Dataset documentation

Tim Barry

2022-05-13

R Markdown

All data for the SCEPTRE2 project are available in the `sceptre2` offsite directory. The `sceptre2` directory has two top-level folders: `data` and `results`. The `data` folder has the following structure:

```
>-- frangieh
|   >-- co_culture
|   |   >-- gene
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_cell_qc.rds
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_cell_qc.rds
|   |   |   >-- metadata_orig.rd
|   |   |   >-- metadata_qc.rds
|   |   >-- protein
|   |       >-- matrix.odm
|   |       >-- metadata_cell_qc.rds
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- control
|   |   >-- gene
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_cell_qc.rds
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_cell_qc.rds
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- protein
|   |       >-- matrix.odm
|   |       >-- metadata_cell_qc.rds
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- ifn_gamma
|       >-- gene
|       |   >-- matrix.odm
|       |   >-- metadata_cell_qc.rds
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- grna
|       |   >-- matrix.odm
|       |   >-- metadata_cell_qc.rds
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- protein
|           >-- matrix.odm
|           >-- metadata_cell_qc.rds
|           >-- metadata_orig.rds
|           >-- metadata_qc.rds
>-- liscovitch
|   >-- experiment_big
|   |   >-- chromatin
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_cell_qc.rds
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |       >-- matrix.odm
|   |       >-- metadata_cell_qc.rds
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- experiment_small
|       >-- chromatin
|       |   >-- matrix.odm
|       |   >-- metadata_cell_qc.rds
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- grna
|           >-- matrix.odm
|           >-- metadata_cell_qc.rds
|           >-- metadata_orig.rds
|           >-- metadata_qc.rds
>-- papalexi
|   >-- eccite_screen
|       >-- gene
|       |   >-- matrix.odm
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- grna
|       |   >-- matrix.odm
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- protein
|           >-- matrix.odm
|           >-- metadata_orig.rds
|           >-- metadata_qc.rds
>-- schraivogel
|   >-- enhancer_screen_chr11
|   |   >-- gene
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |       >-- matrix.odm
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- enhancer_screen_chr8
|   |   >-- gene
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |       >-- matrix.odm
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- ground_truth_perturbseq
|   |   >-- gene
|   |   |   >-- matrix.odm
|   |   |   >-- metadata_orig.rds
|   |   |   >-- metadata_qc.rds
|   |   >-- grna
|   |       >-- matrix.odm
|   |       >-- metadata_orig.rds
|   |       >-- metadata_qc.rds
|   >-- ground_truth_tapseq
|       >-- gene
|       |   >-- matrix.odm
|       |   >-- metadata_orig.rds
|       |   >-- metadata_qc.rds
|       >-- grna
|           >-- matrix.odm
|           >-- metadata_orig.rds
|           >-- metadata_qc.rds
>-- simulated
    >-- experiment_1
        >-- gene
        |   >-- matrix.odm
        |   >-- metadata_qc.rds
        >-- grna
            >-- matrix.odm
            >-- metadata.rds
            >-- metadata_qc.rds
```

The data are organized into three hierarchical levels: (i) paper, (ii) dataset, and (iii) modality. “Paper” is the paper from which a given dataset came (one of `frangieh`, `liscovitch`, `schraivogel`, `papalexi`, and `simulated`); “dataset” is the name of a given dataset (for example, within the `schraivogel` directory, one of `enhancer_screen_chr11`, `enhancer_screen_chr8`, `ground_truth_perturbseq`, and `ground_truth_tapseq`); and “modality” is the name of a given modality within the dataset (one of `gene`, `grna`, `protein`, and `chromatin`). Each leaf node has a file path (relative to the `data` directory) of the following form:

`paper name / dataset name / modality name`

For example, `schraivogel/enhancer_screen_chr11/gene` contains data on the gene modality of the “chromosome 11 enhancer screen” dataset from the Schraivogel paper. Similarly, `papalexi/eccite_screen/grna` contains data on the gRNA modality of the ECCITE-seq screen dataset from the Papalexi paper.

A given leaf node contains (at least) three files: `matrix.odm`, `metadata_orig.rds`, and `metadata_qc.rds`.

`matrix.odm` is a symbolic link to the backing .odm file of the ondisc matrix; `metadata_orig.rds` is a symbolic link to the "raw" (i.e., un-QC'ed) `metadata.rds` file; and `metadata_qc.rds` is the analysis-ready, QC'ed `metadata.rds` file. (This latter file likewise is a symbolic link if the raw `metadata.rds` file and the QC'ed `metadata.rds` file coincide, i.e., no further QC was performed.) Some leaf nodes (e.g., `frangieh/co_culture/gene`) contain a fourth file called `metadata_cell_qc.rds`. This file is an intermediate file and can be ignored.

The `simulated` top-level directory contains a single subdirectory: `experiment_1`. `experiment_1`, in turn, contains `grna` and `gene` subdirectories. Note that the backing `.odm` files in these subdirectories are *not* symbolic links but instead are moderately large files containing the simulated gene and gRNA expression data.

Feature QC

All modalities of all datasets (except for the gRNA modality) were filtered for features expressed in at least 0.005 of cells. No further feature QC was performed.

Cell QC

We describe the cell QC that we performed for each dataset.

1. **Schraivogel**: the Schraivogel data presumably came equipped with mild cell QC; we did not perform any additional cell QC.
2. **Papalexi**: the Papalexi data came equipped with some mild cell QC; we did not perform any additional cell QC.
3. **Liscovitch**: We followed the QC steps of described by Liscovitch to filter for high-quality cells: we required >500 ATAC fragments and >100 gRNA reads per cell. Furthermore, for a given cell, we required the maximum gRNA count divided by the sum of the gRNA counts to exceed a certain threshold, where this threshold was set to 99% for the small experiment and 90% for the large experiment.
4. **Frangieh**: We followed the QC steps described in the paper, which amounted to filtering for cells with exactly one assigned gRNA.

Note that the `metadata_qc.rds` (described above) is the product of *both* the feature and cell QC.

Simulated data

We generated the simulated data as follows.

Gene

Let p denote the number of genes and n the number of cells. We sampled gene-specific means $\mu_1, \ldots, \mu_p \sim \text{Gamma}(1, 2)$ and gene-specific sizes $\theta_1, \ldots, \theta_p \sim \text{Unif}(5, 30)$. Next, for a given gene j with mean μ_j and size θ_j, we sampled expressions $Y_{1,j}, \ldots, Y_{n,j} \sim \text{NBinom}\left(\mu_j, \theta_j\right)$. We defined the gene expression matrix Y as $Y = \{Y_{i,j}\}_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}}$. Finally, we filtered Y for genes expressed in at least 0.005 of cells. Setting $n = 20,000$ and $p = 10,000$ and applying the above procedure, we produced an expression matrix with 9,915 genes and 20,000 cells. We load and print the gene expression matrix below.

```
library(ondisc)
sim_data_dir <- paste0(.get_config_path("LOCAL_SCEPTRE2_DATA_DIR"), "data/simulated/experiment_1/")
gene_odm_fp <- paste0(sim_data_dir, "gene/matrix.odm")
gene_metadata_fp <- paste0(sim_data_dir, "gene/metadata_qc.rds")
```

```
gene_odm <- read_odm(odm_fp = gene_odm_fp, metadata_fp = gene_metadata_fp)
gene_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 9915 features and 20000 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero.
```

gRNA

Let $d = 35$ denote the number of gRNAs (all negative control). For $i \in \{1, \ldots, n\}$, let $g_i \in \{1, \ldots, d\}$ be a draw from the uniform distribution over $\{1, \ldots, d\}$. Next, let $W_1, \ldots, W_n \sim \text{Pois}(100)$, and let $W_i^{\geq 1} = \max\{1, W_i\}$ for all $i \in \{1, \ldots, n\}$. Finally, let X_i be a vector with the value $W_i^{\geq 1}$ in position g_i and 0 elsewhere. We form the gRNA matrix X by concatenating the X_is. We load and print the gRNA count matrix below.

```
gRNA_odm_fp <- paste0(sim_data_dir, "gRNA/matrix.odm")
gRNA_metadata_fp <- paste0(sim_data_dir, "gRNA/metadata_qc.rds")
gRNA_odm <- read_odm(odm_fp = gRNA_odm_fp, metadata_fp = gRNA_metadata_fp)
gRNA_odm
```

```
## A covariate_ondisc_matrix with the following components:
##  An ondisc_matrix with 35 features and 20000 cells.
##  A cell covariate matrix with columns n_nonzero, n_umis.
##  A feature covariate matrix with columns mean_expression, coef_of_variation, n_nonzero, target_type,
```