nature protocols

https://doi.org/10.1038/s41596-025-01257-2

Protocol

Check for updates

Navigating the data processing for cytometry-based single-cell proteomics

Huaicheng Sun[1,2,11], Yuan Zhou[1,11], Ruoyu Jiang[1,3,11], Yuxuan Liu[4], Chengbin Gu[5], Ziqi Pan[1], Minjie Mou[1], Xichen Lian[1], Bohan Chen[6], Tianle Niu[7], Ying Zhang[1], Yintao Zhang[1], Baoliang Zhang[8], Xiuna Sun[4], Hao Yang[7], Xin Shen[1], Yangbo Dai[9], Jiannan Deng[1], Siqi Liu[3], Yang Zhang[7], Mang Xiao[4], Wanqing Xie[8], Qingxia Yang[10], Tingting Fu[1] ✉ & Feng Zhu[1,2] ✉

Abstract

Cytometry-based single-cell proteomics (SCP) has emerged as a powerful technique that greatly advances our understanding of complex biological systems with a new level of granularity. Various methods have been developed to process cytometry-based SCP data. However, it remains extremely challenging to identify the well-performing processing workflows for specific datasets. Here, we develop ANPELA, an out-of-the-box method for navigating the proteomic data processing based on large-scale screening. It enables a comparison among the performances of thousands of the processing workflows in identifying cell subpopulations and inferring pseudo-time trajectories based on machine learning. Several cases are then analyzed, highlighting its ability to identify the optimal ways of data processing for cytometry-based SCP studies. A new package is also deployed to ensure multiscenario usability (such as desktop software, R package and online server), data security (enabling local and open-source execution) and a user-friendly interface (realizing interactive and visualizable applications). Overall, ANPELA can be utilized by a broad audience, including those without coding skills, and is freely accessible and downloadable at https://idrblab.org/anpela/. Its execution time may range from minutes to hours depending on the size of the analyzed data.

Key points

- ANPELA is a tool for evaluating the utility of proteomic data processing workflows and is intended to facilitate the automatic selection of the most appropriate processing methods for single cell proteomic data.
- ANPELA enables a systematic assessment of existing data processing methods for cell subpopulation identification and pseudo-time trajectory inference.

Key references

Zhang, Y. et al. *Adv. Sci.* **10**, e2207061 (2023): https://doi.org/10.1002/advs.202207061

Jurburg, S. D. et al. *Microbiome* **10**, 225 (2022): https://doi.org/10.1186/s40168-022-01423-8

Tang, J. et al. *Brief. Bioinform.* **21**, 621–636 (2020): https://doi.org/10.1093/bib/bby127

Tang, J. et al. *Mol. Cell. Proteomics* **18**, 1683–1699 (2019): https://doi.org/10.1074/mcp.RA118.001169

Cui, X. et al. *Front. Pharmacol.* **10**, 127 (2019): https://doi.org/10.3389/fphar.2019.00127

[1]College of Pharmaceutical Sciences, National Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou, China. [2]Department of Pharmacy, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [3]Chu Kochen Honors College, Zhejiang University, Hangzhou, China. [4]Department of Otolaryngology Head and Neck Surgery, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China. [5]Zhejiang University—University of Edinburgh Institute, Zhejiang University School of Medicine, Haining, China. [6]College of Life Sciences, Zhejiang University, Hangzhou, China. [7]School of Pharmacy, Hebei Medical University, Shijiazhuang, China. [8]Department of Intelligent Medical Engineering, School of Biomedical Engineering, Anhui Medical University, Hefei, China. [9]State Key Laboratory for Organic Electronics and Information Displays, Institute of Advanced Materials, Nanjing University of Posts and Telecommunications, Nanjing, China. [10]Zhejiang Provincial Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases, Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China. [11]These authors contributed equally: Huaicheng Sun, Yuan Zhou, Ruoyu Jiang. ✉e-mail: futt@zju.edu.cn; zhufeng@zju.edu.cn

Introduction

Single-cell proteomics (SCP) holds the transformative potentials to unravel intricate mechanisms in cells, with an unprecedented degree of granularity in protein profiling[1,2]. For decades, the SCP studies, especially those based on cytometry technique, have made substantial progress[3–6], which tremendously benefit from the multifaceted characteristics of flow and mass cytometry, including the abilities of high-throughput screening[7], high-sensitive measurement[8] and so on[9–11]. Within a cytometry-based SCP study, a typical pipeline for analyzing raw data include three essential procedures: data quantification, data processing and data interpretation[12–14]. Particularly, the data quantification retrieves protein expression matrix from the raw experimental signals, converting expression information into a readable format[12]; the data processing removes biases unrelated to experimental condition, helping to adjust the data distribution for the downstream analysis[13]; and the data interpretation mines biological information within the dataset, acquiring the profound insights into cell heterogeneity[14]. Because of the inherent difficulty of these procedures, there is an urgent need for powerful and well-validated tools assisting cytometry-based SCP analyses[15].

Till now, a variety of tools have been developed to promote both the data quantification and data interpretation in cytometry-based SCP study[16,17]. For data quantification, sophisticated tools have been developed in a targeted manner to accommodate different cytometer models, which lays the foundations for subsequent procedure[18]. For data interpretation, existing tools (such as FlowSOM[19] and FLOW-MAP[20]) enable the extraction of critical biological insights through statistical analysis. Owing to the widespread application of these tools, standardized protocols for data quantification and data interpretation have been constructed[21–23]. Despite of the great efforts devoted into the above two procedures, notable challenges remain in the data processing of current cytometry-based SCP studies[24–26]. Specifically, in cytometry-based SCP study, >3,000 data processing workflows can be generated by randomly combining the methods among processing steps[27], and there is a clear data-dependent nature in the performance of selected data processing workflow[28]. In other words, it is extremely challenging to identify the workflow of optimal performance specifically for a studied set of cytometry-based SCP data, which asks for the development of protocol not only comprehensively covering all steps in data processing but also effectively navigating the selection of the optimal workflow[27,28].

In this Protocol, we provided the usage of ANPELA for navigating data processing (that is, assisting users in selecting the optimal workflow) for any cytometry-based SCP study[27,29,30]. A user-centric and application-oriented protocol was proposed here to make ANPELA easily accessible for user without any programming skill. In the 'Anticipated results' section, we showed how it facilitated data processing on ten benchmark datasets and assist the users to acquire reliable biological insights. Moreover, a variety of critical improvements were realized comparing with the previous ANPELA[27,29,30], such as developing desktop software and providing open-source R package.

Development of the Protocol

To meet the demands of proteomic communities for data processing tool, we construct ANPELA (https://idrblab.org/anpela/)[29], an online platform that is freely accessible, user-friendly and regularly updated[27]. The original version of the ANPELA[29] offers, for the first time, an evaluation system for proteomics study, enabling the identification of the optimal workflow. Compared with other methods, ANPELA can not only automatically detect the diverse data formats produced by different quantification tools but also provide the most comprehensive set of processing methods among available tools. In a recent upgrade[27] to extend its capabilities to cytometry-based SCP studies, we incorporated a performance evaluation from multiple perspectives to identify the optimal workflow for specific project from thousands of available processing workflows. Particularly, ANPELA now enables a systematic assessment facilitating both cell subpopulation identification (CSI) and pseudo-time trajectory inference (PTI) using independent criteria[27]. CSI indicated the assignment of cell types/subtypes for single-cell

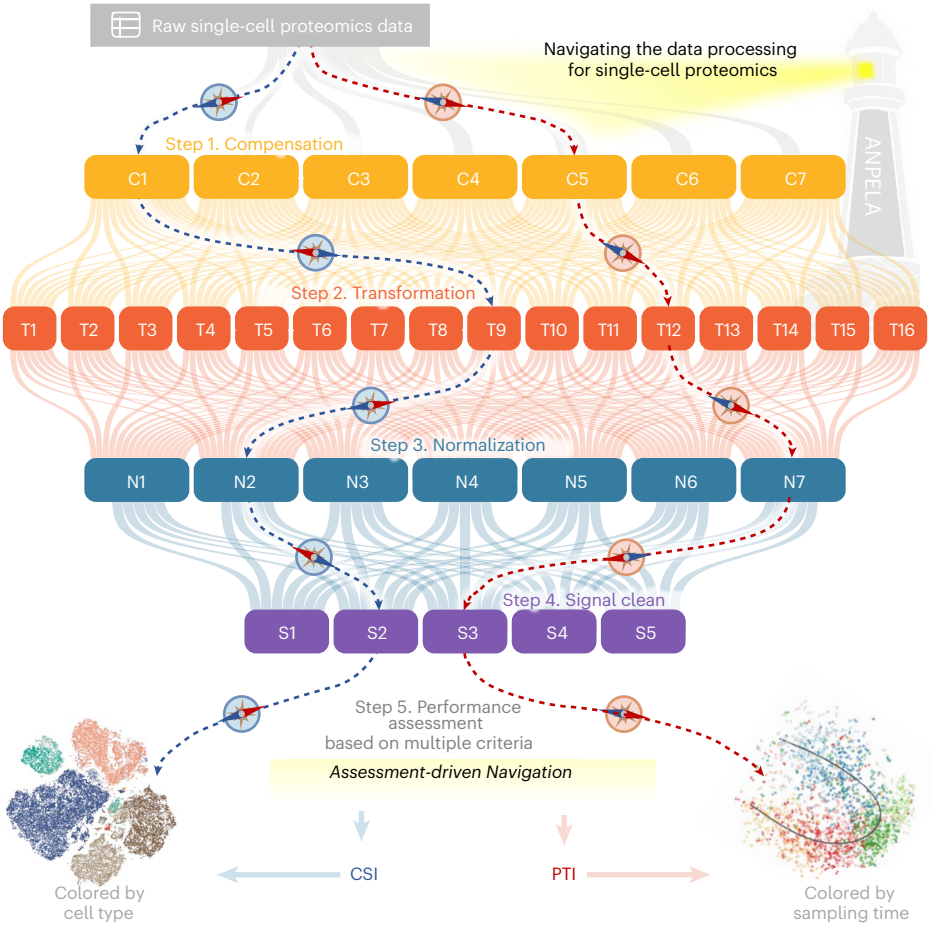


Fig. 1 | The schematic overview of ANPELA. ANPELA enabled the navigation of data processing for cytometry-based single-cell proteomic study. Particularly, the raw data underwent sequential processing steps (1) data compensation, (2) data transformation, (3) data normalization and (4) signal clean. The processed data then underwent step 5 of performance assessment based on multiple criteria. To identify the optimal processing workflow, an assessment-driven navigating strategy was tailored for the two critical tasks in single-cell proteomics: CSI and PTI. More details about the strategy were explicitly provided in the 'Experimental design' section. Each method used in the processing steps was denoted by an abbreviation consisting of the initial letter of the corresponding step and a number. Methods derived from the same step shared the same background color. A comprehensive introduction to all processing methods is available in Supplementary Table 1. Two dot plots at the bottom exemplified the outcome of two critical tasks: CSI (left) and PTI (right), with dots colored by cell types and sampling time, respectively.

experiment, which was one fundamental step in single-cell analyses[31]. PTI referred to the inference of trajectory among cells, and was essential for the exploration of biological processes with continuous dynamics[32].

The latest ANPELA (as shown in Fig. 1) described in this protocol consists of five steps: (1) data compensation, (2) data transformation, (3) data normalization, (4) signal clean and (5) performance assessment based on multiple criteria. In each step, a variety of processing methods existed. The details about all those existing methods were explicitly offered in Supplementary Table 1. For step 5, two series of criteria from multiple perspectives were proposed for CSI and PTI, and the details about those criteria were explicitly described in Supplementary Method 1.

To achieve truly comprehensive navigation of cytometry-based SCP data processing, the latest ANPELA introduces substantial upgrades in this Protocol. Particularly, (1) the number of processing workflows increases from ~1,000 to >3,000, covering the most comprehensive set

of processing methods currently available in research community; (2) the assessment process now supports cell type annotations and multi-branch trajectory inference, rather than being limited to cell clustering and single-lineage trajectory analysis; and (3) a standalone desktop application and an R package are developed, along with substantial upgrade to the web-based platform, endowing ANPELA with multiple usability advantages (such as: multiscenario deployments, data security, open-source accessibility and user-friendly interface) for serving a broader user base. Besides, both online and local versions of ANPELA were provided in this protocol to attract the broad research interests from bioinformaticians, clinicians, system biologists, pharmacists, pathologists, computational biologists and so on, both of which could be readily accessible at https://idrblab.org/anpela/. To accelerate the execution of ANPELA, an R package realizing parallel computing along with comprehensive documentation and tutorial case was also developed, which empowered the users to tailor its functionality according to their customized requirements, and even incorporate it into their analytical pipelines.

Applications of the method

ANPELA has accumulated extensive interests and widespread applications in modern proteomic studies[33–41]. Particularly, ANPELA has been adopted by various proteomics research to adjust the distribution of the raw data and minimize the impact of nonexperimental variables and has been frequently considered to 'aid in the discovery of markers', 'provide well-established criteria for performance evaluation', 'enhance precision, accuracy and reproducibility' and so on[33,37–40]. Furthermore, a single-cell proteomic database SingPro[9] was developed by our group, 20 datasets from which were randomly collected to assess the performances of our newly proposed protocol. All collected datasets could be smoothly analyzed, and each analysis was completed in a limited time frame (if running on a laptop of 2.10 GHz, 32GB RAM and Intel Core i7-13700F, all analyses were completed in 4 h). Meanwhile, the navigating strategy of ANPELA is recognized as 'novel' and 'increasingly important', with some researchers advocating for its application to other research domains[34,37–40]. So far, it has received over 100 visits per day from distinct IP addresses and accumulated over 100,000 visits since its release in 2020. Because of its current impacts, we are optimistic to the sustainable future contributions of ANPELA to our proteomics communities. In the future, ANPELA will undergo continuous updates to expand its capabilities into additional research domains. With the rapid technological advancements, ANPELA may help to open a new chapter. Taking the spatial proteomic technique as an example, notable challenges in this field (such as the extremely high nonexperimental biases) were reported[42,43]. Due to the technological resemblance between spatial proteomics and cytometry-based SCP technique[44,45], the application of ANPELA to this and other emerging research domains may hold substantial potentials.

Overview of the ANPELA protocol

As shown in Fig. 1, ANPELA was a data processing tool for cytometry-based SCP study. It enabled user to execute over 3,000 data processing workflows and employ the assessment-based navigating strategy for discovering the optimal one. The aim of this Protocol was to help users acquire reliable biological insights by picking out the optimal workflow using ANPELA. Particularly, data processing steps, described in Fig. 1, of ANPELA included: (1) data compensation, (2) data transformation, (3) data normalization and (4) signal clean. The details about the data processing steps were provided in Box 1, which highlighted the comprehensiveness of the processing methods offered by ANPELA. Moreover, the assessment-based navigating strategy was specifically tailored for two of the most critical problems (CSI and PTI) in SCP research. Such strategy could evaluate all data processing workflows based on multiple independent criteria and eventually give the overall rankings.

In this Protocol, we provided two independent procedures (Procedures 1 and 2, shown in Fig. 2) to meet the diverse needs of different users. Procedure 1 involved the uses of ANPELA via desktop software/web platform of graphical user interface (GUI), asking for no prior programming expertise. Procedure 2 used ANPELA R package in R environment. Compared with the desktop version, R package achieved faster speeds through parallel computing and enabled the adjustment of parameters by user, offering greater flexibility. Details about the

BOX 1

Brief introduction to the processing steps

Data compensation aims to reduce the signal spillover between the different signal channels[77], which may generated from instrument-related error, fluorescence spillover in flow cytometry[15], isotopic impurities in mass cytometry[16] and so on. Since the signal spillover shows an approximately linear relationship with the original signal, a compensation matrix can be applied to correct the interference between channels based on this linear relationship.

Data transformation converts the data that illustrate a strongly skewed distribution into a more symmetric and normal distribution, meanwhile corrects the heteroscedasticity of data[78,79]. Data transformation can aid in distinguishing different subpopulations and enhance the reliability of downstream analyses, making it a critical step for the cytometry-based SCP studies[79,80].

Data normalization is a critical step that mitigates the impact of technical artifacts arising from non-experimental factors, such as variability in instruments, operators and the collection times. By accounting for the sources of technical artifacts, normalization can extensively improve the comparability and consistency of cytometry-based SCP datasets from different conditions[4,57,81]. This processing step is particularly necessary for large-scale and longitudinal researches, where the accuracy and reliability of data interpretation are affected by technical artifacts.

Signal clean removes erroneous cells in the data, which are generated by some interferences of instrument, such as bubbles, large particulates, fluctuations in flow rate, voltage instability and unstable sample uptake. Removing the erroneous cells manually is extremely time-consuming. Consequently, various automated tools are developed to offer an efficient approach[55,60,82].

Specific data processing methods included in the above processing steps, along with the related information, had been explicitly described in Supplementary Table 1.

R functions in the ANPELA R package were provided in Supplementary Table 2. Two procedures shared the identical key functionality, including four steps: data preparation, data processing, performance assessment and workflow ranking (offered in Fig. 2). The desktop version offered the user-friendly GUI with step-by-step interactive tutorial (Extended Data Fig. 1), making it accessible to user without prior programming expertise. By contrast, the R package version enabled parallel computing and showed considerably faster processing speed, but lacked the user-friendly GUI, making it suitable for the users with a background in R programming.

Experimental design

The preparation of single-cell proteomic data

There were two types of input for ANPELA. The first type was the essential data, which included flow cytometry standard (FCS) files (that is, raw data files) generated by cytometry-based SCP experiment and a metadata file elucidating the correlation between raw data and experimental condition. It was key to know that there was requirement for the FCS files provided by users: for a PTI analysis, a minimum of two time points was required, with at least one FCS file per point; for CSI analyses, two experimental conditions were needed, each with a minimum of two FCS files. Moreover, strict correspondence between file names in the metadata file and names of the FCS files (excluding the '.fcs' extension) was required. The second type was the optional data, which described additional information for processing methods or offered prior knowledge for assessment. Prior knowledge included known markers for CSI assessment and pathway hierarchy file describing the sequential order of protein expression for PTI evaluation. The details about the data preparation could be found in Box 2.

The scanning of the data processing workflows

As shown in Fig. 1, ANPELA comprised four data processing steps with multiple choices in each step (7 for data compensation, 16 for data transformation, 7 for data normalization and 5 for signal clean; '`NON`' means the absence of method in specific step). By randomly combining all methods among four steps, a total of 3,328 workflows could be finally generated. In particular,

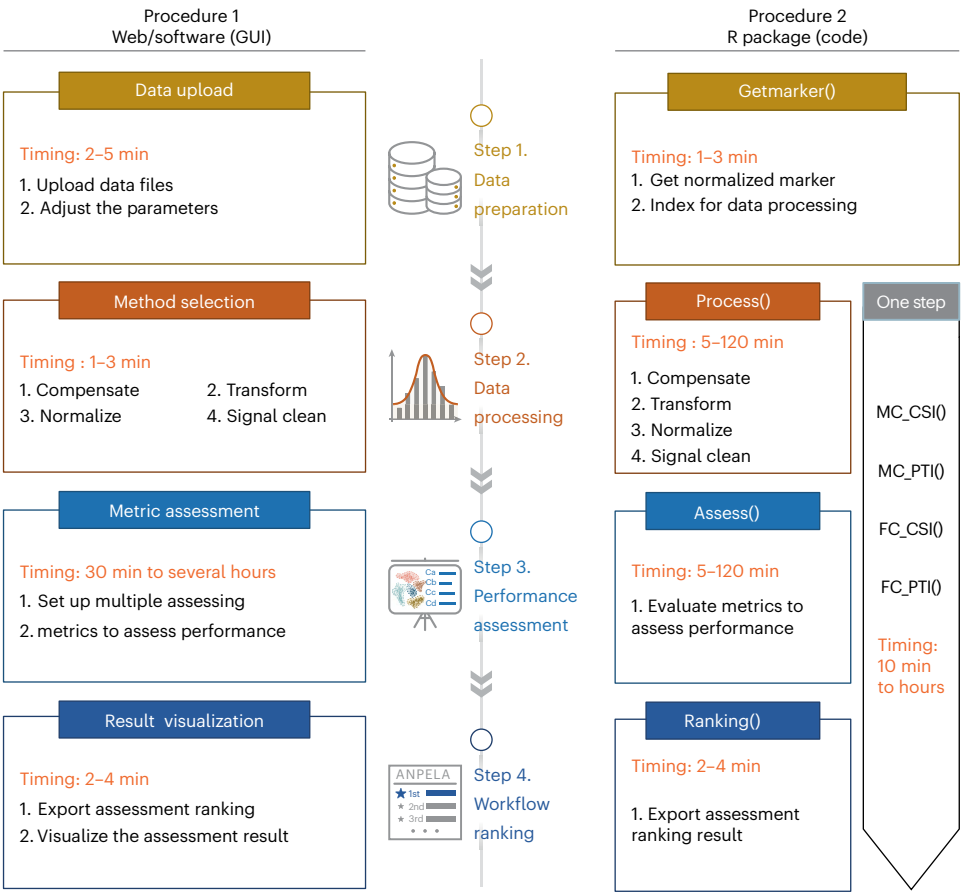


Fig. 2 | Overview of the Protocol procedures. Our ANPELA gave two procedures: Procedure 1 based on graphical user interface (GUI), where users could interact with ANPELA by web- and desktop-based platforms and Procedure 2 based on the *R* package, where users accessed ANPELA's functionality through R function (provided in Supplementary Table 2). Both procedures shared the identical core functions, containing four steps: data preparation, data processing, performance assessment and workflow ranking. Moreover, Procedure 1 allowed a direct interaction by GUI, while Procedure 2 utilized the R functions to implement the same process. All R functions were followed by parentheses, such as 'Getmarker()'. The time required varies considerably depending on several factors, including the number of workflows executed, the number of cores used in parallel and the number of cells involved in the analysis.

1,920 and 1,792 workflows were provided by ANPELA for processing those SCP data from flow cytometry and mass cytometry, respectively, and the detail information was explicitly described in Supplementary Method 2. In other words, ANPELA allowed the users to employ nearly all workflows for cytometry-based SCP data processing, laying the groundworks for identifying the optimal one. Moreover, users were also allowed to customize parameters on the basis of the own needs. For some data processing methods requiring additional data input, ANPELA would not run these methods by default unless the users provided the required additional data input files. Moreover, in some cases, it may not be essential to apply method at every data processing step (that is, selecting the 'NON' option in specific step is allowed in ANEPLA).

The navigating strategy using multiple criteria

As illustrated in Fig. 3, two sets of well-designed criteria were constructed to assess the workflow performance for CSI and PTI. All criteria in a set were independent. A comprehensive evaluation of all workflows was achieved through the collective consideration of the evaluating results from multiple independent perspectives, which aimed to obtain the reliable outcomes.

BOX 2

Brief introduction to the data preparation

Raw data files (mandatory) refer to the raw data files generated from the cytometry-based SCP experiments, with '.fcs' extension. The files give the information on protein expression in cells.

Metadata file (mandatory) is a CSV file titled 'metadata.csv', containing two columns. For CSI analysis, the two columns should be labeled as 'filename' and 'condition'. For PTI analysis, the two columns should be labeled as 'filename' and 'timepoint'. Specifically, 'filename' offers the file name of raw data files, which should be consistent with the file names of raw data files but without the '.fcs' extension. The 'condition' refers to the binary experimental conditions in CSI analyses, such as control/experiment, with at least two raw data files under each condition. The 'timepoint' indicates the sampling times in PTI analyses, requiring at least two time points. Notably, for PTI analyses, if users have prior knowledge suggesting that distinct conditions lead to distinct cellular trajectories, the metadata file should incorporate a 'condition' column. This column will denote the different branching conditions relevant to the PTI analyses.

Known marker(s) (optional), the prior knowledge for CSI analyses, is the known differentially expressed protein(s) between different conditions. In the desktop software of ANPELA, known markers can be selected through a checkbox. In the R package of ANPELA, known marker can be assigned to the parameter 'DEP' that is essential for the function of Assess().

Pathway hierarchy file (optional), the prior knowledge for PTI analyses, is a CSV file offering the sequential order of protein expression. In this file, the first protein in each column is known to reach its expression peak earlier than the other proteins in the same column. If the sequential orders are described, the user should fill them into multiple columns in CSV file. The format of the CSV file can be found in the example data for PTI analysis, which can be downloaded from https://idrblab.org/anpela/PTI-example.zip. For ANPELA desktop software (Procedure 1), this CSV file can be upload through 'Browse' bottom. For ANPELA R package (Procedure 2), this CSV file should be assigned to the 'pathwayhierarchy' parameter for the Assess() function.

Known cell type file (optional) serves as prior knowledge for CSI analyses. It is a CSV file that provides ground truth cell type labels. This file consists of two columns. The 'index' column represents cell IDs, which are formed by combining the file name and the cell index with an underscore. The 'celltype' column indicates the ground truth cell type labels. The format of the CSV file can be found in the case data for CSI analysis, which can be downloaded from https://idrblab.org/anpela/ANPELA_exampledata.zip. As described in Supplementary Table 2, the path of this file should be specified in the 'known_celltype_path' parameter.

Annotation marker file (optional) is also prior knowledge for CSI analyses. It can be either a CSV or XLSX file that offers key protein marker information for cell annotation. This file contains three columns: 'cellName', which indicates the cell type name; 'posGene', which refers to the genes that are highly expressed in the corresponding cell type; and 'negGene', which represents the genes that are lowly expressed in the cell type. The format of the file can be found in the case data for CSI analysis, which can be downloaded from https://idrblab.org/anpela/ANPELA_exampledata.zip. As detailed in Supplementary Table 2, the path of this file should be specified in the 'marker_path' parameter.

For CSI analysis (as provided in Fig. 3a), four criteria were used for evaluating the performances of workflows, which include:

(Ca) Accuracy reflects the impact of workflow on the classification accuracy of machine learning classifiers in distinguishing cells across different conditions within each cell subpopulation (default metric: area under the curve)[46].

(Cb) Tightness describes the effect of the workflow on the compactness of all cell clusters, based on the assumption that an ideal clustering should exhibit high intracluster similarity and high intercluster heterogeneity (default metric: silhouette coefficient)[47].

(Cc) Robustness demonstrates the influence of workflow on the consistency of identified biomarkers across different randomly sampled subsets of specific clusters, measured by the number of overlapped biomarkers (default metric: relative weighted consistency)[48].

(Cd) Correspondence shows the effect of workflow on the correspondence between the results obtained from analysis and prior knowledge, including differentially expressed genes or known cell type annotations in the dataset (default metric: sensitivity, recall)[47].

For PTI analysis (as provided in Fig. 3b), four criteria were used for evaluating the performances of workflows, which contained:

(Ca) Conformance assess the influence of workflow on the consistency between the inferred pseudotime and the actual sampling time, with the assumption that the sampling time could nearly reflect the pseudo-time hierarchy of cells (default metric: time conformance score)[49].

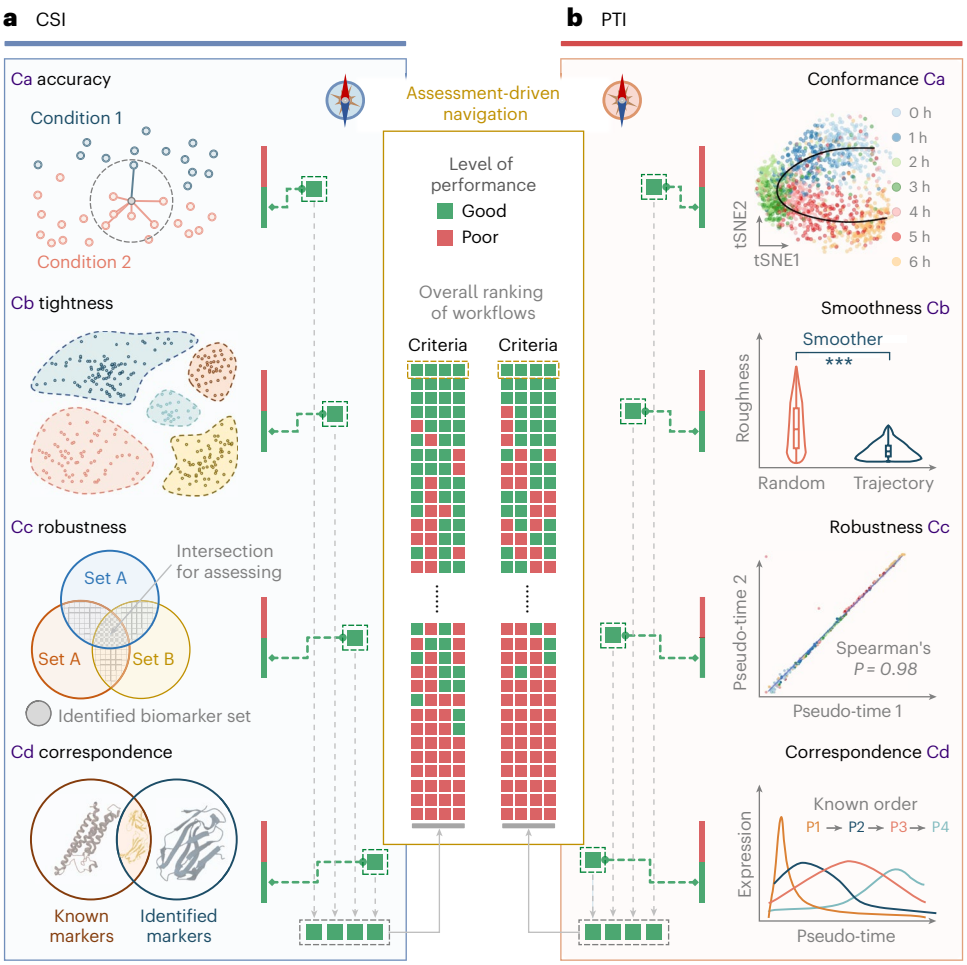


Fig. 3 | Assessment-driven navigation based on multiple criteria. ANPELA systematically assessed all data processing workflows from multiple perspectives. **a**, For CSI, accuracy, tightness, robustness and correspondence were used for measuring. **b**, for PTI, conformance, smoothness, robustness and correspondence were adopted for assessing. The performance of the workflows could be categorized to 'good' or 'poor' based on the well-established cutoff value for each criterion. A comprehensive assessment of workflow was completed by considering all assessing results from multiple perspectives.

(Cb) Smoothness measures the effect of workflow on the smoothness of protein expression dynamics along the inferred trajectory, under the expectation that cell state transitions occur smoothly in pseudo-temporal order (default metric: smoothness score)[50].

(Cc) Robustness evaluates the impact of workflow on the robustness of the inferred trajectories in response to perturbations in the input data, specially through random subsampling of the dataset (default metric: Spearman rank correlation coefficients)[51].

(Cd) Correspondence reflects the effect of workflow on the consistency between the analytical results and prior knowledge, including the known protein activation sequences or established trajectory branching conditions (default metric: true positive rate)[52].

Detail information on those criteria above was shown in Supplementary Method 1. Following the performance assessment, ANPELA ranks all workflows by employing a method of multiple steps. Step 1, a workflow is classified into 'good' or 'poor' based on its performance under an analyzed criterion. This classification was enabled by adopting those established cutoffs previously reported for each criterion (detailed in Supplementary Method 1). Step 2, all workflows are ranked on the basis of the number of 'good' performances assessed by multiple criteria. If one workflow scores all 'good' under four criteria, it will be ranked higher than that the one scores 'good' under less than four criteria. This process is designed to avoid a workflow receiving a high

overall ranking but having poor score under individual criterion. Step 3, for those workflows scoring same number of 'good', an additional ranking process is further performed by collectively considering their scores under all assessing criteria. Particularly, those workflows were ranked using their cumulative scores across all criteria, with a higher score corresponding to better rankings. Finally, overall ranking and processing outcome for all workflows were provided.

Comparison with available methods

A variety of tools have been developed to facilitate the data processing for cytometry-based SCP. Most of those tools focus on one particular step in data processing, such as AutoSpill[15], cytofkit[53], flowAI[54], flowClean[55], flowCut[56], flowStats[57], flowTrans[58], flowVS[59], PeacoQC[60] and spillR[61]. Some other tools cover two to three steps, such as CATALYST[16], Cytobank[62], FCSTrans[63] and flowCore[64]. Only ANPELA and FlowJo provide a comprehensive processing workflow. However, FlowJo is a commercial software with much fewer processing methods (about 16 methods, with the potential of producing less than 200 workflows) than that of ANPELA (a total of 35 methods, resulting in 3,328 workflows). In summary, compared with the available tools, ANPELA was UNIQUE in (a) comprehensively covering all steps of data processing (from compensation to transformation to normalization to signal clean finally to performance assessment), (b) offering the largest number of workflows (>3,000) among existing tools and (c) enabling the guidance of data processing for cytometry-based SCP studies by integrating machine learning, performance ranking, and parallel computation. All these features empowered our ANPELA to navigate data processing.

Expertise needed to implement the protocol

In Procedure 1, the users can easily communicate with the desktop software of ANPELA via an intuitive GUI, which requires no prior programming expertise, enabling researcher to effortlessly apply ANPELA. Moreover, both Textual Tutorial and Interactive Tutorial were embedded in the desktop software to facilitate the application of our protocol (shown in Extended Data Fig. 1). In Procedure 2, users can leverage the R package of ANPELA to realize parallel computing. A mastery of *R* language at a rudimentary level is the sole prerequisite, making it accessible even to a beginner for the *R* programming. For the users hoping to customize ANPELA's functionality, the expertise in *R* language was required. Before utilizing ANPELA R package, it is essential to read the user manual as shown in Supplementary Method 3, and the explicit details about all *R* functions integrated into ANPELA protocol could be found in Supplementary Table 2.

Advantages and limitations of the Protocol

Advantages of our ANPELA Protocol

Our Protocol facilitated the user in selecting the optimal data processing workflow and presented several advantages that made the ANPELA user-centric and application-oriented. Our ANPELA was distinguished by four key advantages: (1) multiscenarios deployment: it provided a desktop software version, an R package version, an online platform version and *R* language source code, aiming to minimize the barrier to use ANPELA and meet the diverse needs of users; (2) data security: both the R package and desktop software of ANPELA did not ask for the online upload of user's private data, which guarantees the security and privacy of user data; (3) open source: all ANPELA codes were made open-source on GitHub, structured into modular components on the basis of functionality and supplemented with appropriate comments to assist users in understanding these codes, thereby facilitating users' editing; (4) user-friendly interface: both the local and online ANPELA featured an intuitive GUI, incorporating a 'Textual Tutorial' and a 'Interactive Tutorial' to facilitate the effortless learning of our ANPELA protocol.

Limitations of our ANPELA Protocol

First, ANPELA protocol should be regularly updated to incorporate newly developed processing methods to ensure its effectiveness in guiding cytometry-based SCP data processing. Second, it is time-consuming to assess thousands of workflows. Taking four datasets shown in Supplementary Table 3 as examples, the application of this protocol varies from minutes to several hours, influenced by the number of selected workflows, input file quantity, cell number and so on. To reduce time costs, the R package of ANPELA employs parallel computing for a substantial reduction

in processing time, which is correlated with the number of parallel cores, with the R package capable of accelerating the efficiency by ~5–20×. Notably, parallel computing requires adequate device memory, and users should choose the number of cores based on to their data and device. Third, the applications of ANPELA are limited by the number of cells in raw data file. Particularly, the robustness (Cc) criterion asks for a cell sampling, where a limited number of sampled cells may lead to substantial accidental error, affecting the credibility of robustness. To ensure the reliability of the assessment results, it is advised for each FCS file to contain at least 200 cells when using ANPELA. Fourth, the desktop version of ANPELA is currently only available for the Windows operating system. Fifth, the assessment criteria may introduce potential biases. Specifically, (1) the 'criterion Ca accuracy' for CSI analysis (calculated as the mean of areas under the curve) may favor processing results that reveal moderate phenotypic differences across cell subpopulations, rather than those that detect highly pronounced difference within a single subpopulation. This suggests that in studies aiming to identify the most distinct cell types under different conditions, a high 'criterion Ca accuracy' score may not always be necessary. (2) The 'criterion Ca conformance' for PTI study is designed to measure consistency between the inferred pseudo- and actual-time (from 0 to 1). A higher 'criterion Ca conformance' value generally indicates better performances. However, a perfect value of 1 may denote an over-fitting or biologically unrealistic trajectories. Therefore, results with very high 'Ca conformance' score should be interpreted with caution.

Materials

Equipment

Hardware

- Minimum: operating system: Windows 32/64-bit, macOS, Linux; Processor: 2.0 GHz; Memory: 8 GB RAM; storage: 2 GB available hard-disk space
- Recommended: operating system: Windows 32/64-bit, macOS, Linux; processor: 4.0 GHz; memory: 32 GB RAM; storage: 8 GB available hard-disk space
- An internet connection is necessary to download the ANPELA desktop software or R packages (including the ANPELA R package and other prerequisite R packages)

Software

If following Procedure 1, using the desktop software of ANPELA protocol:

- No software is needed

If following Procedure 2, using the R Package of ANPELA protocol:

- R v4.0 or later. The installation file of R language, named '*R-x.y.z*.tar.gz' (where *x*, *y* and *z* represent software version), retrieved from CRAN R-project (https://cran.r-project.org), which is compatible with user's operating system (ANPELA was tested by 'R 4.4.1')
- RStudio. The RStudio installation file 'RStudio-*x.y.z*.zip' (where *x*, *y* and *z* represent the version) from RStudio (https://www.rstudio.com/) compatible with operating system
- RTools. The RTools installation file 'rtools*x-y-z*.zip' (where *x*, *y* and *z* denote the version) can be obtained from RTools (https://cran.r-project.org/bin/windows/Rtools/), ensuring version compatibility between the installed R and RTools
- Prerequisite R packages. This protocol needs several R packages (shown in Box 3), with certain packages from CRAN (https://cran.r-project.org), others from Bioconductor (https://bioconductor.org/), the remaining from GitHub (https://github.com/)

Benchmark datasets

As described in Table 1, ten cytometry-based SCP benchmark datasets were collected, which exhibited considerable diversity in analytical techniques (such as photomultiplier tube, avalanche photodiode and time-of-flight mass spectrometry, which covered almost all popular detectors in the cytometry-based SCP research) and study types (five are PTI analyses, and the remaining five are CSI analyses).

BOX 3

Installation of the dependent R packages

Installed from CRAN (can also from other repositories): dplyr, doParallel, rstan, Rtsne, pastecs, cowplot, ggpubr, gridExtra, MLmetrics, fossil, clusterCrit, VennDiagram, stringr, bbmle, mc2d, doSNOW, foreach, igraph, mclust, pheatmap, magrittr, withr, sampling, tree, ellipse, BBmisc, ggpointdensity, openxlsx and R.rsp.

Installation commands:

```
> CRAN_packages <- c("dplyr", "doParallel", "rstan", "Rtsne", "pastecs", "cowplot",
"ggpubr", "gridExtra", "MLmetrics", "fossil", "clusterCrit",
"VennDiagram", "stringr", "bbmle", "mc2d", "doSNOW", "foreach",
"igraph", "mclust", "pheatmap", "magrittr", "withr", "sampling",
"tree", "ellipse", "BBmisc", "ggpointdensity", "openxlsx","R.rsp")
> install.packages(CRAN_packages, dependencies = TRUE)
```

Installed from Bioconductor (can also from other repositories): flowCore, limma, SCORPIUS, slingshot, destiny, flowStats, flowAI, flowCut, flowClean, spillR, PeacoQC, systemPipeR, flowTrans, flowVS and BiocStyle.

Installation commands:

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
> Bioconductor_packages <- c("flowCore", "limma", "SCORPIUS", "slingshot", "destiny",
"flowStats", "flowAI", "flowCut", "flowClean", "spillR",
"PeacoQC", "systemPipeR", "flowTrans", "flowVS",
"BiocStyle")
> BiocManager::install(Bioconductor_packages, ask = FALSE)
```

Installed from GitHub (can also from other repositories): SDMTools, Rtsne.multicore, Rphenograph, CytoSpill, autospill, vite and FLOWMAP.

Installation commands:

```
> if (!requireNamespace("devtools", quietly = TRUE))
install.packages("devtools")
> if (!requireNamespace("remotes", quietly = TRUE))
install.packages("remotes")
> remotes::install_version("SDMTools", "1.1-221.2")
> devtools::install_github("RGLab/Rtsne.multicore")
> devtools::install_github("JinmiaoChenLab/Rphenograph")
> devtools::install_github("qmiao19/CytoSpill")
> devtools::install_github("carlosproca/autospill")
> devtools::install_github("ParkerICI/vite")
> devtools::install_github("zunderlab/FLOWMAP")
```

Equipment setup

Installation and initialization of the desktop software (for Procedure 1)

The desktop software of ANPELA can be downloaded at https://idrblab.org/anpela/ANPELA-Setup.exe, and affiliated web-server can be accessed online at https://idrblab.org/anpela/. Upon executing the installation file, a desktop shortcut will be automatically created. Users can use this desktop software without the requirement for registering or connecting to internet.

Installation and configuration of the R package (for Procedure 2)

▲ **CRITICAL** This section provides a comprehensive guide of the installation and configuration of ANPELA R package. It outlines essential tools required to ensure a smooth use of the protocol.

Table 1 | Ten single-cell proteomic datasets adopted for benchmarking this Protocol

Dataset ID[9] (Ref.)	Analytical technique	Conditions/time points	Brief description of the samples involving in the studied dataset
☐ **Five benchmark datasets for CSI**			
SCP80719 (*Science* **369**, 1210, (2020)[65])	Mass cytometry, detected by TOF	Patients with COVID-19 and healthy individuals	A single-cell Phospho-CyTOF dataset consisting of 22 cell surface proteins and 12 intracellular proteins collected from PBMC cells, to differentiate the immune responses in 36 patients with COVID-19 and that in 45 healthy people, which are sampled from the Hong Kong cohort
SCP57021 (*Acta Neuropathol*. **141**, 901, (2021)[69])	Flow cytometry, detected by APD	Patients with MG and healthy individuals	A single-cell proteomic dataset with 23 proteins collected from fresh thymus tissue, to identify the pathogenic T cell signatures associated with the MG by analyzing three patients with MG undergoing elective thymectomy and 6 healthy controls samples
SCP47065 (*PLoS Comput Biol*. **9**, e1003215, (2013)[70])	Flow cytometry, detected by UCD	Patients with uveitides and healthy individuals	A single-cell proteomic dataset based on flow cytometry containing 14 proteins collected from fresh peripheral blood samples, to investigate the diagnosis of a premature aging disorder in 14 patients diagnosed with non-infectious uveitides and 8 healthy individuals
SCP37430 (*Cell Rep*. **28**, 819, (2019)[71])	Mass cytometry, detected by TOF	Patients with GvHD and non-GvHD controls	A single-cell proteomic dataset involving 32 proteins of human PBMC cells from bone marrow transplant recipients, to identify immune signatures in bone marrow transplantation-associated graft-versus-host disease in three patients with GvHD and four without complications
SCP11272 (*J Autoimmun*. **107**, 102361, (2020)[72])	Mass cytometry, detected by TOF	T_{tolDC} cells and T_{mDC} cells	A single-cell proteomic dataset of 35 surface proteins of two antigen specific T cell lines, to find T cell signature of tolerogenic modulation by evaluating five samples of tolerogenic DC-stimulated T cells (T_{tolDC}) and five samples of mature inflammatory myeloid DCs-stimulated T cells (T_{mDC})
☐ **Five benchmark datasets for PTI**			
SCP77365 (*Nat Protoc*. **15**, 398, (2020)[22])	Mass cytometry, detected by TOF	12 Different time points	A single-cell proteomic temporal dataset using 32 proteins that represents differentiation of mESCs into the three germ layers: endoderm, mesoderm and ectoderm by capturing 12 time points (0, 1, 2, 2.5, 3, 4, 5, 6, 7, 8, 9, 10 and 11 d), with one to three FCS files per point
SCP43132 (*BMC Bioinform*. **22**, 138, (2021)[73])	Flow cytometry, detected by PMT	6 Different time points	A single-cell proteomic time-course dataset consisting of ten cell surface proteins obtained from the human embryonic stem cell line, to interpret the induction of the differentiation process of HUES9 cells by capturing six sequential time points (0, 2, 4, 6, 8 and10 d) across six FCS files
SCP36391 (*eLife* **10**, e64653, (2021)[74])	Flow cytometry, detected by APD	4 Different time points	A single-cell proteomic temporal dataset with 23 markers of PBMCs obtained by longitudinal sampling of healthy volunteers who were challenged intranasally with rhinovirus (RV-A16) to track $CD4^+$ T cell responses across four time points (−14, 0, 7 and 28 d) with four FCS files
SCP93731 (*Nat. Biotechnol*. **30**, 858, (2012)[75])	Mass cytometry, detected by TOF	8 Different time points	A gated single-cell proteomic dataset involving 14 intracellular proteins in $CD8^+$ T cells stimulated with PMA/ionomycin, to study cellular state perturbations induced by small-molecule regulators by measuring across 8 time points (0, 1, 5, 15, 30, 60, 120 and 240 min) with eight FCS files
SCP96723 (*Sci. Transl. Med*. **6**, 255ra131, (2014)[76])	Mass cytometry, detected by TOF	5 Different time points	A single-cell proteomic dataset profiling 33 proteins from whole blood samples of a patient undergoing primary hip arthroplasty, to link clinical recovery with single-cell immune signatures by collecting samples at five time points (BL, 1 h, 24 h, 72 h, 6 week) with 5 FCS files

The dataset IDs were assigned using the SingPro[9] IDs of all studied benchmark datasets. The SingPro ID for downloading raw data files, corresponding references, analytical techniques with specific detectors, the metadata of conditions or time points for data processing and performance assessment in research and brief introduction of the samples incorporated into each dataset were provided. As demonstrated, all the datasets were diverse in their analytical techniques, such as photomultiplier tube (PMT), avalanche photodiode (APD), time-of-flight mass spectrometry (TOF) and unclear detector (UCD), which covered almost all the popular detectors in cytometry-based SCP research. DC, dendritic cell; graft-versus-host disease (GvHD); human embryonic stem cell 9 (HUES9); mass cytometry by time of flight (CyTOF); myasthenia gravis (MG); peripheral blood mononuclear cells (PBMCs); phorbol 12-myristate 13-acetate (PMA); rhinovirus A16 (RV-A16); T cells stimulated with mature inflammatory myeloid dendritic cells (T_{mDC}); T cells stimulated with tolerogenic dendritic cells (T_{tolDC}).

- The R language, RStudio and RTools. Install based on those installation files compatible with user's operating system, which have been downloaded in 'Equipment' section
- The dependent R packages. Install various dependent R packages through sequentially executing the commands that have been explicitly described in Box 3
- The R package of ANPELA. Install by running the following command in RStudio:
  > devtools::install_github("idrblab/ANPELA", build_vignettes = TRUE)
  ▲ **CRITICAL** When installing ANPELA, an error message of 'ERROR: Dependent Package Not Available' indicates that some of the required R packages are not successfully installed. A user should refer to the reinstallation instruction shown in Box 3 to resolve this issue and ensure the proper installation of the missing package(s) before the subsequent setup processes.

Procedure 1: user-friendly desktop software

Preparing the input data

● **TIMING 1–3 min**

▲ **CRITICAL** Two example datasets, one for CSI analysis and another for PTI analysis, were made downloadable.

1. Download example datasets. Two example datasets (each with various raw data files & one metadata file) were made downloadable from: https://idrblab.org/anpela/CSI-example.zip and https://idrblab.org/anpela/PTI-example.zip for CSI and PTI, respectively. As offered in Box 2, users are advised to prepare their data by following the formats of these two datasets.
2. Prepare required files. Generate a CSV file titled 'metadata.csv' and ensure it is in the same directory as the raw data files (in FCS format). This file should contain two columns: for the CSI analysis, two columns in the file should be labeled as 'filename' and 'condition'; for the PTI analysis, those two columns should be labeled as 'filename' and 'timepoint'.
   - The 'filename' represents the file name of the raw data, which should correspond exactly to the file name of the raw data files but without the '.fcs' extension;
   - The 'condition' refers to the binary experimental conditions in the CSI experiments, such as control or experimental, each condition should have at least two FCS files;
   - The 'timepoint' indicates the sampling time in the PTI experiments, requiring at least two time points, for example, 1, 3, 5 and 7 d.

Initializing the software environment

● **TIMING 1–2 min**

3. Start ANPELA software. Click the ANPELA shortcut (installed in the 'Equipment' section). For utilizing the online platform, the users can visit https://idrblab.org/anpela/.
   ▲ **CRITICAL STEP** To enhance users' experience and to reduce the learning curve, our software offers comprehensive user tutorials. As provided in Extended Data Fig. 1, user can start a step-by-step demonstration by clicking on the 'Interactive Tutorial' panel on ANPELA home page and enable a full guidance download by clicking on the 'Textual Tutorial' panel.
   ◆ **TROUBLESHOOTING**
   (Find troubleshooting details in Table 2)
4. Upload data. Open the 'Single-cell Proteomics' panel, choose 'Upload User Data' and then specify 'Study Type' and 'Cytometry Type'. Click the 'Browse' to choose raw data files and metadata file for upload. To advance to the next stage, user can click 'Upload Data'.
   ▲ **CRITICAL STEP** When uploading FCS files, please select all files to upload at a time instead of one after another. Users can select 'Load Example Data' to learn the use of software.
5. Visualize raw data. After data upload, preview of the uploaded data is then provided on panel's right side: 'Expression of Proteins (columns) in Different Cells (rows)' and 'Stacked Density Plot for Different Samples'. User can click on 'NEXT' to proceed to the next step.

Traversing the processing workflows

● **TIMING 1–3 min**

6. Generate processing workflows. Processing cytometry-based SCP data involves a workflow comprising compensation, transformation, normalization and signal clean (provided in Box 1). By default, ANPELA automatically selects all methods that do not require additional file input (shown in Supplementary Table 1). Users can also select their preferred methods by checking the boxes in front of these methods (as shown in Extended Data Fig. 2). For the method requiring additional inputs, users must correctly upload the required files. Moreover, the parameters of each method can be adjusted by users according to their preferences.
   ▲ **CRITICAL STEP** Subsequent data processing and performance assessments will be conducted on the basis of the methods selected and parameters adjusted in this step.

7. Select protein index. Select the columns offering protein expression information. Within the uploaded FCS files, there are columns unrelated to protein expressions which do not need to be included into the analysis. It is recommended to unselect these columns to avoid potential influence. Unselected columns are excluded from the analysis, but are retained in the output data. After the column selection, user can proceed to the next step by clicking 'NEXT'.
   ◆ **TROUBLESHOOTING**
   (Find troubleshooting details in Table 2)

Navigating the data processing

● **TIMING 30 min to several hours**

8. Determine the downstream analysis method. ANPELA provided diverse analytical methods for the user to choose from based on their intended downstream analysis approaches. In CSI analysis, the users can choose their preferred clustering methods (including 'FlowSOM' and 'PhenoGraph'). In PTI analysis, users can choose their favorite trajectory inference methods (including 'slingshot_PCA', 'slingshot_tSNE', 'slingshot_diffMaps', 'scorpius_distPear', 'scorpius_distSpear' and 'scorpius_distManh').
9. Configure assessment criteria. Four independent criteria (as shown in Extended Data Fig. 2) are utilized to evaluate the performance of workflow. Each criterion has a representative metric selected as the default method, along with well-defined cutoffs. It is recommended to utilize those default metrics, unless specific requirements dictate otherwise.
   ▲ **CRITICAL STEP** Among all those four criteria, 'criterion Cd' is special for its requirements of prior knowledge. In CSI analysis, 'Cd' measures correspondence between discovered markers and prior references. In PTI analysis, 'Cd' assesses correspondence between inferred changes and prior knowledge. The details can be found in Box 2 and Supplementary Method 1.
10. Process data and assess performance. Click 'ASSESS' button to begin data processing and performance assessment. Progress bars will be displayed, indicating the ongoing progress of data processing and performance assessment. ANPELA first runs all selected workflows and generates an RData file that records outcomes. All workflows are then assessed, resulting in an overall ranking, and the processed results are accessible for any studied workflow.

Identifying the optimal workflow

● **TIMING 2–4 min**

11. Download the assessment result. Obtain the evaluating results by downloading the CSV file entitled 'Ranking_Table.csv' and the PDF file entitled 'Ranking_Figure.pdf'. As provided in Extended Data Fig. 2b, the CSV file gives the overall ranking of workflows together with the outcomes under each criterion, while the PDF file displays the ranking of workflows.
12. Identify the optimal processing workflow. Open 'Ranking_Table.csv' describing the overall ranking of all workflows. Those top-ranked workflows are considered to be well-performing in downstream analysis, which are capable of removing the impact of experimental artifacts, and, in the meantime, retaining the information of meaningful biological backgrounds.
    ▲ **CRITICAL STEP** It is recommended to utilize the processing results of the optimal processing workflow. However, at times, the workflows of slightly-worse performance may outperform the optimal one in downstream task, but those workflows of bottom-ranked performance are not recommended for use, as they may completely obscure the biological information.
13. Download processed data. The user can download the processing outcomes of a workflow or all workflows. By clicking the 'Download' button on the left side of the processing panel, an RData file that contains all processing outcomes can be freely downloaded. Alternatively, selecting certain workflow and clicking the 'Download' button on the right side will provide the access to the results of the studied workflow in the formats of both FCS and CSV (Box 4).

BOX 4

Output files of the ANPELA R package

The functions mentioned in Procedure 2 will generate output files in the directory specified by parameter 'savepath', to record the outcomes of data processing and performance assessments. Descriptions of these output files generated by different functions are provided below:

- Process(), FCprocess(), MCprocess()
  These functions produce the output file(s) to record the data processing results, with the format determined by parameter 'save_processed_res'. There are two options: **1**) 'one_folder' denotes that the processed results from different workflows will be saved into multiple RData files; and **2**) 'one_RData' indicates that all processed results will be saved into a RData file.
- Assess(), CSIassess(), PTIassess()
  These functions generate the output file 'assess.RData' to document the results of performance assessments. This RData file contains two tables, which describe exact values and performance levels under each criterion for all studied data processing workflows.
- Ranking()
  The output files generated by Ranking() are 'Ranking_Table.csv' and 'Ranking_Figure.pdf'. The 'Ranking_Table.csv' includes the overall ranking and assessing results of all workflows, where the 'Rank' column represents overall ranking, and the 'Value' columns give exact values under each criterion. 'Ranking_Figure.pdf' provides the overall rankings and performance levels and uses colors to represent performance levels: 'good' and 'poor'.
- FC_CSI(), MC_CSI(), FC_PTI(), MC_PTI()
  These functions generate all output files that produced by the functions above.
- Get_CSIres(), Get_PTIres()
  The Get_CSIres() and Get_PTIres() functions generate 'CSIres.RData' and 'PTIres.RData', respectively, to record the results of CSI and trajectory inference for downstream analysis. When the 'Plot' parameter is set to 'T', these functions also produce reduced-dimension plots ('CSI.png' for cell annotation and 'PTI.png' for trajectory inference).
- Visualize()
  The Visualize() function generates plot files that display the visualization results of specific criteria for data processed by specified workflows. The filenames are formatted to end with 'downstream task_criteria.png', for example, 'CSI_Ca.png'.
- Additional output files
  The 'log.txt' and 'info_saved.Rdata' files will be generated when using ANPELA. The 'log.txt' records specific details during the application of Assess(), CSIassess(), PTIassess() and the four one-step functions. The 'log.txt' can also provide essential information for troubleshooting, and the 'info_saved.Rdata' file records the data related to 'metadata' and 'excludedColumn'.

Procedure 2: open-source and editable R package

Preparing the input data

● **TIMING 1–3 min**

▲ **CRITICAL** Two example datasets, one for CSI analysis and another for PTI analysis, were made downloadable to show the users the required formats of input data before their application of desktop software.

1. Download example datasets. Two example datasets (each with various raw data files and one metadata file) were made downloadable from https://idrblab.org/anpela/CSI-example.zip and https://idrblab.org/anpela/PTI-example.zip for CSI and PTI, respectively. As offered in Box 2, users are advised to prepare their data by following the formats of these two datasets.
2. Prepare required files. Create a CSV file titled 'metadata.csv' and make sure it is within the same directory as those raw data files (in FCS format). This CSV file contains two columns: for the CSI analysis, two columns in the file should be labeled as 'filename' and 'condition'; for the PTI analysis, those two columns should be labeled as 'filename' and 'timepoint'. The detailed information about these columns has been offered in Steps 1–2 'Preparing the input data' part in the 'Procedure 1: user-friendly desktop software' section.

Setting the R environment

● **TIMING 10–20 min**

3. Run RStudio. For the desktop version of RStudio, the user needs to click on RStudio icon to get started. For the server version of RStudio, the user needs to log in RStudio-server via their local computer by accessing the website http://localserverIP:Port/auth-sign-in.

The 'localserverIP' and 'Port' indicate the IP address and ethernet port of accessed RStudio-server, respectively. A username and corresponding password (that is set when configuring the RStudio) are then requested for logging into RStudio-server using user's account.

4. Set working directory. Create a new folder named 'ANPELA' in the preferred directory and set this folder as the R working directory by executing the following command:

```
> setwd("Your Preferred Directory/ANPELA/")
```

▲ **CRITICAL STEP** Under R environment, a forward slash (/) is adopted to indicate the file path, unlike the CMD command in the operating system of Windows utilizing a backslash (\).

5. Position input files. Position all input files (including raw data files and metadata file) to the working directory. If the additional files for data processing (as provided in Supplementary Table 1) or the prior knowledge for the performance assessment (as described in Box 2) is available, the corresponding file(s) also should be positioned to the working directory.

6. Load ANPELA R package. Load ANPELA R package via running the following command:

```
> library(ANPELA)
```

◆ **TROUBLESHOOTING**
(Find troubleshooting details in Table 2)

7. Specify input files. Assign the working directory to a parameter 'datapath', which is adopted to denote the folder which stores all raw data files and the metadata file. In CSI analysis, the absolute path of the additional 'known marker file' used as the prior knowledge of 'criterion Cd' is assigned to the object 'DEP'. In PTI analysis, the absolute path of additional 'pathway hierarchy file' adopted as prior knowledge is assigned to the object 'pathwayhierarchy'. The assignment of working directory is accomplished by executing the following command:

```
> datapath <- "Your Preferred Directory/ANPELA/"
```

▲ **CRITICAL STEP** If the prior knowledges are available for assessing, the user can assign these knowledges using the following commands. The details are offered in Box 2. Especially, the corresponding parameters are explicitly described in Supplementary Table 2.

- For CSI analysis, execute the following command in RStudio:

```
> DEP <- "Your Preferred Directory/ANPELA/Known_Marker.csv"
```

- For PTI analysis, execute the following command in RStudio:

```
> pathwayhierarchy <- "Your Preferred Directory/ANPELA/Pathway_Hierarchy.csv"
```

8. Exclude column unrelated to protein expression. In a FCS file, there are columns unrelated to protein expression which do not need to be included into the analyses. As indicated by func1 in Supplementary Table 2, users can apply the corresponding function Getmarker() to retrieve all column names of raw data and then select those required an exclusion, such as 'Time' and 'Cell Length', and assign them to the parameter of 'excludedColumn'.

- For CSI analysis, execute the following command in RStudio:

```
> Getmarker(datapath)
> excludedColumn <- "gate_source(gate_source), cell_id(cell_id), sample_id(sample_id)"
```

- For PTI analysis, execute the following command in RStudio:

```
> Getmarker(datapath)
> excludedColumn <- "Time(Time), Cell_length(Cell_length), DNA-1(DNA.1.Ir191.Dd)"
```

▲ **CRITICAL STEP** The output of Getmarker() consists of multiple column names in the format of 'description(name)', which are separated on the basis of comma and break line. The parameter 'excludedColumn' is generated by conjugating column names using comma and space.

Traversing the processing workflows

● **TIMING 10 min to several hours**

9. Process data using parallel computing. As provided by func2 in Supplementary Table 2, the user can apply the corresponding function Process() to process those raw data. ANPELA is developed to scan a total of 3,328 workflows (1,920 are tailored for flow cytometry-based SCP and 1,792 are for mass cytometry-based SCP). By default, the Process() function is used to scan all workflows which requires no extra input. Furthermore, specific methods can also be processed and, if necessary, parameters are specified (provided in Supplementary Table 2).
   - For CSI analysis, the users can run the following command in RStudio (the 'technique' is determined by the analytical technique of studied example dataset downloaded from https://idrblab.org/anpela/CSI-example.zip; when analyzing other dataset, user should specify the corresponding technique: FC for flow cytometry and MC for mass cytometry):

     > Process(datapath = datapath, technique = "FC", studytype = "CSI", excludedColumn = excludedColumn)

   - For PTI analysis, the users can run the following command in RStudio (the 'technique' is determined by the analytical technique of studied example dataset downloaded from https://idrblab.org/anpela/PTI-example.zip; when analyzing other dataset, users should specify the corresponding technique: FC for flow cytometry, MC for mass cytometry):

     > Process(datapath = datapath, technique = "MC", studytype = "PTI", excludedColumn = excludedColumn)

     ◆ **TROUBLESHOOTING**
     (Find troubleshooting details in Table 2)
     ▲ **CRITICAL STEP** Parallel computing is utilized to extensively shorten the execution time with the elevation of computing efficiency depending on the number of parallel cores. By default, our ANPELA utilizes half of the cores of user's computer to support the parallel computing. Moreover, users are suggested to monitor memory usage and select proper number of parallel cores by setting the parameter 'cores'. As shown by func3 and func4 in Supplementary Table 2, the user can execute the corresponding functions FCprocess() and MCprocess() to process the raw data obtained from flow cytometry and mass cytometry, respectively.

Navigating the data processing

● **TIMING 2–30 min**

10. Assess performance from multiple perspectives. As shown by func5 in Supplementary Table 2, the user can use Assess() function to assess the performance of all workflows.
    - For CSI analysis, run the following command in RStudio:

      > assess_res <- Assess(respath = "./ANPELA_res", studytype = "CSI")

    - For PTI analysis, run the following command in RStudio:

      > assess_res <- Assess(respath = "./ANPELA_res", studytype = "PTI")

      ▲ **CRITICAL STEP** The Assess() function uses two distinct evaluating systems for CSI and PTI analyses, each with four criteria. In CSI analysis, the 'criterion Cd' measures correspondences between discovered markers and known ones. In PTI analysis,

BOX 5

Brief introduction to the one-step functions

One-step functions, including FC_CSI(), MC_CSI(), FC_PTI() and MC_PTI(), can accomplish data processing and performance evaluation in just one step, resembling a combination of three functions: Process(), Assess() and Ranking(). These functions are entitled based on the specific scenarios for which they are designed. Particularly, 'FC' and 'MC' mean that raw data files are acquired from flow cytometry and mass cytometry, respectively, and 'CSI' and 'PTI' represent that assessment is for CSI and PTI analyses, respectively. The user is suggested to choose the appropriate function based on their specific datasets. The applications of one-step functions are also based on Procedure 2. The user can apply the one-step functions to realize data processing and performance assessment in one step by executing the following command in RStudio:

```
> MC_PTI(datapath = datapath, excludedColumn = excludedColumn)
```

▲ **CRITICAL** Similar to Process() and Assess(), four one-step functions require additional inputs when certain data processing methods and 'criterion Cd' are involved. Four functions are shown by func9, func10, func11 and func12 in Supplementary Table 2. The output files are in Box 4.

'criterion Cd' assesses correspondence between inferred changes and prior knowledge. The details can be found in Box 2 and Supplementary Method 1. If prior knowledges are available, the user can assign them to parameter 'DEP' and 'pathwayhierarchy'. As shown by func6 and func7 in Supplementary Table 2, the user can run the corresponding functions CSIassess() and PTIassess() to evaluate.

BOX 6

Brief Introduction to the *Get_CSIres*(), *Get_PTIres*() and *Visualize*() Functions

The Get_CSIres(), Get_PTIres() functions (described by func13 and func14 in Supplementary Table 2) are designed to assist users in retrieving the results of CSI and PTI for downstream analyses. The 'marker_path' parameter specifies the data path of the file containing key protein marker information that is utilized for cell annotation. Meanwhile, the 'workflow' parameter allows users to specify the studied data processing workflow. Users can select one or multiple workflows from the 'Ranking_Table.csv' file, which is generated by ANPELA (as described in Box 4) and assign it to the 'workflow' parameter. The 'None_Arcsinh Transformation_None_None' workflow is provided here as an example:

- For CSI analysis, execute the following command in RStudio to apply Get_CSIres():

```
> Get_CSIres(respath = "./ANPELA_res",
workflow = "None_Arcsinh Transformation_None_None",
marker_path = "Your Marker Path/marker_list.xlsx", plot = T)
```

- For PTI analysis, execute the following command in RStudio to apply Get_PTIres():

```
> Get_PTIres(respath = "./ANPELA_res",
workflow = "None_Arcsinh Transformation_None_None", plot = T)
```

Visualize() enables users to generate visualization results in batch for data processed by specified workflows, consistent with the visualization outputs provided on the web server (described by func15 in Supplementary Table 2). The 'plot_metric' parameter determines which specific criteria users want to visualize. In the following code examples, 'Ca_metric' serves merely as an example of the 'plot_metric' parameter:

- For CSI analysis, the users can run the following command in RStudio:

```
> Visualize(studytype = "CSI", respath = "./ANPELA_res",
workflow = "None_Arcsinh Transformation_None_None",
plot_metric = "Ca_metric")
```

- Or PTI analysis, the users can run the following command in RStudio:

```
> Visualize(studytype = "PTI", respath = "./ANPELA_res",
workflow = "None_Arcsinh Transformation_None_None",
plot_metric = "Ca_metric")
```

Identifying the optimal workflow

● **TIMING 2–4 min**

11. Provide the overall ranking of workflows. As described by func8 in Supplementary Table 2, the user can use Ranking() to rank all workflows by executing the following command:

    > Ranking(data = assess_res)

12. Utilize the one-step function (optional). As described by func9, func10, func11 and func12 in Supplementary Table 2, users can conduct data processing, workflow assessments and overall ranking in one step. The usages of these four functions are detailed in Box 5.
13. Identify the optimal processing workflow. Obtain the assessing result by checking the CSV file entitled 'Ranking_Table.csv' and the PDF file entitled 'Ranking_Figure.pdf', which are generated by Ranking() function. As shown in Extended Data Fig. 2b, the CSV file gives the overall ranking of workflows together with the outcomes under each criterion, while the PDF file displays the ranking of workflows. The top-ranked workflows are considered to be well-performing for the downstream analysis, which can remove the impact of experimental artifacts and meanwhile retain the information of meaningful biological backgrounds.
    ▲ **CRITICAL STEP** It is recommended to utilize the processing results of the optimal processing workflow. However, at times, the workflows of slightly-worse performance may outperform the optimal one in downstream task, but those workflows of bottom-ranked performance are not recommended for use, as they may completely obscure the biological information.
14. Acquire processed data. Obtain the processed data saved in RData files in the folder of 'process_res'. ANPELA suggests to utilize the data processed using the optimal workflow. In addition, intermediate results and visualizations can be obtained through the corresponding R functions (as described in Box 6).

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting Table

Step	Problem	Possible Reason	Solution
Troubleshooting advice for Procedure 1			
3	Problem description: a warning appears in the download tool while downloading the software	The .exe format of the installation file may prompt a warning on the computer systems	Ignore the warning and proceed with the download, as the ANPELA desktop software is free from any malicious content and can be safely downloaded
3	Problem description: The interactive tutorial panel fails to start when clicked by user	(1) The initialization needs a delay (2) Frequent clicking could lead to program unresponsiveness	(1) Wait for 10 s after clicking the button (2) If problems persist, it is recommended to restart the software for solving this problem
7	Problem description: the 'NEXT' button becomes gray and cannot be clicked by user	The processing method(s) requiring additional input has been selected and await the required input	(1) Provide the additional input required (2) If unable to offer the additional input, unselect the method(s), which requires additional input
Troubleshooting advice for Procedure 2			
6	Console error message: ERROR: dependency package_name is not available for package ANPELA	The dependency package named package_name is not installed in user's R environment	Install the missing dependency packages through referring the detailed information about the installation of dependent packages in Box 3
9	Console error message: the filenames of the FCS files are inconsistent with those of metadata files	The 'filename' column in metadata file does not match the filenames of user's raw data files	Ensure that the 'filename' column of the metadata.csv matches exactly with the filenames of raw data files (excluding the .fcs extension)
9	Console error message: the number of FCS file(s) is not enough for the subsequent analysis	There are too few raw data files to meet the basic requirements for ANPELA analysis	For CSI: ensure two conditions in metadata, with each condition possesses at least two FCS files; for PTI: ensure a minimum of two time points

The problems, affiliated step, possible reasons and proposed solutions were described.

Timing

The time required for each step of Procedures 1 and 2 is described below. For specific steps, the required time will be greatly affected by the number of processing workflows executed, specific parameter settings and device configurations (CPU, memory, bandwidth and so on).

Procedure 1 for the user-friendly desktop software
Steps 1–2, preparing the input data: 1–3 min
Steps 3–5, initializing the software environment: 1–2 min
Steps 6–7, traversing the processing workflows: 1–3 min
Steps 8–10, navigating the data processing: 30 min to several hours
Steps 11–13, identifying the optimal workflow: 2–4 min

Procedure 2 for open-source and editable R package
Steps 1–2, preparing the input data: 1–3 min
Steps 3–8, setting the R environment: 10–20 min
Step 9, traversing the processing workflows: 10 min to several hours
Step 10, navigating the data processing: 2–30 min
Steps 11–14, identifying the optimal workflow: 2–4 min

Anticipated results

The output of ANPELA comprises the overall ranking of all executed processing workflows and the resulting data processed by the corresponding workflow, which are explicitly provided in the section of 'Procedure' and Box 4. Our ANPELA protocol aims at assisting user in identifying the optimal processing workflow through using an assessment-based navigation strategy, facilitating the acquisition of reliable biological insights in the cytometry-based SCP studies. In this Protocol, ten benchmark datasets (as provided in Table 1) were utilized to evaluate the performance of our proposed protocol. Particularly, eight of the ten benchmarks were systematically processed using ANPELA and their evaluating outcomes using representative workflows were provided in Table 3. In the meantime, the remaining two datasets, for which their biological insights of downstream analysis were known, were also thoroughly analyzed in the following sections as case studies.

Data processing using the assessment-based navigation strategy
Eight benchmarks were systematically processed using ANPELA and their evaluating outcomes based on representative workflows are described in Table 3. Particularly, the evaluating results of four benchmarks (SCP57021 and SCP47065 based on flow cytometry and SCP11272 and SCP37430 based on mass cytometry) performing the CSI are illustrated in Table 3a, and the measuring outcomes of additional four sets of data (SCP43132 and SCP36391 based on flow cytometry and SCP93731 and SCP96723 based on mass cytometry) performing PTI are offered in Table 3b. For each benchmark, five representative workflows are described. As shown, there was workflow always ranked as 'good' under all criteria (for instance, the first for each dataset in Table 3), and there was other workflow keeping ranked as 'poor' under all three criteria (for example, the last workflow of some datasets in Table 3). Furthermore, it was also clear that a workflow that worked well in one dataset might perform poorly in another. In other words, a workflow could give greatly different performances for different benchmarks. For instance, the '`NON+LGT+MEA+NON`' workflow performed well in conducting mass cytometry-based CSI analyses (SCP11272 and SCP37430; 'good' under all criteria), while it worked poorly in some mass cytometry-based PTI studies (SCP93731; 'poor' under all criteria). This underscored the influences of data dependency and task specificity on the performance of a workflow, emphasizing the importance of identifying suitable workflows for particular dataset.

Table 3 | Assessing outcomes based on five representative workflows for each benchmark dataset

	CSI analysis[a]					PTI analysis[b]				
Flow cytometry (FC)	Data	Workflow	Accuracy	Tightness	Robustness	Data	Workflow	Conformance	Smoothness	Robustness
	SingPro ID: SCP57021	NON+BEP+MEA+PQC	0.938 Good	0.686 Good	0.326 Good	SingPro ID: SCP43132	FLC+CLR+MMN+FAI	0.781 Good	0.994 Good	0.995 Good
		NON+BOX+MEA+FAI	0.816 Good	0.614 Good	0.293 Good		NON+SST+GSN+FCL	0.778 Good	0.928 Good	0.830 Good
		NON+SST+WPS+FCL	0.114 Poor	0.731 Good	0.739 Good		NON+LGT+MEA+NON	0.557 Poor	0.811 Good	0.747 Good
		NON+LGT+MEA+NON	0.656 Poor	0.576 Good	0.267 Good		NON+LGT+MMN+PQC	0.408 Poor	0.780 Poor	0.960 Good
		NON+ANN+MMN+PQC	0.494 Poor	0.554 Good	0.145 Poor		NON+CLR+GSN+FCL	0.413 Poor	0.780 Poor	0.496 Poor
	SingPro ID: SCP47065	FLC+FVS+MMN+PQC	0.910 Good	0.587 Good	0.417 Good	SingPro ID: SCP36391	NON+ANN+WPS+FAI	0.617 Good	1.000 Good	0.995 Good
		FLC+FVS+GSN+FCL	0.832 Good	0.629 Good	0.298 Good		NON+LGT+MEA+FAI	0.603 Good	1.000 Good	0.750 Good
		NON+TRU+WPS+FCL	0.401 Poor	0.631 Good	0.857 Good		NON+LGT+MEA+NON	0.431 Poor	0.999 Good	0.904 Good
		NON+LGT+MEA+NON	0.680 Poor	0.580 Good	0.179 Good		NON+BOX+ZSC+PQC	0.594 Poor	1.000 Good	0.459 Poor
		NON+LGT+GSN+PQC	0.661 Poor	0.605 Good	0.143 Poor		NON+HPL+MEA+FCL	0.471 Poor	0.971 Good	0.394 Poor
Mass cytometry (MC)	Data	Workflow	Accuracy	Tightness	Robustness	Data	Workflow	Conformance	Smoothness	Robustness
	SingPro ID: SCP11272	NON+LGT+MEA+PQC	0.902 Good	0.570 Good	0.365 Good	SingPro ID: SCP93731	CTS+SST+WPS+FCU	0.640 Good	0.987 Good	0.903 Good
		NON+LGT+MEA+NON	0.833 Good	0.574 Good	0.365 Good		NON+HPL+MMN+PQC	0.603 Good	0.989 Good	0.731 Good
		CTS+BOX+MEA+FAI	0.073 Poor	0.753 Good	0.611 Good		CTS+BEP+GSN+FAI	0.445 Poor	0.936 Good	0.973 Good
		CTS+HPL+NON+FAI	0.693 Poor	0.566 Good	0.299 Good		NON+BOX+WPS+FAI	0.506 Poor	0.781 Poor	0.950 Good
		CTS+QUA+MEA+PQC	0.369 Poor	0.482 Poor	0.222 Good		NON+LGT+MEA+NON	0.549 Poor	0.779 Poor	0.392 Poor
	SingPro ID: SCP37430	NON+LGT+MEA+NON	0.721 Good	0.584 Good	0.402 Good	SingPro ID: SCP96723	CTS+BOX+GSN+FCU	0.857 Good	0.999 Good	0.805 Good
		CTS+LGT+MEA+FCU	0.823 Good	0.594 Good	0.290 Good		NON+HPL+MEA+PQC	0.654 Good	1.000 Good	0.747 Good
		NON+LGT+WPS+FCU	0.684 Poor	0.565 Good	0.330 Good		NON+LGT+MEA+NON	0.518 Poor	1.000 Good	0.999 Good
		CTS+SST+MEA+FCU	0.698 Poor	0.581 Good	0.277 Good		NON+QUA+ZSC+PQC	0.506 Poor	0.961 Good	0.794 Good
		CTS+ACS+MMN+PQC	0.672 Poor	0.585 Good	0.144 Poor		CTS+QUA+ZSC+FAI	0.533 Poor	0.965 Good	0.357 Poor

[a]Assessing results of four benchmarks derived from CSI analysis, including two flow cytometry-based and two mass cytometry-based datasets. [b]Assessing results of four datasets derived from PTI analysis, including two flow cytometry-based and two mass cytometry-based datasets. The workflows were represented using the combination of the abbreviations of the methods (shown in Supplementary Table 1). The number under each criterion indicated the exact values of the assessing result, which were used to classify the workflow performance into 'good' and 'poor' as discussed in Supplementary Method 1.

Meanwhile, it also highlighted the critical value of scanning a large number of workflows, before choosing which one a user might select to use for any studied dataset.

Navigating data processing for CSI

To demonstrate ANPELA's ability to navigate the data processing in CSI analysis, a benchmark[65] offering mass cytometry-based SCP data (SCP80719 in Table 1) collected from 36 patients with coronavirus disease 2019 (COVID-19) and 45 healthy individuals was analyzed, and all workflows were systematically assessed. As a result, 'NON+LGT+MEA+FCU', 'CTS+ARN+MMN+NON' and 'NON+FVS+GSN+FCU' were ranked to the 1st, 10th and 50th by ANPELA. As described in Fig. 4a, left, the cell dimensionality reduction plots were illustrated for these workflows, and the plots were then colored on the basis of the experimentally validated cell annotations (indicated as 'True Label' in this study). As shown, the cell subpopulations identified by the 1st-ranked NON+LGT+MEA+FCU could largely differentiate the experimentally validated annotations and that by the 10th- and 50th-ranked workflows also gave certain level of capacity in clustering different colors. However, the discriminations among cell subpopulations by the 10th- and 50th-ranked workflows were poorer compared with that by the 1st-ranked one, as there were much clearer boarders among the experimentally validated cell annotations (shown by different colors) generated by the 1st-ranked workflow compared with that produced by the 10th- and 50th-ranked one.

To quantitatively assess the performance of three workflows in identifying cell subpopulation, the bubble-Sankey plots are given in Fig. 4a, middle. As demonstrated, the circles colored in blue and red were used to indicate the correct and wrong identifications, respectively, and the diameters of all circles in one row (a row denoted an experimentally validated annotation) represented the relative cell proportions. A quantitative comparison

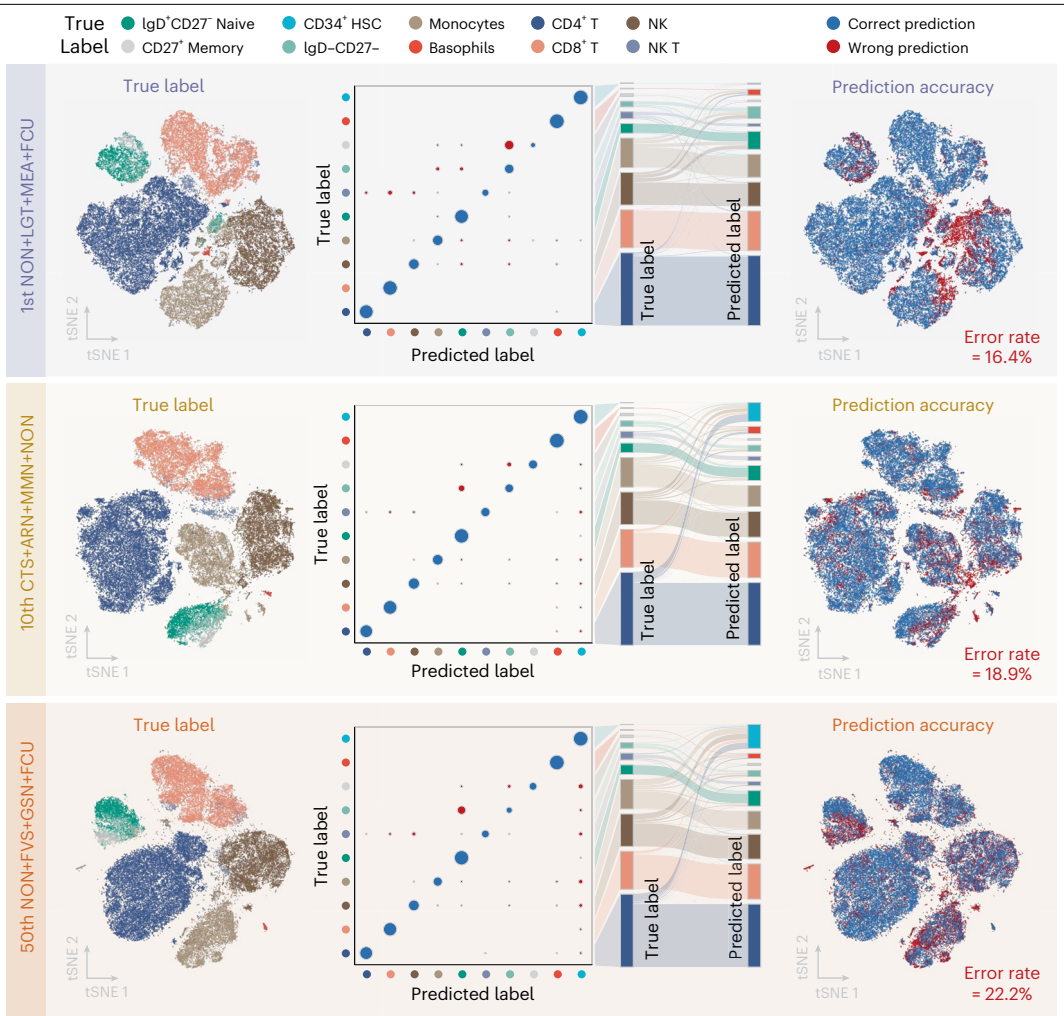


Fig. 4 | Navigating data processing for CSI using ANPELA. The performances of the 1st-ranked 'NON+LGT+MEA+FCU', 10th-ranked 'CTS+ARN+MMN+NON' and the 50th-ranked 'NON+FVS+GSN+FCU' workflows discovered by ANPELA were shown. Left: the cell dimensionality reduction plots are illustrated for these workflows, and the plots were then colored using experimentally validated cell annotations (denoted as 'True Label'). The discriminations among cell subpopulations by the 10th- and 50th-ranked workflows were found slightly poorer compared with that by the 1st-ranked one, as there were much clearer boarders among the experimentally validated cell annotations (denoted in different colors) generated by the 1st-ranked workflow compared with that produced by the 10th- and 50th-ranked one. Middle: the bubble-Sankey plots are depicted. The circles colored in blue and red are used to denote the correct and wrong identifications, respectively, and the circle diameters in one row (a row indicated an experimentally validated annotation) represents the relative cell proportion. It was clear that the 1st-ranked workflow gave fewer red circles and less misclassifications between 'True Label' and 'Predicted Label' compared with the 10th- and 50th-ranked ones. Right: the correct- and wrong-classifications of cells are denoted using blue and red dots, respectively, which resulted in distinct error rates (16.4%, 18.9% and 22.2%) for the 1st-, 10th- and 50th-ranked workflows. NK, natural killer cell; HSC, hematopoietic stem cell.

between the 'True Label' and the 'Predicted Label' was also described. It was apparent that the 1st-ranked workflow gave less RED circles together with less misclassifications between the 'True Label' and the 'Predicted Label' compared with the 10th- and 50th-ranked workflow. As shown on the right section of Fig. 4a, the correct and wrong classifications of cells were further indicated using blue and red dots, respectively, which resulted in a gradual increase in the error rate from 16.4% (1st-ranked workflow) to 18.9% (10th-ranked workflow) and then to 22.2% (50th-ranked workflow). In sum, the result showed that the 1st-ranked workflow could give a better performance than the 10th- and 50th-ranked ones, not to mention significantly better than thousands of the bottom-ranked workflows (as depicted in Supplementary Figs. 1a and 2a). Moreover, to have comprehensive understanding of how performance changes along with ranking results, the performances of top- and bottom-ranked workflows were systematically

analyzed in Supplementary Fig. 1a. As shown, the error rate tended to increase with the decrease of workflow ranking, and significant differences (*P* value <0.05) between error rates of top-ranked (colored in green) and bottom-ranked (colored in red) workflows were identified. Furthermore, an in-depth analysis on the detail error rates of the top-ten ranked workflows (illustrated in Supplementary Fig. 1b) revealed that while all top-ten ranked workflows generally produce good outcome, the error rates do not strictly follow a monotonic trend but rather fluctuate within relatively narrow range.

Moreover, it is noteworthy that the top-ranked workflow is not necessarily the one with the lowest error rate. As shown in Supplementary Fig. 1, the error rates tend to exhibit a fluctuating upward trend rather than a strictly monotonic pattern. In this analysis, for instance, the workflows ranked 8th and 20th achieved lower error rates (15.9% and 12.3%, respectively) than the 1st-ranked workflow. However, it is important to emphasize that ANPELA ranks workflows on the basis of multiple criteria, including not only accuracy but also robustness and other criteria. Although the 20th-ranked workflow achieved the lowest error rate across all workflows, its robustness was found poor performing compared with that of the 1st-ranked workflow. In other words, the first-ranked workflow could obtain an optimal performance across multiple perspectives of data processing. Based on our analysis, the workflow with the lowest error rate may not necessarily be ranked first by ANPELA, but it is typically among the top-ranked candidates. If the user wants to identify the workflow of the lowest error rate, we would suggest them to conduct comprehensive analyses across all workflows, which can be accomplished using ANPELA.

Navigating data processing for PTI

To demonstrate ANPELA's ability to navigate the data processing in PTI analysis, a benchmark[22] of cytometry-based SCP data (SCP77365 in Table 1) showing cell development was studied, and workflows were assessed. As a result, '`NON+FVS+NON+FAI`', '`NON+BOX+NON+NON`' and '`CTS+ARN+ZSC+FCU`' were ranked to the 1st, 10th and 50th by ANPELA, respectively. As reported, pseudo-time trajectories can assume various topologies, including linear, bifurcating, tree-like, cyclic or disconnected graphs, depending on the underlying biological context[66]. As shown in Fig. 5a, the mouse embryonic stem cells (mESCs) were reported to differentiate into three germ layers (mesoderm, ectoderm and endoderm). Accordingly, the inferred trajectories in this dataset were expected to exhibit three or more branches.

A comparison of the abilities of three workflows (1st-ranked, 10th-ranked and 50th-ranked) to infer real time trajectory is depicted in Fig. 5b. As offered on the top section, the PTI scatter plots were illustrated for both workflows, and these plots were colored on the basis of the sampling times of cells (from day 0 to day 11). As shown, the pseudo-time trajectory inferred by the 1st-ranked `NON+FVS+NON+FAI` could accurately reproduce the real time trajectory, and 10th-ranked one partially aligned with the real trajectory, whereas the one inferred by the 50th-ranked workflow largely deviated from the real development direction of mESCs shown in Fig. 5a. To quantitatively assess the performance of workflows, the Sankey plots were given in the middle of Fig. 5b. As provided, a quantitative comparison between the 'True Label' and the 'Predicted Label' was described. The correct- and wrong-assignments of cells were illustrated by green and red bars, respectively, which led to distinct error rates (10.3%, 12.2% and 59.0%) for the 1st-, 10th- and 50th-ranked workflows.

To quantitatively assess the ability of three workflows in identifying cell fate, the PTI scatter plots are further drawn at the bottom of Fig. 5b. As shown, the cells of different types (mESCs, mesoderm, ectoderm and endoderm) are highlighted by light red, purple, green and blue, respectively. It was apparent that the first-ranked workflow not only successfully reproduced the developmental direction of *mESCs* but also effectively captured cells' differentiation into three germ layers. By contrast, the tenth-ranked workflow failed to identify one of the endpoints, which might come from the difficulty in discriminating mesoderm (colored in PURPLE) from ectoderm (colored in GREEN). And the 50th-ranked workflow completely obscured the biological insights of this studied process. All in all, above findings showed that the 1st-ranked workflow could produce much better performance than the 10th- and 50th-ranked one, not to mention significantly better than thousands of the bottom-ranked workflows (as explicitly illustrated in Supplementary Fig. 2b).

To further explore how trajectory inference performance varies with ANPELA ranking, we conducted additional analyses. As shown in Supplementary Fig. 4a, the error rate of inferred trajectories generally increased with ranking, and top-ranked workflows generally yielded significantly better results compared to bottom-ranked ones. Supplementary Fig. 4b illustrates

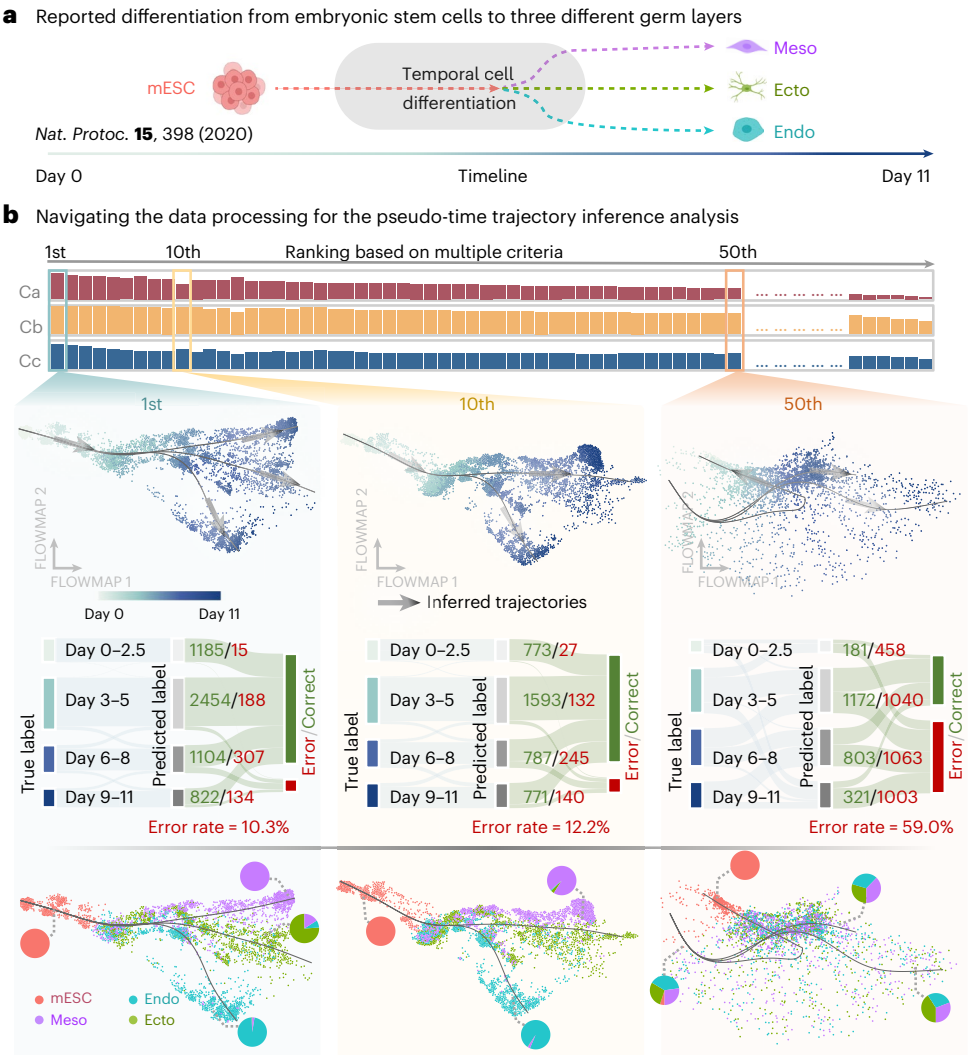


Fig. 5 | Navigating data processing for PTI Using ANPELA. a, Reported cell differentiations of mESCs. The mESCs were reported to differentiate into three germ layers (Meso, mesoderm; Ecto, ectoderm; Endo, endoderm) from day 0 to day 11 (ref. 22). **b**, The performances of 1st-ranked '`NON+FVS+NON+FAI`', 10th-ranked '`NON+BOX+NON+NON`' and 50th-ranked '`CTS+ARN+ZSC+FCU`' workflows in PTI analysis using ANPELA. Top: the PTI scatter plots are colored on the basis of sampling times of cells (from day 0 to 11). The pseudo-time trajectory inferred by the 1st-ranked workflows accurately reproduced the real trajectory, and 10th-ranked one partially aligned with the real trajectory, whereas the one inferred by the 50th-ranked workflow largely deviated from the real development direction of mESC as described in Fig. 5a. Middle: Sankey plots were drawn to provide a quantitative comparison between the 'True Label' and the 'Predicted Label', and the correct and wrong assignments of cells are shown by green and red bars, respectively, which resulted in distinct error rates (10.3%, 12.2% and 59.0%) for the 1st-, 10th- and 50th-ranked workflows. Due to the possibility that a cell may be assigned to a different number of trajectory branches by Slingshot across various inference results, the number of 'True Labels' and 'Predicted Labels' could vary between workflows. Bottom: the PTI scatter plots are drawn, and the cells of different types (mESCs, Meso, Ecto and Endo) are highlighted by light red, purple, green and blue, respectively. Pie charts are included to illustrate the relative proportions of each cell type within the starting and ending clusters, with detailed information shown in Supplementary Fig. 3. It was clear that the 1st-ranked workflow could not only successfully reproduce the developmental direction of mESCs but also effectively capture cells' differentiation into three germ layers. By contrast, the tenth-ranked workflow failed to identify one of the endpoints, and the 50th-ranked workflow completely obscured biological insights.

the detailed error rates for the top-ten workflows, most of which attained relatively low error rates (with only the ninth-ranked workflow exceeding 20%) and exhibited a fluctuating but overall increasing trend in error rates with ranking. This pattern aligned with the results observed in the CSI analysis, reinforcing that ANPELA effectively prioritized high-performing processing workflows, with performance generally declining at lower rankings.

Moreover, an interesting observation in this benchmark dataset is that the ectoderm is positioned between the mesoderm and endoderm in the dimensionality reduction results of both the first and tenth-ranked workflows. This may appear counterintuitive to researchers, as it does not align with the relative anatomical positions during embryonic development. To investigate this further, we reviewed the original publication from which the dataset was derived[22]. The findings reported in that study were consistent with ours, showing the ectoderm situated between the mesoderm and endoderm. In our opinion, although the mesoderm is anatomically located between the ectoderm and endoderm during embryogenesis, the spatial arrangement in dimensionality reduction plots reflects similarities in gene expression profiles rather than physical spatial relationships. The fact that both the top-ranked and tenth-ranked workflows placed the ectoderm between the mesoderm and endoderm may indicate that the gene expression profile of the endoderm is more closely related to that of the ectoderm than to the mesoderm. A literature review supported that certain genes (for example, *KLF7*) exhibited more similar expression patterns between the endoderm and ectoderm compared with mesoderm[67]. However, to the best of our knowledge, no existing study has definitively explained why the ectoderm is positioned between the mesoderm and endoderm in such analyses, leaving this as an open question for future investigation.

Exploring method performances based on multiple benchmarks

The identification of high- or low-performing methods across multiple datasets is indeed of great practical relevance. To identify high- or low-performing workflows, an empirical analysis of ten benchmarks (five each for CSI and PTI, as offered in Table 1) were therefore conducted. In each benchmark, the top-50 ranked workflows were defined here as high-performing (hereafter 'top-workflows') and the bottom-50 as low-performing ('bottom-workflows').

For the compensation step, it is known that most compensation methods need the extra spillover information[68], but only two out of the ten datasets studied in this analysis provided their additional spillover data, which thus greatly limited the application of most compensation methods. In other words, the frequency of methods in top/bottom-workflows (for the CSI and PTI) are dramatically affected by (highly dependent on) the analyzed dataset. For instance, the `SPR` (a compensation method that requires additional spillover of single-stained beads controls) could not be evaluated due to the absence of required spillover data in all five benchmarks, resulting its absences in both bottom- and top-workflows, while the `CTS` (does not require additional input) frequently appears in both bottom-workflows (30%) and top-workflows (26%) in the CSI studies. Similar trend was observed in the PTI analyses. Due to the dependency of method's frequencies in top- and bottom-workflows, it is difficult to discover the methods that are high/low-performing, but the ones that do not require spillover data were found frequently appearing in bottom- and top-workflows.

For the transformation step, in CSI analysis (shown in Supplementary Fig. 5), `ARN` appeared most frequently in the top-workflow (23%), while `QUA` was among the most frequently observed one in bottom-workflow (14%). Similar results were found for PTI (as shown in Supplementary Fig. 6), `ARN` was again the most popular one for top-workflow (10%), while `QUA` remained the most frequent one in bottom-workflows (10%), although the advantage of `ARN` in PTI was less pronounced than in CSI. For the normalization step, in CSI analyses (shown in Supplementary Fig. 5), `MMN` was the most frequent one in top-workflows (24%), while `MEA` was most observed in bottom-workflows (35%). In PTI study (illustrated in Supplementary Fig. 6), `GSN` showed most often in top-workflows (30%), while `MMN` was found to be the most frequent one in bottom-workflows (38%). For the signal clean step, in CSI studies (Supplementary Fig. 5), `NON` gave the highest frequency in top-workflows (33%), while `PQC` had the highest frequency for bottom-workflows (34%). In PTI analysis (Supplementary Fig. 6), `FAI` and `PQC` topped in top- (30%) and bottom-workflows (30%), respectively. Overall, the empirical analyses above revealed some

methods that are high/low performing. The most notable ones among them include ARN and QUA transformation methods. The ARN frequently exhibited good performances in both CSI and PTI tasks, while the QUA frequently demonstrated poor performances in both CSI and PTI tasks. This result aligned with literature survey showing that ARN was recommended by the well-established software PhenoGraph[17] as a default setting, while QUA has been seldom used in cytometry-based SCP analysis for the past 5 years. We also observed that high/low-performing methods varied between CSI and PTI studies, reflecting data- and task-dependent nature of processing outcome[27]. This finding underscored the importance of providing the user with dataset-dependent processing workflow recommendation, which is the core motivation for developing ANPELA protocol.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets that were analyzed within this protocol had been made downloadable on the website https://idrblab.org/anpela/ANPELA_exampledata.zip. These datasets were also accessible in the SingPro database (https://idrblab.org/singpro/) through IDs SCP57021, SCP11272, SCP43132, SCP77365, SCP80719, SCP47065, SCP37430, SCP36391, SCP96723 and SCP93731.

Code availability

All source codes of this protocol are available for use under a GPL v3 license and can be acquired via GitHub at https://github.com/idrblab/ANPELA. ANPELA web platform is freely available for academic purposes at https://idrblab.org/anpela. ANPELA desktop software is available for academic use at https://idrblab.org/anpela/ANPELA-Setup.exe.

Received: 13 January 2025; Accepted: 30 July 2025;
Published online: 16 October 2025

References

1. Doerr, A. Single-cell proteomics. *Nat. Methods* **16**, 20 (2019).
2. Vistain, L. F. & Tay, S. Single-cell proteomics. *Trends Biochem. Sci.* **46**, 661–672 (2021).
3. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* **20**, 363–374 (2023).
4. Hartmann, F. J. & Bendall, S. C. Immune monitoring using mass cytometry and related high-dimensional imaging approaches. *Nat. Rev. Rheumatol.* **16**, 87–99 (2020).
5. Labib, M. & Kelley, S. O. Single-cell analysis targeting the proteome. *Nat. Rev. Chem.* **4**, 143–158 (2020).
6. Su, Y., Shi, Q. & Wei, W. Single cell proteomics in biomedicine: high-dimensional data acquisition, visualization, and analysis. *Proteomics* **17**, 1600267 (2017).
7. Kanno, H. et al. High-throughput fluorescence lifetime imaging flow cytometry. *Nat. Commun.* **15**, 7376 (2024).
8. Fürstenau, M. et al. High resolution assessment of minimal residual disease (MRD) by next-generation sequencing (NGS) and high-sensitivity flow cytometry (hsFCM) in the phase 3 GAIA (CLL13) trial. *Blood* **138**, 72 (2021).
9. Lian, X. et al. SingPro: a knowledge base providing single-cell proteomic data. *Nucleic Acids Res.* **52**, D552–D561 (2024).
10. Hutton, C. et al. Single-cell analysis defines a pancreatic fibroblast lineage that supports anti-tumor immunity. *Cancer Cell* **39**, 1227–1244 (2021).
11. Arnett, L. P. et al. Reagents for mass cytometry. *Chem. Rev.* **123**, 1166–1205 (2023).
12. Bandura, D. R. et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
13. Liechti, T. et al. An updated guide for the perplexed: cytometry in the high-dimensional era. *Nat. Immunol.* **22**, 1190–1197 (2021).
14. Kröger, C. et al. Unveiling the power of high-dimensional cytometry data with cyCONDOR. *Nat. Commun.* **15**, 10702 (2024).
15. Roca, C. P. et al. AutoSpill is a principled framework that simplifies the analysis of multichromatic flow cytometry data. *Nat. Commun.* **12**, 2890 (2021).
16. Chevrier, S. et al. Compensation of signal spillover in suspension and imaging mass cytometry. *Cell Syst.* **6**, 612–620 (2018).
17. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
18. Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165**, 780–791 (2016).
19. Van Gassen, S. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. A* **87**, 636–645 (2015).
20. Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell* **16**, 323–337 (2015).
21. Quintelier, K. et al. Analyzing high-dimensional cytometry data using FlowSOM. *Nat. Protoc.* **16**, 3775–3801 (2021).
22. Ko, M. E. et al. FLOW-MAP: a graph-based, force-directed layout algorithm for trajectory mapping in single-cell time course datasets. *Nat. Protoc.* **15**, 398–420 (2020).
23. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
24. Liu, X. et al. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol.* **20**, 297 (2019).
25. Brooks, T. G., Lahens, N. F., Mrčela, A. & Grant, G. R. Challenges and best practices in omics benchmarking. *Nat. Rev. Genet.* **25**, 326–339 (2024).
26. Aghaeepour, N. et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–238 (2013).
27. Zhang, Y., Sun, H., Lian, X., Tang, J. & Zhu, F. ANPELA: significantly enhanced quantification tool for cytometry-based single-cell proteomics. *Adv. Sci.* **10**, e2207061 (2023).

28. Navarro, P. et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
29. Tang, J. et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief. Bioinform.* **21**, 621–636 (2020).
30. Tang, J. et al. Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell Proteom.* **18**, 1683–1699 (2019).
31. Liu, P. et al. Comprehensive evaluation and practical guideline of gating methods for high-dimensional cytometry data: manual gating, unsupervised clustering, and auto-gating. *Brief. Bioinform.* **26**, bbae633 (2024).
32. Keeler, A. B. et al. A developmental atlas of somatosensory diversification and maturation in the dorsal root ganglia by single-cell mass cytometry. *Nat. Neurosci.* **25**, 1543–1558 (2022).
33. Lammel, D. R., Meierhofer, D., Johnston, P., Mbedi, S. & Rillig, M. C. The effects of arbuscular mycorrhizal fungi (AMF) and *Rhizophagus irregularis* in soil microorganisms accessed by metatranscriptomics and metaproteomics. Preprint at *bioRxiv* https://doi.org/10.1101/860932 (2019).
34. Jurburg, S. D. et al. The community ecology perspective of omics data. *Microbiome* **10**, 225 (2022).
35. Shen, S. et al. High-quality and robust protein quantification in large clinical/pharmaceutical cohorts with IonStar proteomics investigation. *Nat. Protoc.* **18**, 700–731 (2023).
36. Wang, S. et al. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res.* **48**, e83 (2020).
37. Cui, X. et al. Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics. *Front. Pharmacol.* **10**, 127 (2019).
38. Islam, M. A., Majumder, M. Z. H., Miah, M. S. & Jannaty, S. Precision healthcare: a deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction. *Comput. Biol. Med.* **176**, 108432 (2024).
39. Andersen, T. O., Kunath, B. J., Hagen, L. H., Arntzen, M. & Pope, P. B. Rumen metaproteomics: closer to linking rumen microbial function to animal productivity traits. *Methods* **186**, 42–51 (2021).
40. Zhang, T. et al. Block design with common reference samples enables robust large-scale label-free quantitative proteome profiling. *J. Proteome Res.* **19**, 2863–2872 (2020).
41. Louta, M., Banti, K. & Karampelia, I. Emerging technologies for sustainable agriculture: the power of humans and the way ahead. *IEEE Access* **12**, 98492–98529 (2024).
42. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
43. Mund, A., Brunner, A. D. & Mann, M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol. Cell* **82**, 2335–2349 (2022).
44. Chang, Q. et al. Imaging mass cytometry. *Cytom. A* **91**, 160–169 (2017).
45. Rius Rigau, A. et al. Characterization of vascular niche in systemic sclerosis by spatial proteomics. *Circ. Res.* **134**, 875–891 (2024).
46. Phongpreecha, T. et al. Single-cell peripheral immunoprofiling of Alzheimer's and Parkinson's diseases. *Sci. Adv.* **6**, eabd5575 (2020).
47. Lee, H. C., Kosoy, R., Becker, C. E., Dudley, J. T. & Kidd, B. A. Automated cell type discovery and classification through knowledge transfer. *Bioinformatics* **33**, 1689–1695 (2017).
48. Somol, P. & Novovicová, J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1921–1939 (2010).
49. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
50. Ahmed, S., Rattray, M. & Boukouvalas, A. GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* **35**, 47–54 (2019).
51. Qiu, P. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
52. Verrou, K. M., Tsamardinos, I. & Papoutsoglou, G. Learning pathway dynamics from single-cell proteomic data: a comparative study. *Cytom. A* **97**, 241–252 (2020).
53. Chen, H. et al. Cytofkit: a bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS Comput. Biol.* **12**, e1005112 (2016).
54. Monaco, G. et al. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* **32**, 2473–2480 (2016).
55. Fletez-Brant, K., Špidlen, J., Brinkman, R. R., Roederer, M. & Chattopadhyay, P. K. flowClean: automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytom. A* **89**, 461–471 (2016).
56. Meskas, J., Yokosawa, D., Wang, S., Segat, G. C. & Brinkman, R. R. flowCut: an R package for automated removal of outlier events and flagging of files based on time versus fluorescence analysis. *Cytom. A* **103**, 71–81 (2023).
57. Hahne, F. et al. Per-channel basis normalization methods for flow cytometry data. *Cytom. A* **77**, 121–131 (2010).
58. Finak, G., Perez, J. M., Weng, A. & Gottardo, R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinform.* **11**, 546 (2010).
59. Azad, A., Rajwa, B. & Pothen, A. flowVS: channel-specific variance stabilization in flow cytometry. *BMC Bioinform.* **17**, 291 (2016).
60. Emmaneel, A. et al. PeacoQC: peak-based selection of high quality cytometry data. *Cytom. A* **101**, 325–338 (2022).
61. Guazzini, M., Reisach, A. G., Weichwald, S. & Seiler, C. spillR: spillover compensation in mass cytometry data. *Bioinformatics* **40**, btae337 (2024).
62. Chen, T. J. & Kotecha, N. Cytobank: providing an analytics platform for community cytometry data analysis and collaboration. *Curr. Top. Microbiol. Immunol.* **377**, 127–157 (2014).
63. Qian, Y. et al. FCSTrans: an open source software system for FCS file conversion and data transformation. *Cytom. A* **81**, 353–356 (2012).
64. Hahne, F. et al. flowCore: a bioconductor package for high throughput flow cytometry. *BMC Bioinform.* **10**, 106 (2009).
65. Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210–1220 (2020).
66. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
67. Cui, G. et al. Spatial molecular anatomy of germ layers in the gastrulating cynomolgus monkey embryo. *Cell Rep.* **40**, 111285 (2022).
68. Fu, J. et al. Optimization of metabolomic data processing using NOREVA. *Nat. Protoc.* **17**, 129–151 (2022).
69. Ingelfinger, F. et al. Single-cell profiling of myasthenia gravis identifies a pathogenic T cell signature. *Acta Neuropathol.* **141**, 901–915 (2021).
70. Candia, J. et al. From cellular characteristics to disease diagnosis: uncovering phenotypes with supercells. *PLoS Comput. Biol.* **9**, e1003215 (2013).
71. Hartmann, F. J. et al. Comprehensive immune monitoring of clinical trials to advance human immunotherapy. *Cell Rep.* **28**, 819–831 (2019).
72. Suwandi, J. S. et al. Multidimensional analyses of proinsulin peptide-specific regulatory T cells induced by tolerogenic dendritic cells. *J. Autoimmun.* **107**, 102361 (2020).
73. Dai, Y. et al. CytoTree: an R/Bioconductor package for analysis and visualization of flow and mass cytometry data. *BMC Bioinform.* **22**, 138 (2021).
74. Barone, S. M. et al. Unsupervised machine learning reveals key immune cell subsets in COVID-19, rhinovirus infection, and cancer therapy. *eLife* **10**, e64653 (2021).
75. Bodenmiller, B. et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.* **30**, 858–867 (2012).
76. Gaudillière, B. et al. Clinical recovery from surgery correlates with single-cell immune signatures. *Sci. Transl. Med.* **6**, 255ra131 (2014).
77. Bagwell, C. B. & Adams, E. G. Fluorescence spectral overlap compensation for any number of flow cytometry parameters. *Ann. NY Acad. Sci.* **677**, 167–184 (1993).
78. Folcarelli, R. et al. Transformation of multicolour flow cytometry data with OTflow prevents misleading multivariate analysis results and incorrect immunological conclusions. *Cytom. A* **101**, 72–85 (2022).
79. den Braanker, H., Bongenaar, M. & Lubberts, E. How to prepare spectral flow cytometry datasets for high dimensional data analysis: a practical workflow. *Front. Immunol.* **12**, 768113 (2021).
80. Weber, L. M., Nowicka, M., Soneson, C. & Robinson, M. D. diffcyt: differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun. Biol.* **2**, 183 (2019).
81. Sibbertsen, F. et al. Phenotypic analysis of the pediatric immune response to SARS-CoV-2 by flow cytometry. *Cytom. A* **101**, 220–227 (2022).
82. Wang, S. & Brinkman, R. R. Data-driven flow cytometry analysis. *Methods Mol. Biol.* **1989**, 245–265 (2019).

Acknowledgements

We acknowledge the National Natural Science Foundation of China (grant nos. 22220102001, 82373790, and 82404511); Natural Science Foundation of Zhejiang (grant no. RG25H300001); National Key R&D Programs of China (grant no. 2024YFA1307503); Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

Author contributions

F.Z. conceived the idea and designed the entire research. H.C.S., Y. Zhou., R.Y.J. and Y.X.L. wrote codes. H.C.S., Y. Zhou., C.B.G. and Z.Q.P. conducted benchmark studies. H.C.S., Y. Zhou., M.J.M., X.C.L., B.H.C., T.L.N., Y. Zhang., Y.T.Z., X.N.S., H.Y., X.S., W.Q.X and B.L.Z. finished statistical analysis. Y.B.D., J.N.D., S.Q.L., T.T.F., Y. Zhang., M.X. and Q.X.Y. visualized the results. F.Z. and T.T.F. wrote the manuscript.

Competing interests

The authors declare no competing interests.

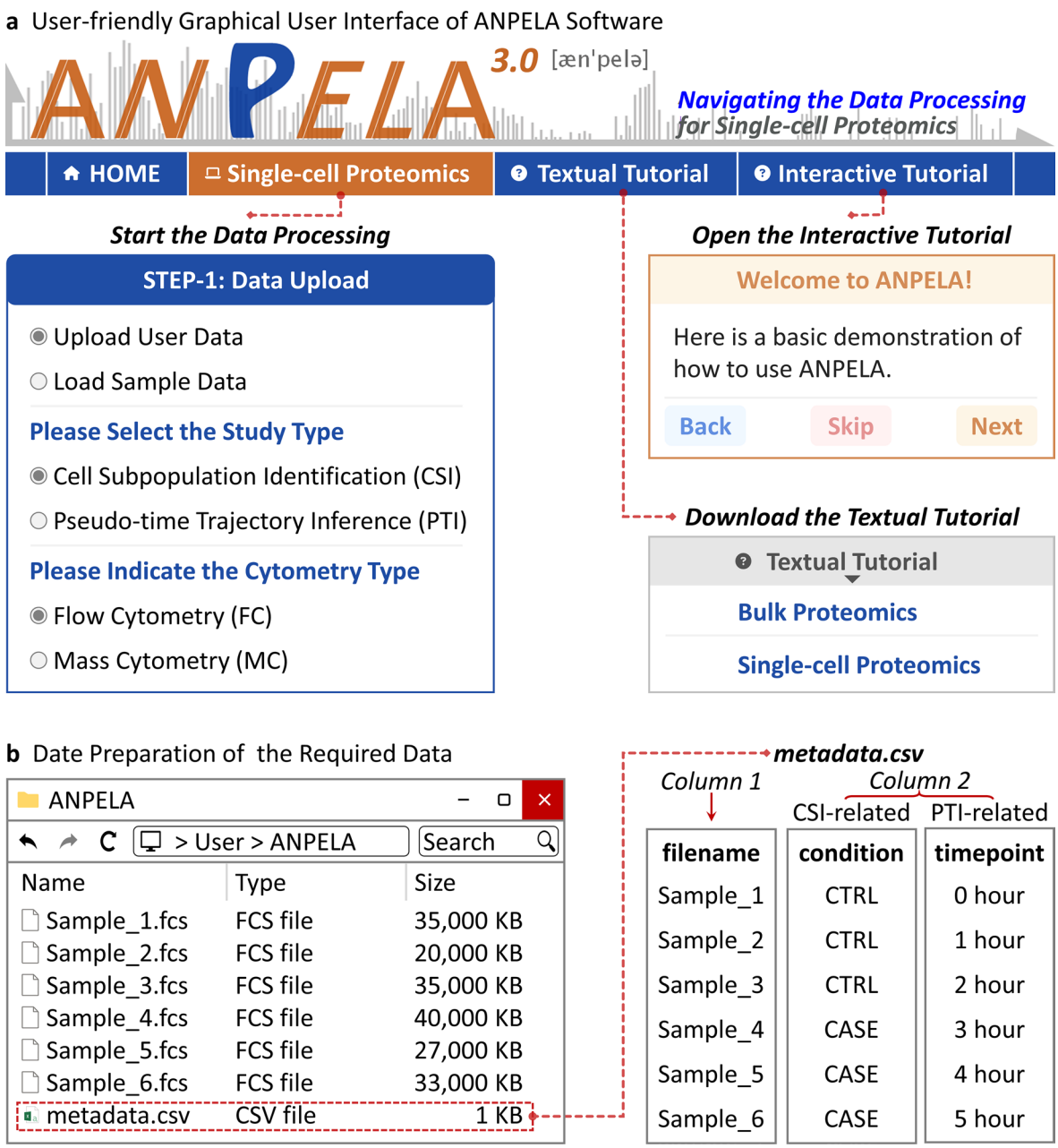
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41596-025-01257-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-025-01257-2.

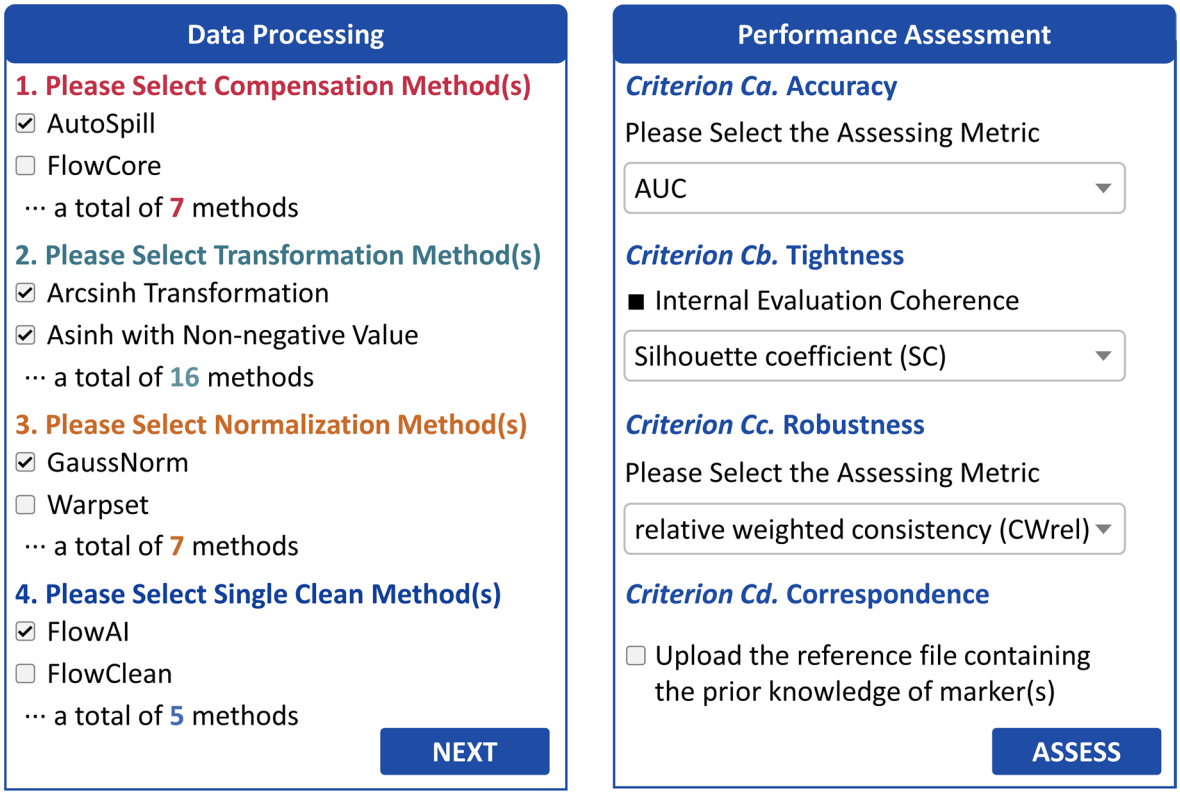
Correspondence and requests for materials should be addressed to Tingting Fu or Feng Zhu.

Peer review information *Nature Protocols* thanks Florian Ingelfinger and the other, anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Graphical User Interface of ANPELA and Preparation of Required Data. (***a***) The navigation bar of ANPELA software included 'HOME', 'Single-cell Proteomics', 'Textual Tutorial', and 'Interactive Tutorial'. Clicking on 'Single-cell Proteomics' initiated data upload and processing. Clicking on 'Textual Tutorial' permitted downloading the textual tutorial. Clicking on 'Interactive Tutorial' opened a step-by-step interactive tutorial. (***b***) The essential data required by ANPELA included FCS files (i.e., raw data files) generated from cytometry-based SCP experiments and a metadata file describing the correlation between the raw data and experimental conditions. The metadata file was a user-created file named 'metadata.csv', containing key information in two columns.

a Interface of Data Processing and Performance Assessment

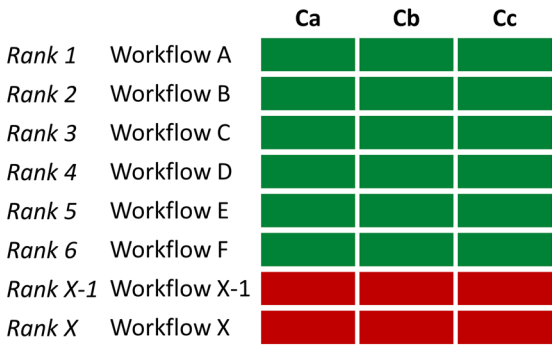


b Outcomes of Performance Assessment

Ranking_Table.csv

	Rank	Ca	Cb	Cc
Workflow A	1	0.937	0.857	0.586
Workflow B	2	0.915	0.843	0.503
Workflow C	3	0.940	0.671	0.429
Workflow D	4	0.793	0.791	0.332
Workflow E	5	0.719	0.728	0.315
...		...		
Workflow X-1	N-1	0.524	0.312	0.149
Workflow X	N	0.519	0.286	0.121

Ranking_Figure.pdf



Extended Data Fig. 2 | GUI of Processing & Assessment and Outcomes of the Assessment. (***a***) Simplified GUI for data processing and performance assessment in the desktop software of ANPELA. (***b***) Results of performance assessment consist of 'Ranking_Table.csv' and 'Ranking_Figure.pdf', which respectively recorded criteria values and performance levels for all executed data processing workflows.

nature portfolio

Corresponding author(s): Feng Zhu

Last updated by author(s): Jul 1, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data used in this study was downloaded from public sources. All data were processed using R script.
Data analysis	Data analysis used R v4.4.1; ANPELA v1.0.0; data.table v1.17.0; doParallel v1.0.17; foreach v1.5.2; ggplot2 v3.5.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets that were analyzed within this protocol had been made downloadable on the website: https://idrblab.org/anpela/ANPELA_exampledata.zip. These datasets were also accessible in the SingPro database (https://idrblab.org/singpro/) through IDs: SCP57021, SCP11272, SCP43132, SCP77365, SCP80719, SCP47065, SCP37430, SCP36391, SCP96723, & SCP93731.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available data were used for benchmark.
Data exclusions	N/A
Replication	Code and method details were carefully checked for completeness and replicability.
Randomization	The random seeds have been fixed in the code, ensuring that the analysis results are not affected by randomness.
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A