Package 'Alleloscope'

October 21, 2020

Type Package

Title Allele-specific copy number genotyping for single cell sequencing data

Version 1.0.0

Author Chi-Yun Wu, Nancy Zhang

Maintainer Chi-Yun Wu <chiyunwu@pennmedicine.upenn.edu>

Description Alleloscope is a method for allele-specific copy number estimation that can be applied to single cell DNA and ATAC sequencing data (separately or in combination), allowing for integrative multi-omic analysis of allele-specific copy number and chromatin accessibility for the same cell.

License GPL-2

Encoding UTF-8

LazyData true

Imports Matrix, matrixStats, ggplot2, stringr, caTools, pheatmap, cowplot, rhdf5, HiddenMarkov, cluster

Depends R (¿= 3.6.0), Matrix, rtracklayer

biocViews Software

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

R topics documented:

`AssignClones_ref` *Using marker regions to assign each cell into c reference subclones*

Description

rhohats, thetahats, snpCoverages are n by m matrices for n cell and m marker regions. The genotypes (rho, theta) are: 1.(0.5,0); 2.(0.5,1); 3.(1,0); 4.(1,0.5); 5.(1,1); 6.(1.5,0); 7.(1.5,1/3); 8.(1.5,2/3); 9.(1.5,1); 10.(2,0); 11.(2,1/4); 12.(2,2/4); 13.(2,3/4); 14.(2,4/4) 15.(2.5,0); 16.(2.5,1/5); 17.(2.5,2/5); 18.(2.5,3/5); 19.(2.5,4/5); 20.(2.5,1); 21.(3,0); 22.(3,1/6); 23.(3,2/6); 24.(3,3/6); 25.(3,4/6); 26.(3,5/6); 27.(3,6/6)

Usage

```
AssignClones_ref(
  rhohats,
  thetahats,
  snpCoverages,
  priorCloneProbs = NULL,
  clone.genotypes,
  sigma.rho
)
```

Arguments

`rhohats:`	n by m matrix of rho_hat values for each cell across the regions.
`thetahats:`	n by m matrix of theta_hat values for each cell across the regions.
`snpCoverages:`	n by m matrix of total read counts covering SNPs for each cell across the regions.
`priorCloneProbs:`	A numeric vector indicating prior prior probability of each subclone.
`clone.genotypes:`	c by m matrix of numbers representing different genotypes for each clone and each maker region (known from scDNA-seq).
`sigma.rho:`	Numeric. Standard deviation of the rho_i values under normal distribution.

`Createobj` *Generate Alleloscope object for analysis*

Description

Generate Alleloscope object for analysis

Usage

```
Createobj(
  alt_all = NULL,
  ref_all = NULL,
  var_all = NULL,
  samplename = "sample",
  genome_assembly = "GRCh38",
  dir_path = "./",
  barcodes = NULL,
  size = NULL,
  assay = "scDNAseq"
)
```

Arguments

alt_all	A SNP by cell read count matrix/ spare matrix for the alternative alleles.
ref_all	A SNP by cell read count matrix/ spare matrix for the reference alleles.
samplename	Sample name for the data.
genome_assembly	The genome assembly used for sequencing alignment. (ex: "GRCh38" or "GRCh37")
dir_path	Path of the output directory.
barcodes	A matrix/ data.frame with barcodes for each cell in the first column.
size	A numeric vector for the size (bp) of different chromosomes (with the names indicating which chromosome from 1 to 22)
assay	A character indicating the type of sequencing data. (ex: "scDNAseq" or "scATACseq")
vcf_all	A matrix/ data.frame of the vcf format for SNP information. (The length and order are the same as nrow(alt_all) and nrow(alt_all))

Value

A Alleloscope object including the necessary information.

EM *Iterative phasing and theta_hat estimation*

Description

Iterative phasing and theta_hat estimation

Usage

```
EM(ref_table, alt_table, max_iter = max_iter, seed = 2020)
```

Arguments

ref_table	A SNP by cell read count matrix/ spare matrix for the reference alleles.
alt_table	A SNP by cell read count matrix/ spare matrix for the alternative alleles.
max_iter	An integer of maximum iteration number.
seed	An integer of random seed number for EM initialization.

Value

A list of estimated indicators (I_hat) for each SNP and estimated major haplotype proportion (theta_hat) for each cell in one region. I_hat is the phasing result indicating whether reference allele is on the major haplotype for each SNP. Theta_hat represents the CNV states for each cell. A cell is considered as a CNV carrier if its theta_hat depart from 0.5.

`Est_regions` *Perform EM iterations on the filtered cells with barcodes, and plot the results for each region.*

Description

Perform EM iterations on the filtered cells with barcodes, and plot the results for each region.

Usage

```
Est_regions(
  Obj_filtered = NULL,
  max_nSNP = 30000,
  plot_stat = TRUE,
  min_ncell = 20,
  rds_path = NULL,
  cont = FALSE,
  max_iter = 50,
  phases = NULL
)
```

Arguments

`Obj_filtered`	A Alleloscope object with allele and segment information for estimating cell major haplotype proportion (theta_hat) for each region.
`max_nSNP`	Integer. Maximum SNP number used for estimating theta_hat for a region.
`plot_stat`	Logical (TRUE/ FALSE). Whether or not to plot the statistics and EM results for each region.
`min_ncell`	Integer. Filter out the cells with reads ¡ min_ncells.
`rds_path`	The path for saving the rds files for the estimated results for each region.
`cont`	Logical (TRUE/FALSE). Whether or not to skip the regions with the rds files in the specified rds_path.
`max_iter`	Integer. Maximum numbers of EM iterations.
`phases`	List. The estimated phase indicators (I_j) of each SNP across all regions.

Value

A "rds_list" of the estimated SNP phases (I_hat), estimated cell major haplotype proportion (theta_hat) for all regions.

`Genotype` *Genotype each cell for each region and plot the genotypes.*

Description

Genotype each cell for each region and plot the genotypes.

Usage

```
Genotype(Obj_filtered = NULL, xmax = NULL, plot_path = NULL, ref_gt = NULL)
```

Arguments

`Obj_filtered`	A Alleloscope object with a n cell by (m region * 2) genotype_values matrix and seg_table_filtered matrix. Every 2 columns in the genotype_table matrix are (rho_hat, theta_hat) of each region.
`xmax`	An integer for the x-axis maximum limit.
`plot_path`	The path for saving the plot.
`ref_gt`	A reference "genotypes" (from scDNA-seq) to help with genotype estimation.

Value

A list of ggplot objects of the genotyping results for all the regions.

`Genotype_value` *Normalize coverage using identified/ specified normal cells and one normal region and generate a table with (rho_hat, theta_hat) of each cell for all regions.*

Description

rho_hat: Relative coverage change for each cell in a region theta_hat: Major haplotype proportion fir each cell in a region

Usage

```
Genotype_value(
  Obj_filtered = NULL,
  type = "tumor",
  raw_counts = NULL,
  ref_counts = NULL,
  cov_adj = 1,
  ref_gtv = NULL
)
```

Arguments

Obj_filtered	A Alleloscope object with theta_hat info in the rds_list and identified/ specified normal cells and a normal region
type	Specify whethere the sample is a "tumor" or "cellline". If "type" is a "cellline", param "ref_counts" needs to be specified for normal sample.
raw_counts	(required) A large binned coverage matrix (m1 bin by n1 cell) with values being read counts for all chromosomal regions of tumor sample.
ref_counts	(required only when type = "cellline") A binned coverage matrix (m2 bin by n2 cell) with values being read counts for all chromosomal regions of normal sample. n2 can be 1 for bulk sample.
cov_adj	An integer for coverage adjustment for tumor cells.
ref_gtv	A reference "genotype_values" (from scDNA-seq) to help with rho_i estimation.

Value

(rho_hat, theta_hat) of each cell for all region in the "genotype_values". Every 2 columns in the genotype_table are (rho_hat, theta_hat) of each region. Each row is a cell.

Lineage_plot *Generate genotype plot (scatter plot) for each region and save in the plot directory.*

Description

Generate genotype plot (scatter plot) for each region and save in the plot directory.

Usage

```
Lineage_plot(
  Obj_filtered = NULL,
  nSNP = 2000,
  clust_method = "ward.D2",
  nclust = 5,
  plot_path = NULL
)
```

Arguments

Obj_filtered	A Alleloscope object with a n cell by (m region * 2) genotype_values matrix and seg_table_filtered matrix. Every 2 columns in the genotype_values matrix are (rho_hat, theta_hat) of each region.
nSNP	An integer for the minimum number of SNPs across segments. Segments with the number of SNPs ¡ nSNP are excluded.
clust_method	Method for clustering. Please refer to the "pheatmap" function.
nclust	An integer for the number of subclones gapped in the plot.
plot_path	The path for saving the plot.

Value

A lineage tree plot constructed using cell-level genotypes across all regions.

Matrix_filter	*Filter object based on cell number for each SNP, SNP number for each cell, SNP variant allele frequency, and exclude the centromere and telomere regions.*

Description

Filter object based on cell number for each SNP, SNP number for each cell, SNP variant allele frequency, and exclude the centromere and telomere regions.

Usage

```
Matrix_filter(
  Obj = NULL,
  cell_filter = 5,
  SNP_filter = 10,
  min_vaf = 0,
  max_vaf = 1,
  centro = NULL,
  telo = NULL,
  plot_stat = TRUE,
  plot_vaf = TRUE
)
```

Arguments

Obj	A Alleloscope object.
cell_filter	An integer of minimum cell number for SNP selection.
SNP_filter	An integer of minimum SNP number for cell selection.
min_vaf	A numerical value in the range (0,1) of minimum SNP variant allele frequency in the pseudo bulk for SNP selection.
max_vaf	A numerical value in the range (0,1) of mzsimum SNP variant allele frequency in the pseudo bulk for SNP selection.
centro	A Matrix/ data.frame of centromere information.
telo	A Matrix/ data.frame of telomere information.
plot_vaf	Logical (TRUE/FALSE). Whether or not to plot the variant allele frequency for the pseudo bulk for all the chromosomes.

Value

A Alleloscope object after the filtering.

Rundf_dna *Run all steps for scDNA-seq data*

Description

Run all steps for scDNA-seq data

Usage

```
Rundf_dna(
  alt_all = NULL,
  ref_all = NULL,
  var_all = NULL,
  samplename = "sample",
  genome_assembly = "GRCh38",
  dir_path = "./",
  barcodes = NULL,
  size = NULL,
  assay = "scDNAseq",
  raw_counts = NULL,
  ref_counts = NULL,
  type = "tumor",
  cell_filter = 5,
  SNP_filter = 10,
  min_vaf = 0,
  max_vaf = 1
)
```

Arguments

alt_all	A SNP by cell read count matrix/ spare matrix for the alternative alleles.
ref_all	A SNP by cell read count matrix/ spare matrix for the reference alleles.
samplename	Sample name for the data.
genome_assembly	The genome assembly used for sequencing alignment. (ex: "GRCh38" or "GRCh37")
dir_path	Path of the output directory.
barcodes	A matrix/ data.frame with barcodes for each cell in the first column.
size	A numeric vector for the size (bp) of different chromosomes (with the names indicating which chromosome from 1 to 22)
assay	A character indicating the type of sequencing data. (ex: "scDNAseq" or "scATACseq")
raw_counts	A large binned coverage matrix (m1 bin by n1 cell) for all chromosomal regions of tumor sample.
ref_counts	A large binned coverage matrix (m2 bin by n2 cell) for all chromosomal regions of normal sample.
type	Specify whethere the sample is a "tumor" or "cellline". If "type" is a "cellline", param "ref_counts" needs to be specified for normal sample.

`cell_filter`	An integer of minimum cell number for SNP selection.
`SNP_filter`	An integer of minimum SNP number for cell selection.
`min_vaf`	A numerical value in the range (0,1) of minimum SNP variant allele frequency in the pseudo bulk for SNP selection.
`max_vaf`	A numerical value in the range (0,1) of mzsimum SNP variant allele frequency in the pseudo bulk for SNP selection.
`vcf_all`	A matrix/ data.frame of the vcf format for SNP information. (The length and order are the same as nrow(alt_all) and nrow(alt_all))

Value

A Alleloscope object including the necessary information.

`Segmentation` *HMM segmentation based on coverage matrix for paired tumor and normal sample.*

Description

If there is no paired normal, other normal sample with the same genome coordinate also works.

Usage

```
Segmentation(
  Obj_filtered = NULL,
  raw_counts = NULL,
  ref_counts = NULL,
  plot_seg = TRUE
)
```

Arguments

`Obj_filtered`	A Alleloscope object.
`raw_counts`	A binned coverage matrix (m1 bin by n1 cell) with values being read counts in DNA sequencing data for all chromosomal regions of tumor sample. n1 can be 1 for bulk sample.
`ref_counts`	A binned coverage matrix (m2 bin by n2 cell) with values being read counts in DNA sequencing data for all chromosomal regions of normal sample. n2 can be 1 for bulk sample. Numbers of bins (rows) should be the same in the paired chromosomal regions for the paired samples
`plot_seg`	Logical (TRUE/ FALSE). Whether or not to plot the segmentation result.

Value

A Alleloscope object with "seg_table" added.

`Segments_filter` *Select the segments in the "seg_table" with more than nSNP*

Description

Select the segments in the "seg_table" with more than nSNP

Usage

```
Segments_filter(Obj_filtered = NULL, nSNP = 2000)
```

Arguments

`Obj_filtered`	A Alleloscope object with SNP info and raw segmentation table "seg_table".
`nSNP`	An integer of minimum number of SNPs for region selecgtion.

Value

A Alleloscope object with "seg_table_filtered" added.

`Select_normal` *Identify candidate normal cells and normal regions for cell coverage normalization*

Description

Identify candidate normal cells and normal regions for cell coverage normalization

Usage

```
Select_normal(Obj_filtered = NULL, raw_counts = NULL, cell_nclust = 5)
```

Arguments

`Obj_filtered`	A Alleloscope object with major haplotype proportion (theta_hat) for each cell of each region in the "rds_list".
`raw_counts`	A large binned coverage matrix (bin by cell) with values being read counts for all chromosomal regions of tumor sample.

Value

A Alleloscope object with a "select_normal" list added. A "select_normal" list includes "barcode_normal": Barcodes of the identified normal cells in the tumor sample. "region_normal": A vector of ordered potential normal regions for selection. (1st is the most possible.) "region_normal_rank": A table with the potential "normal regions" for the k clusters from hierarchical clustering. "k_normal": An integer indicates the kth clsuter that is idenfied as "normal cells"