Package 'COHCAP'

April 9, 2015

Type Package

Title CpG Island Analysis Pipeline for Illumina Methylation Array and Targeted BS-Seq Data

Version 1.4.0

Date 2014-4-15

Author Charles Warden

Maintainer Charles Warden <cwarden45@gmail.com>

Depends WriteXLS, COHCAPanno

Description This package provides a pipeline to analyze single-nucleotide resolution methylation data (Illumina 450k methylation array, targeted BS-Seq, etc.). It provides QC metrics, differential methylation for CpG Sites, differential methylation for CpG Islands, integration with gene expression data, and visualization of methylation values.

License GPL-3

LazyLoad yes

biocViews DNAMethylation, Microarray, MethylSeq, Epigenetics, DifferentialMethylation

R topics documented:

COHCAP.annotate *Annotation for CpG Sites*

Description

Provides annotations (chromosome, location, gene, and CpG island) for CpG sites from a specified annotation file.

Annotations for common platforms are provided with COHCAP (with respect to hg19). Custom annotation files can also be provided.

Output files will be created in the "Raw_Data" subfolder.

Usage

```
COHCAP.annotate(beta.file, project.name, project.folder,
platform, annotation.file = NULL,
output.format = "xls")
```

Arguments

beta.file	Table of beta / percentage methylation values. CpG sites are represented in rows. Samples are represented in columns.
project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files
platform	Annotation file to be used. Enter "450k-UCSC" for UCSC CpG Islands for 450k array probes, "450k-HMM" for HMM CpG Islands for 450k array probes, "27k" for UCSC CpG Islands for 27k array probes. If none of these pre-defined annotations are acceptable, please enter "custom" for the platform and provide an annotation file.
annotation.file	Annotation file to be used for a custom platform. This variable is not used for common, pre-defined annotation files. The annotation file should be a tab-delimited text file with the header "SiteID Chr Loc Gene Island", with columns respectively specifying the CpG identifier (must match beta / percent methylation file), chromosome for CpG site, position for CpG site (preferably in hg19 coordinates), nearest gene mapping for CpG site, nearest CpG island mapping for CpG site.
output.format	Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file

Value

Data frame of beta values (must be between 0 and 1) or percentage methylation values (must be between 0 and 100).

Just like the input table, the first column specifies the SiteID, CpG sites are represented on rows, samples are represented in samples (starting with the 6th column). Additionally, the 2nd column

now specifies the CpG site chromosome, the 3rd column now specifies the CpG site position (in hg19 coordinates, for pre-defined annotation files), the 4th column lists the nearest gene mapping, and the 5th column lists the nearest CpG island mapping.

This data frame is used for quality control and differential methylation analysis.

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
project.folder <- getwd()
project.name <- "450k_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
```

COHCAP.avg.by.island *CpG Island Differential Methylation Analysis (Average by Island Workflow).*

Description

Provides statistics for CpG islands as well as a list of differentially methylated sites. CpG Island statistics are calculated by averaging beta values among samples per site and comparing the average beta values across groups (considering the pairing between sites).

List of differentially methylated islands will be created in the "CpG_Island" folder. Table of statistics for all CpG islands will be created in the "Raw_Data" folder.

Usage

```
COHCAP.avg.by.island(sample.file, site.table, beta.table, project.name,
project.folder, methyl.cutoff=0.7, unmethyl.cutoff = 0.3,
delta.beta.cutoff = 0.2, pvalue.cutoff=0.05, fdr.cutoff=0.05,
num.groups=2, num.sites=4, plot.box=TRUE, paired=FALSE, ref="none",
output.format = "xls", gene.centric=TRUE)
```

Arguments

sample.file	Tab-delimited text file providing group attributions for all samples considered for analysis.

`beta.table`	Data frame with CpG sites in columns (with DNA methylation represented as beta values or percentage methylation), samples in columns, and CpG site annotations are included (in columns 2-5). The COHCAP.annotate function automatically creates this file.
`site.table`	Data frame with CpG site statistics (one row per CpG site) and CpG site annotations (in columns 2-5). The COHCAP.site function automatically creates this file.
`project.name`	Name for COHCAP project. This determines the names for output files.
`project.folder`	Folder for COHCAP output files
`methyl.cutoff`	Minimum beta or percentage methylation value to be used to define a methylated CpG site. Default is 0.7 (used for beta values), which would correspond to 70 Used for either 1-group or 2-group comparison.
`unmethyl.cutoff`	Minimum beta or percentage methylation value to be used to define an unmethylated CpG site. Default is 0.3 (used for beta values), which would correspond to 30 Used for either 1-group or 2-group comparison.
`delta.beta.cutoff`	The minimum absolute value for delta-beta values (mean treatment beta - mean reference beta) to define a differentially methylated CpG site. Only used for 2-group comparison.
`pvalue.cutoff`	Maximum p-value allowed to define a site as differentially methylated. Used only for comparisons with at least 2 groups (with 3 replicates per group)
`fdr.cutoff`	Maximum False Discovery Rate (FDR) allowed to define a site as differentially methylated. Used only for comparisons with at least 2 groups (with 3 replicates per group)
`num.groups`	Number of groups described in sample description file. COHCAP algorithm differs when analysing 1-group, 2-group, or >2-group comparisons. COHCAP cannot currently analyze continuous phenotypes.
`num.sites`	Minimum number of differentially methylated sites to define a differentially methylated CpG island.
`ref`	Reference group used to define baseline methylation levels. Only used for 2-group comparison
`plot.box`	A logical value: Should box-plots be created to visualize CpG island differential methylation?
`paired`	A logical value: Is there any special pairing between samples in different groups? If so, the pairing variable must be specified in the 3rd column of the sample description file. Used for p-value calculation, so this only applies to comparisons with at least 2 groups.
`output.format`	Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file
`gene.centric`	Should CpG islands not mapped to genes be ignored? Default: TRUE (Recommended setting for integration with gene expession data)

Value

Data frame of average beta (or percentage methylation) values across differentially methylated sites within a differentially methylated CpG island.. This data frame can be used for integration analysis.

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
project.name <- "450k_avg_by_island_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
filtered.sites <- COHCAP.site(sample.file, beta.table, project.name,
project.folder, ref="parental")
filtered.islands <- COHCAP.avg.by.island(sample.file, filtered.sites,
beta.table, project.name, project.folder, ref="parental")
```

COHCAP.avg.by.site *CpG Island Differential Methylation Analysis (Average by Site Workflow).*

Description

Provides statistics for CpG islands as well as a list of differentially methylated sites. CpG Island statistics are calculated by averaging beta values among samples per site and comparing the average beta values across groups (considering the pairing between sites).

List of differentially methylated islands will be created in the "CpG_Island" folder. Table of statistics for all CpG islands will be created in the "Raw_Data" folder.

Usage

```
COHCAP.avg.by.site(site.table, project.name, project.folder,
methyl.cutoff=0.7, unmethyl.cutoff = 0.3,
delta.beta.cutoff = 0.2, pvalue.cutoff=0.05,
fdr.cutoff=0.05, num.groups=2, num.sites=4,
output.format = "xls")
```

Arguments

site.table	Data frame with CpG site statistics (one row per CpG site) and CpG site annotations (in columns 2-5). The COHCAP.site function automatically creates this file.
project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files

`methyl.cutoff` Minimum beta or percentage methylation value to be used to define a methylated CpG site. Default is 0.7 (used for beta values), which would correspond to 70 Used for either 1-group or 2-group comparison.

`unmethyl.cutoff` Minimum beta or percentage methylation value to be used to define an unmethylated CpG site. Default is 0.3 (used for beta values), which would correspond to 30 Used for either 1-group or 2-group comparison.

`delta.beta.cutoff` The minimum absolute value for delta-beta values (mean treatment beta - mean reference beta) to define a differentially methylated CpG site. Only used for 2-group comparison.

`pvalue.cutoff` Maximum p-value allowed to define a CpG island as differentially methylated.

`fdr.cutoff` Maximum False Discovery Rate (FDR) allowed to define CpG island as differentially methylated.

`num.groups` Number of groups described in sample description file. COHCAP algorithm differs when analysing 1-group, 2-group, or >2-group comparisons. COHCAP cannot currently analyze continuous phenotypes.

`num.sites` Minimum number of differentially methylated sites to define a differentially methylated CpG island.

`output.format` Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file

Value

Data frame of average beta (or percentage methylation) statistics and/or p-value / false discovery rate statistics (per CpG island).

The content of the data frame depends upon the number of groups specified for analysis. All workflows provide p-values and FDR values. 1 and 2 group comparisons provide counts for methylated and unmethylated sites as well as an overall methylation status per island. >2 group comparisons only provide counts for the total number of differentially methylated sites.

This data frame can be used for integration analysis.

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
project.name <- "450k_avg_by_site_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
```

```
filtered.sites <- COHCAP.site(sample.file, beta.table, project.name,
project.folder, ref="parental")
filtered.islands <- COHCAP.avg.by.site(filtered.sites, project.name,
project.folder)
```

`COHCAP.BSSeq.preprocess`

Preprocessing for Targeted BS-Seq data

Description

Creates custom annotation file as well as COHCAP input file (for COHCAP.annotate).

This function is not necessary for Illumina methylation array analysis.

Output files will be created in specified locations

Usage

```
COHCAP.BSSeq.preprocess(methyl.folder=getwd(),
cohcap.inputfile = file.path(getwd(),"BS_Seq_combined.txt"),
gene.table = file.path(getwd(),"GENCODE_Genes.bed"),
targeted.regions = file.path(getwd(),"UCSC_CpG_Islands.bed"),
annotation.file = file.path(getwd(),"COHCAP.targeted.BSSeq.anno.txt"),
shore.length=2000)
```

Arguments

`methyl.folder`	Folder containing .bed files created using genome_methylation_bismark2bedGraph_v3.pl (following Bismark alignment)
`cohcap.inputfile`	Output file containing a tab-delimited table of percentage methylation values. This table will compatible with the custom annotation file created by this function (annotation.file)
`gene.table`	.bed file containing gene names and coordinates
`targeted.regions`	.bed file containing regions selected for targeted BS-Seq
`annotation.file`	Custom annotation file providing gene and targeted region mappings for CpG sites specifically covered in your Bismark alignment
`shore.length`	Length of shores considered to be part of the CpG island (in bp upstream and downstream of targeted region coordinates)

Value

This function creates two tab-delimited text files.

One is to be used to as a custom annotation file (annotation.file).

The other is used to create an appropriate input file for COHCAP (cohcap.inputfile).

This function will likely take several hours to run. However, it only needs to be run once.

See Also

Useful Example Files: http://sourceforge.net/projects/cohcap/files/COHCAP_BSSEQ_anno.zip/download *Default settings utilize these files (in current working directory) *These files were created using the UCSC Genome Browser (build hg19)

Raw Data for Demo Dataset: http://www.ncbi.nlm.nih.gov/sra/SRX084504 Full, Formatted Demo Dataset (in standalone package): http://sourceforge.net/projects/cohcap/

Bismark: http://www.bioinformatics.babraham.ac.uk/projects/bismark/

UCSC Genome Browser: http://genome.ucsc.edu/cgi-bin/hgTables?command=start

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
bed.folder <- file.path(dir,"BSSeq")
gene.table <- file.path(dir,"GENCODE_Genes_truncated.bed")
targeted.regions <- file.path(dir,"UCSC_CpG_Islands_truncated.bed")

output.folder <- getwd()
annotation.file <- file.path(output.folder,"COHCAP.targeted.BSSeq.anno.txt")
cohcap.inputfile <- file.path(output.folder,"BS_Seq_combined.txt")

COHCAP.BSSeq.preprocess(bed.folder, cohcap.inputfile, gene.table,
targeted.regions,annotation.file)
```

COHCAP.integrate.avg.by.island

Integration with Gene Expression Data(Average by Island Workflow).

Description

Provides lists of genes with a significant negative correlation between DNA methylation and gene expression data.

A table of normalizated intensity / expression values is provided in the gene expression table and a table of filtered beta values is provided for the DNA methylation data.

Lists of genes with negative expression trends will be created in the "Integrate" folder, along with scatter plots (if descired). All correlation stats are provided in the "Raw_Data" folder.

Usage

```
COHCAP.integrate.avg.by.island(beta.table, project.name, project.folder,
expr.file, sample.file, cor.pvalue.cutoff=0.05,
cor.fdr.cutoff = 0.05, cor.cutoff = -0.2, plot.scatter=TRUE,
output.format = "xls")
```

Arguments

beta.table	Data frame with beta values averaged across differentially methylated sites (islands in rows, samples in columns). This table is already filtered for differentially methylated CpG islands. The COHCAP.avg.by.island function automatically creates this file.
project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files
expr.file	Table of normalized expression or intensity values (can be for either microarray or RNA-Seq data). Sample IDs (listed in column header) must match the sample IDs used for the DNA methylation data (e.g. those listed in beta.table)
sample.file	Tab-delimited text file providing group attributions for all samples considered for analysis. Only used if plot.scatter=TRUE
cor.cutoff	The minimum negative correlation coefficient to define a differentially expressed gene.
cor.pvalue.cutoff	Maximum p-value allowed to define a gene as differentially expressed.
cor.fdr.cutoff	Maximum False Discovery Rate (FDR) allowed to define a gene as differentially expressed.
plot.scatter	A logical value: Create scatter plot for genes with a significant negative correlation?
output.format	Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
expression.file <- file.path(dir,"expression-Average_by_Island_truncated.txt")
project.name <- "450k_avg_by_island_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
filtered.sites <- COHCAP.site(sample.file, beta.table, project.name,
project.folder, ref="parental")
filtered.islands <- COHCAP.avg.by.island(sample.file, filtered.sites,
beta.table, project.name, project.folder, ref="parental")
COHCAP.integrate.avg.by.island(filtered.islands, project.name,
project.folder, expression.file, sample.file)
```

COHCAP.integrate.avg.by.site

Integration with Gene Expression Data(Average by Site Workflow).

Description

Provides lists of genes with an inverse CpG island methylation trend (Methylation Down, Expression Up and Methylation Up, Expression Down).

Lists of genes with negative expression trends will be created in the "Integrate" folder.

The "Average by Site" workflow requires that genes already have fold-change, p-value, and FDR values calculated. There many tools available for this type of analysis (limma, sRAP, DEseq, edgeR, etc.)

This function will only work for 2-group comparisons.

Usage

```
COHCAP.integrate.avg.by.site(island.table, project.name, project.folder,
expr.file, expr.pvalue=0.05, expr.fdr = 0.05, expr.fc = 1.5,
output.format = "xls")
```

Arguments

island.table	Data frame with CpG island statistics (one row per CpG island) for differentially methylated CpG islands. The COHCAP.avg.by.site function automatically creates this file.
project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files
expr.file	Table of differential expression statistics. Gene symbols must be in the first column, fold-change values must be in the second column, p-values must be in the third column, and false discovery rate (FDR) values must be in the fourth column. These statistics must be calculated outside of COHCAP. Duplicate gene symbols are OK - statistics will be averaged among duplicate gene symbols.
expr.fc	The minimum absolute value for fold-change values (treatment versus reference) to define gene as differentially expressed (from gene expression table). Only used for 2-group comparison. Fold-change is expected to be on a linear scale.
expr.pvalue	Maximum p-value allowed to define a gene as differentially expressed (from gene expression table).
expr.fdr	Maximum False Discovery Rate (FDR) allowed to define a gene as differentially expressed (from gene expression table)
output.format	Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

sRAP: http://www.bioconductor.org/packages/release/bioc/html/sRAP.html

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
expression.file <- file.path(dir,"expression-Average_by_Site_truncated.txt")
project.name <- "450k_avg_by_site_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
filtered.sites <- COHCAP.site(sample.file, beta.table, project.name,
project.folder, ref="parental")
filtered.islands <- COHCAP.avg.by.site(filtered.sites, project.name,
project.folder)
COHCAP.integrate.avg.by.site(filtered.islands, project.name, project.folder,
expression.file)
```

COHCAP.qc *DNA Methylation Quality Control Statistics*

Description

Provides descriptive statistics (median, top/bottom quartiles, mininum,maximum), sample histograms, sample dendrogram, principal component analysis plot.

Output files will be created in the "QC" subfolder.

Usage

```
COHCAP.qc(sample.file, beta.table, project.name, project.folder,
plot.legend=TRUE, color.palette = c("red","blue",
"green","orange","purple","cyan","pink","maroon",
"yellow","grey","black",colors()))
```

Arguments

sample.file	Tab-delimited text file providing group attributions for all samples considered for analysis.
beta.table	Data frame with CpG sites in columns (with DNA methylation represented as beta values or percentage methylation), samples in columns, and CpG site annotations are included (in columns 2-5). The COHCAP.annotate function automatically creates this file.

project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files
plot.legend	A logical value: Should legend be plotted within QC figures?
color.palette	Colors for primary variable (specified in the second column of the sample file). Remember, COHCAP can only analyze discrete variables categoried with groups (preferably with replicates).

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
project.name <- "450k_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
COHCAP.qc(sample.file, beta.table, project.name, project.folder)
```

COHCAP.site *CpG Site Differential Methylation Analysis*

Description

Provides statistics for CpG sites as well as a list of differentially methylated sites. Can also provide .wig files for visualization in IGV, UCSC Genome Browser, etc.

List of differentially methylated sites and .wig files will be created in the "CpG_Site" folder. Table of statistics for all CpG sites will be created in the "Raw_Data" folder.

Usage

```
COHCAP.site(sample.file, beta.table, project.name, project.folder,
methyl.cutoff=0.7, unmethyl.cutoff = 0.3,
delta.beta.cutoff = 0.2, pvalue.cutoff=0.05,
fdr.cutoff=0.05, ref="none", num.groups=2,
create.wig = TRUE, paired=FALSE, output.format = "xls")
```

Arguments

sample.file	Tab-delimited text file providing group attributions for all samples considered for analysis.
beta.table	Data frame with CpG sites in columns (with DNA methylation represented as beta values or percentage methylation), samples in columns, and CpG site annotations are included (in columns 2-5). The COHCAP.annotate function automatically creates this file.
project.name	Name for COHCAP project. This determines the names for output files.
project.folder	Folder for COHCAP output files
methyl.cutoff	Minimum beta or percentage methylation value to be used to define a methylated CpG site. Default is 0.7 (used for beta values), which would correspond to 70 Used for either 1-group or 2-group comparison.
unmethyl.cutoff	Minimum beta or percentage methylation value to be used to define an unmethylated CpG site. Default is 0.3 (used for beta values), which would correspond to 30 Used for either 1-group or 2-group comparison.
delta.beta.cutoff	The minimum absolute value for delta-beta values (mean treatment beta - mean reference beta) to define a differentially methylated CpG site. Only used for 2-group comparison.
pvalue.cutoff	Maximum p-value allowed to define a site as differentially methylated. Used only for comparisons with at least 2 groups (with 3 replicates per group)
fdr.cutoff	Maximum False Discovery Rate (FDR) allowed to define a site as differentially methylated. Used only for comparisons with at least 2 groups (with 3 replicates per group)
ref	Reference group used to define baseline methylation levels. Only used for 2-group comparison
num.groups	Number of groups described in sample description file. COHCAP algorithm differs when analysing 1-group, 2-group, or >2-group comparisons. COHCAP cannot currently analyze continuous phenotypes.
create.wig	A logical value: Create .wig files (using average beta and delta-beta values)? In the standalone version of COHCAP, this was only an option when using the "Average by Site" workflow (because that was the only situation where the analysis method matched the visualization). .wig files are defined with respect to hg19 (for pre-defined annotation files) and can be visualized using IGV, UCSC Genome Browser, etc.
paired	A logical value: Is there any special pairing between samples in different groups? If so, the pairing variable must be specified in the 3rd column of the sample description file. Used for p-value calculation, so this only applies to comparisons with at least 2 groups.
output.format	Format for output tables: 'xls' for Excel file, 'txt' for tab-delimited text file

Value

Data frame of average beta (or percentage methylation) statistics and/or p-value / false discovery rate statistics.

The content of the data frame depends upon the number of groups specified for analysis (avg.beta only for 1-group; avg.beta, delta.beta, p-value, and FDR for 2-group; p-value and FDR only for >2 groups).

This data frame is used for CpG island analysis.

See Also

COHCAP Discussion Group: http://sourceforge.net/p/cohcap/discussion/general/

Examples

```
library("COHCAP")

dir <- system.file("extdata", package="COHCAP")
beta.file <- file.path(dir,"GSE42308_truncated.txt")
sample.file <- file.path(dir,"sample_GSE42308.txt")
project.folder <- getwd()
project.name <- "450k_test"

beta.table <- COHCAP.annotate(beta.file, project.name, project.folder,
platform="450k-UCSC")
filtered.sites <- COHCAP.site(sample.file, beta.table, project.name,
project.folder, ref="parental")
```

Index