Package 'COHCAP'

September 23, 2020

Type Package

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

Version 1.34.1

Date 2020-04-30

Author Charles Warden <cwarden@coh.org>, Yate-Ching Yuan <yyuan@coh.org>, Xiwei Wu <xwu@coh.org>

Maintainer Charles Warden <cwarden@coh.org>

Depends WriteXLS, COHCAPanno, RColorBrewer, gplots

Imports Rcpp, RcppArmadillo, BH

LinkingTo Rcpp, BH

Description COHCAP (pronounced ``co-cap'') provides a pipeline to analyze single-nucleotide resolution methylation data (Illumina 450k/EPIC methylation array, targeted BS-Seq, etc.). It provides differential methylation for CpG Sites, differential methylation for CpG Islands, integration with gene expression data, with visualizaton options. Discussion Group: https://sourceforge.net/p/cohcap/discussion/bioconductor/

License GPL-3

SystemRequirements Perl

LazyLoad yes

biocViews DNAMethylation, Microarray, MethylSeq, Epigenetics, DifferentialMethylation

git_url https://git.bioconductor.org/packages/COHCAP

git_branch RELEASE_3_11

git_last_commit 5882753

git_last_commit_date 2020-04-30

Date/Publication 2020-09-23

R topics documented:

COHCAP.annotate 2
COHCAP.avg.by.island 3
COHCAP.avg.by.site 6
COHCAP.BSSeq.preprocess 8

COHCAP.BSSeq_V2.methyl.table 9
COHCAP.denovo 11
COHCAP.integrate.avg.by.island 12
COHCAP.integrate.avg.by.site 14
COHCAP.qc 15
COHCAP.reformatFinalReport 16
COHCAP.site 17

Index **21**

`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 = "txt")
```

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, 'csv' for comma-separated file, or '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 = tempdir()#you may want to use getwd() or specify another folder
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, plot.heatmap=TRUE,
paired=FALSE, ref="none", lower.cont.quantile=0, upper.cont.quantile=1,
max.cluster.dist = NULL, alt.pvalue="none",
output.format = "txt", gene.centric=TRUE, heatmap.dist.fun="Euclidian")
```

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 differentially methylated 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 island. 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 island. Default is 0.3 (used for beta values), which would correspond to 30 Used for either 1-group or 2-group comparison.
max.cluster.dist	Update annotations by running de-novo clustering within each set of provided annotations. This is the maximum distance (in bp) between filtered sites with a consistent delta-beta sign. This can produce more than one cluster per pre-existing annotation. Set to NULL by default, to run standard COHCAP algorithm. If you would like to test this function, I would recommend a value between 50 and 500 bp. Only valid for 2-group or continuous comparison.
delta.beta.cutoff	The minimum absolute value for delta-beta values (mean treatment beta - mean reference beta) to define a differentially methylated CpG island. Only used for 2-group comparison (and continuous comparison, where delta beta is max - min beta, with sign based upon correlation coefficient).
pvalue.cutoff	Maximum p-value allowed to define an island 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 an island 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. Set to "continuous" for a continuous primary variable. Otherwise, only used for 2-group comparison.
plot.box	Logical value: Should box-plots be created to visualize CpG island differential methylation? If using a continuous primary variable, line plots are provided instead of box plots.

`plot.heatmap` Logical value: Should heatmap be created to visualize CpG island differential methylation?

`lower.cont.quantile`
For continuous analysis, what beta quantile should be the lower threshold for calculating delta-beta values? Default = 0 (minimum)

`upper.cont.quantile`
For continuous analysis, what beta quantile should be the upper threshold for calculating delta-beta values? Default = 1 (maximum)

`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.

If you have a secondary continuous variable (like age), you can set paired to "continuous". COHCAP will then perform linear-regression analysis (converting primary categorical variable into continuous variable, if necessary)

`alt.pvalue` Use alternative strategies for p-value calculations.

Be careful that the workflow matches the p-value calculation.

For 'rANOVA.1way', use ANOVA (R function) instead of t-test (for 1-variable, 2-group comparison). Might be helpful when SD is 0.

For 'cppANOVA.1way', use ANOVA (C++ code) instead of t-test (for 1-variable, 2-group comparison). Helps decrease run-time relative to R-code.

For 'cppANOVA.2way', use C++ code instead of R function for t-test (for 2-variable, 2-group comparison). Helps decrease run-time relative to R-code, but p-value may be different. For this implementation, I require having at replicates for each interaction term (such as having replicate treatmentswith multiple backgrounds, cell lines, etc.). If each sample has exactly one pair (as may be the case with tumor-normal pairs), and you need to decrease the COHCAP run-time, you may consider using 'cppPairedTtest'.

For 'cppWelshTtest', use C++ code instead of R function for t-test (for 1-variable, 2-group comparison). Helps decrease run-time relative to R-code, but p-value will be different than t.test() function (C++ code assumes unequal variance between groups).

For 'cppPairedTtest', use C++ code instead of R function for t-test (for 2-variable, 2-group comparison). Helps decrease run-time relative to R-code. T-test not usually paired in COHCAP, so p-value will be different than for 1-variable test. May be useful if you need to speed up code and all measurements are paired.

For 'cppLmResidual.1var', use C++ code instead of R function for linear regression (t-test for residuals). Only valid for continuous analysis with 1 variable. WARNING: This code may less sensitive than normal lm() or ANOVA with smaller sample sizes (such as n=6).

For 'RcppArmadillo.fastLmPure', use 'fastLmPure' function within RcppArmadillo for linear regression, with R pt() t-distribution for p-value calculation. This can help decrease the run-time, relative to the lm() function that is used by default.

Can be 'none','rANOVA.1way','RcppArmadillo.fastLmPure', 'cppANOVA.1way', 'cppANOVA.2way','cppWelshTtest', or 'cppPairedTtest'.

`heatmap.dist.fun`
Distance metric for clustering in heatmap.

Can be 'Euclidian' or 'Pearson Dissimilarity'.

output.format	Format for output tables: 'xls' for Excel file, 'csv' for comma-separated file, or '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

List (island.list) with 2 data frames to be used for integration analysis:

beta.table = data frame of average beta (or percentage methylation) values across differentially methylated sites within a differentially methylated CpG island filtered.island.stats = differential methylation results for CpG Islands

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 = tempdir()#you may want to use getwd() or specify another folder
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")
island.list = 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,
max.cluster.dist = NULL,
output.format = "txt")
```

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 analyze continuous phenotypes using the Average-by-Site workflow.
num.sites	Minimum number of differentially methylated sites to define a differentially methylated CpG island.
max.cluster.dist	Update annotations by running de-novo clustering within each set of provided annotations. This is the maximum distance (in bp) between filtered sites with a consistent delta-beta sign. This can produce more than one cluster per pre-existing annotation. Set to NULL by default, to run standard COHCAP algorithm. If you would like to test this function, I would recommend a value between 50 and 500 bp. Only valid for 2-group or continuous comparison.
output.format	Format for output tables: 'xls' for Excel file, 'csv' for comma-separated file, or '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 = tempdir()#you may want to use getwd() or specify another folder
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

WARNING: This function was designed to work with genome_methylation_bismark2bedGraph_v3.pl .bed files (from an earlier version of Bismark) (so, it likely doesn't work with other file formats, such as the coverage files - please see COHCAP.BSSeq_V2.methyl.table() documentation)

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 = tempdir()#you may want to use getwd() or specify another folder
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.BSSeq_V2.methyl.table

Create Percent Methylation Table from Bismark Coverage files for Targeted BS-Seq data

---

Description

Creates percent methylation for COHCAP input (for COHCAP.annotate).

This function is not necessary for Illumina methylation array analysis.

NOTE: This does not create an annotation file. There is a script that can create an annotation file using the output of this function and an Ensembl .gtf (for hg19): *inst/extdata/Perl/downloaded_Ensembl_annotation.pl*

However, it is possible that the downloaded_Ensembl_annotation.pl script may need to be edited for other builds / organisms.

Usage

```
COHCAP.BSSeq_V2.methyl.table(cov.files, sampleIDs, percent.table.output,
min.cov = 10, min.percent.observed = 0.75,
chr.index = 1, pos.index = 2, percent.index = 4,
methylated.count.index = 5, unmethylated.count.index = 6,
read.gz=FALSE)
```

Arguments

`cov.files`	Array of Bismark coverage files created using 'bismark_methylation_extractor' (following Bismark alignment) Files can be compressed. However, to double-check files, please set read.gz to TRUE if using the original, compressed files.
`sampleIDs`	Array of sample IDs to be used as column heads in percent methylation table. Paired determined by order (which needs to be the same) in 'cov.files' and 'sampleIDs'
`percent.table.output`	Tab-delimited text output file for this function. This can be used as the input for COHCAP.annotate(), but an additional annotation file is also required for that function.
`min.cov`	Minimum coverage to list methylation at site for a given samples. Default is 10 reads. If site is present at this coverage in other samples, this value will be represented as an "NA"
`min.percent.observed`	Minimum percent of samples with at least 'min.cov' at a particular site. Default is 0.75 (75%)
`chr.index`	1-based index for chromosome in input file (element within 'cov.files' array). So, if this value is changed, function can potentially be applied to any table with all necessary values.
`pos.index`	1-based index for chromosome position in input file (element within 'cov.files' array). So, if this value is changed, function can potentially be applied to any table with all necessary values.
`percent.index`	1-based index for percent methylation in input file (element within 'cov.files' array). So, if this value is changed, function can potentially be applied to any table with all necessary values. Not strictly necessary if methylated and unmethylated counts also present, but this is currently required to be defined.
`methylated.count.index`	1-based index for methylated counts in input file (element within 'cov.files' array). So, if this value is changed, function can potentially be applied to any table with all necessary values.
`unmethylated.count.index`	1-based index for unmethylated counts in input file (element within 'cov.files' array). So, if this value is changed, function can potentially be applied to any table with all necessary values.

read.gz	Logical: are all files in 'cov.files' compressed?
	One value, so compression status must be the same for all files.
	May not be strictly necessary, but probably good to check input file type.

Value

This function creates a tab-delimited text containing methylation values for samples present in at least 'min.percent.observed' samples.

This table would be comparable to the beta input file for COHCAP.annotate().

The 1st column from this table could also be used to create an annotation file.

See Also

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

Raw Data for Demo Dataset (different samples within study described in COHCAP paper): https://www.ncbi.nlm.nih.gov https://www.ncbi.nlm.nih.gov/sra/SRR096438

BioPerl: https://bioperl.org/

Ensembl: https://ensembl.org

For hg19, there are steps to find and download the Ensembl .gtf are within: inst/extdata/Perl/downloaded_Ensembl_annot

Examples

```
library("COHCAP")

dir = system.file("extdata/BSSeq", package="COHCAP")
rep1 = file.path(dir,"SRR096437_truncated.bismark.cov")
rep2 = file.path(dir,"SRR096438_truncated.bismark.cov")

cov.files = c(rep1,rep2)
sampleIDs = c("MCF7.Rep1","MCF7.Rep2")
percent.table="BS_Seq_combined_V2.txt"

COHCAP.BSSeq_V2.methyl.table(cov.files, sampleIDs, percent.table, read.gz=FALSE, min.percent.observed = 0.45)
```

COHCAP.denovo *De-Novo Identification of Differentially Methylated CpG Site Clusters*

Description

Identifies differentially methylated CpG sites that form clusters.

Only works with 2-group comparison result. Does not map regions to genes or integrate with gene expression data.

Usage

```
COHCAP.denovo(site.table, project.name, project.folder,
min.sites = 4, max.dist = 500, output.format = "txt")
```

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.
min.sites	Minimum number of differentially methylated sites to define a differentially methylated CpG island.
max.dist	Maximum distance between ordered differentially methylated CpG sites to be considered in a cluster.
output.format	Format for output tables: 'xls' for Excel file, 'csv' for comma-separated file, or 'txt' for tab-delimited text file.

Value

Returns a table of clusters of differentially methylated sites, defined using criteria specified to COHCAP.site()

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 = tempdir()#you may want to use getwd() or specify another folder
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")
COHCAP.denovo(filtered.sites, project.name, project.folder)
```

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(island.list, 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 = "txt", ref="none")
```

Arguments

island.list	list with two data frames: island.list$filtered.island.stats = differential methylation results for CpG Islands island.list$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, 'csv' for comma-separated file, or 'txt' for tab-delimited text file
ref	Describes reference setting for upstream analysis. If creating scatterplots, checks if ref="continuous" (changing plots from discrete to continuous color scale)

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 = tempdir()#you may want to use getwd() or specify another folder
```

```
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")
island.list = COHCAP.avg.by.island(sample.file, filtered.sites,
beta.table, project.name, project.folder, ref="parental")
COHCAP.integrate.avg.by.island(island.list, 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, DEseq2, 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 = "txt")
```

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, 'csv' for comma-separated file, or '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 = tempdir()#you may want to use getwd() or specify another folder
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 = tempdir()#you may want to use getwd() or specify another folder
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.reformatFinalReport

Prepare COHCAP beta file from Illumina GenomeStudio Final Report

---

Description

Reformats FinalReport file into format for COHCAP.annotate() function.

Optionally allows user to re-name samples, using a table where the first column is the chipID and the 2nd column is the name that should appear in heatmaps, etc.

Usage

```
COHCAP.reformatFinalReport(FinalReport, beta.file, renaming.file=NULL,
detection.pvalue.cutoff=0.01)
```

Arguments

FinalReport	FinalReport from Illumina GenomeStudio. "AVG_Beta" and "Detection Pval" columns should be exported for "Sample Methylation Profile". The goal of this function is to create a single table with beta values, with the Detection P-value being used to remove measurements with low intensities.
beta.file	Table of beta / percentage methylation values. CpG sites are represented in rows. Samples are represented in columns. Samples with high detection p-values will be censored as NA values.
renaming.file	If you would like to provide more descriptive names for your samples, uprovide a table where the first column is the chipID and the 2nd column is the name that should appear in heatmaps, etc. Please make sure your new names don't begin with numbers, or contain certain special characters (such as dashes, spaces, etc) that will be reformated as periods in column names when imported into R.
detection.pvalue.cutoff	Maximum p-value to keep beta values in reformated table.

Value

Function does not return value. Instead, it writes the censored beta table for COHCAP input as a text file (specified as 'beta.file').

See Also

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

Examples

```
library("COHCAP")

dir = system.file("extdata", package="COHCAP")
FinalReport = file.path(dir,"EPIC_DEMO.txt")
beta.file = "COHCAP_beta.txt"
sample.relabel.file = file.path(dir,"EPIC_DEMO_samples.txt")

COHCAP.reformatFinalReport(FinalReport, beta.file, renaming.file=sample.relabel.file)

project.folder = tempdir()#you may want to use getwd() or specify another folder
project.name = "EPIC_FinalReport_test"
annotation.file = file.path(dir,"EPIC_DEMO_mapping.txt")
beta.table = COHCAP.annotate(beta.file, project.name, project.folder,
platform="custom",annotation.file=annotation.file,
output.format = "txt")
```

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, paired=FALSE,
delta.beta.cutoff = 0.2, pvalue.cutoff=0.05,
fdr.cutoff=0.05, ref="none", num.groups=2,
lower.cont.quantile=0, upper.cont.quantile=1,
create.wig = "avg", alt.pvalue="none",
plot.heatmap=TRUE, output.format = "txt", heatmap.dist.fun="Euclidian")
```

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 percent methylation. 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 percent methylation. 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. Set to "continuous" for a continuous primary variable. Otherwise, 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. Not used if "ref" is set to "continuous" (for linear-regression of a continuous variable).

`create.wig` Set to "avg" to create average beta (per group) and delta-beta values. Set to "sample" to create .wig files for each sample. Set to "avg.and.sample" to create average, delta-beta, and per-sample .wig files. Seto to "none" to avoid creating .wig files 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.

`plot.heatmap` Logical value: Should heatmap be created to visualize CpG site differential methylation? For best comparison to island heatmap, please use same parameters at site and island level.

`lower.cont.quantile`
For continuous analysis, what beta quantile should be the lower threshold for calculating delta-beta values? Default = 0 (minimum)

`upper.cont.quantile`
For continuous analysis, what beta quantile should be the upper threshold for calculating delta-beta values? Default = 1 (maximum)

`alt.pvalue` Use alternative strategies for p-value calculations.

Be careful that the workflow matches the p-value calculation.

For 'rANOVA.1way', use ANOVA (R function) instead of t-test (for 1-variable, 2-group comparison). Might be helpful when SD is 0.

For 'cppANOVA.1way', use ANOVA (C++ code) instead of t-test (for 1-variable, 2-group comparison). Helps decrease run-time relative to R-code.

For 'cppANOVA.2way', use C++ code instead of R function for t-test (for 2-variable, 2-group comparison). Helps decrease run-time relative to R-code, but p-value may be different. For this implementation, I require having at replicates for each interaction term (such as having replicate treatmentswith multiple backgrounds, cell lines, etc.). If each sample has exactly one pair (as may be the case with tumor-normal pairs), and you need to decrease the COHCAP run-time, you may consider using 'cppPairedTtest'.

For 'cppWelshTtest', use C++ code instead of R function for t-test (for 1-variable, 2-group comparison). Helps decrease run-time relative to R-code, but p-value will be different than t.test() function (C++ code assumes unequal variance between groups).

For 'cppPairedTtest', use C++ code instead of R function for t-test (for 2-variable, 2-group comparison). Helps decrease run-time relative to R-code. T-test not usually paired in COHCAP, so p-value will be different than for 1-variable test. May be useful if you need to speed up code and all measurements are paired.

For 'cppLmResidual.1var', use C++ code instead of R function for linear regression (t-test for residuals). Only valid for continuous analysis with 1 variable. WARNING: This code may less sensitive than normal lm() or ANOVA with smaller sample sizes (such as n=6).

For 'RcppArmadillo.fastLmPure', use 'fastLmPure' function within RcppArmadillo for linear regression, with R pt() t-distribution for p-value calculation. This can help decrease the run-time, relative to the lm() function that is used by default.

Can be 'none','rANOVA.1way','RcppArmadillo.fastLmPure', 'cppANOVA.1way', 'cppANOVA.2way','cppWelshTtest', or 'cppPairedTtest'.

`heatmap.dist.fun`
Distance metric for clustering in heatmap.

Can be 'Euclidian' or 'Pearson Dissimilarity'.

`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.

If you have a secondary continuous variable (like age), you can set paired to "continuous". COHCAP will then perform linear-regression analysis (converting primary categorical variable into continuous variable, if necessary)

`output.format` Format for output tables: 'xls' for Excel file, 'csv' for comma-separated file, or '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 = tempdir()#you may want to use getwd() or specify another folder
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

COHCAP.annotate, 2
COHCAP.avg.by.island, 3
COHCAP.avg.by.site, 6
COHCAP.BSSeq.preprocess, 8
COHCAP.BSSeq_V2.methyl.table, 9
COHCAP.denovo, 11
COHCAP.integrate.avg.by.island, 12
COHCAP.integrate.avg.by.site, 14
COHCAP.qc, 15
COHCAP.reformatFinalReport, 16
COHCAP.site, 17