Introduction

In this section, we provide an overview of the data management challenges specific to single-cell sequencing experiments. Single-cell sequencing enables the analysis of gene expression at the individual cell level, leading to unique data management requirements due to the high dimensionality and complexity of the data. Single-cell sequencing encompasses a diverse array of techniques to sequence transcriptomics, epigenomics and combinations of multiple modalities (multiomics) at single-cell resolution, each presenting its own set of challenges and considerations. Additionally, the field embraces a wide range of data analysis approaches, further compounding the complexity. Consequently, addressing the standardized description and storage of data and associated metadata becomes paramount in this context. To ensure the reproducibility and reliability of research findings, it is crucial to proactively identify the specific steps in the data workflow that should be preserved. Moreover, decisions regarding data formats must be made collectively to facilitate seamless data sharing, collaboration, and long-term data preservation within the single-cell user community. This document aims to elucidate these challenges and provide practical guidance for navigating them effectively.

Preprocessing and Quality Control

Data preprocessing and quality control are integral components of the single-cell data analysis workflow, playing a pivotal role in ensuring the integrity of the data and facilitating accurate downstream analysis. These challenges encompass a spectrum of tasks aimed at enhancing the quality, reliability, and interpretability of the data, making them conducive for subsequent analysis. By comprehensively addressing data preprocessing and quality control, we aim to provide researchers with a robust framework for navigating these critical stages of the single-cell sequencing process. This includes strategies for addressing technical variability, identifying and mitigating low-quality cells or outliers, managing batch effects, and other sources of variability that may arise within and between datasets.

Data analysis rational

Description

Preprocessing encompasses tasks such as the removal of empty droplets, quality control, batch correction, data normalization, and transformation to mitigate technical variations. These steps aim to ensure that the data is in a suitable state for downstream analysis. Then, the next step's central objectives include the identification of individual cells within the dataset, the assignment of gene expression profiles to each cell, and the generation of count matrices that represent the expression levels of thousands of genes across all cells. To do so, tools like CellRanger (for 10x data) or STARsolo (a more generic and open-source tool that supports various droplet- and plate-based data) are used to facilitate the crucial process of cell and gene assignment. These tools are specifically designed to take the raw sequencing data and process it into quantifiable and interpretable information. This transformation of raw data into structured, cell-by-gene matrices is fundamental for downstream analyses, such as clustering cells by similar gene expression profiles, identifying cell types or inferring cell evolution trajectories. In essence, CellRanger and STARsolo play a pivotal role in converting large and complex sequencing data into a format that researchers can subsequently explore to extract these aforementioned biological insights. Following the execution of cell and gene assignment, post-processing steps come into play. These post-processing stages involve activities like efficient clustering of cells and biologically relevant annotation of clusters. By carefully orchestrating both pre- and post-processing phases, researchers can enhance the quality, reliability, and interpretability of their single-cell sequencing data, ultimately leading to more accurate and biologically meaningful insights.

Considerations

Pre-cell/gene assignation

- **Low-Quality Cell Detection**: Explore methods for identifying and removing low-quality cells or outliers from the dataset.
- **Normalization and Transformation**: Determine how to effectively normalize and transform the data to account for technical variability.

Post-cell/gene assignation

- **Batch Effects Handling**: Develop strategies to mitigate batch effects and other sources of variability within and between datasets.
- **Efficient Clustering**: Consider techniques to achieve efficient and meaningful clustering of single-cell data.
- **Biological Annotation**: Determine how to annotate the identified cell clusters with biologically relevant labels.

Solutions

- **Normalization and Transformation**: Consider using established methods such as shifted logarithm, variance stabilizing transformation (sctransform) or cell pool-based size factor estimators (scran) to address differences in sequencing depth and monitor data quality. Alternative normalization methods such as term frequency-inverse document frequency (TF-IDF) are well-suited for scATAC-seq data.
- **Low-Quality Cell Detection**: Evaluate metrics like the number of detected genes per cell, mitochondrial gene content, and UMI counts to define quality criteria. The threshold for data quality acceptability is variable depending of several factors like the number of replicates or the type of organism (prokaryote, plant, animal...) used.
- **Batch Effects Handling**: Examining your data to check that the most important elements for the clustering/cell comparison are biological and not technical. Exploring batch correction methods like Harmony can help reduce technical biases in data integration.
- **Biological Annotation**: Use known marker genes or reference-based annotation to assign cell types or states to clusters. A database of known cell markers (like CellMarker) can be helpful.

Each of these elements needs to be provided with a comprehensive description. Including details on the normalization techniques applied, outlier removal strategies, and batch correction methods employed to enhance data quality and reliability.

Data Integration and Analysis Across Experiments

Description

The analysis of single-cell sequencing data frequently requires the integration and comparative examination of data stemming from various experiments. Combining datasets to gain a broader perspective or comparing results from distinct experiments, navigating the intricacies of data integration, harmonization, and interpretation is essential for extracting meaningful insights from single-cell sequencing data. This section addresses these considerations and provides solutions to facilitate the effective analysis and interpretation of integrated data, allowing researchers to draw comprehensive conclusions from diverse experimental sources.

Considerations

- **Data Integration**: How can we integrate data from different experiments while accounting for differences in experimental conditions? (eg. sample WT vs KO with several time points)
- **Data Comparison**: What approaches can be used to identify shared cell types and biological signals across datasets? (eg. sample WT vs KO comparison scATAC- and scRNA-seq)
- **Annotation Consistency**: How should we manage metadata and annotations to ensure consistent interpretation across experiments?

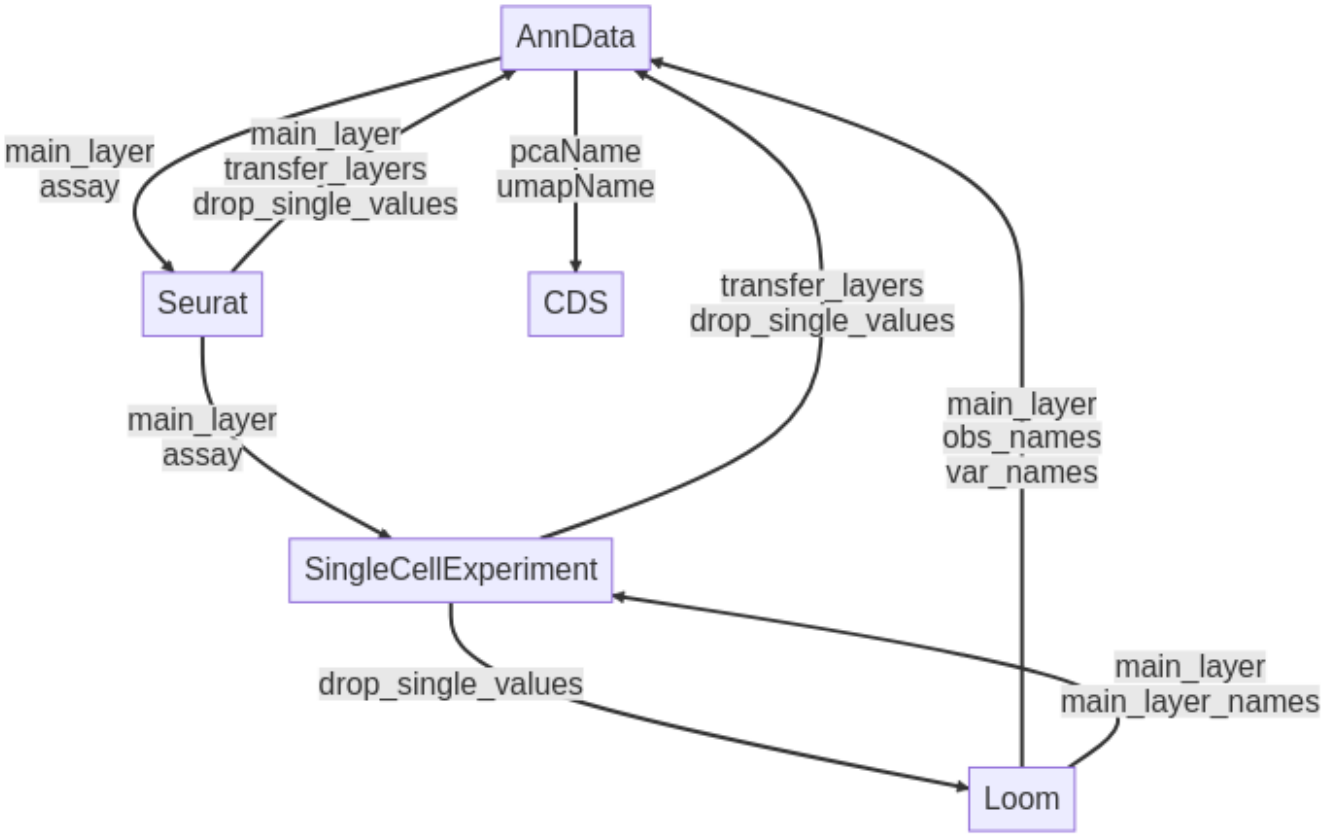
Solutions

- **Data Integration & Data Comparison**: Use a built-in method for data integration & comparison (like Seurat or Scanpy), including normalization, batch correction method and dimensionality reduction techniques to see their effect. Here the difficulty is to make sure the integration/comparison is reliable, meaning being careful that the cell type annotations are consistent with previous knowledge and that the number/cell repartition is relevant.
- **Annotation Consistency**: Consistent metadata and annotation practices are needed, including standardized naming and format usage. Re-using terms from UniProt or Gene Ontology should be considered.

Datatype Consistency and Interoperability Across Formats

Description

Single-cell sequencing data is encoded into many different competing formats, with HDF5-compatible formats such as AnnData and Loom, as well as other commonly-used formats such as SeuratObject, CellDataSet (CDS) and SingleCellExperiment (SCE). Each of these formats is favoured by their respective analysis suites; Scanpy, Seurat, Monocle and Scater.



The image above depicts the conversion routes of a popular conversion tool SCEasy, which demonstrates the limited conversion potential between the different formats. Indeed, data are stored in a matrix composed of different layers, converting the format may lead to the loss of some of them as described in the image. Some of these formats use different programming languages to perform the conversion, such as the Loom format which requires a Python component.

Considerations

- **Datatype Preferences**: Which datatypes should be actively maintained and supported, and which ones should be discouraged? (e.g. popularity, complexity of format, stability between versions)
- **Datatype Support**: Which datatypes do we actively support via bioinformatic cloud pipelines and tutorials?

Solutions

- **Datatype Preferences**: The most common formats are AnnData and SeuratObject. There is waning support for Loom, CDS and SCE, though SCE is an important format on BioConductor and is a common datatype for sharing single-cell experiments in publications.
- **Datatype Support**: Seurat and ScanPy are popular analysis workflows in Galaxy, and it might be important to ensure that there is consistent and stable conversion potential between the two.

Long-Term Data Storage and Accessibility

Description

Ensuring the long-term storage and accessibility of single-cell sequencing data pose distinct challenges that demand attention. This section delves into the critical considerations for effectively storing and making single-

cell sequencing data accessible over an extended period of time.

Considerations

- **Effective Archiving**: What are the best practices for archiving and safeguarding extensive single-cell sequencing datasets to ensure their long-term preservation?
- **Ethical Data Handling**: How can we guarantee data privacy and adhere to ethical guidelines when sharing sensitive single-cell data with the research community?
- **Collaborative Platforms**: Which platforms or repositories are suitable for simplifying data sharing and encouraging collaboration among researchers?
- **Enhancing Reproducibility**: What specific steps and formats should be employed to enable reproducibility in single-cell sequencing experiments?

Solutions

- **Effective Archiving**: Use established data repositories like GEO (Gene Expression Omnibus) or ArrayExpress for storage of experimental descriptive metadata and processed data such as count matrices. The corresponding raw sequencing data can be optimally archived at Sequence Read Archive or European Nucleotide Archive.

- **Ethical Data Handling**: Emphasize the importance of informed consent and ethical considerations in data-sharing agreements.

- **Collaboration Platforms**: Explore version control systems (e.g., Git), data sharing platforms (e.g., Zenodo), data analysis platforms (e.g., Galaxy), and domain-specific repositories (e.g., Single Cell Portal) to facilitate efficient data sharing, analysis, and collaboration.

- **Enhancing Reproducibility**: Guide on enhancing reproducibility, including the use of containerization technologies like Docker to encapsulate analysis environments to ensure analysis can be reproduced with the exact same tools version. Particularly, BioContainers comes in handy when dealing with bioinformatics tools. Emphasize the importance of documenting analysis workflows, code, and metadata using standardized formats and sharing them in version-controlled repositories. Galaxy provides a solution for containerization, versioning, workflow management and reproducibility for novice users.

Analysis step description and format proposal

- **Raw Sequencing Data**:

    - *Data Type*: Raw FASTQ files for sequencing reads.
    - *Format*: Compressed FASTQ format (*.fastq.gz).
    - *Explanation*: Raw sequencing data is typically stored in compressed FASTQ format (*.fastq.gz). This format retains the original sequencing reads and is space-efficient. Compressed files reduce storage requirements while preserving data integrity.

- **Cell-Gene Assignment**:

    - *Data Type*: Cell-gene assignment matrix indicating gene expression levels per cell. Additionally, gene and cell annotations (eg. gene symbols or batches, time points, genotypes) are added.
    - *Format*: Standardized data matrix format, such as h5, h5ad or CSV.

    - *Explanation*: The cell-gene assignment matrix, representing gene expression per cell, is best stored in a standardized format like h5 h5ad or CSV as it will allow the modification needed for the next step while being readable by most single-cell tools.

- **Dimensionality Reduction and Clustering**:

    - *Data Type*: Reduced-dimension representations (e.g., PCA, t-SNE) and cell clusters.
    - *Format*: Include plots and files in common data visualization formats (e.g., PDF, PNG).
    - *Explanation*: Visual formats like PDF and PNG allow easy sharing and visualization.

- **Annotation and Biological Interpretation**:

    - *Data Type*: Annotated cell types, differential gene expression results, and any other biologically meaningful annotations.
    - *Format*: Structured and standardized annotation files, such as Excel spreadsheets, CSV or JSON, alongside visualizations like heatmaps or volcano plots in common visualization formats.
    - *Explanation*: Biologically meaningful annotations, including cell types and differential gene expression results, should be stored in structured formats. Visualizations like heatmaps or volcano plots should be included in standard visual formats for easy interpretation.

- **Analysis Code and Environment**:

    - *Data Type*: All code and scripts used for data preprocessing, analysis, and visualization.
    - *Format*: Version-controlled repositories using Git, or container files to capture analysis environments should be used. Detailed documentation for code execution should be provided.
    - *Explanation*: Analysis code and scripts should be version-controlled using Git repositories. Additionally, capturing the analysis environment using Docker or Singularity container files helps to ensure reproducibility. Detailed documentation of code execution is essential for transparency and re-usability.

- **Metadata**:

    - *Data Type*: Comprehensive metadata describing experimental conditions, sample information, and data processing steps.
    - *Format*: Structured metadata files in widely accepted formats like JSON, CSV or Excel spreadsheets, following community-specific metadata standards if available.
    - *Explanation*: Following community-specific metadata standards, if available, ensures consistency and compatibility with other datasets.

By preserving these steps and data in standardized and accessible formats, researchers can enhance the reproducibility of single-cell sequencing experiments, facilitate collaboration, and ensure that others can validate and build upon their work effectively. Other additional files can be kept if useful for the interpretation (e.g., for scATAC, the results files containing sequence fragments or mapping can be important, they should be kept in standardized format: tsv, bam, bed or bigwig).

Relevant Tools and Resources

- Review of single cell best practices: Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023) https://doi.org/10.1038/s41576-023-00586-w

- The single cell best practice book by the single-cell best practices consortium: https://www.sc-best-practices.org/preamble.html