Package 'SClineager'

March 11, 2020

Type Package

Title Tracing lineage of single cells via genetic variants based on a Bayesian Hierarchical modeling

Version 1.10

Date 2020-03-10

Author Seongoh Park [aut, cre], Tao Wang [aut, cre], Tianshi Lu [aut]

Maintainer Seongoh Park <seongohpark6@gmail.com>

Description We developed a Bayesian hierarchical model that performs lineage tracing of single cells based on genetic markers. The single cell variants are called from single cell sequencing data. Single cell variant calling has two inherent issues: (1) low coverage on many positions in many cells and (2) allelic bias due to true monoallelic expression in single cells or due to sampling bias. This algorithm infers genetic trajectories of cells by taking these two issues into account. More details about the structure of the data can be found in the example dataset that goes along with this R package. The details of the Bayesian model can be found in our upcoming paper.

License GPL-3

Encoding UTF-8

Imports Rcpp (>= 1.0.2), MCMCpack, vioplot, gplots

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.1.1

NeedsCompilation yes

R topics documented:

explore_sclineager *Exploratory analysis by visualization*

Description

Explore data before applying SClineager.

Usage

```
explore_sclineager(folders, coverage_cutoff, file)
```

Arguments

folders	A path where each subfolder has the raw output from one read_sclineager call (for cells of one sample), which contains the "cleaned.RData" file.
coverage_cutoff	Coverage below this cutoff is considered not observable.
file	A file name (extension pdf) at which outputs (figures) are saved. For example, "figure_eda.pdf".

Value

This returns nothing, but figures in `file`.

read_sclineager *Read datsets for SClineager*

Description

Read datsets for SClineager

Usage

```
read_sclineager(runinfo, coverage_cutoff, coverage_percentage,
  cell_percentage, folder, artefact_percentage, species = "hg38",
  keep_other = FALSE)
```

Arguments

runinfo	A data frame with "Path" and "Cell". "Path" is the path to the raw mutation output folder for each cell. "Cell" is the name of each single cell.
coverage_cutoff	Coverage below this cutoff is considered not observable.
coverage_percentage	A numeric value in [0,1]. If a mutation has more than `coverage_percentage` of cells whose coverages are below `coverage_cutoff`, then it will not considered in modeling.

`cell_percentage`
: A numeric value in [0,1]. If a cell has more than `cell_percentage` of mutations not observable, then it will not considered in modeling.

`folder`
: A folder at which outputs are saved.

`artefact_percentage`
: A numeric value in [0,1]. Only keep mutations that appeared in more than `artefact_percentage` of cells.

`species`
: A character value that indicates germline_mutations. Default is "hg38" and other possibilities are "hg19" and "mm10".

`keep_other`
: A logical value whether to keep intronic and intergenic mutations. Default is `FALSE`.

Value

"summary.pdf" and "results.RData" are saved under `folder`. The RData file contains a list with components (also returned):

`mutations_mat`
: The variant allele frequency (VAF) matrix, whose rows are mutations and columns are samples.

`coverage_mat`
: The total sequencing coverage matrix, whose rows are mutations and columns are samples.

`runinfo`
: Same as `runinfo` in the input, but trimmed to contain only cells that appeared in `mutations_mat`, in the same order.

`annotation`
: Annotation information of the variants in `mutations_mat`.

Examples

```
# See the github page https://github.com/inmybrain/SClineager.
```

---

run_sclineager *An wrapper function for running SClineager*

Description

This function wraps the main function sclineager_internal to process data and return outputs.

Usage

```
run_sclineager(file_in, folder, categories, max_iter, keep_genes = NULL,
  mask_genes = NULL, vaf_offset, dfreedom, skip_common, psi = NULL,
  control = NULL, save = FALSE)
```

Arguments

`file_in`
: A path pointing the input file of the .RData file generated by read_sclineager. . It must have a list named "results" with components "runinfo", "mutations_mat", "annotation", and "coverage_mat".

`folder`
: A folder at which outputs are saved. If it does not exist, it will be automatically created.

categories	A character vector consisting of any combination of "frameshift substitution", "nonframeshift substitution", "nonsynonymous SNV", "synonymous SNV", "splicing", "UTR5", "UTR3", "stopgain", and "stoploss".
max_iter	The number of posterior samples. Half of samples are discarded as burn-in steps.
keep_genes	(Optional) A character vector of genes. These genes will only be considered. This can be useful for, for example, limiting the estimation to only the genes that are known to be abundantly expressed in all or the majority of the sequenced cells
mask_genes	(Optional) A character vector of genes. Genes in this vector will be masked from the analyses. It is probably a good idea to mask all HLA genes from the analysis.
vaf_offset	A small number (e.g. 1e-2) for converting VAF to logit VAF.
dfreedom	(Hyperparameter) the degrees of freedom used in inverse Wishart distribution of psi. Should be larger than ncol(mutations_mat)-1.
skip_common	For the variants in consideration, if this flag is set to TRUE, the variants that are common in the human population (esp6500siv2_all>0.01) will be skipped.
psi	(Hyperparameter) covariance matrix of VAF across samples. If NULL (default), it will be set to an identity matrix.
control	Control parameters for determining the variance function relating the observed VAF to the true VAF (see Details). If NULL (default), then a=1, b=0.5, c=1, d=20, e=0.4.
save	Logical. If TRUE, posterior samples are saved and returned. Default is FALSE.

Value

"imputation_results.pdf" and "results.RData" are saved in folder. The .RData file contains a list with components (also returned):

genotype_mat	(Parameter) an estimated VAF matrix, obtained by averaging last half of posterior samples.
genotype_mat_all	A list of posterior samples for VAF. NULL is returned if save is FALSE.
genotype_mat_orig	Same as mutations_mat, but contains the raw VAF data.
sigma	(Parameter) an estimated covariance matrix of VAF across samples, obtained by averaging last half of posterior samples.
sigma_all	A list of posterior samples for a covariance matrix. NULL is returned if save is FALSE.
k_mat	A matrix whose entry is a variance function computed by empirical estimates.
loglike	A vector of posterior likelihoods.
runinfo	Same as runinfo in the input, but trimmed to contain only cells that appeared in mutations_mat, in the same order.
annotation	Annotation information of the variants in mutations_mat.

Examples

```
# See the github page https://github.com/inmybrain/SClineager.
```

sclineager_gibbs *An internal function for Gibbs sampling*

Description

An internal function for Gibbs sampling wrapped by sclineager_internal

Usage

```
sclineager_gibbs(psi, k_mat, transform_mat, mu, dfreedom, sigma,
  max_iter = 100L, save = FALSE)
```

See Also

sclineager_internal

sclineager_internal *The Gibbs sampler for SClineager*

Description

Perform the MCMC sampling to estimate parameters in SClineager.

Usage

```
sclineager_internal(mutations_mat, coverage_mat, max_iter, vaf_offset,
  dfreedom, psi, control = NULL, save = FALSE)
```

Arguments

mutations_mat	(Data) The variant allele frequency (VAF) matrix, whose rows are mutations and columns are samples. Each value is in [0,1] and NA value is for unobservable cells.
coverage_mat	(Data) The total sequencing coverage matrix, whose rows are mutations and columns are samples. Each value is positive real number and NA value is for unobservable cells.
max_iter	See run_sclineager.
vaf_offset	See run_sclineager.
dfreedom	See run_sclineager.
psi	See run_sclineager.
control	See run_sclineager.
save	See run_sclineager.

Value

A list with components:

`genotype_mat` See run_sclineager.

`genotype_mat_all` See run_sclineager.

`genotype_mat_orig` See run_sclineager.

`sigma` See run_sclineager.

`sigma_all` See run_sclineager.

`k_mat` See run_sclineager.

`loglike` See run_sclineager.

See Also

run_sclineager

Examples

```
# See the github page https://github.com/inmybrain/SClineager.
```

Index