Package 'hsegHMM'

October 3, 2017

Title Hidden Markov Model-based Allele-specific Copy Number Alteration Analysis Accounting for Hypersegmentation

Version 0.0.4

Date 2017-10-02

Author Hyoyoung Choo-Wosoba, Paul S. Albert, and Bin Zhu

Maintainer Bill Wheeler <wheelerb@imsweb.com>

Description An R package for the hsegHMM model of allele-specific SCNA analysis.

Suggests facets

License GPL-2

NeedsCompilation no

R topics documented:

`hsegHMM` *Hidden Markov Model-based Allele-specific Copy Number Alteration Analysis Accounting for Hypersegmentation*

Description

An R package for the hsegHMM model of allele-specific SCNA analysis

Details

Somatic copy number alternation (SCNA) is a common feature of the cancer genome and is associated with cancer etiology and prognosis. The allele-specific SCNA analysis of a tumor sample aims to identify the allele-specific copy numbers of both alleles, adjusting for the ploidy and the tumor purity. Next generation sequencing platforms produce abundant read counts at the base-pair resolution across the exome or whole genome which is susceptible to hypersegmentation, a phenomenon where numerous regions with very short length are falsely identified as SCNA. This

package employs a robust hidden Markov model approach that accounts for hypersegmentation for allele-specific SCNA analysis, and an efficient E-M algorithm procedure that uses a forward-backward algorithm for evaluating the E-step. The main functions that perform this method are hsegHMM_N for normally distributed log(ratio) values, and hsegHMM_T for t-distributed log(ratio) values.

Author(s)

Hyoyoung Choo-Wosoba, Paul S. Albert, and Bin Zhu <bin.zhu@nih.gov>

References

Choo-Wosoba, H., Albert, P.S., Zhu, B. hsegHMM: Hidden Markov Model-based Allele-specific Copy Number Alteration Analysis Accounting for Hypersegmentation

---

`hsegHMM_N` *The hsegHMM procedure for normally distributed log(Ratio) values*

Description

A hidden Markov model approach that accounts for hypersegmentation for allele-specific SCNA analysis for normally distributed log(Ratio) values

Usage

```
hsegHMM_N(logR, logOR, purity=0.8, ploidy=1.5, logR.var=0.5, logOR.var=0.5,
                    genoStates=c("", "A", "AA", "AB", "AAB", "AAA", "AAAB",
                                 "AABB", "AAAA", "AAAAB", "AAABB", "AAAAA"),
                prob0=NULL, transProb=NULL, maxiter=100, stopTol=0.01, minLogOR2=1e-6,
                    optim.control=list(trace=0))
```

Arguments

`logR`	Vector of (non-missing) log(Ratio) values. No default
`logOR`	Vector of log(Odds Ratio) values. Note that this vector can have missing values and will be squared inside the function. No default.
`purity`	Initial value for the tumor purity. The default is 0.8.
`ploidy`	Initial value for the ploidy. The default is 1.5.
`logR.var`	Initial value for the variance component of logR. The default is 0.5.
`logOR.var`	Initial value for the variance component of logOR. The default is 0.5.
`genoStates`	Character vector of hidden genotype states. The default is c("", "A", "AA", "AB", "AAB", "AAA", "AAAB", "AABB", "AAAA", "AAAAB", "AAABB", "AAAAA").
`prob0`	NULL or a vector of initial probabilities for `genoStates`. If NULL, then it will be set to `rep(1/length(genoStates), length(genoStates))`. The default is NULL.
`transProb`	NULL or a matrix of transition probabilities. If NULL, then it will be set to `matrix(c(rep(c(1-(J-1)/5000,rep(1/5000,J)),(J-1)),1-(J-1)/5000),J,J)`, where `J = length(genoStates)`. The default is NULL.

maxiter	Maximum number of iterations for the algorithm. The default is 100.
stopTol	Stopping tolerance for the algorithm. The algorithm will stop when two successive log-likelihood values differ by less than stopTol. The default is 0.01.
minLogOR2	Minimum value for logOR^2 to prevent numerical difficulties in the algorithm. The default is 1e-6.
optim.control	List for the control option in the optim function. The default is list(trace=0).

Details

Missing values are allowed for logOR values as long as logR is observed. The optimization is performed using the L-BFGS-B method in the optim function. See the reference for details of the algorithm.

Value

A list with the following names and descriptions:

- converged Convergence status
- loglike Final value of the log-likelihood
- allele1 Allele 1
- allele2 Allele 2
- alleleFreq1 Frequency of allele 1
- alleleFreq2 Frequency of allele 2
- copyNumber Copy number
- post.prob Matrix of posterior probabilities for each genotype state
- which.max.post.prob Indices for the genotype status which gives the maximum posterior probability.
- logR_hat The expected value of logR
- logOR_hat The expected value of logOR
- purity_hat The expected value of purity
- ploidy_hat The expected value of ploidy
- logR.var_hat The expected value of logR.var
- logOR.var_hat The expected value of logOR.var
- genoStates The genotype states
- prob0 The initial probability of the genotype states
- transProb The matrix of transition probabilities
- AIC Akaike information criterion
- BIC Bayesian information criterion
- covariance Covariance matrix for all parameters

Author(s)

Hyoyoung Choo-Wosoba, Paul S. Albert, and Bin Zhu <bin.zhu@nih.gov>

References

Choo-Wosoba, H., Albert, P.S., Zhu, B. hsegHMM: Hidden Markov Model-based Allele-specific Copy Number Alteration Analysis Accounting for Hypersegmentation

See Also

hsegHMM_T

Examples

```
data(hseg_data, package="hsegHMM")

hsegHMM_N(lr, logor)
```

---

hsegHMM_T *The hsegHMM procedure for t-distributed log(Ratio) values*

Description

A hidden Markov model approach that accounts for hypersegmentation for allele-specific SCNA analysis for t-distributed log(Ratio) values

Usage

```
hsegHMM_T(logR, logOR, purity=0.8, ploidy=1.5, logR.var=0.5, logOR.var=0.5, df=3,
                genoStates=c("", "A", "AA", "AB", "AAB", "AAA", "AAAB",
                             "AABB", "AAAA", "AAAAB", "AAABB", "AAAAA"),
            prob0=NULL, transProb=NULL, maxiter=100, stopTol=0.01, minLogOR2=1e-6,
                df.min=0.0001, df.max=100, optim.control=list(trace=0))
```

Arguments

logR	Vector of (non-missing) log(Ratio) values. No default
logOR	Vector of log(Odds Ratio) values. Note that this vector can have missing values and will be squared inside the function. No default.
purity	Initial value for the tumor purity. The default is 0.8.
ploidy	Initial value for the ploidy. The default is 1.5.
logR.var	Initial value for the variance component of logR. The default is 0.5.
logOR.var	Initial value for the variance component of logOR. The default is 0.5.
df	Initial value for the degrees of freedom. The default is 3.
genoStates	Character vector of hidden genotype states. The default is c("", "A", "AA", "AB", "AAB", "AAA", "AAAB", "AABB", "AAAA", "AAAAB", "AAABB", "AAAAA").
prob0	NULL or a vector of initial probabilities for genoStates. If NULL, then it will be set to rep(1/length(genoStates), length(genoStates)). The default is NULL.
transProb	NULL or a matrix of transition probabilities. If NULL, then it will be set to matrix(c(rep(c(1-(J-1)/5000,rep(1/5000,J)),(J-1)),1-(J-1)/5000),J,J), where J = length(genoStates). The default is NULL.
maxiter	Maximum number of iterations for the algorithm. The default is 100.
stopTol	Stopping tolerance for the algorithm. The algorithm will stop when two successive log-likelihood values differ by less than stopTol. The default is 0.01.

minLogOR2	Minimum value for logOR^2 to prevent numerical difficulties in the algorithm. The default is 1e-6.
df.min	Lower bound for the df parameter in the optimization. The default is 0.0001.
df.max	Upper bound for the df parameter in the optimization. The default is 100.
optim.control	List for the control option in the optim function. The default is list(trace=0).

Details

Missing values are allowed for logOR values as long as logR is observed. The optimization is performed using the L-BFGS-B method in the optim function. To prevent numerical errors in the gamma function, df.min should be a positive value and df.max a finite positive value. For the other parameters, the lower and upper bounds can be infinite. See the reference for details of the algorithm.

Value

A list with the following names and descriptions:

- converged Convergence status
- loglike Final value of the log-likelihood
- allele1 Allele 1
- allele2 Allele 2
- alleleFreq1 Frequency of allele 1
- alleleFreq2 Frequency of allele 2
- copyNumber Copy number
- post.prob Matrix of posterior probabilities for each genotype state
- which.max.post.prob Indices for the genotype status which gives the maximum posterior probability.
- logR_hat The expected value of logR
- logOR_hat The expected value of logOR
- purity_hat The expected value of purity
- ploidy_hat The expected value of ploidy
- logR.var_hat The expected value of logR.var
- logOR.var_hat The expected value of logOR.var
- df_hat The expected value of df
- genoStates The genotype states
- prob0 The initial probability of the genotype states
- transProb The matrix of transition probabilities
- AIC Akaike information criterion
- BIC Bayesian information criterion
- covariance Covariance matrix for all parameters

Author(s)

Hyoyoung Choo-Wosoba, Paul S. Albert, and Bin Zhu <bin.zhu@nih.gov>

References

Choo-Wosoba, H., Albert, P.S., Zhu, B. hsegHMM: Hidden Markov Model-based Allele-specific Copy Number Alteration Analysis Accounting for Hypersegmentation

See Also

hsegHMM_N

Examples

```
data(hseg_data, package="hsegHMM")

hsegHMM_T(lr, logor)
```

---

hseg_data *Data for examples*

Description

Data for examples.

Details

The object contains logR and logOR values for the hsegHMM_T and hsegHMM_N examples.

See Also

hsegHMM_T, hsegHMM_N

Examples

```
data(hseg_data, package="hsegHMM")

# Display some of the data
lr[1:10]
logor[1:10]
```

Index