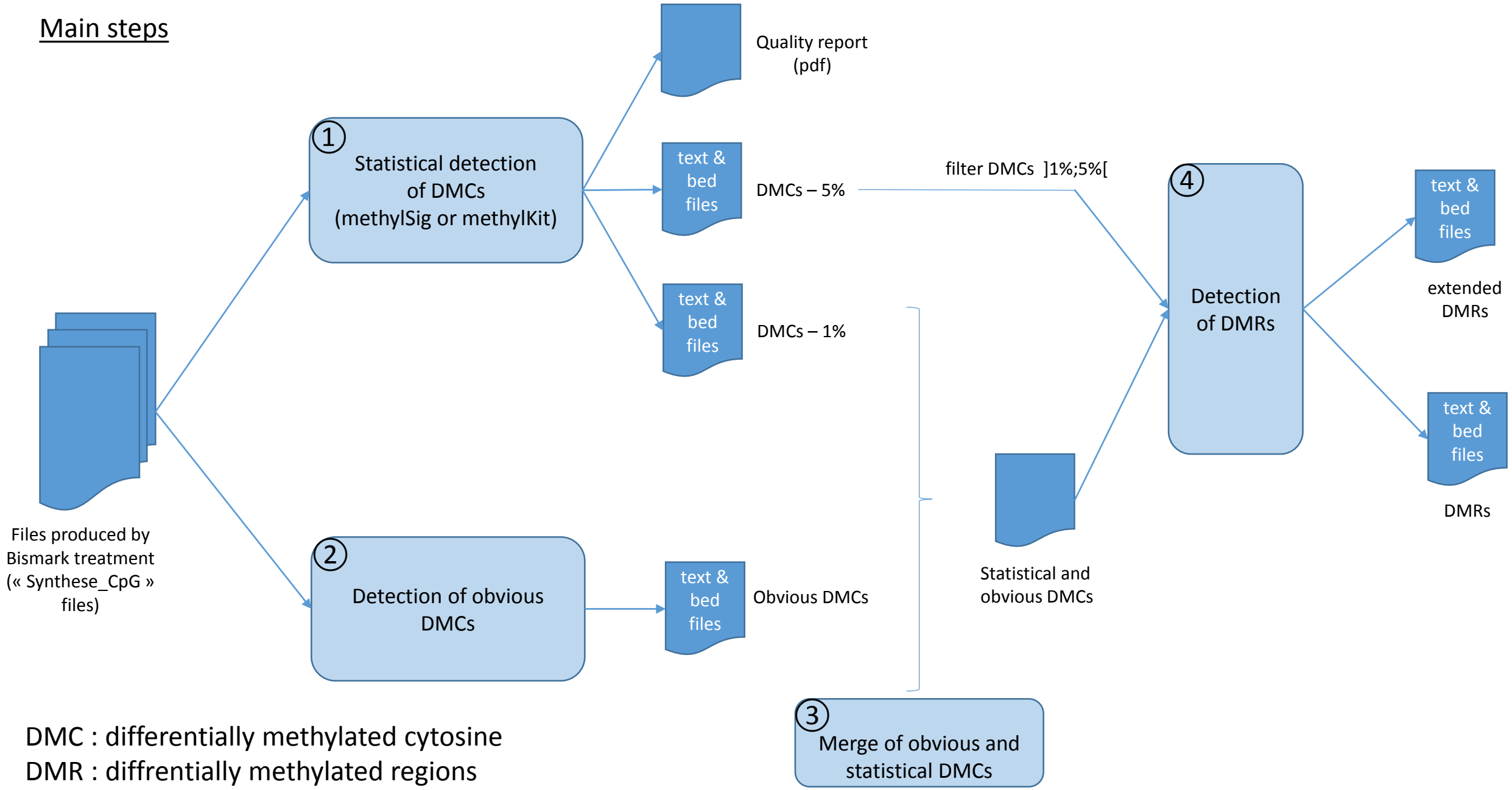
Analysis of DNA methylation differences between two groups of samples

Main steps



DMC : differentially methylated cytosine

DMR : diffrentially methylated regions

Figure comments

① : Detection of CpGs significantly differentially methylated between two conditions
This analysis requires to have replicates within group of each condition.
User can choose between analysis conducted by either methylSig or methylKit packages.
The analysis produces a quality report file and two {text;bed} files containing the significant DMCs for two p-value/q-value thresholds (default: <1% and <5%). Bed files can be imported in genome browers (e.g. IGV)

② : Detection of CpGs obviously differentially methylated between two conditions.
This analysis does not require replicates within the conditions compared.
If parameter **methdiff_threshold2** is set to 0.95, all samples in condition C1 must have a % of methylation >=95%,
while samples in condition C2 must have a % of methylation <=5% (idem with inversion of C1 and C2 conditions).

③ : DMCs statistically significants (statistical value: pValue or qValue<1%) detected in step ① are merged with obvious DMCs detected in step ②

④ : Adjacents DMCs collected in step ③ are grouped in DMRs (DMRs output file).
DMCs detected in step ① with a p-value/q-value ∈]1%;5%[and adjacent to DMRs are used to extend these DMRs (« extended DMRs » output file)

Requirements

Before to launch analysis one need to:

- have a set of files produced by bismark (synthese_CpG.txt), one for each replicate from each condition
- prepare a configuration file describing the parameters used for the analysis (e.g : analysis_config.txt)

To launch analysis, type following command :

```
RRBS_HOME/Scripts/Differential_analysis/get_methylation_differences.sh <relative or absolute path to config file>
```

Structure of analysis config file :

```
#Analysis parameters:
#output_dir         ./out
#title              Male vs Female
#...
#parameter_n        value
Sample  File    Condition
M1      M1/extract/synthese_CpG.txt      Male
M2      M2/extract/synthese_CpG.txt      Male
F1      F1/extract/synthese_CpG.txt      Female
F2      F2/extract/synthese_CpG.txt      Female
```

Part dedicated to analysis parametrization
(parameters and values are separated by one or several <TAB> character)

Definition of the two groups to compare and localization of analysis input files
(fileds in this table are separated by a <TAB> character)

'`output_dir`' and '`title`' parameters are the two global parameters needed for the analysis.
All other parameters are step specific and will be presented hereafter.

Several examples of config files and their corresponding analysis results are available in `Differential_analysis/analysis_examples` directory.

Explanations on configuration parameters

① Detection of DMCs

```
#MethylSig/Kit parameters:
#-------------------------
#stat_method                  methylKit
#min_coverage1                10
#max_coverage1                500
#min_per_group                2
#stat_value                   pvalue
#stat_threshold1              0.01
#methdiff_threshold1          0.25
```

`stat_method`:
Either `'methylSig'` or `'methylKit'` (no default value)

`min_coverage1`, **`max_coverage1`** and **`min_per_group`** :
Minimal and maximal coverage for a CpG to be taken into account for the analysis.
A minimal number of samples (**`min_per_group`**) in each group must satisfy this coverage range.
(default values : 10 for **`min_coverage1`**, no limit for **`max_coverage1`**, smallest group size for **`min_per_group`**).

`stat_value`:
Either `'pValue'` or `'qValue'` depending whether one wants to use the raw or the adjusted p-value for significant result selection.
(default value : qValue)

`stat_threshold1`:
Threshold used for significant result selection (used in conjunction with **`stat_value`** parameter) (default value : 0.01)

`methdiff_threshold1`:
Minimum value accepted for the absolute difference between average methylation in the two groups (default value : 0.25)

Explanations on configuration parameters

② Detection of obvious DMCs

```
#Obvious DMCs parameters:
#------------------------
#min_coverage2                10
#max_coverage2                500
#methdiff_threshold2          0.95
```

`min_coverage2, max_coverage2` :
Minimal and maximal coverage for a CpG to be taken into account for the analysis. All samples must satisfy these criteria.
(default value : 10 for min, 500 for max).

`methdiff_threshold2`:
Minimum value accepted for the methylation in all samples of one condition, while all samples in the other condition will have a maximal methylation of 1 - **`methdiff_threshold2`**
(default value : 1)

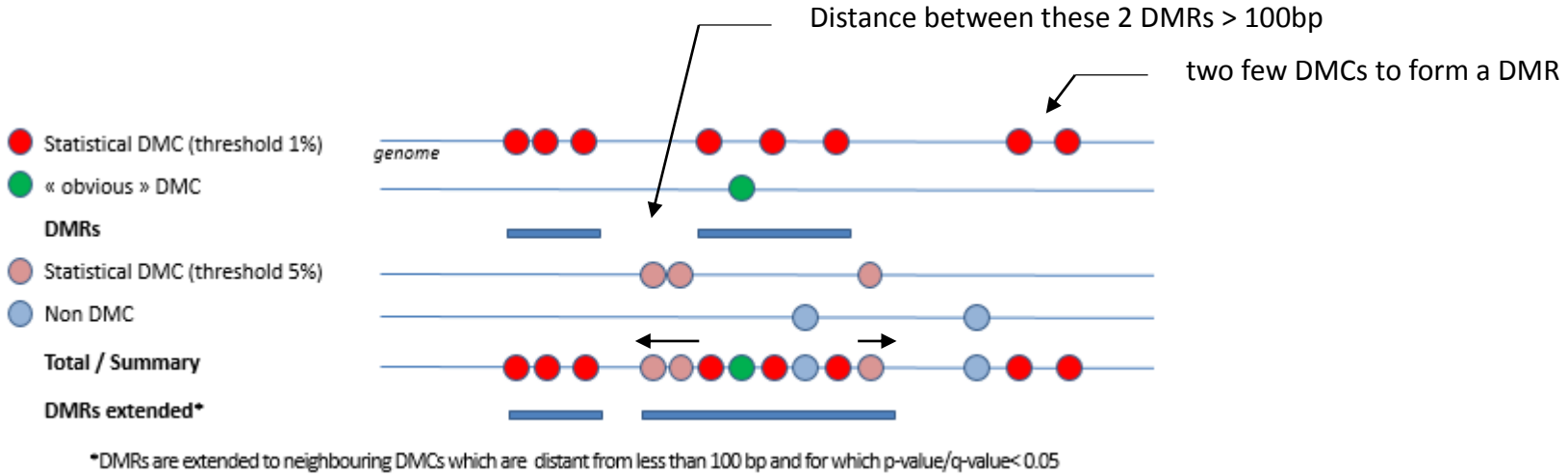
Explanations on configuration parameters

③ Merge of significant and obvious DMCs

No specific parameter is required for this step.

④ Detection of DMRs

```
#DMCs -> DMRs  parameters:
#-------------------------
#nb_min_DMCs_in_DMRs          3
#max_distance_between_DMCs    100
#stat_threshold2              0.05
```



`max_distance_between_DMCs`:
Maximum distance accepted between two DMCs to decide that they are neighbours (default value : 100)

`nb_min_DMCs_in_DMRs` :
Minimal number of neighbours DMCs needed to get a DMRs (default value : 3).

`stat_threshold2`:
Used to select DMCs detected in step ①. These DMCs will be used to try to extend DMRs.

Possible analyses

1°) More than one sample per condition

Please refer to :

`DMCs_config_methylKit.txt`

or :

`DMCs_config_methylSig.txt`

2°) Only one sample per condition

Please refer to :

`DMCs_config_nostat.txt`

3°) Generation of DMRs from a list of DMCs

Please refer to :

`DMCs_config_DMRonly.txt`

(these configuration files are available in `RRBS_HOME/Scripts/Differential_analysis/analysis_examples` directory)

Comparing analyses results

The script `RRBS_HOME/Scripts/Differential_analysis/venn_DMCs_sets.py` allows you to compare a list of DMCs.

Command used to launch the comparison :

```
python venn_DMCs_sets.py file_DMC_set1.txt file_DMC_set2.txt file_DMC_set3.txt > output_file
```

You can compare 2 sets or 3 sets of results.

Input file must have following format :

Chromosome	**Start**	**<any other columns>**	**Difference in methylation**	**Methylation state**

Second to last column (**Difference in methylation**) should be filled with a value corresponding to difference in methylation between the two conditions compared.

Example :

Chromosome	**Start**	**<any other columns>**	**Difference in methylation**	**Methylation state**
1	123456	….	-22.345	hypometh
1	345678	….	50.3461	hypermeth

Output file must format (example for a comparison of two files (A and B) :

difference in methylation found in A and B

Chromosome	**Start**	**End**	**Only in A**	**Only in B**	**Common A B**	**A**	**B**
1	123456	123457	*			-22.345	
1	234567	234568		*			44.8277
1	345678	345679			*	50.3461	50.1234