Package 'MAFsnp'

November 11, 2016

Title MAFsnp: A multi-sample accurate and flexible SNP caller using next- generation sequencing data

Version 1.0

Author Jiyuan Hu

Maintainer Jiyuan Hu <jiyuan_hu@fudan.edu.cn>

Description MAFsnp is an R package for detecting SNPs using next-generation sequencing data from multiple samples.

Depends R (>= 2.14)

License GPL (>= 2)

Encoding UTF-8

LazyData true

URL http://github.com/zhanghfd/MAFsnp

BugReports http://github.com/zhanghfd/MAFsnp/issues

RoxygenNote 5.0.1

R topics documented:

MAFsnp-package *MAFsnp: A multi-sample accurate and flexible SNP caller using next-generation sequencing data*

Description

This package calls SNPs using next-generation sequencing data from multiple samples. MAFsnp is based on a frequentist framework, which provides p-values with or without FDR correction for each nucleotide locus. MAFsnp outputs called SNPs at any given nominal FDR level. This package also provides a function for simulating read count data.

Details

Package:	MAFsnp
Type:	Package
Version:	1.0
Date:	2015-06-04
License:	GPL (>= 2)

Author(s)

Jiyuan Hu

Maintainer: Jiyuan Hu <jiyuan_hu@fudan.edu.cn>

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N =5; SNPtype = 3; M=1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
d = eLRT(d);
d = est.eLRT.distribution(d);
d = pvalues(d);
d = SNP.calling(d,FDR=0.05);
```

eLRT *Estimated likelihood ratio test statistics*

Description

This function calculates the estimated likelihood ratio test statistics.

Usage

```
eLRT(d)
```

Arguments

d	This is an MAFsnp object.

Value

T	The eLRT test statistics for those loci with at least one variant read
position.with.variation	The positions of those loci with at least one variant read

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M = 1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
d = eLRT(d);
```

---

est.eLRT.distribution *Estimation of the finite sample null distribution of eLRT*

Description

This function estimates the finite sample null distribution of eLRT, which is a mixure of zero and a scaled chi-squared distribution.

Usage

```
est.eLRT.distribution(d, cutoff = 15)
```

Arguments

d	This is an MAFsnp object.
cutoff	A cutoff used to reduce the impact of outliers on evaluating the null distribution.

Value

a.k	A list with two elements, the first element 'a.hat' is the proportion of zero part in the null distribution, and the second element 'k.hat' is the expectation of the non-zero part in the null distribution.

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M = 1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
d = eLRT(d);
d = est.eLRT.distribution(d);
```

`gen.count.data` *Generation of simulated read count data*

Description

Generate a MAFsnp object, which is a list containing a read count matrix 'count', a vector of the true minor allele frequencies 'p', a matrix of true genotypes 'genotype', a vector of sequencing/mapping error probabilities 'error', a vector of true SNP statuses 'flag', and a vector of nucleotide locus positions 'position'.

Usage

```
gen.count.data(ErrorProb, nSample, N, SNPtype, M = 1000, r = 100)
```

Arguments

`ErrorProb`	Mean sequencing/mapping error probability, which is usually >= 0.001 and <= 0.01.
`nSample`	The number of samples.
`N`	Mean read coverage across all nucleotide loci.
`SNPtype`	An indicator for various uniform distribution of allele frequency, 1 for U(0.001,0.01), 2 for U(0.01,0.05), and 3 for U(0.05,0.1).
`M`	The number of simulated SNPs, with a default value of 1000.
`r`	The ratio of number of non-SNPs versus SNPs, with a default value of 100 (100M non-SNPs).

Value

`count`	A (r+1)*M rows by 2*nSample matrix for (r+1)*M nucleotide loci of nSample samples. The first nSample columns are for the variant count with the last nSample column corresponding to the coverage count of each individual. Each row corresponds to the read count data for a locus
`p`	The true minor allele frequencies

genotype	A matrix of true genotypes. Columns 1, 2, 3: the numbers of samples with homogeneous reference genotype RR, heterogeneous variant genotype Rr, and homegeneous variant genotype rr, respectively
error	The mean sequencing/mapping error probabilities.
flag	The true statuses of SNPs (1 for a SNP and 0 for a non-SNP)
position	The positions of nucleotide loci

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M = 1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
```

MAFobj *MAFsnp object*

Description

Creation of an MAFsnp object.

Usage

```
MAFobj(d)
```

Arguments

d	This is list containing the read count matrix, positions of nucleotide locus, and true SNP statuses obtained from dbSNP or any other public database.

Value

X	Variant count matrix
N	Coverage count matrix
Xsum	Number of variants (across all samples) for each locus

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M = 1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
```

`pvalues` *P-value calculation*

Description

This is a function for calculating the p-values with or without FDR correction for each nucleotide locus.

Usage

```
pvalues(d)
```

Arguments

`d`	This is an MAFsnp object

Value

`p.value`	P-values for those nucleotide loci with at least one variant read.
`p.fdr`	FDR corrected p-values for those nucleotide locci with at least one variant read.

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M = 1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
d = eLRT(d);
d = est.eLRT.distribution(d);
d = pvalues(d);
```

SNP.calling *Calling SNPs*

Description

This is a function for calling SNPs for a given FDR threshold.

Usage

```
SNP.calling(d, FDR = 0.01)
```

Arguments

d	This is an MAFsnp object.
FDR	The FDR threshold used to call SNPs, with a default value of 0.01.

Value

snps	A matrix of 3 columns for positions, p-values, and FDR adjusted p-values, respectively

Author(s)

Jiyuan Hu

References

Hu J, Li T, Xiu Z, Zhang H. (2015) MAFsnp: A Multi-sample Accurate and Flexible SNP Caller Using Next-generation Sequencing Data. PLoS ONE 10(8): e0135332. doi:10.1371/journal.pone.0135332.

Examples

```
e = 0.01; nSample = 50; N = 5; SNPtype = 3; M=1e3;
d = gen.count.data(ErrorProb=e,nSample=nSample,N=N,SNPtype=SNPtype,M=M);
d = MAFobj(d);
d = eLRT(d);
d = est.eLRT.distribution(d);
d = pvalues(d);
d = SNP.calling(d,FDR=0.05);
```

Index