Manuscript Title

This manuscript (permalink) was automatically generated from AlexsLemonade/ScPCA-manuscript@231954c on March 1, 2024.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

Introduction

Since the introduction of single-cell RNA-seq technology, the number of studies that utilize single-cell RNA-seq has grown rapidly[1]. Unlike its predecessor, bulk RNA-seq, which averages the profiles of all cells within a sample, single-cell technology quantifies gene expression in individual cells. Tumors are known to be transcriptionally heterogeneous, so many studies have highlighted the importance of using single-cell RNA-seq in studying tumor samples [2]. Researchers can use tumor single-cell RNA-seq to analyze and identify individual cell populations that may play important roles in tumor growth, resistance, and metastasis [3]. Additionally, single-cell RNA-seq data provides insight into how tumor cells may be interacting with normal cells in the tumor microenvironment[4].

With the growing number of single-cell RNA-seq datasets, efforts have emerged to create central, harmonized sources for datasets. Harmonized data resources allow researchers to leverage more samples from various biological contexts to complete their analysis and elucidate previously unknown similarities across samples and disease types. The Human Cell Atlas (HCA) and Human Tumor Atlas Network (HTAN) are two of many such examples. The HCA, which aims to use single-cell genomics to provide a comprehensive map of all cell types in the human body [5], contains uniformly processed single-cell RNA-seq data obtained from normal tissue with few samples derived from diseased tissue. The HTAN also hosts a collection of genomic data collected from tumors across multiple cancer types, including single-cell RNA-seq [6].

Existing resources have focused on making large quantities of harmonized data from normal tissue or adult tumor samples publicly available, but there are considerably fewer efforts to harmonize and publicize data from pediatric tumors. Pediatric cancer is much less common than adult cancer, so the number of available samples from pediatric tumors is smaller compared to the number of adult tumors [7]. Additionally, not every institution has access to data from pediatric tumors. Thus, it is imperative to provide harmonized data from pediatric tumors to all pediatric cancer researchers [8]. To address this unmet need, Alex's Lemonade Stand Foundation and the Childhood Cancer Data Lab developed and maintain the Single-cell Pediatric Cancer Atlas (ScPCA) Portal (https://scpca.alexslemonade.org/), an open-source data resource for single-cell and single-nuclei RNA sequencing data of pediatric tumors.

The ScPCA Portal holds uniformly processed summarized gene expression from 10X Genomics' droplet-based single-cell and single-nuclei RNA-seq for over 500 samples from a diverse set of over 50 types of pediatric cancers. Originally comprising data from 10 projects funded by Alex's Lemonade Stand Foundation, the Portal has since expanded to include data contributed by pediatric cancer research community members. In addition to gene expression data from single-cell and single-nuclei RNA-seq, the Portal includes data obtained from bulk RNA-seq, spatial transcriptomics, and feature barcoding methods, such as ADT/CITE-seq and cell hashing. All data provided on the portal are available in formats ready for downstream analysis, such as SingleCellExperiment or AnnData, with objects containing normalized gene expression counts, dimensionality reduction results, and cell type annotations.

To ensure that all current and future data on the Portal are uniformly processed, we created scpca-nf, a Nextflow-based open-source pipeline (https://github.com/AlexsLemonade/scpca-nf). Using a consistent pipeline for all data increases transparency and allows users to perform analysis across multiple samples and projects without having to do any re-processing. The scpca-nf workflow uses alevin-fry [9] for fast and efficient quantification of gene expression for all samples on the Portal, including single-cell RNA-seq data and any associated ADT/CITE-seq or cell hash data, spatial

transcriptomics data, and bulk RNA-seq data. The scpca-nf pipeline also serves as a resource for the community, allowing others to process their own samples for comparison to samples available on the Portal and allowing us to accept uniformly processed community contributions.

Here, we present the Single-cell Pediatric Cancer Atlas as a resource for all pediatric cancer researchers. The ScPCA Portal provides downloads ready for immediate use, allowing researchers to skip time-consuming data re-processing and wrangling steps. We provide comprehensive documentation about data processing and the contents of files on the portal, including a guide to getting started working with an ScPCA dataset (https://scpca.readthedocs.io/). The ScPCA Portal helps advance pediatric cancer research by accelerating researchers' ability to answer important biological questions.

Results

The Single-cell Pediatric Cancer Atlas Portal

In March of 2022, the Childhood Cancer Data Lab launched the Single-cell Pediatric Cancer Atlas (ScPCA) Portal to make uniformly processed, summarized single-cell and single-nuclei RNA-seq data and de-identified metadata from pediatric tumor samples available for download. Data available on the Portal was obtained using two different mechanisms: raw data was accepted from ALSF-funded investigators and processed using our open-source pipeline, `scpca-nf`, or investigators processed their raw data using `scpca-nf` and submitted the output for inclusion on the Portal.

All samples on the Portal include a core set of metadata obtained from investigators, including age, sex, diagnosis, subdiagnosis (if applicable), tissue location, and disease stage. Some investigators submitted additional metadata, such as treatment and tumor stage, which can also be found on the Portal. All submitted metadata was standardized to maintain consistency across projects before adding to the Portal. In addition to providing a human-readable value for the submitted metadata, we also provide an ontology term ID, if applicable. Submitted metadata was mapped to an associated ontology term IDs obtained from HsapDV (age) [10], PATO (sex) [11], NCBI taxonomy (organism) [12], MONDO (disease) [13], UBERON (tissue) [14], and Hancestro (ethnicity, if applicable) [15]. Including ontology term IDs for each sample provides users with standardized metadata terms that can be used across all projects.

The Portal contains data from 500 samples and over 50 tumor types. The total number of samples for each diagnosis is shown in Figure 1A, along with a breakdown of the proportion of samples from each disease stage within a diagnosis group. Figure 1A summarizes all samples from patient tumors or patient-derived xenografts currently available on the Portal. Most samples found on the Portal were obtained from patients with leukemia (n = 192). The Portal also includes samples from brain and central nervous system tumors (n = 154), sarcoma and soft tissue tumors (n = 68), and a variety of other solid tumors (n = 87). Most samples were collected at initial diagnosis (n = 424), with a smaller number of samples collected either at recurrence (n = 64), during progressive disease (n = 10), or post-mortem (n = 2). Along with the patient tumors, the Portal contains a small number of human tumor cell line samples (n = 4).

Each of the available samples contains summarized gene expression data from either single-cell or single-nuclei RNA sequencing. However, some samples also include additional data, such as quantified expression data from tagging cells with Antibody-derived tags (ADT), like CITE-seq antibodies [16], or multiplexing samples with hashtag oligonucleotides (HTO)[17] prior to sequencing. Out of the 500 samples, 96 have associated CITE-seq data, and 19 have associated multiplexing data. In some cases, multiple libraries from the same sample were collected for additional sequencing,

either for bulk RNA-seq or spatial transcriptomics. Specifically, 118 samples on the Portal were sequenced using bulk RNA-seq and 94 samples were sequenced using spatial transcriptomics. A summary of the number of samples with each additional modality is shown in Figure 1B, and a detailed summary of the total samples with each sequencing method broken down by project is available in Supplemental Table 1.

Samples on the Portal are organized by project, where each project is a collection of similar samples from an individual lab. Users can filter projects based on diagnosis, included modalities (e.g., CITE-seq, bulk RNA-seq), 10X Genomics version (e.g., 10Xv2, 10Xv3), and whether or not a project includes samples derived from patient-derived xenografts or cell lines. The project card displays an abstract, the total number of samples included, a list of diagnoses for all samples included in the Project, and links to any external information associated with the project, such as publications and links to external data, such as SRA or GEO (Figure 1C). The project card will also indicate the type(s) of sequencing performed, including the 10X Genomics kit version, the suspension type (cell or nucleus), and if additional sequencing is present, like bulk RNA-seq or multiplexing.

Uniform processing of data available on the ScPCA Portal

All data available on the Portal was uniformly processed using `scpca-nf`, an open-source and efficient Nextflow[18] workflow for quantifying single-cell and single-nuclei RNA-seq data. Using Nextflow as the backbone for the `scpca-nf` workflow ensures both reproducibility and portability. All dependencies for the workflow are handled automatically, as each process in the workflow is run in a Docker container. Nextflow is compatible with various computing environments, including high-performance computing and cloud-based computing, allowing users to run the workflow in their preferred environment. Setup requires organizing input files and updating a single configuration file for your computing environment after installing Nextflow and either Docker or Singularity. Nextflow will also handle parallelizing sample processing as allowed by your environment, minimizing run time. The combination of being able to execute a Nextflow workflow in any environment and run individual processes in Docker containers makes this workflow easily portable for external use.

When building `scpca-nf`, we sought a fast and memory-efficient tool for gene expression quantification to minimize processing costs. We expected many users of the Portal to have their own single-cell or single-nuclei data processed with Cell Ranger[19], due to its popularity. Thus, selecting a tool with comparable results to Cell Ranger was also desirable. In comparing `alevin-fry` [9] to Cell Ranger, we found `alevin-fry` had a lower run time and memory usage (Supplemental Figure 1A), while retaining comparable mean gene expression for all genes (Supplemental Figure 1B), total UMIs per cell (Supplemental Figure 1C), or total genes detected per cell (Supplemental Figure 1D). (All analyses comparing gene expression quantification tools are available in a public analysis repository[20].) Based on these results, we elected to use `salmon alevin` and `alevin-fry` [9] in `scpca-nf` to quantify gene expression data.

`scpca-nf` takes FASTQ files as input (Figure 2A). Reads are aligned using the selective alignment option of `salmon alevin` to an index with transcripts corresponding to spliced cDNA and intronic regions, denoted by `alevin-fry` as a `splici` index. The output from `alevin-fry` includes a gene by cell count matrix for all barcodes identified, even those that may not contain true cells. This unfiltered counts matrix is stored in a `SingleCellExperiment` object[21] and output from the workflow to a `.rds` file with the suffix `_unfiltered.rds`.

`scpca-nf` performs filtering of empty droplets, removal of low-quality cells, normalization, dimensionality reduction, and cell type annotation (Figure 2A). The unfiltered gene by cell counts matrices are filtered to remove any barcodes that are not likely to contain cells using `DropletUtils::emptyDropsCellRanger()` [22], with all cells that pass being saved to a

`SingleCellExperiment` object and `.rds` file with the suffix `_filtered.rds`. Then, low-quality cells are identified and removed with `miQC` [23], which jointly models the proportion of mitochondrial reads and detected genes per cell and calculates a probability that each cell is compromised. The remaining cells are normalized [24] and undergo dimensionality reduction using both principal component analysis (PCA) and UMAP. Finally, cell types are classified using two automated methods, `SingleR` [25] and `CellAssign` [26]. The results from this analysis are stored in a processed `SingleCellExperiment` object saved to a `.rds` file with the suffix `_processed.rds`.

To make downloading from the Portal convenient for R and Python users, downloads are available as either `SingleCellExperiment` or `AnnData` [27] objects. All `SingleCellExperiment` objects saved as `.rds` files are converted to `AnnData` objects and saved as `.hdf5` files in `scpca-nf` (Figure 2A). Downloads contain the unfiltered, filtered, and processed objects from `scpca-nf` to allow users to choose to perform their own filtering and normalization or to start their analysis from a processed object.

All downloads from the Portal include a quality control (QC) report with a summary of processing information (e.g., `alevin-fry` version), library statistics (e.g., the total number of cells), and a collection of diagnostic plots for each library (Figure 2B-G). The knee plot includes all droplets (i.e., before removing empty droplets) sorted based on the total number of UMIs, and those retained after filtering empty droplets are indicated in the plot (Figure 2B). For each cell that remains after filtering empty droplets, the number of total UMIs, genes detected, and mitochondrial reads are calculated and summarized in a scatter plot (Figure 2C). We include plots showing the `miQC` model and which cells are kept and removed after filtering with `miQC` (Figure 2D-E). A UMAP plot with cells colored by the total number of genes detected and a faceted UMAP plot where cells are colored by the expression of a top highly variable gene are also available (Figure 2F-G).

Processing samples with additional modalities

`scpca-nf` includes modules for processing samples with sequencing modalities beyond single-cell or single-nuclei RNA-seq data: corresponding ADT or CITE-seq data [16], multiplexed data via cell hashing [17], spatial transcriptomics, or bulk RNA-seq.

Antibody-derived tags

To process ADT libraries, the ADT FASTQ files were provided as input into `scpca-nf` and quantified using `salmon alevin` and `alevin-fry` (Supplemental Figure 2A). Along with the FASTQ files, `scpca-nf` takes a tab-separated values (TSV) file with one row for each ADT – containing the name used for the ADT and associated barcode – required to build an ADT-specific index for quantifying ADT expression with `alevin-fry`. The output from `alevin-fry` is the unfiltered ADT by cell counts matrix. The ADT by cell counts matrix is read into R alongside the gene by cell counts matrix and saved as an alternative experiment (`altExp`) within the main `SingleCellExperiment` object containing the unfiltered RNA counts. This `SingleCellExperiment` object containing both RNA and ADT counts is output from the workflow to a `.rds` file with the suffix `_unfiltered.rds`.

`scpca-nf` does not filter any cells based on ADT expression or remove cells with low-quality ADT expression. Any cells removed after filtering empty droplets based on the unfiltered RNA counts matrix are also removed from the ADT counts matrix. The workflow calculates QC statistics for ADT counts using `DropletUtils::cleanTagCounts()` that are stored alongside the ADT by cell counts matrix in the filtered `SingleCellExperiment` object. The `SingleCellExperiment` object

containing the filtered RNA and ADT counts matrix and associated ADT QC statistics is saved to an `.rds` file with the suffix `_filtered.rds`.

The ADT by cell counts matrix is normalized by first determining the ambient profile and then using that profile to calculate median size factors with `scuttle::computeMedianFactors()` [28,29]. We skip normalization for cells with low-quality ADT expression, as indicated by `DropletUtils::cleanTagCounts()`. Although `scpca-nf` normalizes ADT counts, the workflow does not perform any dimensionality reduction of ADT data; only the RNA counts data is used as input for dimensionality reduction. The normalized ADT data is saved as an `altExp` within the processed `SingleCellExperiment` containing the normalized RNA data and is output to a `.rds` file with the suffix `_processed.rds`. All `.rds` files containing `SingleCellExperiment` objects and associated `altExp` objects, are converted to `AnnData` objects and exported as separate RNA (`_rna.hdf5`) and ADT (`_adt.hdf5`) AnnData objects.

If a library contains associated ADT data, the QC report output by `scpca-nf` will include an additional section with a summary of ADT-related statistics, such as how many cells express each ADT, and ADT-specific diagnostic plots (Supplemental Figure 2B-D).
As mentioned above, `scpca-nf` uses `DropletUtils::cleanTagCounts()` to calculate QC statistics for each cell using ADT expression but does not filter any cells from the object. We include plots summarizing the removal of low-quality cells based on RNA and ADT counts in the QC report (Supplemental Figure 2B). The first quadrant indicates which cells would be kept if the object was filtered on both RNA and ADT. The other facets highlight which cells would be removed if filtering was done using only RNA counts, only ADT counts, or both. The top 4 ADTs with the most variable expression are also identified and visualized using density plots to show the normalized ADT expression across all cells (Supplemental Figure 2C) and UMAPs – calculated from RNA data – with cells colored by ADT expression (Supplemental Figure 2D).

Multiplexed libraries

To process multiplexed libraries, the HTO FASTQ files are input to `scpca-nf` and quantified using `salmon alevin` and `alevin-fry` (Supplemental Figure 2C). Along with the FASTQ files, `scpca-nf` requires two TSV files to process multiplexed data: one to build an HTO-specific index for quantifying HTO expression with `alevin-fry` and a second indicating which HTO was used for which sample when multiplexing the library. The unfiltered HTO by cell counts matrix output from `alevin-fry` is saved as an alternative experiment (`altExp`) within the main `SingleCellExperiment` containing the unfiltered RNA counts. This `SingleCellExperiment` object containing both RNA and HTO counts is output from the workflow to a `.rds` file with the suffix `_unfiltered.rds`.

As with ADT data, `scpca-nf` does not filter any cells based on HTO expression, and any cells removed after filtering empty droplets based on the unfiltered RNA counts matrix are also removed from the HTO counts matrix and saved to an `.rds` file with the `_filtered.rds` suffix. `scpca-nf` does not perform any additional filtering or processing of the HTO by cell counts matrix, so the same filtered matrix is saved to the processed `.rds` file with the `_processed.rds` suffix.

Although `scpca-nf` quantifies the HTO data and includes an HTO by cell counts matrix in all objects, `scpca-nf` does not demultiplex the samples into one sample per library. Instead, `scpca-nf` applies multiple demultiplexing methods, including demultiplexing with `DropletUtils::hashedDrops()` [30], demultiplexing with `Seurat::HTODemux()` [17], and genetic demultiplexing when bulk RNA-seq data is available. `scpca-nf` uses the genetic demultiplexing method described in Weber et al. [31], which uses bulk RNA-seq as a reference for the

expected genotypes found in each single-cell RNA-seq sample. The results from all available demultiplexing methods are saved in the filtered and processed `SingleCellExperiment` objects.

If a library has associated HTO data, an additional section is included in the QC report output by `scpca-nf`. This section summarizes HTO-specific library statistics, such as how many cells express each HTO. No additional plots are produced, but a table summarizing the results from all three demultiplexing methods is included.

Bulk and spatial transcriptomics

Multiple libraries were collected for some samples, with the additional libraries being used for bulk RNA-seq and/or spatial transcriptomics. Both of these additional sequencing methods are supported by `scpca-nf`. `scpca-nf` takes FASTQ from bulk RNA-seq as input, trims reads using `fastp` [32], and then aligns reads with `salmon` (Supplemental Figure 3A) [33]. The output is a single TSV file with the gene by sample counts matrix for all samples in a given ScPCA project. This gene by sample matrix is only included with project downloads on the Portal.

To quantify spatial transcriptomics data, `scpca-nf` takes the RNA FASTQ and slide image as input (Supplemental Figure 3B). As there is not yet support for spatial transcriptomics with `alevin-fry`, `scpca-nf` uses Space Ranger to quantify all spatial transcriptomics data [34]. The output includes the spot by gene matrix along with a summary report produced by Space Ranger.

Downloading projects from the ScPCA Portal

On the Portal, users can select to download data from individual samples or all data from an entire ScPCA project. When downloading data for an entire project, users can choose between receiving the individual files for each sample (default) or one file containing the gene expression data and metadata for all samples in the project. Users also have the option to choose their desired format and receive the data as `SingleCellExperiment` (`.rds`) or `AnnData` (`.hdf5`) objects.

For downloads with samples as individual files, the download folder will include a sub-folder for each sample in the project (Figure 3A). Each sample folder contains all three object types (unfiltered, filtered, and processed) as either `SingleCellExperiment` (`.rds`) or `AnnData` (`.hdf5`) objects and the QC report for all libraries from the given sample. The objects house the summarized gene expression data and associated metadata for the library indicated in the filename.

All project downloads include a metadata file, `single_cell_metadata.tsv`, containing relevant metadata for all samples, and a `README.md` with information about the contents of each download, contact and citation information, and terms of use for data downloaded from the Portal (Figure 3A-B). If the ScPCA project includes samples with bulk RNA-seq, two additional files are included: a gene by sample counts matrix (`bulk_quant.tsv`) with the quantified gene expression data for all samples in the project and a metadata file (`bulk_metadata.tsv`).

Merged objects

Providing data for all libraries within a single file makes it easier for users to perform joint gene-level analyses, such as differential expression or gene set enrichment analyses, on multiple samples simultaneously. Therefore, we make a single, merged object available for each project containing all raw and normalized gene expression data and metadata for all single-cell and single-nuclei RNA-seq libraries within a given ScPCA project. The data in the merged object has simply been combined, and no batch-corrected or integrated data is included. If downloading data from a ScPCA project as a

single, merged file, the download will include a single `.rds` or `.hdf5` file, a summary report for the merged object, and a folder with all individual QC and cell type reports for each library found in the merged object (Figure 3B).

To build the merged objects, we created an additional stand-alone workflow for merging the output from `scpca-nf`, `merge.nf` (Figure 3C). `merge.nf` takes as input the processed `SingleCellExperiment` objects output by `scpca-nf` for all single-cell and single-nuclei libraries included in a given ScPCA project. The gene expression data stored in all `SingleCellExperiment` objects are then merged to produce a single merged gene by cell counts matrix containing all cells from all libraries and all shared genes. The genes available in the merged object will be the same as those in each individual object, as all objects on the Portal were quantified using the same index. Any metadata found in the individual processed `SingleCellExperiment` objects are also merged (e.g., `colData`, `rowData`, and `metadata`). The merged normalized counts matrix is then used to select high-variance genes in a library-aware manner before performing dimensionality reduction with both PCA and UMAP. `merge.nf` outputs the merged and processed object as a `SingleCellExperiment` object.

We also account for additional modalities in `merge.nf`. If at least one library in a project contains ADT data, the raw and normalized ADT data are also merged and saved as an `altExp` in the merged `SingleCellExperiment` object. If any libraries in a project are multiplexed, the HTO data is not merged and is not included in the merged object. All merged `SingleCellExperiment` objects are converted to `AnnData` objects and exported as `.hdf5` files. If the merged object contains an `altExp` with merged ADT data, two `AnnData` objects are exported to create separate RNA (`_rna.hdf5`) and ADT (`_adt.hdf5`) objects.

`merge.nf` outputs a summary report for each merged object, which includes a set of tables summarizing the types of samples and libraries included in the project, such as types of diagnosis, and a faceted UMAP showing all cells from all libraries. In the UMAP, each panel represents a different library included in the merged object, with all cells from the specified library shown in color, while all other cells are gray. An example of this UMAP showing a subset of libraries from a ScPCA project is available in Figure 3D.

Materials and Methods

Data generation and processing

Raw data and metadata were generated and compiled by each lab and institution contributing to the Portal. Single-cell or single-nuclei libraries were generated using one of the commercially available kits from 10x Genomics. For bulk RNA-seq, RNA was collected and sequenced using either paired-end or single-end sequencing. For spatial transcriptomics, cDNA libraries were generated using the Visium kit from 10x Genomics. All libraries were processed using our open-source pipeline, `scpca-nf`, to produce summarized gene expression data.

Processing single-cell and single-nuclei RNA-seq data with alevin-fry

To quantify RNA-seq gene expression for each cell or nucleus in a library, `scpca-nf` uses `salmon alevin` [35] and `alevin-fry` [9] to generate a gene by cell counts matrix. Prior to mapping, we generated an index using transcripts from both spliced cDNA and unspliced cDNA sequences, denoted as the `splici` index [9]. The index was generated from the human genome, GRCh38, Ensembl version 104. `salmon alevin` was run using selective alignment to the `splici` index with

the `--rad` option to generate a reduced alignment data (RAD) file required for input to `alevin-fry`.

The RAD file was used as input to the recommended `alevin-fry` workflow, with the following customizations. At the `generate-permit-list` step, we used the `unfiltered-pl` option to provide a list of expected barcodes specific to the 10x kit used to generate each library. The `quant` step was run using the `cr-like-em` resolution strategy for feature quantification and UMI de-duplication.

Post alevin-fry processing of single-cell and single-nuclei RNA-seq data

The output from running `alevin-fry` includes a gene by cell counts matrix, with reads from both spliced and unspliced reads for all potential cell barcodes. This output is read into R to create a `SingleCellExperiment` using the `fishpond::load_fry()` function. The resulting `SingleCellExperiment` contains a `counts` assay with a gene by cell counts matrix where all spliced and unspliced reads for a given gene are totaled together. We also include a `spliced` assay that contains a gene by cell counts matrix with only spliced reads. These matrices include all potential cells, including empty droplets, and are provided in the "unfiltered" objects included in downloads from the Portal.

Each droplet was tested for deviation from the ambient RNA profile using `DropletUtils::emptyDropsCellRanger()` and those with an FDR ≤ 0.01 were retained as likely cells. If a library did not have a sufficient number of droplets and `DropletUtils::emptyDropsCellRanger()` failed, cells with fewer than 100 UMIs were removed. Gene expression data for any cells that remain after filtering are provided in the "filtered" objects.

In addition to removing empty droplets, `scpca-nf` also removes cells from downstream analysis that are likely to be compromised by damage or low-quality sequencing. `miQC` was used to calculate the probability of each cell being compromised [23]. Any cells with a likelihood of being compromised greater than 0.75 and fewer than 200 genes detected were removed before further processing. The gene expression counts from the remaining cells were log-normalized using the deconvolution method from Lun, Bach, and Marioni [24]. `scran::modelGeneVar()` was used to model gene variance from the log-normalized counts and `scran::getTopHVGs` was used to select the top 2000 high-variance genes. These were used as input to calculate the top 50 principal components using `scater::runPCA()`. Finally, UMAP embeddings were calculated from the principal components with `scater::runUMAP()`. The raw and log-normalized counts, list of 2000 high-variance genes, principal components, and UMAP embeddings are all stored in the "processed" object.

Quantifying gene expression for libraries with CITE-seq or cell hashing

All libraries with antibody-derived tags (ADTs) or hashtag oligonucleotides (HTOs) were mapped to a reference index using `salmon alevin` and quantified using `alevin-fry`. The reference indices were constructed using the `salmon index` command with the `--feature` option. References were custom-built for each ScPCA project and constructed using the submitter-provided list of ADTs or HTOs and their barcode sequences.

The ADT by cell or HTO by cell counts matrix produced by `alevin-fry` were read into R as a `SingleCellExperiment` object and saved as an alternative experiment (`altExp`) in the same `SingleCellExperiment` object with the unfiltered gene expression counts data. The `altExp` within the unfiltered object contains all identified ADTs or HTOs and all barcodes identified in the RNA-seq gene expression data. Any barcodes that only appeared in either ADT or HTO data were

discarded, and cell barcodes that were only found in the gene expression data (i.e., did not appear in the ADT or HTO data) were assigned zero counts for all ADTs and HTOs. Any cells removed after filtering empty droplets were also removed from the ADT and HTO counts matrices and before creating the filtered `SingleCellExperiment` object.

Processing ADT expression data from CITE-seq

The ADT count matrix stored in the unfiltered object was used to calculate an ambient profile with `DropletUtils::ambientProfileEmpty()`. This ambient profile was used to calculate quality-control statistics with `DropletUtils::cleanTagCounts()` for all cells remaining after removing empty droplets. Any negative or isotype controls were taken into account when calculating QC statistics. Cells with a high level of ambient contamination or negative/isotype controls were flagged as having low-quality ADT expression, but we did not remove any cells based on ADT quality from the object. The filtered and processed objects contain the results from running `DropletUtils::cleanTagCounts()`.

ADT data was then normalized by calculating median size factors using the ambient profile with `scuttle::computeMedianFactors()`. If median-based normalization failed for any reason, ADT counts were log-transformed after adding a pseudocount of 1. Normalized counts are only available for any cells that would be retained after ADT filtering, and any cells that would be filtered out based on `DropletUtils::cleanTagCounts()` are assigned `NA`. The normalized ADT data is available in the `altExp` of the processed object.

Processing HTO data from multiplexed libraries

Although we did not perform any demultiplexing of samples within a multiplexed library, we did apply three different demultiplexing methods. Results from all three methods are included in the filtered and processed `SingleCellExperiment` objects along with the HTO counts data.

Genetic demultiplexing

If all samples in a multiplexed library were also sequenced using bulk RNA-seq, we performed genetic demultiplexing using genotype data from both bulk RNA-seq and single-cell or single-nuclei RNA-seq [31]. If bulk RNA-seq was not available, no genetic demultiplexing was performed.

Bulk RNA-seq reads for each sample were mapped to a reference genome using `STAR` [**doi10.1093/bioinformatics/bts635?**] and multiplexed single-cell or single-nuclei RNA-seq reads were mapped to the same reference genome using `STARsolo` [36]. The mapped bulk reads were used to call variants and assign genotypes with `bcftools mpileup` [37]. `cellsnp-lite` was then used to genotype single-cell data at the identified sites found in the bulk RNA-seq data [38]. Finally, `vireo` was used to identify the sample of origin [38].

HTO demultiplexing

For all multiplexed libraries, we performed demultiplexing using `DropletUtils::hashedDrops()` and `Seurat::HTODemux()`. For both methods, we used the default parameters and only performed demultiplexing on the filtered cells present in the filtered object. The results from both these methods are available in the filtered and processed objects.

Quantification of spatial transcriptomics data

- Use of space ranger

Quantification of bulk RNA-seq data

- Use of salmon

Cell type annotation

- Implementation of SingleR and CellAssign
- Description of metrics used (e.g., what is the delta median and where does the probability come from)

Generating merged data

- combining counts data and metadata

Converting SingleCellExperiment objects to AnnData objects

- use of zellkonverter

Code and data availability

Figure Titles and Legends

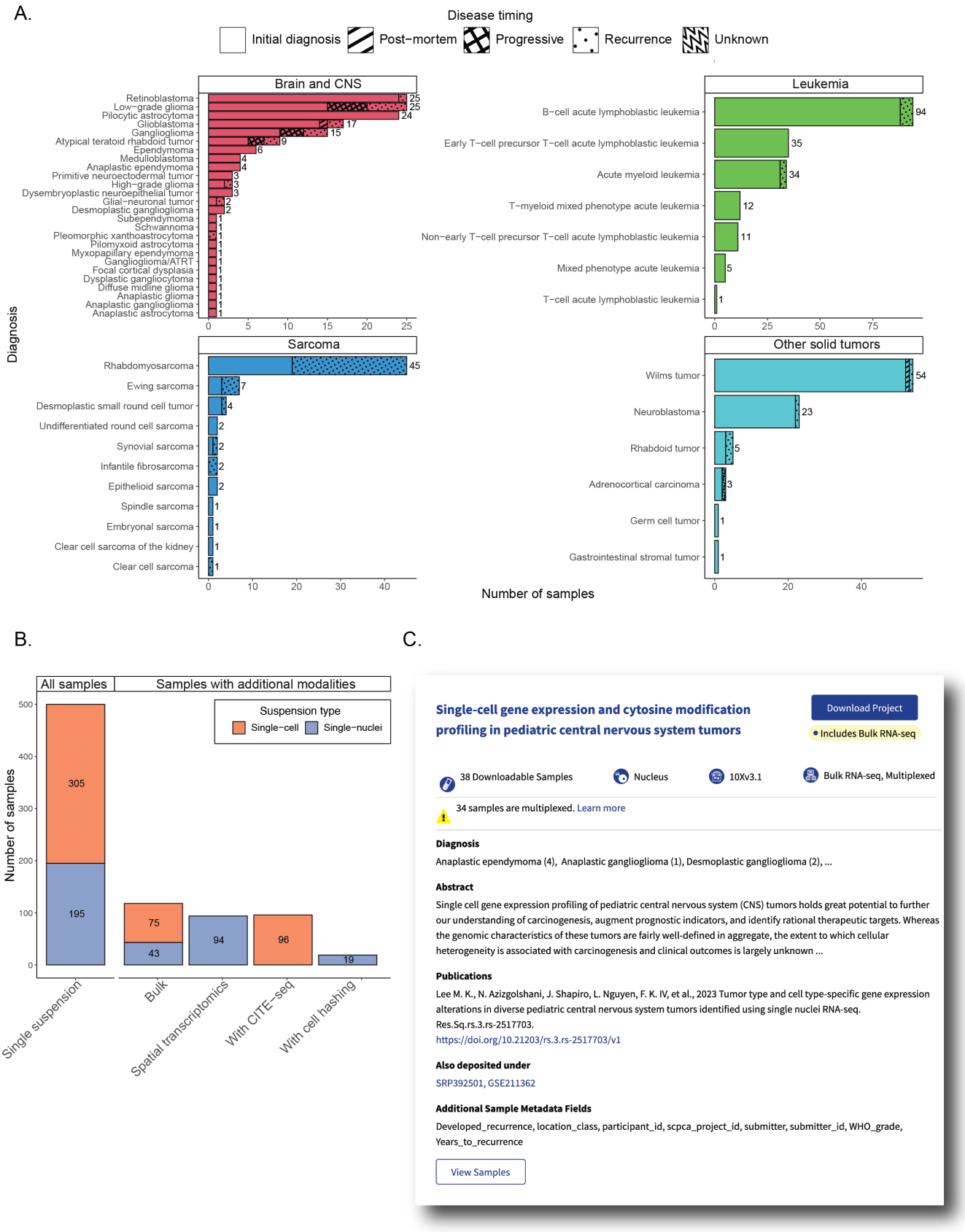

Figure 1: Figure 1. Overview of ScPCA Portal contents.

A. Barplots showing sample counts across four main cancer groupings in the ScPCA Portal, with each bar displaying the number of samples for each cancer type. Each bar is shaded based on the number of samples with each disease timing, and total sample counts for each cancer type are shown to the right of each bar.

B. Barplot showing sample counts across types of modalities present in the ScPCA Portal. All samples in the portal are shown under the “All Samples” heading. Samples under the “Samples with additional modalities” heading represent a subset of the total samples with the given additional modality. Colors shown for each additional modality indicate the suspension type that the single-cell or single-nuclei sample is associated with. For example, 75 single-cell samples and 43 single-nuclei samples have accompanying Bulk RNA-seq data.

C. Example of a project card as displayed on the “Browse” page of the ScPCA Portal. This project card is associated with project `SCPCP000009`. Project cards include information about the number of samples, technologies and modalities, additional sample metadata information, submitter-provided diagnoses, as well as submitter-provided abstract. Where available, submitter-provided citation information as well as other databases where this data has been deposited are also provided.

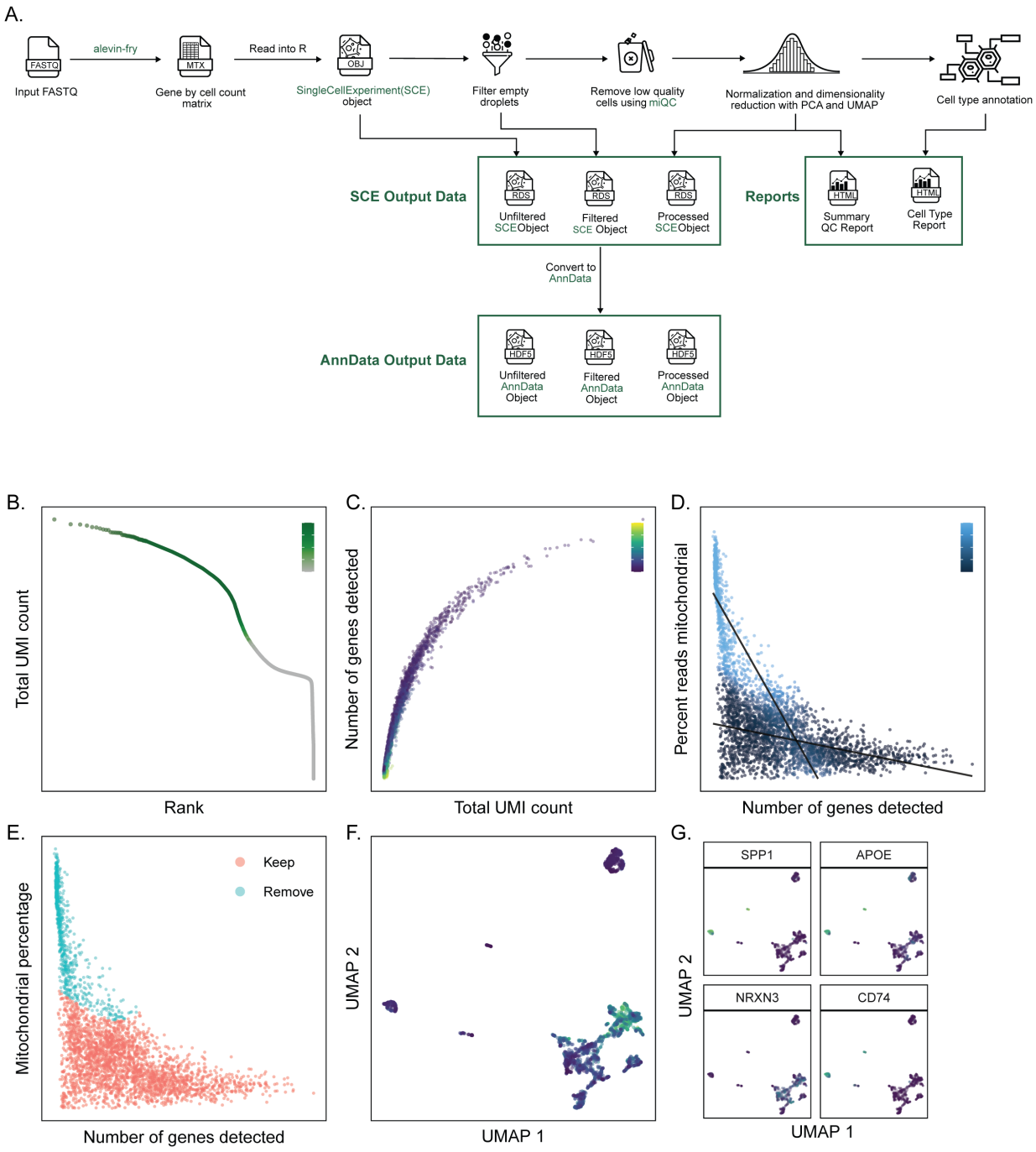

Figure 2: Figure 2. Overview of the `scpca-nf` workflow.

A. An overview of `scpca-nf`, the primary workflow for processing single-cell and single-nuclei data for the ScPCA Portal. Mapping is first performed with `alevin-fry` to generate a gene-by-cell count

matrix, which is read into `R` and converted into a `SingleCellExperiment` (`SCE`) object. This `SCE` object is exported as the `Unfiltered SCE Object` before further post-processing. Next, empty droplets are filtered out, and the resulting `SCE` is exported as the `Filtered SCE Object`. The filtered object undergoes additional post-processing, including removing low-quality cells, normalizing counts, and performing dimension reduction including principal components analysis and UMAP calculation. The object undergoes cell type annotation and is exported as the `Processed SCE Object`. A summary QC report and a supplemental cell type report are prepared and exported. Finally, all `SCE` files are converted to `AnnData` format and exported. Panels B-G show example figures that appear in the summary QC report, shown here for `SCPCL000001`, as follows.

B. The total UMI count for each cell in the `Filtered SCE Object`, ordered by rank. Points are colored by the percentage of cells that pass the empty droplets filter.

C. The number of genes detected in each cell passing the empty droplets filter against the total UMI count. Points are colored by the percentage of mitochondrial reads in the cell.

D. `miQC` model diagnostic plot showing the percent of mitochondrial reads in each cell against the number of genes detected in the `Filtered SCE Object`. Points are colored by the probability that the cell is compromised as determined by `miQC`.

E. The percent of mitochondrial reads in each cell against the number of genes detected in each cell. Points are colored by whether the cell was kept or removed, as determined by both `miQC` and a minimum unique gene count cutoff, prior to normalization and dimensionality reduction.

F. UMAP embeddings of log-normalized RNA expression values where each cell is colored by the number of genes detected.

G. UMAP embeddings of log-normalized RNA expression values for the top four most variable genes, colored by the given gene's expression. In the actual summary QC report, the top 12 most highly variable genes are shown.

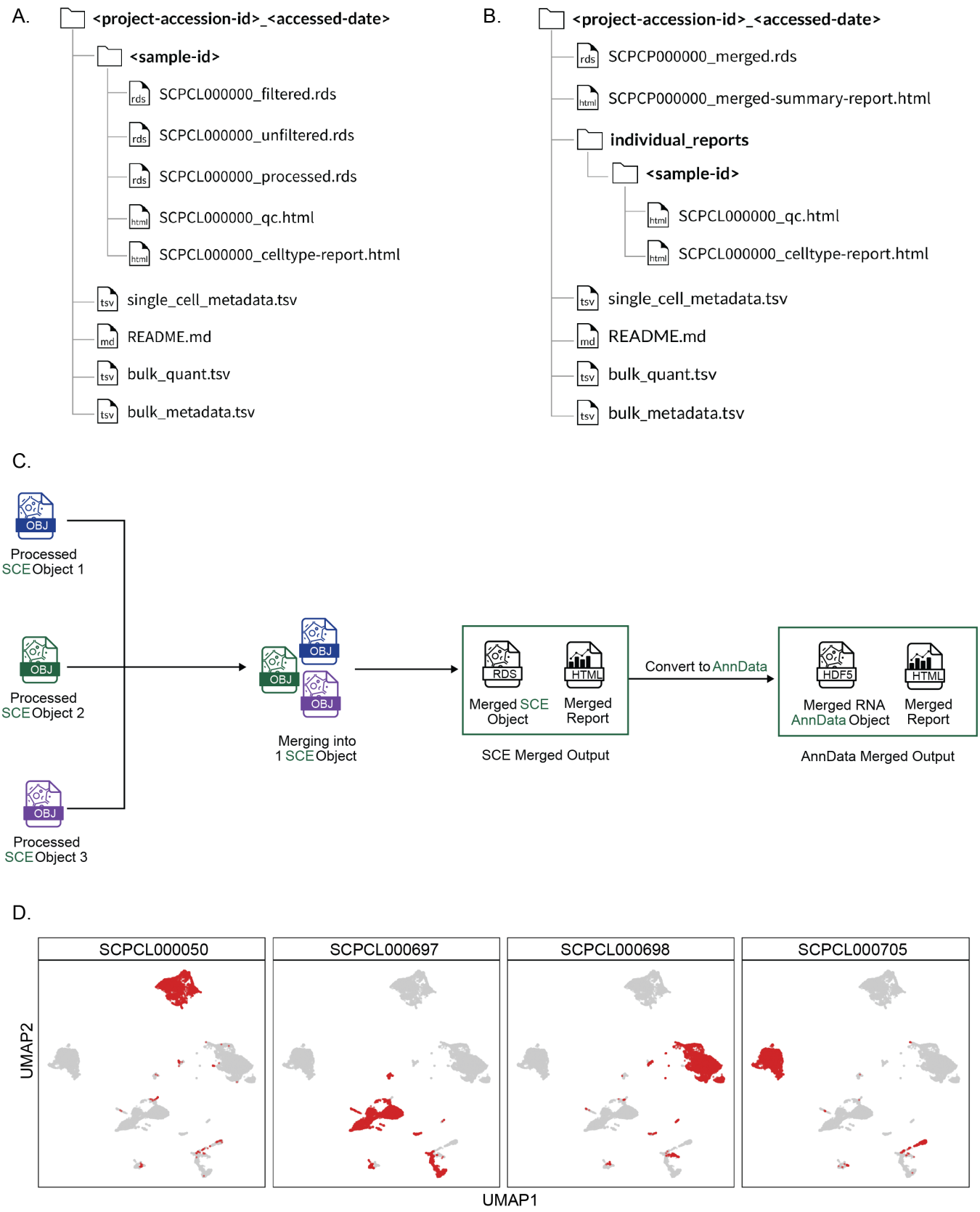

Figure 3: Figure 3. ScPCA Portal project download file structure and merged object workflow.

A. File download structure for an ScPCA Portal project download in `SingleCellExperiment` (`SCE`) format. The download folder is named according to both the project ID and the date it was downloaded. Download folders contain one folder for each sample ID, each containing the three versions (unfiltered, filtered, and processed) of the expression data as well as the summary QC report and cell type report all named according to the ScPCA library ID. The `single_cell_metadata.tsv` file contains sample metadata for all samples included in the download. The `README.md` file provides information about the contents of each download file, additional contact and citation information,

and terms of use for data downloaded from the ScPCA Portal. The files `bulk_quant.tsv` and `bulk_metadata.tsv` are only present for projects that also have bulk RNA-Seq data and contain, respectively, a gene by sample matrix of raw gene expression as quantified by `salmon`, and associated metadata for all samples with bulk RNA-Seq data.

B. File download structure for an ScPCA Portal merged project download in `SCE` format. The download folder is named according to both the project ID and the date it was downloaded. Download folders contain a single merged object containing all samples in the given project as well as a summary report briefly detailing the contents of the merged object. All summary QC and cell type reports for each individual library are also provided in the `individual_reports` folder arranged by their sample ID. As in panel (A), additional files `single_cell_metadata.tsv`, `bulk_quant.tsv`, `bulk_metadata.tsv`, and `README.md` are also included.

C. Overview of the merged workflow. Processed `SCE` objects associated with a given project are merged into a single object, including ADT counts from CITE-seq data if present, and a merged summary report is generated. Merged objects are available for download either in `SCE` or `AnnData` format.

D. Example of UMAPs as shown in the merged summary report. A grid of UMAPs is shown for each library in the merged object, with cells in the library of interest shown in red and all other cells belonging to other libraries shown in gray. The UMAP is constructed from the merged object such that all libraries contribute an equal weight, but no batch correction was performed. The libraries pictured are a subset of libraries in the ScPCA project `SCPCP000003`.

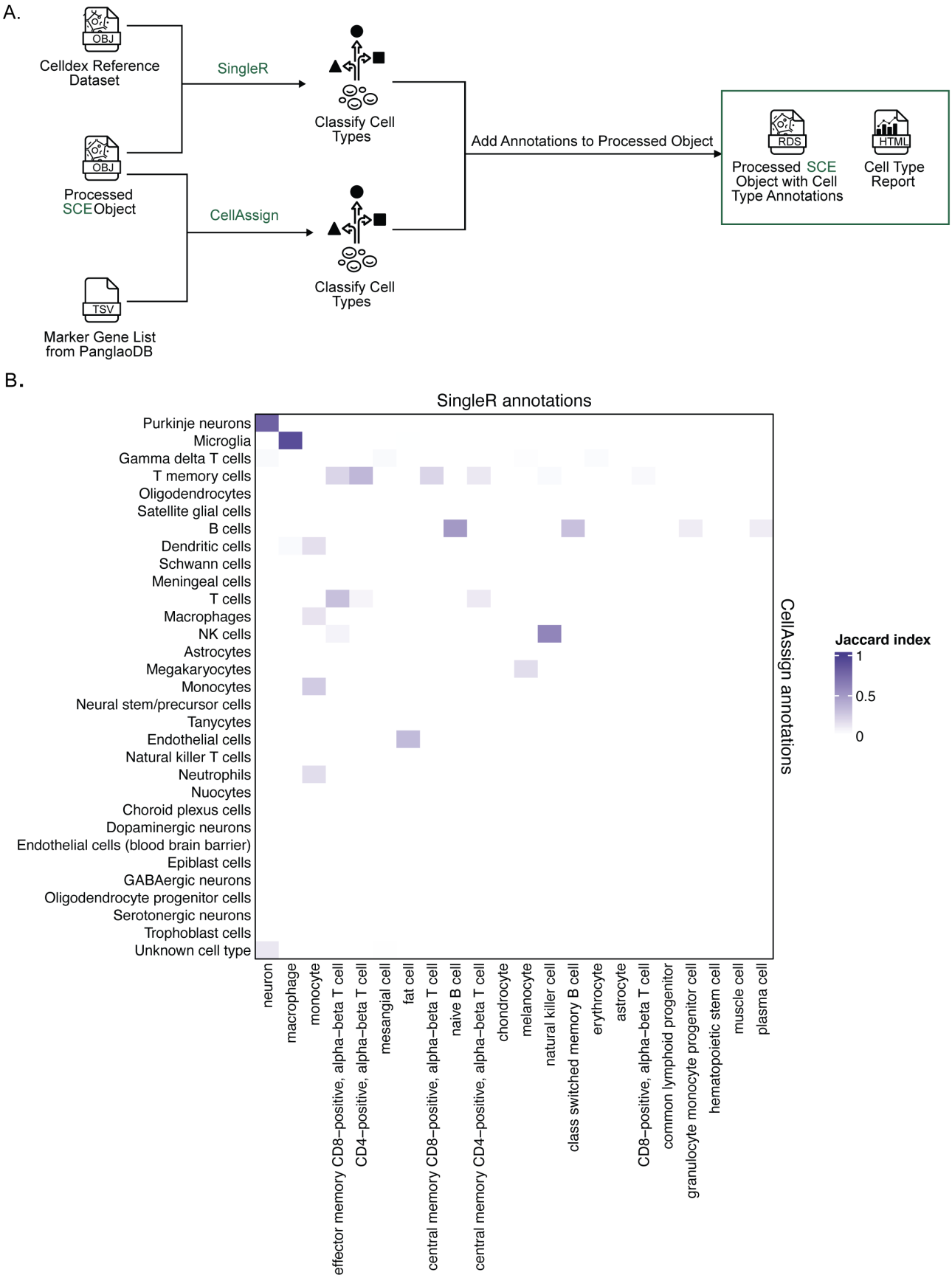

Figure 4: Figure 4. Cell type annotation in `scpca-nf`.

A. Expanded view of the process for adding cell type annotations within `scpca-nf`, as introduced in Figure 2A. Cell type annotation is performed on the `Processed SCE Object`. A `celldex` [25] reference dataset with ontology labels is used as input for annotation with `SingleR` [25], and a list of marker genes compiled from `PanglaoDB` [39] is used as input for annotation with `CellAssign` [26]. Results from cell type annotation are then added to the `Processed SCE Object`, and a cell

type summary report with information about reference sources, comparisons among cell type annotation methods, and diagnostic plots is created. Although not shown in this panel, cell type annotations are also included in the `Processed AnnData Object` created from the `Processed SCE Object` (Figure 2A).

B. Example heatmap as shown in the cell type summary report comparing annotations with `SingleR` and `CellAssign`. Heatmap cells are colored by the Jaccard similarity index. A value of 1 means that there is complete overlap between which cells are annotated with the two labels being compared, and a value of 0 means that there is no overlap between which cells are annotated with the two labels being compared. The heatmap shown is from library `SCPCL000498`.

References

1. **Exponential scaling of single-cell RNA-seq in the past decade**
Valentine Svensson, Roser Vento-Tormo, Sarah A Teichmann
Nature Protocols (2018-03-01) https://doi.org/gc5ndt
DOI: 10.1038/nprot.2017.149 · PMID: 29494575

2. **Defining cell types and states with single-cell genomics**
Cole Trapnell
Genome Research (2015-10) https://doi.org/f7st9g
DOI: 10.1101/gr.190595.115 · PMID: 26430159 · PMCID: PMC4579334

3. **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**
Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, … Bradley E Bernstein
Science (2014-06-20) https://doi.org/gdm4dv
DOI: 10.1126/science.1254257 · PMID: 24925914 · PMCID: PMC4123637

4. **Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment**
Dalia Barkley, Reuben Moncada, Maayan Pour, Deborah A Liberman, Ian Dryg, Gregor Werba, Wei Wang, Maayan Baron, Anjali Rao, Bo Xia, … Itai Yanai
Nature Genetics (2022-08) https://doi.org/gqtn64
DOI: 10.1038/s41588-022-01141-9 · PMID: 35931863 · PMCID: PMC9886402

5. **The Human Cell Atlas**
Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, …
eLife (2017-12-05) https://doi.org/gcnzcv
DOI: 10.7554/elife.27041 · PMID: 29206104 · PMCID: PMC5762154

6. **The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution**
Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E Rood, Orr Ashenberg, Ethan Cerami, Robert J Coffey, Emek Demir, … Xiaowei Zhuang
Cell (2020-04) https://doi.org/ggtkzd
DOI: 10.1016/j.cell.2020.03.053 · PMID: 32302568 · PMCID: PMC7376497

7. **Cancer in Children and Adolescents - NCI** (2023-09-29)
https://www.cancer.gov/types/childhood-cancers/child-adolescent-cancers-fact-sheet

8. **Use case driven evaluation of open databases for pediatric cancer research**
Fleur Jeanquartier, Claire Jean-Quartier, Andreas Holzinger
BioData Mining (2019-01-15) https://doi.org/ggjv7q
DOI: 10.1186/s13040-018-0190-8 · PMID: 30675185 · PMCID: PMC6334395

9. **Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data**
Dongze He, Mohsen Zakeri, Hirak Sarkar, Charlotte Soneson, Avi Srivastava, Rob Patro
Nature Methods (2022-03) https://doi.org/gptg86
DOI: 10.1038/s41592-022-01408-3 · PMID: 35277707 · PMCID: PMC8933848

10. **Ontology Lookup Service (OLS)** https://www.ebi.ac.uk/ols4/ontologies/hsapdv

11. **Ontology Lookup Service (OLS)** https://www.ebi.ac.uk/ols4/ontologies/pato

12. **Home - Taxonomy - NCBI** https://www.ncbi.nlm.nih.gov/taxonomy

13. **Ontology Lookup Service (OLS)** https://www.ebi.ac.uk/ols4/ontologies/mondo

14. **Ontology Lookup Service (OLS)** https://www.ebi.ac.uk/ols4/ontologies/uberon

15. **Ontology Lookup Service (OLS)** https://www.ebi.ac.uk/ols4/ontologies/hancestro

16. **Simultaneous epitope and transcriptome measurement in single cells**
Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, Peter Smibert
Nature Methods (2017-07-31) https://doi.org/gfkksd
DOI: 10.1038/nmeth.4380 · PMID: 28759029 · PMCID: PMC5669064

17. **Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics**
Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck III, Peter Smibert, Rahul Satija
Genome Biology (2018-12) https://doi.org/ggbm6p
DOI: 10.1186/s13059-018-1603-1 · PMID: 30567574 · PMCID: PMC6300015

18. **Nextflow's documentation! — Nextflow 23.10.0 documentation**
https://www.nextflow.io/docs/latest/index.html

19. **Cell Ranger - Official 10x Genomics Support**
10x Genomics
https://www.10xgenomics.com/support/software/cell-ranger/latest

20. **AlexsLemonade/alsf-scpca**
Alex's Lemonade Stand Foundation
(2021-12-21) https://github.com/AlexsLemonade/alsf-scpca

21. **Orchestrating single-cell analysis with Bioconductor**
Robert A Amezquita, Aaron TL Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, … Stephanie C Hicks
Nature Methods (2019-12-02) https://doi.org/ggdxgx
DOI: 10.1038/s41592-019-0654-x · PMID: 31792435 · PMCID: PMC7358058

22. **EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data**
Aaron TL Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, John C Marioni
Genome Biology (2019-03-22) https://doi.org/gfxdhf
DOI: 10.1186/s13059-019-1662-y · PMID: 30902100 · PMCID: PMC6431044

23. **miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data**
Ariel A Hippen, Matias M Falco, Lukas M Weber, Erdogan Pekcan Erkan, Kaiyang Zhang, Jennifer Anne Doherty, Anna Vähärautio, Casey S Greene, Stephanie C Hicks
PLOS Computational Biology (2021-08-24) https://doi.org/gng37g
DOI: 10.1371/journal.pcbi.1009290 · PMID: 34428202 · PMCID: PMC8415599

24. **Pooling across cells to normalize single-cell RNA sequencing data with many zero counts**

Aaron T L. Lun, Karsten Bach, John C Marioni
Genome Biology (2016-04-27) https://doi.org/gfgntn
DOI: 10.1186/s13059-016-0947-7 · PMID: 27122128 · PMCID: PMC4848819

25. **Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage**
Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, … Mallar Bhattacharya
Nature Immunology (2019-01-14) https://doi.org/gfv3p2
DOI: 10.1038/s41590-018-0276-y · PMID: 30643263 · PMCID: PMC6340744

26. **Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling**
Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, … Sohrab P Shah
Nature Methods (2019-09-09) https://doi.org/ggr7ps
DOI: 10.1038/s41592-019-0529-1 · PMID: 31501550 · PMCID: PMC7485597

27. **anndata: Annotated data**
Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, FAlexander Wolf
Cold Spring Harbor Laboratory (2021-12-19) https://doi.org/gst7w6
DOI: 10.1101/2021.12.16.473007

28. **scuttle**
Aaron Lun, Davis McCarthy
Bioconductor (2020) https://doi.org/gtkc7k
DOI: 10.18129/b9.bioc.scuttle

29. **Chapter 12 Integrating with protein abundance | Advanced Single-Cell Analysis with Bioconductor** https://bioconductor.org/books/3.16/OSCA.advanced/integrating-with-protein-abundance.html#cite-seq-median-norm

30. **DropletUtils**
Bioconductor
https://doi.org/gtkc7d
DOI: 10.18129/b9.bioc.dropletutils

31. **Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design**
Lukas M Weber, Ariel A Hippen, Peter F Hickey, Kristofer C Berrett, Jason Gertz, Jennifer Anne Doherty, Casey S Greene, Stephanie C Hicks
GigaScience (2021-09) https://doi.org/gmwhsc
DOI: 10.1093/gigascience/giab062 · PMID: 34553212 · PMCID: PMC8458035

32. **fastp: an ultra-fast all-in-one FASTQ preprocessor**
Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu
Bioinformatics (2018-09-01) https://doi.org/gd9mrb
DOI: 10.1093/bioinformatics/bty560 · PMID: 30423086 · PMCID: PMC6129281

33. **Salmon provides fast and bias-aware quantification of transcript expression**
Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, Carl Kingsford
Nature Methods (2017-03-06) https://doi.org/gcw9f5
DOI: 10.1038/nmeth.4197 · PMID: 28263959 · PMCID: PMC5600148

34. **Space Ranger - Official 10x Genomics Support**
10x Genomics

https://www.10xgenomics.com/support/software/space-ranger/latest

35. **Alignment and mapping methodology influence transcript abundance estimation**
Avi Srivastava, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I Love, Carl Kingsford, Rob Patro
Genome Biology (2020-09-07) https://doi.org/gg98sd
DOI: 10.1186/s13059-020-02151-8 · PMID: 32894187 · PMCID: PMC7487471

36. **STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data**
Benjamin Kaminow, Dinar Yunusov, Alexander Dobin
Cold Spring Harbor Laboratory (2021-05-05) https://doi.org/gqj7ft
DOI: 10.1101/2021.05.05.442755

37. **Twelve years of SAMtools and BCFtools**
Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li
GigaScience (2021-01-29) https://doi.org/gjxzc9
DOI: 10.1093/gigascience/giab008 · PMID: 33590861 · PMCID: PMC7931819

38. **Cellsnp-lite: an efficient tool for genotyping single cells**
Xianjie Huang, Yuanhua Huang
Bioinformatics (2021-05-08) https://doi.org/gqcggs
DOI: 10.1093/bioinformatics/btab358 · PMID: 33963851

39. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**
Oscar Franzén, Li-Ming Gan, Johan LM Björkegren
Database (2019-01-01) https://doi.org/ggkzxr
DOI: 10.1093/database/baz046 · PMID: 30951143 · PMCID: PMC6450036