Polly Verified: v1.0

Bulk RNASeq Data on Polly

Polly by Elucidata

WHAT ARE BULK RNASEQ DATASETS ON POLLY?

Polly offers bulk RNA-Seq datasets that are processed via a uniform upstream processing pipeline. These datasets incorporate all accessible source metadata. Additionally, they include metadata fields meticulously curated by Polly's harmonization engine. These datasets are packaged in a GCT file format.

The details of all the checks performed on every dataset before delivery are listed here. These are performed at both the dataset-level and the sample-level fields on Polly.

QUALITY ASSURANCE AND QUALITY CHECKS PERFORMED

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.
- Ensure the curated metadata passes the following additional checks -
  - Case sensitivity
    - ‘none’ should be lowercase at the dataset and sample-level
    - ‘Normal’, ‘True’, ‘False’ labels should be present in camel case.
  - Singular and plural terms should be verified from the publication and the schema.

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.
- Ensure the following additional checks are performed -
  - Cell lines and species should match.
  - Year and ‘total_num_samples’ should be greater than 0.
  - Re-verification should be performed in case all the curated fields are ‘none’.
  - For the curated fields, ‘none’ or empty values should not be present along with non-none values.

Completeness Checks

- Provide complete metadata from multiple sources (if applicable).
- Ensure none of the metadata fields are empty.
- Ensure all the samples available on source are present on Polly
- Ensure mandatory metadata fields such as source or publication links are present.

Checks for Publication Information

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

114 Sansome Street
San Francisco, CA 94104

QUALITY ASSURANCE AND QUALITY CHECKS PERFORMED

Checks on Raw Counts Matrix

- Ensure raw are scientifically valid and the latest upstream pipeline has been used.
- Ensure that the GCT file is parsable in Python and R.
- Ensure the dataset can open on Phantasus

Checks on the GCT File

- Ensure the row index doesn't have any NaN values.
- Ensure the column index doesn't have any NaN values.
- Ensure the GCT data matrix doesn't have any NaN values.
- Ensure

Other Checks

- Ensure that the datasets' filters, card view and table view function properly on Polly's GUI.
- Ensure the metadata table and sunburst plot function properly on the details page of the dataset.
- Ensure custom-curated fields (if requested) are added and harmonized.
- Ensure Gene ID is unique and standardized, with no mixed gene annotations present.
- Ensure all datasets have consistent row metadata (as per the defined row metadata columns).
- 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).
- Ensure the dataset-level metadata field providing the Polly Verification status is present and correct.