Package 'iASeq'

September 24, 2012

Type Package

Title iASeq: integrating multiple sequencing datasets for detecting allele-specific events

Version 1.0.0

Date 2012-02-13

Author Yingying Wei, Hongkai Ji

Maintainer Yingying Wei <ywei@jhsph.edu>

Description It fits correlation motif model to multiple RNAseq or ChIPseq studies to improve detection of allele-specific events and describe correlation patterns across studies.

Depends R (>= 2.14.1)

Imports graphics, grDevices

License GPL-2

LazyLoad yes

biocViews SNP, RNAseq, ChIPseq, Bioinformatics

R topics documented:

`iASeq-package` *iASeq: integrating multiple sequencing datasets for detecting allele-specific events*

Description

In diploid organisms, certain genes can be expressed, methylated or regulated in an allele-specific manner, corresponding to allele-specific expression, allele-specific methylation and allele-specific binding. These allele-specific events (AS) are of high interest for phenotypic diversity and disease susceptibility. Next generation sequencing technologies provide opportunities to study AS globally. However, little is known about the mechanism of AS. For instance, the patterns of allele-specific binding across different Transcription Factors (TFs) and histone modifications (HMs) are unclear. Moreover, the limited number of reads on heterozygotic SNPs results in low-signal-to-noise ratio when calling AS. Here, we propose a Bayes hierarchical model to study AS by jointly analyzing multiple ChIPseq studies, RNAseq studies or MeDIPseq studies. The model is able to learn the patterns of AS across studies and make substantial improvement in calling AS.

Details

Package:	iASeq
Type:	Package
Version:	0.99.0
Date:	2012-02-13
License:	GPL-2

Author(s)

Yingying Wei, Hongkai Ji

Maintainer: Yingying Wei <ywei@jhsph.edu>

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

See Also

`iASeqmotif`, `plotBIC`, `plotMotif`, `sampleASE`, `ASErawfit`, `singleEMfit`, `sampleASE`

`ASErawfit` *Single Study Based Statistics for Allele Specific Events*

Description

This function produces standard statistics for allele-specific events based on a single RNAseq or ChIPseq study. It first pools replicates within a given study to sum the read counts for the reference allele and the non-reference allele. Then based on the pooled read counts, it calculates naive z statistic, naive Bayes statistic and empirical Bayes statistic.

Usage

```
ASErawfit(exprs,studyid,repid,refid)
```

Arguments

exprs	A matrix, each row of the matrix corresponds to a heterozygotic SNP and each column of the matrix corresponds to the reads count for either the reference allele or non-reference allele in a replicate of a study.
studyid	The group label for each column of exprs matrix. all columns in the same study have the same studyid.
repid	The sample label for each column of exprs matrix. The two columns within the same sample, one for reference allele and the other for non-reference allele, have the same repid. In other words, repid discriminates the different replicates within the same study.
refid	The reference allele label for each column of exprs matrix. Please code 0 for reference allele columns and 1 for non-reference allele columns to make the interpretation of over expressed (or bound) to be skewing to the reference allele. Otherwise, just interpret the other way round.

Details

One should indicate the studyid, repid and refid for each column clearly.

Value

z	Naive z statistic. A matrix, each row of the matrix corresponds to a heteroygpotic SNP of the input matrix ('exprs') and each column corresponds to a study.
b	Naive Bayes statistic. A matrix, each row of the matrix corresponds to a heteroygpotic SNP of the input matrix ('exprs') and each column corresponds to a study.
B	Empirical Bayes statistic. A matrix, each row of the matrix corresponds to a heteroygpotic SNP of the input matrix ('exprs') and each column corresponds to a study.
c0d	α parameter for the null beta prior distribution for pooled counts for each study. A vector whose length equals to the number of studies.
d0d	β parameter for the null beta prior distribution for pooled counts for each study. A vector whose length equals to the number of studies.
p0d	Mean of the null beta prior distribution for pooled counts for each study. A vector whose length equals to the number of studies.
p0dz	Raw mean of the reference allele proportion. A vector whose length equals to the number of studies.

Author(s)

Yingying Wei

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

See Also

sampleASE

Examples

```
data(sampleASE)
raw.fitted<-ASErawfit(sampleASE_exprs,sampleASE_studyid,sampleASE_repid,sampleASE_refid)
```

---

iASeq internal *iASeq Internal Functions*

Description

These functions are not part of the package application programming interface and are not recommended to be used by the users.

Usage

```
f0.loglike
fup.loglike
fdown.loglike
iASeqmotiffit
```

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

---

iASeqmotif *Correlation Motif Fit for Allele Specific Events*

Description

This function fits the Correlation Motif model to multiple RNAseq or ChIPseq studies. It gives the fitted values for the probability distribution of each motif, the fitted values of the given correlation matrix and the posterior probability for each SNP to be allele-specific events (allele-specific expression or allele-specific binding).

Usage

```
iASeqmotif(exprs,studyid,repid,refid,K,iter.max=100,tol=1e-3)
```

Arguments

exprs	A matrix, each row of the matrix corresponds to a heterozygotic SNP and each column of the matrix corresponds to the reads count for either the reference allele or non-reference allele in a replicate of a study.
studyid	The group label for each column of exprs matrix. all columns in the same study have the same studyid.
repid	The sample label for each column of exprs matrix. The two columns within the same sample, one for reference allele and the other for non-reference allele, have the same repid. In other words, repid discriminates the different replicates within the same study.
refid	The reference allele label for each column of exprs matrix. Please code 0 for reference allele columns and 1 for non-reference allele columns to make the interpretation of over expressed(or bound) to be skewing to the reference allele. Otherwise, just interpret the other way round.
K	A vector, each element specifing the number of non-null motifs a model wants to fit.
tol	The relative tolerance level of error.
iter.max	Maximun number of iterations.

Details

For the i^th element of K, the function fits total number of $K[i]+1$ motifs, $K[i]$ non-null motifs and the null motif, to the data. Each SNP can belong to one of the $K[i]+1$ possible motifs according to prior probability distribution, $motif.prior$. For SNPs in motif j ($j >= 1$), the probability that they are over expressed (or bound) for the reference allele in study d is $motif.qup(j,d)$ and the probability that they are under expressed (or bound) is $motif.qdown(j,d)$. One should indicate the studyid, repid and refid for each column clearly.

Value

bestmotif$p.post	The posterior probability for each SNP to be allele-specific event. A vector whose length correpsonds to the number of SNPs.
bestmotif$motif.prior	Fitted values of the probability distribution of the $K[i]+1$ motifs for the best fitted model, the first element specifies the null motif and the 2nd to $K[i]+1$th element correspond to the $K[i]$ non-null motifs.
bestmotif$motif.qup	Fitted values of the over expressed (or bound) correlation motif matrix for the best fitted model. Each row corresponds to a non-null motif and each column corresponds to a study.
bestmotif$motif.qdown	Fitted values of the under expressed (or bound) correlation motif matrix for the best fitted model. Each row corresponds to a non-null motif and each column corresponds to a study.
bestmotif$clustlike	Posterior probability for a SNP to belong to a specific motif based on the best fitted model. Each row corresponds to a SNP and each column corresponds to a motif class.
bestmotif$c0j	α parameter for the null beta prior distribution for each sample.

`bestmotif$d0j` β parameter for the null beta prior distribution for each sample.

`bestmotif$loglike`
The log-likelihood for the best fitted model.

`bic` The BIC values of all fitted models. A matrix whose first column is the same as input motif number vector ('K') and the second column corresponds to the BIC value of model given by the motif number in the first column in the same row.

`loglike` The log-likelihood of all fitted models. A matrix whose first column is the same as input motif number vector ('K') and the second column corresponds to the log likelihood value of the model given by the motif number in the first column in the same row.

Author(s)

Yingying Wei, Hongkai Ji

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji(2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

See Also

`plotBIC`, `plotMotif`, `sampleASE`

Examples

```
data(sampleASE)
#fit 1 to 2 non-null correlation motifs to the data
motif.fitted<-iASeqmotif(sampleASE_exprs,sampleASE_studyid,sampleASE_repid,sampleASE_refid,
K=1:2,iter.max=2,tol=1e-3)
```

---

plotBIC *BIC Plot*

Description

This function plots BIC values for all fitted motif models.

Usage

```
plotBIC(fitted_cormotif)
```

Arguments

`fitted_cormotif`
The object obtained from iASeq.

Author(s)

Yingying Wei

See Also

iASeqmotif, plotMotif, sampleASE

Examples

```
example(iASeqmotif) # compute 'motif.fitted'
plotBIC(motif.fitted)
```

---

plotMotif *Correlation Motif Plot*

Description

This function plots the Correlation Motif patterns, the associated prior probability distributions and the number of SNPs called for each motif based on posterior probability.

Usage

```
plotMotif(bestmotif,title="",cutoff)
```

Arguments

bestmotif	The bestmotif obtained from iASeqmotif.
title	The title on the figure.
cutoff	The posterior probability cutoff for calling a SNP belonging to certain motif.

Details

Each row in all graphs corresponds to one motif pattern. The first graph shows qup, the correlation motif pattern of over expression (binding). The second graph shows $qdown$, the correlation motif pattern of under expression (binding). The grey color scale of cell (k, d) indicates the probability that motif k is over or under expressed in study d. Each row of the two bar charts corresponds to the motif pattern in the same row of the left two pattern graphs. The length of the bar in the first bar chart estimates the number of SNPs of the given pattern in the dataset according to motif frequency, which is equal to $motif.fitted\$bestmotif\$motif.prior$ multiplying the number of total SNPs. The length of the bar in the second bar chart shows the number of SNPs called for the given pattern according to the $cutoff$ of posterior probability.

Author(s)

Yingying Wei, Hongkai Ji

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

See Also

iASeqmotif, plotBIC, sampleASE

Examples

```
example(iASeqmotif) # compute 'motif.fitted'
plotMotif(motif.fitted$bestmotif,cutoff=0.9)
```

`sampleASE` *Example Dataset for iASeq*

Description

Here we present four files needed for the various iASeq fit functions.

Details

sampleASE consists of five ChIP-seq studies from ENCODE GM12878 cell lines with 5504 heterozygotic SNPs. Each study has two replicates. Each replicate's fastq reads file was aligned to hg18 whole genome using MAQ (Version 0.7.1) with default parameters. Uniquely alignments were extracted following the mapping quality above 0. Alignment can also be down using other alignment tools such as Bowtie. The GM12878 genotype data was downloaded from the website http://alleleseq.gersteinlab.org/downloads.html [Rozowsky J et al.]. The reads aligned to each allele of a heterozygotic SNP were counted correspondingly. sampleASE_exprs saves the read counts. sampleASE_studyid prepares the study label for each sample; sample_repid describes the sample label for each column; sample_refid shows whether each column corresponds to the reference allele or the non-reference allele.

Value

`sampleASE_exprs`
: The read count matrix for the example dataset used by iASeq package. Each row of the matrix corresponds to a heterozygotic SNP and each column of the matrix corresponds to the reads count for either the reference allele or non-reference allele in a replicate of a study.

`sampleASE_studyid`
: The group label for each column of sampleASE_exprs matrix. All columns in the same study have the same studyid and there are five ChIP-seq studies in this example.

`sampleASE_repid`
: The sample label for each column of sampleASE_exprs matrix. The two columns within the same sample, one for reference allele and the other for non-reference allele, have the same repid. In other words, repid discriminates the different replicates within the same study. Here each study has two replicates.

`sampleASE_refid`
: The reference allele label for each column of sampleASE_exprs matrix. 0 is coded for reference allele columns and 1 is coded for non-reference allele columns.

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7:522.

See Also

iASeqmotif, plotBIC, plotMotif, ASErawfit, singleEMfit

Examples

```
data(sampleASE)
#fit 1 to 2 non-null correlation motifs to the data
motif.fitted<-iASeqmotif(sampleASE_exprs,sampleASE_studyid,sampleASE_repid,sampleASE_refid,
K=1:2,iter.max=2,tol=1e-3)
plotBIC(motif.fitted)
plotMotif(motif.fitted$bestmotif,cutoff=0.9)
```

---

singleEMfit *Single Study EM Fit for Allele Specific Events*

Description

This function runs an EM algorithm for allele-specific events based on a single RNAseq or ChIPseq study. It first pools replicates within a given study to sum the read counts for the reference allele and the non-reference allele. Then based on the pooled read counts, it fits an EM algorithm with three mixture components, the null distribution, the reference allele over expressed (bound) and under expressed (bound) distributions to the data.

Usage

```
singleEMfit(exprs,studyid,repid,refid,iter.max=100,tol=1e-3)
```

Arguments

exprs	A matrix, each row of the matrix corresponds to a heterozygotic SNP and each column of the matrix corresponds to the reads count for either the reference allele or non-reference allele in a replicate of a study.
studyid	The group label for each column of exprs matrix. All columns in the same study have the same studyid.
repid	The sample label for each column of exprs matrix. The two columns within the same sample, one for reference allele and the other for non-reference allele, have the same repid. In other words, repid discriminates the different replicates within the same study.
refid	The reference allele label for each column of exprs matrix. Please code 0 for reference allele columns and 1 for non-reference allele columns to make the interpretation of over expressed(or bound) to be skewing to the reference allele. Otherwise, just interpret the other way round.
tol	The relative tolerance level of error.
iter.max	Maximun number of iterations.

Value

p.study	The posterior probability for each SNP to be allele-specific event within each study. A matrix where each row corresponds to a SNP and each column corresponds to a study.
motif.qup	Fitted values of probability for the reference allele of each SNP to be over expressed (or bound) within each study. A matrix where each row corresponds to a SNP and each column corresponds to a study.
motif.qdown	Fitted values of probability for the reference allele of each SNP to be under expressed (or bound) within each study. A matrix where each row corresponds to a SNP and each column corresponds to a study.
condlike	A list where each element is a matrix and corresponds to a study. Each row of each matrix corresponds to a SNP. The three column of each matrix represents the posterior probability for a SNP to belong to the null distribution, the over expressed distribution and the under expressed distribution within the given study.

Author(s)

Yingying Wei

References

Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq: integrating multiple ChIP-seq datasets for detecting allele-specific binding.

See Also

sampleASE

Examples

```
data(sampleASE)
singleEM.fitted<-singleEMfit(sampleASE_exprs,sampleASE_studyid,sampleASE_repid,
sampleASE_refid,iter.max=2,tol=1e-3)
```

Index