Package ‘hummingbird’

March 20, 2024

Type  Package
Title  Bayesian Hidden Markov Model for the detection of differentially methylated regions
Version  1.12.0
Description  A package for detecting differential methylation. It exploits a Bayesian hidden Markov model that incorporates location dependence among genomic loci, unlike most existing methods that assume independence among observations. Bayesian priors are applied to permit information sharing across an entire chromosome for improved power of detection. The direct output of our software package is the best sequence of methylation states, eliminating the use of a subjective, and most of the time an arbitrary, threshold of p-value for determining significance. At last, our methodology does not require replication in either or both of the two comparison groups.
License  GPL (>=2)
Depends  R (>= 4.0)
Encoding  UTF-8
LazyData  true
Imports  Rcpp, graphics, GenomicRanges, SummarizedExperiment, IRanges
Suggests  knitr, rmarkdown, BiocStyle
LinkingTo  Rcpp
biocViews  HiddenMarkovModel, Bayesian, DNAMethylation, BiomedicalInformatics, Sequencing, GeneExpression, DifferentialExpression, DifferentialMethylation
VignetteBuilder  knitr
git_url  https://git.bioconductor.org/packages/hummingbird
git_branch  RELEASE_3_18
git_last_commit  8ba9af2
git_last_commit_date  2023-10-24
Repository  Bioconductor 3.18
Date/Publication  2024-03-20
hummingbird-package

**Author**  Eleni Adam [aut, cre],
Tieming Ji [aut],
Desh Ranjan [aut]

**Maintainer**  Eleni Adam <eadam002@odu.edu>

**R topics documented:**

- hummingbird-package
- abnormM
- abnormUM
- exampleHummingbird
- exampleSECase
- exampleSEControl
- hummingbirdEM
- hummingbirdEMinternal
- hummingbirdGraph
- hummingbirdPostAdjustment
- hummingbirdPostAdjustmentInternal
- normM
- normUM
- pos

**Description**

A package for identifying differentially methylated regions (DMRs) between case and control groups using whole genome bisulfite sequencing (WGBS) or reduced representative bisulfite sequencing (RRBS) experiment data.

The hummingbird package uses a Bayesian hidden Markov model (HMM) for detecting DMRs. It fits a Bayesian HMM for one chromosome at a time. The final output of hummingbird are the detected DMRs with start and end positions in a given chromosome, directions of the DMRs (hyper- or hypo-), and the numbers of CpGs in these DMRs.

The hummingbird package implements the algorithm described in the publication below.

**Details**

The main functions of the package are: hummingbirdEM, hummingbirdPostAdjustment and hummingbirdGraph.
abnormM

Author(s)
Eleni Adam, Tieming Ji, Desh Ranjan
Maintainer: Eleni Adam <eadam002@odu.edu>

References

<table>
<thead>
<tr>
<th>abnormM</th>
<th>Sample matrix</th>
</tr>
</thead>
</table>

Description
A matrix containing the methylated read count data of the case group. It is part of the sample dataset exampleHummingbird.

Usage
abnormM

Details
Each column of the matrix represents a replicate and each row represents a CpG position.

Source
Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples
data(exampleHummingbird)
abnormM
Sample matrix

Description

A matrix containing the unmethylated read count data of the case group. It is part of the sample dataset exampleHummingbird.

Usage

abnormUM

Details

Each column of the matrix represents a replicate and each row represents a CpG position.

Source

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples

data(exampleHummingbird)
abnormUM

Sample dataset

Description

Example of input data for the hummingbird package.

The sample dataset is partial data of chromosome 29 in the large offspring syndrome (LOS) study described in Chen Z. et al (2017).

Usage

data("exampleHummingbird")
exampleSECase

**Format**

- **expeirementSEControl**  A SummarizedExperiment object containing the input data for the control group: The two assays: normM, normUM and the CpG position information: pos.

- **experimentSECase**  A SummarizedExperiment object containing the input data for the case group: The two assays: abnormM, abnormUM and the CpG position information: pos.

- **normM**  A matrix containing the methylated read count data of the control group. Each column of the matrix represents a replicate and each row represents a CpG position.

- **normUM**  A matrix containing the unmethylated read count data of the control group. Each column of the matrix represents a replicate and each row represents a CpG position.

- **abnormM**  A matrix containing the methylated read count data of the case group. Each column of the matrix represents a replicate and each row represents a CpG position.

- **abnormUM**  A matrix containing the unmethylated read count data of case group. Each column of the matrix represents a replicate and each row represents a CpG position.

- **pos**  The CpG positions.

**Source**

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

**Examples**

```r
gp library(SummarizedExperiment)
gp data(exampleHummingbird)
```

---

**Description**

A SummarizedExperiment object containing the input data for the case group. It is part of the sample dataset exampleHummingbird.

**Usage**

```r
gp exampleSECase
```

**Details**

It contains the two assays: abnormM, abnormUM and the CpG position information: pos.
Source

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples

```r
library(SummarizedExperiment)
data(exampleHummingbird)
eexampleSECase
```

---

**exampleSEControl**  
*Sample input data*

Description

A SummarizedExperiment object containing the input data for the control group. It is part of the sample dataset exampleHummingbird.

Usage

```r
eexampleSEControl
```

Details

It contains the two assays: normM, normUM and the CpG position information: pos.

Source

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples

```r
library(SummarizedExperiment)
data(exampleHummingbird)
eexampleSEControl
```
**hummingbirdEM**

**EM Algorithm for Fitting the Hidden Markov Model**

**Description**

This function reads input data, sets initial values, executes the Expectation-Maximization (EM) algorithm for the Bayesian HMM and infers the best sequence of methylation states.

**Usage**

```r
hummingbirdEM(experimentInfoControl, experimentInfoCase, binSize)
```

**Arguments**

- `experimentInfoControl`
  - A SummarizedExperiment object containing the input data for the control group: The two assays: normM, normUM and the CpG position information: pos.
  - normM is a matrix containing the methylated read count data of the normal group and normUM is a matrix containing the unmethylated read count data of the normal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- `experimentInfoCase`
  - A SummarizedExperiment object containing the input data for the case group: The two assays: abnormM, abnormUM and the CpG position information: pos.
  - abnormM is a matrix containing the methylated read count data of the abnormal group and abnormUM is a matrix containing the unmethylated read count data of the abnormal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- `binSize`
  - The size of a bin. Default value is: 40.

**Value**

A GenomicRanges object that contains the start and end positions of each bin, the distance between the current bin and the bin ahead of it, the average methylation rate of normal and abnormal groups and the predicted direction of methylation change ("0" means a predicted normal bin; "1" means a predicted hypermethylated bin; "-1" means a predicted hypomethylated bin).

**Examples**

```r
library(GenomicRanges)
library(SummarizedExperiment)
data(exampleHummingbird)

# CpG position vector
pos <- pos[,1]
# Assays for the normal group
assaysControl <- list(normM = normM, normUM = normUM)
```
# Assays for the case group
assaysCase <- list(abnormM = abnormM, abnormUM = abnormUM)

# SummarizedExperiment objects
dataAnalysisControl <- SummarizedExperiment(assaysControl,
   rowRanges = GPos("chr29", pos))
dataAnalysisCase <- SummarizedExperiment(assaysCase,
   rowRanges = GPos("chr29", pos))

emInfo <- hummingbirdEM(experimentInfoControl = dataAnalysisControl,
   experimentInfoCase = dataAnalysisCase, binSize = 40)

hummingbirdEMinternal  EM Algorithm (internal)

---

**Description**

Expectation-Maximization Algorithm for Fitting the Hidden Markov Model. This function reads in methylated and unmethylated read count data, transforms it into logarithm bin-wise data, sets up initial values and implements the EM algorithm to estimate HMM parameters and find the best sequence of hidden states based on model fitting.

**Usage**

```
hummingbirdEMinternal(normM, normUM, abnormM, abnormUM, pos, binSize)
```

**Arguments**

- **normM**
  A matrix containing the methylated read count data of the normal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- **normUM**
  A matrix containing the unmethylated read count data of the normal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- **abnormM**
  A matrix containing the methylated read count data of the abnormal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- **abnormUM**
  A matrix containing the unmethylated read count data of the abnormal group. Each column of a matrix represents a replicate and each row represents a CpG position.

- **pos**
  The CpG position information.

- **binSize**
  The size of a bin.

**Details**

Users do not need to call this function directly. This is a low-level function used by the higher-level function in the hummingbird package, the hummingbirdEM.
**Value**

obs
For each bin: The predicted direction of methylation change ("0" means a predicted normal bin; "1" means a predicted hypermethylated bin; "-1" means a predicted hypomethylated bin). The distance between the current bin and the bin ahead of it, the start and end positions of each bin.

normAbnorm
The average methylation rate of normal and abnormal groups.

**See Also**

Users may call the `hummingbirdEM` function.

**Examples**

```r
library(GenomicRanges)
library(SummarizedExperiment)

# Load sample dataset containing input data
data(exampleHummingbird)

# Run the EM (internal) function
hmmbirdEMinternalOutput <- hummingbirdEMinternal(
  normM = assays(exampleSEControl)[["normM"]],
  normUM = assays(exampleSEControl)[["normUM"]],
  abnormM = assays(exampleSECase)[["abnormM"]],
  abnormUM = assays(exampleSECase)[["abnormUM"]],
  pos = pos, binSize = 40)
```

---

**hummingbirdGraph Observations and Predictions Graphs**

**Description**

This function generates observation and prediction graphs for a user specified region.

**Usage**

```r
hummingbirdGraph(experimentInfoControl, experimentInfoCase, postAdjInfoDMRs, coord1, coord2)
```

**Arguments**

**experimentInfoControl**
A SummarizedExperiment object containing the input data for the control group:
The two assays: normM, normUM and the CpG position information: pos.

**experimentInfoCase**
A SummarizedExperiment object containing the input data for the case group:
The two assays: abnormM, abnormUM and the CpG position information: pos.
The DMRs GenomicRanges object output of the `hummingbirdPostAdjustment` function.

coord1 The start coordinate of the genomic region to plot.

coord2 The end coordinate of the genomic region to plot.

Value

The function outputs two graphs: The Observations graph and the Predictions graph. The observation figure shows bin-wise average methylation rate for case and control groups. The prediction figure shows bin-wise prediction, where "0" denotes a predicted normal bin; "1" denotes a predicted hypermethylated bin; and "-1" denotes a predicted hypomethylated bin.

Examples

```r
library(GenomicRanges)
library(SummarizedExperiment)
data(exampleHummingbird)
emInfo <- hummingbirdEM(experimentInfoControl = exampleSEControl,
                        experimentInfoCase = exampleSECase, binSize = 40)
postAdjInfo <- hummingbirdPostAdjustment(
    experimentInfoControl = exampleSEControl,
    experimentInfoCase = exampleSECase,
    emInfo = emInfo, minCpGs = 10,
    minLength = 100, maxGap = 300)

hummingbirdGraph(experimentInfoControl = exampleSEControl,
                  experimentInfoCase = exampleSECase,
                  postAdjInfoDMRs = postAdjInfo$DMRs,
                  coord1 = 107991, coord2 = 108350)
```

**Description**

This function adjusts HMM output. It enables three additional requirements on DMRs: 1) the minimum length of a DMR, 2) the minimum number of CpGs in a DMR, and 3) the maximum distance (in base pairs) between any two adjacent CpGs in a DMR.

**Usage**

```r
hummingbirdPostAdjustment(experimentInfoControl, experimentInfoCase, emInfo,
                          minCpGs, minLength, maxGap)
```
Arguments

- **experimentInfoControl**
  A SummarizedExperiment object containing the input data for the control group: The two assays: normM, normUM and the CpG position information: pos.

- **experimentInfoCase**
  A SummarizedExperiment object containing the input data for the case group: The two assays: abnormM, abnormUM and the CpG position information: pos.

- **emInfo**
  The output GenomicRanges object of the hummingbirdEM function.

- **minCpGs**
  The minimum number of CpGs contained in a DMR. Default value: 10.

- **minLength**
  The minimum length of a DMR. Default value: 500.

- **maxGap**
  The maximum gap between any two CpGs. Default value: 300.

Value

A list of two GenomicRanges objects, the DMRs and the obsPostAdj.

- **DMRs**
  Contains the detected regions based on the user-defined arguments (minLength, minCpGs, and maxGap). It contains the refined DMRs with the start genomic position, the end genomic position, length of the region, direction of predicted methylation change ("0" indicates no significant change, "1" indicates predicted hyper-methylation, and "-1" indicates predicted hypo-methylation), and the number of CpGs.

- **obsPostAdj**
  The methylation status of each CpG site.

Examples

```r
library(GenomicRanges)
library(SummarizedExperiment)
data(exampleHummingbird)
emInfo <- hummingbirdEM(experimentInfoControl = exampleSEControl,
                         experimentInfoCase = exampleSECase, binSize = 40)
postAdjInfo <- hummingbirdPostAdjustment(
                           experimentInfoControl = exampleSEControl,
                           experimentInfoCase = exampleSECase,
                           emInfo = emInfo, minCpGs = 10,
                           minLength = 100, maxGap = 300)
```

Description

Post Adjustment algorithm for the output of the EM. This function adjusts HMM output such that each detected DMR has a minimum length and a minimum number of CpGs in each DMR.
hummingbirdPostAdjustmentInternal

Usage

hummingbirdPostAdjustmentInternal(em, pos, minCpGs, minLength, maxGap)

Arguments

em The output of the hummingbirdEMinternal function, specifically the obs object.
pos The CpG position information.
minCpGs The minimum number of CpGs contained in a DMR.
minLength The minimum length of a DMR.
maxGap The maximum gap between any two CpGs.

Details

Users do not need to call this function directly. This is a low-level function used by the higher-level function in the hummingbird package, the hummingbirdPostAdjustment.

Value

DMRs The detected regions based on the user-defined arguments (minLength, minCpGs, and maxGap).
It contains the (numbered) refined DMRs with the start genomic position, the end genomic position, length of the region, direction of predicted methylation change ("0" indicates no significant change, "1" indicates predicted hypermethylation, and "-1" indicates predicted hypo-methylation) and the number of CpGs.

obsPostAdj The methylation status of each CpG site.

See Also

Users may call the hummingbirdPostAdjustment function.

Examples

library(GenomicRanges)
library(SummarizedExperiment)

# Load sample dataset containing input data
data(exampleHummingbird)

# Run the EM (internal) function
hmmbirdEMinternalOutput <- hummingbirdEMinternal(  
    normM = assays(exampleSEControl)[["normM"]],  
    normUM = assays(exampleSEControl)[["normUM"]],  
    abnormM = assays(exampleSECase)[["abnormM"]],  
    abnormUM = assays(exampleSECase)[["abnormUM"]],  
    pos = pos, binSize = 40)

# Run the Post Adjustment (internal) function
hmmbirdPAinternalOutput <- hummingbirdPostAdjustmentInternal(  
    DMRs = hmmbirdEMinternalOutput,  
    pos = pos, minCpGs = 2, minLength = 100, maxGap = 50)
Sample matrix  

**Description**  
A matrix containing the methylated read count data of the control group. It is part of the sample dataset exampleHummingbird.

**Usage**  
normM

**Details**  
Each column of the matrix represents a replicate and each row represents a CpG position.

**Source**  
Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

**Examples**  
```r
data(exampleHummingbird)
normM
```

Sample matrix  

**Description**  
A matrix containing the unmethylated read count data of the control group. It is part of the sample dataset exampleHummingbird.

**Usage**  
normUM

**Details**  
Each column of the matrix represents a replicate and each row represents a CpG position.
Source

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples

data(exampleHummingbird)
normUM

table(pos)

<table>
<thead>
<tr>
<th>pos</th>
<th>Sample matrix</th>
</tr>
</thead>
</table>

Description

The CpG positions. It is part of the sample dataset exampleHummingbird.

Usage

pos

Source

Chen et al. (2017) Global misregulation of genes largely uncoupled to DNA methylome epimutations characterizes a congenital overgrowth syndrome. Scientific Reports 7, 12667

The raw FASTQ files of the WGBS experiment from this study are publicly available at Gene Expression Omnibus (GEO) database with accession no. GSE93775.

Examples

data(exampleHummingbird)
pos
Index

* datasets
  abnormM, 3
  abnormUM, 4
  exampleHummingbird, 4
  exampleSECase, 5
  exampleSEControl, 6
  normM, 13
  normUM, 13
  pos, 14

abnormM, 3
abnormUM, 4

exampleHummingbird, 4
exampleSECase, 5
exampleSEControl, 6

hummingbird (hummingbird-package), 2
hummingbird-package, 2
hummingbirdEM, 7, 9
hummingbirdEMinternal, 8
hummingbirdGraph, 9
hummingbirdPostAdjustment, 10, 12
hummingbirdPostAdjustmentInternal, 11

normM, 13
normUM, 13

pos, 14