Polly Verified: v1.0

Author Processed Single Cell Data on Polly

Polly
by Elucidata

WHAT ARE AUTHOR PROCESSED SINGLE CELL DATASETS?

Polly offers author-processed single-cell datasets that are processed using parameters specified in the associated publication. These datasets incorporate all accessible source metadata, including cell type information when provided. Additionally, they include 32 metadata fields meticulously curated by Polly's harmonization engine. These datasets have two associated files:
(i) author-processed H5AD file containing normalized counts and raw counts for filtered cells and genes. Cell annotations are also included in this file based on marker genes documented in the publication.
(ii) The second file consists of pre-QC unfiltered cells with raw count matrix and all associated source and Polly harmonized metadata.

QUALITY ASSURANCE & QUALITY CHECKS PERFORMED

Metadata Quality Checks

Checks for Schema Compliance and Lexical Errors

- Ensure each metadata field follows the allocated format (List/String/Number/Boolean).
- Ensure labels are in a standardized format with consistent default values.
- Ensure there are no extra spaces or special characters.

Ontology Checks

Ensure curated ontology-backed fields follow the corresponding ontologies.

Checks for Logical Errors

- Ensure that the dataset and sample-level metadata are in agreement.
- Ensure the number of samples in the dataset-level metadata matches the count of samples in the dataset.

Completeness Checks

- Ensure none of the metadata fields are empty.
- Ensure that QC metrics calculated during processing are present.
- Ensure mandatory metadata fields such as sample, publication link, and highly variable genes are present.

Checks for Publication Information

Ensure dataset level metadata fields for publication related information are not empty.

Other Checks

- Provide complete metadata from multiple sources (if applicable).
- Ensure custom-curated fields(if requested) are added.
- Ensure the link to the source/publication is added and valid.
- Ensure there are no missing samples.
- Ensure the number of cells in the publication and the H5AD file match.
- Ensure Gene ID is unique and standardized, with no mixed gene annotations present.
- Ensure only RNA-seq samples are captured and there are no mixed samples (In some cases, scATAC-seq and scRNA-seq samples are both added to the same dataset).

QUALITY CHECKS PERFORMED

Data Quality Checks

Checks on Raw Counts Matrix

- Ensure raw and processed counts are scientifically valid and are different.
- Ensure raw counts Anndata layer does not have normalized counts.
- Ensure cell barcodes corresponding to each cell are unique.

Scanpy Quality Metrics

- Ensure QC metrics metadata fields (n_genes_by_counts, total_counts, pct_counts_mt) generated from the scanpy pipeline are present.
- Ensure QC metrics metadata fields not part of the scanpy pipeline are not present or are removed (if present).
- Ensuring the defined threshold for QC filtering is apt.
- Ensuring samples with fewer or low-quality cells are highlighted before downstream analysis.

Analysis Details

Ensure the following parameters are provided in the dataset metadata:
min_gene, min_cell, n_genes_by_counts, total_counts, pct_counts_mt, batch_correction_tool, batch_correction_column.

Checks for UMAP/tSNE Projections

- Ensure either UMAP or tSNE or both projections are present
- Ensure replicability of UMAP/tSNE plots from the publication.

Cluster /information

- Ensure cell cluster information is present.
- Ensure the number of clusters is within a reasonable range (upto 50 for 30K cells)

Batch Effect

- Ensure that there is no batch effect across the samples.

Checks for Cell type Annotation

- Ensure cell types are curated according to author-provided markers in the publication.
- Ensure cell number distribution of each cell type matches with the publication.
- Ensure cell types in the dataset and sample-level metadata are in agreement.
- Ensure curated cell types are the closest ontological terms for author-provided cell types.
- Ensure cell type annotation is performed correctly by mapping of each cluster to a unique cell type.