Methylation Analysis

Methylation Array Harmonization Workflow

Introduction

The Methylation Array Harmonization Workflow uses raw methylation array data from multiple generations of Illumina Infinium DNA methylation arrays, namely Human Methylation 27 (HM27), HumanMethylation 450 (HM450) and EPIC platforms, to measure the level of methylation at known CpG sites as beta values, calculated from array intensities (Level 2 data) as Beta = M/(M+U). This differs from the Methylation Liftover Pipeline in that the raw methylation array data is used instead of submitted methylation beta values, and the data is processed through the software package SeSAMe[1]. Additionally, the analysis results from the Methylation Array Harmonization Workflow are expected to be of higher quality than results from the Methylation Liftover Pipeline.

SeSAMe offers correction to detection failures that occur in other DNA methylation array software commonly due to germline and somatic deletions by utilizing a novel way to calculate the significance of detected signals in methylation arrays. By correcting for these artifacts as well as other improvements to DNA methylation data processing, SeSAMe improves upon detection calling and quality control of processed DNA methylation data. SeSAMe output files include: two Masked Methylation Array IDAT files, one for each color channel, that contains channel data from a raw methylation array after masking potential genotyping information; and a subsequent Methylation Beta Value TXT file derived from the two Masked Methylation Array IDAT files, that displays the calculated methylation beta value for CpG sites.

SeSAMe Methylation Beta Values File Format

Descriptions for fields present in GDC Harmonized Methylation Beta Values File are detailed below:

Field	Definition
Composite Element	A unique ID for the array probe associated with a CpG site
Beta Value	Represents the ratio between the methylated array intensity and total array intensity, falls between 0 (lower levels of methylation) and 1 (higher levels of methylation)

I/O	Entity	Format
Input	Raw Methylation Array	IDAT
Output	Masked Methylation Array	IDAT

I/O	Entity	Format
Output	Methylation Beta Values	TXT

Methylation Liftover Pipeline

Note: as of Data Release 32, Methylation Liftover files are no longer supported and do not appear in the GDC Data Portal.

Introduction

The DNA Methylation Liftover Pipeline uses data from the Illumina Infinium Human Methylation 27 (HM27) and HumanMethylation450 (HM450) arrays to measure the level of methylation at known CpG sites as beta values, calculated from array intensities (Level 2 data) as Beta = M/(M+U).

Using probe sequence information provided in the manufacturer's manifest, HM27 and HM450 probes were remapped to the GRCh38 reference genome [2]. Type II probes with a mapping quality of <10, or Type I probes for which the methylated and unmethylated probes map to different locations in the genome, and/or had a mapping quality of <10, had an entry of '*' for the 'chr' field, and '-1' for coordinates. These coordinates were then used to identify the associated transcripts from GENCODE v22, the associated CpG island (CGI), and the CpG sites' distance from each of these features. Multiple transcripts overlapping the target CpG were separated with semicolons. Beta values were inherited from existing TCGA Level 3 DNA methylation data (hg19-based) based on Probe IDs.

Methylation Liftover Pipeline Table Format

Field	Definition
Composite Element	A unique ID for the array probe associated with a CpG site
Beta Value	Represents the ratio between the methylated array intensity and total array intensity, falls between 0 (lower levels of methylation) and 1 (higher levels of methylation)
Chromosome	The chromosome in which the probe binding site is located
Start	The start of the CpG site on the chromosome
End	The end of the CpG site on the chromosome
Gene Symbol	The symbol for genes associated with the CpG site. Genes that fall within 1,500 bp upstream of the transcription start site (TSS) to the end of the gene body are used.
Gene Type	A general classification for each gene (e.g. protein coding, miRNA, pseudogene)

Field	Definition
Transcript ID	Ensembl transcript IDs for each transcript associated with the genes detailed above
Position to TSS	Distance in base pairs from the CpG site to each associated transcript's start site
CGI Coordinate	The start and end coordinates of the CpG island associated with the CpG site
Feature Type	The position of the CpG site in reference to the island: Island, N_Shore or S_Shore (0-2 kb upstream or downstream from CGI), or N_Shelf or S_Shelf (2-4 kbp upstream or downstream from CGI)

I/O	Entity	Format
Input	Submitted Methylation Beta Values	TXT
Output	Methylation Beta Values or Masked Methylation Array	TXT/IDAT

File Access and Availability

Type	Description	Format
Methylation Beta Value	A table that associates array probes with CpG sites and associated metadata.	TXT
Masked Methylation Array	A data file that contains channel data from a raw methylation array after masking of potential genotyping information.	IDAT

[1]. Zhou, Wanding, Triche Timothy J., Laird Peter W. and Shen Hui. "SeSAMe: Reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions." Nucleic Acids Research. (2018): doi: 10.1093/nar/gky691

[2]. Zhou, Wanding, Laird Peter L., and Hui Shen. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes." Nucleic Acids Research. (2016): doi: 10.1093/nar/gkw967

NIH... Turning Discovery Into Health ®

Version 1.0