ASAP CRN Cloud Postmortem-Derived Brain Sequencing Collection

README - currently release 1.0.0, 10.5281/zenodo.11585274

Overview

The Human Postmortem-derived Brain Sequencing Collection is a harmonized repository comprised of sequencing data (Single-Nucleus RNA-seq and Bulk RNA-seq) contributed by ASAP CRN teams. This release contains 156 harmonized samples of cell ranger FASTQ data.

The current collection will be expanded and improved as additional Human PMDBS data is uploaded into the ASAP CRN Cloud. When complete, the collection will provide sequencing data produced by the following techniques: single-nucleus RNAseq, single-cell RNAseq, bulk RNA-seq, ATAC-seq, long read WGS, and single-nucleus multiome sequencing (paired snRNAseq, snATACseq).

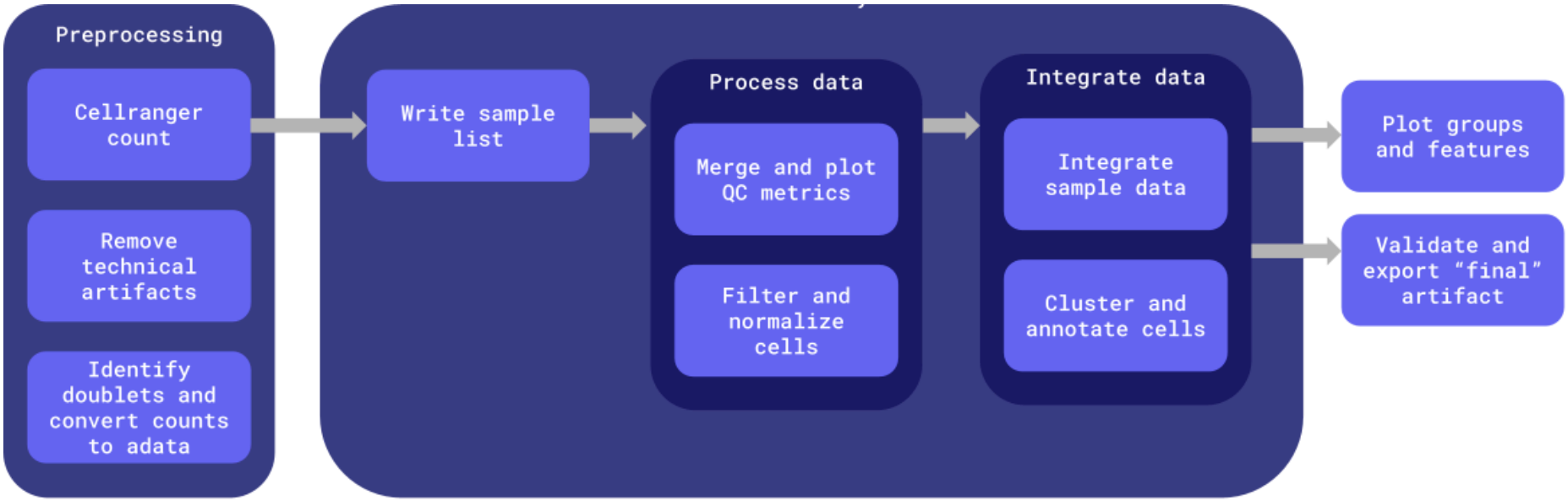
ASAP Teams: Team Hafler, Team Lee, Team Jakobsson, Team Scherzer
Lead Principal Investigators: David Hafler, Johan Jakobsson, Michael Lee, Clemens Scherzer
Co-Principal Investigators: Roger Barker**,** Jose Bras, Xianjun Dong, Molly Gale Hammell, Agnete Kirkeby, Joshua Levin, Le Zhang
Contributors: Andy Henrie, Hampton Leonard, Heather Ward, Chandra Sreeganga, Biqing Zhu, Haowei Wang, Anthony Russo, Jae Min Park, Lee Marshall, Kimberly Paquette, Kaitlyn Westra, Andrew Pyman, Ellison Lopes, Celia Kun-Rodrigues, Rita Guerreiro, Anita Adami, Cole Wunderlich, Talitha Forcier, Raquel Garza, Annelies Quagebeur, Yogita Sharma, Oliver Tam, Serrano Geidy, Thomas Beach, Madison Cline, Zhixiang Liao, Idil Tuncali, Monika Sharma, Jacob Parker, Zechuan Lin, Jie Yuan, Nathan Haywood, Sean Simmons, Maoxuan Lin, Kwanho Kim, Xufei Teng, Daniel El Kodsi, Roger Barker, Jose Bras, Xianjun Dong, Molly Gale Hammell, Agnete Kirkeby, Joshua Levin, Le Zhang, David Hafler, Johan Jakobsson, Michael Lee, Clemens Scherzer

GitHub: https://github.com/ASAP-CRN
Data Release Date: 6/24/24

Release Number: 1.0.0
Release memo: 10.5281/zenodo.11585274

Platformed Data

The raw scRNAseq data has been processed and harmonized to create a harmonized database of gene expression data. The platformed data is summarized below as Raw Data, Curated Data, and QC Summaries below. The complete workflow used for processing the platformed data can be found at the ASAP CRN github repository.



Data File Manifest: ASAP CRN Cloud File Manifest - v1.1

Raw Data

The *raw* data refers to the fastq files transferred by each ASAP CRN Team. These data are available with the caveat that they are stored in "requester pays" buckets on the google cloud platform. A google cloud billing account will need to be registered to access these data. More information is available HERE.

Curated Data

The curated data can be categorized as "*preprocessed*" and "*processed*", and "*integrated*".

Preprocessed. Pre-processing refers to the first stage of sequence alignment from the *raw* fastq files with cellranger for each sample, followed by basic QC. This QC consists count correction due to ambient RNA molecules and random barcode swapping with `cellbender`, and removing doublets with `scrublet` , yielding a series of artifacts: Intermediate data objects (e.g. *adata_object.h5ad) for each sample, and overall QC plots.

Processed. Processing refers to *filtering* the low quality cells – the next level of QC – and *feature selection*, yielding a data object merged across samples (e.g. `*merged_adata_object.h5ad`, `*merged_adata_object_filtered.h5ad`) and as per this workflow. The filtering cutoffs were chosen to be:

- Mitochondrial gene percentage < 10%
- Doublet_scores < 0.2
- Total counts between 500, and 100,000
- Number features per cell between 300 and 10,000

Feature selection found the top genes with the 'seurat_v3' flavor of scanpy's highly_variable_genes method on the pre-processed counts.

Integrated. The integrated or *harmonized* data represents batch-corrected aggregated data. Batch correction and integration of the individual samples into a single count table is the critical step of harmonization. In the workflow we call this entire process the "cohort analysis", and the key steps of integration take place in the "clustering" workflow.
The full *integration* is achieved through the following steps:

1. Batch correction on the aggregate data with scVI method (e.g.`*adata_object.scvi_integrated.h5ad`), and

2. Clustered and annotated by creating embeddings neighbor graph), clustering on the neighbor graph, finding a UMAP embedding for visualization, and annotation with CellAssign (e.g. *adata_object.scvi_integrated.umap_cluster.h5ad, *adata_object.scvi_integrated.umap_cluster.annotate_cells.h5ad) Cell type assignment is done with a very simple taxonomy and marker genes (gs://asap-workflow-dev/workflow-resources/celltype_marker_table.csv)

QC Summaries (Plots)

As advertisements to the harmonized data a family of plots illustrating the Quality Control (QC) statistics and an overall illustration of the dataset. Embedding plots showcasing the distribution of batch labels, doublet scores, cell types, total Count, N Feature, percent-mitochondria, percent-ribosome, and samples are generated. As are a scatter of UMI count vs Gene count, and violin plots for the preprocessing QC metrics.

Final Artifacts

The harmonized data is benchmarked with scib-metrics tools to characterize the quality of biological conservation and batch equalization. A visual summary and table are available (e.g. *..scib_report.csv) The “output” artifacts have “counts”, “normalized counts”, PCs, scVI embeddings, and Harmony equalized PCs available (e.g. *.harmony_integrated.h5ad.)

Metadata & Data Dictionary

The ASAP CRN Common Data Elements (CDE) outlines the variables which are harmonized across the ASAP CRN. These metadata enable the harmonization of the data submitted by multiple teams. The metadata consists of five tables, which are described below, and also detailed as Data Dictionaries for each.

Data Dictionary: ASAP CRN Data Dictionary - v2.1

Metadata refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. Metadata was aggregated using the table-based submission for each team's dataset. The tables - STUDY, PROTOCOL, SUBJECT, SAMPLE, CLINPATH, and DATA - can be thought of as spreadsheets, and each table is prepared as simple .csv files which

constitute the tables outlined in the ASAP CRN CDE (final sheet). These have been aggregated across submissions and unique ASAP IDs generated for each dataset, team, subject and sample. I.e. ASAP_dataset_id, ASAP_team_id, ASAP_subject_id, and ASAP_sample_id.

The metadata can be browsed on ASAP CRN Cloud *Explorer* and exposed in Verily Workbench.

This document.