A partir de données brutes de séquençage, le plateau technique propose une analyse bioinformatique des séquences obtenues par RRBS (*Reduced Representation Bisulfite Sequencing*).

Remarque : ce type d'analyse n'est envisageable que sur des espèces pour lesquelles il existe un génome de référence, et éventuellement une annotation (position des gènes).

Prestation proposée

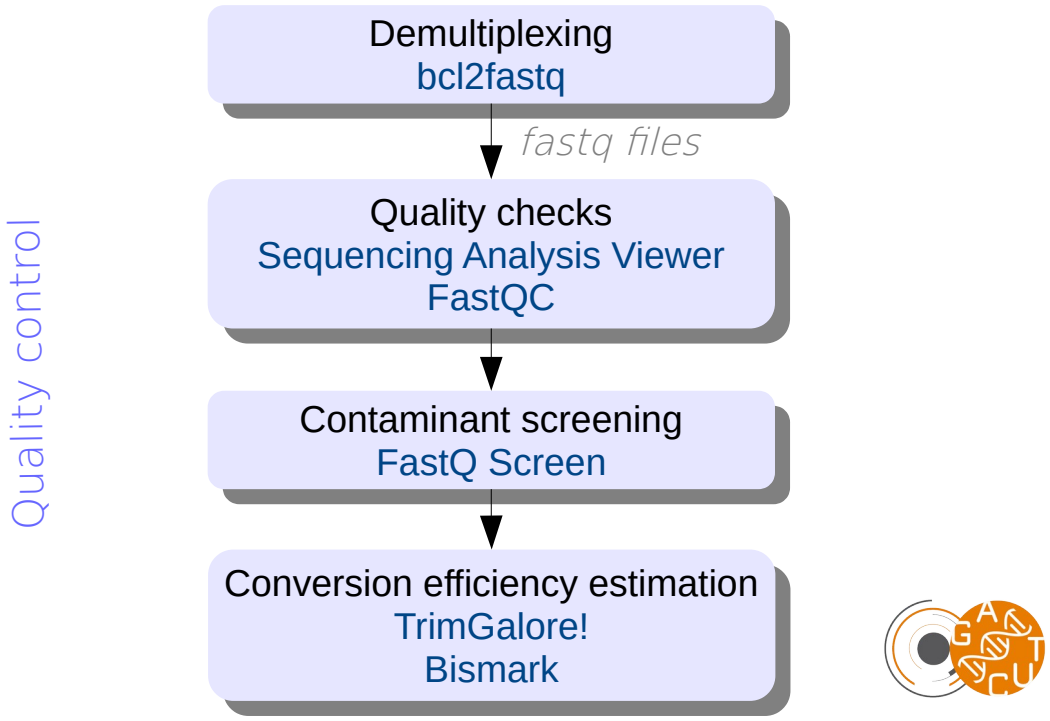
A partir des données brutes de séquençage des échantillons, le plateau technique réalise les étapes suivantes :

Production des fichiers fastq et démultiplexage des données
Le démultiplexage et la production des fichiers fastq est réalisée grâce au logiciel Illumina bcl2fastq.

Contrôle qualité des données
Le contrôle qualité s'appuie sur plusieurs critères :

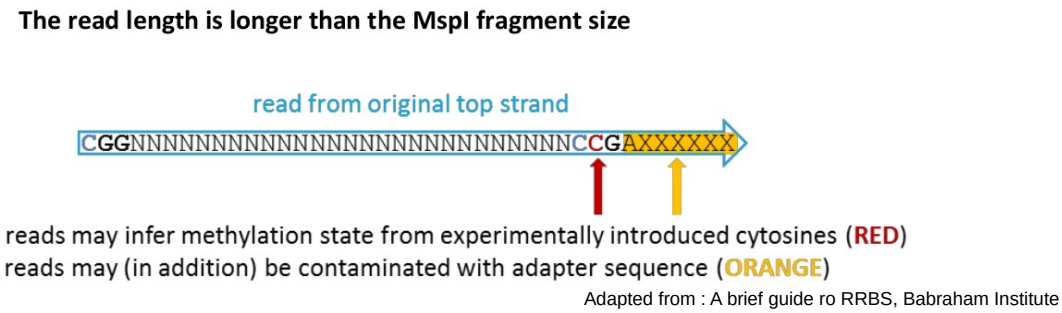
- validation du run en utilisant une série de critères associés,
- distribution des scores de qualité à chaque cycle,
- distribution des scores moyens de qualité par séquence,
- pourcentage de bases "N" par cycle,
- recherche de contaminants.

Avec le kit Diagenode, des contrôles méthylés et non méthylés sont ajoutés sous forme de *spike-in* lors de la préparation des banques. Cela permet d'ajouter une étape supplémentaire de contrôle qualité en calculant le taux de non-conversion, afin de vérifier que l'étape de conversion bisulfite a bien fonctionné.



Ces étapes sont réalisées systématiquement (bioinfo niveau 1). Les étapes suivantes sont réalisées sur demande (bioinfo niveau 2).

Nettoyage des séquences d'adaptateur
Les reads qui sont plus longs que le fragment MspI nécessitent une étape de nettoyage afin de supprimer l'adaptateur et les deux bases ajoutées artificiellement lors de la préparation des banques.

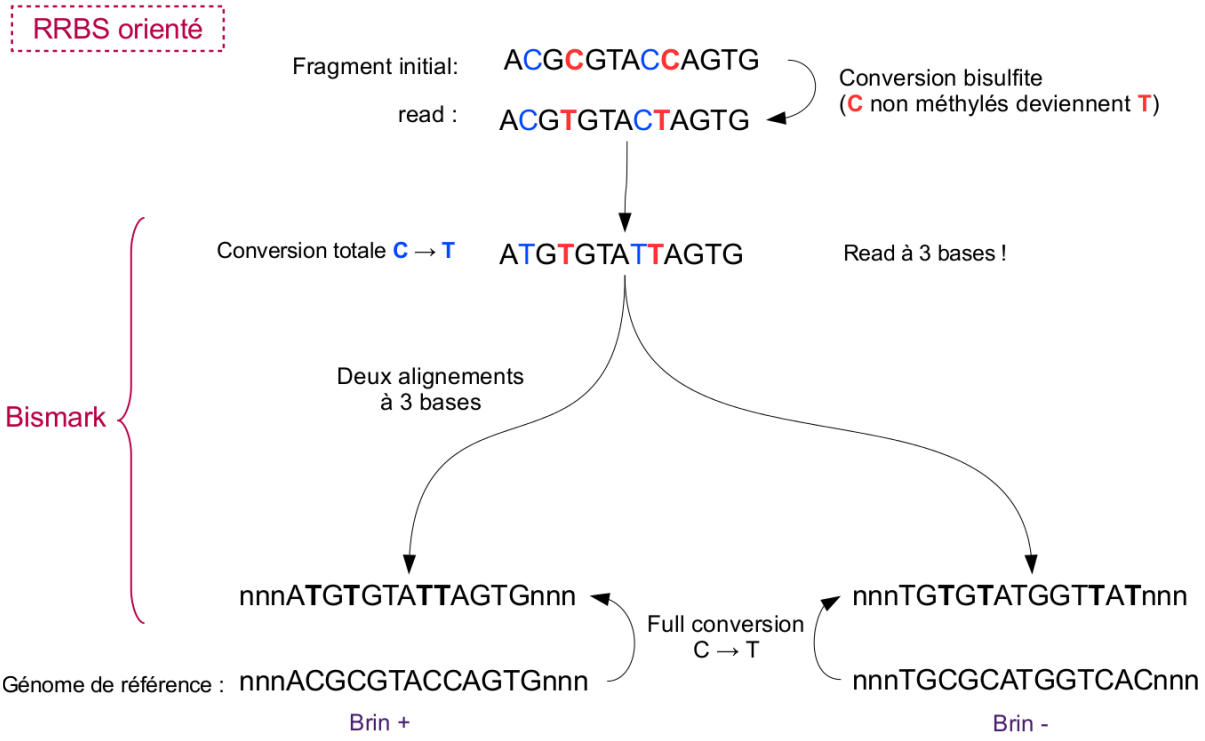


Adapted from : A brief guide ro RRBS, Babraham Institute

Pour réaliser cette étape, le logiciel Trim Galore! est utilisé.

Alignement des *reads* sur le génome de référence
Les séquences obtenues sont alignées sur le génome de référence (la version du génome doit être indiquée par le client) ou sur un ensemble de séquences fourni au format fasta. Nous utilisons le logiciel d'alignement Bismark, qui permet d'aligner des reads traités au bisulfite sur un génome de référence et de déterminer leur état de méthylation en une seule étape.

La particularité des reads issus d'un traitement au bisulfite est que les C initialement non méthylées sont transformées en T. Il n'est donc pas possible de faire un simple alignement sur le génome de référence. La figure suivante illustre le principe de Bismark.

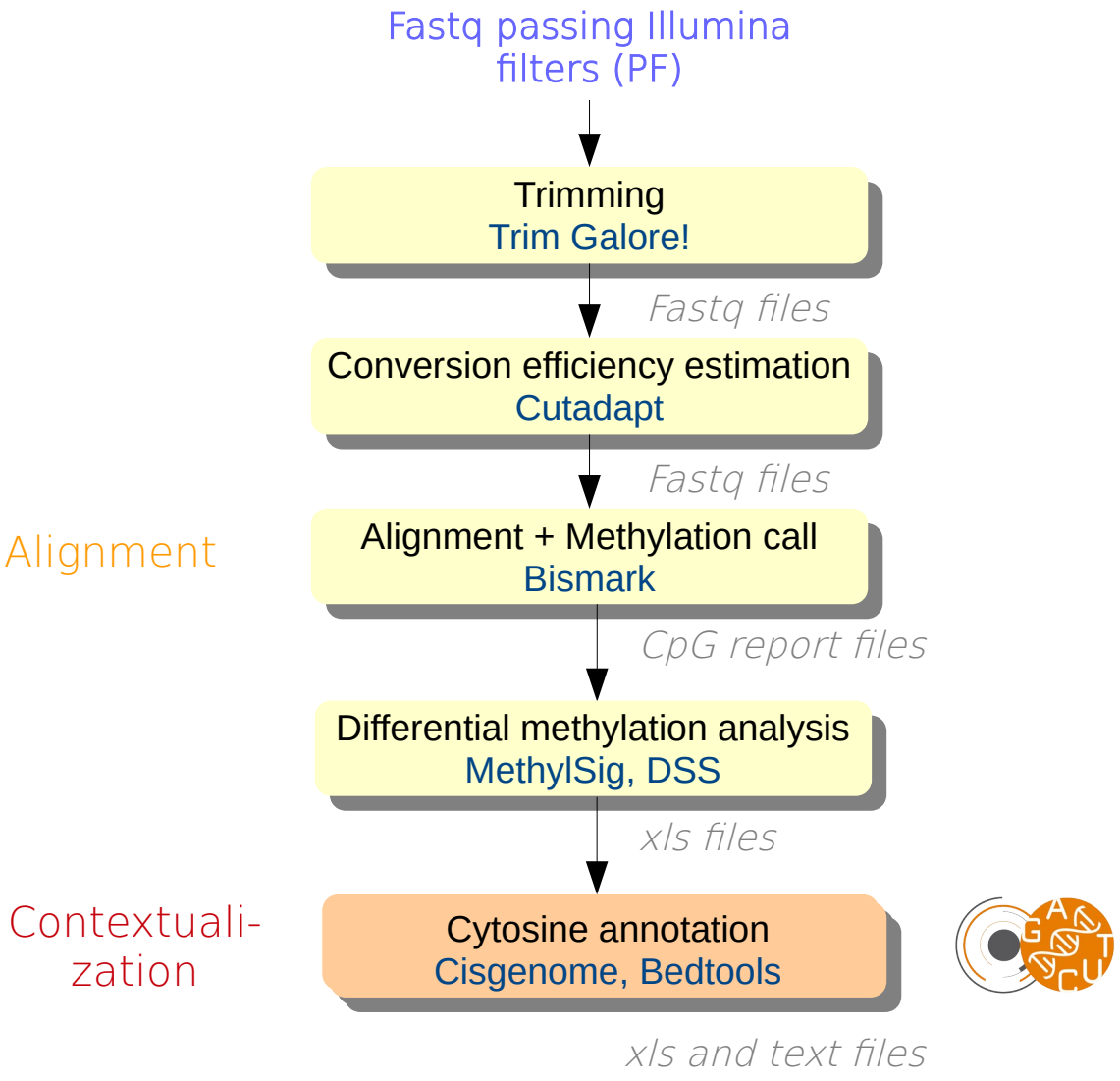


Bismark utilise aussi la séquence génomique initiale et le read pour déterminer l'état de méthylation.

Des fichiers au format BAM sont ainsi produits, ainsi que des fichiers détaillant l'état de méthylation des bases.

Identification des cytosines différentiellement méthylées
Ces analyses visent à identifier des cytosines (ou des régions) différentiellement méthylées, c'est-à-dire des cytosines dont le niveau de méthylation est différent entre deux conditions. Pour cela, nous utilisons les packages R MethylSig et DSS, qui utilisent une méthode statistique basée sur un modèle bêta-binomial et utilisent des réplicats pour estimer la variabilité biologique. DSS autorise l'analyse de données qui ne possèdent qu'un réplicat dans une des conditions biologiques.
A l'issue de l'analyse, un fichier contenant la liste des cytosines (ou des régions) différentiellement méthylées ainsi que les p-values associées est produit.

La figure ci-dessous illustre les étapes d'analyse réalisées sur demande (bioinfo niveau 2, cases en jaune) décrites ci-dessus, ainsi que les prestations complémentaires (bioinfo niveau 3, case en orange) détaillées dans le paragraphe suivant.



Prestations complémentaires

Des analyses complémentaires peuvent être effectuées selon la demande du client (bioinfo niveau 3).

Annotation des cytosines différentiellement méthylées
Cette étape implique qu'une annotation des gènes soit disponible (format gff ou gtf).
Une cytosine ou une région génomique peut être annotée en utilisant le gène le plus proche. Cela est calculé grâce à l'outil Bedtools closest.

Le logiciel Cisgenome est également utilisé pour produire un résumé des positions des cytosines ou régions différentielles (intergéniques, exons, introns etc).

Matériel initial

Les données nécessaires à l'analyse bioinformatique sont directement issues d'un des séquenceurs du plateau technique.

Restitution des résultats

A l'issue des analyses, plusieurs fichiers sont disponibles :

- un rapport d'analyse au format PDF disponible depuis le logiciel de gestion de projet redmine
- un fichier .fastq pour chaque échantillon : fichier texte contenant les séquences nucléotidiques ainsi que les scores de qualité correspondants
- un fichier .bam pour chaque échantillon aligné : le format bam est la version binarisée (compressée) du format sam. Ce fichier contient les résultats de l'alignement, et permet notamment leur visualisation avec un logiciel comme IGV ou SeqMonk.
- des fichiers contenant les niveaux de méthylation par base
- en cas d'analyse différentielle, des fichiers contenant la liste des cytosines (ou des régions) différentiellement méthylées ainsi que les p-values associées
- en cas d'annotation, la liste des cytosines, ou régions, annotées avec le gène le plus proche

L'ensemble de ces fichiers est disponible sur le serveur SFTP du plateau, à partir de la mise en ligne du rapport sur le gestionnaire de projet. Ce serveur est accessible par login et mot de passe, fournis avec le rapport d'analyse.

Durée de conservation des données

Les login et mot de passe pour accéder au gestionnaire de projet et les documents qui y sont mis en ligne n'ont pour l'instant aucune limite de validité.
En revanche, les fichiers qui sont mis en ligne sur le serveur sftp n'y sont hébergés que pour une durée de 10 jours à compter de l'édition du rapport de résultat.
Le dépôt des données brutes dans une base de données peut être demandé avant publication ; nous ne nous engageons pas à conserver ces données.