

HemaScope: A Tool for Analyzing Single-cell and Spatial Transcriptomics Data of Hematopoietic Cells

Zhenyi Wang ^{1, #}, Yuxin Miao ^{2, #}, Hongjun Li ², Wenyan Cheng ¹, Minglei Shi ³,
Gang Lv ¹, Yating Zhu ⁴, Junyi Zhang ¹, Tingting Tan ¹, Jin Gu ², Michael Q. Zhang ⁵,
Jianfeng Li ^{1, *}, Hai Fang ^{1, *}, Zhu Chen ^{1, *}, Saijuan Chen ^{1, *, §}

¹Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Research Unit of Hematologic Malignancies Genomics and Translational Research of Chinese Academy of Medical Sciences, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

²MOE Key Laboratory of Bioinformatics, Beijing National Research Center for Information Science and Technology, Bioinformatics Division, Department of Automation, Tsinghua University, Beijing 100084, China

³Institute of Medical Innovation and Research, Peking University Third Hospital, Beijing 100191, China

⁴School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

⁵Department of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080-3021, USA

*Corresponding authors: lee_jianfeng@sjtu.edu.cn (Li J), fh12355@rjh.com.cn (Fang H), zchen@stn.sh.cn (Chen Z), sjchen@stn.sh.cn (Chen S).

#Equal contribution.

§Current address: Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

Handling Editor: Qian-Fei Wang

Abstract

Single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) techniques hold great value in evaluating the heterogeneity and spatial characteristics of hematopoietic cells within tissues. These two techniques are highly complementary, with scRNA-seq offering single-cell resolution and ST retaining spatial information. However, there is an urgent demand for well-organized and user-friendly toolkits capable of handling single-cell and spatial information. Here, we present HemaScope, a specialized bioinformatics toolkit featuring modular designs to analyze scRNA-seq and ST data generated from hematopoietic cells. It enables users to perform quality control, basic analysis, cell atlas construction, cellular heterogeneity exploration, and dynamical examination on scRNA-seq data. Also, it can perform spatial analysis and microenvironment analysis on ST data. Meanwhile, HemaScope takes into consideration hematopoietic cell-specific features, including lineage affiliation evaluation, cell cycle prediction, and marker gene collection. To enhance the user experience, we have deployed the toolkit in user-friendly forms: *HemaScopeR* (an R package), *HemaScopeCloud* (a web server), *HemaScopeDocker* (a Docker image), and *HemaScopeShiny* (a graphical interface). In case studies, we employed it to construct a cell atlas of human bone marrow, analyze age-related changes, and identify acute myeloid leukemia cells in mice. Moreover, we characterized the microenvironments in angioimmunoblastic T cell lymphoma and primary central nervous system lymphoma, elucidating tumor boundaries. HemaScope is freely available at https://zhenyiwangthu.github.io/HemaScope_Tutorial/.

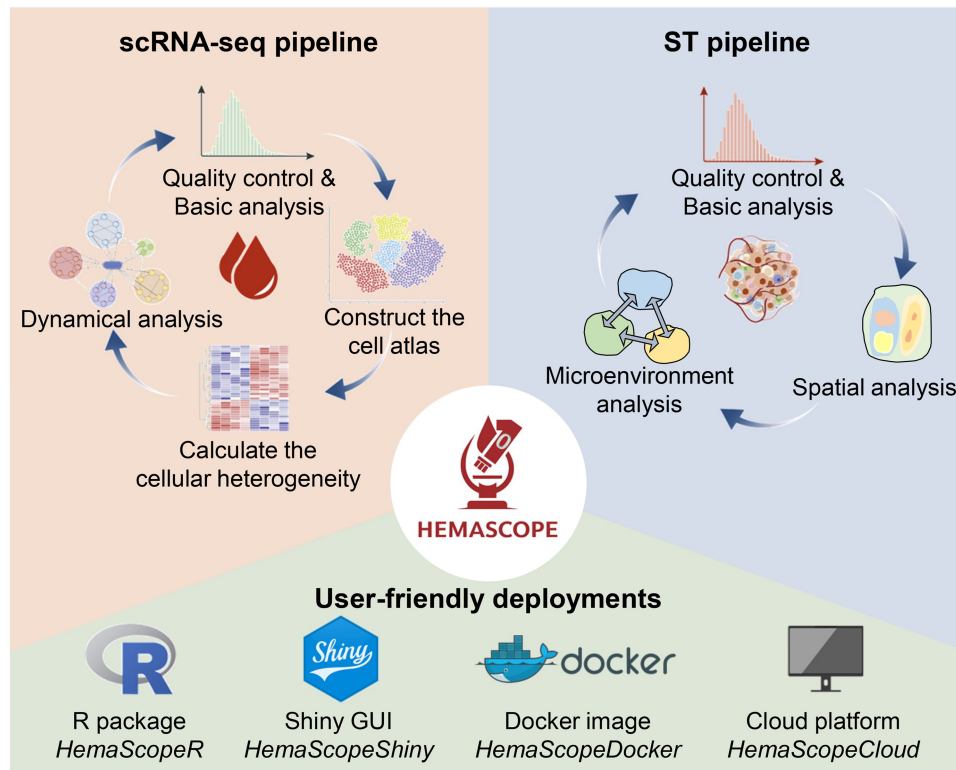
Key words: scRNA-seq; Spatial transcriptomics; R package; Web server; Shiny.

Received: 29 December 2023; Revised: 28 October 2024; Accepted: 18 November 2024.

© The Author(s) 2025. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Graphical abstract



Introduction

Single-cell RNA sequencing (scRNA-seq) [1] and spatial transcriptomics (ST) [2] techniques have emerged as pivotal methods for studying cellular heterogeneity and dynamic processes [3], as well as tissue spatial architecture. scRNA-seq offers high-throughput measurement of messenger RNA (mRNA) abundance in individual cells, allowing for gene expression profiling across numerous cells within a sample. Although ST can detect the transcriptome of specific spots with a defined resolution (e.g., 10X Visium, 55 μm in diameter, and 100 μm center-to-center distance between spots), it sacrifices single-cell resolution. However, it can record spatial locations for each spot, making scRNA-seq data and ST data complementary.

With the advancements in computational methods, there are over 1400 algorithms currently available to analyze scRNA-seq and ST data [4,5]. However, there are many pitfalls and recommendations that should be considered in data analysis practices [6]. These present challenges in choosing appropriate algorithms and designing effective data analysis pipelines. In previous studies, various research scope-specific scRNA-seq and ST analysis toolkits have been reported. For example, Guo et al. introduced scCancer [7], a specialized tool designed for processing scRNA-seq data in cancer research. Prieto et al. developed SingleCAnalyzer [8], a user-friendly cloud-based platform for comprehensive scRNA-seq analysis from FAST-Quality (FASTQ) files. SINCERA [9], developed by Guo et al., focused on processing scRNA-seq data from sorted cells or entire organs. Dries et al. introduced Giotto [10], an open-source toolbox for spatial data analysis, while Palla et al. presented Squidpy [11], a Python framework that merges omics and image analysis for scalable interpretation of spatial molecular data. Despite these popular

toolkits making data analysis in several research scopes more convenient and efficient, there remains a pressing need for toolkits specific to hematopoietic cells that also have user-friendly features adaptable to varying computational skill levels. We compared 20 milestone toolkits in detail with the one presented in this work in Table S1.

In this work, we developed a modular computational toolkit tailored for in-depth analysis of scRNA-seq and ST data from hematopoietic cells. HemaScope embodies the principles of portability, reproducibility, scalability, and user-friendliness. It encompasses quality control for scRNA-seq and ST data, and performs both unsupervised and supervised downstream analyses. For scRNA-seq data, it offers a suite of functionalities, including clustering, cell type identification, copy number variation (CNV) estimation, detection of differentially expressed genes (DEGs), cell cycle analysis, heterogeneity quantification, lineage differentiation potential assessment, cell trajectory prediction, transcription factor (TF) analysis, gene set variation analysis (GSVA), and cell-cell interaction analysis. For ST data, it includes detection of spatially variable features, deconvolution, coexistence analysis, and spatial interaction analysis. To enhance the user experience, we developed an open-source R package named “HemaScopeR” with an intuitive graphical interface named “HemaScopeShiny”, a web server named “HemaScopeCloud”, a Docker image named “HemaScopeDocker”, and comprehensive step-by-step tutorials.

Aside from the technical aspects above, HemaScope also takes special consideration of hematopoietic cell-specific features. First, HemaScope introduces comprehensive quality control metrics. Second, its downstream analysis targets critical issues in hematopoietic cell research, including cellular

heterogeneity, dynamics, hematopoietic hierarchy, and niches. Third, it introduces a lineage score to measure the affiliation levels of cells with various lineages within the hematopoietic hierarchy. Additionally, it introduces a strategy to estimate the cell cycle, and combines hematopoietic marker gene expression, data-driven prediction, and artificial intelligence-based methods to annotate hematopoietic cell types. Furthermore, it integrates the latest visual representation of the cellular hematopoietic hierarchy presented by Dai et al. (<https://github.com/NRCTM-bioinfo/HematoMap>), allowing researchers to visually observe hematopoiesis cellular composition and the cell blockage of acute leukemia in data through intuitive tree diagrams. By leveraging HemaScope, we analyzed five publicly available datasets, systematically demonstrating its efficacy across these analysis tasks.

Method

HemaScope overview

As illustrated in [Figure 1](#), the HemaScope toolkit consists of three modules. The first module covers the analytical workflow for scRNA-seq data, comprising four major components: data quality control and basic analysis, construction of a cell atlas, calculation of cellular heterogeneity, and examination of dynamic characteristics within cell populations. The second module is dedicated to the analytical pipeline for ST data, encompassing data quality control and basic analysis, spatial analysis, and microenvironment analysis. The third module includes four user-friendly features. We developed an R package with a graphical user interface that enables users to execute analyses either by coding command lines or by clicking buttons. Additionally, we distributed a Docker image to streamline the runtime environment setup and software dependencies. To accommodate the diverse needs of researchers, we also developed an online cloud platform. All bioinformatics tools integrated into HemaScope are listed in [Table S2](#).

Modular design 1: scRNA-seq pipeline

This pipeline accepts input from either the processed output files generated by 10X Genomics Cell Ranger, a Seurat [12] object, or a gene expression matrix in which rows denote genes and columns denote cells. It consists of four primary components.

The first part includes quality control and basic analysis

Initially, cells undergo filtering based on specific criteria: (1) cells expressing fewer than the specified “min.feature” number of genes, and (2) cells exceeding the “percent.mt.limit” for mitochondrial count percentage. Only genes detected in at least “min.cells” cells are retained. Subsequently, gene expression values are normalized, variable features are identified for principal component analysis (PCA), and DoubletFinder [13] is employed to identify doublets. Additionally, batch effects are mitigated using the FindIntegrationAnchors strategy from Seurat. We provided pitfalls and recommendations for setting cutoffs in [File S1](#).

The second part is dedicated to constructing the cell atlas

Cells are visualized in 2-dimensional (2D) space using dimensionality reduction methods, including PCA [14], Uniform Manifold Approximation and Projection (UMAP)

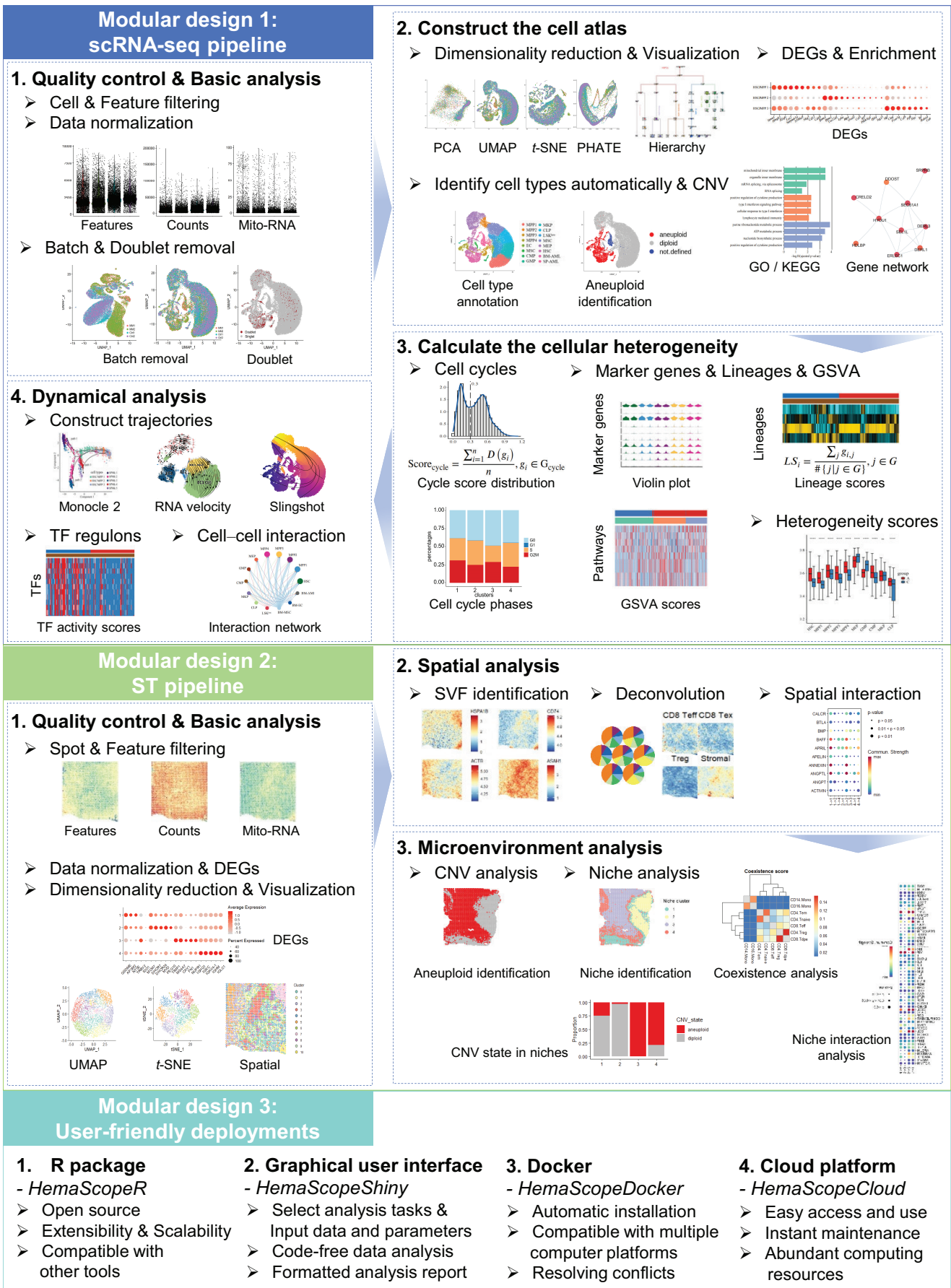
[15], *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) [16], Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) [17], and a tree-like structure by HematoMap (details are provided in [File S1](#) and [Figure S1](#)). Subsequently, the Louvain clustering algorithm [18] is employed for unsupervised cell cluster identification. Next, seven methods are integrated to improve the accuracy of cell type annotation (details are provided in [File S1](#)). For the convenience of readers, we referred to the literature [19–23] and online data resources [24], and provided well-organized marker gene lists for human and mouse hematopoietic cells in [Tables S3](#) and [S4](#). The Wilcoxon rank-sum test is then utilized to identify highly expressed gene sets within each cell type, followed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using clusterProfiler [25], as well as gene network analysis via OpenXGR [26] to explore the features of each cell type from different perspectives.

The third part focuses on calculating cellular heterogeneity

To quantify the heterogeneity of hematopoietic cells, we designed specific parameters and collected gene sets. An emerging theme in hematopoietic stem cell (HSC) and leukemia research is the equilibrium between cell proliferation and quiescence within the HSC pool through the cell cycle [27,28]. In HemaScope, we referred to the work of Dong et al. [20] to further introduce a parameter named cell cycle score ($Score_{cycle}$), combined with a data-driven machine learning method, scran [29], to classify single cells to G0, G1, S, and G2M phases (details are in [File S1](#) and [Table S5](#)). We also designed a parameter called lineage score (LS_i) which can quantify the affiliation levels of individual cells to various lineages within the hematopoietic hierarchy based on the gene expression profiles (details are in [File S1](#); [Tables S6](#) and [S7](#)). LS_i not only indicates the potential of normal hematopoietic stem/progenitor cells to differentiate into different lineages, but also detects cell blockage in leukemia. Then, cellular heterogeneity within cell groups is assessed by using the Spearman correlation coefficient, measuring differences among single cells. The Wilcoxon rank-sum test evaluates the statistical significance of average similarity differences between two sample groups. Next, expert annotated gene sets from the Molecular Signatures Database (MSigDB) and GSVA [30] are employed to explore the enriched biological processes in each cell.

The fourth part involves dynamical analysis

HemaScope is capable of analyzing cellular dynamics from three perspectives: cell trajectory, TF regulation, and cell–cell communication. Monocle 2 [31] is used to analyze cell differentiation trajectories, employing the minimum spanning tree approach. Slingshot [32] is utilized to investigate differentiation trajectories based on PCA. Additionally, scVelo [33] is employed to analyze cell differentiation trajectories based on RNA velocity. The results obtained from these three analytical methods, each grounded in distinct model assumptions, provide complementary insights. We compared these three methods in terms of algorithmic principles in [File S1](#). Following the aforementioned analyses, SCENIC [34] is used to compute activated TF regulons in each cell, followed by CellChat [35], which assesses the cross-talk among different cell populations.



Modular design 2: ST pipeline

The ST pipeline in HemaScope is capable of processing spatial transcriptomes with various resolutions generated by diverse spatial molecular technologies, including 10X Visium, multiplexed error-robust fluorescence *in situ* hybridization (MERFISH), and spatial enhanced resolution omics-sequencing (Stereo-seq). This pipeline comprises three primary parts.

The first part comprises quality control and basic analysis

Upon data loading, basic quality control measures are implemented. Cells with low unique molecular identifier (UMI) counts or gene numbers, as well as genes with minimal occurrence, are filtered out. Subsequently, Seurat [36] is employed to normalize, scale (or apply SCTransform), and perform PCA on the refined dataset, resulting in a low-dimensional representation defined by the first 10 principal components (PCs) by default. These PCs are then used for shared-nearest-neighbor (SNN)-based clustering, as well as for generating UMAP and *t*-SNE visualizations. DEGs across all clusters are identified based on fold changes and adjusted *P* values, determined via the Wilcoxon rank-sum test with Bonferroni correction.

The second part engages in spatial analysis

HemaScope incorporates the detection of spatially variable features (SVFs), spatial interaction analysis, and deconvolution. SVFs are identified using the “FindSpatiallyVariableFeatures” function in Seurat, which employs Global Moran’s I. Communication analysis by optimal transport (COMMOT) [37] is then utilized to analyze intercellular spatial communication. By default, secreted signaling ligand–receptor pairs from the CellChatDB [35] database are utilized to infer spatial interaction strength. These significantly expressed ligand–receptor pairs within and between clusters are visualized using dot plots. In the deconvolution step, HemaScope infers the abundance distribution of various cell types from mixed signals in 10X Visium data using cell2location. The resulting cell type distribution across the slide serves as the basis for microenvironment analysis. Additionally, reference single-cell datasets are utilized to calculate the top 50 DEGs for each cell type as marker genes. These marker genes are used to score each cell type, complementing the results from cell2location (details are provided in File S1).

The third part focuses on microenvironment analysis

It starts with CNV analysis to pinpoint tumor areas in tissues, employing copy number karyotyping of aneuploid tumors (CopyKAT) [38] to distinguish tumor regions from normal ones within the microenvironment. Following CNV analysis, the coexistence of various cell types in the tissue is calculated based on deconvolution results. To assess the global coexistence of cell types, correlation or Wasserstein distance is used to calculate their distributional distance. The Pearson correlation [39] measures the correlation-based coexistence score, while the Wasserstein distance defines the distribution-based coexistence score for cell types (details are provided in File

S1). Next, diverse niches are identified using *K*-means clustering based on the absolute cellular abundances from deconvolution results. The cellular abundance of these niches and the proportion of tumor regions within each niche aid researchers in evaluating niche characteristics. Furthermore, COMMOT infers and visualizes interactions within and between niches.

Modular design 3: user-friendly deployments

As illustrated in Figure 2A, we developed four user-friendly features to enhance the user experience. First, *HemaScopeR* is a well-organized R package. This open-source package ensures high scalability and compatibility with diverse R packages. Second, *HemaScopeShiny* is a Graphical User Interface (GUI) based on Shiny [40]. As illustrated in Figure 2B, researchers can opt for the GUI mode when using *HemaScopeR*, enabling code-free data analysis. Third, for easier installation and deployment on local computers, we developed a Docker image, *HemaScopeDocker*, to automate software downloads, configure necessary toolkits, and ensure compatibility across various computing platforms. Fourth, we developed a cloud platform, *HemaScopeCloud* (Figure 2C), to facilitate researchers in conducting online data analyses. In Table S1, we provided a detailed comparison of our toolkit with 20 other milestones.

Public datasets used for demo

To demonstrate the practical application capabilities of HemaScope in both a healthy population and various hematological diseases, we utilized publicly available scRNA-seq and ST datasets as case studies in the following sections and File S1. These datasets include scRNA-seq data of bone marrow cells from 20 healthy adult human donors across a broad age range [41], 45 acute myeloid leukemia (AML) patients [42,43], and scRNA-seq data of bone marrow cells from mice with *Myc*-driven AML [44], as well as scRNA-seq and ST data from primary central nervous system lymphoma (PCNSL) tumor samples [45] and ST data from angioimmunoblastic T cell lymphoma (AITL) samples [46]. We provide detailed information about these datasets in Table S8.

Results

HemaScope generates a comprehensive single-cell atlas of human bone marrow across a broad age range

We analyzed scRNA-seq data obtained from bone marrow samples of 20 healthy donors aged between 24 and 84 years [41] using HemaScope. Figure 3A demonstrates the presence of 22 cell types, as well as cell counts and percentages, in the cell atlas generated by HemaScope. The statistical results are provided in Table S9. Figure S2A displays the distribution of gene numbers, read counts, and mitochondrial RNA proportion used in the quality control process, while the effective identification and removal of doublets are illustrated in Figure S2B using DoubletFinder within HemaScope.

Figure 1 Continued

scRNA-seq, single-cell RNA sequencing; PCA, principal component analysis; UMAP, Uniform Manifold Approximation and Projection; *t*-SNE, *t*-distributed Stochastic Neighbor Embedding; PHATE, Potential of Heat-diffusion for Affinity-based Trajectory Embedding; DEG, differentially expressed gene; CNV, copy number variation; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSVA, gene set variation analysis; TF, transcription factor; ST, spatial transcriptomics; SVF, spatially variable feature.

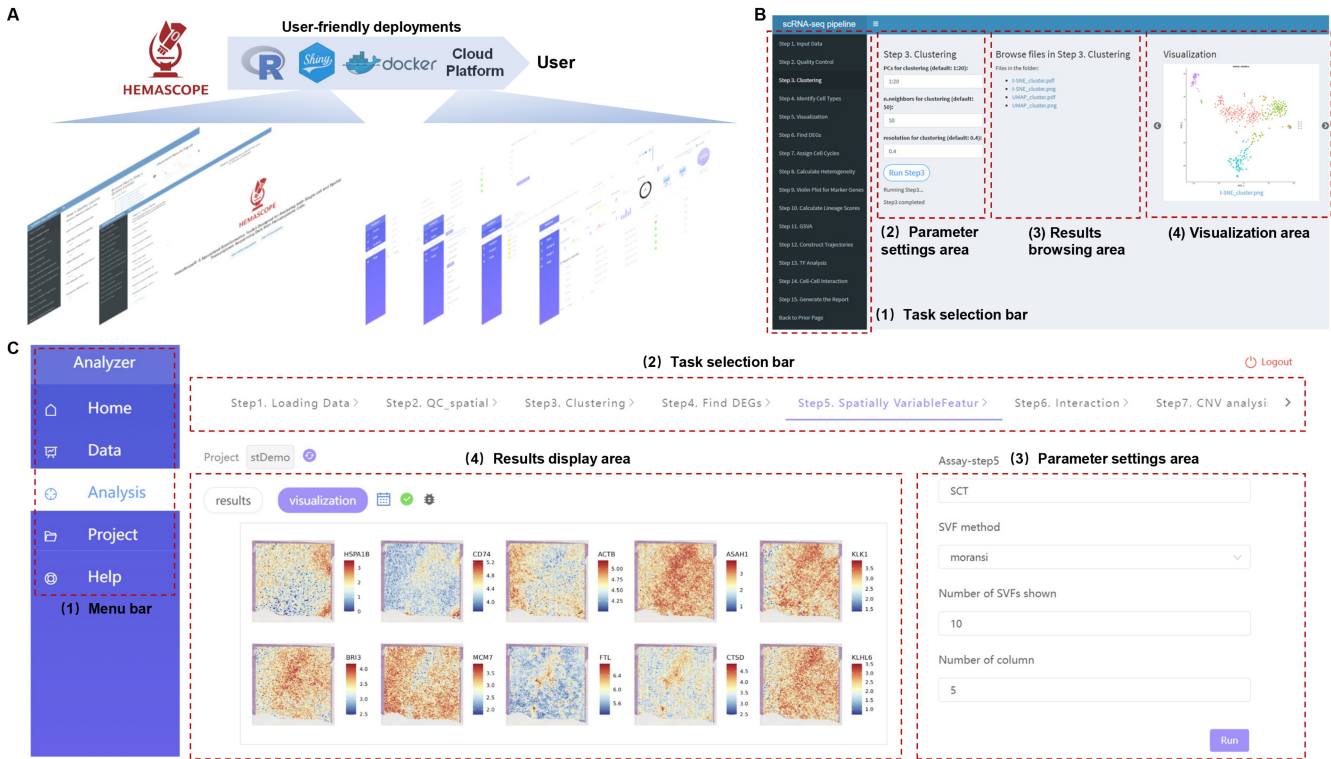


Figure 2 User-friendly deployments of HemaScope

A. HemaScope includes *HemaScopeR* (an R package), *HemaScopeShiny* (a Shiny GUI), *HemaScopeDocker* (a Docker image), and *HemaScopeCloud* (a cloud platform). **B.** A screenshot of *HemaScopeShiny* demonstrates features such as the task selection bar, the parameter settings area, the results browsing area, and the visualization area. **C.** A screenshot of *HemaScopeCloud* illustrates the menu bar, the task selection bar, the parameter settings area, and the results display area. GUI, Graphical User Interface.

Notably, no significant batch effects were observed among the 20 donors' data (Figure 3B and C, Figure S2C).

Identifying cell types in scRNA-seq data analysis is a challenging and open question. Inspired by the AdaBoost method [47] from the field of machine learning, we integrated seven methods to improve the accuracy of cell type prediction. These methods include profiling of hematopoietic cell-specific marker gene expression by various visualization strategies (Figure 3D, Figure S2D–F; marker genes are listed in Table S3), DEGs (Figure S2G; Table S10), GO enrichment, KEGG pathway enrichment, and gene network analysis (e.g., Figure S2H displays the enriched gene network of DEGs in CD10⁺ B cells, i.e., cluster 16 in Figure 3A), in addition to assessing ChatGPT [48] and label transfer (Figure S2I–O). We provided the technical details of the cell type annotation process in File S1. As depicted in Figure S2P, while there is considerable heterogeneity among donors, noticeable alterations in cell type proportions within the bone marrow were observed with increasing age. In particular, a significant increase in CD4⁺ naïve T cells and CD4⁺ memory T cells was evident [41], as shown in the stacked area plot. The computational details and statistical results on cell types, cell counts, and percentages are provided in File S1 and Table S9. Additionally, we employed HemaScope to dissect cellular heterogeneity and dynamics in human bone marrow cells, with further details provided in File S1 and Figure S3.

HemaScope analyzes the microenvironment in PCNSL from a spatial perspective

In this case, we used HemaScope to analyze the tumor microenvironment of four ST samples from PCNSL [45].

Analyzing the cellular composition and interactions in the microenvironment aids in understanding niche function [49]. Therefore, after basic quality control, deconvolution was conducted on the ST data using scRNA-seq data from HematoMap as a reference dataset. The “Hot” sample demonstrates a significantly higher infiltration level of T cells compared to the “Cold” sample (Figure 4A, Figure S4A and B), aligning with the inherent characteristics of each sample. Additionally, the non-tumor area of the invasive margin-excluded (IME) sample exhibits significant enrichment of CD14⁺ and CD16⁺ monocytes (Figure 4B), which is consistent with the macrophage definition in the original study [45].

Next, we categorized microenvironments by applying *K*-means clustering to the cellular abundance of these spatial data. The results highlighted the significance of the *CXCL12–CXCR4* pathway in “Hot” samples (Figure S4C and D), aligning with the findings of the original study [45]. In the IME sample, four microenvironment niches were identified (Figure 4C). Upon assessing the CNV results in the IME sample (Figure 4D), the analysis of spot CNV states across the four clusters indicated that clusters 1 and 2 were predominantly located in the non-tumor area, while clusters 3 and 4 were primarily found in the tumor area (Figure 4E). When examining the microenvironment cell composition within these clusters, clusters 1 and 2 exhibited an enrichment of monocytes and stromal cells, whereas clusters 3 and 4 displayed the opposite pattern (Figure S4E). Additionally, the upregulation of the *ANXA1–FPR1* signaling pathway in clusters 1 and 2 was consistent with the findings of the original study [45] (Figure S4F).

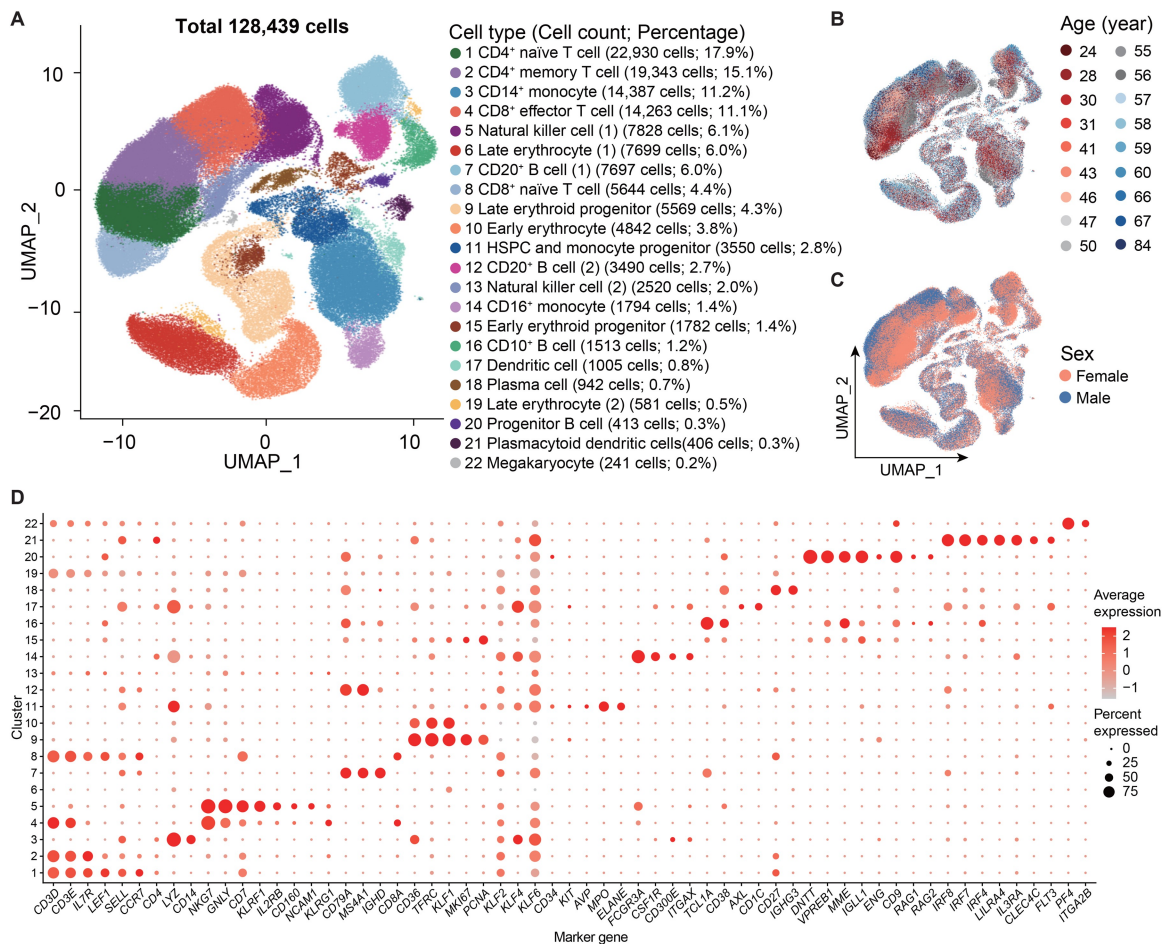


Figure 3 The single-cell atlas of human bone marrow across a broad age range

A. Visualization of cell clusters in a two-dimensional space using UMAP, annotated by cell types, cell counts, and percentages (rounded to one decimal place). **B.** Age labels of 20 donors (ranging from 24 to 84 years) are represented by distinct colors of points. **C.** Sex labels of 20 donors, with red and blue points indicating males and females, respectively. **D.** Dot plot showcasing the marker genes of various cell types in bone marrow. Larger dots indicate a higher proportion of cells expressing the gene, while a deeper red hue indicates increased average expression within the specific cell type.

Furthermore, we performed an in-depth analysis of the interaction between cluster 1 and cluster 4, as they encompassed both tumor and non-tumor regions (Figure 4E). Notably, these microenvironments were situated at the boundary between tumor and non-tumor regions in the IME sample (Figure 4C and D). At this boundary, there was extensive activation of the *SPP1* signaling pathway, especially the *SPP1*–*CD44* signal (Figure 4F). A significant interaction between *SPP1* in cluster 1 and *CD44* in cluster 4 was observed, consistent with the high *SPP1* and *CD44* expression at the tumor boundary (Figure 4G, Figure S4G and H). Gene expression patterns indicated *SPP1* concentration in the non-tumor area, specifically within the region previously identified as containing macrophages. In various cancers, tumor-associated macrophage (TAM) expressing *SPP1* (TAM-*SPP1*) facilitated immune suppression and tumor progression by interacting with *CD44* receptors on tumor cells [50–53]. This suggested potential enrichment of TAM-*SPP1* at the PCNSL boundary, contributing to an immune-suppressive microenvironment. Meanwhile, the activation of the *CXCL13*–*CXCR5* signaling axis was observed within and between clusters 1 and 4 (Figure 4F, Figure S4I), indicating an inflammatory response [54] at the tumor boundary. Utilizing the inflammation

features from CancerSEA [55], we scored the IME sample and revealed higher inflammation scores at the boundary (Figure 4H), corresponding to the activation of the *CXCL13*–*CXCR5* axis. Similar tumor boundary characteristics were also found in the invasive margin-immunosuppressed (IMS) sample (see File S1). In addition, we estimated spatial interactions between B cells and malignant T cells in an AITL sample using HemaScope (see File S1). In summary, we employed HemaScope to comprehensively delineate the characteristics of the tumor microenvironments in PCNSL and AITL.

Discussion

HemaScope is a user-friendly and modular design toolkit tailored for analyzing single-cell and ST data of hematopoietic cells. It automates the analysis workflow by utilizing scRNA-seq or ST matrices and spot coordinates as inputs, generating publication-quality figures, well-formatted tables, and organized analysis reports. Our open-source R package, *HemaScopeR*, is highly extensible and compatible with other toolkits. To enhance the user experience, we developed a GUI (*HemaScopeShiny*) for code-free data analysis, a Docker

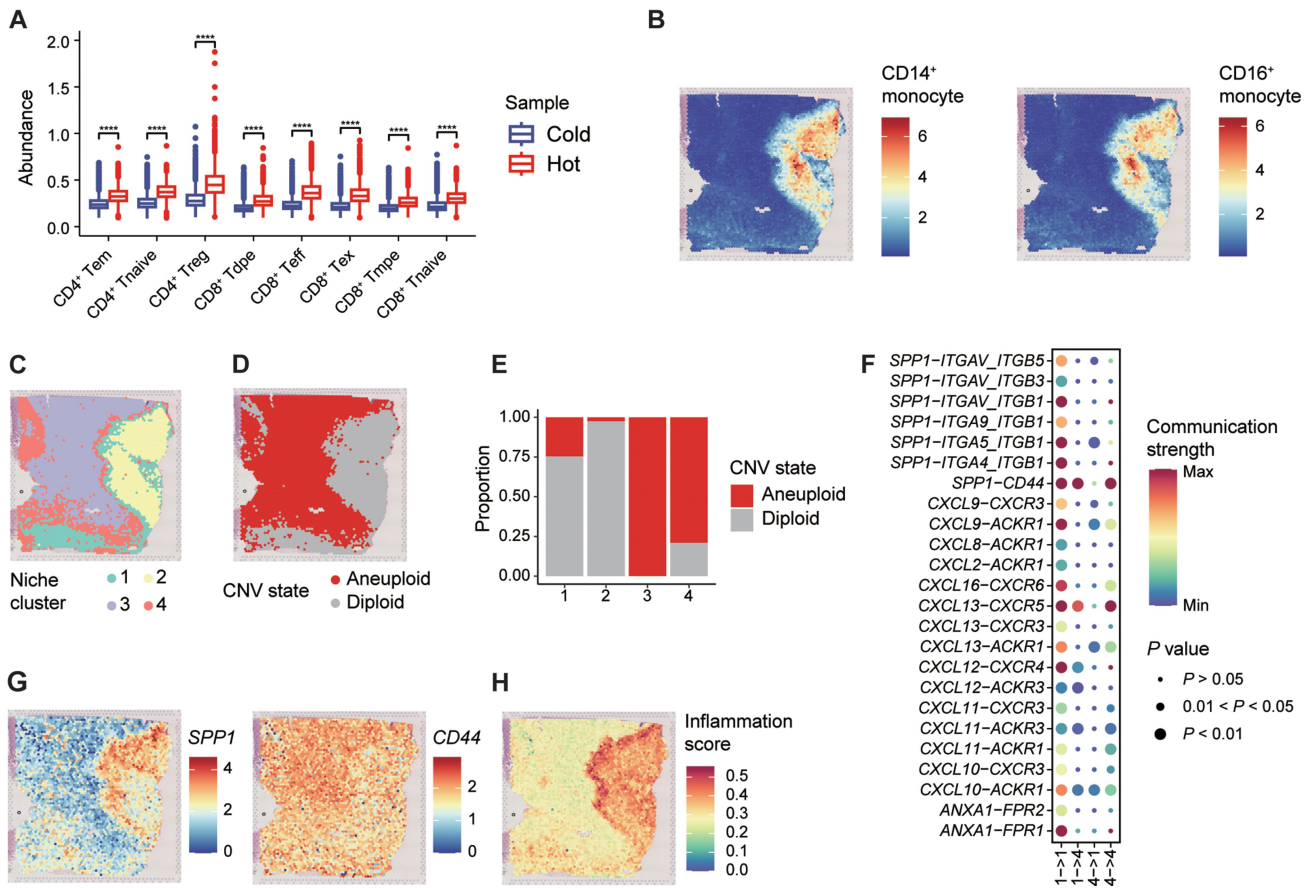


Figure 4 HemaScope analysis of the microenvironment in PCNSL

A. Box plot depicting T cell abundance in “Hot” and “Cold” samples, with P values calculated using the Wilcoxon rank-sum test (****, $P < 0.0001$). **B.** Spatial representation of CD14⁺ and CD16⁺ monocyte abundance in the IME. **C.** Spatial representation of niche clusters. **D.** Spatial representation of CNV states inferred via CopyKAT. **E.** Bar chart presenting proportions of various CNV states within each niche cluster. **F.** Bubble heatmap indicating the strength and statistical significance of interactions within and between niche clusters 1 and 4. **G.** Spatial distribution of expression patterns of *SPP1* and *CD44*. **H.** Spatial representation of the inflammation scores. CD4⁺ Tem, CD4⁺ effector memory T cell; CD4⁺ Tnaive, CD4⁺ naive T cell; CD4⁺ Treg, CD4⁺ regulatory T cell; CD8⁺ Tdpe, CD8⁺ *KLRG1*⁺ *IL7R*⁺ double-positive effector T cell; CD8⁺ Teff, CD8⁺ effector T cell; CD8⁺ Tex, CD8⁺ exhausted T cell; CD8⁺ Tmpe, CD8⁺ memory precursor effector T cell; CD8⁺ Tnaive, CD8⁺ naive T cell; PCNSL, primary central nervous system lymphoma; IME, invasive margin-excluded; CopyKAT, copy number karyotyping of aneuploid tumors.

image (*HemaScopeDocker*) for easy installation, and a web server (*HemaScopeCloud*) for online use.

The scRNA-seq pipeline consists of four primary components: quality control and basic analysis, construction of a cell atlas, evaluation of cellular heterogeneity, and dynamical analysis. First, we conducted quality control, dimensionality reduction, visualization, cell type identification, CNV detection, DEG analysis, GO/KEGG pathway enrichment, and gene network analysis to aid in building a comprehensive cell atlas. Next, we evaluated cellular heterogeneity and dynamics by predicting cell cycle phases (G0, G1, S, and G2M) and quantifying lineage affiliation. We then applied the GSVA algorithm to analyze cell enrichment in gene sets from the MSigDB and used Spearman correlation coefficients to assess intercellular heterogeneity. For the dynamical analysis, we employed Monocle 2, scVeloc, and Slingshot to predict cellular trajectories. Additionally, we used SCENIC to investigate TF regulons and CellChat to explore cell–cell interactions. HemaScope was utilized to construct a cell atlas, track age-related changes, and investigate cellular heterogeneity and dynamics in human bone marrow, as well as to identify

preleukemic and leukemic cells in mouse bone marrow. The effectiveness of this scRNA-seq pipeline is supported by literature validation.

Characterizing the tumor microenvironment from a spatial perspective is essential for understanding tumor progression within the ST pipeline. In addition to core analyses, HemaScope focuses on examining cellular composition, CNV status, and spatial interactions to provide a comprehensive view of the tumor microenvironment. Through the analysis of five spatial transcriptome samples from PCNSL and AITL, HemaScope not only validated the findings of the original study [45] but also offered deeper insights, particularly regarding the tumor boundary. The results suggested a significant accumulation of TAM-*SPP1* at the PCNSL tumor boundary, contributing to an immunosuppressive environment, along with evidence of a marked inflammatory response in this region.

In the future, there are several aspects where HemaScope can be further enhanced. First, single-cell multi-omics technologies [56] enable the simultaneous acquisition of transcriptomic, proteomic, and chromatin accessibility data at

single-cell resolution, highlighting the need for pipelines that can integrate these diverse data types. Second, as the resolution of ST continues to improve [57], toolkits must be developed to scale and adapt to varying resolutions. Third, the rise of spatial multi-omics technologies [58] creates a demand for pipelines capable of handling spatial multi-omics data. Additionally, with the continuous development of new algorithms and the expanding knowledge of hematopoietic cells, these advancements should be incorporated into HemaScope to further optimize the toolkit, ensuring both user-friendliness and flexibility (File S1).

Code availability

The open-source R package “HemaScopeR”, along with the GUI “HemaScopeShiny”, is available online at GitHub (<https://github.com/ZhenyiWangTHU/HemaScopeR/>). The code has also been submitted to BioCode at the National Genomics Data Center (NGDC), China National Center for Bioinformatics (CNCB) (BioCode: BT007725), which is publicly accessible at <https://ngdc.cncb.ac.cn/biocode/tool/7725>. The web server “HemaScopeCloud” is freely accessible at <https://hemascope.hiplot.cn/?home=hemascope>. The Docker image “HemaScopeDocker” can be obtained by pulling from Docker Hub. Detailed tutorials are available online at https://zhenyiwangthu.github.io/HemaScope_Tutorial/.

CRedit author statement

Zhenyi Wang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yuxin Miao:** Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Hongjun Li:** Software. **Wenyan Cheng:** Investigation, Validation, Writing – review & editing. **Minglei Shi:** Investigation. **Gang Lv:** Resources. **Yating Zhu:** Software. **Junyi Zhang:** Investigation. **Tingting Tan:** Software. **Jin Gu:** Methodology. **Michael Q. Zhang:** Methodology. **Jianfeng Li:** Resources, Software, Supervision. **Hai Fang:** Funding acquisition, Supervision, Validation, Writing – original draft, Writing – review & editing. **Zhu Chen:** Supervision, Writing – original draft, Writing – review & editing. **Saijuan Chen:** Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzaf002>).

Acknowledgments

This work was supported by the CAMS Innovation Fund for Medical Sciences (Grant No. CIFMS 2021-I2M-5-010), the China National Postdoctoral Program for Innovative Talents

(Grant No. BX20230229), the China Postdoctoral Science Foundation (Grant No. 2024M752038), the National Natural Science Foundation of China (Grant No. 32470681), and the Innovative Research Team of High-Level Local Universities in Shanghai. The authors would like to thank Prof. Chunyan Ji at the Department of Hematology, Qilu Hospital of Shandong University, Cheeloo College of Medicine, Shandong University, for providing the ST data of GSE181304 and GSE230207 for analysis. The authors would also like to thank Prof. Jian Hou at the Department of Hematology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, for providing the ST data of samples obtained from AITL. The authors would like to thank Xinlei Zhang, Shaoyong Peng, Mingming Su, Yixuan Huang, and Miao Zhou at Beijing ClouDNA Technology Co., Ltd. for their assistance in the development of the HemaScopeCloud platform. Additionally, the authors would like to thank Wei Yin, Yuting Dai, Shixiang Wang, Benben Miao, Wei Dong, and Xiaofeng Chen for their discussions regarding the development of HemaScope. Zhenyi Wang and Yuxin Miao used ChatGPT (v3.5) solely for enhancing readability and language.

ORCID

0000-0002-9407-0135 (Zhenyi Wang)
 0000-0003-1738-2466 (Yuxin Miao)
 0000-0002-8057-6186 (Hongjun Li)
 0000-0003-4347-5981 (Wenyan Cheng)
 0000-0002-1108-4233 (Minglei Shi)
 0000-0003-2573-8813 (Gang Lv)
 0009-0005-6937-1955 (Yating Zhu)
 0009-0001-0725-1000 (Junyi Zhang)
 0000-0001-6268-782X (Tingting Tan)
 0000-0003-3968-8036 (Jin Gu)
 0000-0002-7022-6115 (Michael Q. Zhang)
 0000-0003-2349-208X (Jianfeng Li)
 0000-0003-3961-8572 (Hai Fang)
 0000-0002-2968-7665 (Zhu Chen)
 0000-0003-3789-1284 (Saijuan Chen)

References

- [1] Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. *Nat Methods* 2020;17:11–4.
- [2] Asp M, Bergenstrahle J, Lundeberg J. Spatially resolved transcriptomes-next generation tools for tissue exploration. *Bioessays* 2020;42:e1900221.
- [3] Wang Z, Zhong Y, Ye Z, Zeng L, Chen Y, Shi M, et al. MarkovHC: Markov hierarchical clustering for the topological structure of high-dimensional single-cell omics data with transition pathway and critical point detection. *Nucleic Acids Res* 2022;50:46–56.
- [4] Heumos L, Schaar AC, Lance C, Litnetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;24:550–72.
- [5] Yuan Z, Zhao F, Lin S, Zhao Y, Yao J, Cui Y, et al. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nat Methods* 2024;21:712–22.
- [6] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;15:e8746.
- [7] Guo W, Wang D, Wang S, Shan Y, Liu C, Gu J. scCancer: a package for automated processing of single-cell RNA-seq data in cancer. *Brief Bioinform* 2021;22:bbaa127.

- [8] Prieto C, Barrios D, Villaverde A. SingleCAnalyzer: interactive analysis of single cell RNA-seq data on the cloud. *Front Bioinform* 2022;2:793309.
- [9] Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol* 2015;11:e1004575.
- [10] Dries R, Zhu Q, Dong R, Eng CL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;22:78.
- [11] Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* 2022;19:171–8.
- [12] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. *Cell* 2019;177:1888–902.e21.
- [13] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* 2019;8:329–37.e4.
- [14] Greenacre M, Groenen PJF, Hastie T, D’Enza AL, Markos A, Tuzhilina E. Principal component analysis. *Nat Rev Methods Primers* 2022;2:100.
- [15] Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;37:38–44.
- [16] Laurens VDM, Hinton G. Visualizing data using *t*-SNE. *J Mach Learn Res* 2008;9:2579–605.
- [17] Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;37:1482–92.
- [18] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;36:411–20.
- [19] Zhang X, Song B, Carlino MJ, Li G, Ferchen K, Chen M, et al. An immunophenotype-coupled transcriptomic atlas of human hematopoietic progenitors. *Nat Immunol* 2024;25:703–15.
- [20] Dong F, Hao S, Zhang S, Zhu C, Cheng H, Yang Z, et al. Differentiation of transplanted haematopoietic stem cells tracked by single-cell transcriptomic analysis. *Nat Cell Biol* 2020;22:630–9.
- [21] Giladi A, Paul F, Herzog Y, Lubling Y, Weiner A, Yofe I, et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol* 2018;20:836–46.
- [22] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. *Elife* 2017;6:e27041.
- [23] Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 2015;163:1663–77.
- [24] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, 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 Res* 2023;51:D870–6.
- [25] Yu G, Wang L, Han Y, He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16:284–7.
- [26] Bao C, Wang S, Jiang L, Fang Z, Zou K, Lin J, et al. OpenXGR: a web-server update for genomic summary data interpretation. *Nucleic Acids Res* 2023;51:W387–96.
- [27] Jude CD, Gaudet JJ, Speck NA, Ernst P. Leukemia and hematopoietic stem cells: balancing proliferation and quiescence. *Cell Cycle* 2008;7:586–91.
- [28] Ling VY, Straube J, Godfrey W, Haldar R, Janardhanan Y, Cooper L, et al. Targeting cell cycle and apoptosis to overcome chemotherapy resistance in acute myeloid leukemia. *Leukemia* 2023;37:143–53.
- [29] Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 2016;5:2122.
- [30] Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;14:7.
- [31] Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* 2017;14:309–15.
- [32] Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;19:477.
- [33] Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;38:1408–14.
- [34] Aibar G, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;14:1083–6.
- [35] Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell–cell communication using CellChat. *Nat Commun* 2021;12:1088.
- [36] Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293–304.
- [37] Cang Z, Zhao Y, Almet AA, Stabell A, Ramos R, Plikus MV, et al. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nat Methods* 2023;20:218–28.
- [38] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* 2021;39:599–608.
- [39] Fawcner-Corbett D, Antanaviciute A, Parikh K, Jagielowicz M, Geros AS, Gupta T, et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* 2021;184:810–26.e23.
- [40] Croci O, Campaner S. ChroKit: a Shiny-based framework for interactive analysis, visualization and integration of genomic data. *Nucleic Acids Res* 2023;51:W83–92.
- [41] Oetjen KA, Lindblad KE, Goswami M, Gui G, Dagur PK, Lai C, et al. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* 2018;3:e124928.
- [42] Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun* 2019;10:3660.
- [43] Wu J, Xiao Y, Sun J, Sun H, Chen H, Zhu Y, et al. A single-cell survey of cellular hierarchy in acute myeloid leukemia. *J Hematol Oncol* 2020;13:128.
- [44] Wu B, Chen X, Pan X, Deng X, Li S, Wang Z, et al. Single-cell transcriptome analyses reveal critical roles of RNA splicing during leukemia progression. *PLoS Biol* 2023;21:e3002088.
- [45] Xia Y, Sun T, Li G, Li M, Wang D, Su X, et al. Spatial single cell analysis of tumor microenvironment remodeling pattern in primary central nervous system lymphoma. *Leukemia* 2023;37:1499–510.
- [46] Du J, Qiu C, Li WS, Wang B, Han XL, Lin SW, et al. Spatial transcriptomics analysis reveals that CCL17 and CCL22 are robust indicators of a suppressive immune environment in angioimmunoblastic T cell lymphoma (AITL). *Front Biosci (Landmark Ed)* 2022;27:270.
- [47] Zhu J, Zou H, Rosset S, Hastie T. Multi-class AdaBoost. *Stat Interface* 2009;2:349–60.
- [48] Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;21:1462–65.
- [49] Li R, Xue K, Li J. FGF13 suppresses acute myeloid leukemia by regulating bone marrow niches. *Front Med* 2022;16:896–908.
- [50] He H, Chen S, Fan Z, Dong Y, Wang Y, Li S, et al. Multi-dimensional single-cell characterization revealed suppressive immune microenvironment in AFP-positive hepatocellular carcinoma. *Cell Discov* 2023;9:60.
- [51] Matsubara E, Yano H, Pan C, Komohara Y, Fujiwara Y, Zhao S, et al. The significance of SPP1 in lung cancers and its impact as a

- marker for protumor tumor-associated macrophages. *Cancers (Basel)* 2023;15:2250.
- [52] He C, Sheng L, Pan D, Jiang S, Ding L, Ma X, et al. Single-cell transcriptomic analysis revealed a critical role of SPP1/CD44-mediated crosstalk between macrophages and cancer cells in glioma. *Front Cell Dev Biol* 2021;9:779319.
- [53] Wei J, Chen Z, Hu M, He Z, Jiang D, Long J, et al. Characterizing intercellular communication of pan-cancer reveals SPP1⁺ tumor-associated macrophage expanded in hypoxia and promoting cancer malignancy through single-cell RNA-seq data. *Front Cell Dev Biol* 2021;9:749210.
- [54] Kazanietz MG, Durando M, Cooke M. CXCL13 and its receptor CXCR5 in cancer: inflammation, immune response, and beyond. *Front Endocrinol (Lausanne)* 2019;10:471.
- [55] Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res* 2019;47:D900–8.
- [56] Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;24:695–713.
- [57] Vickovic S, Eraslan G, Salmen F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat Methods* 2019;16:987–90.
- [58] Wu X, Xu W, Deng L, Li Y, Wang Z, Sun L, et al. Spatial multi-omics at subcellular resolution via high-throughput *in situ* pairwise sequencing. *Nat Biomed Eng* 2024;8:872–89.