Package 'SATS'

July 15, 2024

Title SATS (Signature Analyzer for Targeted Sequencing)

Version 0.0.8

Date 2024-07-15

Description This package is created to perform mutational signature analysis for targeted sequenced tumors. Unlike the canonical analysis of mutational signatures, SATS factorizes the mutation counts matrix into a panel context matrix (measuring the size of the targeted sequenced genome for each tumor in the unit of million base pairs (Mb)), a signature profile matrix, and a signature activity matrix. SATS also calculates the expected number of mutations attributed by a signature, namely signature expectancy, for each targeted sequenced tumor.

Imports stats, glmnet, GenomicRanges, IRanges, Biostrings, dplyr, BSgenome.Hsapiens.UCSC.hg19

Depends R (>= 4.1.0)

License GPL-2

NeedsCompilation yes

Author DongHyuk Lee [aut],
Bin Zhu [aut],
Bill Wheeler [cre]

Maintainer Bill Wheeler <wheelerb@imsweb.com>

Contents

SATS-package *SATS (Signature Analyzer for Targeted Sequencing)*

Description

This package is created to perform mutational signature analysis for targeted sequenced tumors. Unlike the canonical analysis of mutational signatures, SATS factorizes the mutation counts matrix into a panel context matrix (measuring the size of the targeted sequenced genome for each tumor in the unit of million base pairs (Mb)), a signature profile matrix, and a signature activity matrix. SATS also calculates the expected number of mutations attributed by a signature, namely signature expectancy, for each targeted sequenced tumor.

Details

This package includes a novel algorithm, SATS, to perform mutational signature analysis for targeted sequenced tumors. The algorithm first applies the signeR algorithm to extract profiles of de novo mutational signatures by appropriately adjusting for various panel sizes. Next, the profiles of identified de novo mutational signatures are mapped to the profiles of catalog signatures of tumor mutation burden (TMB), in the unit of the number of mutations per million base pairs, using penalized non-negative least squares. Then, given the panel sizes and profiles of mapped TMB catalog signatures, signature activities are estimated for all samples simultaneously through the Expectation-Maximization (EM) algorithm. Finally, the expected number of mutations attributed by a signature, namely signature expectancy, is calculated for each targeted sequenced tumor.

The main functions in this package are `EstimateSigActivity`, `CalculateSignatureBurdens`, and `MappingSignature`.

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

References

Lee, D., Hua, M., Wang, D., Song, L., Yu, K., Yang, X., Shi, J., Landi, M., Zhu, B. The mutational signatures of 100,477 targeted sequenced tumors. Submitted.

CalculateSignatureBurdens *CalculateSignatureBurdens*

Description

Estimation of the expected number of mutations attributed by TMB-based catalog signatures (signature expectancy) given the panel size matrix, the catalog signature profile matrix and the signature activities matrix.

Usage

```
CalculateSignatureBurdens(L, W, H)
```

Arguments

L	Panel size matrix or data frame with samples in columns
W	Catalog signature profiles matrix or data frame with signatures in columns
H	Activity matrix or data frame with samples in columns

Details

The panel size matrix L is of size P (the mutation context) by N (the sample size). The catalog signature profile matrix has dimension of P by K (the number of signatures) and the activity matrix H is of size K by N. For single base substitutions (SBS), P is 96. If K is the number of signatures and N is the number of samples, then H must be of dimension K X N, `ncol(L)` = N, and `ncol(W)` = K.

Value

A matrix of dimension K X N, where K is the number of signatures and N is the number of samples.

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

See Also

EstimateSigActivity

Examples

```
data(SimData, package="SATS")

CalculateSignatureBurdens(SimData$L, SimData$TrueW_TMB, SimData$TrueH)
```

EstimateSigActivity *EstimateSigActivity*

Description

Estimation of signature activities given the original mutation type matrix, the panel size matrix, and the catalog signature profile matrix.

Usage

```
EstimateSigActivity(V, L, W, n.start=50, iter.max=5000, eps=1e-5)
```

Arguments

V	Mutation type matrix or data frame with samples in columns
L	Panel size matrix or data frame with samples in columns
W	Catalog signature profiles matrix or data frame with signatures in columns
n.start	Number of initializations. The default is 50.
iter.max	Maximum number iterations in the EM algorithm. The default is 5000.
eps	Stopping tolerance in the EM algorithm. The default is 1e-5.

Details

The panel size matrix L and mutation type matrix V are of size P (the mutation context) by N (the sample size). The catalog signature profile matrix has dimension of P by K (the number of signatures). For single base substitutions (SBS), P is 96. For the objects V, L, and W, we must have dim(V) = dim(L) and ncol(W) = K, where K is the number of signatures. EstimateSigActivity() uses the EM algorithm is used to estimate signature n.start, iter.max and eps control EM part. Because the convergence to a local saddle point can be an issue of the EM algorithm, it would be good practice to try multiple initial values (n.start, the default is 50). For each initial value, the default value of the maximal iteration of the EM algorithm (iter.max) is 5000, and the stopping tolerance (eps) is set to 1e-5.

Value

A list containing the estimated activity matrix `H`, the log-likelihood `loglike`, and the logical value `converged`.

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

See Also

`CalculateSignatureBurdens`

Examples

```
    data(SimData, package="SATS")

  ## Not run:
    EstimateSigActivity(SimData$V, SimData$L, SimData$TrueW_TMB)

## End(Not run)
```

`GeneratePanelSize` *Generate Panel Size Matrix*

Description

Generation of the panel size matrix given the panel information.

Usage

```
GeneratePanelSize(genomic_information, Types)
```

Arguments

`genomic_information`	Data frame of panel information (see details).
`Types`	Mutation type order either one of "COSMIC" or "signeR" (see details).

Details

The first argument 'genomic_information' should contain columns 'Chromosome', 'Start_Position', 'End_Position', 'SEQ_ASSAY_ID'. The column 'Chromosome' contains chromosome number where 'Start_Position' and 'End_Position' columns are start and end positions of the targeted panel. The last column 'SEQ_ASSAY_ID' distinguishes different panels consisting of the result. Please note that the column names of 'genomic_information' identical to 'Chromosome', 'Start_Position', 'End_Position', 'SEQ_ASSAY_ID'. The second argument specifies mutation type order as either one of "COSMIC" or "signeR" where "COSMIC" corresponds to the order from the COSMIC database v3.2 and "signeR" corresponds to the order from the signeR package. **Note**: The result of 'GeneratePanelSize()' may NOT be an 'L' matrix. The 'L' matrix can be constructed by attaching the columns of the function output that correspond to the columns of the 'V' matrix. The resulting augmented matrix can be used as the opportunity matrix for 'signeR()' function, 'L' matrix for 'EstimateSigActivity()' and 'CalculateSignatureBurdens()' functions. Therefore, it is important the mutation type order (row names) should be the same as input matrix (Mutation type matrix 'V'). We highly recommend to confirm that both 'V' and 'L' matrices have the same mutation type order corresponding to one of COSMIC database v3.2 or signeR package (both have the same order but have different expression) to conduct the consistent analysis.

Value

A data frame of 96 by 'S' (the number of panels, 'SEQ_ASSAY_ID') where entries denote the number of trinucleotides per million base pairs.

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

Examples

```
data(SimData, package="SATS")

GeneratePanelSize(genomic_information = SimData$PanelEx, Types = "COSMIC")
GeneratePanelSize(genomic_information = SimData$PanelEx, Types = "signeR")
```

MappingSignature *Find a subset of TMB-based catalog SBS signatures*

Description

This function finds a subset of TMB-based catalog SBS signatures whose linear combination approximate de novo SBS signatures detected by signeR.

Usage

```
MappingSignature(W_hat, W_ref=NULL, niter=100, cutoff.I2=0.1, min.repeats=80)
```

Arguments

W_hat	Matrix or data frame of de novo signatures from signeR
W_ref	NULL or a matrix or data frame of TMB-based catalog signatures. If NULL, then it will default to SimData$W_TMB (see SimData).
niter	Number of iterations. The default is 100.
cutoff.I2	Cutoff value to select signatures. The default is 0.1.
min.repeats	Minimum number of iterations to select signatures with I^2 > cutoff.I2 . The default is 80.

Details

MappingSignature() applies penalized non-negative least squares (pNNLS) for selecting the TMB-based catalog signatures. Specifically, it repeats pNNLS 100 times (niter) to reduce the randomness of cross-validation involved in pNNLS. Then TMB-based catalog signatures are selected with a coefficient greater than 0.1 (cutoff.I2) in more than 80 repeats (min.repeats).

Value

A data frame with column names of W_ref (it returns COSMIC SBS names if COSMIC catlog based reference signatures are used) and freq (the number of repetitions greater than cutoff coefficient values out of niter iterations).

Author(s)

Donghyuk Lee <dhyuklee@pusan.ac.kr> and Bin Zhu <bin.zhu@nih.gov>

RefTMB *Example Data*

Description

The pan-cancer repertoire of reference signatures and the reference TMB (Tumor mutation burden) signature profiles of SBS (Single base substitutions) and DBS (Double base substitutions).

Details

This file consists of the list RefTMB with the following objects:

- SBS_W : The reference SBS TMB signature profiles of size 96 by 76.
- DBS_W : The reference DBS TMB signature profiles of size 78 by 11.
- SBS_refSigs : The list of cancer specific SBS signature names.
- DBS_refSigs : The list of cancer specific DBS signature names.

SimData *Data for examples*

Description

Simulated data as an example

Details

This file consists of the list SimData with the following objects:

- V : Simulated mutation catalog matrix of size 96 by 10027 generated from the Poisson distribution with the mean corresponding to each element of L*(WH), where * is elementwise multiplication, L is the panel size matrix below, W is the profile matrix of the tumor mutation burden (TrueW_TMB below), and H is the signature activity matrix (TrueH below).
- L : Panel size matrix of size 96 by 10027.
- TrueH : Simulated signature activity matrix of size 6 by 10027 used to generate V.
- TrueW_TMB : Tumor mutation burden based signatures size 96 by 6 for 6 SBS signatures (SBS1, SBS2/13, SBS4, SBS5, SBS40, SBS89) used to generate V.
- SingleTumorEx : A single simulated tumor (column singleV) and its sequencing context (column singleL).
- PanelEx: An example data frame of panel information

Index