Manuscript Title

This manuscript (permalink) was automatically generated from AlexsLemonade/ScPCA-manuscript@c750b62 on February 14, 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.

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