# Introduction to scDesign

Wei Vivian Li

2020-12-13

## `design_data`

`design_data` simulates additional scRNA-seq data by estimating gene expression parameters from a real scRNA-seq dataset. When `ngroup = 1`, it each time simulates a single dataset based on user-specified total read number `S` and cell number `ncell`.

```r
realcount1 = readRDS(system.file("extdata", "astrocytes.rds", package = "scDesign"))
simcount1 = design_data(realcount = realcount1, S = 1e7, ncell = 1000, ngroup = 1, ncores = 1)

realcount1[1:3, 1:3]
#>             GSM1657885 GSM1657932 GSM1657938
#> 1/2-SBSRNA4          0          0          0
#> A2M                  0          0         34
#> A2ML1                0          0         25
simcount1[1:3, 1:3]
#>       cell1 cell2 cell3
#> gene1     0     0     0
#> gene2     0     0    68
#> gene3     0     0     1
```

When `ngroup > 1`, it simulates `ngroup` datasets following a specified differentiation path. The key parameters are

- `ngroup` number of cell states
- `S`: total read number for each cell state
- `ncell`: cell number for each state
- `pUp`: proportion of up-regulated genes between two adjacent states
- `pDown`: proportion of down-regulated genes between two adjacent states
- `fU`: upper bound of fold changes of DE genes' expression
- `fL`: lower bound of fold changes of DE genes' expression

```r
simdata = design_data(realcount = realcount1, S = rep(1e7,3), ncell = rep(100,3), ngroup = 3,
                   pUp = 0.03, pDown = 0.03, fU = 3, fL = 1.5, ncores = 1)

# simdata is a list of three elements
names(simdata)
#> [1] "count"     "genesUp"   "genesDown"

# count matrix of the cell state 2
simdata$count[[2]][1:3, 1:3]
#>       C2_1 C2_2 C2_3
#> gene1  132    0    0
#> gene2    6    2    6
#> gene3    0    0    0
```

```
# up-regulated genes from state 1 to state 2
simdata$genesUp[[2]][1:3]
#> [1] "gene1655" "gene614"  "gene6057"

# down-regulated genes from state 1 to state 2
simdata$genesDown[[2]][1:3]
#> [1] "gene1958" "gene4631" "gene4888"
```

If users would like to specify the gene expression mean parameters (e.g., estimated from bulk data) instead of letting scDesign estimate them from the real scRNA-seq data, this can be done by setting the `exprmean` parameter in `design_data`. The provided mean expression should be on the $log10$ scale. Note that `exprmean` should be a named vector and its names should match the gene names (i.e., rownames) of `realcount1`. Please see example code below:

```
realcount1 = readRDS(system.file("extdata", "astrocytes.rds", package = "scDesign"))
simcount1 = design_data(realcount = realcount1, S = 1e7, ncell = 1000,
ngroup = 1, ncores = 1, exprmean = exprmean)
```

## design_sep

`design_sep` assists experimental design by selecting the optimal cell numbers for the two cell states in scRNA-seq, so that the subsequent DE analysis becomes most accurate based on the user-specified criterion. It assumes that cells from the two states are prepared as two separate libraries and sequenced independently. Key parameters include

- `realcount1`: a real count matrix of cell state 1
- `realcount2`: a real count matrix of cell state 2
- `S1`: total number of RNA-seq reads for cell state 1. Default to 1e8
- `S2`: total number of RNA-seq reads for cell state 2. Default to 1e8
- `ncell`: a two-column matrix specifying the candidate numbers of cells

```
realcount1 = readRDS(system.file("extdata", "astrocytes.rds", package = "scDesign"))
realcount2 = readRDS(system.file("extdata", "oligodendrocytes.rds", package = "scDesign"))

# candidate cell numbers for experimental design
ncell = cbind(2^seq(6,11,1), 2^seq(6,11,1))
prlist = design_sep(realcount1, realcount2, ncell = ncell, de_method = "ttest", ncores = 10)

# returns a list of five elements
# precision, recall, TN (true negative rate),
# F1 (harmonic mean of precision and recall),
# F2 (harmonic mean of TN and recall)
names(prlist)
#> precision  recall  TN  F1  F2
prlist$precision
#> p_thre 64vs64 128vs128 256vs256 512vs512 1024vs1024 2048vs2048
#> 0.01   0.332  0.272    0.178    0.121    0.084      0.056
#> 0.001  0.448  0.361    0.231    0.147    0.097      0.063
#> 1e-04  0.532  0.434    0.282    0.175    0.11       0.07
#> 1e-05  0.599  0.491    0.331    0.203    0.124      0.076
#> 1e-06  0.649  0.534    0.375    0.231    0.138      0.083
```

`design_sep` also saves the analysis results to a txt file [REF] and a set of power analysis plots [REF].

## design_joint

design_joint assists experimental design by selecting the optimal (total) cell number for a cell population that contains the two cell states of interest, so that the subsequent DE analysis becomes most accurate based on the user-specified criterion. It assumes that cells from the two states are prepared in the same library and sequenced together. Key parameters include

- realcount1: a real count matrix of cell state 1
- realcount2: a real count matrix of cell state 2
- S: the total number of RNA-seq reads for the cell population. Default to 1e8
- ncell: the (candidate) total number of cells
- prop1: the proportion of state 1 cells in the cell population
- prop2: the proportion of state 2 cells in the cell population

```
# candidate cell numbers for experimental design
ncell = round(2^seq(9,13,1))
prlist = design_joint(realcount1, realcount2, prop1 = 0.2, prop2 = 0.15,
                      ncell = ncell, de_method = "ttest", ncores = 10)

# returns a list of five elements
names(prlist)
#> precision  recall  TN  F1  F2
prlist$recall
#>       512   1024  2048  4096  8192
#> 0.01  0.315 0.33  0.259 0.176 0.111
#> 0.001 0.235 0.281 0.24  0.169 0.108
#> 1e-04 0.176 0.236 0.22  0.162 0.105
#> 1e-05 0.133 0.198 0.2   0.155 0.102
#> 1e-06 0.103 0.166 0.181 0.147 0.099
```