BMC Genomics

SOFTWARE

Open Access

crocketa: an automated Snakemake framework for integrated single-cell transcriptome and immune-repertoire analysis

Gonzalo Soria-Alcaide[1,5], Gonzalo García-Aguilera[1], Marta Portasany-Rodríguez[1,5], Jaanam Lalchandani[1,5], África González-Murillo[1], Elena G. Sánchez[1], Cristina Saiz-Ladera[1], Manuel Ramírez-Orellana[2,1,3] and Jorge Garcia-Martinez[4*]

Abstract

crocketa (single-Cell Repertoire Organization & Combined Kinetics Exploration for Transcriptomic Analysis) is an automated and adaptable *Snakemake* pipeline designed to perform fundamental initial stages of single-cell analysis for both transcriptomic and immune repertoire data, with additional steps for detailed assay characterization. *crocketa* stems from the imperative need to devise an optimal methodology for integrating and analyzing single-cell RNA sequencing (scRNA-seq) alongside single-cell T-cell Receptor (scTCR-seq) or B-Cell Receptor (scBCR-seq) data in an automated way by means of state-of-the-art tools. This pipeline encompasses basic steps from primary & secondary scRNA-seq along with the incorporation and analysis of immune repertoire, capable of accommodating both human and mouse datasets. *crocketa* addresses this need by providing a reproducible and automated framework for integrated analysis of single-cell transcriptomic and immune-repertoire data.

Keywords Pipeline, scRNAseq, TCR, BCR, Immune repertoire, Clonotype, Immunology

Introduction & background

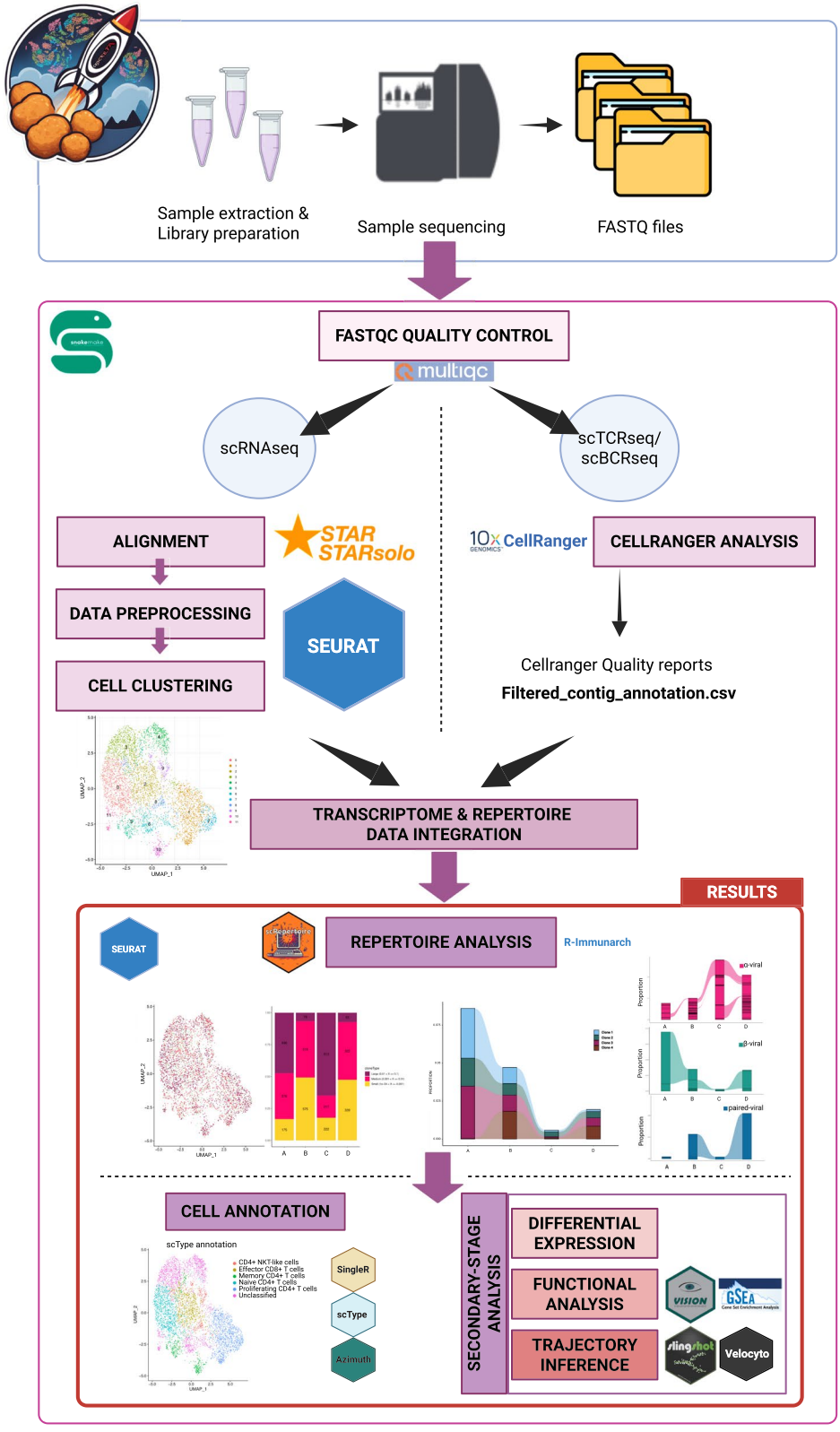
Lymphocytes are central players in the adaptive immune system, and while they have been widely explored from a transcriptional perspective, the detailed coupling of these profiles with lymphocyte clonality has only recently become a focal point of high-resolution immunological research. The immune repertoire comprises a vast diversity of B-cell receptors (BCRs) and T-cell receptors (TCRs). BCRs and TCRs play a central role in the immune system's recognition of pathogens as well as antigens, and their analysis can provide valuable insights into areas such as treatment efficacy in cancer and other diseases [1–3]. An immense clonal diversity arises from *V(D)J* recombination, which results in the hypervariable antigen-binding complementarity-determining region 3 (CDR3). Diversity is further amplified by certain genetic mechanisms such as imprecise gene segment junctions or somatic hypermutation in the case of B-cells. The outcome is a potential diversity of over 10^{20} unique TCR immune receptor sequences and 10^{18} unique BCR sequences [4, 5]. The collection of TCRs & BCRs within a specific individual constitutes its immune repertoire. The immune repertoire reflects both the immune history and

*Correspondence:
Jorge Garcia-Martinez
xurde.garcia.martinez@gmail.com

Full list of author information is available at the end of the article

Graphical Abstract

Schematic representation of *crocketa* workflow. Created in BioRender: https://BioRender.com/mvjsfdh



current status of the individual [6–9]. Adaptive immune responses to threats involve clonal expansion and selection of B-cells and T-cells [10]. The repertoire is highly dynamic and adaptive, and its study sheds light on the disease prognosis and therapeutic outcomes [11]. CDR3 constitutes the identifier of individual lymphocyte clonality: lymphocytes with identical CDR3 regions are defined as belonging to the same clone [12].

Recent years have seen rapid and unprecedented advancements in scRNA-seq technologies, both in technical capabilities and analytical methodologies. Continuous improvements in sensitivity, resolution and scalability have revolutionized transcriptomics, enabling a more precise exploration of cellular heterogeneity [13]. Concurrently, high throughput sequencing technologies have enabled the exponential development of methods for immune repertoire accession and analysis, leading to a significant increase in the demand for optimal approaches to large-scale clonal analysis [14–16]. These advancements now provide the opportunity to integrate both data sources, allowing for single-cell level insights and exploration of new research horizons. The integration of these two technologies facilitates the investigation of immune transcriptional heterogeneity throughout the disease process while accounting for clonality [17].

Recent years have seen significant progress in establishing benchmarking frameworks for core scRNA-seq analytical tasks, such as clustering and feature selection [18, 19]. However, while the field has moved toward more standardized practices for transcriptomic data itself, the integrated analysis of transcriptional states and immune repertoires remains hindered by methodological heterogeneity [20]. Current approaches to immune repertoire analysis often rely on fragmented pipelines, limiting seamless integration between transcriptional profiling and clonal characterization. While tools such as *Bollito* [21], *Platypus* [22], and *scRepertoire* [23] offer partial solutions, none provide a unified workflow that enables efficient integration of single-cell data while maintaining the flexibility to accommodate user-specific analytical strategies. Moreover, most existing tools lack state-of-the-art methodologies for the combined analysis of transcriptional and clonal dynamics.

Here, we introduce *crocketa* (single-Cell Repertoire Organization & Combined Kinetics Exploration for Transcriptomic Analysis), an automated workflow based on *Snakemake* [24] that integrates transcriptional and clonal analysis within a unified framework. By systematically combining established tools rather than introducing new algorithms, *crocketa* ensures reproducibility, scalability, and configurable analysis from raw sequencing data to integrated insights. Crucially, its ability to correlate transcriptional states with clonal expansion has direct implications for cancer immunotherapy, where identifying expanded clones associated with exhaustion or memory phenotypes could inform cellular therapy development. By providing this fully automated, integrative solution, *crocketa* addresses a critical gap in single-cell multi-omic analysis pipelines.

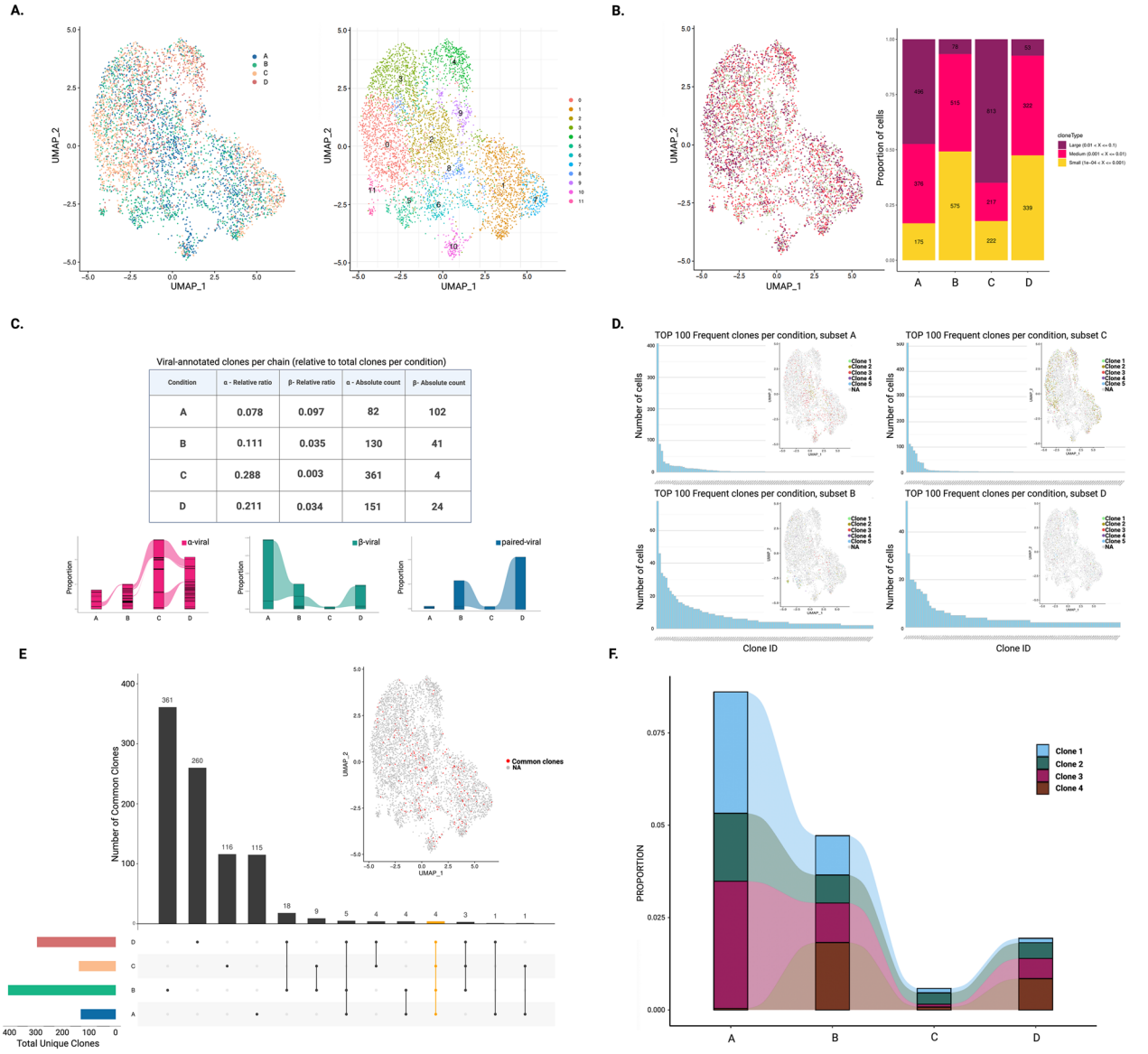
Results

The workflow outlined in the Methods section and illustrated in the Graphical Abstract yields a wide range of results.

crocketa outputs aggregate quality control reports from multiple stages of the processing in a *MultiQC* [25] report. Regarding single-cell analysis, a wide variety of pdf-formatted visualizations are generated for each stage including UMAP visualizations, reporting figures, clonality-descriptive plots and others. Additionally, each analysis stage performed in *Seurat* [26] provides its respective single-cell object in .rds format for downstream processing. Moreover, statistical results in column-formatted files are also created for further inspection. An in-depth and comprehensive description of the results is available at https://github.com/OncologyHNJ/crocketa, while key transcriptomic and clonal findings are presented below.

Repertoire analysis

Once transcriptomics data is properly processed up to the cell clustering step (Fig. 1A), clonal data is integrated and analyzed. Initially, clonal single-cell data is summarized in UMAP coordinates, where clonal cells are ranked according to certain relative or absolute user-defined clonal frequency ranges (default relative values: *Single* < = *10* − *4*< *Small* < = *10* − *3*< *Medium* < = *10* − *2*< *Large* < = *10* − *1* < *Hyperexpanded* < = *1*). Bar plots illustrating clonal diversity results, cell counts, relative proportions and unique number of clones, are generated (Fig. 1B*)*. In the case of TCRs, viral annotation is performed as detailed in the methods section, and results can be extracted and visualized for alpha-chain, beta-chain and paired-chain viral annotated TCRs. As observed in this dataset, viral annotation had a greater impact on conditions C and D than in conditions A and B, in terms of both absolute and relative counts (Fig. 1C). Following TCR viral annotation, the same analyses are applied to both the full and non-viral single-cell object, enabling users to either explore both approaches or select the one best suited for their needs. Some per-sample results can be generated to elucidate the most frequent clones, visualized in either bar plots or UMAP visualizations (Fig. 1D*).* Additionally, results concerning clonal overlap among samples or conditions of interest are reported. Venn diagrams and/or Upset-like plots are provided for one-to-one and all-to-all comparisons (Fig. 1E*)* in order to observe all possible combinations across conditions



Condition	α - Relative ratio	β- Relative ratio	α - Absolute count	β- Absolute count
A	0.078	0.097	82	102
B	0.111	0.035	130	41
C	0.288	0.003	361	4
D	0.211	0.034	151	24

Fig. 1 Immune repertoire analysis. **A** UMAP dimensionality plots colored by condition of interest [left] and cell cluster [right]. A-to-D define different co-culture conditions (see Methods). **B** Immune repertoire dataset. UMAP dimensionality plot colored by clonal frequency [left]; Bar plot showing the proportion and number of cells per clonal frequency range in each condition of interest [right]. **C** TCR viral annotation analysis. Table summarizing viral-annotated TCRs clones per chain, displayed as absolute counts and normalized to the total pool of clones per condition. Alluvial plots illustrate the distribution of relative viral-annotated clones across single-α, single- β and paired α/β categories. **D** Most frequent clones per condition through bar plots (Top 100) and UMAP dimensionality plots (Top 5). **E** Clonal overlapping across conditions of interest through Upset-like plot and UMAP dimensionality plot. **F** Frequency fluctuation of common clones across conditions of interest through alluvial plot. Created in *BioRender*. https://BioRender.com/6ahz4pb

simultaneously through *ggvenn* (https://cran.r-project.org/web/packages/ggvenn/readme/README.html) and *UpsetR* software [27]. Further visualizations are included to identify common clones to all conditions (e.g. UMAP plots, Fig. 1E), in addition to an alluvial plot illustrating frequency fluctuations across conditions for each individual sample/condition *(*Fig. 1F*)*. More results and updates will be provided on the online GitHub documentation.

Post-clustering analysis

Upon completion of the repertoire analysis, *crocketa* proceeds to the final stage of the workflow, which, while not essential, serves as an optional secondary step in the transcriptional analysis. Output for differential expression analysis, functional enrichment, and trajectory inference may be computed. *crocketa* includes an intermediate cell annotation step in order to elucidate cell identity. Several visualizations will be provided for each of the genes of interest specified through *Seurat*-based methods such

as *FeaturePlot* graphics (Fig. 2A). To explore functional enrichment, several methods might be applied to highlight transcriptional activation of biological pathways. According to *Vision* [28], the expression of *E2F*-targets (based on *Hallmarks*) seems to be enriched in clusters 1 and 7 (Fig. 2B), consistent with *MKI67* expression as a proliferation marker and with preliminary *scType* [29] annotation results, identifying these cells as proliferating CD4 + T cells (Fig. 2C*)*.

Automatic-annotation results are generated for each tool and reference applied at a cell or cluster level, with corresponding UMAP plots produced for each approach (Fig. 2C). Specifically, *scType* employs a customizable reference to annotate cell populations; these initial labels are subsequently refined through integration with *scANVI* [30] to resolve previously unclassified cells. While *scType* serves as the default mandatory annotation tool, additional software (detailed in the Methods section) can be optionally applied. Through the combination of all possible annotation results from the specified approaches, user can finally assign a robust label to each of the cell/clusters in the assay.

To further characterize the dynamic nature of the dataset, trajectory inference and RNA velocity analyses can be optionally performed to elucidate cell-state transitions. *Slingshot* [31] is utilized to model global lineages and pseudotime architectures based on cluster connectivity (Fig. 2D). In parallel, to capture the directional momentum of individual cells, RNA velocity can be estimated via *Velocyto* [32]. These velocity vectors are subsequently integrated into a hybrid *CellRank* [33] kernel (60% splicing, 40% gene expression), combining transcriptomic similarity with splicing dynamics to produce refined, high-confidence cell fate predictions (Fig. 2E).

This study presents descriptive results to illustrate the potential of the analysis pipeline, which may gain additional relevance when placed in the context of real-world use cases. Within the *crocketa* analysis framework, multiple visualizations linking T cell states to experimental conditions of interest are generated. As an illustrative example, distinct culture conditions are explored to assess their effect on T cell development and clonal expansion. Using *crocketa*, culture condition C seemed to be associated to a marked expansion of the naïve T cell compartment, which shows the highest representation

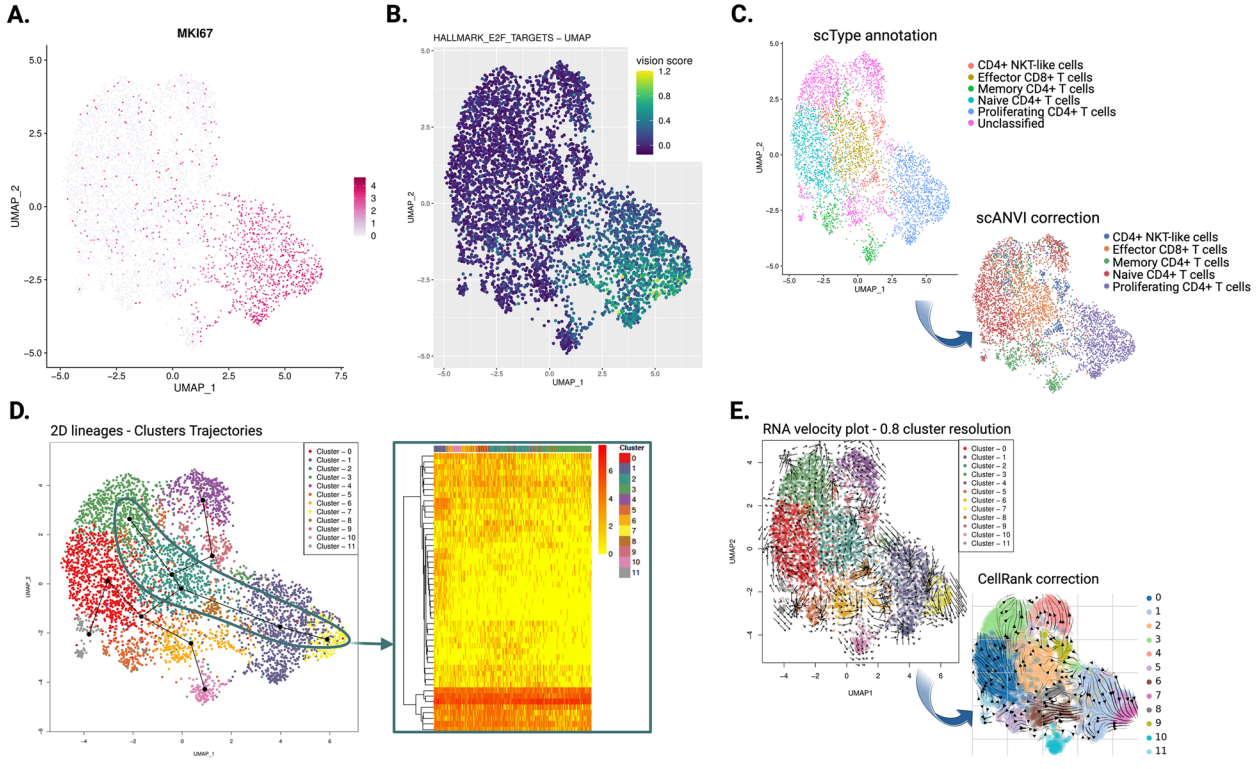


Fig. 2 Cell annotation. **A** Expression of *MKI67* gene, marker of proliferation. **B** *Vision* results for Hallmark_E2F_targets, pathway associated with cell proliferation. **C** Cell annotation results. Identification of cell types through the integration of *scType* and *scANVI*, showing successful resolution of previously unclassified labels across the dataset. **D** *Slingshot* trajectory inference. UMAP projection showing predicted lineages with cells colored by cluster [left]. Heatmap illustrating transcriptional dynamics along a specific trajectory (e.g. transition from cluster 7 to cluster 3) [right]. **E** RNA Velocity and cell fate prediction. UMAP projection showing individual cell future states projected via *Velocyto* alone or integrated with *CellRank* hybrid kernel. Cells colored by cluster. Created in *BioRender*. https://BioRender.com/nbb6yee

among the most frequent clonotypes under this condition. Overall, this analysis enables the association of specific T cell states with expanded clonotypes, thereby identifying conditions that preferentially promote the expansion of particular T cell states or TCRs of interest (*Supplementary Fig. 1)*. These results are intended as an illustrative application of the pipeline and do not represent conclusions derived from real-world experimental validation.

Benchmarking

To evaluate the computational efficiency and functional scope of *crocketa*, we conducted a comparative analysis against two established pipelines for immune repertoire integration: *Immunopipe* [34] and *Dandelion* [35]. A standardized public test dataset of ~ 500 cells (available in the *crocketa* repository) was processed on identical hardware to compare runtime, memory footprint, and workflow completeness.

crocketa stands out as the only benchmarked pipeline that provides a seamless, end-to-end automated workflow from raw FASTQ files to integrated analysis objects, eliminating the need for manual preprocessing. In contrast, both *Immunopipe* and *Dandelion* require external inputs (e.g. count matrices and V(D)J annotations) generated by third-party tools. Notably, for this benchmarking, *crocketa* had to to be executed in first place in order to generate *STAR* and *CellRanger* outputs required as input for the other tools.

A detailed, step-by-step breakdown of performance metrics for comparable analytical stages is provided in Supplementary Tables 1 and 2. Overall, *crocketa* demonstrated a trend toward lower memory usage and faster execution times across equivalent processing stages, such as cell annotation and clonotype integration when compared to *Immunopipe*. While *Dandelion* proved efficient performance for specific, isolated tasks, its nature as a modular toolkit (rather than a full automated pipeline) requires extensive manual scripting to achieve a comparable analytical scope. Consequently, its performance metrics are highly dependent on the user's specific implementation, complicating direct, standardized comparison with the strictly reproducible workflows of *crocketa* and *Immunopipe.*

Due to the nature of the test dataset (a minimal simulated subset of a public B cell reference) biological conclusions are inherently limited. Regarding cell recovery and QC, *crocketa* identified a total number of 579 cells through *STAR* alignment, retaining 514 high-quality cells after filtering. Similarly, when the same count matrix was provided to *Immunopipe* it rescued 525 cells with minimal parameter differences; however, its rigid T/B cell selection step led to a notable loss of information. Specifically, *Immunopipe* misclassified minor subpopulations as T cells (gamma-delta and double negative states) or platelets. While these classifications were likely artifacts of the feature-reduced dataset, the primary concern lies in *Immunopipe*'s subsequent automated discarding of the cells, preventing for further manual inspection or reclassification. In a real-case study, this lack of granular control and automated exclusion of "ambiguous" populations could result in the loss of biologically relevant information such as rare, intermediate cell states. In contrast, *crocketa* yielded a consistent B-cell characterization (primarily in a naïve state) through the previously described annotation workflow, with an unclassified portion of cells which were then cross-validated as B-cells via *SingleR*.

In terms of repertoire analysis, the integrated *CellRanger* module in *crocketa* successfully assigned clonotypes to 509 cells (99.02% recovery). When *Immunopipe* processed the same *CellRanger*-derived annotations, 399 cells (76% recovery) were identified as B-cells with an associated clonotype, direct result of its annotation filtering step.

Beyond this stage, direct comparison is not feasible, as the evaluated tools implement different analytical paradigms and generate heterogeneous outputs.

A feature comparison (Table 1) highlights that *crocketa* offers the most comprehensive and configurable suite for integrated transcriptomic and repertoire analyses. While *Immunopipe* and *Dandelion* include certain specialized features not in *crocketa*'s initial release (e.g., metabolic analysis, clonal networks), *crocketa* provides superior granularity in reporting through detailed statistical tables which are absent in *Immunopipe*. Furthermore, *crocketa* guarantees full reproducibility via its *Snakemake* implementation and supports both T- and B-cell receptors within a unified framework (*Supplementary Table 2*).

Table 1 Comparative functional scope of single-cell immune repertoire analysis pipelines.

Feature	*crocketa*	*Dandelion*	*Immunopipe*
Integrated GEX + VDJ Analysis	Yes	Yes	Yes
GEX-Only Analysis Path	Yes	No	Yes (Limited)
Raw FASTQ Input (End-to-End)	Yes	No	No
Configurable Analytical Parameters	High	Low	Medium
Functional Signature Analysis	Yes	No	Yes
Trajectory Inference	Yes	Yes	No
Metabolic Landscape Analysis	No	No	Yes
Combined TCR & BCR Support	Yes	Yes	No
Viral Clone Annotation	Yes (TCR)	No	No
Clonal Network & Clustering	No	Yes	No

This combination of efficient performance on core tasks, a unique end-to-end automated workflow, and extensive, reproducible functionality positions *crocketa* as a practical and accessible solution for integrated single-cell multi-omic analysis.

Discussion

crocketa integrates a carefully selected set of widely recognized, high-impact tools (e.g., *Seurat*, *Immunarch* [36], *STAR* [37]) that have been benchmarked in peer-reviewed studies, ensuring both methodological rigor and reproducibility. The pipeline is based on well-established and validated approaches, combining flexibility and modularity to accommodate a wide range of research contexts, from fundamental immunology to translational oncology. Researchers can analyze transcriptomic and repertoire data separately or together, with customizable workflows designed to prioritize specific research objectives.

The integrated outputs enable users to interpret clonality patterns (e.g., expansion, overlap) in direct relation to transcriptional states, exploring functional relationships within the unified analysis object. Default parameters for quality control, clonotype frequency bins, and normalization are grounded in current best practices to ensure robust out-of-the-box performance while remaining fully configurable. *Snakemake's* parallelization further enhances the pipeline's scalability, making it suitable for large-scale cohorts. By automating the integration of clonal and transcriptomic data, *crocketa* significantly reduces analytical variability and accelerates hypothesis generation. The pipeline's automated multi-step integration (e.g., clonotype-transcriptome mapping) minimizes manual biases, enabling faster and more reliable hypothesis testing. This approach is valuable across a broad spectrum of applications, from basic immunology studies to the discovery of clinical biomarkers in oncology and immunotherapy research. While the current version provides a ready-to-use analytical framework for bioinformaticians, future development will focus on broadening accessibility through a user-friendly Shiny-based interface and will incorporate advanced tools like *TESSA* [38] for Bayesian modeling of transcriptional and clonal states, enabling more rigorous joint statistical inference.

However, *crocketa* has some constraints. Its reliance on *CellRanger* [39] for BCR/TCR alignment limits repertoire analysis to *10x* Genomics datasets, excluding non-*10x* platforms like SMART-seq, though scRNA-seq analysis remains feasible. Additionally, while *Snakemake's* resource management aids in handling large datasets, the pipeline still inherits the computational demands of single-cell tools, such as *Seurat's* memory-intensive clustering, which requires high-performance computing (HPC) or cloud environments for large-scale analyses (> 1,000,000 cells). Furthermore, the integration of clonotype and transcriptome data assumes high-quality input, meaning that low RNA-seq library complexity or sparse TCR/BCR reads can lead to incomplete or inaccurate cell-clonotype pairings.

Conclusion

crocketa represents a versatile, robust and scalable solution for the integrated analysis of transcriptional and clonal data at the single-cell level. Built on *Snakemake*, it ensures parallelization, reproducibility, and modular customization, addressing key challenges in multi-omic single-cell studies. Unlike fragmented approaches, *crocketa* streamlines the entire workflow, from raw data processing to advanced immune repertoire characterization, facilitating the identification of expanded T- and B-cell clones and their transcriptional states. Its seamless integration of gold-standard tools for clonal annotation, transcriptomic signatures, and visualization makes it a powerful resource for researchers exploring immune dynamics in cancer, autoimmunity, and infectious diseases.

Designed with flexibility and scalability in mind, *crocketa* is compatible with diverse experimental designs and sequencing technologies. The pipeline will undergo continuous updates, incorporating new analytical features and improved compatibility with emerging methodologies. By enabling the correlation of immune clonality with functional states, *crocketa* has the potential to contribute significantly to immunotherapy research, particularly in identifying clonal expansions linked to exhaustion or memory phenotypes.

Materials & methods

Example use-case

crocketa pipeline has been rigorously tested with samples sequenced from over 20 leukemic mice using scRNA-seq data, and more than 10 pediatric oncology patients, including cases of sarcomas and neuroblastomas, with multiomic datasets comprising scRNA-seq, scTCR-seq and scBCR-seq. For the results presented in this study, analysis was focused on tumor-infiltrating lymphocytes (TILs) derived from pediatric oncology patients under four distinct co-culture conditions. Single-cell sequencing was carried out to characterize the transcriptome and immune repertoire at a high resolution. The data generated from these analyses are currently being prepared for publication, and the results hold significant scientific and clinical value, particularly in understanding tumor-immune interactions and identifying potential therapeutic targets. While full clinical data remains restricted due to ongoing research, here we provide partial results for a subset of the sequenced cells cultured from a human patient, visualized in Figs. 1 and 2. Additionally,

an independent public dataset is provided in the GitHub repository for testing purposes.

Although the primary dataset described in the study does not include BCR data, the presented framework is capable of processing BCR and TCR datasets either independently or jointly. The provided GitHub public dataset involves BCR analysis. The only step not implemented yet for BCRs is the viral clone annotation step, which is currently available only for T-cells.

Implementation details

crocketa is an automated and adaptable *Snakemake* pipeline designed to perform fundamental stages of single-cell analysis for both transcriptomics and repertoire data. The pipeline enables the parallelization of analysis and resources whenever feasible. *Snakemake* facilitates the parallel execution of analysis steps while ensuring that the workflow is carried out sequentially, following a predefined order. The pipeline was developed and tested on Ubuntu 24.04 LTS but is designed to run seamlessly on any standard Linux distribution. For Windows users, it can be easily deployed within a Docker container, ensuring compatibility across different operating systems. We chose *Snakemake* as the workflow management system due to its robust scalability, reproducibility, and ease of containerization. Unlike other workflow managers, *Snakemake* natively supports wrapper integration within Docker, allowing all dependencies to be encapsulated in a controlled environment while maintaining high computational efficiency.

Containerizing the pipeline in Docker offers multiple advantages, including isolation from system dependencies, cross-platform compatibility, and streamlined deployment in cloud or HPC environments. Creating a Dockerized version of the pipeline is straightforward:

1. Define a Dockerfile specifying the required environment (e.g., base image, dependencies, and *Snakemake* installation).
2. Build the image using: docker build -t pipeline_image.
3. Run the pipeline within the container using: docker run --rm -v $(pwd):/workspace pipeline_image snakemake --cores all.

This approach ensures that the workflow remains portable, scalable, and reproducible across diverse computational infrastructures.

The pipeline encompasses multiple stages, including sequencing and cell-specific quality control (QC), alignment, dimensionality reduction, cell clustering, cluster marker extraction, cell annotation, differential expression analysis, functional enrichment, trajectory inference, and RNA velocity. *crocketa* incorporates several widely used single-cell resources, such as *CellRanger*, *STAR*, *DoubletFinder* [40], *Seurat*, *scANVI*, *SingleR* [41], *Azimuth* [42], and *Immunarch*, among others, and stands out in terms of parallelization, flexibility, reproducibility, and customization. The pipeline establishes an environment for each step through *bioconda* and *conda-forge* repositories [43], reducing complexity and enhancing reproducibility.

Pipeline setup

Workflow configuration relies on three files that allow users to set up the analysis from the outset and specify input files and experimental details. These configuration files include samples.tsv (which contains information about the samples to be analyzed), units.tsv (which provides paths to the input data), and config.yaml (which contains all pipeline-customizable parameters). Input data can be provided in various formats, including FASTQ files and expression matrices.

The pipeline workflow is summarized in the graphical abstract, and each stage of the analysis is detailed below.

Quality control

When *FASTQ* files are provided as input and *crocketa* is initiated, a quality control (QC) analysis step is executed to assess quality of the sequencing using widely employed tools such as *FastQC* [44]. Additional *MultiQC* reports are obtained summarizing *FastQC* and *FastQ screen* [45] information and integrating this data with other software such as *RSeQC* [46]. When matrix input is employed, this step is skipped.

Alignment, data preprocessing & clustering (transcriptional primary analysis)

Sequence alignment and quantification are performed only when FASTQ files are provided as input. This step is carried out using *STAR* & *STARsolo*, ensuring high-performance alignment for single-cell sequencing data. The alignment parameters are fully configurable via the config.yaml file, which serves as the master configuration file for all tools and software within the pipeline, ensuring flexibility and adaptability to diverse experimental setups. While default standard parameters are provided to accommodate common sequencing technologies (e.g., *10x* Genomics, Drop-seq), users can easily modify these settings to tailor the alignment process to their specific dataset. The configuration supports various chemistry versions and multi-omic configurations, including 5' Chemistry for multi-omic data and 3' Chemistry for scRNA-seq data exclusively.

Once these steps are performed, all analyses are carried out in *R* environments through *Seurat*.

Regarding primary analysis of transcriptional data, initial preprocessing is required. This preprocessing includes a cell-level QC, in which certain thresholds are

user-specified in configuration file. Specifically, users can define stringent threshold for minimal and maximal UMI counts (default > 200–500), molecule counts, and maximum mitochondrial/ribosomal gene expression (< 20% and < 40%, respectively). Rare genes are similarly filtered based on a minimum cell-expression frequency ($n > 3$). This step allows for the removal of low-quality, empty droplets, or apoptotic cells. In addition, *DoubletFinder* algorithm can be applied to identify and remove hypothetical multiplet artifacts in each analyzed sample.

While recommended settings are provided and further detailed in the GitHub repository and fully documented at the configuration file (config.yaml), *crocketa* maintains flexibility, allowing experienced users to tune these parameters according to their specific experimental requirements.

Retained cells will then undergo normalization (via standard methods or SCTransform), dimensionality reduction (PCA, UMAP), clustering explored across an editable, wide resolution range (typically 0.1–1.2) & visualization. At the data normalization stage, *crocketa* offers flexible handling of batch effects, which is critical given the pipeline's design to automate and process large-scale, multi-sample datasets. Users can apply standard correction methods or even integrate datasets when batch effects are particularly pronounced.

Repertoire data extraction, integration & analysis

If only scRNA-seq data is provided, this step can be skipped if wished.

When input repertoire data is FASTQ formatted, sequencing data are subjected to a second alignment through *CellRanger* pipeline. This enables clonotype data extraction from scTCR-seq or scBCR-seq technologies. Consequently, a *filtered_contig_annotations.csv* output file is generated, containing clonotype data for each cell across all analyzed samples. This file is essential for repertoire integration to single-cell object. In addition, *CellRanger* provides additional QC reports complementary to previously obtained *MultiQC* reports in which we can check TCR/BCR-seq quality.

If matrix files are provided as input, contig csv- formatted file must be provided and specified in the corresponding section of the configuration files.

Data integration is accomplished through *scRepertoire R software*. Moreover, based on the specific research focus or the studied disease, users might only be interested in clones not associated with viruses (labeled as non-viral) as happens in a tumoral context. Therefore, clonotype analysis includes a preliminary annotation of TCR clones as viral or non-viral, according to the *VDJ* database (*VDJdb*) [47] through the *Immunarch* R package. Once completed, two different *Seurat* objects are generated: one containing all available information and another containing only non-viral cells, filtering out cells associated with clones matching *VDJdb*. Analysis is conducted on both objects.

At this stage, several repertoire metrics can be obtained in order to elucidate clonal expansion, clonal diversity, clonal overlap and other metrics within each single sample, while also enabling comparisons across all of them.

Post-clustering analysis

Finally, post-clustering analyses are applied if enabled by the user to the single-cell object. Analysis starts with a cell annotation step, in which *crocketa* aims to elucidate cell identity, and avoid any bias regarding computational software or cell references. With this purpose, two approaches are carried out:

- Automatic annotation: The pipeline integrates *scType* as the primary annotation tool, utilizing a customizable reference framework to identify cell populations. For enhanced precision, these labels are further refined with *scANVI*, specifically to reclassify ambiguous or unclassified cells. Additionally, the pipeline supports optional annotation through *SingleR* and *Azimuth*. In contrast to *scType*'s user-editable reference, these tools rely on pre-defined reference datasets, enabling advanced users to cross-validate cell identities, selecting specialized reference sets for high-resolution benchmarking.
- Manual data explore: A gene-set of interest is user-provided and expression for every gene is assessed in our single-cell object through *Seurat*. This information is complementary to automatic annotation approach to confirm annotation results. If there is no gene-set specified, a general list of marker genes regarding bone-marrow subpopulations is provided.

Results are extracted at both single cell level and cluster level. The main point is to integrate all cell- label information in order to provide a robust cell annotation.

Furthermore, this post-clustering analysis will enable population marker extraction through differential expression analysis, up/down-regulated biological pathways from functional enrichment analysis, and different cell states in a pseudotime space from trajectory inference analysis. These analyses are performed using rigorously validated software including *Vision*, *GSEA-Preranked* method from *MSigDB* [48], *Slingshot*, *Velocyto* and *CellRank*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12767-y.

Supplementary Material 1.

Acknowledgements
The authors sincerely appreciate the work of the whole FIBHNJS staff.

Authors' contributions
GSA: conceptualization, data curation, methodology, software, visualization, writing - original draft, writing - review & editing. GGA: resources, supervision, writing - review & editing. MPR: conceptualization, software, supervision, writing - review & editing. JL: conceptualization, supervision, writing - review & editing. AGM: resources, conceptualization, supervision, writing - review & editing. EGS: resources, conceptualization, supervision, writing - review & editing. CSL: conceptualization, supervision, writing - review & editing. MRO: conceptualization, resources, supervision, writing - original draft, writing - review & editing. JGM: conceptualization, methodology, software, writing - original draft, writing - review & editing.

Funding
This work was supported by Instituto de Salud Carlos III, ***ISCIII*** (grant number PMP21/00088, PI23/00703) and Fundación Familia Alonso. MRO's lab is supported by Asociación Pablo Ugarte, Asociación NEN, Fundación Neuroblastoma and Fundación Unoentrecienmil.

Data availability
The datasets generated and analysed during the current study are not publicly available due to ongoing research and publication constraints. However, a public dataset is provided in the official GitHub repository for testing and validation purposes. For detailed information regarding the dataset structure and its application, please refer to the "Example Use-Case" subsection under Materials and Methods. https://github.com/OncologyHNJ/crocketa.

Code availability
Project name: ***crocketa***.
Project home page: https://github.com/OncologyHNJ/crocketa.
Operating system(s): ***Platform independent (Linux recommended)***.
Programming language: ***Snakemake***, ***Python***, ***R***, ***Bash***.
Other requirements: ***Following installation instructions as described in github – Snakemake***, ***python***, ***java11***, ***conda/mamba***, ***pandas***, ***numpy = 1.23.5***, ***CellRanger 7.2.0***, ***Docker/Singularity optional***.
License: ***MIT License***.
Restrictions for non-academics: ***No restrictions; open-source software available for academic and non-academic use***.

Declarations

Ethics approval and consent to participate
All procedures involving human participants were approved by Ethic Committee under protocol number R-0076/19, and written informed consent was obtained in accordance with institutional guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Fundación para la Investigación Biomédica Hospital Infantil Universitario Niño Jesús (FIBHNJS), Madrid, Spain
[2]Hospital Infantil Universitario Niño Jesús (HUNJ), Madrid, Spain
[3]Instituto de Investigación Sanitaria Hospital de la Princesa (IISLP), Madrid, Spain
[4]Fundación para la Investigación Biomédica Hospital Universitario la Princesa (FIBHUP), Madrid, Spain
[5]Escuela de Doctorado, Universidad Autónoma de Madrid (EDUAM), Madrid, Spain

Received: 24 November 2025 / Accepted: 16 March 2026
Published online: 21 March 2026

References

1. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. 2015. Available from: www.sciencemag.org.
2. Rohaan MW, Borch TH, van den Berg JH, Met Ö, Kessels R, Geukes Foppen MH, et al. Tumor-Infiltrating Lymphocyte Therapy or Ipilimumab in Advanced Melanoma. N Engl J Med. 2022;387:2113–25.
3. Nakahara Y, Matsutani T, Igarashi Y, Matsuo N, Himuro H, Saito H, et al. Clinical significance of peripheral TCR and BCR repertoire diversity in EGFR/ALK wild-type NSCLC treated with anti-PD-1 antibody. Cancer Immunol Immunother. 2021;70:2881–92.
4. Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. Front Immunol. 2013;4(485):4.
5. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566:393–7.
6. Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multi-modal single-cell RNA sequencing. ImmunoInformatics. 2022;5:100009.
7. Gao S, Wu Z, Arnold B, Diamond C, Batchu S, Giudice V et al. Single-cell RNA sequencing coupled to TCR profiling of large granular lymphocyte leukemia T cells. Nat Commun. 2022;13(1):1982.
8. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Front Immunol 2018;9:224.
9. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302(5909):575–81.
10. Burnet FM. Theories of immunity. Perspect Biol Med. 1960;3:447–58.
11. Landsverk OJB, Snir O, Casado RB, Richter L, Mold JE, Réu P, et al. Antibody-secreting plasma cells persist for decades in human intestine. J Exp Med. 2017;214:309–17.
12. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends Immunol. 2015;36:738–49.
13. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K et al. High- throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10(1):3120.
14. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: Strategies and applications. Trends Immunol. 2014;35:581–90.
15. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014;32:158–68.
16. Wardemann H, Busse CE. Novel Approaches to Analyze Immunoglobulin Repertoires. Trends Immunol. 2017;38:471–82.
17. Frank ML, Lu K, Erdogan C, Han Y, Hu J, Wang T, et al. T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. Clin Cancer Res. 2023;29:994–1008.
18. Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022;23(1):49.
19. Zappia L, Richter S, Ramírez-Suástegui C, Kfuri-Rubens R, Vornholz L, Wang W, Dietrich O, Frishberg A, Luecken MD, Theis FJ. Feature selection methods affect the performance of scRNA-seq data integration and querying. Nat Methods. 2025;22(4):834–44.
20. Luecken MD, Gigante S et al. Defining and benchmarking open problems in single-cell analysis. Res Sq. 2024;rs.3.rs-4181617. [Preprint].
21. Garcia-Jimeno L, Fustero-Torre C, Jimenez-Santos MJ, Gomez-Lopez G, Di Domenico T. Al- Shahrour F. bollito: a flexible pipeline for comprehensive single-cell RNA-seq analyses. Bioinformatics. 2022;38:1155–6.
22. Yermanos A, Agrafiotis A, Kuhn R, Robbiani D, Yates J, Papadopoulou C et al. Platypus: An open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. NAR Genom Bioinform. 2021;3(2):lqab023.

23. Borcherding N, Bormann NL, scRepertoire. An R-based toolkit for single-cell immune receptor analysis. F1000Res. 2020;9:47.
24. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.
25. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
26. 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 [Internet]. 2024;42:293–304.
27. Conway JR, Lex A, Gehlenborg N, UpSetR. An R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40.
28. DeTomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. Functional interpretation of single cell similarity maps. Nat Commun. 2019;10(1):4376.
29. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246.
30. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17(1):e9620.
31. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.
32. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560:494–8.
33. Lange M, Bergen V, Klein M, et al. CellRank for directed single-cell fate mapping. Nat Methods. 2022;19:159–70.
34. Wang P, Yu Y, Dong H, Zhang S, Sun Z, Zeng H et al. Immunopipe: a comprehensive and flexible scRNA-seq and scTCR-seq data analysis pipeline. NAR Genomics Bioinf. 2025;7(2):lqaf063.
35. Chenqu Suo K, Polanski E, Dann RGH, Lindeboom et al. Roser Vilarrasa-Blasi, Roser Vento-Tormo, Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origin. Nature Biotechnology. 2024;42:40–51.
36. Nazarov VI, Tsvetkov VO, Fiadziushchanka S, Rumynskiy E, Popov AA. immunarch: Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires. https://immunarch.com/, https://github.com/immunomind/immunarch2023.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
38. Zhang Z, Xiong D, Wang X, et al. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat Methods. 2021;18:92–9. https://doi.org/10.1038/s41592-020-01020-3.
39. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
40. McGinnis CS, Murrow LM, Gartner ZJ, DoubletFinder. Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8:329–e3374.
41. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72.
42. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–e358729.
43. Grüning B, Dale R, Sjamp A, Chapman BA, Rowe J, Tomkins-Tinch CH et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. Available from: https://github.com/gentoo/sci.
44. Babraham Bioinformatics. FastQC a Quality Control Tool for High Throughput Sequence Data. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [cited 8 Nov 2023].
45. Wingett SW, Andrews S. Fastq screen: A tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.
46. Wang L, Wang S, Li W, RSeQC. Quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.
47. Goncharov M, Bagaev D, Shcherbinin D, Zvyagin I, Bolotin D, Thomas PG, et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. Nat Methods. 2022;19:1017–9.
48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. 2005. Available from: https://www.pnas.org/doi/10.1073/pnas.0506580102.

Publisher's Note