DKOsimR

Yue Gu

2025-12-22

Tutorial: How to generate synthetic CRISPR data using DKOsimR?

Tunable paramters:

Abbreviations: %, percentage; GI, genetic interaction; std. dev., standard deviation.

- 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

Laboratory Data Pattern Approximation