Manuscript Title

This manuscript (permalink) was automatically generated from AlexsLemonade/ScPCA-manuscript@0528587 on February 26, 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

1. History and overview of the Portal

- In 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 available for download
- The Portal currently holds X amount of samples from X amount of tumor types
- Data available on the Portal was obtained using two mechanisms - accepting raw data from ALSF-funded investigators and investigators who used our open-source pipeline to produce summarized gene expression data for inclusion on the portal.
- In addition to providing summarized gene expression data, we collect a core set of metadata that is provided on the Portal for all samples including, age, sex, diagnosis, subdiagnosis (if applicable), tissue location, and disease stage.
- All metadata that is provided by the submitter is reviewed to standardize as much as possible. We also utilize ontology ID's where possible.
- Fig. 1A shows how many samples we have from each type of tumor. For each diagnosis, we also indicate what proportion of the samples come from each disease stage (e.g., initial diagnosis, recurrence, post-mortem).
- The samples obtained on the portal are mostly from patient tumors, although some are from patient-derived xenografts and human cell lines
- In addition to single-cell and single-nuclei RNA-seq, many samples have associated bulk RNA-seq, ADT data (CITE-seq), cell hashing, or spatial transcriptomics.
- Fig. 1B summarizes the total number of samples that are single-cell vs. single-nuclei. Additionally, we show how many of the samples on the portal also have either bulk, CITE, cell hashing, or spatial data.
- Supplemental Table 1 shows a breakdown of how many of each modality is found in each project.

2. Obtaining additional project information

- On the Portal, samples are organized by project. Each project is a collection of similar samples from a single investigator.
- To select projects of interest, users can filter based on diagnosis, modality included, single-cell or single-nuclei and 10X version. Additionally, users will be able to filter based on if the project includes cell line samples or xenografts.
- A summary of each project, including a list of samples found in each project, is displayed on the Portal.
- Fig.1C shows an example of this summary which include an abstract, links to any external information about the projects such as any associated publication information, and links to

external places where data may be stored such as SRA or GEO.
- If a project includes bulk, CITE, spatial, or multiplexing, this will also be indicated on the project card.

Uniform processing of data available on the ScPCA Portal

1. Processing data with scpca-nf and alevin-fry

- All data available on the portal was uniformly processed using scpca-nf, an open-source and efficient Nextflow workflow for quantifying single-cell and single-nuclei RNA-seq data.
- The workflow uses `salmon alevin` and `alevin-fry` to quantify gene expression data and outputs both raw and normalized counts stored as `SingleCellExperiment` and `AnnData` objects.
- In building the workflow we sought to look for a tool that was fast and memory efficient with comparable results to other popular tools, like `Cell Ranger`.
- 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.
- We compared quantification of single-cell and single-nuclei samples with `alevin-fry` and `Cell Ranger` and observed a decrease in both run time and memory usage in `alevin-fry` compared to `Cell Ranger` (FigS1A).
- When comparing the total UMIs per cell, total genes detected per cell, and mean gene expression, there was no observable difference between `alevin-fry` and `Cell Ranger` (FigS1B-D).
- By utilizing `alevin-fry` in the `scpca-nf` workflow we can process multiple samples at a fraction of the time and cost.

2. Post-processing of quantified gene expression data (Fig 2A)

- In addition to quantification of gene expression, `scpca-nf` also performs filtering, normalization, dimensionality reduction, and cell type annotation.
- 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 matrix is stored in a `SingleCellExperiment` and output from the workflow as an `_unfiltered.rds` file.
- The unfiltered gene by cell counts matrices are then filtered using `DropletUtils::emptyDropsCellRanger()` to remove any barcodes that are not likely to contain cells. All cells that pass this filtering are saved to a filtered `SingleCellExperiment` object and `_filtered.rds` file.
- This filtered object is used as input to the post-processing part of the workflow. This includes removal of low-quality cells using `miQC`, normalization, and dimensionality reduction. The final step of the post-processing performed in `scpca-nf` is classification of cell types using automated methods, `SingleR` and `CellAssign`. The results from this analysis are stored in a processed object saved to a `processed.rds`.
- By providing all three files, unfiltered, filtered, and processed this allows users to perform their own filtering and normalization or to skip those steps and use the already processed objects.
- Finally, all `SingleCellExperiment` objects saved as `.rds` files are converted to `AnnData` objects and saved as `.hdf5` files to allow for downstream processing in either R or Python.
- On the Portal, users can choose to download data as either `SingleCellExperiment` or `AnnData` objects and all downloads will contain all three objects output from `scpca-nf`, the unfiltered, filtered, and processed objects (do we include the download illustrations in the figure to display this?)

3. QC report (Fig 2B)

- Along with outputting the uniformly processed data files, `scpca-nf` also includes a step to create a quality control report for each library.
- This report includes a summary of processing information and library statistics, e.g., the total number of mapped reads, total number of cells, and relevant versions of tools used within the workflow like `salmon` and `alevin-fry`.
- Each report also includes summarized plots showing the quality of each library.
- The knee plot shown in the report ranks the total number of UMIs in each droplet and indicates cells that remained after filtering out empty droplets.
- For each cell that passes filtering out empty droplets, the number of total UMIs, genes detected, and mitochondrial reads is calculated. These cell metrics are summarized in a single plot.
- To remove low-quality cells from the counts matrices, `scpca-nf` applies `miQC`, a data driven approach to filtering cells. The `miQC` model and a plot showing which cells are kept and removed when filtering with `miQC` are shown in the QC report.
- Finally, remaining cells are normalized and undergo dimensionality reduction. The QC report includes a single UMAP where cells are colored by the total number of genes detected and a faceted UMAP where cells are colored by the expression of a top highly variable gene.

4. Benefits of scpca-nf/ Nextflow allows for reproducibility and portability (Does this fit here or should it be earlier before describing the workflow?)

- Using Nextflow as the backbone for the `scpca-nf` workflow ensures reproducibility and portability for users on other systems.
- The scpca-nf workflow can be run in almost any environment including slurm, torque, AWS batch, etc (https://www.nextflow.io/docs/latest/executor.html). This allows users to run this workflow in the environment that they are comfortable in with minimal set-up of dependencies.
- Nextflow handles all dependencies automatically and set up generally requires only organizing input files and configuring Nextflow to run in your environment.
- Each process in the workflow is run in a docker container, so users only need to install Nextflow and docker to be able to use this workflow.
- Nextflow also handles parallelizing processing based on your environment and will configure processing so that run time is minimal.

Making samples with additional modalities available on the Portal

1. Processing samples with additional modalities

- In addition to samples that have single-cell and single-nuclei RNA-seq, we also received samples from submitters with additional sequencing modalities, including CITE-seq, cell hashing, spatial transcriptomics, and bulk RNA-seq.
- To make all the data that we received available, we included additional modules in `scpca-nf` that would accommodate these additional sequencing modalities.
- For a full summary of the libraries and samples available with additional modalities, see supplemental Table 1.
- `scpca-nf` is capable of processing samples that are from a mix of sample types. This means that libraries with and without ADT data or any of the modalities discussed here can be processed together in a single run. Nextflow will handle the parallel processing such that each sample type is run through the correct processes for that modality type.

2. CITE-seq (Fig. S2A-B)

- For all libraries with associated ADT or CITE-seq data, we provide both the RNA and ADT gene expression data in the files available for download on the portal.

- Both FASTQ from single-cell/single-nuclei and from ADT were input into `scpca-nf` and quantified using `salmon alevin` and `alevin-fry`.
- We required a barcodes file from each submitter that contained the ADT labels and the associated barcode. This was used to build the index used for quantification of the ADT FASTQ and creation of the cell by ADT matrix.
- Unlike with RNA counts, we do not perform any filtering of cells due to low quality ADT expression. However, we do include the results from running `DropletUtils::cleanTagCounts()` in both the filtered and processed objects produced by `scpca-nf`.
- Similar to RNA counts, we do normalize ADT data and provide the normalized counts matrix in the processed SCE object, but we do not provide any dimensionality reduction of ADT data, only the RNA data is used for dimensionality reduction.
- ADT data can be found in the `altExp` slot of each `SingleCellExperiment` object or as a separate `_adt.hdf5` file for `AnnData` objects.
- This section includes a summary of statistics such as how many cells express each ADT.
- For libraries with ADT, we also include additional plots.
- As mentioned above, we include the results from `DropletUtils::cleanTagCounts()`, but do not filter any ADTs or cells from the object. Instead we include a column in the `colData` of the processed SCE object that indicates if it is recommended to remove ADTs or not. In the QC report, we summarize filtering taking into account removal of cells because of low quality RNA or ADT. The plot shown in the report highlights which cells would be removed if only filtering using RNA, only ADT, or both.
- Similar to the UMAPs for the RNA data from single-cell/single-nuclei that highlight the top variable genes, the report includes UMAPs highlighing the 4 most variable ADTs in the data. This is shown with UMAPs and ridge plots.

3. Cell hashing (Fig. S2C)

- Similar to ADT data, if any libraries were multiplexed and have an associated cell hashing library, both the RNA and HTO FASTQ are provided as input to the workflow and quantified with `salmon alevin` and `alevin-fry`. We also include a library pools file which indicates which libraries contain which samples and the associated tags used to label each sample.
- Although we quantify the HTO data and include the cell by HTO counts matrix in all objects, we do not demultiplex the samples so that there is one sample per library. Instead, we apply multiple demultiplexing methods including genetic demultiplexing, demultiplexing with `DropletUtils::hashedDrops()`, and demultipilexing with `Seurat::HTODemux()`. The results from these three methods are included in the filtered and processed objects.
- Add some more details about how we do genetic demultiplexing, using vireo and bulk RNA-seq
- If a library has associated HTO data, an additional section is added to the QC report included on the portal and output by `scpca-nf`.
- This section includes a summary of statistics such as how many cells express each HTO.
- For HTO, we do not include any additional plots, but we do show a table summarizing how many cells belong to each sample included in the multiplexed library using each of the demultiplexing methods mentioned.

4. Bulk and Spatial (Fig S3)

- Some samples underwent sequencing using both single-cell/single-nuclei and an additional method like bulk RNA-seq or spatial transcriptomics.
- `scpca-nf` is able to quantify both of these additional sequencing methods.
- Bulk RNA FASTQ are first trimmed using `fastp` and then aligned using `salmon`. The bulk output is a single tsv file with the sample by gene matrix for all samples in that project.
- For spatial transcriptomics, the spatial RNA FASTQ and slide image are input into `scpca-nf` and quantified using `spaceranger`. The output includes the spot by gene matrix along with a summary report, produced by `spaceranger`.

Downloading projects from the ScPCA Portal

1. Users can download all samples for a given project together

- The portal has two different options to allow users to download data for all samples in a given ScPCA Project, either as invididual files for each sample or as a single merged file.
- By default, when downloading a project, the download will include a folder for each sample that is included in the project.
- That folder will contain all individual `SingleCellExperiment` objects as `.rds` files or `AnnData` objects as `.hdf5` files, depending on the file format chosen by the user (Fig. 3A).
- Each of these objects contains the gene expression data and metadata for a single library.
- If a given project has associated bulk RNA-seq, then a sample by gene counts matrix, `bulk_quant.tsv`, including the quantified gene expression data for all samples in a project with associated bulk RNA-seq will be included.

2. Merged objects

- Providing all data from all libraries withing a single file makes it easier for users to perform joint analysis on multiple samples at the same time. Specifically, these objects can be useful for comparing gene-level metrics across multiple samples, such as differential expression analysis and gene set enrichment analysis.
- Therefore, we make a single, merged `SingleCellExperiment` or `AnnData` object (Fig. 3B) available for each project (without batch-correction or integration).
- This file contains one object with all raw and normalized gene expression data and metadata for all single-cell and single-nuclei RNA-seq libraries within a given ScPCA project
- If downloading a project that contains at least one library with CITE-seq, the quantified CITE-seq expression data will also be merged. In SCEs this is provided as an `altExp` within the main object, but for `AnnData` objects, the quantified CITE-seq data is provided as a separate file.

2. The merged object workflow (Fig. 3C and 3D)

- To create the merged objects, we created an additional stand-alone workflow for merging the output from `scpca-nf`, `merge.nf` (Fig. 3C).
- Following processing of each `SingleCellExperiment` object with `scpca-nf`, all processed objects from all libraries and samples within a project are input to the merge workflow, which combines all input data into a single merged object.
- The merged object contains raw and normalized gene expression counts for all cells in all libraries. The same index was used for processing all individual libraries, so the genes found will be the same as in an invididual object.
- After merging, the top 2000 high-variance genes are calculated by modeling variance within each library included in the merged object.
- These high-variance genes are used to calculate new PCA coordinates using `batchelor::multiBatchPCA()` and specifying librares as batches.
- The top 50 PCs were selected and used as input to calculate new UMAP embeddings on the merged object.
- Similar to `scpca-nf`, the merged `SingleCellExperiment` object is converted to a merged `AnnData` object and both formats are provided as download options on the Portal.
- Along with the merged objects, for each project, a merged summary report is created and output.
- This report includes a brief summary of the samples and libraries included in the merged object, including a summary of the type of libraries (e.g., single-cell, single-nuclei, with CITE-seq) and sample diagnoses included in the object.
- The report also contains a UMAP showing all cells from all libraries included in the merged object. For each library, a separate panel is shown, and cells from that library are colored while all other

cells are gray (Fig. 3D).

Materials and Methods

Data generation

- how data was generated in different labs using 10X and then sent to the Data Lab

Data processing (do we need this section?)

- Mention that all data was processing using `scpca-nf` either by us or external submitters

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

- Use of salmon alevin and alevin-fry to process all raw FASTQ files
- Information on index used
- Parameter choices for alevin-fry

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

- filtering of empty droplets
- removal of low quality cells
- normalization
- HVG selection
- PCA and UMAP calculation

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

- How we used alevin-fry to quantify ADT and HTO libraries

Processing CITE-seq expression data

- Filtering low quality cells based on ADT data
- Normalization of ADT data

Genetic demultiplexing

- Use of vireo and matching bulk RNA-seq

HTO demultiplexing

- Seurat
- DropletUtils

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. 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.

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