COHCAP: City of Hope CpG Island Analysis Pipeline

Charles Warden

April 26, 2022

1 Introduction

COHCAP (City of Hope CpG Island Analysis Pipeline) is an algorithm 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.

COHCAP is currently available as a standalone program (click here). Here are potential advantages and disadvantages to using the standalone versus Biocondcutor COHCAP packages

Potential Advantages:

- The Bioconductor package no longer uses separate functions for Illumina array vs. Targeted BS-Seq data. This simplifies the number of parameters you need to understand in order to run COHCAP and simplifies software maintenance.

- For 450k and 27k array data, you do not need to provide an annotation file to COHCAP (but you can always specify custom annotations with the Bioconductor package, which is useful for EPIC array and BS-Seq data).

- COHCAP output files are now better organized into subfolders for easier interpretation.

- Additional functions (such as filter for delta-beta values, clustering of sites within regions, etc.) has been added in the Bioconductor package.

- The Bioconductor package doesn't require users to install Java

- The Bioconductor package doesn't require users to point Perl to their Rscript executable file

- The Bioconductor package provides a feature to automatically create a targeted BS-Seq annotation file (and creates necessary Bioconductor input file with Illumina array formatting)

Potential Disadvantages:

- The Bioconductor version uses a greater percentage of R code for functions, so it is slower than the standalone version.
- The Bioconductor pacakge doesn't contain a GUI.
- There are formatting differences between the two COHCAP packages, so standalone documentation does not completely apply to the Bioconductor package.

So, if you are comfortable with writing code in R, then you may prefer using the COHCAP Biocondcutor package. If you are not confortable with any programming (and you have a relatively small sample), you may prefer using the standalone version of COHCAP. Large patient cohorts will most likely need to be run on a powerful computer (such as a Linux cluster).

There is a separate discussion group for Bioconductor issues at on the Source-Forge page.

EPIC Array: EPIC array data can be analyzed using the "custom" CpG island mapping in the Bioconductor package.

In both cases, users should cite the following article when using COHCAP:

Charles D. Warden, Heehyoung Lee, Joshua D. Tompkins, Xiaojin Li, Charles Wang, Arthur D. Riggs, Hua Yu, Richard Jove, Yate-Ching Yuan. (2013) COHCAP: An Integrative Genomic Pipeline for Single-Nucleotide Resolution DNA Methylation Analysis. Nucleic Acids Research. 41 (11): e117

2 Data

Beta values from the Human Methylation 450k array and expression values from the Affymetrix Human Gene 1.0 ST Array. The DNA methylation data corresponds to GSE42308, the gene expression data corresponds to GSE42307. This example dataset contains two groups, each with 3 replicates. Fold-change, p-values, and FDR values for gene expression data were calculated as described in the Warden et al. 2013 NAR publication listed above. The dataset is significantly truncated for testing purposes.

```
> library("COHCAP")
> dir = system.file("extdata", package="COHCAP")
> beta.file = file.path(dir,"GSE42308_truncated.txt")
> sample.file = file.path(dir,"sample_GSE42308.txt")
```

```
> expression.file = file.path(dir,"expression-Average_by_Island_truncated.txt")
> project.folder = tempdir()
> project.name = "450k_avg_by_island_test"
```

The code for this example assumes all files should be created in the current
working directory. However, you can specify the input and output files in any
location (using the complete file path).

3 Data Annotation

To normalize DNA Methylation beta or percentage methylation values, run the
following function

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

[1] 173    7
[1] 172    5
[1] 172   11
```

The output is standard data frame with samples in rows and genes in columns,
which is also saved as an Excel file in the "Raw_Data" folder.

4 CpG Site Statistics

To display calculate CpG site statistics, filter for differentially methylated sites,
and/or create .wig files, run the following function

```
> filtered.sites = COHCAP.site(sample.file, beta.table, project.name,
+                                 project.folder, ref="parental")

[1] "Reading Sample Description File...."
[1] 172    6
[1] 172    6
[1] "Differential Methylation Stats for 2 Groups with Reference"
[1] 172    5
[1] 172   10
[1] 172   10
[1] 34 10
[1] 34 10
```

The filtered list of CpG sites are created in the "CpG_Site" folder, which is
also the data frame returned by the function. Statistics for all CpG sites are
located in the "Raw_Data" folder. If .wig files are created (default setting),
they can be found within the "CpG_Site/wig" folder (in the subfolder with the
corresponding project name).

5 CpG Island Analysis

To display calculate CpG island statistics, filter for differentially methylated islands, and/or create box-plots for differentially methylated islands, run the following function

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

[1] "Reading Sample Description File...."
[1] 172    6
[1] 172    6
[1] "Group: mutant"    "Group: parental"
[1] "Checking CpG Site Stats Table"
[1] 34 10
[1] 29 10
[1] 5
[1] "Average CpG Sites per CpG Island"
[1] "Differential Methylation Stats for 2 Groups with Reference"
[1] 4 8
[1] 4 8
[1] "There are 4 differentially methylated islands"
[1] 4 8
[1] 4 8
[1] "Plotting Significant Islands Box-Plots.."
```

The filtered list of CpG islands (with box-plots, if applicable) are created in the "CpG_Island" folder. Box-plots can only be created using the "Average by Island" workflow (shown above) Statistics for all CpG islands (meeting the cutoff for minimum number of CpG sites) are located in the "Raw_Data" folder.

   The function returns a data frame that can be used for integration with gene expression data. The format of data frame depends upon the workflow used for analysis. CpG island statistics will be provided by the "Average by Site" workflow (COHCAP.avg.by.site) whereas a table of beta values for differentially methylated islands will be provided by the "Average by Island" workflow (COHCAP.avg.by.island, shown above)

6 Integration with Gene Expression

To identify genes inverse expression changes (with scatterplots for visualization), please run the following function:

```
> COHCAP.integrate.avg.by.island(island.list, project.name, project.folder,
+                                 expression.file, sample.file)

[1] 4 8
[1] 4 8
```

```
[1] 4
[1] 1.777472e-06 1.706586e-06 2.418011e-06 2.014309e-07
[1] 4
[1] 4
[1] "4 significant correlations"
[1] "Plotting Correlated Genes...."
[1]  4 14
[1] "mutant"   "parental"
```

The function doesn't return any value and represents the last possible step in the COHCAP pipeline.

Integration can only be performed in the "Average by Site" workflow with a 2-group comparison. This results in two tables (Methylation Up, Expression Down and Methylation Down, Expression Up) in the "Integrate" folder. The gene expression file for the "Average by Site" workflow includes population-level fold-change, p-values, and FDR values (which must be calculated outside of COHCAP).

Integration can be performed using the "Average by Island" workflow with any number of groups. This results in a single list of genes with negative expression correlations in the "Integrate" as well as a folder for all correlation statistics in the "Raw_Data" folder. If desired, scatter plots will also be produced for genes with significant negative correlations between methylation and gene expression data. The gene expression file for the "Average by Island" workflow is a table of normalized intensity / expression values (can be from either microarray or RNA-Seq data).

7   Acknowledgements