

ASAP CRN Cloud PMDBS scRNAseq Collection

README - v3.1.2 [10.5281/zenodo.20403941](https://doi.org/10.5281/zenodo.20403941)

Overview

The **PMDBS scRNAseq Collection** has been updated to **v3.1.2** to reflect the updated Datasets' metadata to conform with the **v4.4** Common Data Elements (CDE) with the **v5.0.0** [ASAP CRN Cloud Release](#).

ASAP Teams: Team Hafler, Team Lee, Team Jakobsson, Team Scherzer, Team Hardy, Team Sulzer

See [CRN Cloud Authorship List](#) for the full list of contributing investigators/researchers by dataset.

ASAP CRN GitHub: github.com/ASAP-CRN

Data Release Date: 2026-06-15

ASAP CRN Cloud Release Number: v4.1.0

ASAP CRN Release DOI: [10.5281/zenodo.8384742](https://doi.org/10.5281/zenodo.8384742)

Curated Data File Manifest: [ASAP CRN Cloud File Manifest - v3.1](#)

PMDBS scRNAseq Collection

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. Code, explanations, and schematics for processing the platformed data can be found at the [ASAP-CRN sc-rnaseq-wf](#) repo.

Collection details

scRNAseq Workflow Release - v4.0.0

- Release date: 2025-12-04
- GitHub release [link](#)
- Pipeline version: v4.0.0

Collection Datasets list

- hafler-pmdbs-sn-rnaseq-pfc, DOI:[10.5281/zenodo.15490150](https://doi.org/10.5281/zenodo.15490150)
- lee-pmdbs-sn-rnaseq, DOI:[10.5281/zenodo.16744323](https://doi.org/10.5281/zenodo.16744323)
- hardy-pmdbs-sn-rnaseq, DOI:[10.5281/zenodo.16749080](https://doi.org/10.5281/zenodo.16749080)
- scherzer-pmdbs-sn-rnaseq-mtg, DOI: [10.5281/zenodo.16751625](https://doi.org/10.5281/zenodo.16751625)
- jakobsson-pmdbs-sn-rnaseq, DOI:[10.5281/zenodo.15162834](https://doi.org/10.5281/zenodo.15162834)
- sulzer-pmdbs-sn-rnaseq, DOI:[10.5281/zenodo.17612853](https://doi.org/10.5281/zenodo.17612853)
- cohort-pmdbs-sc-rnaseq, DOI:[10.5281/zenodo.20403941](https://doi.org/10.5281/zenodo.20403941) (collection doi)

Curated Data

Curated data is the result of processing the raw dataset contributions by ASAP CRN Workflows.

For the scRNAseq Datasets 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 of count correction due to ambient RNA molecules and random barcode swapping with `cellbender`, and calculating doublet metrics with `scrublet`. By merging all samples, and filtering the low quality cells we derive the QC-ed gene expression matrix for all cells in the dataset. 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

Processed. Processing refers to *feature selection*, yielding two key artifacts: an unprocessed (but QC-ed) data object containing all cells from all samples (e.g. `*merged_adata_object.h5ad`) and a final processed anndata object containing the gene expression of highly variable genes for all cells (e.g. `*final.h5ad`) and as per the [ASAP-CRN pmdbs-sc-rnaseq-wf](#). Prior to feature selection the [MapMyCells](#) is run on the SEA-AD reference to make an assumption free inference to brain cell types from the full gene-expression count matrices. Feature selection found the top genes with the [highly_variable_genes method from Pearson's residuals](#) 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](#), and
2. Assigning remaining “UNKNOWN” cell-types (due to poor MMC fidelity) using scANVI.
3. Clustered and annotated by creating embeddings neighbor graph), clustering on the neighbor graph at multiple Leiden resolutions, finding a UMAP embedding for visualization
4. An additional Harmony integration step is performed to make alternative batch equalization to `scVI`, which also enables quantitative assessment of the batch related variance with respect to preservation of biological variance (via `scib-metrics`)

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. `*.final.h5ad.`)

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](#). `gs://asap-raw-<team_name>-<source>-<dataset_name>`

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.

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**, **DATA**, etc. - 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](#). 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](#).