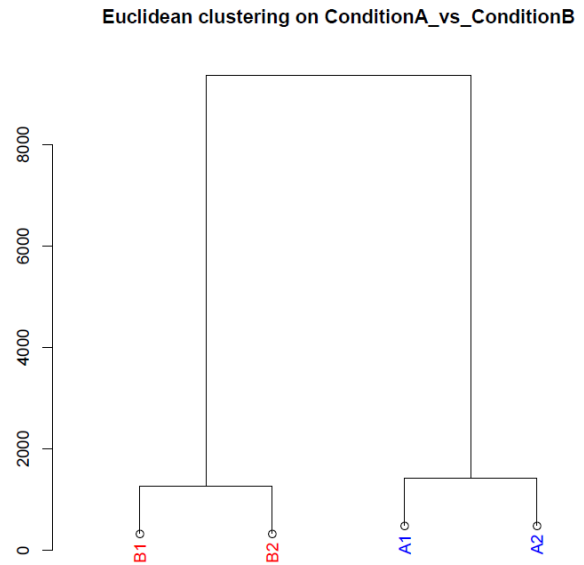
Descriptive analysis of RRBS samples

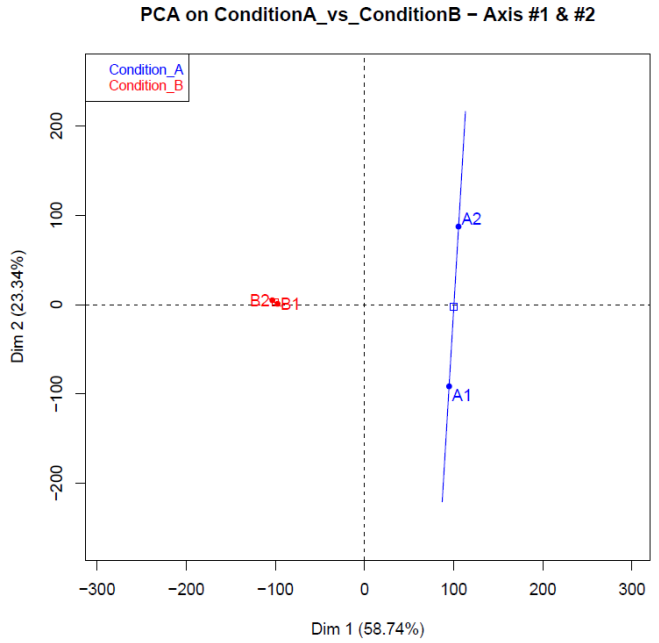
Descriptive analysis takes as input a set of files produced by the **Bismark_methylation_call** pipeline (one file for each samples).
Each one of these files contains , per CpG, the coverage and the methylation percentage observed.

It produces a unique table with CpG in lines and samples in columns (methylation percentage at the intersection) and performs a hierarchical clustering (HC) and a principal component analysis (PCA) on this table.
These two analyses aim at describing the relative similarities/differences between samples.

At the end of the process, user can obtain :

- the aggregate table of methylation percentage per CpG and samples
- a pdf document displaying results of HC and PCA.





Prerequisites and command used to launch the analysis

To run a descriptive analysis, you need :

- a set of files produced by **Bismark_methylation_call** pipeline (`<sample analysis directory>/extract/synthese_CpG.txt`)
- a configuration file explaining where to find the input files, which analyses to produce, …

Following slides will explain in detail how to setup this configuration file.

Once you have all these prerequisites satisfied, you simply launch your analysis :

```
RRBS_HOME/Descriptive_analysis/descriptive_analysis.sh      <pathname to your configuration file>
```

Configuration file

Configuration file is divided in two parts.
Part ① is dedicated to analysis parameters. All lines in this part must begin by character '#'. Parameter name and its value must be separated by a <TAB> character.

Part ② describe samples used in the analysis. The first line is the header containing the name of the columns, then each following lines describe each sample with : its name, the path to the file containing percentage of methylation and coverage for each CpG, the group to which belongs this sample, the colour used to identify samples of the same group (for a list of possible colour name, please refer to http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf)
On these lines, fields must be separated by a <TAB> character.

```
#Global parameters:
#----------------------
#title  ConditionA_vs_ConditionB
#output_dir     ./out
#
#min_samples_per_condition      2
#min_coverage   10
#max_coverage   500
#output_table_file      aggregated_methylation_results.txt
#produce_hclust Yes
#produce_pca    Yes
#pca_scaling    Yes
#sampling_factor        0.9
Sample  File    Condition       Colour
A1      ../Differential_analysis/analysis_examples/data/A1_syntheseCpG_chr1.txt Condition_A     blue
A2      ../Differential_analysis/analysis_examples/data/A2_syntheseCpG_chr1.txt Condition_A     blue
B1      ../Differential_analysis/analysis_examples/data/B1_syntheseCpG_chr1.txt Condition_B     red
B2      ../Differential_analysis/analysis_examples/data/B2_syntheseCpG_chr1.txt Condition_B     red
```

① ②

Analysis parameters

Parameter name	Comment
title	Used to give a name to your analysis. This name will appear in the title of the graphics and is used to name output files
output_dir	Path to the directory where the pdf (and possibly table of methylation) will be produced (Default : "." i.e. directory from which descriptive analysis is launched)
min_coverage and max_coverage	Minimal and maximal number of reads per CpG required to take a CpG into account for the analysis. For a given sample and a CpG, if the coverage does not satisfy above conditions, its methylation proportion is set to 'NA' (this will noticeably increase Principal Component Analysis - PCA - computation). (Default value : 10 for min and 500 for max)
min_samples_per_condition	Minimal number of samples per group satisfying above coverage conditions. (Default value : no minimum)
keep_NA	Whether to keep rows containing at least one NA value (Possible values : Yes/No - Default : Yes)
output_table_file	Name of the file used to save methylation level table. If no path is specified before the name, the file will be produced into directory specified in output_dir
produce_hclust	Flag used to indicate if user wants to produce hierarchical clustering. (Possible values : Yes/No - Default : Yes)
produce_pca	Flag used to indicate if user wants to produce principal component analysis. (Possible values : Yes/No - Default : Yes)
pca_scaling	Flag indicating if data should be scaled to unit variance before computation of PCA. See FactoMineR documentation. (Possible values : Yes/No - Default : Yes)
with_sample_labels	Flag indicating if sample labels should be displayed on PCA axes projection
sampling_factor	Proportion of the # of CpG satifying above conditions used to compute analysis This parameter is used to reduce execution time of the PCA (Value between 0 and 1 ; Default : 1)