DKOsimR Tutorial

Tutorial: How to generate synthetic CRISPR data using DKOsimR?

Abbreviations: SKO, single knockout; DKO, double knockout; %, percentage; GI, genetic interaction; std. dev., standard deviation.

1. Introduction

This tutorial introduces DKOsimR package for generating synthetic CRISPR double knockout data. It mainly includes:

- Full list of all tunable parameters to initialize the simulation.
- Descriptions on how users may generate synthetic CRISPR data in two modes:
  - default simulation, with 4 gene class: negative, wild-type, non-targeting control, positive
  - simulation applicable to approximate real lab data, with 3 gene class: negative (essential), unknown, non-targeting control
- Examples of applying genetic interaction detection methods on simulated datasets.

2. Install and Load DKOsimR

Simply install and load the package into your R session with following commands:

```
if(!requireNamespace("devtools", quietly = TRUE))
  install.packages("devtools")
devtools::install_github("yuegu-phd/DKOsimR", quiet = TRUE)
#> Installing 1 packages: data.table
library(DKOsimR)
```

Make sure all required dependencies are installed using `devtools::install(dependencies = TRUE)`

3. List of tunable parameters:

Let's go through the designed parameters in simulation framework. Below list out all tunable parameters and how it is represented in DKOsimR.

- Initialized Library Parameters
  - `sample_name`: name of the simulation run
  - `coverage`: cell representation per guide
  - `n`: number of unique single gene
  - `n_guide_g`: number of guide per gene
  - `moi`: multiplicity of infection - % of cells that are transfected by any virus
  - `sd_freq0`: dispersion of initial counts distribution
- GI Parameters:
  - `p_gi`: % of genetic interaction presence
  - `sd_gi`: std. dev. of re-sampled phenotype with gi presence
- Guide Parameters:
  - `p_high`: % of high-efficacy guides
  - `mode`: CRISPR mode:
    - use `CRISPRn-100%Eff` if need 100% efficient guides without randomization

    * use `CRISPRn` if need high-efficient guides draw from distribution
- Gene Class Parameters:
  1. % of theoretical phenotype to each gene class - make sure they add up to 1
  - `pt_neg`: % negative
  - `pt_pos`: % positive
  - `pt_wt`: % wild-type
  - `pt_ctrl`: % non-targeting control
  2. Mean and std. dev. of theoretical phenotype
  - `mu_neg`: mean of negative genes
  - `sd_neg`: std. dev. of negative genes
  - `mu_pos`: mean of positive genes
  - `sd_pos`: std. dev. of positive genes
  - `sd_wt`: std. dev. of wild-type genes
- Bottleneck Parameters:
  - `size.bottleneck`: bottleneck size - threshold indicating the ceiling of cell growth
  - `n.bottlenecks`: number of bottleneck encounters - how many times do we encountering bottlenecks?
  - `n.iterations`: number of maximum doubling cycles, by default, we assume a maximum of 30 doublings if we didn't encounter bottleneck
- Randomization Parameter:
  - `rseed`: values used for random number generator - use same number to control same sets of genes having GI

4. Running Simulation

To generate synthetic double knockout data, by default, values parameters of simulated CRISPR screens are set based on empirical assumptions as follows:

- Initialized Library Parameters
  - `coverage`: 100
  - `n_guide_g`: 3
  - `moi`: 0.3
  - `sd_freq0`: 1/3.29 (chosen by setting a 10-fold difference between 95th and 5th percitiles of SKO counts distribution)
- GI Parameters:
  - `p_gi`: 0.03
  - `sd_gi`: 1.5
- Guide Parameters:
  - `p_high`: 1
  - `mode`: CRISPR mode: `CRISPRn-100%Eff`
- Gene Class Parameters:
  1. % of theoretical phenotype to each gene class - make sure they add up to 1
  - `pt_neg`: 0.15
  - `pt_pos`: 0.05
  - `pt_wt`: 0.75
  - `pt_ctrl`: 0.05
  2. Mean and std. dev. of theoretical phenotype
  - `mu_neg`: -0.75
  - `sd_neg`: 0.1
  - `mu_pos`: 0.75
  - `sd_pos`: 0.1
  - `sd_wt`: 0.25
- Bottleneck Parameters:
  - `size.bottleneck`: 2
  - `n.bottlenecks`: 1

  - `n.iterations`: 30
- Randomization Parameter:
  - `rseed`: 888

To run simulation by default, simply name your simulation by `sample_name` and specify the number of single genes by `n`. Be cautious that number of genes in each gene class has to be an integer. A quick Simulation Settings Summary would be returned for each run, and the Run Time in the unit of hours would be collected after one successful run. An example running code is

```
dkosim(sample_name="test", n=60)
#>
#> # ------------------------------------------------------------
#> # Simulation Settings Summary:
#> # ------------------------------------------------------------
#> # Sample Name: test
#> # Number of Genes: 60
#> # Cell Library Size (Initial): 1611000
#> # Coverage: 100 x
#> # Number of Single Knockout(SKO): 60
#> # Number of Double Knockout(DKO): 1770
#> # Number of Guides per Gene: 3
#> # Number of Constructs: 16110
#> # Variance of Initialized Counts: 0.09
#>
#> # Genetic Interactions (GI):
#> ## Proportion of GI(%): 3
#> ## Number of Interacting Gene Pairs: 53
#> ## Variance of re-sampled phenotypes w/ GI: 2.25
#>
#> # Proportion of Each Initialized Gene Class (by theoretical phenotypes):
#> ## Negative(%): 15 ~ TN( -0.75 , 0.01 ,-1,-0.025)
#> ## Positive(%): 5 ~ TN( 0.75 , 0.01 ,0.025, 1)
#> ## Wild-Type(%): 75 ~ TN(0, 0.0625 ,-0.025, 0.025)
#> ## Non-Targeting Control(%): 5 ~ Delta(0)
#>
#> # Proportion of Guides (by efficacy):
#> ## High-efficacy(%): 100 ~ 1
#> ## Low-efficacy(%): 0 ~ TN(0.05, 0.0049, 0, 1)
#>
#> # Multiplicity of Infection (MOI): 0.3
#> # Percentage of viral particles delivered in cells during transfection(%): 22.22 ~ Poisson( 0.3 )
#> # Resampling Size based on MOI (Passage Size): 716075
#> # Bottleneck Size ( 2 x Initial Guide-Level Library Size): 3222000
#> # Number of Bottleneck Encounters (Number of Passages): 1
#> # Total Available Doublings: 30
#> # Number of Replicates: 2
#> # Pseudo-count: 5e-07
#>
#> # ------------------------------------------------------------
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
```

```
#>
#>     intersect, setdiff, setequal, union
#> Loading required package: iterators
#> Loading required package: parallel
#> Warning: package 'data.table' was built under R version 4.3.3
#>
#> Attaching package: 'data.table'
#> The following objects are masked from 'package:dplyr':
#>
#>     between, first, last
#> [1] "repA completed" "repB completed"
#> [1] "number of cores 14"
#> [1] "Run Time (hrs):  0.00793916666666667"
```

You may change the values to any tunable parameters as desired, but please make sure your input on percentage of each gene class add up to 1 for all classes, and each initialized number of genes is an integer. An example running code is

```
dkosim(sample_name="test",
       coverage=100,
       n=60,
       n_guide_g=3,
       sd_freq0 = 1/3.29,
       moi = 0.3,
       p_gi=0.03,
       sd_gi=1.5,
       p_high=1,
       mode="CRISPRn-100%Eff",
       pt_neg=0.15,
       pt_pos=0.05,
       pt_wt=0.75,
       pt_ctrl=0.05,
       mu_neg=-0.75,
       sd_neg=0.1,
       mu_pos=0.75,
       sd_pos=0.1,
       sd_wt=0.25,
       size.bottleneck = 3,
       n.bottlenecks= 2,
       n.iterations = 30,
       rseed = 111)
#>
#> # ------------------------------------------------------------
#> # Simulation Settings Summary:
#> # ------------------------------------------------------------
#> # Sample Name: test
#> # Number of Genes: 60
#> # Cell Library Size (Initial): 1611000
#> # Coverage: 100 x
#> # Number of Single Knockout(SKO): 60
#> # Number of Double Knockout(DKO): 1770
#> # Number of Guides per Gene: 3
#> # Number of Constructs: 16110
#> # Variance of Initialized Counts: 0.09
```

```
#>
#> # Genetic Interactions (GI):
#> ## Proportion of GI(%): 3
#> ## Number of Interacting Gene Pairs: 53
#> ## Variance of re-sampled phenotypes w/ GI: 2.25
#>
#> # Proportion of Each Initialized Gene Class (by theoretical phenotypes):
#> ## Negative(%): 15 ~ TN( -0.75 , 0.01 ,-1,-0.025)
#> ## Positive(%): 5 ~ TN( 0.75 , 0.01 ,0.025, 1)
#> ## Wild-Type(%): 75 ~ TN(0, 0.0625 ,-0.025, 0.025)
#> ## Non-Targeting Control(%): 5 ~ Delta(0)
#>
#> # Proportion of Guides (by efficacy):
#> ## High-efficacy(%): 100 ~ 1
#> ## Low-efficacy(%): 0 ~ TN(0.05, 0.0049, 0, 1)
#>
#> # Multiplicity of Infection (MOI): 0.3
#> # Percentage of viral particles delivered in cells during transfection(%): 22.22 ~ Poisson( 0.3 )
#> # Resampling Size based on MOI (Passage Size): 1074112
#> # Bottleneck Size ( 3 x Initial Guide-Level Library Size): 4833000
#> # Number of Bottleneck Encounters (Number of Passages): 2
#> # Total Available Doublings: 30
#> # Number of Replicates: 2
#> # Pseudo-count: 5e-07
#>
#> # ------------------------------------------------------------
#>
#> [1] "repA completed" "repB completed"
#> [1] "number of cores 14"
#> [1] "Run Time (hrs):  0.0125233333333333"
```

5. Laboratory Data Pattern Approximation

6. Applying Genetic Interaction Detection Methods on simulated data

Reference

Gu, Y., Hart, T., Leon-Novelo, L., Shen, J.P.. Double-CRISPR Knockout Simulation (DKOsim): A Monte-Carlo Randomization System to Model Cell Growth Behavior and Infer the Optimal Library Design for Growth-Based Double Knockout Screens. bioRxiv 2025.09.11.675497. DOI: 10.1101/2025.09.11.675497.