Simulated dataset documentation

Tim Barry

2022-05-03

This note documents the simulated dataset used in SCEPTRE 2.

Directory structure

The simulated dataset is stored in `sceptre2/data/simulated`. There are two subdirectories: `gRNA` and `gene`.

```
-- gRNA
    -- expressions.odm
    -- metadata.rds
-- gene
    -- expressions.odm
    -- metadata.rds
```

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 $Y_{1,j}, \ldots, Y_{n,j} \sim \text{NBinom}(\mu_j, \theta_j)$. 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/")
gene_odm_fp <- paste0(sim_data_dir, "gene/expressions.odm")
gene_metadata_fp <- paste0(sim_data_dir, "gene/metadata.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.

```
library(ondisc)
sim_data_dir <- paste0(.get_config_path("LOCAL_SCEPTRE2_DATA_DIR"), "data/simulated/")
gRNA_odm_fp <- paste0(sim_data_dir, "gRNA/expressions.odm")
gRNA_metadata_fp <- paste0(sim_data_dir, "gRNA/metadata.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,
```