

Package ‘Alleloscope’

October 20, 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@penndmedicine.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	2
Createobj	2
EM	3
Est_regions	4
Genotype	5
Genotype_value	5
Lineage_plot	6
Matrix_filter	7
Rundf_dna	8
Segmentation	9
Segments_filter	10
Select_normal	10

AssignClones_ref	<i>Using marker regions to assign each cell into c reference sub-clones</i>
------------------	---

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	<i>Generate Alleloscope object for analysis</i>
-----------	---

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

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

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

<code>ref_table</code>	A SNP by cell read count matrix/ spare matrix for the reference alleles.
<code>alt_table</code>	A SNP by cell read count matrix/ spare matrix for the alternative alleles.
<code>max_iter</code>	An integer of maximum iteration number.
<code>seed</code>	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 (θ_{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. θ_{hat} represents the CNV states for each cell. A cell is considered as a CNV carrier if its θ_{hat} depart from 0.5.

Est_regions	<i>Perform EM iterations on the filtered cells with barcodes, and plot the results for each region.</i>
-------------	---

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 (θ_{hat}) for each region.
max_nSNP	Integer. Maximum SNP number used for estimating θ_{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 \leq 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 (θ_{hat}) for all regions.

Genotype	<i>Genotype each cell for each region and plot the genotypes.</i>
----------	---

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	<i>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.</i>
----------------	---

Description

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

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

<code>Lineage_plot</code>	<i>Generate genotype plot (scatter plot) for each region and save in the plot directory.</i>
---------------------------	--

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

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

Value

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

Matrix_filter	<i>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.</i>
---------------	--

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 maximum 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

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

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

Value

A Alleloscope object including the necessary information.

Segmentation	<i>HMM segmentation based on coverage matrix for paired tumor and normal sample.</i>
--------------	--

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

<code>Obj_filtered</code>	A Alleloscope object.
<code>raw_counts</code>	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.
<code>ref_counts</code>	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
<code>plot_seg</code>	Logical (TRUE/ FALSE). Whether or not to plot the segmentation result.

Value

A Alleloscope object with "seg_table" added.

Segments_filter	<i>Select the segments in the "seg_table" with more than nSNP</i>
-----------------	---

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	<i>Identify candidate normal cells and normal regions for cell coverage normalization</i>
---------------	---

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 cluster that is identified as "normal cells"