illumına®

Data Analysis: DNA Methylation

An overview of data processing using the BaseSpace® Correlation Engine for DNA methylation analysis.

Introduction

The Illumina BaseSpace Correlation Engine contains biosets generated by mining the vast amounts of publicly available data and/or summary statistics through a systematic screening, curation, and data analysis process. This technical note provides an overview of the DNA methylation data processing pipeline (Figure 1).

Required Input Files

- Platform definition file: a list of array elements with chromosomal coordinates
- Sample annotation file: sample names with specific disease conditions, treatments, and/or cell line descriptions
- Sample measurement files: array elements with a relative measure of methylation. This measurement is calculated as the ratio of methylated probe signal over the total locus signal.

Data Pre-Processing

Platform Definition File

If necessary, probe chromosomal coordinates are converted to coordinates of the current build using the LiftOver tool.

Sample Annotation File

This file defines groups in all samples. Each group of samples shares annotation attributes, and is analyzed together to generate one bioset. All further analysis is done within each group.

Sample Measurement Files

All methylation data imported thus far are preprocessed data. The data have been screened for probes and samples that violate quality control indicators, and the data have been normalized. To ensure data quality, the BaseSpace Engine performs classical multidimensional scaling of the data matrix and removes obvious outliers. In addition, if detection p-values are supplied, samples containing a large percentage of reporters showing a lack of measurement are removed from the analysis.

Currently, the BaseSpace Engine processes data from bisulfite DNA methylation studies that use the Illumina Infinium® HD Assay and/or GoldenGate® Methylation Assay. It also processes data from affinity enrichment DNA methylation studies.

Analysis

For each probe, a linear model is fit to the intensity values across all samples using the R function lmFit. The estimated group means and standard errors are then used to compute a moderated t-statistic and p-value indicating the level of differential methylation signals between groups. These statistics are computed by empirical Bayes shrinkage of the standard errors toward a common value. This analysis is performed in R by the function eBayes. Both lmFit and eBayes are defined in the R package Limma.[1] This analysis pipeline is mindful of and guarded about the current challenges facing DNA methylation analysis.[2] For each probe, the percent differential between groups is calculated from the estimated group means returned by lmFit.
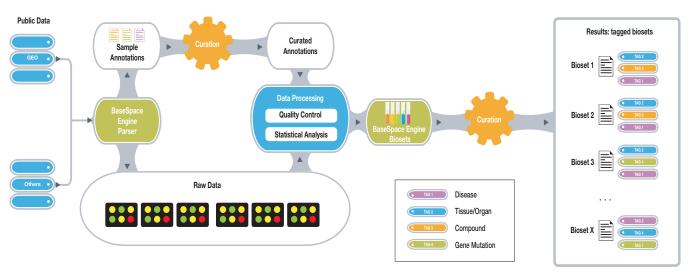

Figure 1: DNA Methylation Data Processing Pipeline—The data processing pipeline includes systematic screening, curation, and data analysis of publicly available data and/or summary statistics.

Data Post-Processing

Probes having p-values greater than 0.05, or absolute percent differential between groups less than 5, are removed from the reported bioset.

Platform probes are sorted by absolute percent DNA methylation differential between the 2 study groups.

Tags are assigned to each study.

References

1. Smyth GK. Limma: linear models for microarray data analysis. *Meth Mol Biol*. 2005;224:111-136.
2. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Genet Rev*. 2010;11:191-203.

For Research Use Only. Not for use in diagnostic procedures.