MAY 12, 2020

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.bfhbj5n

External link:
<https://www.biorxiv.org/content/10.1101/2020.04.12.038281v1.full>

Protocol Citation: Fang Wang, Qihan Wang, Vakul Mohanty, Shaoheng Liang, Jinzhuang Dou, Jincheng Han, Darlan Conterno Minussi, Ruli Gao, Li Ding, Nicholas Navin, Ken Chen 2020. Minimal Event Distance Aneuploidy Lineage Tree (MEDALT) inference based on single cell copy number profile.
protocols.io
<https://dx.doi.org/10.17504/protocols.io.bfhbj5n>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

Created: Apr 23, 2020

Minimal Event Distance Aneuploidy Lineage Tree (MEDALT) inference based on single cell copy number profile

In 1 collection

Fang Wang¹, Qihan Wang², Vakul Mohanty¹, Shaoheng Liang¹, Jinzhuang Dou¹, Jincheng Han¹, Darlan Conterno Minussi¹, Ruli Gao³, Li Ding⁴, Nicholas Navin¹, Ken Chen⁵

¹The University of Texas MD Anderson Cancer center; ²Rice University;

³Houston Methodist Research Institute;

⁴McDonnell Genome Institute Washington University School of Medicine;

⁵The University of Texas MD Anderson Cancer Center

Fang Wang

ABSTRACT

This protocol describes two innovative algorithms:

- 1) A minimal event distance aneuploidy lineage tree (MEDALT) inference algorithm allows implementing genetically meaningful distances and is scalable to current single-cell datasets containing thousands of cells, and
- 2) A statistical routine, Lineage Speciation Analysis (LSA), enables prioritization of CNAs and genes that are non-randomly associated with the observed lineage expansion and thereby are potentially functionally important.

Last Modified: May 12,
2020

PROTOCOL integer ID:
36111

Keywords: single cell
technology, tumor evolution,
copy number alteration

- 1 Install Python 2.7 and R 3.5
Download MEDALT tool from <https://github.com/KChen-lab/MEDALT.git>

Software

MEDALT

NAME

Fang Wang and Qihan Wang

DEVELOPER

Extract input dataset

Dataset

Single cell copy number profile generated by single cell DNA seq

NAME

<https://github.com/KChen-lab/MEDALT/blob/master/example/scDNA.CNV.txt>

LINK

Dataset

Single cell copy number profile inferred from single cell RNA se

NAME

<https://github.com/KChen-lab/MEDALT/blob/master/example/scRNA.CNV.txt>

LINK

- 2 Decompress gzipped files (MEDALT-1.0.tar.gz)

Command

```
tar -zxvf MEDALT-1.0.tar.gz
cd MEDALT-1.0
```

```
#help document
python scTree.py -h
```

Usage: python scTree.py <-P path> <-I input> <-D datatype>

Input integer copy number profile. Columns correspond to chromosomal position.
Rows correspond to cells.

Options:

- version   show program's version number and exit
- h, --help   Show this help message and exit.
- P PATH, --Path=PATH Path to script
- I INPUT, --Input=INPUT
                      Input file
- G GENOME, --Genome=GENOME
                      Genome version hg19 or hg38
- O OUTPUT, --Output=OUTPUT
                      Output path
- D DATATYPE, --Datatype=DATATYPE
                      The type of input data. Either D (DNA-seq)
                      or R (RNA-seq).
- W WINDOWS, --Windows=WINDOWS
                      the number of genes you want to merge when
                      you input copy number profile inferred from
                      scRNA-seq. Default 30.
- R PERMUTATION, --Permutation=PERMUTATION
                      Whether reconstructed permuted tree (T) or
                      not (F). If not, permuted copy number
                      profile will be used to perform LSA. Default
                      value is F due to time cost.

3 Run the example data generated based on single cell DNA sequencing technology

Command

```
python scTree.py -P ./ -I ./example/scDNA.CNV.txt -D D -G hg19 -O ./example/outputDNA
```

Transfer data to segmental level

Inferring MEDALT.

MEDALT inference finish.

Performing LSA.

Loading required package: BiocGenerics

Loading required package: parallel

Attaching package: 'BiocGenerics'

The following objects are masked from 'package:parallel':

clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
clusterExport, clusterMap, parApply, parCapply, parLapply,
parLapplyLB, parRapply, parSapply, parSapplyLB

The following objects are masked from 'package:stats':

IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

anyDuplicated, append, as.data.frame, basename, cbind, colMeans,
colnames, colSums, dirname, do.call, duplicated, eval, evalq,
Filter, Find, get, grep, grepl, intersect, is.unsorted, lapply,
lengths, Map, mapply, match, mget, order, paste, pmax, pmax.int,
pmin, pmin.int, Position, rank, rbind, Reduce, rowMeans, rownames,
rowSums, sapply, setdiff, sort, table, tapply, union, unique,
unsplit, which, which.max, which.min

Loading required package: S4Vectors

Loading required package: stats4

Attaching package: 'S4Vectors'

The following object is masked from 'package:base':

expand.grid

Loading required package: IRanges

Loading required package: GenomicRanges

Loading required package: GenomInfoDb

Loading required package: Biostrings

Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:base':

strsplit

Loading required package: BSgenome

Loading required package: rtracklayer

Loading required package: GenomicFeatures

Loading required package: AnnotationDbi

Loading required package: Biobase

Welcome to Bioconductor

Vignettes contain introductory material; view with
'browseVignettes()'. To cite Bioconductor, see
'citation("Biobase")', and for packages 'citation("pkgname")'.

Loading required package: VariantAnnotation

Loading required package: SummarizedExperiment

Loading required package: DelayedArray

Loading required package: matrixStats

Attaching package: 'matrixStats'

The following objects are masked from 'package:Biobase':

anyMissing, rowMedians

Loading required package: BiocParallel

Attaching package: 'DelayedArray'

The following objects are masked from 'package:matrixStats':

colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

The following object is masked from 'package:Biostrings':

type

The following objects are masked from 'package:base':

aperm, apply

Loading required package: Rsamtools

Attaching package: 'VariantAnnotation'

The following object is masked from 'package:base':

tabulate

Loading required package: GenomicAlignments

There were 20 warnings (use warnings() to see them)

Attaching package: 'igraph'

The following objects are masked from 'package:DelayedArray':

path, simplify

The following objects are masked from 'package:rtracklayer':

blocks, path

The following object is masked from 'package:Biostrings':

union

The following object is masked from 'package:GenomicRanges':

union

The following object is masked from 'package:IRanges':

union

The following object is masked from 'package:S4Vectors':

union

The following objects are masked from 'package:BiocGenerics':

normalize, path, union

The following objects are masked from 'package:stats':

decompose, spectrum

The following object is masked from 'package:base':

union

Warning message:

package 'igraph' was built under R version 3.5.2

Attaching package: 'DescTools'

The following object is masked from 'package:igraph':

%c%

Warning message:

package 'DescTools' was built under R version 3.5.2

[1] Visualization MEDALT!

null device

1

[1] LSA segmentation!

[1] Calculating CFL

[1] Calculating permutation CFL

[1] Estimate emperical p value

[1] Estimate parallel evolution

null device

1

Done!

Note

R packages (igraph, HelloRanges and DescTools) are loaded.

Expected result

Three text files are expected:

(1) CNV.tree.txt which is an rooted directed tree including three columns: parent node, child node and distance.

☐ CNV.tree.txt

(2) segmental.LSA.txt which includes broad CNAs significantly associated with lineage expansion.

☐ segmental.LSA.txt

(3) gene.LSA.txt which includes focal (gene) CNAs significantly associated with lineage expansion.

☐ gene.LSA.txt

Two figures are also expected:

(1) singlecell.tree.pdf which is a visualization of MEDALT by igraph. You also can input CNV.tree.txt into Cytoscape to generate preferred visualization.

Each node represents a cell, each edge represents a kinship between two cells, arrows point towards younger cells, and the root represents a normal diploid cell.

(2) LSA.tree.pdf which is a visualization of identified CNAs by igraph.

Note

We run the example data only through permuting copy number profile instead of reconstructing tree based on permuted copy number profile. The setting can be changed via -R T.

4 Run the example data inferred using inferCNV based on single cell RNA sequencing technology

☐ scRNA.CNV.txt

Command

```
python scTree.py -P ./ -I ./example/scRNA.CNV.txt -D R -G hg19 -O ./example/outputRNA
```

Transfer data to segmental level

The number of genes which are merger into the bin is default value 30. If you want change it

please specify the value through -W

Inferring MEDALT.

MEDALT inference finish.

Performing LSA.

Loading required package: BiocGenerics

Loading required package: parallel

Attaching package: 'BiocGenerics'

The following objects are masked from 'package:parallel':

clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
clusterExport, clusterMap, parApply, parCapply, parLapply,
parLapplyLB, parRapply, parSapply, parSapplyLB

The following objects are masked from 'package:stats':

IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

anyDuplicated, append, as.data.frame, basename, cbind, colMeans,
colnames, colSums, dirname, do.call, duplicated, eval, evalq,
Filter, Find, get, grep, grepl, intersect, is.unsorted, lapply,
lengths, Map, mapply, match, mget, order, paste, pmax, pmax.int,
pmin, pmin.int, Position, rank, rbind, Reduce, rowMeans, rownames,
rowSums, sapply, setdiff, sort, table, tapply, union, unique,
unsplit, which, which.max, which.min

Loading required package: S4Vectors

Loading required package: stats4

Attaching package: 'S4Vectors'

The following object is masked from 'package:base':

expand.grid

Loading required package: IRanges

Loading required package: GenomicRanges

Loading required package: GenomInfoDb

Loading required package: Biostrings

Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:base':

strsplit

Loading required package: BSgenome

Loading required package: rtracklayer

Loading required package: GenomicFeatures

Loading required package: AnnotationDbi

Loading required package: Biobase

Welcome to Bioconductor

Vignettes contain introductory material; view with
'browseVignettes()'. To cite Bioconductor, see
'citation("Biobase")', and for packages 'citation("pkgname")'.

Loading required package: VariantAnnotation

Loading required package: SummarizedExperiment

Loading required package: DelayedArray

Loading required package: matrixStats

Attaching package: 'matrixStats'

The following objects are masked from 'package:Biobase':

anyMissing, rowMedians

Loading required package: BiocParallel

Attaching package: 'DelayedArray'

The following objects are masked from 'package:matrixStats':

colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

The following object is masked from 'package:Biostrings':

type

The following objects are masked from 'package:base':

aperm, apply

Loading required package: Rsamtools

Attaching package: 'VariantAnnotation'

The following object is masked from 'package:base':

tabulate

Loading required package: GenomicAlignments

There were 20 warnings (use warnings() to see them)

Attaching package: 'igraph'

The following objects are masked from 'package:DelayedArray':

path, simplify

The following objects are masked from 'package:rtracklayer':

blocks, path

The following object is masked from 'package:Biostrings':

union

The following object is masked from 'package:GenomicRanges':

union

The following object is masked from 'package:IRanges':

union

The following object is masked from 'package:S4Vectors':

union

The following objects are masked from 'package:BiocGenerics':

normalize, path, union

The following objects are masked from 'package:stats':

decompose, spectrum

The following object is masked from 'package:base':

union

Warning message:

package 'igraph' was built under R version 3.5.2

Attaching package: 'DescTools'

The following object is masked from 'package:igraph':

%c%

Warning message:

package 'DescTools' was built under R version 3.5.2

[1] Visualization MEDALT!

null device

1

[1] LSA segmentation!

[1] Calculating CFL

[1] Calculating permutation CFL

[1] Estimate empirical p value

[1] Estimate parallel evolution

null device

1

Done!

Expected result

Three text files are expected:

(1) CNV.tree.txt which is an rooted directed tree including three columns: parent node, child node and distance.

☐ CNV.tree.txt

(2) segmental.LSA.txt which includes broad CNAs significantly associated with lineage expansion.

☐ segmental.LSA.txt

(3) gene.LSA.txt which includes focal (gene) CNAs significantly associated with lineage expansion.

gene.LSA.txt

Two figures are also expected:

(1) singlecell.tree.pdf which is a visualization of MEDALT by igraph.

(2) LSA.tree.pdf which is a visualization of identified CNAs by igraph.