UNIVERSITÄT BERN

Methyl-seq pipeline Documentation

20.05.2021 Irene Keller

SIB Swiss Institute of Bioinformatics

Preprocessing

UNIVERSITÄT BERN

We use the nf-core methylseq Bismark workflow for mapping the reads to the reference genome, deduplicate, extract methylation calls and produce a quality report.

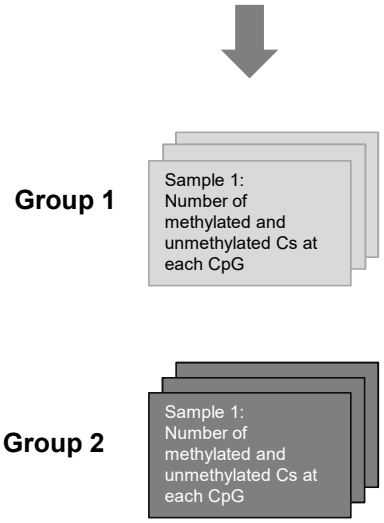
For more background: https://nf-co.re/methylseq/1.1

For details on tools and versions used: documentation/software_versions.csv in your results archive

Overview of analyses

UNIVERSITÄT BERN

nf-core methylseq
(see previous slide)



For each **pairwise comparison** among experimental groups, test for differential methylation (DM)

Individual CpGs

MethylKit
https://www.bioconductor.org/packages/release/bioc/html/methylKit.html

Annotation

Homer
http://homer.ucsd.edu/homer/ngs/annotation.html

De novo identification of DM regions

metilene
https://www.bioinf.uni-leipzig.de/Software/metilene/

dmrseq
http://www.bioconductor.org/packages/release/bioc/html/dmrseq.html

Create union

Homer
http://homer.ucsd.edu/homer/ngs/annotation.html

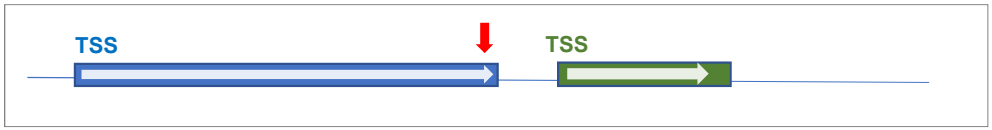
Annotation

UNIVERSITÄT BERN

Homer annotatePeaks.pl is used to annotate differentially methylated CpGs and regions. The annotation includes the following information:

Annotation	Location within the gene (e.g. exon, intron etc)
Refseq.Gene	RefSeq ID
external_gene_name	Gene symbol
ensembl_gene_id	Ensembl Gene ID
Detailed.Annotation	Similar to "Annotation" but with more detail in some cases
Distance.to.TSS	Distance to nearest transcription start site (negative values mean upstream of the TSS, positive values mean downstream)
Nearest.PromoterID	Various IDs and info on the gene
Entrez.ID	
Nearest.Unigene	
Nearest.Refseq	
Nearest.Ensembl	
Gene.Name	
Gene.Alias	
Gene.Description	
Gene.Type	

The fields shaded in blue refer to the gene with which the CpG or the **centre** of the DMR overlaps. The fields shaded in green, on the other hand, give information about the gene whose transcription start site (TSS) is closest to the CpG or centre of the DMR. Often, the blue and green fields will refer to the same gene but this is not always the case.
In the example below, the CpG of interest (indicated by red arrow) falls within an exon of the blue gene, and the blue fields in the table will contain information about this blue gene. The nearest TSS, however, is that of the green gene, and the green fields in the table refer to the green gene.



Often, a given position will overlap with multiple possible annotations. In this case, annotations are prioritised as explained in the Homer manual:
http://homer.ucsd.edu/homer/ngs/annotation.html

Overview of files in your zip archive

QC files (subfolder «qc»)

UNIVERSITÄT BERN

- **multiqc_report.html**: Summarises various stats from the nfcore pipeline, e.g. read quality, mapping rates etc.
- **NbrReadsPerCpG.pdf**: Histograms showing number of reads per CpG in each sample
- **PercentMethylationPerCpG.pdf**: Histograms showing % methylation for each CpG and sample. Here, we typically expect a bimodal distribution with two peaks at 0 and 100%

Clustering of samples on CpG methylation levels:

- **dendrogram.pdf**
- **pca.pdf**

Subfolder «CpGs»

UNIVERSITÄT BERN

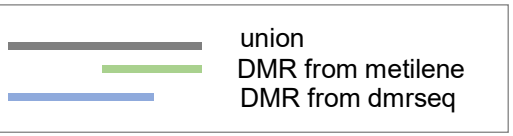
- CpG-level results from MethylKit, annotated with Homer
- One table with all significantly DM CpGs in a given pairwise contrast. A CpG is considered significant if the FDR-adjusted q-value is <0.05 and the difference in average methylation is at least 25%

- Annotation columns: refer to slide 4
- For each sample, there are three columns with the total number of reads (coverage), the number of reads with a C (numCs) and the number of reads with a T (numTs).
- Test results from methylKit are in the last 3 columns:
    - pvalue: Unadjusted for multiple testing
    - qvalue: False discovery rate adjusted value. This is the one to consider.
    - meth.diff: % methylation difference between the 2 groups

Subfolder «regions»

- Union of the differentially methylated regions from metilene and dmrseq. In each tool, a threshold of 5% (after FDR adjustment) is used to identify significant regions
- Annotation columns: refer to slide 4
- Remaining columns:

Example 1:



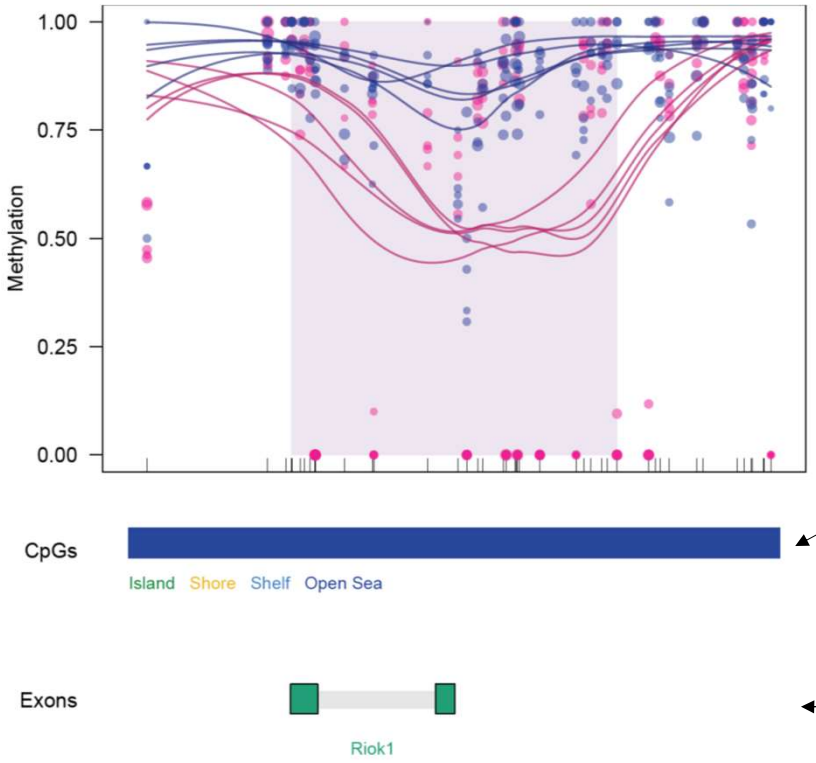
There will be 1 row in the table with the coordinates for these 3 intervals in the grey, green and blue fields, respectively.

Example 2:

There will be 2 row in the table, one for each metilene interval. The grey and blue intervals will be the same for both entries

seqnames start end	Chromosome, start and end positions of the **merged** interval
DMR from Metilene	
met_chr	Chromosome
met_start	Start
met_end	End
met_q_value	FDR-adjusted P-value
met_mean_diff	Mean methylation difference between groups
met_nbr_cpg	Number of CpGs in DMR
met_mean_group1	Mean methylation in group 1 (first in filename)
met_mean_group2	Mean methylation in group 2 (second in filename)
DMR from dmrseq	
dmr_chr	Chromosome
dmr_start	Start
dmr_end	End
dmr_width	Width of DMR
dmr_nbr_cpg	Number of CpGs in DMR
dmr_q_value	FDR-adjusted P-value
dmr_meth_diff	Mean methylation difference between groups

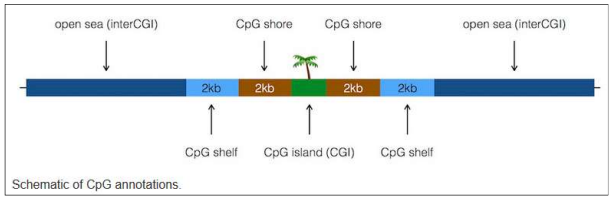
DMR plot



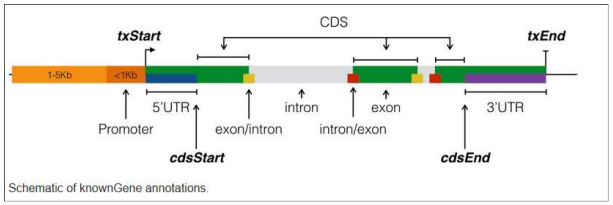
Each point represents one CpG in one sample, with a smoothed line for each sample. The position of CpGs is also indicated by the marks along the bottom edge of the plot. Experimental groups are indicated by colours. The size of the points is proportional to the coverage.

The region shaded in pink corresponds to the DMR.

Two annotation tracks are provided for the region:



http://bioconductor.org/packages/release/bioc/vignettes/annotatr/inst/doc/annotatr-vignette.html#cpg-annotations



http://bioconductor.org/packages/release/bioc/vignettes/annotatr/inst/doc/annotatr-vignette.html#genic-annotations