COHCAP: City of Hope CpG Island Analysis Pipeline

Charles Warden

April 19, 2014

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.

- Common annotations are pre-installed with the Bioconductor package. For HM450k array, 27k array, and selected targeted BS-Seq data, you do not need to provide an annotation file to COHCAP.

- COHCAP output files are now better organized into subfolders for easier interpretation (compared to the standalone v1.2.4).

- A filter for delta-beta values has been added in the Bioconductor package (compared to the standalone v1.2.4).

- 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:

- Because COHCAP functions are now universal in the Bioconductor package, the COHCAP Bioconductor package doesn't filter Illumina array data based upon detection p-values. In most cases, lack of a detection p-value filter will probably not be a problem. However, very poor quality data will show a high proprotion of high detection p-values, so you may want to run additional QC analysis on your samples to ensure you are working with good quality data.
- The Bioconductor pacakge doesn't contain a GUI.
- There are minor formatting differences between the two COHCAP packages, so standalone documentation may 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).

Most of the information on the COHCAP FAQ page will apply to both versions. Questions for both versions should be directed to the COHCAP Discussion Group, with bioconductor users posting questions under "Bioconductor: [Problem Description]".

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 <- getwd()
> 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  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 Quality Control

To display QC metrics for DNA methylation data, run the following function

```
> COHCAP.qc(sample.file, beta.table, project.name, project.folder)

[1] 172   6
[1] 172   6
[1] "Group: mutant"   "Group: parental"
[1] "Color: red"  "Color: blue"
[1] "Calculating Sample Statistics..."
[1] "Creating Sample Histogram..."
[1] "Creating Dendrogram..."
[1] "Creating PCA Plot..."
null device
          1
```

The QC plots are created in the "QC" folder.

5 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] "Group: mutant"    "Group: parental"
[1] "Differential Methylation Stats for 2 Groups with Reference"
[1] 172    5
[1] 172   10
[1] 172   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).

6   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

```
> filtered.islands <- 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 7
[1] 4 7
[1] "There are 4 differentially methylated islands"
[1] 4 8
[1] 4 8
[1] "Plotting Significant Islands...."
```

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)

7 Integration with Gene Expression

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

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

[1] "4 significant correlations"
[1] "Plotting Correlated Genes...."
[1]  4 14
```

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 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).

8 Targeted BS-Seq Analysis

The method of reading BS-Seq data is notably different than the standalone version of COHCAP.

In the standalone version, users directly read .bed files created from the extended Bismark pipeline. Because the input file is standardized in the Bioconductor version of COHCAP, users must first produce a file meeting this new standard requirement.

Accordingly, a function unique to targeted BS-Seq analysis (COHCAP.BSSeq.preprocess) has been added to the Bioconductor package to faciliate this step.

Additionally, this function also creates a custom annotation file that is uniquely optimized for the user's own targeted sequencing data.

The old UCSC CpG Island file is still available on sourceforge (click here to download). This should technically work with any BS-Seq dataset. However, we strongly encourage users to use the file created from COHCAP.BSSeq.preprocess, especially since this step will likely be required to create the COHCAP input file.

Users should be warned that this step will likely take several hours to run. However, it luckily only has to be run once.

Detailed usage instructions can be found in the help description, called via help(COHCAP.BSSeq.preprocess).