# Analysis of Bisulfite Sequencing PCR

# User guide

Version 1.0.0

Marie Denoulet and Chann Lagadec

May 7, 2022

# Contents

# List of Figures



# List of Code Listings

# 1 General information

## 1.1 What ABSP does do?

ABSP, standing for "Analysis of Bisulfite Sequencing PCR", is an R based tool to analyze methylation profiles using data from Bisulfite Sequencing PCR (BSP) experiment results. It was developed to help researchers to estimate and compare methylation percentages of a DNA region studied using BSP experiments. It provides a complete automated workflow, from trace file sequencing results to data visualization and statistics.

## 1.2 What are the advantages of using ABSP?

- A complete workflow. ABSP uses as input the chromatogram trace files as the sequencing results, and through a two-steps analysis, it (1) computes the methylation percentages of individual samples after validating the sequencing quality and (2) gathers the methylation levels from all samples to summarize methylation data, generate publication-ready figures and perform comparative statistics to answer to the experiment hypothesis on the DNA methylation differences between conditions.
- A fully automated process. ABSP uses a shiny app on R to provide a user-friendly interface. To launch the analytic process the user is guided to provide the required inputs and can launch the desired analysis with one click. For each analysis, an HTML report file is generated to visualize the results and keep a record of them. Additionally, output files, such as tables and figures, are automatically saved in the corresponding result folders. For an even more automated use, several analyses can be launched with the help of pre-filled input tables (spreadsheet document to fill) in the special tab "Multiple analyses", which is useful for large amounts of samples.
- Analyses of both direct-BSP and cloning-BSP sequencing data. ABSP can analyze results from both BSP methods. No existing tool is currently able to analyze both. It allows continuity in the experiment analytic process, as the direct-BSP approach can be performed before cloning the PCR products to have preliminary insights on DNA methylation, and then further confirmed/validated using cloning-BSP.
- Accessible and flexible. ABSP is coded using R, a crossed-platform tool language increasingly used in biology research, making it very accessible to any researchers. Additionally, for researchers accustomed to R coding, as the entire scripts are provided, ABSP is fully upgradeable. Also, we provide specific guidelines to easily modify some features to adapt ABSP to experiment needs, as adjusting quality thresholds or changing graphical parameters (see section 3.3 Code modifications at page 36).

## 1.3 Bisulfite Sequencing PCR

The Bisulfite Sequencing PCR (BSP) is an experimental technique aiming to estimate methylation levels of CpG sites on a specific DNA region of interest, among a population of DNA molecules. The method was originally developed by Frommer *et al.* in 1992[1] and Clark *et al.* in 1994[2] and was named BSP in opposition to the methylation specific PCR (MSP) method by Li *et al.* in 2002.[3]

This method is composed of three steps, described in figure 1:

1. A DNA bisulfite conversion
2. A PCR amplification and an optional cloning
3. A sequencing of either PCR products or individual subclones

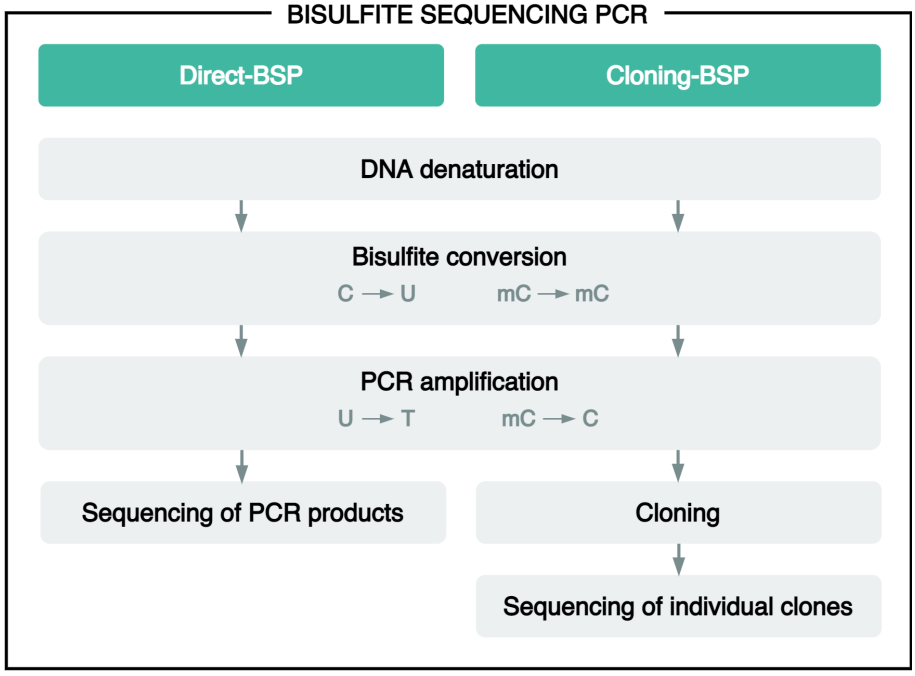


Figure 1. Bisulfite Sequencing PCR experimental principle.

Two approaches of BSP could be used (figure 1). The *direct-BSP* method is characterized by the direct sequencing of PCR products, whereas the *cloning-BSP* consists in cloning PCR products within a specific vector, use it to transform and select bacteria, select several clones, and sequence the individual clones.

[1] M Frommer et al. "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." In: *Proceedings of the National Academy of Sciences* 89.5 (1992), pp. 1827–1831. ISSN: 0027-8424. DOI: 10.1073/pnas.89.5.1827.

[2] S J Clark et al. "High sensitivity mapping of methylated cytosines." In: *Nucleic acids research* 22.15 (1994), pp. 2990–7. ISSN: 0305-1048. DOI: 10.1093/nar/22.15.2990.

[3] Long-Cheng Li and Rajvir Dahiya. "MethPrimer: designing primers for methylation PCRs". In: *Bioinformatics* 18.11 (2002), pp. 1427–1431. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/18.11.1427.

As described in figure 2, the analysis strategies of these two sub-methods are different on some points.

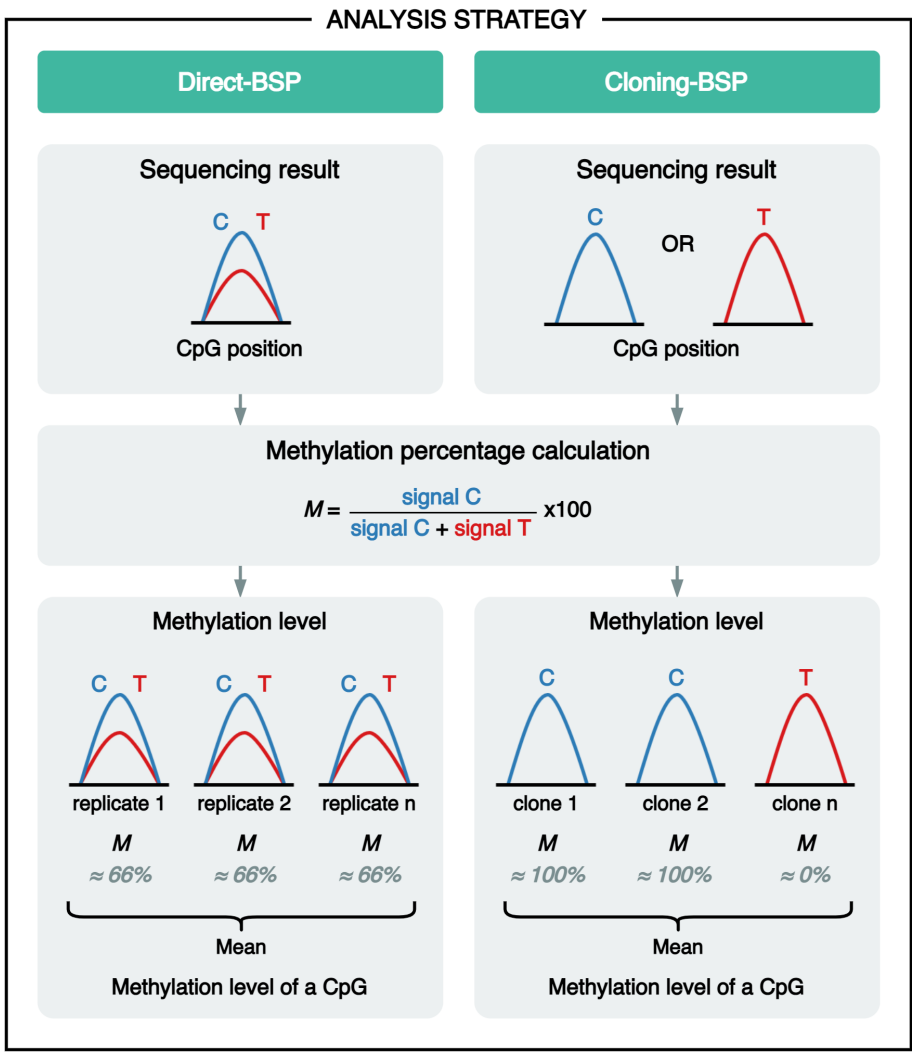


Figure 2. Bisulfite Sequencing PCR analysis strategies for both direct-BSP and cloning-BSP.

First, in **direct-BSP**, mix of DNA molecules with different unknown methylation statuses are sequenced, thereby, at each CpG site, two base signals can co-exist: the methylated signal (C) and the unmethylated signal (T). By calculating the signal ratio, the methylation level of a CpG in the DNA population can directly be estimated, but for reproducibility and statistical significance purposes, it still needs to be repeated in several biological replicates to obtain the final methylation level of a CpG (direct-BSP results are considered less quantitative than cloning-BSP ones).

Secondly, in **cloning-BSP**, as each clone represents only one PCR product, the methylation status of a CpG can be either methylated (C) or unmethylated (T),

thereby only one of these two signals can exist. Accordingly, the signal ratio can only give either 0% or 100% of methylation (partial methylation is considered biased results), revealing the methylation status. For each CpG, the proportion of clones with a methylated status reveals the methylation level of the CpG in the original DNA population, in general, a minimum of 10 clones is recommended to have a 10% accuracy of the methylation level.

## 1.4 How does ABSP work?

As a first step, each BSP sequencing result is defined by a combination of experiment information (figure 3):

- **Sequence.** The sequence identifier refers to a unique amplicon sequence produced by the BSP experiment, using a unique set of primer. For example, if several regions of a gene are analyzed by BSP, as each region corresponds to a unique amplicon they must have distinct sequence name (e.g. *CDH1promoter* and *CDH1exon1* ; *CDH1-1*, *CDH1-2*, and *CDH1-3*). Make sure the sequence name is strictly identical for all samples of the same sequence.
- **Collection.** The collection corresponds to a separation of samples above groups. Samples from different collections can not be compared, even if they belong to the same group. For example, collections can be different cell lines, organs, or patients, in which the same groups are compared but not between the different collections. To compare these types of samples, consider them as groups. Make sure the collection name is strictly identical for all samples of the same collection.
- **Group.** The group corresponds to the condition that will be compared with other groups/conditions in the grouped analysis. For example, groups can be the "control" and "treated" conditions. Make sure the group name is strictly identical for all samples of the same group.
- **Replicate.** Information has to be provided only when using the direct-BSP approach. The replicate number refers to the repetition identifier number of the sequencing. In order to have robust and reproducible data and to perform comparative statistics, each sample needs to be sequenced at least three times (in both directions).
- **Clones.** Information has to be provided only when using the cloning-BSP approach. In order to estimate the methylation levels among the DNA population, the methylation statuses of several individual clones needs to be sequenced. The ratio of methylated and unmethylated clones for a CpG position will gives the methylation level estimation.
- **Sequencing files.** The sequencing of each sample is performed in both directions, using a forward primer and a reverse primer.
- **Experiment.** The term "experiment (data)" refers to the unique combination

of collection, group and replicate or clone, for a specific sequence, corresponding to the sample information for the sequencing.

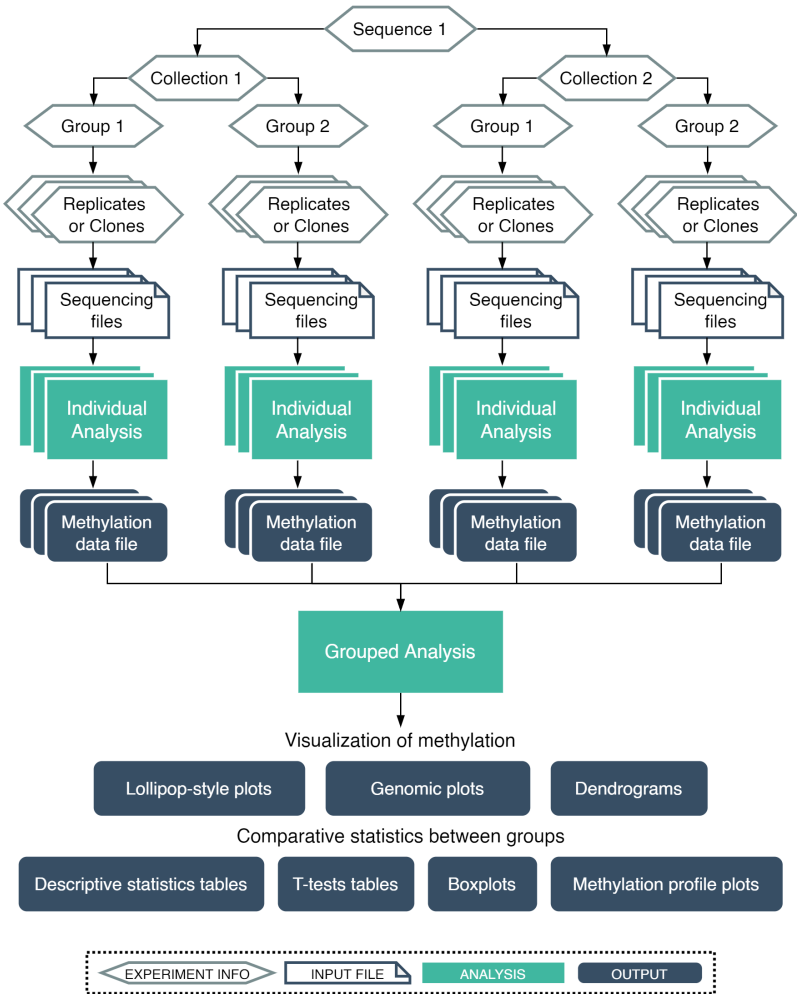


Figure 3. Workflow of the ABSP analytic process.

Then, the ABSP analysis is divided into two steps, corresponding to two scripts (R markdown scripts), as illustrated in the workflow in figure 3.

- Individual analysis. For each individual experiment point, two sequencing .ab1 files (one from forward direction, one from reverse direction) are used as input for the individual analysis. First, the sequencing results are trimmed based on quality to get the correct sequence for alignments with the reference DNA input. Then, results from the alignments go through a quality control step to check for mismatches, gaps, length of aligned sequences, and bisulfite conversion rates (calculated on cytosines outside CpG that should be thymines). If the results are defined as correct, the methylation levels of CpG can be calculated and visualized on a genomic plot. Several output files (e.g. chromatograms, sequences, tables) are saved in folders, especially the methylation data file as a result of the individual sequencing experiment

analyzed.

- Grouped analysis. All methylation data files from the same sequence are gathered by the grouped analysis. First, a preprocessing step is performed to organize data. Then, visualization plots, lollipop-style plots, and genomic plots (with associated clustering dendrograms) are generated to view methylation data differences. Finally, a statistical analysis is performed, descriptive statistics tables and Student's t-test p-values tables are generated, as well as boxplots with t-test p-values and methylation profile plots with Kruskal–Wallis test p-values, to display the significant methylation differences.

For more details, the inputs, processes, and outputs of these two steps, individual analysis, and grouped analysis, are displayed in the figure 4, where the 3 tabs of ABSP are represented: "**Individual analysis**", "**Grouped analysis**", and "**Multiple analyses**".

The two steps, individual analysis and grouped analysis, require the manual entry of input data in the corresponding tab, and therefore only one analysis can be launched at the same time (figure 5). An additional tab has been implemented to launch multiple analyses all at once. This multiple analyses tab can be used to launch either several individual analyses and/or several grouped analyses at the same time, by using tables (*.xlsx* files or *.csv* files) as input instead of the manual entry of input for unique analysis.

## 1.5 License

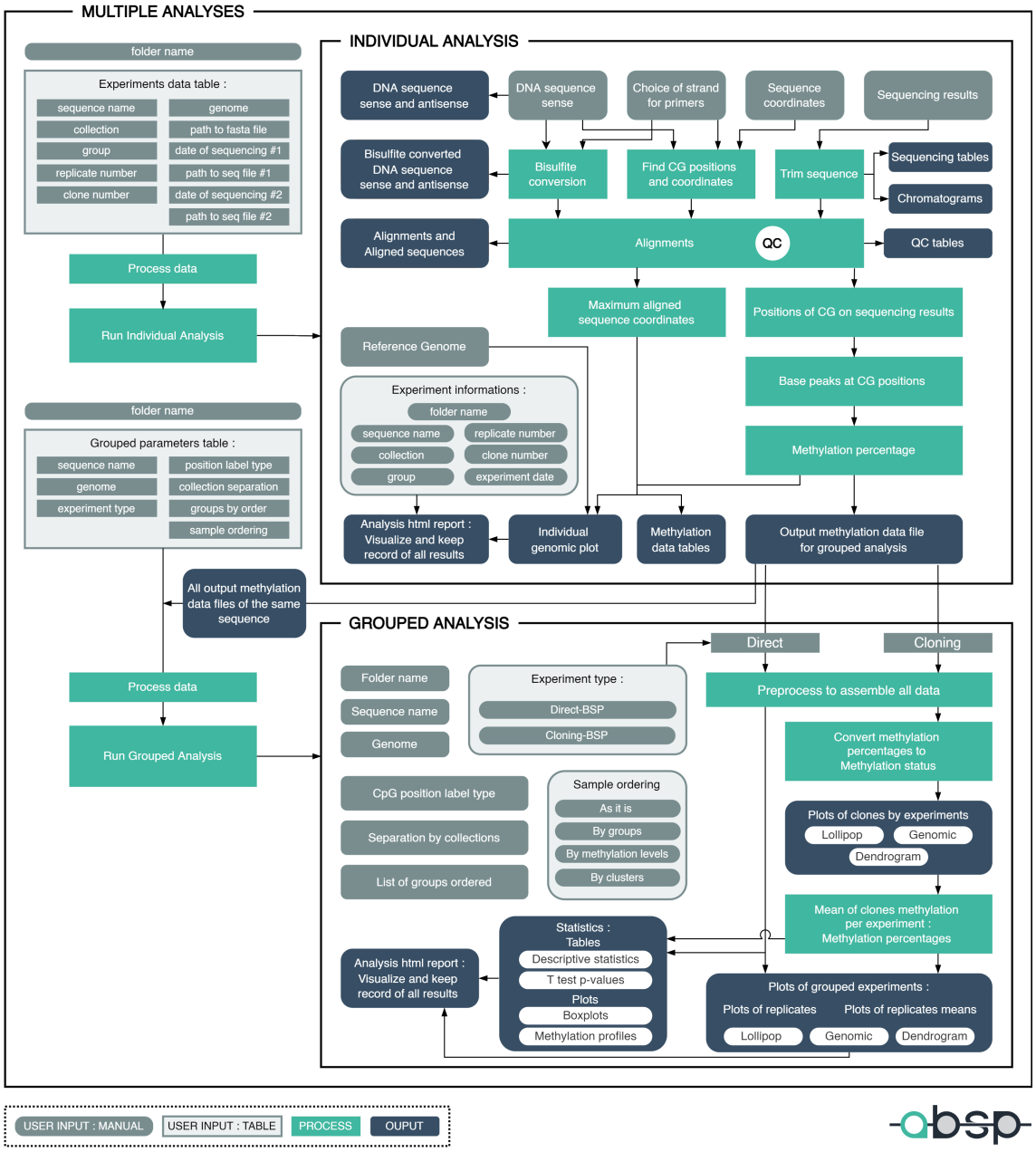


Figure 4. Detailed workflow of the ABSP analytic process.

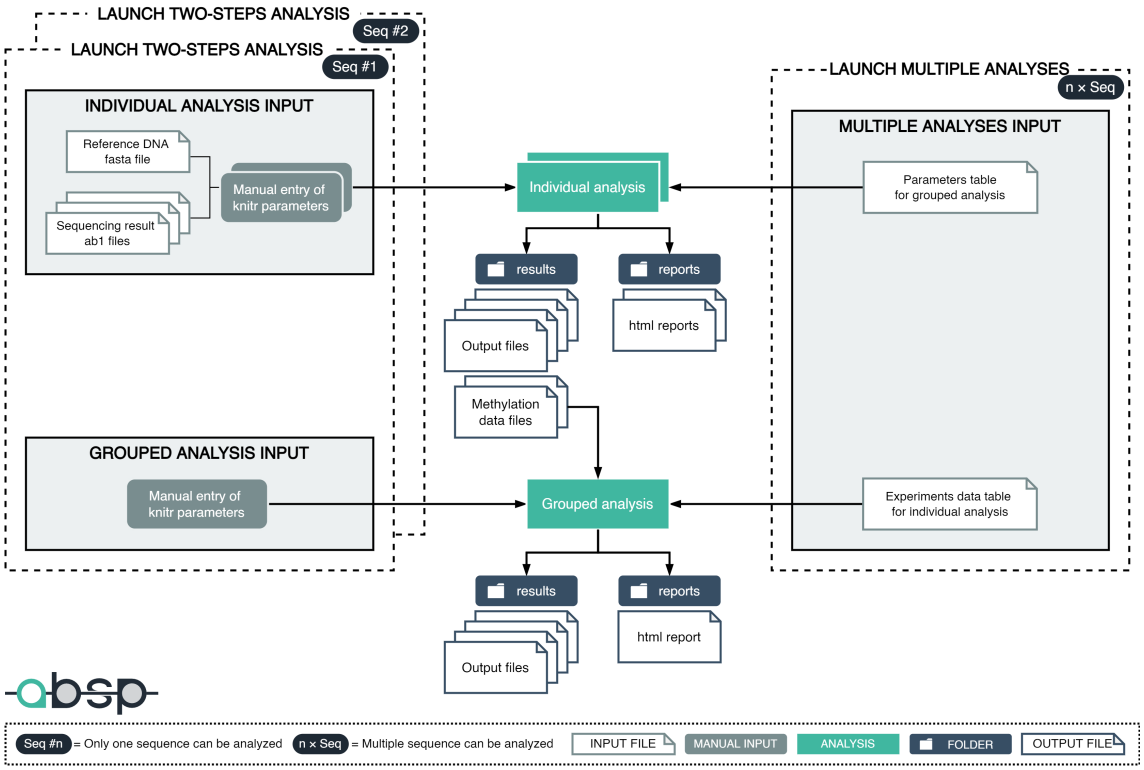


Figure 5. Diagram of the possible ways to launch ABSP analyses.

# 2 How to proceed with analysis using ABSP?

## 2.1 Open ABSP for the first time

### 2.1.1 Download and installations

- Install R: https://www.r-project.org/
- Install RStudio: https://www.rstudio.com/
- Download the ABSP files on github: https://github.com/ABSP-methylation-tool/ABSP

### 2.1.2 Content of files

The main ABSP folder content is:

- documents *folder for documents available to the user*
  - ABSP User Guide.pdf *reference manual to use ABSP*
  - List of BSgenomes.xlsx *file listing the available genomes*
  - multiple_grouped_parameters_table.xlsx *table of inputs to launch multiple grouped analyses*
  - multiple_grouped_parameters_table.ods *table of inputs to launch multiple grouped analyses*
  - multiple_individual_analyses_table.xlsx *table of inputs to launch multiple individual analyses*
  - multiple_individual_analyses_table.ods *table of inputs to launch multiple individual analyses*
- renv *folder for R environment and packages*
  - library *folder for the private project library*
  - activate.R *activation script run by the project .Rprofile*
- reports *folder for analysis reports*
- results *folder for analysis results (data, tables, graphics...)*
- scripts *folder for scripts and associated files required to run analysis*
  - ABSP_functions.R *R script providing the necessary functions for ABSP*
  - ABSP_grouped_analysis.RMD *R markdown script of grouped analysis*
  - ABSP_individual_analysis.RMD *R markdown script of individual analysis*
  - custom.css *CSS script for custom theme settings of the .html report files*
  - logo.svg *ABSP logo vector image*

- www *folder for files necessary for the shiny app*
  - ABSP - Analysis.svg *diagram of ABSP analysis strategy*
  - ABSP - BSP.svg *diagram of BSP experiment principle*
  - ABSP - fasta file.png *image of a reference sequence .fasta file example*
  - ABSP - Launch analysis.svg *diagram of the different ways to launch ABSP analyses*
  - ABSP - Workflow simple.svg *diagram of the BSP workflow*
  - custom_app.css *CSS script for custom theme settings of the app interface*
  - logo.svg *ABSP logo vector image*
- ABSP RProject.Rproj *R project file*
- app.R *shiny app file*
- renv.lock *lockfile describing the state of the project's library at some point in time*

For ABSP to function properly, all the aforementioned files must be downloaded and present in the ABSP main folder with the same structure.

Make sure not to rename, move or delete the provided folders and files. If you want to reorganize files or folders it is better to copy to other directories than to modify the files. However, new folders can be added to the ABSP main folder without causing issues.

### 2.1.3 Launch the app

- Open the *ABSP Rproject.Rproj* file with RStudio.
- Open the *app.R* file with RStudio.
- Find the *"Run App"* button in the upper right corner, click on the arrow right next to it and select *"Run external"* (for a better display).
- Click on the *"Run App"* button to launch the app.
- A pop-up window should appear if the shiny package was not already installed on your device, click on *"Yes"* to accept the Shiny installation.

Once this procedure is done, the package installation can start and it might take a few minutes until the app can be opened in the default web browser.

## 2.2 Open ABSP (not for the first time)

### 2.2.1 Launch the app

- Open the *ABSP Rproject.Rproj* file with RStudio.
- Open the *app.R* file with RStudio.
- Click on the *"Run App"* button to launch the app.

To launch the different analyses refer to the following sections below which describes the individual analysis, grouped analysis and multiples analyses (figure 5).

## 2.3 Individual analysis

### 2.3.1 Input files requirements

Sequencing result *.ab1* files As input, ABSP requires the chromatogram trace file (*.ab1*) from the sequencing run (Sanger) using the bisulfite converted DNA PCR products as templates. It is highly recommended to have both directions sequenced, by a primer on each side of the PCR product: a forward primer and a reverse primer. Although, the analysis can be run using only one trace file instead of both. The directions do not need to be specified as the analysis will determine it automatically.

Reference DNA sequence and information *.fasta* file ABSP also requires a *.fasta* file containing information about the reference DNA. A *.fasta* file is composed of a **header** and a **body**. The header must contain both the **genomic coordinates** of this sequence and the **choice of strand**, the one used for primer design, the one that will be amplified by PCR. Indeed, as both strands are no longer complementary after bisulfite conversion, the primers have to be designed on only one bisulfite converted strand as template DNA. The body must contain the **nucleotide sequence** of the reference DNA from the **plus strand** (upper/sense strand) of the genome.

Formats for the *.fasta* file header:

- The **genomic coordinates** must be written in the format *chr#:######-######* (e.g. *chr16:68771087-68771462*).
- The **choice of strand amplified** must be either *"primers=plus"* or *"primers=minus"*. If none of these characters strings are present in the *.fasta* file, by default the plus strand is chosen.

An example of a *.fasta* file content is depicted in figure 6. Note that any other information in the header, such as the sequence name, for example, can be added without consequences if they do not interfere with the previously described formats.

```
> CDH1 chr16:68770900-68771299 primers=plus
CCAAGTGTAAAAGCCCTTTCTGATCCCAGGTCTTAGTGAGCCACCGGCGGGGCTGGGATTCGAACCCAGTGGAATCAGAAC
CGTGCAGGTCCCATAACCCACCTAGACCCTAGCAACTCCAGGCTAGAGGGTCACCGCGTCTATGCGAGGCCGGGTGGGCGG
GCCGTCAGCTCCGCCCTGGGGAGGGGTCCGCGCTGCTGATTGGCTGTGGCCGGCAGGTGAACCCTCAGCCAATCAGCGGTA
CGGGGGGCGGTGCCTCCGGGGCTCACCTGGCTGCAGCCACGCACCCCCTCTCAGTGGCGTCGGAACTGCAAAGCACCTGTG
AGCTTGCGGAAGTCAGTTCAGACTCCAGCCCGCTCCAGCCCGGCCCGACCCGACCGCACCCGGCGCCTGCCCTCGC
```

Figure 6. Example of *.fasta* file for the reference DNA input required for ABSP individual analysis.

As a recommendation, to help with the creation of the *.fasta* file, the IGV (Integrative Genomics Viewer)[4] software can be easily used to navigate the genome and to get the nucleotide sequence of a specific region, alongside with its genomic coordinates. After navigating on the genome, the current viewed region can be added as a region of interest by selecting the *"Regions"* > *"Region Navigator"* > *"Add"* (figure 7). In this *"Regions of Interest"* panel, the added region will appear in the list of regions, its coordinates can be adjusted and it can be annotated with a description. By right-clicking on a region from this list, a context menu appears and two options can be selected, *"Copy Sequence"* or *"Copy Details"*. The first one copies to the clipboard the nucleotide sequence of the region, and the second one copies the genomic coordinates (in the correct format for the *.fasta* file) as well as the description associated with the region. Make sure to properly verify the first and last nucleotides, as there can be a one nucleotide difference between the coordinates and the actual sequence.

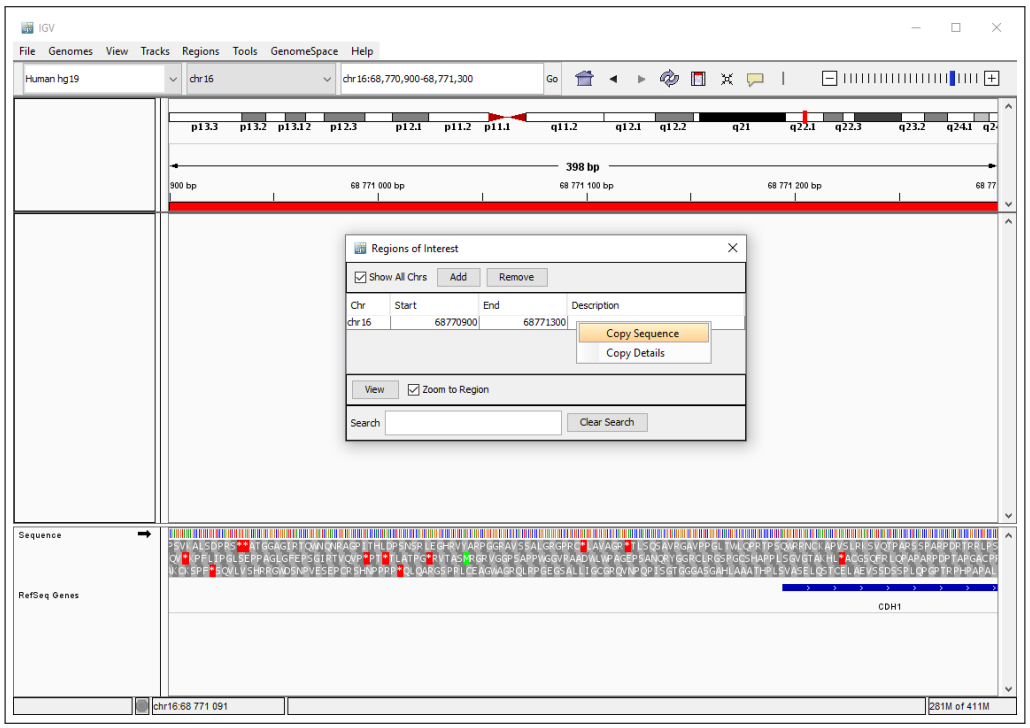


Figure 7. IGV (Integrative Genomics Viewer) software window.

### 2.3.2 Procedure

In the **individual analysis tab**, the left panel is to launch analysis, and the right panel provides entries information.

Experiment information

[4] https://software.broadinstitute.org/software/igv/

- **Select an existing folder or enter a new folder name.** Located in the *ABSP/results* folder, all of the analysis results will be saved in this folder. Having different folders of results can be used to separate the different analyses by projects, experiments, or users. Note that the six first letters of the folder name will appear in the report file name.
- **Select an existing sequence folder or enter a new sequence name.** The sequence folder is located in the previously selected folder. All of the analysis results will be saved in this folder corresponding to the same sequence. Note that this sequence name will be used in output files (tables, plots...) to refer to the sequence.
- **Enter the collection name.** The collection corresponds to a separation of samples above groups *(details in section 1.4 How does ABSP work? at page 7)*. Make sure the collection name is strictly identical for all samples of the same collection.
- **Enter the group name.** The group corresponds to the condition to compare *(details in section 1.4 How does ABSP work? at page 7)*. Make sure the group name is strictly identical for all samples of the same group.
- **For direct-BSP only: Enter replicate number.** In the case of direct sequencing of PCR products only, the replicate number corresponds to the repetition identifier of the sample *(details in section 1.4 How does ABSP work? at page 7)*.
- **For cloning-BSP only: Enter clone number.** In the case of clone sequencing only, the clone number corresponds to the identifier number of each clone from the same condition *(details in section 1.4 How does ABSP work? at page 7)*.

Reference DNA sequence

- **Select the reference genome.** It will only be used to display the genomic sequence in the genomic plot. Make sure to click on the *"Pre-install genome"* button if the selected genome is used for the first time. Only a short list of genomes is displayed in the drop-down list but more genome assemblies are available. Go to section *3.3 Code modifications at page 36* to get information on how to add another genome in the drop-down list. The complete list of available genomes can be found in the *"List of BSgenomes.xlsx"* file in the *"ABSP/documents"* folder.
- **Select *.fasta* file of reference DNA sequence.** As described above in the Input files requirements section, the *.fasta* file of the reference DNA sequence needs to be selected from your folders.

### 2.3.3 Output report

The HTML report file of the analysis is automatically saved in the *reports* folder in the ABSP directory.

Header First, in the top panel, the information about the sample experimental conditions is displayed in a table.

- Folder name
- Sequence name
- Collection
- Group
- Replicate number (for direct-BSP only)
- Cloning number (for cloning-BSP only)
- Date of sequencing #1
- Date of sequencing #2
- Date of analysis
- Prefix of output files

Reference DNA This tab summarizes all the data computed from the reference DNA sequence *.fasta* file.

- **Reference DNA sequence** General information on the sequence (name, strand used for primer design, length, and genomic coordinates) and genomic sequence from plus strand (given by the reference DNA *.fasta* file) and minus strand (reverse complement).
- **Localization of CG dinucleotides** Detection of CpG sites on the reference DNA (on plus and minus strand) with attribution of the CG number, from 1 to n on the plus strand.
- **Bisulfite converted sequences** Sequences of reference DNA after theoretical bisulfite conversion (CpG sites considered as methylated). The bisulfite conversion is performed on the strand used for PCR primer design, as only this strand is amplified during PCR. The PCR regenerates the opposite strand, corresponding to the reverse complement of the bisulfite converted DNA template.

Sequencing trimming This tab summarizes the trimming of sequencing results based on quality. Two parameters are used: the Phred quality score of each base retrieved from the sequencing file, and the mixed base peak ratio.

- **Summary** In the first tab, the default thresholds used to trim the sequencing results are displayed.
  - Minimum length of the trimmed sequence (default is 30 bp)
  - Minimum Phred quality score (default is 30, corresponding to a base-calling error probability of 0.001%)
  - Minimum ratio of primary peak (default is 0.75)
  - Minimum percentage of non-mixed positions (default is 75%)

  Below the threshold table, the trimming summary for both sequencing results is displayed and indicates whether or not the trimming was successful

(correct trimmed sequence quality) or failed (incorrect trimmed sequence quality).

- **Details per sequencing**
  - **Raw sequence** The sequence, chromatogram, and data table of the sequencing results are displayed.
  - **Quality report** The first trimming is based on the base-calling quality as it uses the Phred quality scores of each base to find the best sequence to trim. This step is provided by the SangeranalyseR package[5,6] . The thresholds and the results of this quality trimming are displayed.
  - **Mixed base peak report** The second trimming is based on the primary peak ratio over the other peaks for each position. At each position, the signal ratio of the primary peak is computed using the peak height values for each base, with formula: if $peak_C > \{peak_A, peak_T, peak_G\}$

$$Primary\ peak\ ratio = \frac{peak_C}{peak_A + peak_T + peak_G + peak_C}$$

    If the ratio is above the threshold (default is 0.75), the position is considered non-mixed; if the ratio is below the threshold, the position is considered mixed. All the possible trimmed sequences are obtained by selecting the sequence between n (from 3 to 15) consecutive non-mixed positions. For each one of the possible trimmed sequences, the percentage of non-mixed positions is calculated. Among those, the trimmed sequence which is selected corresponds to the one with a percentage of non-mixed positions above the threshold (default is 75%) with the minimum of consecutive non-mixed positions at extremities (this number is displayed).
  - **Trimming plot** The two previous report steps give two different trimmed sequences that can be viewed on the trimming plot (figure 8). The top panel represents a dot plot of the Phred quality score per position, values in green are above the threshold and values in red below, in which the start and end positions of the trimmed sequence are represented by orange vertical lines. The second panel is also a dot plot but it represents the primary peak ratio per position, values in green are above the threshold (considered as non-mixed) and values in red below (considered as mixed), in which the start and end positions of the trimmed sequence are represented by cyan vertical lines. In the last panel, the two trimmed sequences are represented in the same color, orange, and cyan. The raw sequence is displayed in red. The overlapping of the two previous trimmed sequences gives the final trimmed sequence, in green, which is the one kept for the rest of the analysis and corresponds to the

[5] https://sangeranalyser.readthedocs.io/en/latest/index.html

[6] Kuan-Hao Chao et al. "sangeranalyseR: simple and interactive analysis of Sanger sequencing data in R". in: *bioRxiv* (2020), p. 2020.05.18.102459. DOI: 10.1101/2020.05.18.102459.

information given in the summary tab and the final trimmed sequence tab.

– **Final trimmed sequence** The sequence, chromatogram, and data table of the final trimmed sequencing results are displayed.

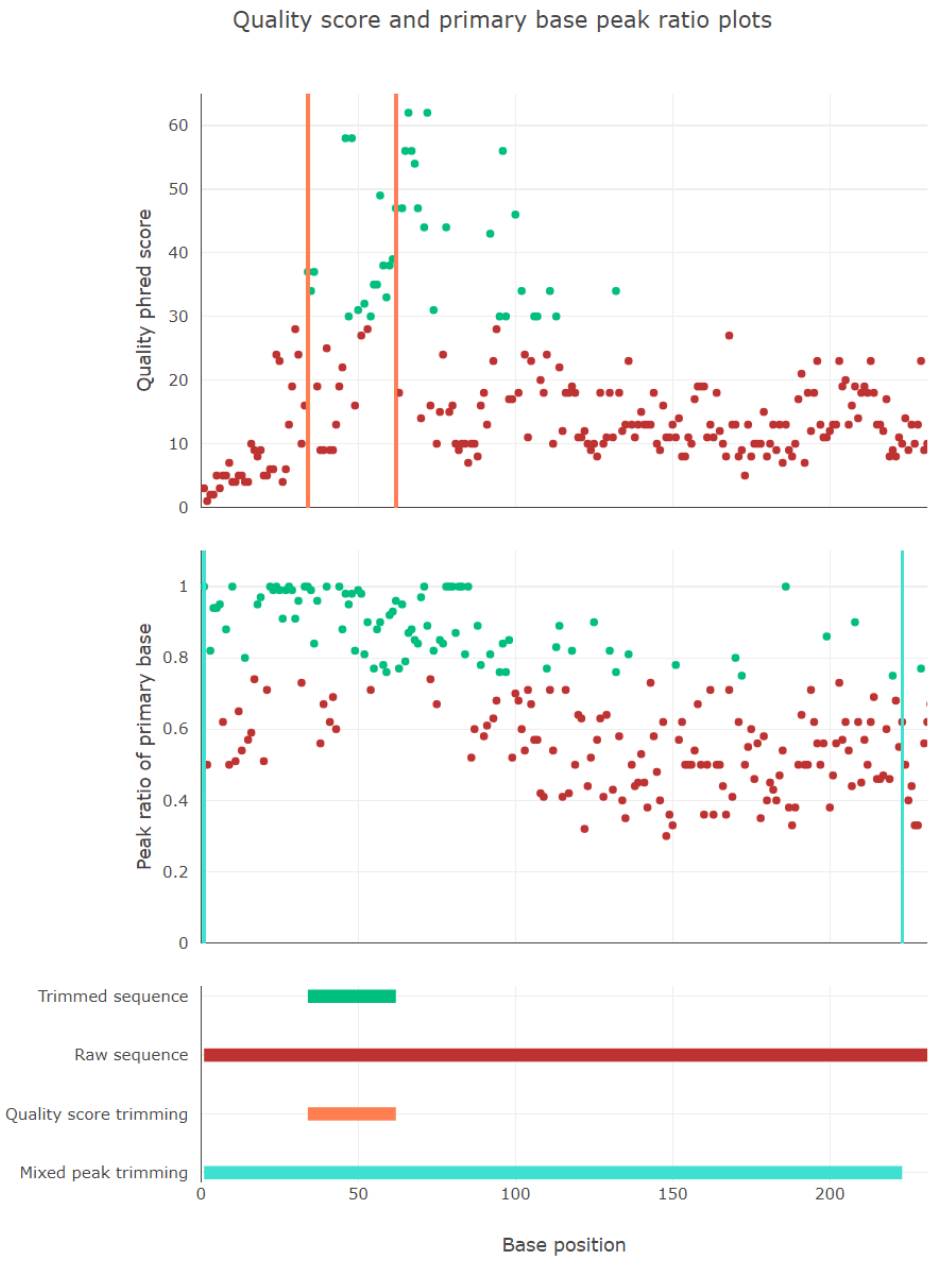


Figure 8. Example of trimming plot from the output report.

Alignments In the first two tabs, each sequencing is aligned with either the sense sequence (bisulfite converted sequence from the template strand) as if it is a forward sequencing, or the antisense sequence (reverse complement of bisulfite converted sequence from the template strand) as if it is a reverse sequencing. The direction of each sequencing is determined based on the aligned sequence length: the alignment which gives the longest aligned sequence is considered the correct one. If for one sequencing the aligned sequences are equal between alignment as forward and as reverse, the direction determination depends on the other one. The two last tabs display the correct alignments and aligned sequences.

Quality control This tab summarizes the quality of the trimmed sequencing result aligned with the reference DNA.

- **Summary** In the first tab, the default thresholds are used to control the quality of the aligned sequencing result.
  - Minimum length of the aligned sequence (default is 30 bp)
  - Minimum identity percentage of alignment (default is 75%)
  - Minimum of bisulfite conversion rate mean (default is 0.90, corresponding to 90% conversion efficacy)

  Below the threshold table, the quality summaries for both sequencing results are displayed and indicate whether or not the aligned sequencing results have a sufficient quality (correct) or insufficient (incorrect) relative to thresholds.
- **Mismatch positions** For both sequencing, a table indicates the mismatched positions and nucleotides, on both the sequencing result and reference DNA sequence.
- **Insertions/deletions** For both sequencing, a table indicates the insertions/deletions (gaps) found in either the sequencing result or the reference DNA sequence.
- **Conversion rates** For both sequencing, a table indicates the bisulfite conversion rate for each cytosine outside a CpG in the aligned sequence. The first column corresponds to the identifier number of the cytosine on the reference DNA sequence, in the second one is displayed its position on the reference DNA sequence, in the third one its position on the trimmed sequencing result, and in the fourth its position on the raw sequencing result. The position matching is obtained thanks to the alignment of sequences. For each position on the raw sequencing result, the peak height values of signals are extracted and used to compute the conversion rate, with the following formulas (for forward and reverse sequencing respectively):

$$\textit{Bisulfite conversion rate} = \frac{peak_T}{peak_C + peak_T}$$

$$\textit{Bisulfite conversion rate} = \frac{peak_A}{peak_G + peak_A}$$

  For each position a bisulfite conversion is obtained. The mean of rates from

all positions is indicated in the summary table in the first tab, as well as standard deviation.

- **Maximum aligned sequence** The maximum aligned sequence corresponds to the sequence covered by at least one of the two sequencing results. Its information such as its length, its coordinates, and its nucleotide sequence are displayed.

Methylation

- For both sequencing, the tables of computed **methylation percentages** are displayed. The first column corresponds to the CpG site identifier number on the reference DNA sequence (list of all CpG sites in the *Reference DNA* tab, *Localization of CG dinucleotides* tab). In the next three columns, its coordinates are specified. Then, the position of the methylated cytosine (the C in forward, the G in reverse) is displayed, as well as its position in the raw and trimmed sequencing results. The position matching is obtained thanks to the alignment of sequences. For each position on the raw sequencing result, the peak height values of signals are extracted and used to compute the methylation percentage, with the following formulas, for forward and reverse sequencing respectively:

$$\textit{Methylation percentage} = \frac{peak_C}{peak_C + peak_T} \times 100$$

$$\textit{Methylation percentage} = \frac{peak_G}{peak_G + peak_A} \times 100$$

- **Combined** Methylation results from both sequencing results are then combined in a unique table with the calculation of the average methylation and standard deviation per position.
- **Individual methylation plot** Finally, a plot is generated to visualize results relative to the genomic sequence.

Output data

- **Directories** A diagram of output files directories is displayed (figure 9).
- **Files** A list of all the output files with links to local folders is displayed.
- **Methylation data file preview** The output methylation data file which will be used as input for the grouped analysis is displayed as a table (The *"alg_coord_start"* and *"alg_coord_end"* columns contain a unique value corresponding to the start and end coordinates of the maximum aligned sequence for the individual analysis).

### 2.3.4 Output files

All the output files are located in the *results* directory, as depicted by the diagram in figure 9.

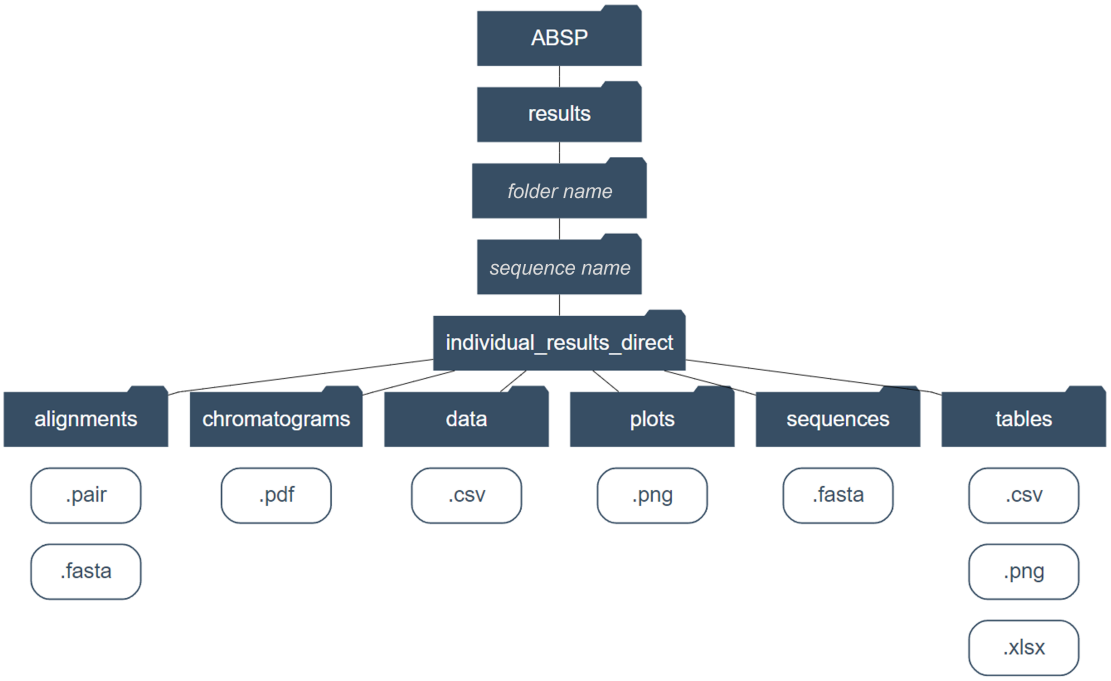


Figure 9. Diagram of output directories to locate output files from the individual analysis. The *"folder name"* and *"sequence name"* depend on the input entries when launching the analysis.

As for the same sequence (same primer set) results from both direct-BSP and cloning-BSP can be generated, the two types of outputs are separated into two subfolders: *"individual_results_direct"* and *"individual_results_cloning"*.

- **alignments** In subfolders specific to each individual analysis, the .pair files of the alignments and the *.fasta* files of the aligned sequence are saved.
- **chromatograms** In subfolders specific to each individual analysis, the chromatograms of raw sequencing and the chromatograms of the trimmed sequencing are saved as .pdf files.
- **data** It contains all the methylation data *.csv* files for the grouped analysis.
- **plots** It contains all the individual methylation plots as well as legends for plots as *.png* image files.
- **sequences** The reference DNA sequences from plus and minus strands and the bisulfite converted sequence of the template strand are saved as *.fasta* files.
- **tables** In subfolders specific to each individual analysis, the data tables of raw sequencing results, the data tables of trimmed sequencing results, the summary table of sequencing trimming, the summary table of quality control, the bisulfite conversion rates table and the methylation percentages tables

are saved as *.csv* *.png* and *.xlsx* files.

## 2.4 Grouped analysis

### 2.4.1 Procedure

In the **grouped analysis tab**, the left panel is to launch analysis, and the right panel provides entries information.

Experiment information

- **Select an existing folder.** Located in the *"ABSP/results"* folder, all of the analysis results will be saved in this folder. Having different folders of results can be used to separate the different analyses by projects, experiments, or users. Note that the six first letters of the folder name will appear in the report file name.
- **Select an existing sequence folder.** Located in the *"ABSP/results/previous folder"* folder, all of the analysis results will be saved in this folder for this sequence. This sequence name will be used in output files (tables, plots...) to refer to the sequence.
- **Select the reference genome.** It will only be used to display the genomic sequence in the genomic plot. See section 3.3 Code modifications at page 36 to add another genome in the drop-down list. The complete list of available genomes can be found in the *"List of BSgenomes.xlsx"* file in the *"ABSP/documents"* folder.
- **Select the experiment type.** The choice of the experiment can be either Direct-BSP or Cloning-BSP. The correct experiment type entry is essential to retrieve the methylation data files either in the *"individual_results_direct"* or *"individual_results_cloning"* folders.

Plot parameters

- **Select position labels for plots.** The CpG positions on plots can be referred to by different label types:
  - The **CpG coordinates** label type displays the genomic coordinates of the CpG site in the format *chr#:######-######* (e.g. *"chr16:68771230-68771231"*).
  - The **CpG numbers** label type displays the CpG site identifier number on the represented sequence, from 1 to n.
  - The **None** label type displays blank labels, which can be a suitable alternative in case of extremely close CpG positions as labels may overlap.
- **Choose to separate plots by collections.** For lollipop plots, genomic plots, and boxplots and only for display purposes (it does not affect data). If this parameter is not ticked (default), all samples from different collections will be displayed on the same plot. If this parameter is ticked, a plot is generated

by collection, displaying samples from one collection only.

- **Indicate the order of groups for display.** For this input to work, the folder, sequence, and experiment type have to be selected and correct. In this case, the group names are extracted from methylation data files corresponding to the previously selected entries. The groups must all be selected, in the desired order for display.
- **Select the types of sample ordering for plots.** Four different sample ordering (ordinate axis ordering) are available for visualization plots, each provides a specific way of ordering samples on the ordinate axis based on different parameters. At least one must be selected, up to the four of them.
  - **As it is** arranges samples by alphabetic order of collections. If none or one collection is present, this order is equivalent to the *By groups* one.
  - **By groups** arranges samples by the provided group order above.
  - **By methylation levels** arranges samples depending on their methylation mean.
  - **By clusters** arranges samples depending on the hierarchical clustering calculated and represented by an associated dendrogram.

### 2.4.2 Output report

The HTML report *.html* file of the analysis is automatically saved in the *"reports"* folder in the ABSP directory.

Header First, in the top panel, the information about the sample experimental conditions are displayed in a table.

- Folder name
- Sequence name
- Reference genome
- Type of experiment
- Group order
- Date of analysis

Files content This tab summarizes all the data that have been used for this analysis.

- **Data files content** General information about the data is listed: sequence name, collections, groups and replicates or clones.
- **Data files paths** Paths of the methylation files that were found and used for the analysis are listed.

Methylation data This tab regroups the methylation data in tables from the retrieved files.

- **Methylation data of replicates/clones** For each replicate or clone, depending on the experiment type, the methylation percentages of each CpG sites

are displayed in a table, with the mean of all positions and the associated standard deviation by replicate/clone.
For clones: Methylation percentages calculated from sequencing results are converted to 0% or 100% methylation status.
  - A CpG site is considered unmethylated (0%) when methylation percentage is between 0% and the defined threshold (default is 20%).
  - A CpG site is considered methylated (100%) when methylation percentage is between the defined threshold (default is 80%) and 100%.
  - A CpG site partially methylated, with methylation percentage between 20% and 80%, is removed and annotated as NA (Not Available).
  - For one clone, if more than a threshold percentage (default is 20%) of CpG sites are partially methylated, the clone is considered as defective and all of its CpG sites are annotated as NA (Not Available).

  The thresholds can be modified in the script, please refer to section 3.3 Code modifications at page 36.

- **Methylation data of groups** For each group, the mean of methylation percentages of each CpG sites are displayed in a table, with the mean of all positions and the associated standard deviation by groups.

Plots of replicates Only for direct-BSP. This tab provides plots to visualize the methylation data of replicates.
Lollipop plots (condensed and proportional), genomic plot and cluster dendrogram plot are as illustrated in figure 10.

Plots of clones Only for cloning-BSP. This tab provides plots to visualize the methylation data of clones from each sample in separated plots.
Lollipop plots (condensed and proportional), genomic plot and cluster dendrogram plot are illustrated in figure 11.

Plots of groups This tab provides plots to visualize the methylation data of groups per collection. For each sample, the mean of replicates or clones is calculated per CpG and results are displayed in this tab.
Lollipop plots (condensed and proportional), genomic plot and cluster dendrogram plot are illustrated in figure 12.

Statistics This tab aims to compare methylation between groups, to find statistical significant differences, either by comparing CpG site per CpG site, or by comparing the all region.

- **Descriptive statistics of groups** Two tables of methylation data descriptive statistics are displayed: one for methylation data by CpG positions, and one for methylation data of all CpG positions for each sample.
- **Student's T test** Two tables display T tests p-values between groups 2 by 2, one table for methylation data by CpG positions, and one for methylation data of all CpG positions for each sample.

- **Boxplots** Boxplots of methylation data with T tests p-values as numbers or symbols are generated, one plot details data of each CpG position (figure 13 A for direct-BSP data and figure 13 D for cloning-BSP), and another one data from methylation mean of the region (figure 13 B for direct-BSP data and figure 13 E for cloning-BSP).
- **Methylation profile plots** Methylation profile plots with Kruskal-Wallis tests p-values are generated for each collection (figure 13 C and F).

The Methylation plotter[7,8] tool served as a base to develop the different types of plots with different sample ordering.

Output data

- **Directories** A diagram of output files directories is displayed (figure 14).
- **Files** A list of all the output files with links to local folders is displayed.

### 2.4.3 Output files

All the output files are located in the *results* directory, as depicted by the diagram in figure 14.

As for the same sequence results from both direct-BSP and cloning-BSP can be generated, the two types of outputs are separated into two subfolders: *"grouped_results_direct"* and *"grouped_results_cloning"*.

- **boxplots** Boxplots are saved as *.png* files.
- **dendro_plots** Cluster dendrograms are saved as *.png* files.
- **genomic_plots** Genomic plots are saved as *.png* files, in subfolders for plots of replicates/clones and plots of groups.
- **lollipop_plots** Lollipop plots are saved as *.png* files, in subfolders for plots of replicates/clones and plots of groups.
- **meth_profile_plots** Methylation profile plots are saved as *.png* files
- **tables** The methylation data tables, the descriptive statistics tables of positions or means (means of all positions), and the Student's T test tables are saved as *.csv* and *.xlsx* files.

[7] http://maplab.imppc.org/methylation_plotter/index.html

[8] Izaskun Mallona, Anna Díez-Villanueva, and Miguel A Peinado. "Methylation plotter: a web tool for dynamic visualization of DNA methylation data". In: *Source Code for Biology and Medicine* 9.1 (2014). Methylation plotter, p. 11. DOI: 10.1186/1751-0473-9-11.

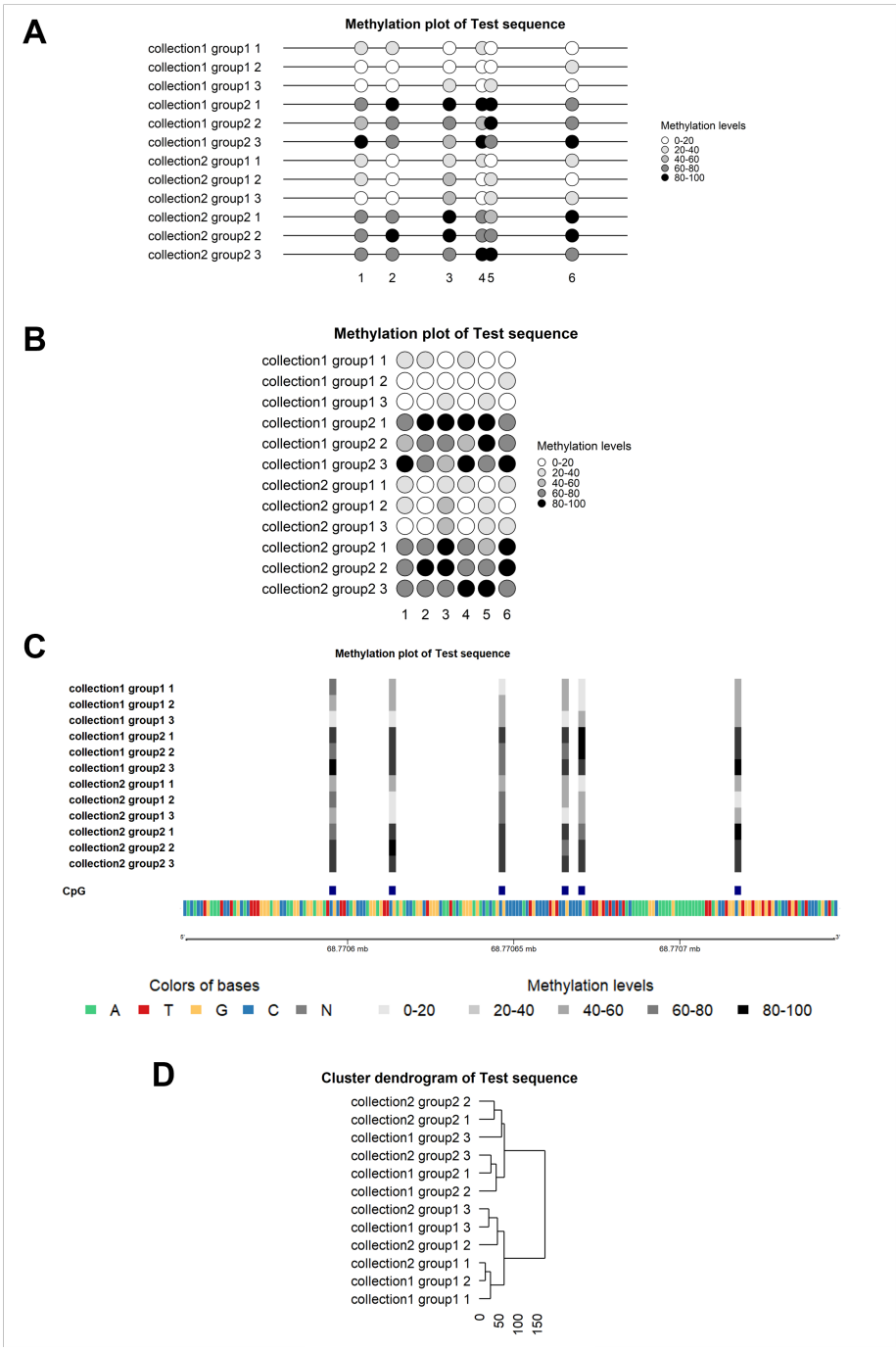


Figure 10. Visualization plots of all replicates (direct-BSP). The plots were generated based on mock methylation data for a test sequence. Three replicates (1, 2, and 3) per sample are represented on the plots, from two groups (group1 and group2) and two collections (collection1 and collection2) (sample ordering *as it is*). The methylation levels are given as percentages and corresponds directly to the methylation output data from the individual analyses. **A.** Proportional lollipop plot. **B.** Condensed lollipop plot. **C.** Genomic plot. **D.** Cluster dendrogram (for sample ordering *by clusters*).

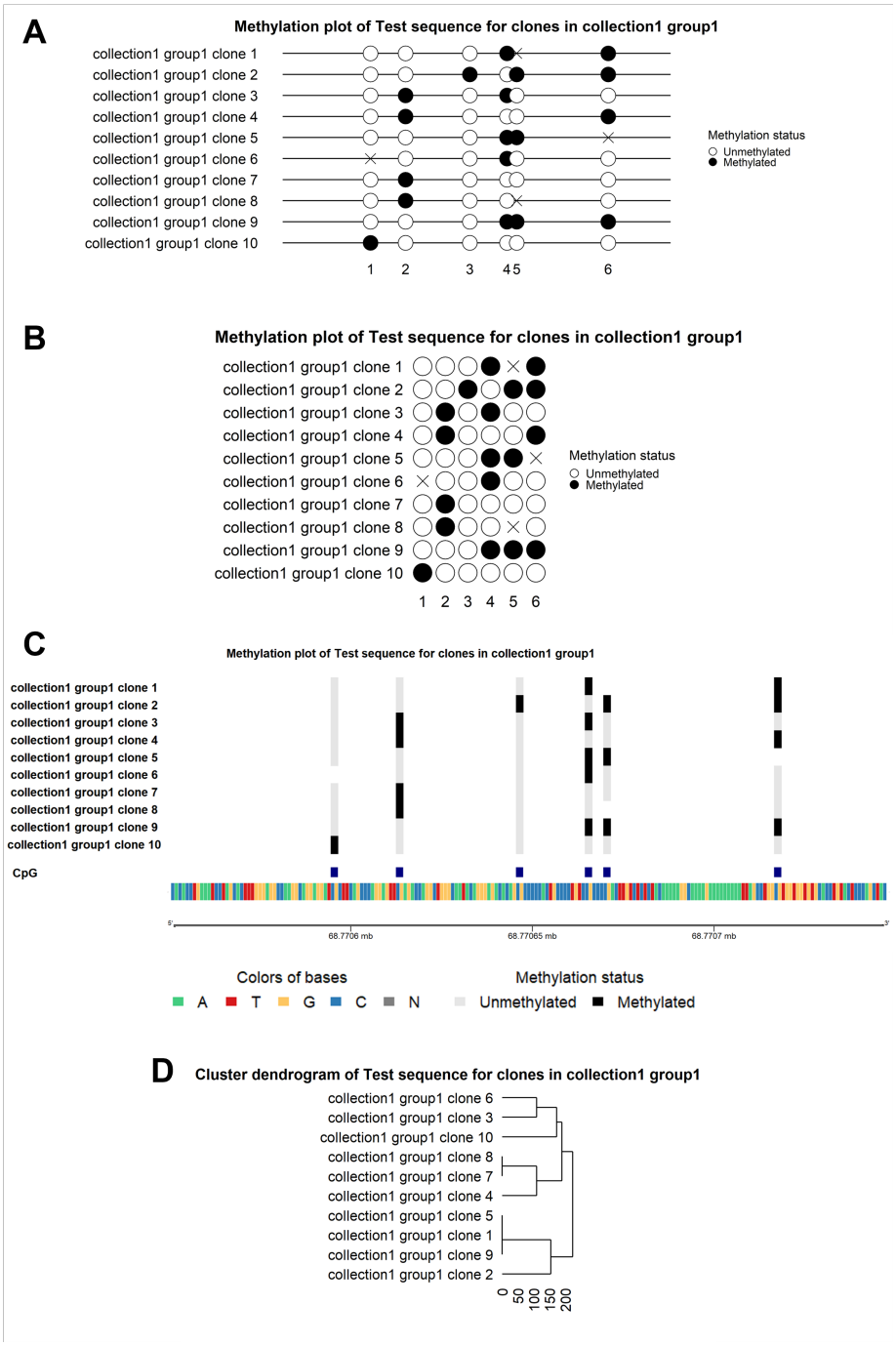


Figure 11. Visualization plots of all clones from one sample (cloning-BSP). The plots were generated based on mock methylation data for a test sequence. Ten clones (from 1 to 10) for the sample *"collection1 group1"* are represented on the plots (sample ordering *as it is*). The methylation status corresponds to the conversion of methylation percentages from the individual analyses into unmethylated (0%) or methylated (100%) (or not available) methylation statuses. **A.** Proportional lollipop plot. **B.** Condensed lollipop plot. **C.** Genomic plot. **D.** Cluster dendrogram (for sample ordering *by clusters*).

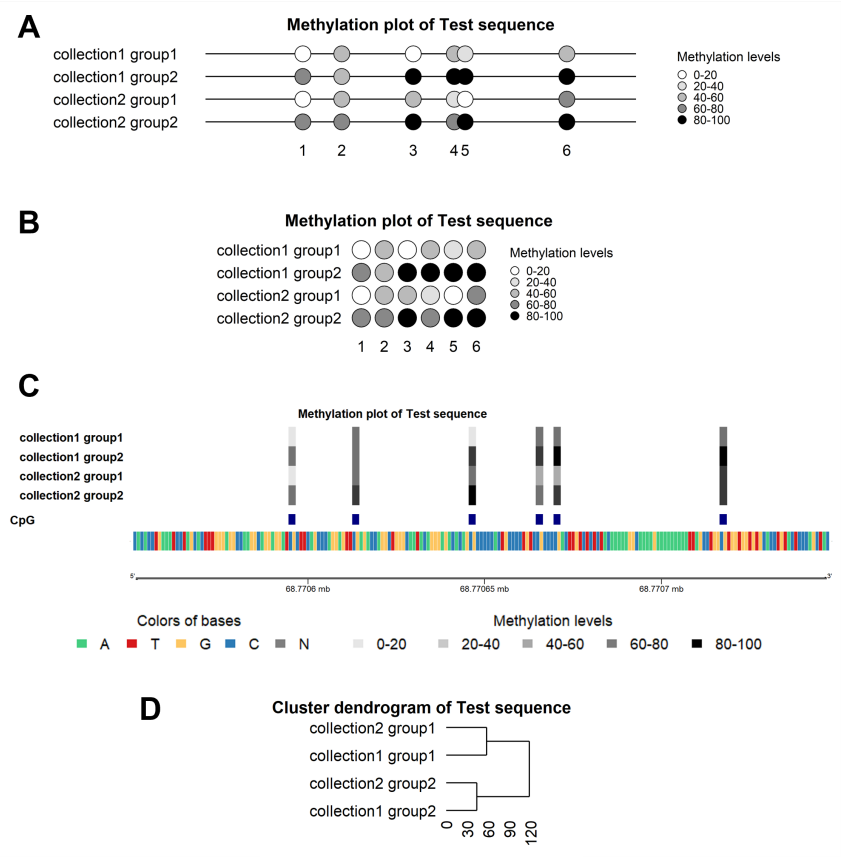


Figure 12. Visualization plots of groups (means of replicates/clones per samples). The plots were generated based on mock methylation data for a test sequence. Plots represents the means of methylation percentages of ten clones per sample and per CpG position (sample ordering a*s it is*). **A.** Proportional lollipop plot. **B.** Condensed lollipop plot. **C.** Genomic plot. **D.** Cluster dendrogram (for sample ordering *by clusters*).

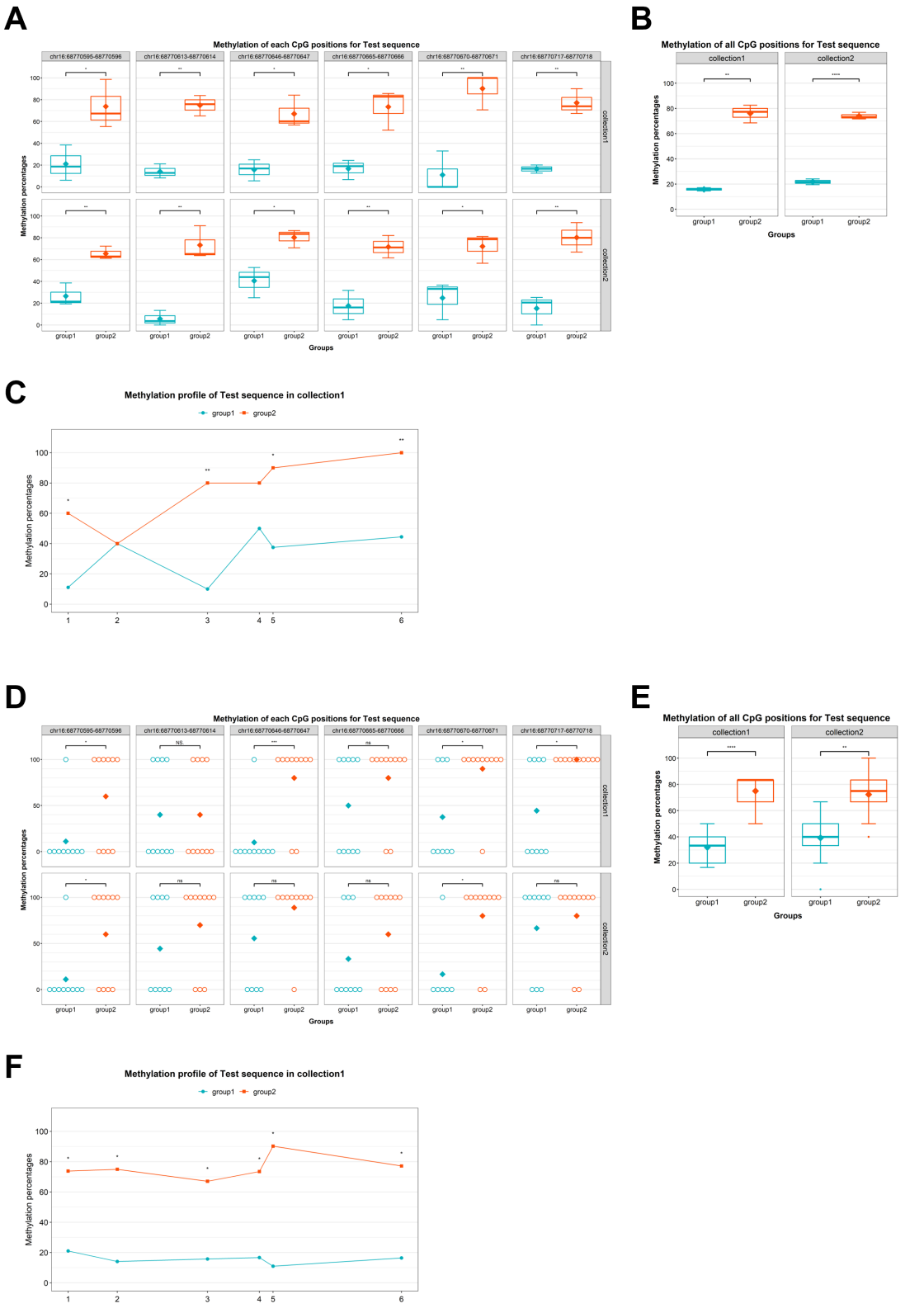
A
Methylation of each CpG positions for Test sequence
chr16:68770595-68770596
chr16:68770613-68770614
chr16:68770646-68770647
chr16:68770665-68770666
chr16:68770670-68770671
chr16:68770717-68770718
collection1
collection2
Methylation percentages
group1
group2
Groups
B
Methylation of all CpG positions for Test sequence
collection1
collection2
Methylation percentages
group1
group2
Groups
C
Methylation profile of Test sequence in collection1
group1
group2
Methylation percentages
D
Methylation of each CpG positions for Test sequence
NS
ns
collection1
collection2
Methylation percentages
group1
group2
Groups
E
Methylation of all CpG positions for Test sequence
collection1
collection2
Methylation percentages
group1
group2
Groups
F
Methylation profile of Test sequence in collection1
group1
group2
Methylation percentages

Figure 13. Boxplots and methylation profile plots. The plots were generated based on mock methylation data for a test sequence. **A.** Boxplots of direct-BSP methylation results for each CpG site. **B.** Boxplots of direct-BSP methylation results for means of all CpG sites. **C.** Methylation profile plot of direct-BSP methylation results for one of the two collections. **D.** Boxplots of cloning-BSP methylation results for each CpG site. As methylation levels of CpG from clones can only be either 0% or 100% thereby boxes can't be drawn, instead, each clone is represented in the plot by a circle. **E.** Boxplots of cloning-BSP methylation results for means of all CpG sites. **F.** Methylation profile plot of cloning-BSP methylation results for one of the two collections. In boxplots, symbols represents significance levels of Student's T test p-values (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$). In methylation profile plots, symbols represents significance levels of Kruskal-Wallis test p-values (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$).

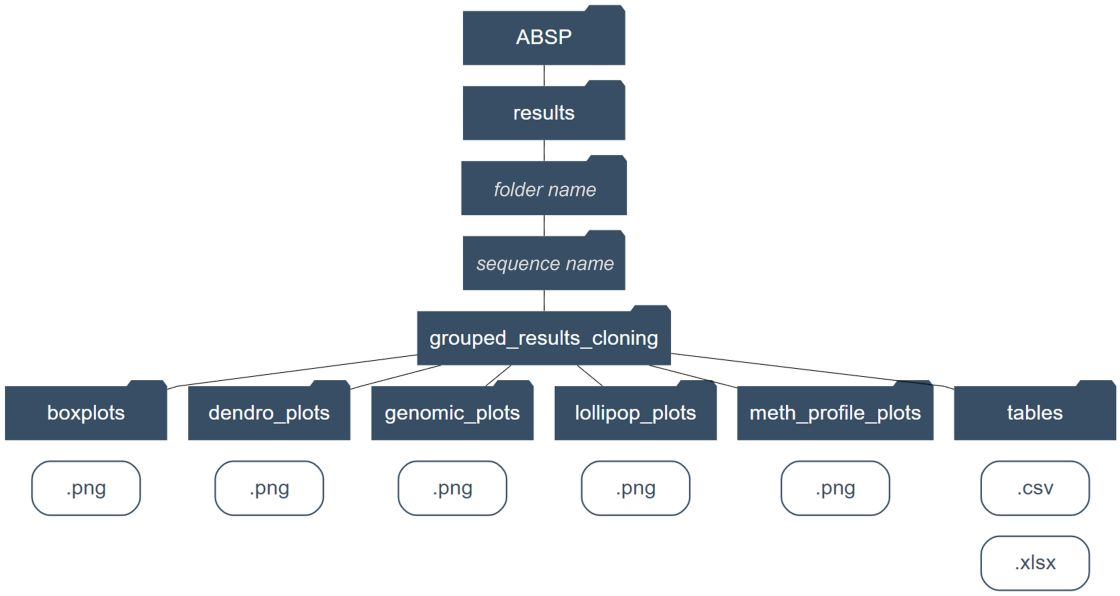


Figure 14. Diagram of output directories to locate output files from the individual analysis. The *"folder name"* and *"sequence name"* depend on the input entries when launching the analysis.

## 2.5 Multiple analyses

As described in the figure 5, multiple analyses can be launched at the same time using data tables as input with all the required information. Both individual and grouped analyses can be launched, at the same time or separately.

### 2.5.1 Input files requirements

Two files are provided in the *"documents"* folder.They must be filled with the desired input entries, all information about how to fill the documents are also indicated within the documents as notes.

The documents are provided in the *.xlsx* (Microsoft Excel Open XML Format Spreadsheet) and *.ods* (OpenDocument Spreadsheet) formats, but the input file format must be either *.xlsx* or *.csv* (Comma-Separated Values) (*.ods* files must be converted to one of those formats).

- **Experiments data table for individual analyses:** *"multiple_individual_analyses_table.xlsx"*.
  - SEQUENCE NAME. The sequence name should be unique and consistent for each amplicon. It must not contain any special character.
  - COLLECTION. The collection refers to a separation of samples above the groups/conditions. Leave empty if you do not want to indicate a collection.
  - GROUP. The group refers to the condition that is to be compared. It must not contain any special character.
  - REPLICATE NUMBER. The replicate number refers to the experiment repetition identifier. It must be an integer $\geq$1. Only for direct-BSP experiments, leave empty otherwise.
  - CLONE NUMBER. The clone number refers to the individual clone identifier. It must be an integer $\geq$1. Only for cloning-BSP experiments, leave empty otherwise.
  - GENOME. The genome refers to the reference genome assembly used for coordinates and for plots displaying the genomic sequence. Only a short list of genomes are displayed in the cells, but you can use another available genome. Please refer to the provided document: *"List of BSgenomes"* to get the list of available genomes assemblies and the correct spelling.
  - PATH TO FASTA FILE OF REFERENCE DNA. Path to the *.fasta* file of the reference DNA sequence on your computer. On Windows, you can copy the file path by holding down shift then right-clicking on the file and selecting *"Copy as path"* in the menu. On MacOS, you can copy the file path by right-clicking on the file to display the menu then holding down the option key and selecting *"Copy ... as Pathname"*.
  - DATE SEQUENCING #1. Date of the sequencing result #1 for traceability. The date format must be YYYY-MM-DD and the cell format must be set to *"date"*.
  - PATH TO SEQUENCING FILE #1. Path to sequencing result #1 .ab1 file.
  - DATE OF SEQUENCING #2. Date of the sequencing result #2 for traceability.

The date format must be YYYY-MM-DD and the cell format must be set to *"date"*.

- PATH TO SEQUENCING FILE #2. Path to sequencing result #2 .ab1 file.

- **Parameters table for grouped analyses:** *"multiple_grouped_analyses_table.xlsx"*.
  - SEQUENCE NAME. The sequence name should be unique and consistent for each amplicon. It must not contain any special character.
  - GENOME. The genome refers to the reference genome assembly used for coordinates and for plots displaying the genomic sequence. Only a short list of genomes are displayed in the cells, but you can use another available genome. Please refer to the provided document: *"List of BSgenomes"* to get the list of available genomes assemblies and the correct spelling.
  - EXPERIMENT TYPE. The experiment type can be either Direct-BSP or Cloning-BSP.
  - CpG POSTION LABEL TYPE. The CpG position label type refers to the displayed element corresponding to the CpG position on plots.
  - SEPARATION OF COLLECTIONS. This parameter refers to the generation of plots: if FALSE, all samples from all collections will be displayed on the same plot and if TRUE, samples will be split in different plots, one plot per collection.
  - LIST OF GROUPS ORDERED. The cell must contain all of the groups within the experiment in the order you want them to be displayed. Type your groups separated by commas. It must not contain any special character.
  - TYPE OF SAMPLE ORDERING. The type of sample ordering refers to the parameter used for ordering samples on plots. Four types of ordering are available: As it is, By groups, By methylation levels and By clusters. Multiple ordering can be chosen by typing the ordering names separated by commas.

### 2.5.2 Procedure

- **Select an existing folder** within the ABSP results folder to locate all of the analyses results. To create a new folder, select the *"Create new folder"* entry and enter the name of the new folder in the text input. Note that the six first letters will appear in the report file name.
- **Select the filled table as input.** Both the experiments data table and the grouped parameters table can be provided at the same time to launch individual analyses followed by grouped analyses, or only one of the two tables can be provided and will launch the corresponding analyses, either individual analyses or grouped analyses.
- **Launch the analyses** by clicking on the bottom button *"Run analyses"*.

### 2.5.3 Output files

The reports and output files of analyses are saved in the *reports* and *results* folders, in the same way as for manual launch of individual and grouped analyses, see

sections 2.3 Individual analysis at page 14 and 2.4 Grouped analysis at page 23.

# 3 Complementary information

## 3.1 Some recommendations for the BSP experiment

The length of the PCR products should be exceed 350-400 bp, as the bisulfite treatment causes DNA strand breakages long amplicon can not be properly amplified, and ABSP plot display for sequences above 400 bp with numerous CpG is not optimal.

Several tools can be used to design BSP primers, several are listed in the Methtools[9] list of tools, such as Methprimer[10] to design primers and BiSearch[11] to check for unintended PCR products, on bisulfite treated DNA.

## 3.2 Detailed workflow of ABSP individual analysis

In figure 15, the individual analysis input, steps and outputs are described with more details.

[9] http://bigd.big.ac.cn/methbank/methTool/list/
[10] https://www.urogene.org/methprimer/
[11] http://bisearch.enzim.hu/

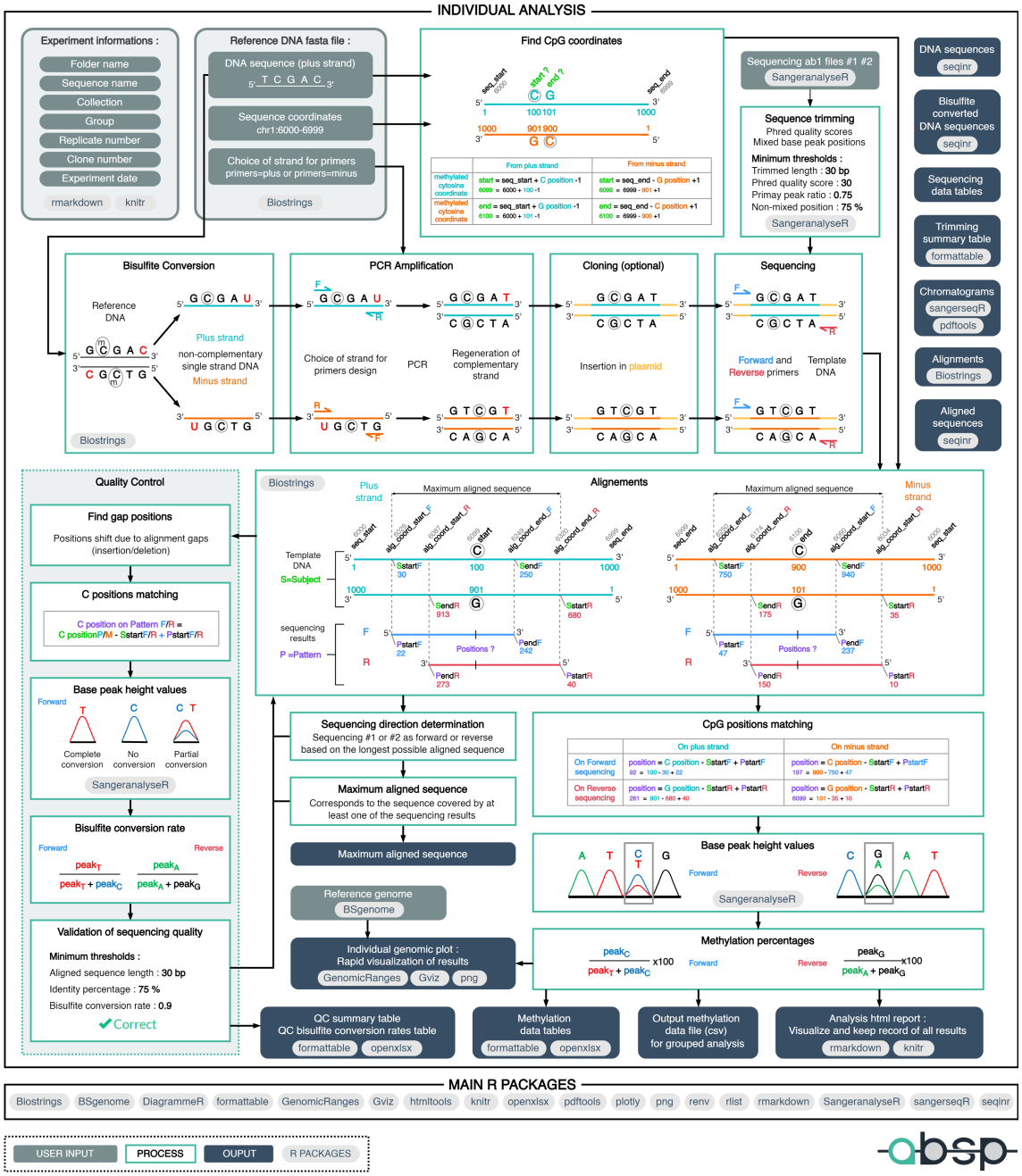


Figure 15. Detailed workflow of the individual analysis.

## 3.3 Code modifications

### 3.3.1 List of reference genomes

In the individual analysis and grouped analysis tabs of the app, the drop-down list to select the reference genome is limited. If your reference genome does not appear it doesn't mean that it is not available, and can manually added. The complete list of genomes assemblies[12] with the correct spelling can be found in the *"List of BSgenomes.xlsx"* file in the *"ABSP/documents"* folder.

To modify the displayed drop-down list items, the *app.R* script have to be modify. In the listing 1, the *"list_genomes"* object corresponds to the vector of character strings listing the displayed reference genome. Any of the other reference genome can be added to the list.

Listing 1. List of reference genomes displayed in the drop-down lists as selectable inputs for analyses. From the *"app.R"* script.

```
29 # ********************************
30
31 # Here, to simplify, a short list of genomes is displayed but all BSgenome
       can be used
32 # To get the list of all genomes of BSgenome package run : 'BSgenome::
      available.genomes()', more information on genomes at https://genome.
      ucsc.edu/cgi-bin/hgGateway
33 # A new genome can be added to the list displayed just below :
34 list_genomes <- c(
35     "BSgenome.Hsapiens.UCSC.hg19", "BSgenome.Hsapiens.UCSC.hg38", "
      BSgenome.Mmusculus.UCSC.mm10", "BSgenome.Mmusculus.UCSC.mm39",
36     "BSgenome.Rnorvegicus.UCSC.rn6", "BSgenome.Rnorvegicus.UCSC.rn7", "
      BSgenome.Cfamiliaris.UCSC.canFam3", "BSgenome.Mmulatta.UCSC.rheMac8",
37     "BSgenome.Ggallus.UCSC.galGal6" , "BSgenome.Drerio.UCSC.danRer11", "
      BSgenome.Celegans.UCSC.ce11", "BSgenome.Dmelanogaster.UCSC.dm6")
38
39 # ********************************
```

### 3.3.2 Modify the default thresholds

#### Individual analysis

The default thresholds used for the individual analysis can be modified (listing 2).

#### Grouped analysis

The default thresholds used for the grouped analysis, for clones methylation percentage conversion into methylation status (0% or 100%), can be modified (listing 3).

[12] https://genome.ucsc.edu/cgi-bin/hgGateway

Listing 2. Default thresholds in the individual analysis script. From the *"ABSP_individual_analysis.Rmd"* script.

```
```{r Thresholds, include=F}

# Thresholds

# Maximum base-calling error probability (value per position):
th_quality_error <- 0.001
# Minimum phred quality score, logarithmically linked to error probability (value per position):
th_quality_phred <- (-10*log(th_quality_error,10))

# Minimum ratio of primary peak, corresponding to the primary peak value over the total of peak value, to consider a position as non-mixed (value per position):
th_mixed_position <- 0.75
# Minimum percentage of non-mixed positions in the trimmed sequence to be considered as non-mixed (value for the total trimmed sequence):
th_mixed_perc <- 75 # %

# Minimum length of trimmed sequences
th_min_trim <- 30 # bp

# Minimum length of aligned sequences
th_min_alg <- 30 # bp
```
```

Listing 3. Default thresholds in the grouped analysis script. From the *"ABSP_grouped_analysis.Rmd"* script.

```
```{r Thresholds, include=F}

# Thresholds (used for cloning only) :

# unmethylated clones : methylation between 0% and 20%
th_unmethylated_max <- 20

# methylated clones : methylation between 80% and 100%
th_methylated_min <- 80

# maximum proportion of partial positions allowed : 20% of CpG positions
th_partialpos_ratio <- 0.2

clone_thresholds <- c(th_unmethylated_max, th_methylated_min, th_partialpos_ratio)
```
```

### 3.3.3 Modify the plots colors and point shapes

For the grouped analysis, the plot colors and point shapes parameters can be modified. As depicted in listing 4, the color of bases can be changed in the *bases_colors* object, the plot colors for each group in the *plot_colors* object and the point shapes for each group in the *plot_shapes* object.

Listing 4. Colors and shapes setting for plots in the grouped analysis script. From the *ABSP_grouped_analysis.Rmd* script

```
322 ```{r colors and shapes settings, include=F}
323 # Colors of sequence track for genomic plots
324 bases_colors <- c(A="#43CD80", T="#D7191C", G="#FDC661", C="#2C7BB6", N="#
        7F7F7F")
325
326 # Colors of groups for plots
327 plot_colors <- c("aqua"="#00AFBB",
328                  "tangerine"="#FC4E07",
329                  "sun"="#E7B800",
330                  "berry"="#d30446",
331                  "lime"="#90c613",
332                  "grape"="#7839de",
333                  "flamingo"="#d12a97",
334                  "jade"="#00b673",
335                  "ink"="#1221ed",
336                  "terracotta"="#a93b2c")
337
338 # Shapes of groups for methylation profile plots
339 plot_shapes <- c("round"=19,
340                  "square"=15,
341                  "triangle"=17,
342                  "round_border"=21,
343                  "square_border"=22,
344                  "triangle_border"=24,
345                  "diamond"=18,
346                  "small_round"=20,
347                  "reverse_triangle_border"=25,
348                  "diamond_border"=23)
349 ```
```

In the *ABSP_grouped_analysis.Rmd* script, the code chunk below the one depicted in the listing 4 gives examples of different palette colors and points shapes, with functions to visualize them.

# 4 Troubleshooting guide

## 4.1 Individual analysis

| Error | Cause | Solution |
|---|---|---|
| **Analysis report aborted after the reference DNA step.** *Error: Length of reference DNA sequence does not match with the provided genomic coordinates. Please verify concordance between the reference DNA sequence and genomic coordinates.* | The provided reference sequence in the fasta file has not the same length as the one calculated based on genomic coordinates provided in the fasta file header. So the sequence and coordinates might not match. | Check the reference and its coordinates. If the IGV (Integrative Genomics Viewer) software is used to get them, make sure to properly verify the first and last nucleotides, as there can be a one nucleotide difference between the coordinates and the actual sequence. |
| **Analysis report aborted after the sequencing trimming step.** *Error: Analysis has been stopped as none of the sequencing results are of sufficient quality to be used.* | The provided sequencing results are not of good quality, the trimming step was not able to find a trimmed sequence long enough (below length threshold) and/or with poor quality (below quality score and/or non-mixed positions thresholds). | Lowering a bit the trimming thresholds might solve the issue, but if not, no solution can be provided. The best recommendation is to perform the sequencing run one more time. If the sequencing is not of good quality again, then there must be an experimental issue with BSP samples or sequencing runs. |
| **Analysis report aborted after the alignment step.** *Error: Analysis has been stopped as none of the possible alignments are of sufficient length (< N bp) to be used.* | Even if the sequencing results were of good quality and successfully trimmed, the alignments of them with the reference DNA sequence gives too short aligned sequences to pursue the analysis. | Check if the reference DNA sequence is the correct one. Check if the sequencing results passed the trimming step with values just above thresholds that might explain the alignment results. In this case, lowering some thresholds may solve the issue. If not, the best option is to provide new results from a new sequencing run. |

| | | |
|---|---|---|
| **Analysis report aborted after the alignment step.** *Error: Analysis has been stopped as sequencing result direction (forward or reverse) could not be found.* | This error occurs when the determined direction for both sequencing results happened to be the same, or when only one sequencing has been successfully trimmed and aligned, but alignments as forward or reverse have the same length. | For the first case, the results might have passed the alignment step with values just below thresholds, explaining the identical length of aligned sequence in both direction, still the best option should be to provide new sequencing results to have better trimmed and aligned results. For the second case, lowering a bit some thresholds might help for the other sequencing to pass the trimming and alignment steps, still the best option should be to provide new sequencing results to have better trimmed and aligned results. |
| **Analysis report aborted after the alignment step.** *Error: Analysis has been stopped as none of the provided sequencing are of sufficient quality to be used.* | The provided sequencing results are not of good quality, they passed the trimming step but not the alignment step. The aligned sequences were not long enough (below length threshold) or direction determination failed. | Lowering a bit the trimming and alignment thresholds might solve the issue, but if not, no solution can be provided. The best recommendation is to perform the sequencing run one more time. If the sequencing is not of good quality again, then there must be an experimental issue with BSP samples or sequencing runs. |
| **Analysis report aborted after the Quality Control (QC) step.** *Error: Analysis has been stopped as sequencing results are defined as incorrect by Quality Control. The analysis can not be performed.* | The provided sequencing results are not of good quality when compared to the reference DNA; they passed the trimming and alignment steps but the quality control found the sequencing results incorrect: either the identity percentage or the mean bisulfite conversion rates is are below the threshold. | Lowering a bit the thresholds might solve the issue, but if not, no solution can be provided. The best recommendation is to perform the sequencing run one more time. If the sequencing is not of good quality again, then there must be an experimental issue with BSP samples or sequencing runs. |

| | | |
|---|---|---|
| **Analysis report aborted after the Quality Control (QC) step.**<br>*Error: Analysis has been stopped as no CpG sites were found covered by sequencing results.* | Even if the sequencing results passed the trimming, alignment and quality control steps, the aligned results is not long enough and does not cover any CpG sites on the sequence, thereby methylation levels can not be computed. | Lowering a bit the thresholds might solve the issue to get longer aligned sequences, but if not, no solution can be provided. The best recommendation is to perform the sequencing run one more time. If the sequencing is not of good quality again, then there must be an experimental issue with BSP samples or sequencing runs. |
| **Analysis failed and no report has been generated** | Unexpected error. | Please contact us and send us the error message appearing in the RStudio Console. |

## 4.2 Grouped analysis

| Error | Cause | Solution |
|---|---|---|
| **Analysis report aborted after the header.** *Error: No file methylation data files found. Check inputs: folder name and sequence name, and check the 'data' directory in the individual results folder.* | No methylation data files from the individual analysis, corresponding to the input information, were able to be retrieved. | Check if the individual analysis have already been run, if the files were not moved in another folder, thy should be in the 'data' directory in the individual results folder, or if the input information as folder name, sequence name, and experiment type are correct. If the issue is not solved by these recommendations, please contact us. |
| **Analysis failed and no report has been generated.** | Unexpected error. | Please contact us and send us the error message appearing in the RStudio Console. |

# 5 Acknowledgements

An important part of the plot generation work for the grouped analysis was based on the R code provided by the Methylation plotter[13,14] tool.

Main R packages used:

- arrangements
- BiocManager
- Biostrings
- BSgenome
- compareGroups
- DiagrammeR
- dplyr
- formattable
- GenomeInfoDb
- GenomicRanges
- ggdendro
- ggplot2
- ggpubr
- Gviz
- htmltools
- htmlwidgets
- knitr
- openxlsx
- pdftools
- plotly
- png
- purrr
- RColorBrewer
- readr
- renv
- rlist
- rmarkdown
- Rmisc
- rstatix
- sangeranalyseR
- sangerseqR
- seqinr
- shiny
- shinythemes
- webshot

[13] http://maplab.imppc.org/methylation_plotter/index.html

[14] Mallona, Díez-Villanueva, and Peinado, "Methylation plotter: a web tool for dynamic visualization of DNA methylation data".