figure 5A,B, Figures S5A,B and S6A,B). As a complementary validation, we evaluated the spatial autocorrelation [50] of identified genes in the PBMC3K dataset. Consistent with the AUC findings, CFMF exhibited significantly higher autocorrelation scores compared to other methods (Figure S6C). Collectively, these results underscore the effectiveness of CFMF in prioritising biologically meaningful genes in both normal and malignant single-cell transcriptomic contexts.

Given that genes exhibiting elevated expression in specific cell types generally yield higher AUC values, we next investigated whether CFMF-selected genes enhance the discrimination across cell types. To this end, we applied two complementary

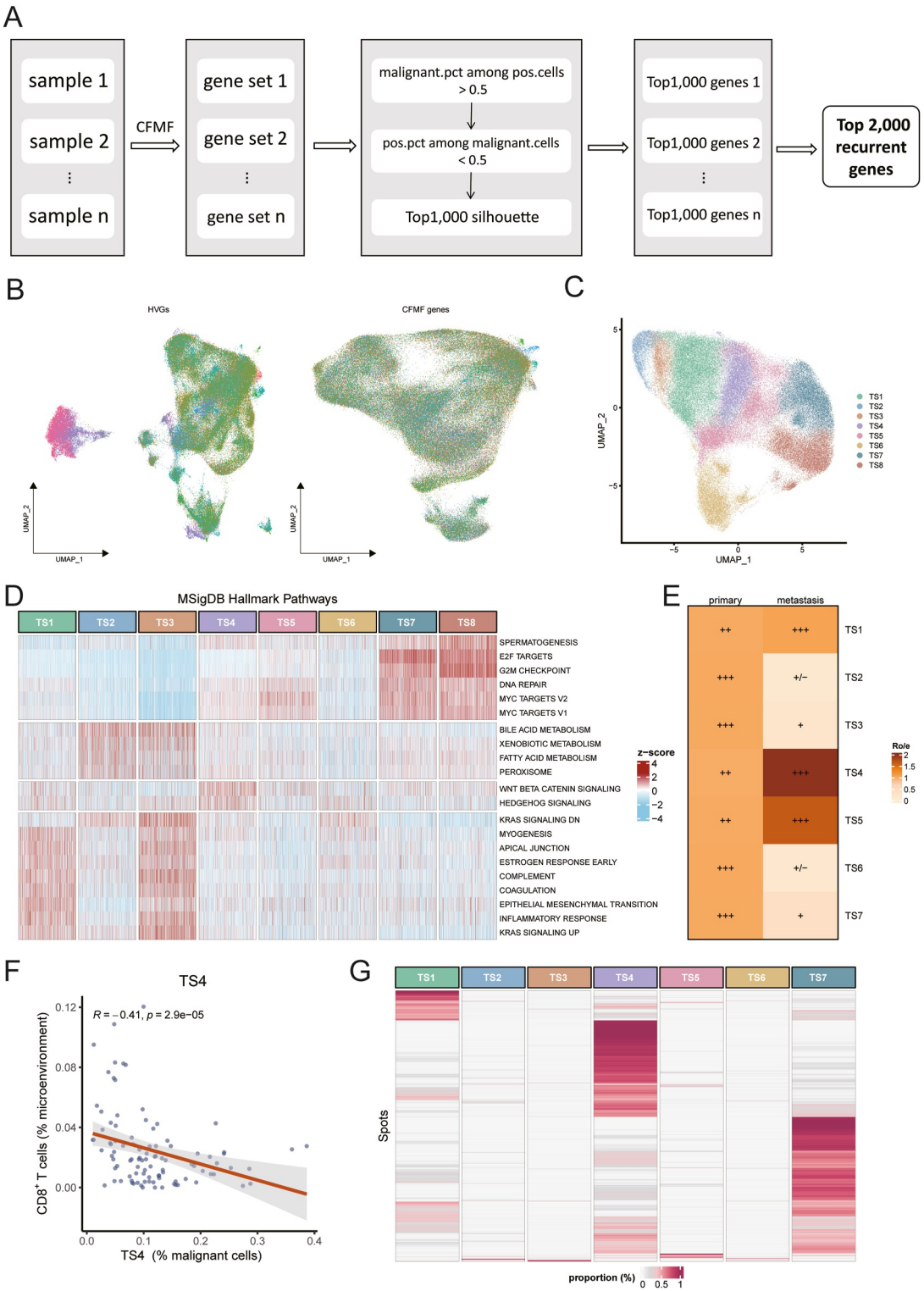

FIGURE 4 | CFMF as a powerful tool for delineating tumour heterogeneity. (A) Flow chart summarising the pipeline used to identify and prioritise markers of malignant cells in CRC. For an identified gene *g*, malignant.pct among pos.cells denotes the proportion of malignant cells among all cells expressing gene *g*, and pos.pct among malignant.cells denotes the proportion of cells expressing gene *g* among all malignant cells. The top 1000 genes refer to the 1000 genes with the highest silhouette scores. (B) UMAP plots demonstrating the integration of 109 CRC samples using top 2000 HVGs (left) and CFMF-selected genes (right). Different colours represent different samples. (C) UMAP plot of 8 distinct CRC transcriptional states. (D) Heatmap of MSigDB hallmark pathway activities across CRC transcriptional states; 1000 cells were randomly sampled per transcriptional state. (E) Primary versus metastasis prevalence of CRC transcriptional states measured by $R_{o/e}$. $R_{o/e}$ values were categorised as follows to reflect the degree of enrichment: +++, $R_{o/e} > 1$; ++, $0.8 < R_{o/e} \leq 1$; +, $0.2 \leq R_{o/e} \leq 0.8$; +/−, $0 < R_{o/e} < 0.2$; −, $R_{o/e} = 0$. (F) Scatterplots showing Spearman correlation between the fraction of TS4 among malignant cells and the fraction of CD8+ T cells in the tumour microenvironment. Spearman's correlation and *p*-value are shown. (G) Heatmap showing the proportion of transcriptional states in malignant spots across 26 CRC metastatic samples.

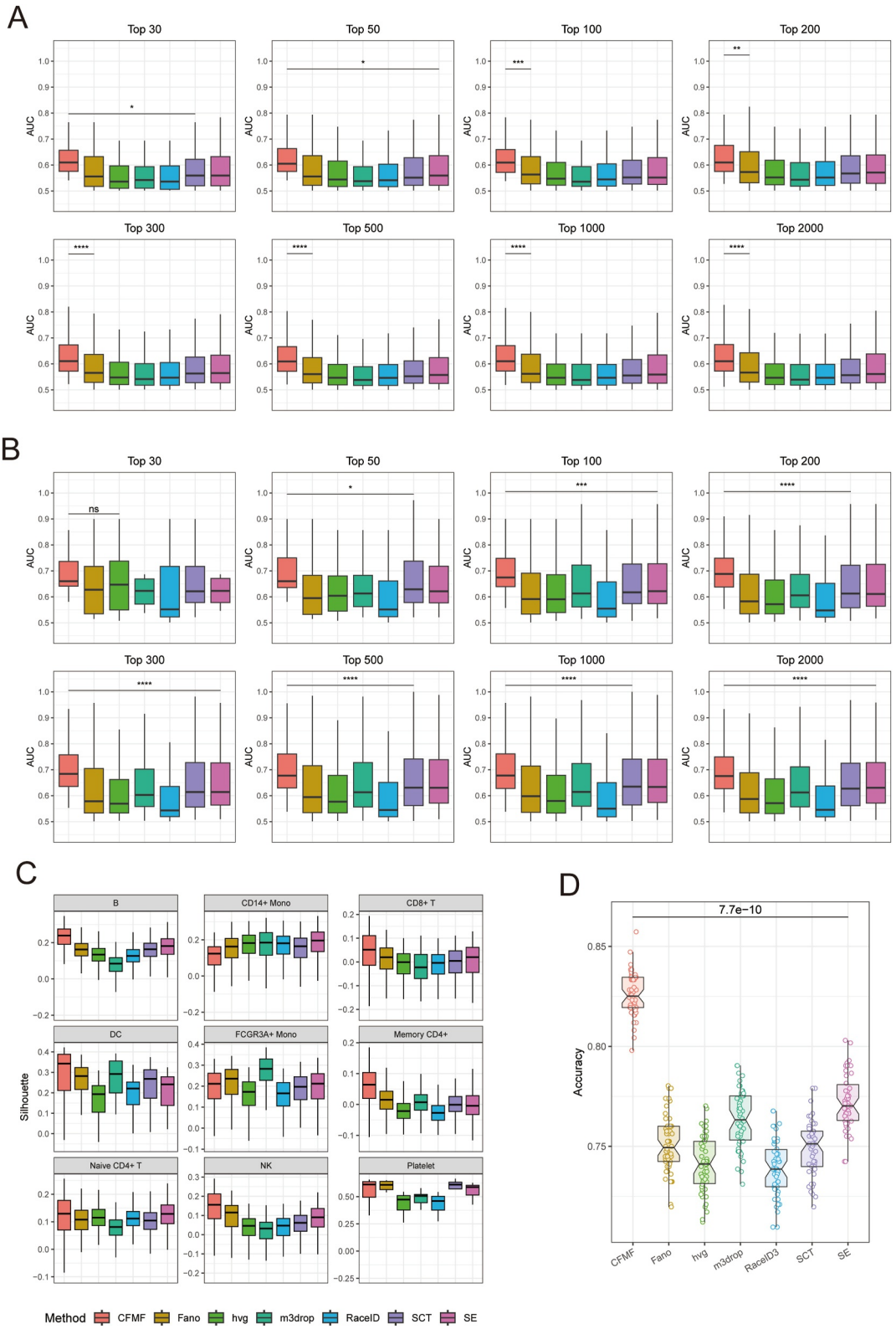

FIGURE 5 | Comparison of CFMF with established methods. (A, B) Boxplots of AUC values for genes selected by different methods across a range of number of genes (top 30-2000) in the PBMC3K dataset (A) and a representative GBM sample (B). (C) Boxplots of silhouette coefficients of different cell types calculated by different methods in the PBMC3K dataset. (D) Boxplot illustrating SingleR annotation accuracy across cell types using genes identified by various methods in the PBMC3K dataset across 50 iterations. *p*-values were calculated by paired two-sided wilcoxon rank sum test. Significance: $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$.

validation strategies. First, we assessed the clustering performance of genes identified by different methods, using silhouette coefficients [30] as the quantitative metric. Specifically, these genes were substituted for HVGs in the standard Seurat clustering workflow. Strikingly, CFMF-selected genes yielded higher silhouette scores across multiple cell types in the PBMC3K dataset compared with those obtained from alternative approaches (Figure 5C), indicating that CFMF-selected genes provide enhanced discrimination among distinct cell types.

Second, we investigated whether marker genes identified by CFMF exhibited superior performance in automatic cell type annotation. We randomly sampled 70% of the cells from the PBMC3K dataset to identify marker genes using various methods. These marker genes were then used to predict the cell identities of the remaining 30% of the cells using the SingleR annotation framework [28]. Notably, CFMF-selected genes achieved significantly higher classification accuracy than those identified by alternative approaches (Figure 5D). Collectively, these results demonstrate that CFMF-selected genes encompass a larger proportion of biologically informative genes reflecting cell type-specific signatures than conventional methods, thereby facilitating more precise discrimination of distinct cell types.

4 | Discussion

In this study, we introduce CFMF, a clustering-independent computational framework designed to identify candidate marker genes from scRNA-seq datasets. The method is built on the assumption that if a gene is a marker gene of a specific cell type, then cells expressing this gene should exhibit relatively higher transcriptomic similarity. To evaluate its performance, we applied CFMF across multiple independent datasets and consistently observed robust results. For instance, in the PBMC3K dataset, CFMF recovered well-established canonical marker genes such as *CD14* for classical monocytes and *LEF1* for $CD4^+$ naïve T cells, while also uncovering novel but highly cell type-specific marker genes, such as *AKR1C3* for NK cells. Similarly, in the GBM dataset, CFMF successfully identified previously reported GBM cellular state signature genes, including *MLC1* for AC signature, and *TNR* for OPC signature. Based on this validation, we further examined the utility of CFMF across three distinct analytical contexts, thereby highlighting its versatility and broad applicability to diverse single-cell transcriptomic studies.

In our view, one important application of CFMF is to detect rare cell types. A persistent challenge in clustering-based cell classification lies in the selection of clustering parameters. If the resolution is too low, rare cells are merged with other populations; conversely, if the resolution is set too high, a single homogeneous cell type may be erroneously fragmented into multiple clusters. Both scenarios ultimately lead to misleading interpretations. Established tools for rare cell type identification, such as CellSIUS [51], RaceID [5], and GiniClust [52], largely rely on clustering or related computational strategies, and therefore remain susceptible to these parameter-dependent pitfalls. By contrast, CFMF adopts a fundamentally distinct paradigm. A notable strength of CFMF is its capacity to capture intra-cluster heterogeneity, thereby revealing the potential existence of previously unrecognised subtypes or rare cell types. For instance, in normal lung scRNA-seq datasets, CFMF successfully identified rare cell type marker genes such as *ASCL3* and *BSND* for ionocytes, providing molecular evidence for the presence of these rare cell types. Thus, unlike tools explicitly developed for rare cell type detection, the primary utility of CFMF lies in its auxiliary role in highlighting rare or atypical populations, particularly in cases where clustering resolution is insufficient to delineate subtle heterogeneity.

Large-scale integrative analysis of scRNA-seq datasets has become increasingly prevalent in recent research, yet one of its major challenges lies in effective batch correction [18, 53, 54]. In conventional workflows for batch effect correction, an essential step is the selection of appropriate genes for dimensionality reduction. Ideally, these genes should capture genuine biological variation rather than technical variability introduced by batch effects. CFMF was applied in each individual sample, which mitigates the effect of batch on gene expression. Consequently, the expression variation of CFMF-selected genes is more likely to reflect underlying biology rather than technical artefacts. Although we cannot completely exclude the possibility of residual technical variation, integration leveraging recurrent CFMF-selected genes demonstrates markedly greater robustness compared with HVG-based PCA. For example, when HVGs were used for PCA followed by Harmony integration, the $LGR5^+$ and $TOP2A^+$ transcriptional state was dispersed across clusters (Figure S7A). In contrast, substituting HVGs with CFMF-selected genes enabled robust integration and precise delineation of this biologically relevant transcriptional state, thereby underscoring the pivotal role of gene selection in PCA-based analyses.

Taken together, these findings demonstrate that CFMF-selected marker genes mitigate the impact of technical variation while retaining the capacity to capture biologically relevant heterogeneity, thus providing a powerful framework for resolving tumour transcriptional states and dissecting the complexity of cancer ecosystems.

The high sensitivity of CFMF entails higher computational costs compared to HVG-based methods (Table S3), owing to its exhaustive neighbourhood analysis. However, this could be mitigated by parallel processing capabilities. Since marker gene identification is typically a one-time step, we consider the increased computational expense a justified trade-off for the substantial gains of performance in marker gene identification.

5 | Conclusion

In summary, CFMF offers a versatile and robust framework for single-cell transcriptomic analysis, distinguished by four key advantages. First, as a clustering-independent method, it eliminates the reliance on resolution parameters and distributional assumptions, thereby preventing the misclassification of ambiguous cell states. Second, it exhibits superior sensitivity in capturing rare cell types that are often masked in conventional workflows. Third, CFMF effectively delineates continuous biological heterogeneity, particularly within complex malignant

ecosystems. Finally, by prioritising biologically recurrent features over technical artefacts, CFMF facilitates robust data integration across diverse samples. Collectively, these attributes position CFMF as a powerful tool for advancing the discovery of novel cell types and dissecting cellular diversity.

Author Contributions

S.Y. and K.L. conceived and designed the study. All authors contributed to writing, reviewing, and editing the manuscript and approved the manuscript.

Acknowledgements

The authors would like to thank all the public data sources involved in the present study.

Funding

The research is supported by National Natural Science Foundation of China (Grants 32370715, 82330108), Science Fund for Distinguished Oversea Young Scholars of Shandong Province (Grant 2023HWYQ-015), Taishan Young Scholar Program of Shandong Province (Grant tsqn202312020), and Cheeloo Young Scholar Program of Shandong University.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data analyzed during the current study are available in public repositories as specified in the Methods section, reference number [14–26]. This study did not generate any new raw data. *Code Availability Statement*: All R scripts used for data processing and figure generation are available on GitHub (https://github.com/ShuminYin/project_CFMF).

References

1. K. Hrovatin, L. Sikkema, V. A. Shitov, et al., "Considerations for Building and Using Integrated Single-Cell Atlases," *Nature Methods* 22, no. 1 (2025): 41–57, https://doi.org/10.1038/s41592-024-02532-y.

2. M. P. Snyder, S. Lin, A. Posgai, et al., "The Human Body at Cellular Resolution: The NIH Human Biomolecular Atlas Program," *Nature* 574, no. 7777 (2019): 187–192, https://doi.org/10.1038/s41586-019-1629-x.

3. K. J. Travaglini, A. N. Nabhan, L. Penland, et al., "A Molecular Cell Atlas of the Human Lung From Single-Cell RNA Sequencing," *Nature* 587, no. 7835 (2020): 619–625, https://doi.org/10.1038/s41586-020-2922-4.

4. X. P. Han, Z. Zhou, L. Fei, et al., "Construction of a Human Cell Landscape at Single-Cell Level," *Nature* 581, no. 7808 (2020): 303–309, https://doi.org/10.1038/s41586-020-2157-4.

5. D. Grün, A. Lyubimova, L. Kester, et al., "Single-Cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types," *Nature* 525, no. 7568 (2015): 251–255, https://doi.org/10.1038/nature14966.

6. M. D. Luecken and F. J. Theis, "Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial," *Molecular Systems Biology* 15, no. 6 (2019): e8746, https://doi.org/10.15252/msb.20188746.

7. T. Abdelaal, L. Michielsen, D. Cats, et al., "A Comparison of Automatic Cell Identification Methods for Single-Cell RNA Sequencing Data," *Genome Biology* 20, no. 1 (2019): 194, https://doi.org/10.1186/s13059-019-1795-z.

8. C. Hu, T. Li, Y. Xu, et al., "CellMarker 2.0: An Updated Database of Manually Curated Cell Markers in Human/Mouse and Web Tools Based on scRNA-Seq Data," *Nucleic Acids Research* 51, no. D1 (2023): D870–D876, https://doi.org/10.1093/nar/gkac947.

9. C. E. Meacham and S. J. Morrison, "Tumour Heterogeneity and Cancer Cell Plasticity," *Nature* 501, no. 7467 (2013): 328–337, https://doi.org/10.1038/nature12624.

10. I. Dagogo-Jack and A. T. Shaw, "Tumour Heterogeneity and Resistance to Cancer Therapies," *Nature Reviews Clinical Oncology* 15, no. 2 (2018): 81–94, https://doi.org/10.1038/nrclinonc.2017.166.

11. R. B. Patterson-Cross, A. J. Levine, and V. Menon, "Selecting Single Cell Clustering Parameter Values Using Subsampling-Based Robustness Metrics," *BMC Bioinformatics* 22, no. 1 (2021): 39, https://doi.org/10.1186/s12859-021-03957-4.

12. Y. Ouadah, E. R. Rojas, D. P. Riordan, S. Capostagno, C. S. Kuo, and M. A. Krasnow, "Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch," *Cell* 179, no. 2 (2019): 403–416.e423, https://doi.org/10.1016/j.cell.2019.09.010.

13. M. Roulis, A. Kaklamanos, M. Schernthanner, et al., "Paracrine Orchestration of Intestinal Tumorigenesis by a Mesenchymal Niche," *Nature* 580, no. 7804 (2020): 524–529, https://doi.org/10.1038/s41586-020-2166-3.

14. G. X. Y. Zheng, J. M. Terry, P. Belgrader, et al., "Massively Parallel Digital Transcriptional Profiling of Single Cells," *Nature Communications* 8, no. 1 (2017): 14049, https://doi.org/10.1038/ncomms14049.

15. C. Neftel, J. Laffy, M. G. Filbin, et al., "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma," *Cell* 178, no. 4 (2019): 835–849, https://doi.org/10.1016/j.cell.2019.06.024.

16. N. Abdelfattah, P. Kumar, C. Wang, et al., "Single-Cell Analysis of Human Glioma and Immune Cells Identifies S100A 4 as an Immunotherapy Target," *Single Cell Portal* 13, no. 1 (2022): 767, https://doi.org/10.1038/s41467-022-28372-y.

17. S. Darmanis, S. A. Sloan, D. Croote, et al., "Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the M Igrating Front of Human Glioblastoma," *Cell Reports* 21, no. 5 (2017): 1399–1410, https://doi.org/10.1016/j.celrep.2017.10.030.

18. L. Sikkema, C. Ramírez-Suástegui, D. C. Strobl, et al., "An Integrated Cell Atlas of the Lung in Health and Disease," *Nature Medicine* 29, no. 6 (2023): 1563–1577, https://doi.org/10.1038/s41591-023-02327-2.

19. H.-O. Lee, Y. Hong, H. E. Etlioglu, et al., "Lineage-Dependent Gene Expression Programs Influence the Immune Landscape of Colorectal Cancer," *Nature Genetics* 52, no. 6 (2020): 594–603, https://doi.org/10.1038/s41588-020-0636-z.

20. K. Pelka, M. Hofree, J. H. Chen, et al., "Spatially Organized Multicellular Immune Hubs in Human Colorectal Cancer," *Cell* 184, no. 18 (2021): 4734–4752, https://doi.org/10.1016/j.cell.2021.08.003.

21. J. Qian, S. Olbrecht, B. Boeckx, et al., "A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling," *Cell Research* 30, no. 9 (2020): 745–762, https://doi.org/10.1038/s41422-020-0355-0.

22. F. Uhlitz, P. Bischoff, S. Peidli, et al., "Mitogen-Activated Protein Kinase Activity Drives Cell Trajectories in Colorectal Cancer," *EMBO Molecular Medicine* 13, no. 10 (2021): e14123, https://doi.org/10.15252/emmm.202114123.

23. J. R. Fleischer, A. M. Schmitt, G. Haas, et al., "Molecular Differences of Angiogenic Versus Vessel Co-Opting Colorectal Cancer Liver Metastases at Single-Cell Resolution," *Molecular Cancer* 22, no. 1 (2023): 17, https://doi.org/10.1186/s12943-023-01713-1.

24. L.-H. Che, J. W. Liu, J. P. Huo, et al., "A Single-Cell Atlas of Liver Metastases of Colorectal Cancer Reveals Reprogramming of the Tumor Microenvironment in Response to Preoperative Chemotherapy," *Cell Discovery* 7, no. 1 (2021): 80, https://doi.org/10.1038/s41421-021-00312-y.

25. L. Wang, H. Babikir, S. Müller, et al., "The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation," *Cancer Discovery* 9, no. 12 (2019): 1708–1719, https://doi.org/10.1158/2159-8290.CD-19-0329.

26. C.-K. Mo, J. Liu, S. Chen, et al., "Tumour Evolution and Microenvironment Interactions in 2D and 3D Space," *Nature* 634, no. 8036 (2024): 1178–1186, https://doi.org/10.1038/s41586-024-08087-4.

27. Y. H. Hao, S. Hao, E. Andersen-Nissen, et al., "Integrated Analysis of Multimodal Single-Cell Data," *Cell* 184, no. 13 (2021): 3573–3587, https://doi.org/10.1016/j.cell.2021.04.048.

28. D. Aran, A. P. Looney, L. Liu, et al., "Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage," *Nature Immunology* 20, no. 2 (2019): 163–172, https://doi.org/10.1038/s41590-018-0276-y.

29. A. P. Patel, I. Tirosh, J. J. Trombetta, et al., "Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma," *Science* 344, no. 6190 (2014): 1396–1401, https://doi.org/10.1126/science.1254257.

30. L. Lovmar, A. Ahlford, M. Jonsson, and A. C. Syvänen, "Silhouette Scores for Assessment of SNP Genotype Clusters," *BMC Genomics* 6, no. 1 (2005): 35, https://doi.org/10.1186/1471-2164-6-35.

31. A. T. L. Lun, D. J. McCarthy, and J. C. Marioni, "A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data With Bioconductor," *F1000Research* 5 (2016): 2122, https://doi.org/10.12688/f1000research.9501.2.

32. S. Aibar, C. B. González-Blas, T. Moerman, et al., "SCENIC: Single-Cell Regulatory Network Inference and Clustering," *Nature Methods* 14, no. 11 (2017): 1083–1086, https://doi.org/10.1038/Nmeth.4463.

33. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. Mesirov, and P. Tamayo, "The Molecular Signatures Database Hallmark Gene Set Collection," *Cell Systems* 1, no. 6 (2015): 417–425, https://doi.org/10.1016/j.cels.2015.12.004.

34. I. Vastrik, P. D'Eustachio, E. Schmidt, et al., "Reactome: A Knowledge Base of Biologic Pathways and Processes," *Genome Biology* 8, no. 3 (2007): R39, https://doi.org/10.1186/gb-2007-8-3-r39.

35. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, et al., "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update," *Nucleic Acids Research* 44, no. W1 (2016): W90–W97, https://doi.org/10.1093/nar/gkw377.

36. L. Zhang, X. Yu, L. Zheng, et al., "Lineage Tracking Reveals Dynamic Relationships of T Cells in Colorectal Cancer," *Nature* 564, no. 7735 (2018): 268–272, https://doi.org/10.1038/s41586-018-0694-x.

37. K. Yoshihara, M. Shahmoradgoli, E. Martínez, et al., "Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data," *Nature Communications* 4, no. 1 (2013): 2612, https://doi.org/10.1038/ncomms3612.

38. D. M. Cable, E. Murray, L. S. Zou, et al., "Robust Decomposition of Cell Type Mixtures in Spatial Transcriptomics," *Nature Biotechnology* 40, no. 4 (2022): 517–526, https://doi.org/10.1038/s41587-021-00830-w.

39. C. Neftel, J. Laffy, M. G. Filbin, et al., "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma," *Cell* 178, no. 4 (2019): 835–849.e21, https://doi.org/10.1016/j.cell.2019.06.024.

40. A. Nguyen, W. H. Khoo, I. Moran, P. I. Croucher, and T. G. Phan, "Single Cell RNA Sequencing of Rare Immune Cell Populations," *Frontiers in Immunology* 9 (2018): 1553, https://doi.org/10.3389/fimmu.2018.01553.

41. H. J. Wu, Y. Kirita, E. L. Donnelly, and B. D. Humphreys, "Advantages of Single-Nucleus Over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis," *Journal of the American Society of Nephrology* 30, no. 1 (2019): 23–32, https://doi.org/10.1681/Asn.2018090912.

42. A. Moorman, E. K. Benitez, F. Cambulli, et al., "Progressive Plasticity During Colorectal Cancer Metastasis," *Nature* 637, no. 8047 (2024): 947–954, https://doi.org/10.1038/s41586-024-08150-0.

43. F. de Sousa e Melo, A. V. Kurtova, J. M. Harnoss, et al., "A Distinct Role for Lgr5+ Stem Cells in Primary and Metastatic Colon Cancer," *Nature* 543, no. 7647 (2017): 676–680, https://doi.org/10.1038/nature21713.

44. M. Fujii and T. Sato, "Defining the Role of Lgr5 Stem Cells in Colorectal Cancer: From Basic Research to Clinical Applications," *Genome Medicine* 9, no. 1 (2017): 66, https://doi.org/10.1186/s13073-017-0460-y.

45. P. Brennecke, S. Anders, J. K. Kim, et al., "Accounting for Technical Noise in Single-Cell RNA-Seq Experiments," *Nature Methods* 10, no. 11 (2013): 1093–1095, https://doi.org/10.1038/nmeth.2645.

46. T. S. Andrews and M. Hemberg, "M3Drop: Dropout-Based Feature Selection for scRNASeq," *Bioinformatics* 35, no. 16 (2019): 2865–2867, https://doi.org/10.1093/bioinformatics/bty1044.

47. D. Grün, M. Muraro, J. C. Boisset, et al., "De Novo Prediction of Stem Cell Identity Using Single-Cell Transcriptome Data," *Cell Stem Cell* 19, no. 2 (2016): 266–277, https://doi.org/10.1016/j.stem.2016.05.010.

48. C. Hafemeister and R. Satija, "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression," *Genome Biology* 20, no. 1 (2019): 296, https://doi.org/10.1186/s13059-019-1874-1.

49. B. Liu, C. Li, Z. Li, D. Wang, X. Ren, and Z. Zhang, "An Entropy-Based Metric for Assessing the Purity of Single Cell Populations," *Nature Communications* 11, no. 1 (2020): 3155, https://doi.org/10.1038/s41467-020-16904-3.

50. D. DeTomaso and N. Yosef, "Hotspot Identifies Informative Gene Modules Across Modalities of Singl e-Cell Genomics," *Cell Systems* 12, no. 5 (2021): 446–456, https://doi.org/10.1016/j.cels.2021.04.005.

51. R. Wegmann, M. Neri, S. Schuierer, et al., "CellSIUS Provides Sensitive and Specific Detection of Rare Cell Populations From Complex Single-Cell RNA-Seq Data," *Genome Biology* 20, no. 1 (2019): 142, https://doi.org/10.1186/s13059-019-1739-7.

52. L. Jiang, H. D. Chen, L. Pinello, and G. C. Yuan, "GiniClust: Detecting Rare Cell Types From Single-Cell Gene Expression Data With Gini Index," *Genome Biology* 17, no. 1 (2016): 144, https://doi.org/10.1186/s13059-016-1010-4.

53. J. Q. Ma, Y. Wu, L. Ma, et al., "A Blueprint for Tumor-Infiltrating B Cells Across Human Cancers," *Science* 384, no. 6695 (2024): eadj4857, https://doi.org/10.1126/science.adj4857.

54. Y. Chu, E. Dai, Y. Li, et al., "Pan-Cancer T Cell Atlas Links a Cellular Stress Response State to Immunotherapy Resistance," *Nature Medicine* 29, no. 6 (2023): 1550–1562, https://doi.org/10.1038/s41591-023-02371-y.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Figure S1: Validating CFMF with public scRNA-seq datasets. (A) UMAP plots of PBMC3K dataset showing annotated cell types and canonical markers identified by CFMF. **(**B) UMAP plots of PBMC3K dataset showing novel markers identified by CFMF. (C) UMAP plots of the GBM dataset showing four GBM cellular state signature scores and expression of the corresponding genes identified by CFMF. **Figure S2:** CFMF fuels the identification of rare cell types. (A–D) UMAP plots of

four representative normal lung samples showing annotation of ionocytes and neuroendocrine cells (left), seurat clusters (middle) and marker expression (right). (A) UMAP and violin plots of per-cell total unique molecular identifier counts (nCount) and number of detected features (nFeature). (B) Heatmap of Reactome pathway activities across CRC transcriptional states. For each state, 1000 cells were randomly sampled. **Figure S3:** CRC transcriptional states detection and quality metrics. (A) UMAP and violin plots of per-cell total unique molecular identifier counts (nCount) and number of detected features (nFeature). (B) Heatmap of Reactome pathway activities across CRC transcriptional states. For each state, 1000 cells were randomly sampled. **Figure S4:** Molecular and cellular characteristics of CRC transcriptional states. (A) UMAP showing CFMF-identified genes of distinct CRC transcriptional states. (B) Dot plots of marker expression across CRC transcriptional states. (C) Scatterplots showing the Spearman correlation between the proportion of each transcriptional state among malignant cells and the proportion of $CD8^+$ T cells in the microenvironment. Spearman's correlations and *P*-values are shown. **Figure S5:** Comparison of CFMF with established methods in two GBM datasets. (A, B) Boxplots of AUC values for genes identified by different signature gene selection methods with different numbers of genes (top 30-2000) in Abdelfattah2022 (A) and the Wang2019 dataset (B). **Figure S6:** Comparison of CFMF with established methods in an additional GBM dataset and the PBMC3K dataset. (A) Boxplots of AUC values for genes identified by different signature gene selection methods with different numbers of genes (top 30-2000) in the Neftel2019 dataset(A) and the Wang2019 dataset (B). (C) Boxplots of spatial autocorrelation hotspot Z-scores for genes selected by different signature gene selection methods (top 2,000) in the PBMC3K dataset. **Figure S7:** CFMF improves tumor sample integration. (A) UMAPs showing expression of *LGR5* and *TOP2A* when dimensionality reduction is performed using HVGs versus CFMF-selected genes. **Table S1:** Overview of datasets used in this study. **Table S2:** Differentially expressed genes in the PBMC3K dataset. **Table S3:** Computational costs across PBMC datasets of different sizes.