Mise au point d'un pipeline bioinformatique de détection de méthylation de l'ADN chez deux espèces d'arbres : le peuplier & le chêne

Abdeljalil Senhaji Rachik

AIX-MARSEILLE UNIVERSITÉ

MASTER BIO-INFORMATIQUE

Parcours type : Développement Logiciel et
Analyse des Données

2021-2022

Mise au point d'un pipeline bioinformatique de détection de méthylation de l'ADN chez deux espèces d'arbres : le peuplier & le chêne.

Auteur : Abdeljalil Senhaji Rachik

Encadré par : Odile Rogier
Isabelle Lesur Kupin
Abel Garnier

CEA - Institut de Biologie Francois Jacob - Centre National de la Recherche
en Génomique Humaine Laboratory for Epigenetics and Environment (LEE)
2 rue Gaston Crémieux CP5721 91057 Evry

Remerciements

Ce rapport de master 2 vient clôturer une belle expérience scientifique qui a été réalisée au sein de laboratoire de l'épigénétique et l'environnement CEA. Pour cela, je souhaiterai remercier :

- **Odile Rogier**, qui m'a dirigé, encadré, et guidé dans ce travail de recherche ; Son encadrement et ses qualités scientifiques m'ont permis de structurer mon travail.
- **Isabelle Lesur Kupin** pour votre encadrement, vos immenses savoir et vos précieux conseils, merci pour votre disponibilité et votre gentillesse.
- **Abel Garnier**, pour le soutien la motivation et surtout les aides au long de cette durée de stage, Merci pour votre capacité à faire partager votre savoir, merci également pour votre disponibilité.
- **Pr. Jorg Tost** qui m'a accueilli au sein de son laboratoire et permet de mener à bien mon stage. Je tiens aussi a remercié tous les membres de l'équipe.
- **Pr. Marc Villar**, le directeur de l'unité BioForA de m'avoir accueilli durant ce stage dans l'unité BioForA de l'INRAE, val-de-loire.
- Tous les collaborateurs du projet EpiTree et en particulier Pr. **Stéphane Maury**, pour leur motivation et leur bienveillance.

- **Nicolas Wiart** et **Elise Larsonneur** grand merci pour la meilleure formation « Calcul Haute Performance au TGCC/CCRT appliqué aux données génomiques ».

Lexique

° C : Degré Celsius

A : Adénine

DAG : Directed Acyclic Graph

ACP : Analyse en composantes principales

ADN : Acide désoxyribonucléique

ANR : Agence nationale de la recherche

BAM : Carte d'alignement binaire

BSC : Bisulfite Crick

BSCR : Complément inverse de BSC

BS-seq : Séquençage au bisulfite

BSW : Bisulfite Watson

BSWR : Complément inverse de BSW

C : Cytosine

CEA : Commissariat à l'Energie Atomique

CO2 : dioxyde de carbone

CP : Composante principal

CPU : Unités centrales de traitement

CSV : Valeurs séparées par des virgules

GIEC : Groupe d'Experts Intergouvernemental sur l'Évolution du Climat

HPC : Calcul haute performance

Mb : Mégabase

mC : Méthylation de cytosine

MC-seq : Séquençage de capture de méthyle

NGS : Séquençage de nouvelle génération

PCR : Réaction de polymérisation en chaîne

RRBS : Séquençage du bisulfite à représentation réduite

SNP : Polymorphisme nucléotidique unique

WGBS : Séquençage du bisulfite du génome entier

YAML (Yet Another Multicolumn Layout) : Encore une nouvelle structure multi-colonnes

Table des matières

Introduction 1
1. L'influence du changement climatique sur les arbres 1
2. L'Épigénétique chez les plantes 3
3. Présentation du projet d'étude 5
4. Objectifs du stage 7

Matériel et méthodes 9
1. Optimisation des conditions expérimentales MC-Seq 9
A. Échantillons utilisés 9
B. Description des conditions / paramètres testés 9
2. Mise en place du pipeline d'analyse de méthylation d'ADN 11
A. Présentation globale du pipeline 11
B. Les étapes du pipeline 13
3. Analyses statistiques des tests d'optimisation du MC-Seq 19
A. Structuration des données 19
B. Analyse des séquences capturées 19
C. Analyses statistiques de methylome 21

Résultats 21
1. Validation du Pipeline 23
A. Résultats « bruts » de sortie 23
B. Fonctionnalités du pipeline 25
2. Exploitation des données testant les conditions expérimentales MC-Seq 27

Discussion 29
Perspectives 37

Conclusion 39

Références : 41

Introduction

1. L'influence du changement climatique sur les arbres

Depuis la formation de la Terre, les modifications du climat sont un facteur clé dans son histoire, influençant l'évolution des espèces et l'extinction de certaines d'entre elles. Le réchauffement climatique est un dérèglement du climat au niveau planétaire dû à une augmentation de la concentration des gaz à effet de serre (généralement le CO2) dans l'atmosphère. Ceci aboutit à un changement de température globale de la planète. Depuis 1990, la production mondiale de gaz à effet de serre a augmenté de près de 40%. En 2013, le GIEC (le Groupe d'Experts Intergouvernemental sur l'Évolution du Climat) confirme le réchauffement observé depuis 1950. Il est extrêmement probable (à 90%) que l'influence humaine sur le climat ait été la cause majeure du réchauffement observé depuis le milieu du vingtième siècle [1]. Le GIEC prévoit une augmentation significative du réchauffement de 1,5°C (2,7°F) entre 2030 et 2052 si la température globale continue à augmenter au rythme actuel [2]. Il a aussi mentionné, en 2017, que l'augmentation des températures avait dépassé le seuil de 1°C par rapport à la période préindustrielle (figure 1) [3]. Le réchauffement climatique a des conséquences sur la fonte des glaces et sur l'augmentation de la fréquence et de l'intensité des phénomènes météorologiques extrêmes tels que des vagues de chaleur, des sécheresses plus importantes ainsi que des précipitations plus fortes [4]. Ces derniers impactent la biodiversité et réduisent la productivité des écosystèmes naturels [5].

Dans ce contexte, les changements environnementaux, en particulier l'augmentation de la température et de la sécheresse, affectent directement l'écosystème et deviennent un enjeu majeur dans le développement et l'entretien des forêts et des arbres. En effet, l'impact des changements climatiques est particulièrement dévastateur sur les arbres forestiers alors qu'ils jouent un rôle écologique clé dans l'environnement terrestre [6]. Une méta-analyse publiée en 2016 montre que la mort des forêts est un phénomène mondial et qu'elle est principalement liée à la sécheresse et à l'élévation de la température (figure 2) [7]. Étant des espèces à longue, voire très longue espérance de vie, les arbres ne peuvent pas migrer et c'est uniquement lorsque leur pollen ou leurs graines sont dispersés par le vent ou les animaux que leur descendance est déplacée. Ils doivent donc subir ces changements d'environnement tout en cherchant à s'adapter. Il a été démontré que les arbres survivants peuvent réagir et s'adapter à ces changements environnementaux grâce à des mécanismes génétiques complexes [8]. En raison de leur grande variabilité génétique et de leur plasticité phénotypique importante, ils sont très sensibles aux variations environnementales mais ont un fort potentiel d'adaptation [9,10].

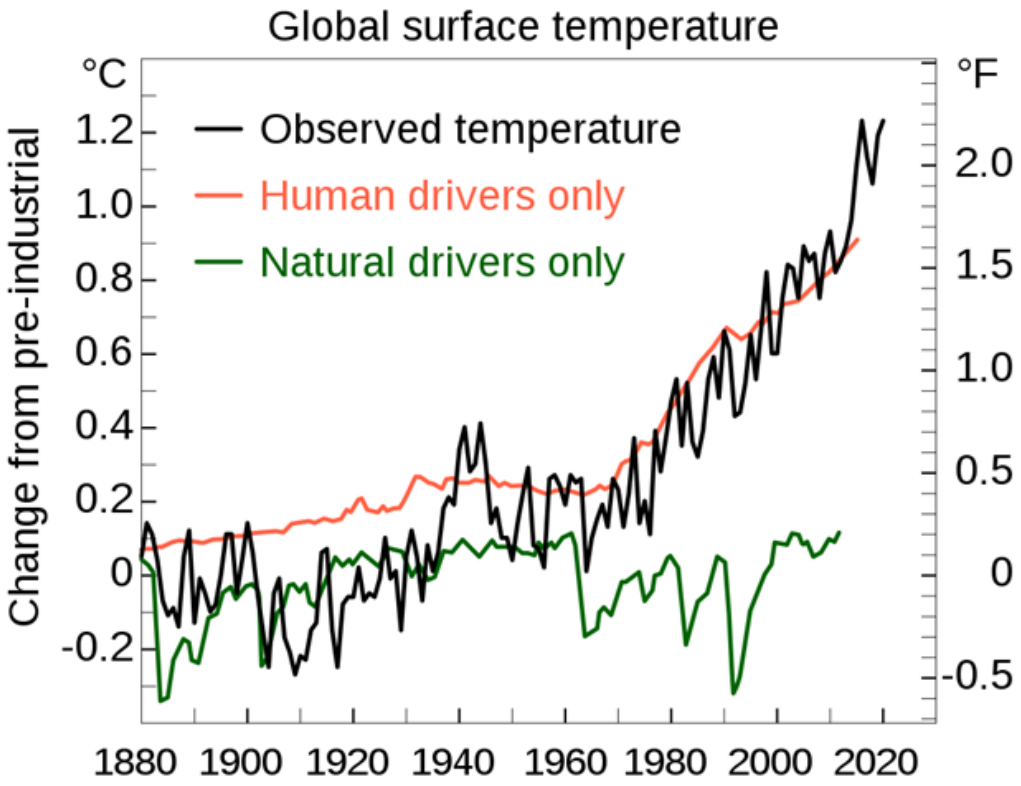


Figure 1 : Températures observées par la NASA de 1880 à nos jours. La moyenne des températures de 1850–1900 est utilisée comme température de référence préindustrielle [3].

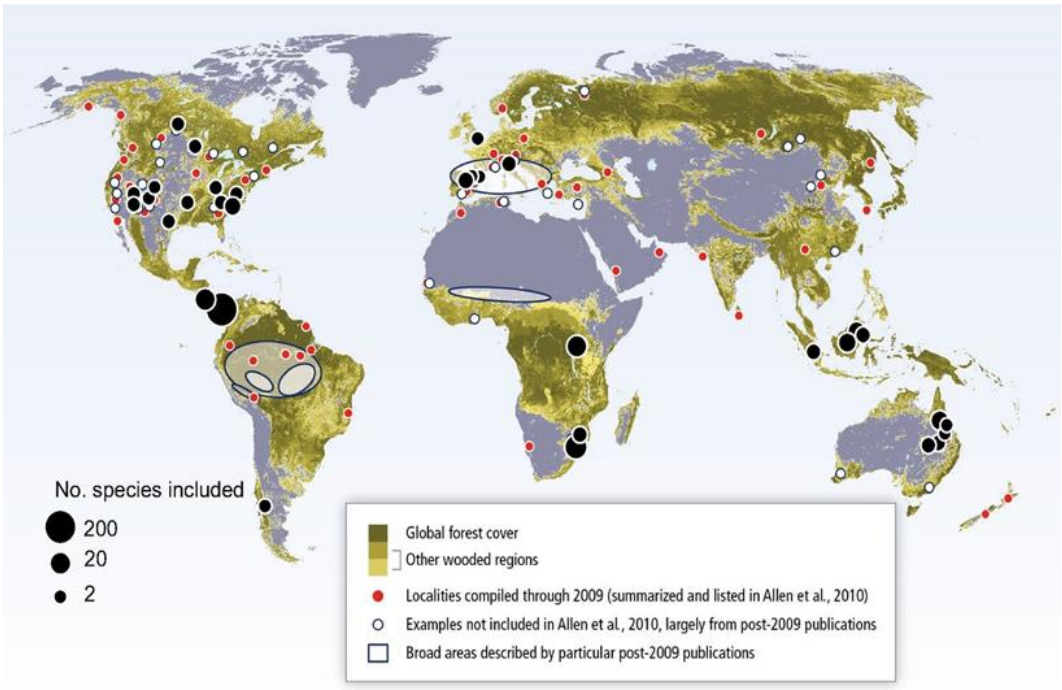


Figure 2: Événements régionaux de mortalité des arbres due à la sécheresse et à la chaleur dans le monde [7].

Les peupliers sont des plantes pérennes du groupe des angiospermes, de la famille des Salicacées et du genre *Populus*. Ce genre regroupe 35 espèces de peupliers, qui sont considérées comme espèces pionnières, car possédant une croissance extrêmement rapide avec une forte capacité de multiplication végétative [13]. Avec une répartition très large dans l'écosystème, elles couvrent la quasi-totalité de l'hémisphère Nord. Dans le cadre du réchauffement climatique, les peupliers ont de grands besoins en eau et une faible tolérance à la sécheresse. Le peuplier *Populus trichocarpa* a été la première espèce d'arbre dont le génome a été entièrement séquencé. Le génome a été publié en 2006 [14]. La dernière version de l'assemblage (V.4.0) contient 392,2 Mb (millions de bases) réparties sur 19 chromosomes et 41335 gènes. Plusieurs versions qui diffèrent par la qualité d'assemblage de son génome sont publiées. [15]

Les chênes (genre Quercus) appartiennent à la famille des Fagacées. Ils se répartissent dans les régions tempérées de l'hémisphère Nord. On dénombre 465 espèces de chênes. Ces dernières font partie des espèces forestières les plus polymorphes au niveau génétique, d'où leur forte capacité d'adaptation [16]. Le génome a une taille de 740 Mb avec 12 paires de chromosomes et il possède 25,808 gènes [17].

Grâce à leur diversité génétique et leur potentiel d'adaptation aux changements climatiques, ces deux espèces sont considérées comme des modèles d'étude pertinents.

La plupart des changements de conditions environnementales ont une incidence sur la fréquence des recombinants homologues et des réarrangements alléliques. On observe également des modifications sur un grand nombre de marques épigénétiques telles que la méthylation de l'ADN (Acide DésoxyriboNucléique) qui peuvent affecter l'expression des gènes et l'activité des éléments transposables et autres séquences répétées, ce qui altère la plasticité phénotypique des plantes [18] [19].

2. L'Épigénétique chez les plantes

L'épigénétique est l'étude des processus qui influencent l'expression des gènes au niveau moléculaire sans modifier la séquence d'ADN et qui peuvent générer des modifications potentiellement héritables dans le phénotype. L'épigénomique consiste à étudier des mécanismes moléculaires altérant l'expression des gènes et l'activité des éléments transposables dans le cadre de la cellule ou de l'organisme (Figure 3) [20, 21].

L'épigénétique agit essentiellement au niveau de la chromatine. L'ADN peut prendre deux formes : l'euchromatine (structure relâchée) et l'hétérochromatine (structure compacte) [22]. Selon l'état de la chromatine, la transcription des gènes peut être différente car les facteurs de transcription ont assez de place pour intervenir (structure relâchée) ou pas (structure compacte) [23, 24].

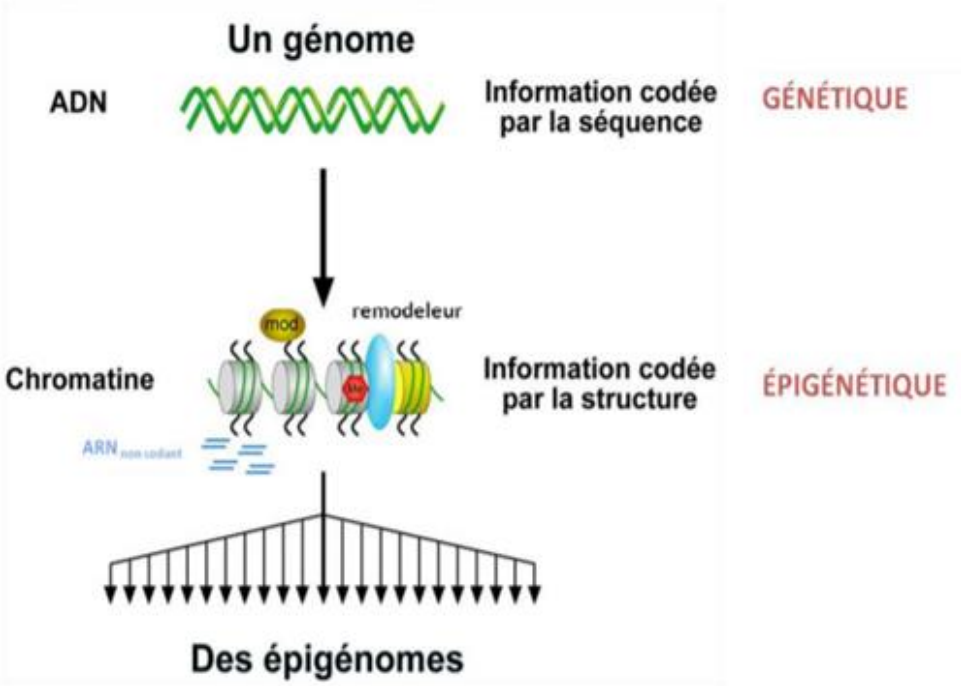


Figure 3 : Schéma représentant le concept moderne de l'épigénétique. Avec un génome, plusieurs états de la chromatine peuvent exister [19].

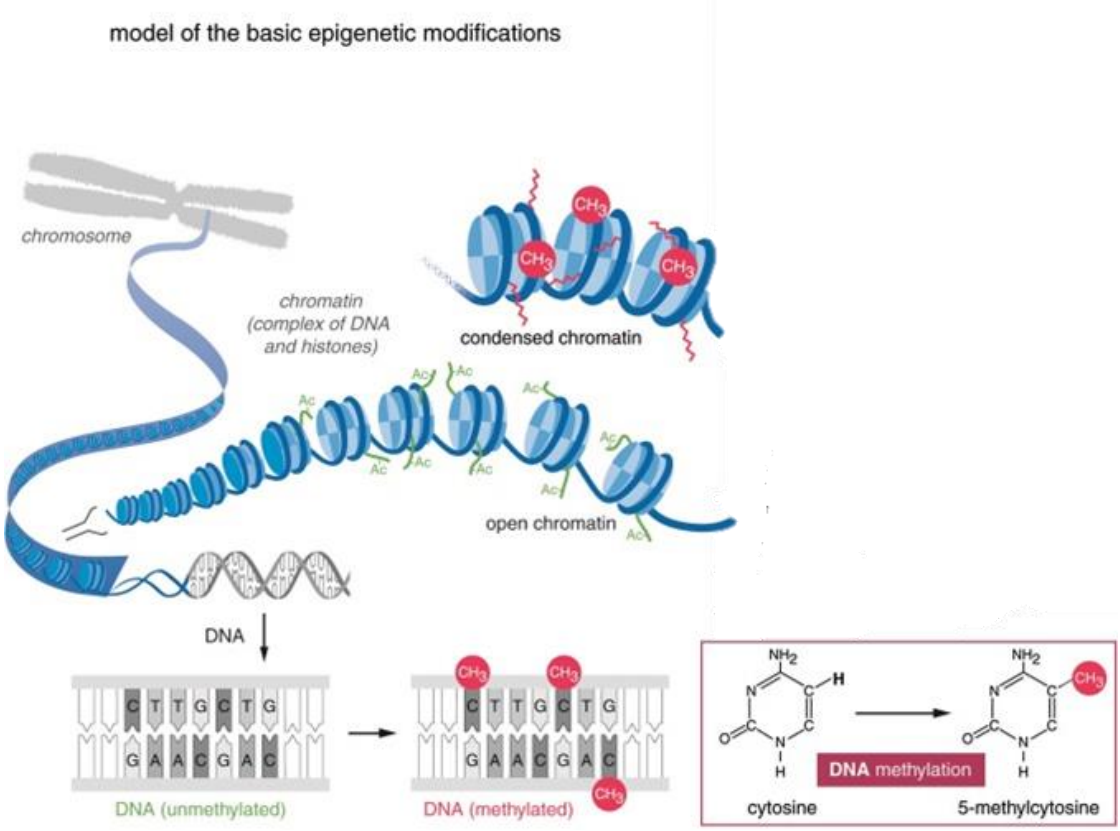


Figure 4 : Modèle de modifications épigénétiques de la chromatine par méthylation de l'ADN, modification des histones [25].

Une autre marque épigénétique correspond à la méthylation de l'ADN, qui correspond à l'ajout d'un groupe méthyl (CH3) sur le carbone de la position 5' d'une cytosine au lieu d'un atome d'hydrogène pour former la 5-méthylcytosine (5mC) comme cinquième base de l'ADN (Figure 4) [25]. On retrouve la méthylation de l'ADN chez les plantes dans trois contextes nucléotidiques différents : CG, CHG et CHH (où H correspond soit à A, C ou T) [27, 28, 29, 30]. Les plantes se distinguent par une alternance de domaines méthylés et non méthylés, existant à la fois au niveau des gènes et des régions intergéniques, ce qui constitue un profil de « mosaïque de méthylation » [30]. La méthylation des sites CG est la plus abondante, ces sites pouvant atteindre un niveau de méthylation de 80 à 100%, tandis que la méthylation des sites CHH excède rarement 10%. Le niveau de méthylation des sites CHG varie de 20 à 100%. Ce taux de méthylation dans chaque contexte de méthylation est variable entre les espèces [31]. La méthylation de l'ADN est une modification épigénétique qui joue un rôle important dans la régulation de l'expression des gènes et donc dans un large éventail de processus cellulaires [32, 33]. Une forte méthylation dans la zone promotrice d'un gène est souvent liée à une faible expression de ce gène.

De nos jours, il existe différentes techniques pour étudier la méthylation de l'ADN : MBDcap-seq, MeDIP-chip, MeDIP-seq, RRBS, WGBS, MC-Seq, ... (Figure 5) [32].

Parmi elles, le WGBS (whole-genome bisulfite sequencing) ou BS-Seq (Bisulfite-Sequencing) est une technologie de séquençage de nouvelle génération du génome complet qui permet de déterminer l'état de méthylation de l'ADN à l'aide de bisulfite de sodium (figure 6).

Une seconde approche, le MC-Seq, permet de ne séquencer qu'une région spécifique du génome avec un traitement au bisulfite de sodium préalable. Moins chère qu'un séquençage complet de tout le génome, cette technique permet à la fois d'analyser plus d'individus tout en augmentant la couverture de séquençage pour un coût global équivalent [34]. Cette approche par capture nécessite la synthèse de sondes qui vont cibler les régions génomiques souhaitées.

3. Présentation du projet d'étude

L'épigénétique jouerait un rôle important dans la plasticité phénotypique et l'adaptation des arbres [35, 36, 37]. De ce fait, ces dernières années, plusieurs publications ont étudié le rôle que l'épigénétique pourrait jouer dans la réponse adaptative des arbres aux variations environnementales [38, 39, 40, 41].

C'est dans ce contexte que le projet ANR EpiTree (https://www6.inra.fr/epitree-project/) a vu le jour. Il a pour objectif d'étudier la réponse des arbres aux changements climatiques en identifiant les liens qui existent entre les marques de méthylation de l'ADN (épigénétique), l'expression génique (transcriptomique) et la variation allélique (SNP) (Figure 7) [42].

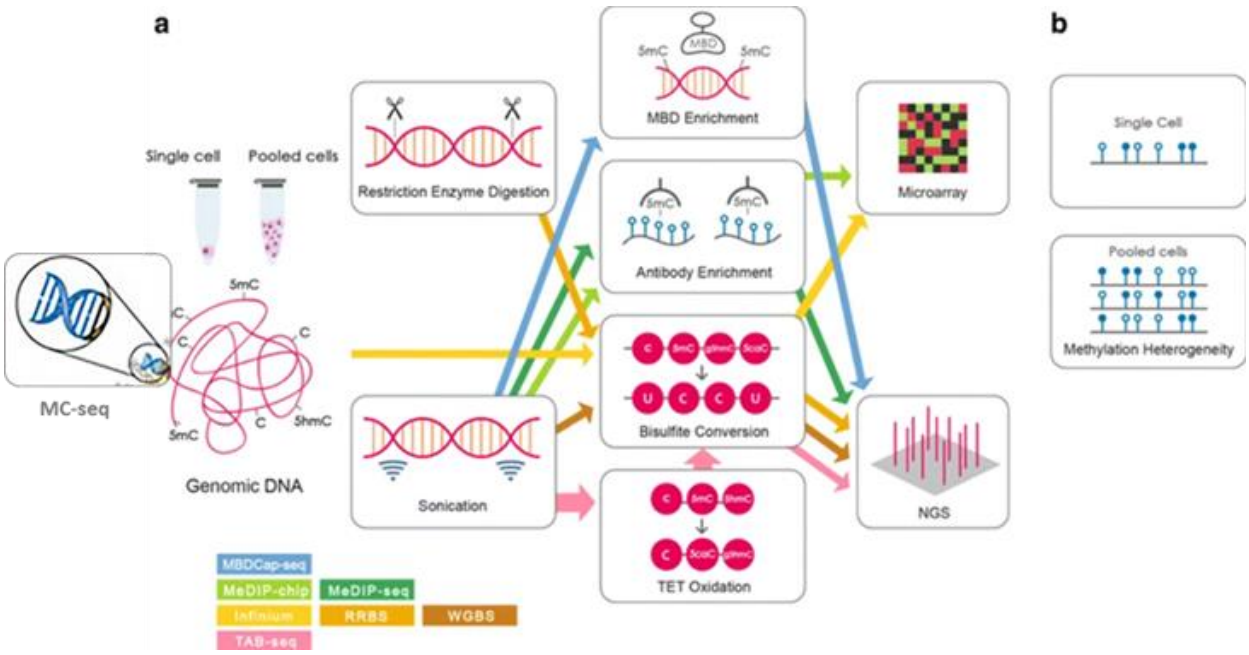


Figure 5: Méthodes couramment utilisées pour analyser la méthylation de l'ADN à l'échelle du génome. (a) Les procédures incluent une étape de fragmentation de l'ADN génomique avant que les méthylations soient détectées et quantifiées par micro puce (Microarray) ou séquençage NGS. (b) La quantification des méthylations et leur répartition sur l'ADN sont analysées au niveau cellulaire ou au niveau d'une population de cellules. (Figure issue de [32] et adapté pour MC-Seq)

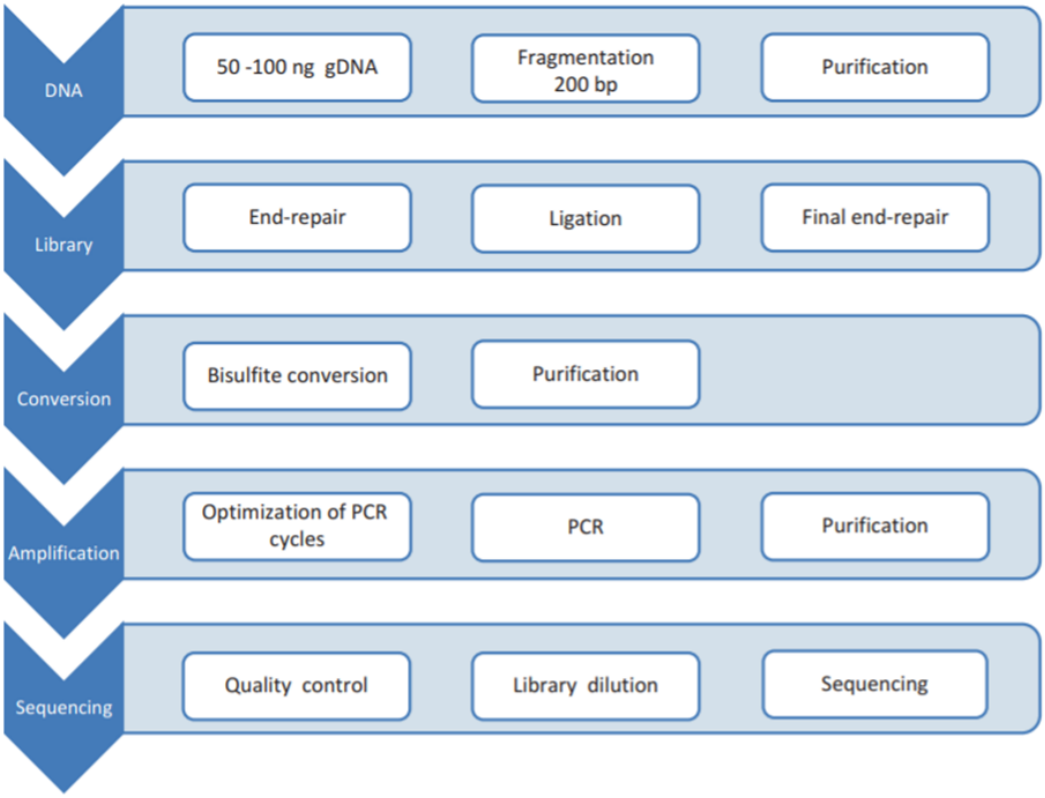


Figure 6 : Flux de travail du protocole de séquençage du bisulfite pour le génome entier [32].

Dans ce projet, deux espèces d'arbres possédant de grandes ressources génomiques et des avantages économiques et écologiques remarquables sont étudiées : le peuplier et le chêne. Deux principales contraintes environnementales sont considérées : la température (impact sur le modèle du chêne) et l'humidité du sol (impact sur le modèle du peuplier).

Un des objectifs majeurs du projet est d'étudier l'effet de la méthylation de l'ADN sur l'adaptation au milieu d'environ 500 individus peupliers et chênes. Compte tenu du nombre important d'individus à étudier, la technique de WGBS n'est pas possible. En effet, son coût serait trop important pour étudier les niveaux de méthylation de l'ADN à l'échelle du génome entier pour autant d'individus. Pour cette raison, il a été décidé de se concentrer sur une partie du génome seulement, constituée de gènes d'intérêt et de régions montrant une grande variabilité au niveau de la méthylation entre individus, en utilisant la technique MC-Seq ou séquençage par capture. Pour identifier ces zones, au début du projet, un séquençage WGBS a été effectué sur 20 peupliers et 10 chênes. Il a permis l'identification de zones de méthylation très variables entre ces individus. A ces zones d'intérêt s'ajoutent des gènes candidats impliqués dans la résistance à la sécheresse, par exemple. L'ensemble de ces régions correspond à 24Mb de séquences d'intérêt qui vont être capturées et séquencées selon le protocole MC-Seq chez environ 250 peupliers et 250 chênes.

De plus, afin d'optimiser les coûts, les partenaires du projet EpiTree ont mis en place une expérience préalable à l'analyse des 500 individus. L'objectif est ici de tester des conditions non-optimales et de les comparer à celles recommandées par le fabricant de réactifs Agilent, ce qui nous permettrait encore de diminuer le coût de l'étude. Un même individu (peuplier) a été testé dans plusieurs conditions expérimentales. Parmi les paramètres que nous avons fait varier, notons la dilution des sondes, la qualité de l'ADN, le nombre de cycles de PCR ou encore la technique de fragmentation de l'ADN, l'objectif étant d'identifier l'impact de chaque paramètre sur les résultats.

4. Objectifs du stage

Mon stage s'inscrit donc dans le cadre du projet EpiTree et plus particulièrement dans le traitement des données de méthylations du peuplier. Lors de l'analyse des 30 WGBS, la détection des sites méthylés a été un peu fastidieuse car elle a été réalisée sous l'environnement Galaxy et a nécessité plusieurs étapes successives de traitement des données, chaque étape nécessitant une intervention humaine. Cependant, une telle organisation ne semblait pas réaliste pour plus de 500 individus.

La définition et la mise en œuvre d'un pipeline d'analyse bio-informatique adapté à l'étude de la méthylation de l'ADN représente donc un réel besoin pour les partenaires du projet EpiTree. Ce pipeline doit présenter deux caractéristiques principales.

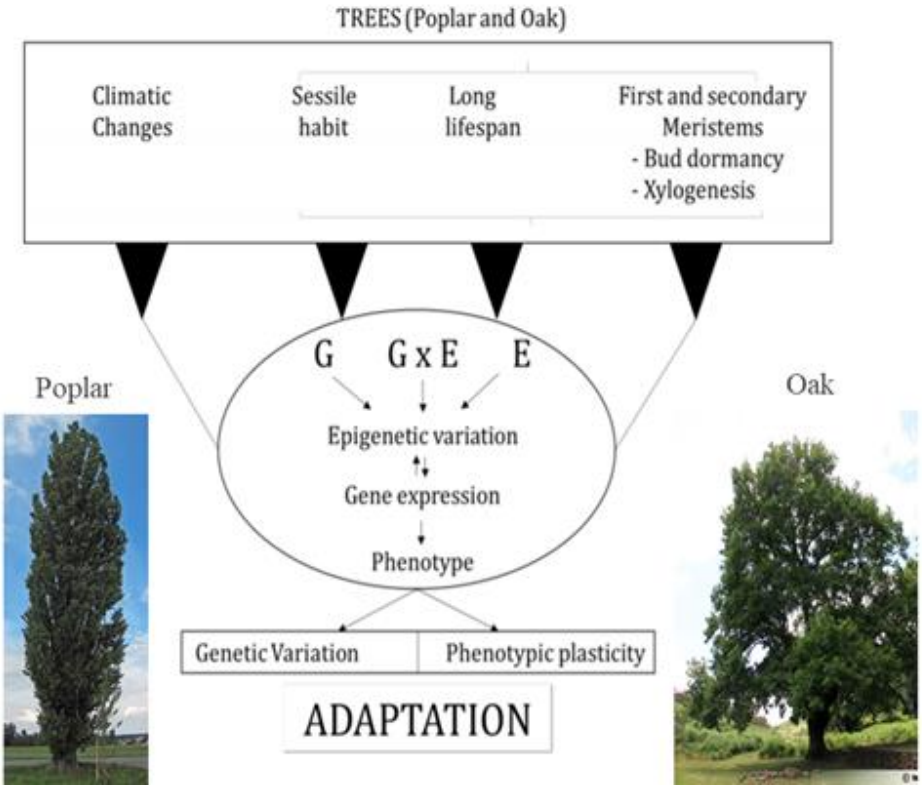


Figure 7 : Un aperçu du projet EpiTree qui étudie les liens entre l'environnement de 2 espèces d'arbres, le peuplier et le chêne et leur variabilité génétique, avec la variation épigénétique en particulier qui a un rôle dans l'adaptation de ces espèces à leur environnement [42].

Il doit, dans un premier temps, permettre aux utilisateurs d'enchaîner les différentes étapes d'analyse des données de séquençage de façon automatisée tout en permettant de réaliser ces analyses sur un grand nombre d'individus. Il doit ensuite être le plus robuste et automatisable possible, avec des performances optimisées sur le cluster de calcul du CEA (Commissariat à l'Énergie Atomique). De plus, pour la poursuite du projet, il est nécessaire de vérifier la faisabilité de l'approche MC-Seq et de définir les conditions expérimentales répondant à nos critères. Je dois donc, dans un premier temps, mettre en place un pipeline bioinformatique permettant d'enchaîner toutes les étapes de détection des méthylations dans les données de séquençage via un outil dédié à la gestion des pipelines (snakemake). Je vais, dans un deuxième temps, utiliser ce pipeline sur les données de séquençage MC-Seq afin de comparer les résultats obtenus (sur le même individu) à la fois par WGBS et par MC-Seq. Enfin, il faudra comparer entre elles les modalités expérimentales (dilution des sondes, fragmentation ADN, quantité d'ADN, …) afin d'identifier les conditions non-optimales qui permettent toutefois d'obtenir un séquençage suffisant pour détecter les niveaux de méthylation dans les zones d'intérêt.

Matériel et méthodes

1. Optimisation des conditions expérimentales MC-Seq

Nous souhaitons donc tester plusieurs conditions expérimentales qui diffèrent du protocole officiel d'Agilent. Notre objectif est double : 1/ vérifier que les niveaux de méthylations calculés sur les zones d'intérêt MC-Seq sont équivalents à ceux calculés sur ces mêmes zones mais avec un séquençage WGBS (génome complet) et 2/ identifier les conditions expérimentales optimisables qui nous permettraient, tout en garantissant la qualité des résultats, de réduire les coûts de l'expérience.

A. Échantillons utilisés

Les données analysées dans ce test sont issues du séquençage de tissus de jeune xylème et cambium d'un même arbre DRA-038, un peuplier noir (*P. nigra*). Cet individu de référence a été séquencé à la fois en WGBS et en MC-Seq. Pour le séquençage par capture, plusieurs conditions expérimentales sont testées (décrites ci-dessous) menant à 18 échantillons (14 échantillons de DRA-038 et 4 échantillons d'autres peupliers) (figure 8).

B. Description des conditions / paramètres testés

La société Agilent (qui produit les sondes) a plusieurs recommandations concernant le volume de sondes, la quantité d'ADN à utiliser (3000 ng), la quantité d'ADN à obtenir après fragmentation (350 ng) ... Compte tenu du coût des réactifs, il nous est impossible d'analyser la totalité de nos échantillons en WGBS. Cependant, dans les conditions préconisées par Agilent, le coût des expériences MC-Seq reste trop élevé. Nous avons donc décidé de tester plusieurs conditions expérimentales (figure 8) pour

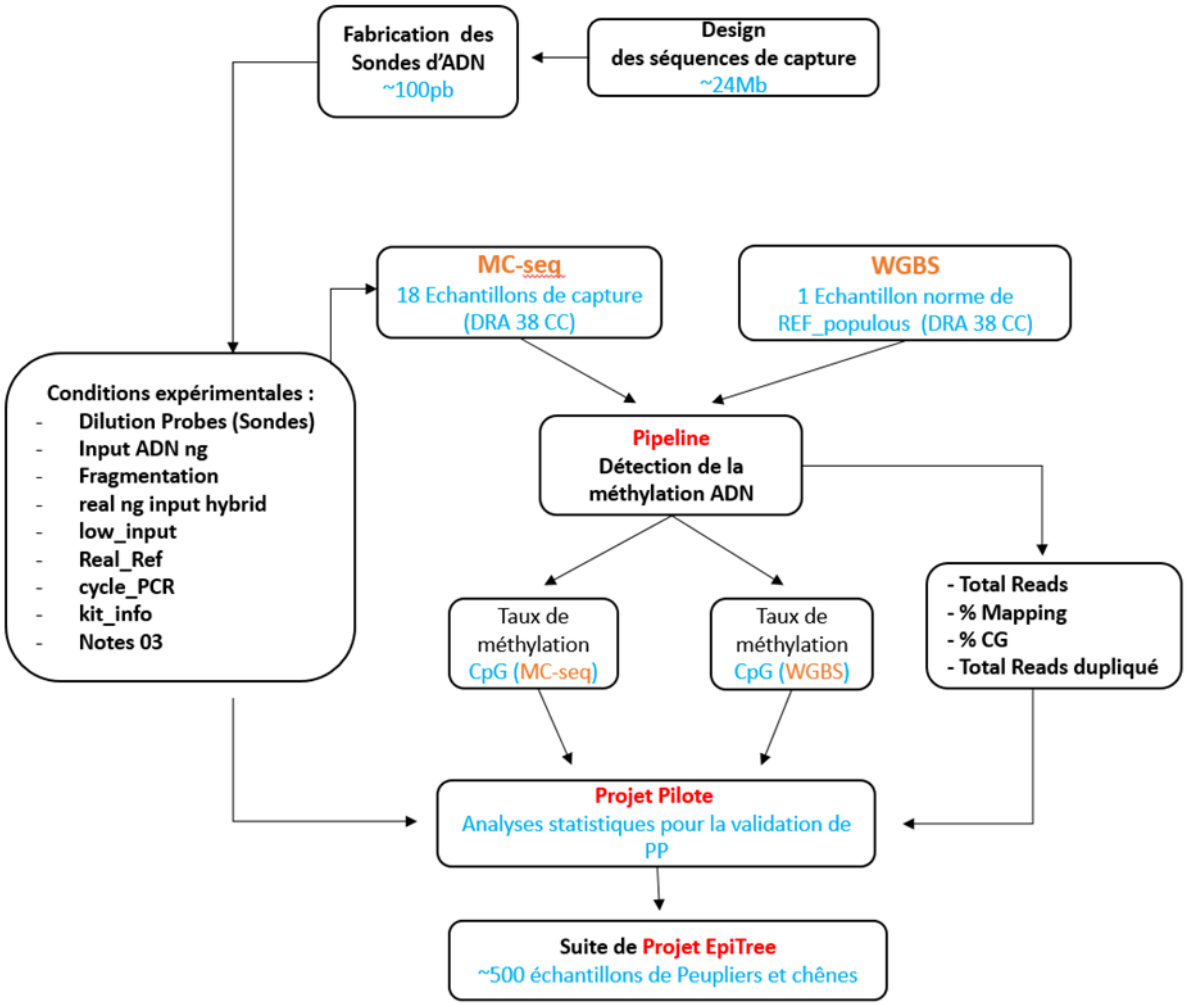


Figure 8 : Procédure d'analyse des données produites pour l'optimisation des conditions environnementales MC-Seq.

réduire le coût de l'approche MC-Seq en agissant sur la dilution des sondes, la technique de fragmentation, la quantité initiale d'ADN, la qualité de l'ADN, la quantité d'ADN après fragmentation, … (Annexe 1).

La dilution des sondes représente une condition clé pour l'optimisation du coût de préparation, la fabrication des sondes étant l'étape la plus coûteuse. Quatre conditions de dilution ont été testées : 1, 1/8, 1/10, 1/16. La qualité de l'ADN constitue également un paramètre qui doit être testé car nous ne sommes pas sûrs d'avoir assez d'ADN de bonne qualité pour tous les individus que nous voulons étudier. Plusieurs quantités initiales d'ADN ont été testées : 3000, 1000, 750, 600 et 500 ng. La fragmentation de l'ADN consiste à couper l'ADN en morceaux qui seront ensuite hybridés avec les sondes Agilent. Deux modes de fragmentation d'ADN ont été testés : la sonication Covaris et la fragmentation enzymatique. De la même façon, l'impact de la quantité d'ADN fragmenté utilisée lors de l'hybridation avec les sondes a été testée. Enfin, 4 individus ont été rajoutés en plus de l'individu DRA-038 car ils ont été récoltés depuis plus longtemps que notre individu de référence : leur ADN congelé pourrait donc être moins intègre et sera représentatif d'une partie des individus qui seront séquencés par la suite et dont l'ADN sera très certainement dégradé. Des informations complémentaires (répétition & low input, kit périmé et cycle de PCR) ont été rajoutées dans le tableau récapitulatif des conditions pour voir si elles avaient une incidence sur les résultats.

2. Mise en place du pipeline d'analyse de méthylation d'ADN

Afin de pouvoir enchaîner les étapes d'analyse des données de séquençage obtenues par la technique du MC-Seq, j'ai mis en place un pipeline d'analyse de données de méthylation de l'ADN. Étant donné qu'un grande nombre d'individus sera étudié avec cette approche, il est important de pouvoir automatiser les analyses.

A. Présentation globale du pipeline :

Je suis parti des outils précédemment utilisés sous Galaxy par les équipes du projet pour analyser les données WGBS (ayant permis la mise au point de la capture). Pour lancer de manière automatique tous ces outils, le pipeline est implémenté via le gestionnaire de flux d'analyses (workflow) snakemake (figure 9) [43] [44]. La mise en place de notre pipeline a été réalisée sur le cluster de calcul du CEA qui utilise l'ordonnanceur SLURM. Ce dernier gère le lancement de chaque tâche en choisissant les nœuds de calcul les plus appropriés tout en évitant que les ressources soient surchargées. Développer notre pipeline sur un cluster de calcul nous permet de paralléliser les tâches et donc de diminuer le temps des analyses. Nous avons développé notre pipeline afin d'optimiser l'utilisation des ressources disponibles ou Central Processing Units (CPU) sur le cluster du CEA en parallélisant les tâches. Nous avons aussi voulu rendre son utilisation la plus intuitive possible afin de permettre son utilisation par le plus grand nombre d'utilisateurs.

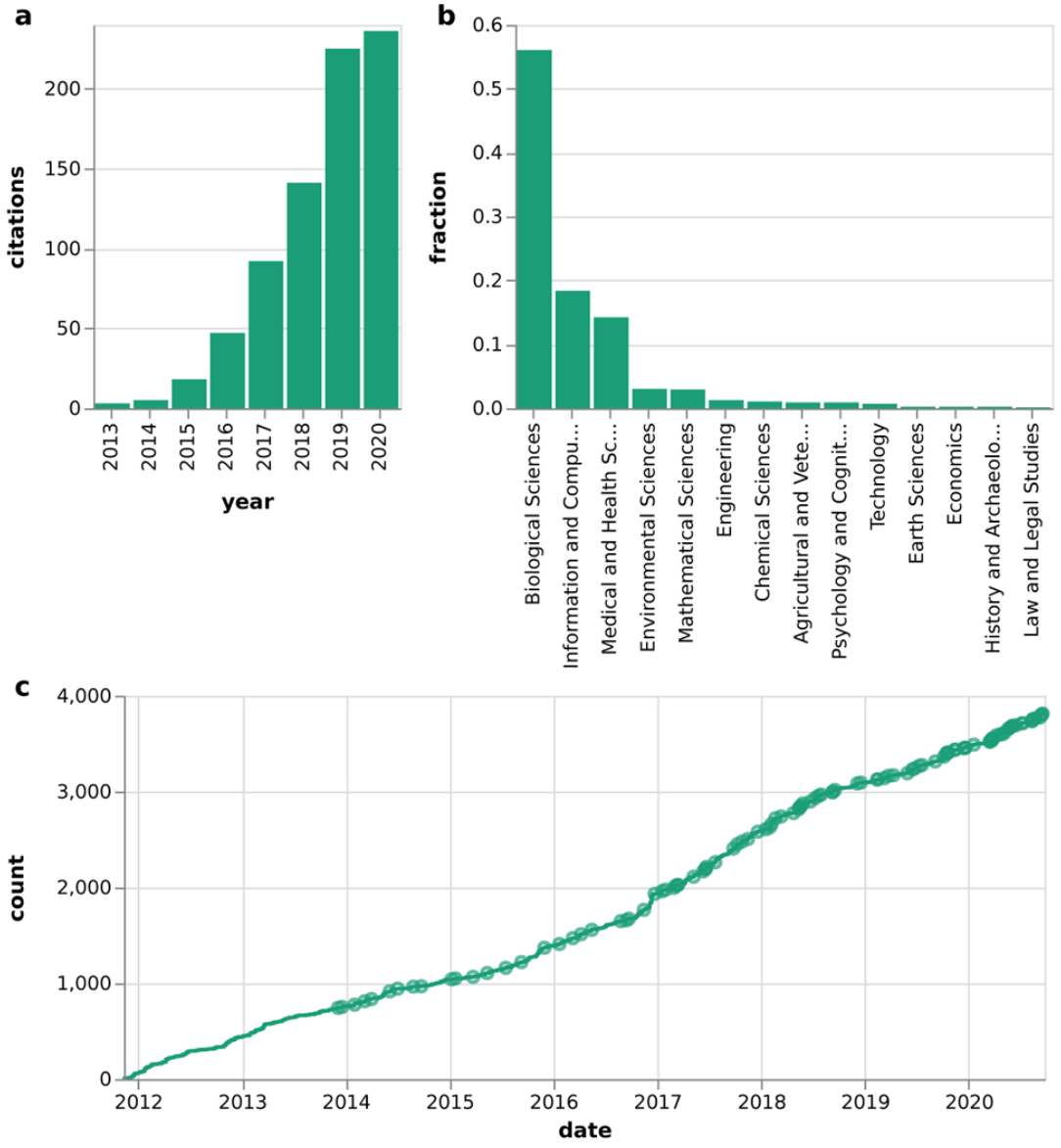


Figure 9: Etudes publiées utilsantSnakemake. (a) Evolution du nombre d'études citant l'article original de Snakemake [43] entre 2013 et 2020. (b) Répartition des articles citant Snakemake par discipline scientifique. (c) Evolution du nombre cumulé de commits git entre 2012 et 2020 [44].

Un flux d'analyse snakemake est défini en spécifiant des règles dans un Snakefile et ses paramètres de lancement dans un fichier de configuration. Les règles décomposent le flux d'analyse par étapes (briques) et spécifie les fichiers à utiliser en entrée et ceux à produire en sortie de chaque brique. Le fichier de configuration, écrit en YAML (plus lisible que Json), définit les divers paramètres des outils (comme l'emplacement du fichier contenant le génome de référence et les données brutes ou le chemin du répertoire utilisé pour le processus d'analyse) (figure 10). Une fois ces paramètres renseignés, le pipeline peut être lancé à travers un script bash qui contient lui-même une configuration du cluster et la ligne de commande permettant le lancement du pipeline. Ce dernier effectue alors l'analyse des données en 5 étapes décrites ci-dessous.

B. Les étapes du pipeline

Chaque étape du pipeline correspond à une brique décrite dans le fichier YAML. Elle est constituée soit de commandes Shell soit de commandes R qui vont permettre de réaliser une analyse. Pour chaque brique, le nombre de fichiers nécessaires en entrée, le nombre de fichiers générés et leurs formats sont différents. Le pipeline de détection de méthylation d'ADN comporte six étapes principales : (1) le nettoyage des données brutes suivi d'un contrôle qualité, (2) l'alignement contre un génome de référence, (3) la suppression des duplications, (4) la détection des cytosines méthylées (mC) dans les 3 contextes de méthylation, (5) leur extraction par contexte et (6) quelques analyses statistiques de base sur la détection des méthylations (figure 11).

B.1 Nettoyage des données brutes suivi d'un contrôle qualité

La première étape du pipeline consiste à nettoyer les données brutes de séquençage. Les fichiers d'entrée sont au format FASTQ : pour chaque base séquencée, un score de qualité (score Phred) est associé. Il est basé sur la probabilité qu'une base soit incorrecte. Les nucléotides avec un score Phred supérieur à 30 sont considérés comme de très bonne qualité car équivalent à une précision de 99,99% (ou la probabilité que la base soit incorrecte est de 1 sur 1000). Des séquences de mauvaise qualité peuvent entraîner un alignement incorrect contre le génome de référence et donc fausser les résultats. Cette étape est réalisée par l'outil Trim Galore (version 0.6.5) [45] qui produit des fichiers FASTQ nettoyés ainsi qu'un fichier de statistiques du nettoyage. Les adaptateurs utilisés (standards d'Illumina soient "AAATCAAAAAAAC" et "AGATCGGAAGAGC") sont excisés aux extrémités 5' et 3' de chaque reads. De plus, les bases de qualité inférieure à 20 sont enlevées.

Pour vérifier la qualité du nettoyage, un contrôle qualité est réalisé avec l'outil FastQC (v.0.11.9) [46]. Cet outil donne notamment des indications sur la qualité des bases des reads, des statistiques sur les reads (longueur, % de duplication…) et sur la présence d'adaptateurs. Tous les résultats des statistiques seront intégrés par la suite dans l'outil MultiQC [47] qui regroupe tous les échantillons et génère un rapport de résultats.

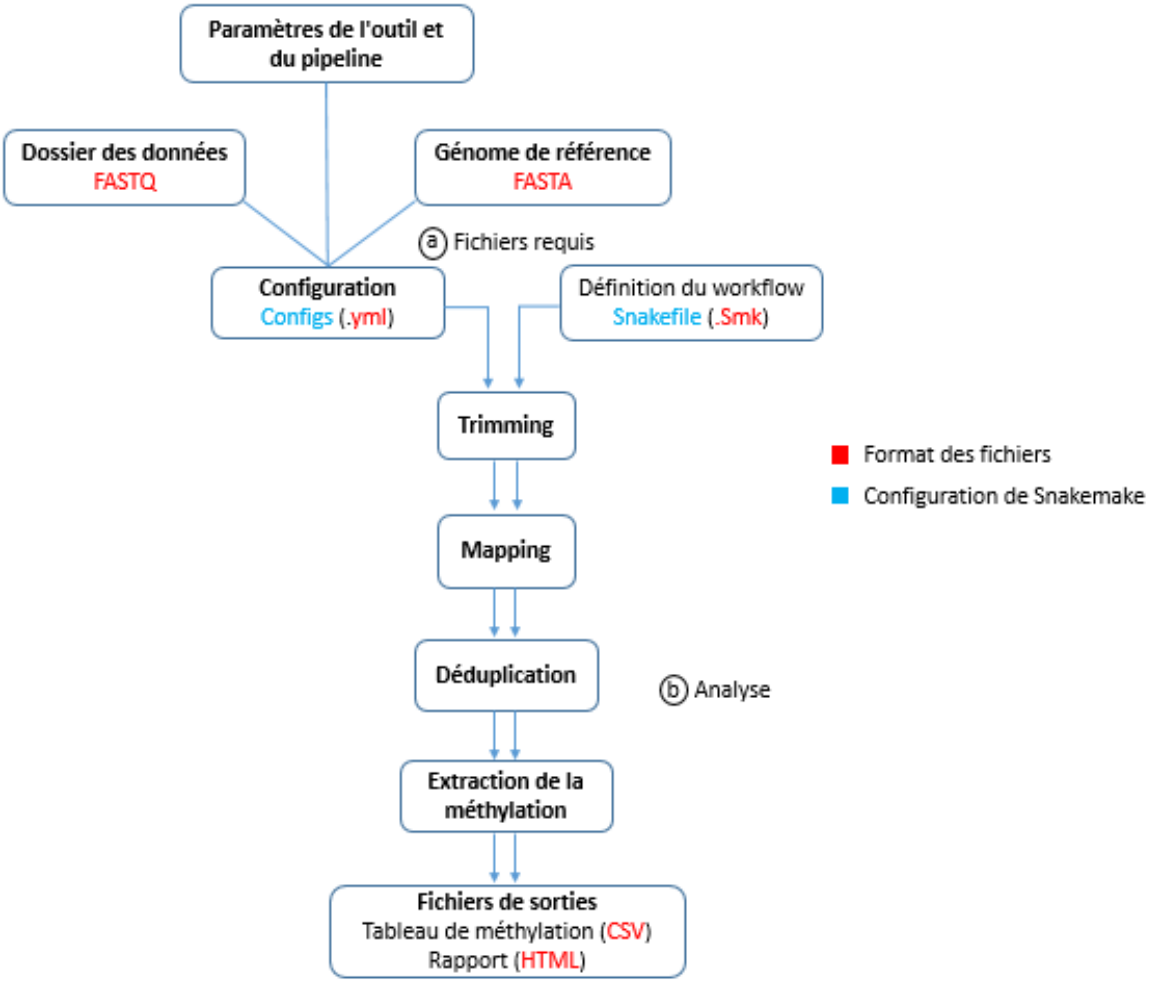


Figure 10: Description de notre pipeline réalisé sous Snakemake (a) les fichiers d'entrée nécessaires au pipeline. (b) les 5 principales étapes d' analyse du pipeline.

B.2 Alignement des séquences

L'étape suivante consiste à aligner les séquences nettoyées (au format FASTQ) contre un génome de référence. Attention, un alignement de séquences suite à un traitement bisulfite est différent d'un alignement standard. Tout d'abord, l'alignement T/C est asymétrique : le T dans les reads traitées au bisulfite pourrait être aligné avec un C ou un T dans la référence, mais pas l'inverse (Figure 13) [48]. Ensuite, chaque reads doit être comparée à quatre séquences (Bisulfite Watson (BSW), bisulfite Crick (BSC), complément inverse de BSW (BSWR) et complément inverse de BSC (BSCR) au lieu des deux habituelles (Watson & Crick) (Figure 12) [49].

Pour cette étape, nous avons utilisé l'outil BSMAPz (v.1.1.3) [50]. BSMAPz prend comme entrée les fichiers FASTQ pairés nettoyés et les aligne contre le génome de référence du peuplier (*Populus trichocarpa* v4.1). La sortie de l'alignement est au format BAM (figure 11). Il comprend les informations suivantes: ID de reads, séquence de reads, score de qualité, longueur de reads, indicateur d'alignement, chromosome aligné, position, brin, nombre d'alignement.

Pour vérifier la qualité des alignements, l'outil SAMtools Stats (v.1.11) [51] est lancé sur les fichiers BAM pour créer un fichier de statistiques du nombre de reads traitées, alignées, non alignées… Ce fichier sera intégré par la suite dans MultiQC.

B.3 Suppression des duplications

Ensuite, il est important de retirer les duplicats générés lors de l'étape de PCR (Polymerase Chain Reaction) du séquençage NGS (next-generation sequencing). Ces séquences correspondent à la même séquence nucléotidique initiale et sont produites lors des étapes de PCR utilisées lors de la création des librairies. Les conserver pourrait conduire à une fausse estimation de la méthylation à une position donnée. Ces séquences sont identifiables car elles commencent par le même enchaînement de nucléotides mais ont une longueur différente.

Cette étape prend en entrée les fichiers d'alignement (BAM) et élimine les reads dupliquées. Pour cela, plusieurs outils de la suite SAMtools [51] (v.1.11) sont utilisés : « sort » (pour trier le fichier BAM par coordonnées), puis « Fixmate » pour corriger les défauts dans l'appariement de reads qui aurait pu être introduits par l'aligneur BSMAPz et enfin « Markdup » pour marquer les alignements multiples et les supprimer. L'outil fournit un fichier d'alignement (BAM) ne contenant plus de reads dupliquées ainsi qu'un fichier de statistiques sur les duplications, intégré par la suite dans MultiQC (figure 10).

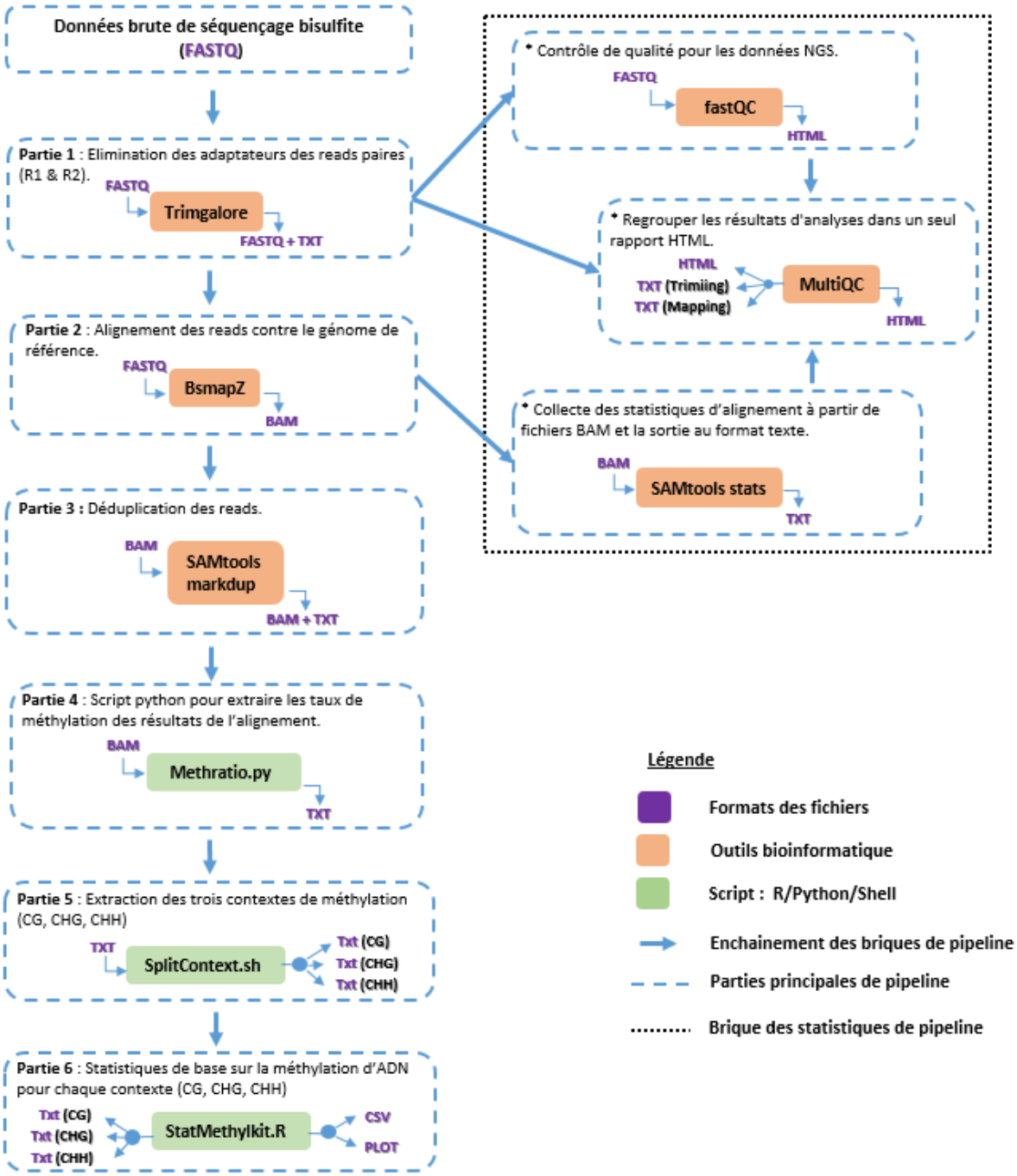


Figure 11 : Schéma détaillé de la suite d'analyse de mon pipeline indiquant pour chaque étape d'analyse (brique) l'outil ou le script utilisé et les formats d'entrée et de sortie.

B.4 Détection des méthylations dans les 3 contextes

L'étape suivante consiste en l'identification des sites de méthylation à partir des fichiers d'alignement (BAM). Un script python nommé « Methratio.py » (fourni avec BSMPAz) permet d'extraire, à partir des fichiers d'alignement produits par BSMAPz, le taux de méthylation pour chaque cytosine (C) analysée. La sortie correspond à un fichier texte (.txt) tabulé avec, à chaque position donnée de C, des indications sur le type de contexte (CG, CHG ou CHH), le rapport de méthylation (ratio), le brin de référence avec C méthylés (+) / C non méthylés (-), le nombre de comptes C totaux sur le locus, le nombre de comptes C + T totaux sur le locus et d'autres informations.

B.5 Extraction des taux de méthylation par contexte

Pour la suite de l'analyse, il est important d'étudier les méthylations par contexte (CG, CHG ou CHH) : il faut donc un fichier par contexte. J'ai donc écrit un script en BASH « SplitContext.sh » qui prend comme entrée le fichier global des méthylations fourni par le programme « methratio.py » et qui va séparer ce fichier selon les 3 contextes de méthylation (CG, CHG, CHH) au sein de 3 fichiers distincts (un par contexte) (figure 11).

B.6 Premières analyses statistiques sur les méthylations

Les 3 fichiers contenant les méthylations par contexte (par individu) restent de grande taille et difficilement interprétables. Il nous a donc semblé nécessaire d'ajouter une étape supplémentaire dans notre pipeline qui renseigne l'utilisateur sur la couverture des sites méthylés par contexte et qui génère des représentations graphiques de ces informations (figure 11). Pour cela, nous avons d'abord utilisé le package MethylKit [52] du logiciel R (v.3.12) permettant de réaliser plusieurs analyses statistiques à partir d'un ensemble de fichiers texte correspondant à différents individus mais pour un contexte de méthylation donné (CG, CHH ou CHG). J'ai écrit un script en R qui a été intégré dans le pipeline grâce à l'objet « script » de snakemake. Ce script possède deux fonctionnalités. Il permet de générer un fichier texte tabulé (CSV) regroupant tous les individus (MC-Seq et WGBS) par contexte de méthylation à l'aide de la fonction *merge* de R. Ensuite, il permet de créer des histogrammes représentant les taux de méthylation mais aussi la couverture d'alignement grâce aux fonctions *getMethylationStats()* et *getCoverageStats()* .

Ces 3 fichiers CSV sont donc les fichiers de sortie de mon pipeline. Ils serviront de fichiers initiaux pour les collaborateurs du projet qui pourront étudier la variabilité de la méthylation sur le génome en lien avec l'adaptation des individus aux conditions environnementales. Ils seront aussi essentiels pour comparer les 2 techniques de séquençage : par WGBS (génome complet) et par MC-Seq (partie du génome), ainsi que pour la validation des conditions testées avec la méthode MC-Seq. Enfin, le pipeline fournit plusieurs fichiers de statistiques sur les différentes étapes qui permettront de vérifier que les analyses se sont bien déroulées et qui sont intégrés dans MultiQC.

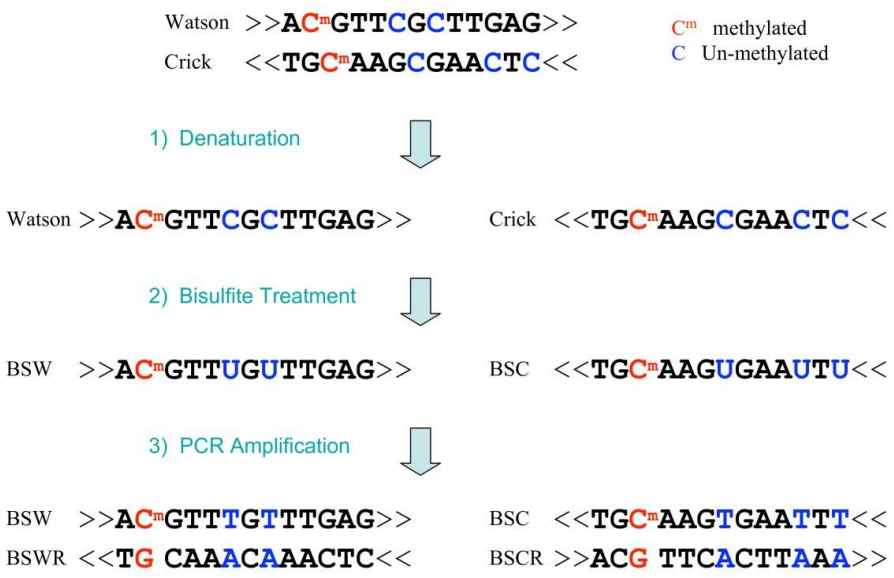


Figure 12 : Les étapes du séquençage bisulfite. 1) Dénaturation: séparation des brins Watson et Crick; 2) Traitement au bisulfite: conversion de cytosines non méthylées (en bleues) en uraciles; les cytosines méthylées (en rouge) restent inchangées; 3) L'amplification par PCR des séquences traitées au bisulfite conduit à la formation de quatre brins distincts: Bisulfite Watson (BSW), bisulfite Crick (BSC), complément inverse de BSW (BSWR) et complément inverse de BSC (BSCR) [48].

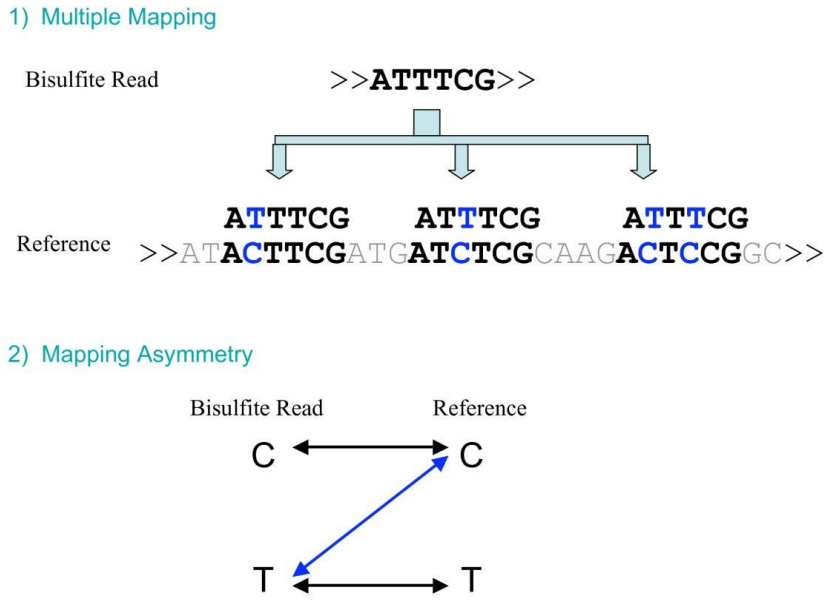


Figure 13: Méthode d'alignement des reads de séquençage obtenus après traitement bisulfique de l'ADN sur un génome de référence. 1) Multiples alignements possibles en raison de la conversion cytosine-thymine dans le traitement au bisulfite. 2) Asymétrie de l'alignement : les thymines dans les reads de bisulfite peuvent être alignées avec les cytosines de la référence (illustrées en bleu) mais pas l'inverse [49].

3. Analyses statistiques des tests d'optimisation du MC-Seq

Pour pouvoir comparer les résultats fournis par les différents échantillons de mon étude, j'ai réalisé des analyses statistiques à l'aide d'outils qui ne sont pas intégrés dans le pipeline mais qui sont nécessaires à la validation de la technique MC-Seq.

A. Structuration des données

En sortie de pipeline, nous avons donc trois fichiers (un par contexte) contenant les niveaux de méthylations de tous nos échantillons réunis. Nous disposons aussi d'un tableau qui regroupe l'ensemble des conditions expérimentales correspondant à nos échantillons ainsi qu'un fichier (BED) contenant les zones du génome ciblées par la capture. Les conditions expérimentales de tous les échantillons ont été regroupées au sein d'un même tableau (Annexe 1). L'échantillon analysé par WGBS n'ayant pas subi le même traitement que les autres, les variables discontinues ont été annotées « WGBS » pour marquer leur singularité. Les variables continues, quant à elles, ont été annotées en tant que valeurs manquantes. Le tableau (fourni par les collaborateurs) comportait des colonnes « Note_01 » et « Note_02 » dans lesquelles le technicien qui a réalisé les librairies, a indiqué des remarques. Le contenu de ces colonnes faisait référence à un ensemble de conditions sans relations. Afin de lever cette ambiguïté, j'ai créé de nouvelles colonnes distinctes regroupant une information unique dans chacune (répétition, « low input », kit périmé et cycle PCR complémentaire) (Annexe 2).

Pour les analyses qui suivent, j'ai filtré le fichier de méthylations selon plusieurs critères. D'abord, j'ai supprimé les lignes contenant des valeurs manquantes afin de ne garder que les positions ayant un taux de méthylations connu pour tous les échantillons. De plus, pour pouvoir comparer les méthylations trouvées par WGBS et MC-Seq, j'ai extrait uniquement les méthylations de l'échantillon WGBS se trouvant dans nos régions d'intérêt (extraites à partir du fichier BED des séquences capturées). Enfin, dans le cadre de mon étude et pour simplifier l'interprétation des résultats, j'ai préféré analyser ensemble uniquement les 14 échantillons correspondant à l'individu de référence (DRA-038) également séquencé en WGBS. J'ai donc retiré du tableau les conditions et les données de taux de méthylation correspondant aux 4 individus autre que l'individu de référence.

B. Analyse des séquences capturées

Pour pouvoir vérifier que la technique par MC-Seq est pertinente, il est nécessaire de vérifier que les séquences obtenues correspondent bien aux zones choisies du génome (séquences « On-target »). Pour cela, j'ai lancé un script (écrit par Abel Garnier) utilisant l'outil Samtools *view* et le fichier BED contenant les séquences capturées. On obtient ainsi le pourcentage de reads qui s'alignent (après l'étape de déduplication) dans nos régions d'intérêt.

Echantillons	Total des pb traitées (R1)	Total filtré (R1)	Total des pb traitées R2	Total filtré (R2)
E01	3,702,210,600bp	3,475,068,738bp	3,702,210,600bp	3,543,193,597bp
E02	4,067,846,400bp	3,810,115,304bp	4,067,846,400bp	3,894,446,805bp
E03	3,422,934,600bp	3,310,710,629bp	3,422,934,600bp	3,269,025,496bp
E04	3,297,908,250bp	3,199,317,816bp	3,297,908,250bp	3,148,776,843bp
E05	3,576,018,150bp	3,393,305,833bp	3,576,018,150bp	3,409,805,675bp
E06	3,075,562,350bp	2,966,794,151bp	3,075,562,350bp	2,933,717,211bp
E07	3,278,193,600bp	3,170,001,566bp	3,278,193,600bp	3,132,264,890bp
E08	2,867,369,850bp	2,770,541,070bp	2,867,369,850bp	2,739,489,754bp
E09	3,228,581,400bp	3,114,653,858bp	3,228,581,400bp	3,086,564,995bp
E10	3,610,252,650bp	3,492,895,459bp	3,610,252,650bp	3,450,426,756bp
E11	3,288,248,850bp	3,175,449,757bp	3,288,248,850bp	3,143,636,408bp
E12	3,162,306,000bp	3,066,885,043bp	3,162,306,000bp	3,019,655,693bp
E13	3,760,269,750bp	3,639,737,088bp	3,760,269,750bp	3,590,560,324bp
E27	4,556,111,700bp	4,427,278,192bp	4,556,111,700bp	4,351,968,191bp
E28	3,109,808,250bp	3,026,002,584bp	3,109,808,250bp	2,968,080,336bp
E29	2,914,136,400bp	2,835,431,331bp	2,914,136,400bp	2,781,706,894bp
E30	3,362,842,350bp	3,266,266,545bp	3,362,842,350bp	3,210,732,690bp
E31	3,221,807,400bp	3,132,855,273bp	3,221,807,400bp	3,075,160,287bp
DRA_38	32,646,919,062bp	30,289,829,913bp	32,646,919,062bp	28,596,875,799bp

Tableau 1 : le total des paires de bases avant et après trimming

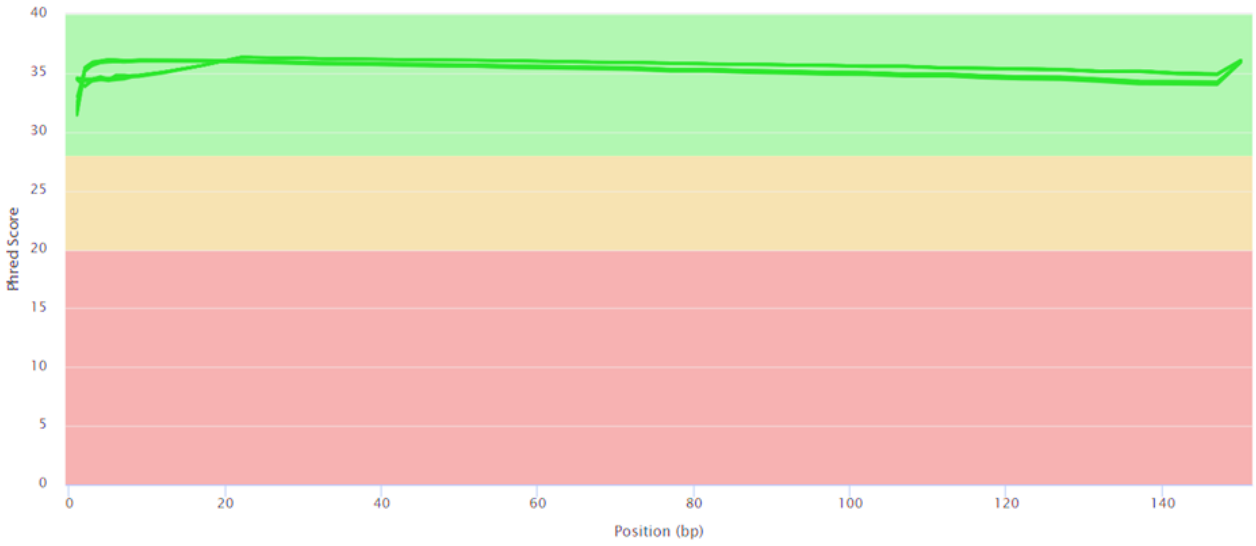


Figure 14: Score de qualité des séquences après trimming, la partie verte (>30) indique une très bonne qualité des reads. Entre 20 et 28 la qualité des reads est moyenne et si la qualité est inférieure à 20, les reads sont de mauvaise qualité. La proportion de chaque base dans les reads est indiquée.

C. Analyses statistiques de methylome

C.1 Corrélation

Afin de vérifier la similarité des taux de méthylation entre les échantillons MC-Seq et ceux de l'échantillon WGBS, un calcul de corrélation a été réalisé en utilisant la méthode de Spearman. Les résultats ont été représentés graphiquement à l'aide du package R *PerformanceAnalytics*. Le choix a été fait de limiter cette analyse aux positions de méthylation incluses dans les 100 premières régions du fichier BED (régions ciblées par la capture) car ce type de représentation est très consommatrice de mémoire vive qui est par ailleurs limitée lors de l'utilisation de l'environnement de programmation. Nous souhaitions aussi garder la lisibilité des résultats d'analyse.

C.2 Analyse en Composantes Principales (ACP)

Notre objectif est d'identifier les conditions expérimentales exerçant le plus d'influence sur le taux de méthylation, leurs effets ou bien l'absence d'effet. Une analyse en composantes principales (ACP) a donc été réalisée sur les taux de méthylation en lien avec les conditions expérimentales utilisées. L'ACP a comme principe une transformation de variables (éventuellement) corrélées en d'autres variables non corrélées appelées composantes principales (CP). Les composantes résultantes de cette transformation sont définies de telle manière que la première composante principale présente la variance la plus élevée et explique le pourcentage le plus élevé de la variabilité des données. Afin d'identifier les composantes expliquant le plus les variations observées, on a utilisé la méthode d'Elbow [52] qui se base sur la recherche du point à partir duquel, l'ajout d'une nouvelle composante augmente peu l'explication de la variation totale des données.

Pour réaliser cette ACP, nous avons utilisé le package R *PCAtools* [53] avec plusieurs fonctionnalités : (1) « ScreePlot » pour déterminer le nombre optimal de composantes principales à conserver ainsi que pour expliquer la variation de chacune d'entre elles, (2) l'outil graphique « bi-plot » pour une simple représentation de la répartition des échantillons selon les composantes retenues et (3) un graphique de corrélation présentant la corrélation entre les différentes conditions expérimentales et les composantes de notre ACP.

Résultats :

J'ai développé un pipeline de détection des niveaux de méthylations sur des données de séquençage WGBS et MC-Seq sur le cluster de calcul du CEA. Après avoir vérifié la fonctionnalité de chaque étape d'analyse, j'ai vérifié le fonctionnement global du pipeline. Enfin, j'ai analysé, à l'aide de ce pipeline, les jeux de données de séquençage WGBS et MC-Seq obtenus sur 18 échantillons de peuplier noir (*P. nigra*).

Echantillons ID	Total des Reads	Alignement (%)
E01	28311331	45.9
E02	31397407	46.9
E03	26752606	45.3
E04	26111601	46.1
E05	27345856	43.0
E06	24026040	43.4
E07	26031231	46.1
E08	22546694	45.8
E09	24895439	44.1
E10	28350316	45.2
E11	25729274	45.3
E12	24846345	45.6
E13	29686867	46.1
E27	36867205	47.9
E28	24751793	46.8
E29	23183272	46.7
E30	27505283	48.8
E31	25874987	47.2

Tableau 2: Nombre des reads conservées après le nettoyage pour chaque échantillon et pourcentage de ces reads alignées sur le génome de référence de *P. trichocarpa*.

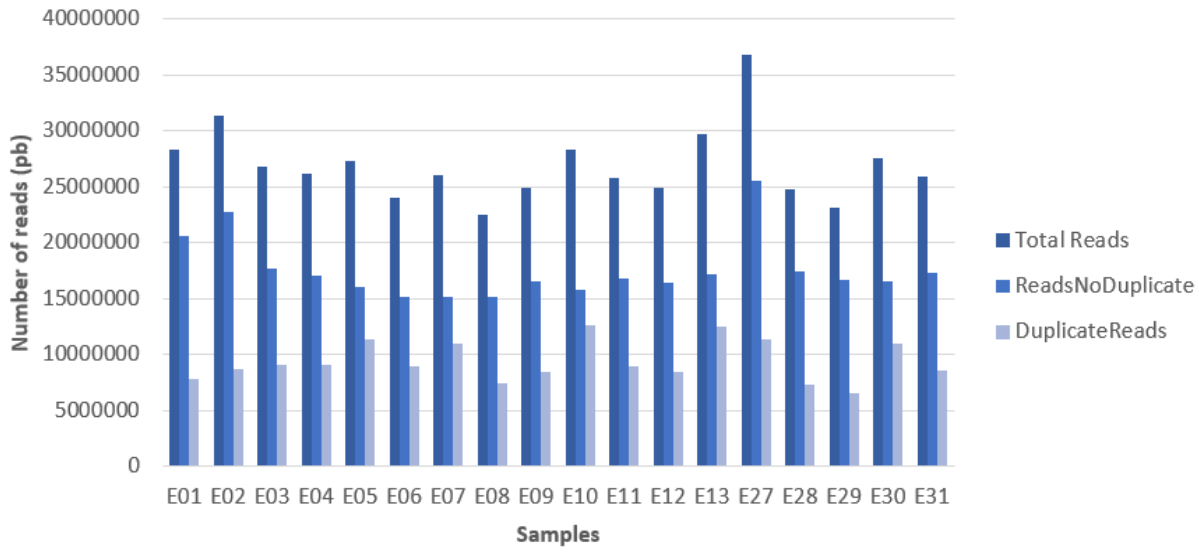


Figure 15: Nombre de reads dupliquées et non dupliquées pour chaque échantillon.

1. Validation du Pipeline

Dans un premier temps, il a fallu vérifier que les briques constituant le pipeline produisaient les sorties attendues et que les différentes étapes d'analyses s'enchaînaient correctement.

A. Résultats « bruts » de sortie

Je vais ici vous présenter les différents fichiers obtenus grâce aux briques du pipeline.

Les premiers résultats (Tableau 1) concernent la première étape du pipeline : le nettoyage des reads. Pour chaque individu séquencé en MC-Seq, le nombre moyen de reads brutes obtenues est de 26,900,752, représentant un total de 3,702,210,600 pb. Après retrait des adaptateurs et des séquences de mauvaise qualité, il reste en moyenne 3,227,694,766 pb par individu. En moyenne, seuls 12.82% des données brutes ont été retirées, ce qui montre la bonne qualité des données. Sur les 32,646,919,062 bp initiales de l'individu de référence séquencé en WGBS, 30,289,829,913 bp ont été conservées, soient 92.78% des données initiales.

Je vais ici vous présenter les différents fichiers obtenus grâce aux briques du pipeline.

Ensuite, un contrôle de qualité a été réalisé. Cette étape a montré (par plusieurs graphiques) la bonne qualité de nos données et de notre nettoyage avec un score de qualité supérieure à 30 pour l'ensemble des séquences (figure 14). Du plus, l'outil FastQC a détecté la présence de duplications dans tous les échantillons.

L'alignement des reads MC-Seq contre le génome de *Populus trichocarpa* (v4.1) indique un pourcentage de séquences appariées alignées qui varie de 43.0% (pour l'échantillon E05) à 48.8% (pour E30) (Tableau 2). Concernant l'individu séquencé en WGBS, la taux d'alignement est de 44%.

Le pourcentage de reads dupliquées est compris entre 27,4% et 44,5% (Figure 15) pour l'ensemble des 18 échantillons MC-Seq et de 13,7% pour l'individu séquencé en WGBS.

Le nombre total de méthylations détectées par échantillon oscille entre 16,183,391 et 20,354,557. On observe, pour chaque contexte de méthylation (CG, CHG et CHH), des variations importantes du nombre de cytosines méthylées (mC) qui correspondent, en moyenne, à 76 % pour CHH, 14,3 % pour CHG et enfin 9,6 % pour CG (Figure 16).

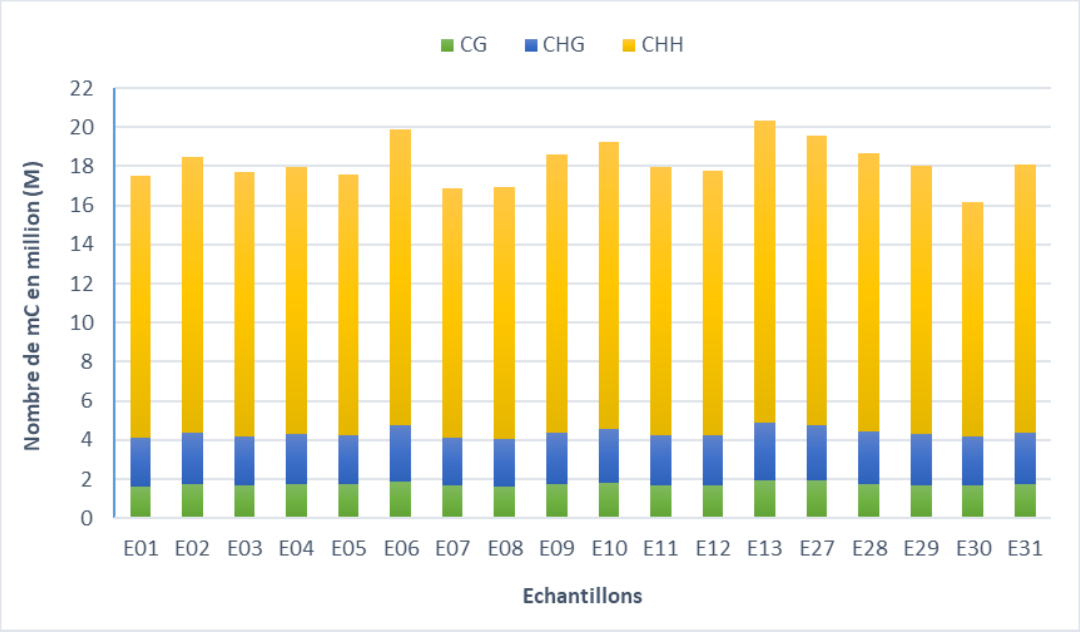


Figure 16: Répartition des nombres des cytosines méthylées (mC) dans chaque contexte de méthylation (CG, CHG & CHH) pour chaque échantillon.

Echantillons	CG	CHG	CHH
E01	9,40%	14,10%	76,50%
E02	9,50%	14,10%	76,40%
E03	9,50%	14,20%	76,30%
E04	9,60%	14,40%	76,00%
E05	9,80%	14,50%	75,70%
E06	9,60%	14,40%	76,10%
E07	9,90%	14,50%	75,70%
E08	9,60%	14,30%	76,00%
E09	9,50%	14,10%	76,40%
E10	9,50%	14,30%	76,20%
E11	9,50%	14,20%	76,30%
E12	9,50%	14,20%	76,20%
E13	9,50%	14,60%	75,80%
E27	9,80%	14,40%	75,80%
E28	9,50%	14,30%	76,20%
E29	9,50%	14,30%	76,20%
E30	10,60%	15,20%	74,20%
E31	9,70%	14,40%	75,80%
DRA_38	8,42%	14,11%	77,46%

Tableau 3: les pourcentages des cytosines méthylées dans chaque contexte de méthylation pour les échantillons MC-Seq et WGBS.

Concernant l'individu séquencé en WGBS, sur les régions du génome correspondant aux séquences capturées en MC-Seq, le taux de méthylation des cytosines est équivalent à celui des échantillons séquencés en MC-Seq. Les proportions de méthylations CHH, CHG et CG sont de 79,8%, 12,7% et 7,3%, respectivement, ce qui est similaire aux résultats trouvés en séquençage MC-Seq (Tableau 3).

Le nombre de mC identifiées en contexte CHH va de 12,013,675 (pour E30) à 15,437,966 (pour E13). Pour le contexte CHG, ce nombre varie entre 2,430,118 (E08) et 2,975,150 (E13). Enfin, en contexte CG, seulement 1,631,743 de mC sont identifiées chez E08 tandis que le maximum de mC est détecté chez E13 avec 1,941,441 méthylations (Tableau 3).

L'intégration des méthylations des 18 échantillons de MC-Seq et de notre échantillon WGBS nous a permis d'obtenir des statistiques de base sur la distribution des données de méthylation des trois contextes (CG, CHG, CHH) telles que la couverture et le pourcentage de méthylation. Celles-ci ont montré que la distribution du pourcentage de méthylation est bimodale pour les deux contextes CG et CHG, où la plupart des positions ont une méthylation élevée ou faible. Pour CHH, nous n'avons pas de distribution bimodale (Figure 17).

B. Fonctionnalités du pipeline

B.1 Performances du pipeline

J'ai effectué une série de tests afin d'évaluer les performances du pipeline. J'ai d'abord voulu vérifier la reproductibilité des résultats en exécutant les briques, qui constituent le pipeline, de façon indépendante et j'ai comparé les résultats obtenus avec ceux obtenus en exécutant le pipeline. J'ai confirmé l'obtention de résultats identiques pour chaque étape d'analyse. Notre pipeline permet donc d'effectuer des analyses reproductibles.

L'utilisation de snakemake assurant un lancement en parallèle des différents jobs du pipeline, j'ai ensuite vérifié que les ressources informatiques demandées étaient exploitées de façon optimale. Les ressources demandées pour l'analyse des 18 échantillons de MC-Seq (avec des fichiers d'entrée de 1,5 à 2,8 Go) sont de 1 nœud de calcul possédant un processeur avec 36 cœurs et 10705 Mo de RAM par cœur. Notre objectif est que l'analyse en parallèle de plusieurs échantillons MC-Seq dure moins de 24h pour pouvoir utiliser la queue de calcul normale du CEA plus facilement disponible. Pour l'analyse des échantillons MC-Seq, l'exécution du pipeline a pris 2h 18min en moyenne (Tableau 4). L'utilisation des ressources dépend principalement des besoins des logiciels en termes de CPU ainsi que du nombre d'échantillons analysés simultanément.

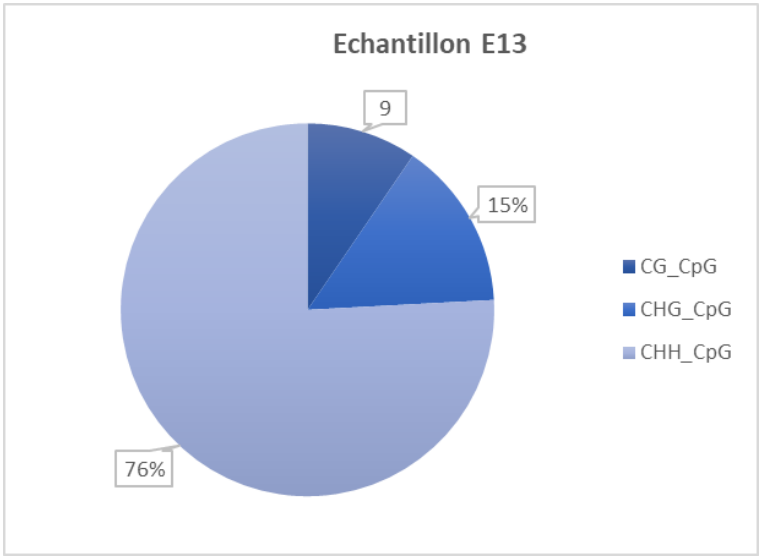


Figure 17: Diagramme à secteurs montrant la distribution des Cm par contexte de méthylation (CG, CHH, CHG) pour l'échantillon (E13)

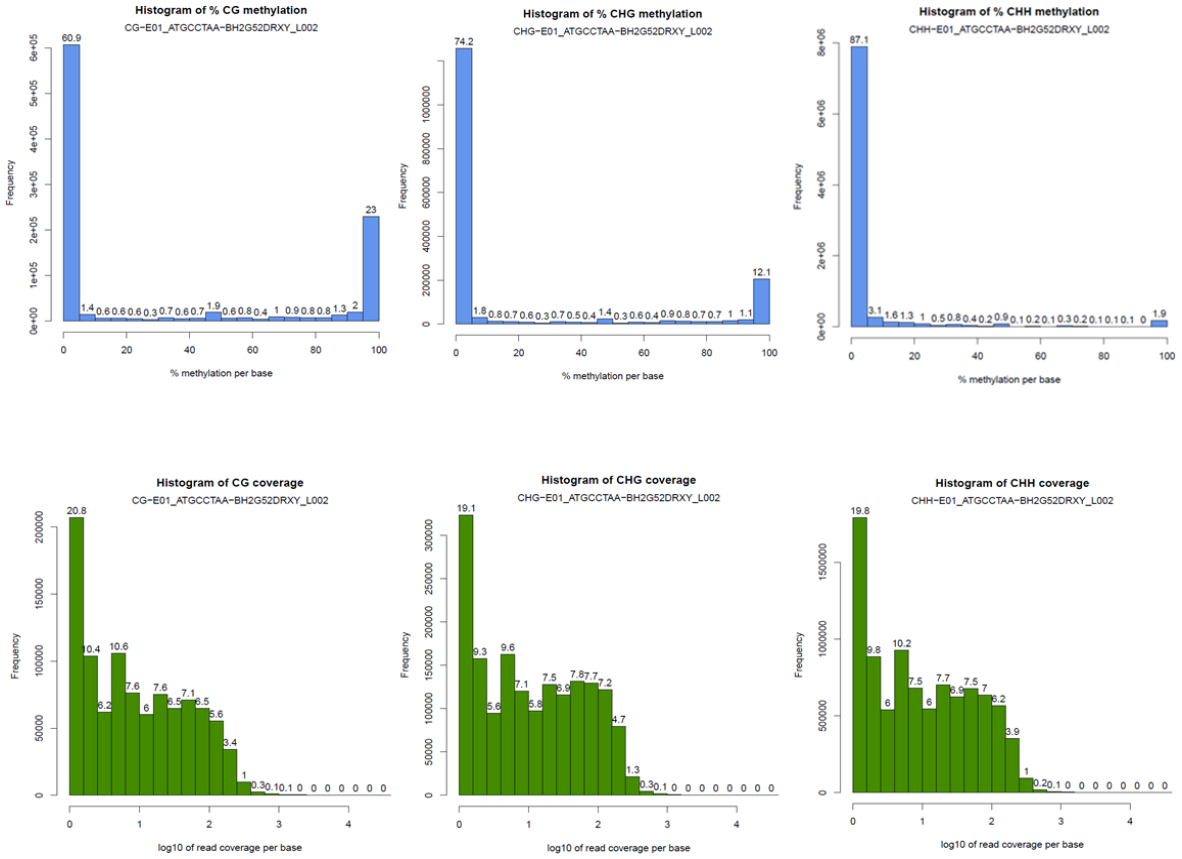


Figure 17: Pourcentages de méthylation et profondeurs de séquençage (couvertures) par base de l'échantillon de référence de WGBS.

L’étape d’alignement (BsmapZ) est l’étape la plus chronophage (~1h 12min) par rapport aux autres briques, suivie de l’étape de nettoyage (Trimgalore) qui dure en moyenne 35min, puis de l’étape d’analyses statistiques (Methylkit) (~12min) et celle de l’extraction des méthylations par contexte (~8min) (Tableau 4).

Un graphique de visualisation des tâches et de leurs dépendances sous forme de DAG (Directed Acyclic Graph) (figure 18) a été généré par snakemake pour contrôler la progression de l’exécution. Sur la figure 18, les étapes qui ont déjà été exécutées sont affichées en lignes pointillées et les étapes qui vont l'être sont entourées de lignes pleines. Le graphique montre également la manière dont snakemake organise et parallélise les différentes tâches. Les premières briques du pipeline (Trimgalore, FastQC, MultiQC, SAMtools stats, Bsmapz, SAMtools makdup, Methratio, Spliting-context) sont lancées en parallèle. Par contre, à partir de l'étape d'extraction des méthylations par contexte, toutes les opérations précédentes doivent être complétées pour que l'exécution se poursuive. En effet, la brique suivante (Methylkit) a besoin des fichiers contenant les niveaux de méthylation (par contexte) de tous les échantillons pour s’exécuter (figure 18).

2. Exploitation des données testant les conditions expérimentales MC-Seq

Pour les 18 échantillons MC-Seq, on obtient un pourcentage de séquences correspondant à nos zones capturées compris entre 51% (pour l’échantillon E30) et 63,9% (E01).

Dans la suite du manuscrit, je vais vous présenter les résultats obtenus pour le contexte CG. J’ai obtenu des résultats similaires pour CHG (Annexe 4) mais légèrement moins bons. Pour ce qui est du contexte CHH, du fait du grand nombre de méthylations nécessitant un temps de calcul long, je n’ai pas encore pu finaliser les résultats.

La comparaison des échantillons a été réalisée en les considérant par paires et en calculant les coefficients de corrélation entre les taux de méthylation des 14 échantillons MC-Seq (individu DRA-038) et l’échantillon WGBS (figure 19). Dans le cadre de cette analyse, une couverture minimale de 1X a été considérée. On observe que les coefficients de corrélation sont similaires entre la plupart des échantillons MC-Seq et l’échantillon WGBS. Les échantillons MC-Seq sont fortement corrélés au WGBS : les taux de corrélation allant de 82 % (E05) à 90%. De plus, les échantillons MC-Seq sont aussi très bien corrélés entre eux (81% à 92%). Toutes les distributions des valeurs de méthylation des échantillons MC-Seq et de l'échantillon WGBS sont bimodales (figure 19), à part celle de l’échantillon E05, qui est unimodale. Les résultats de l'ACP présentés ci-dessous concernent uniquement les profils des taux de méthylation (pour le contexte CG) des 14 échantillons associés

Briques	Durée (h:m:s)	Max_MEM	CPU usage	Nb Nœuds	Temps de calcul
Trimgalore	00:35:28	117.93	6	1	**2h 18min**
FastQC	00:03:06	284.67	3	1	
MultiQC	00:00:23	66.40	6	1	
BsmapZ	01:13:06	1986.69	6	1	
Samtools-Stats	00:01:16	7.98	6	1	
Samttols Markdup	00:05:57	1648.20	6	1	
Methratio	00:08:12	1641.11	6	1	
Spliting-context	00:00:24	5.20	1	1	
Methylkit	00:12:04	8772.90	1	1	

Tableau 4: Ressources informatiques utilisées par le pipeline pour analyser les 18 échantillons de MC-Seq. Pour chaque brique du pipeline lancé en parallèle, on indique les ressources informatiques utilisées, la durée de calcul par échantillon, le maximum de mémoire utilisée par échantillon, le nombre de CPU utilisé par brique et le nombre de nœuds mobilisés.

aux différentes conditions expérimentales et le WGBS. Tout d'abord, pour choisir le nombre optimal de composantes principales (CP) à conserver, la figure 20 (ScreePlot) présente le pourcentage de variation des données expliqué par chacune des composantes principales. Généralement, on considère qu'une CP est pertinente, si elle explique au moins 10% de la variation observée. Ici, la première composante est la seule expliquant un pourcentage significatif (25%) de la variation du taux de méthylation en fonction des conditions expérimentales du MC-Seq tandis que les 13 autres expliquent chacune 6,5 % de la variabilité. Le résultat de la méthode Elbow conforte notre première impression car elle ne retient que les deux premières composantes qui sont les plus représentatives pour décrire les données de mC en lien avec les conditions expérimentales de MC-Seq car elles expliquent environ 31,5% de la variation observée.

Ensuite, le graphique d'Eigoncor (Figure 21) présente les corrélations de Pearson entre les différentes conditions expérimentales et les composants de notre ACP. On note qu'au niveau de la première composante (CP1), la seule corrélation significative est avec la méthode de fragmentation (-0,7). Il y a aussi quelques valeurs de corrélation significatives avec les autres composants : CP5 avec la quantité d'ADN fournie à l'hybridation (-0,55), CP7 avec la quantité initiale d'ADN (-0,78) et CP10 avec plusieurs conditions. Cependant, pour rappel, ces autres CP ne représentent que 6,5% de la variation observée.

Enfin, sur la figure 22 présentant l'ACP (en fonction de la PC1 et de la PC2), on peut noter que la majorité des échantillons MC-Seq sont regroupés entre eux et avec l'échantillon WGBS. Si l'on considère la première composante de l'ACP, l'échantillon E05 est très éloigné des autres échantillons. Par contre, sur l'axe 2, ce sont les échantillons E08 et E29 qui sont excentrés des autres échantillons.

Discussion :

Le réchauffement climatique actuel est associé à la modification de conditions environnementales, telles que l'augmentation de la température de la fréquence des épisodes de sécheresse, ce qui signifie que les arbres doivent s'adapter pour survivre. Le projet EpiTree a été mis en place par des experts de différents domaines de la biologie pour étudier le rôle de l'épigénétique (méthylation de l'ADN) dans les mécanismes d'adaptation chez 2 espèces d'arbres : le peuplier et le chêne. C'est dans cet objectif que j'ai développé un pipeline bioinformatique d'analyses de données de méthylation de l'ADN. Je l'ai validé sur une expérience visant à établir les conditions expérimentales pour l'étude de la méthylation de zones d'intérêt du génome chez le peuplier.

Il existait déjà d'autres pipelines d'analyse de données de méthylation de l'ADN : nf-core/methylseq [54], gemBS [55], Bicycle [56], BSseeker3 [57] ou MethylStar [58]. Cependant, aucun ne répond à l'intégralité de nos besoins. En effet, ils utilisent des outils qui ne sont pas adaptés à l'étude des plantes.

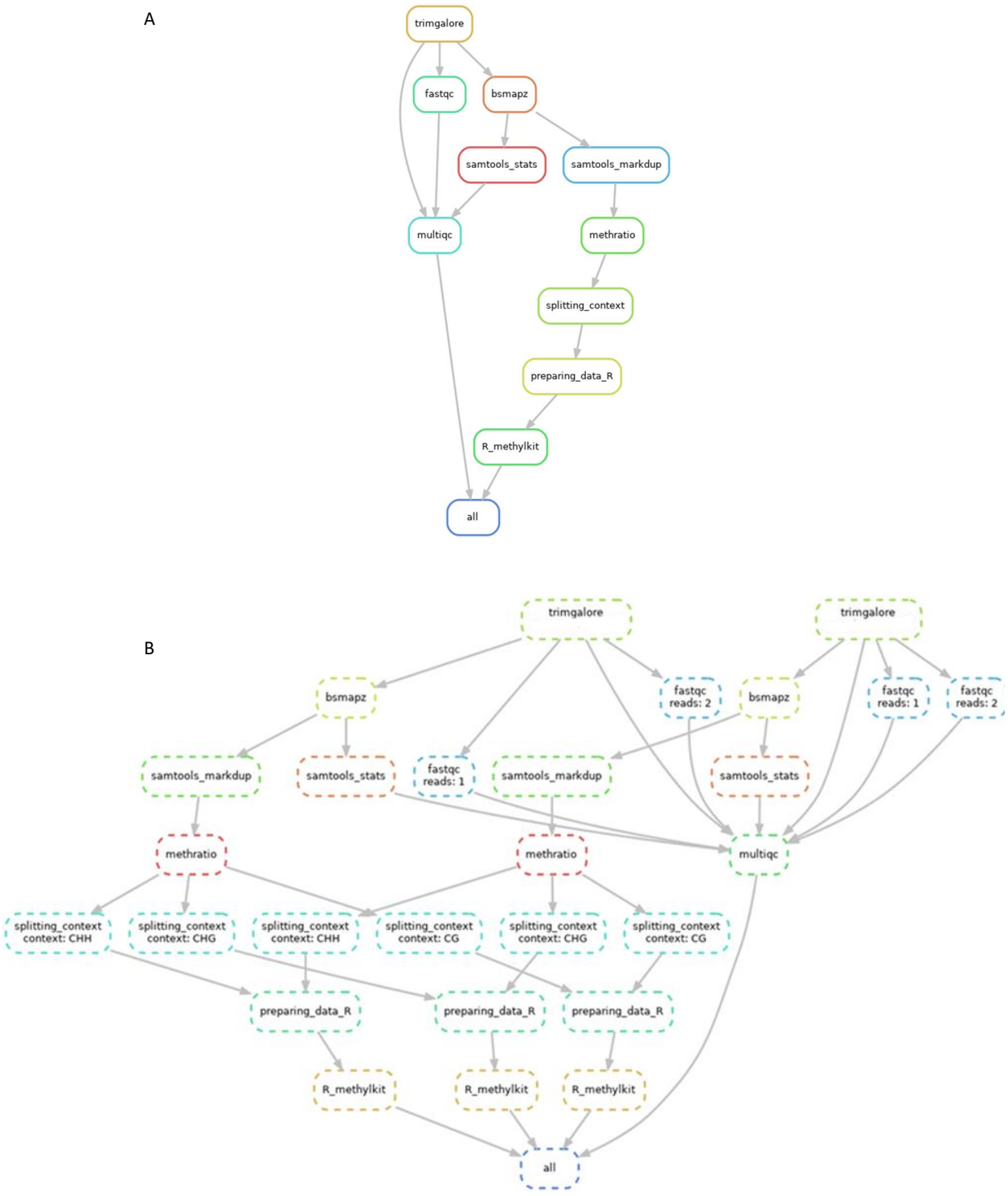


Figure 18: DAG (Directed Acyclic Graph) représentant graphique de l'enchaînement des briques du dans Snakemake. (A) DAG général des outils de pipeline. (B) DAG représente l'analyse du pipeline et la façon dont snakemake crée des étapes de pipeline parallèles.

Ces pipelines ont aussi des contraintes en termes de type de données utilisées et le nombre de fichiers d'entrée, le mode d'installation. Nous avions aussi des contraintes techniques liées à l'infrastructure (le cluster de calcul du CEA) sur laquelle devait être développé le pipeline. Un des avantages de notre pipeline est qu'il peut être exporté facilement sur les clusters de calculs des collaborateurs, notamment la plateforme Genotoul à Toulouse (http://bioinfo.genotoul.fr/). Au niveau du gestionnaire de flux d'analyses, nous avons choisi snakemake plutôt que Nextflow [59]. Son premier avantage est qu'il est basé sur Python, ce qui rend plus pratique. Il permet également de reprendre l'analyse là où elle a été arrêtée lorsqu'un problème technique survient, ce qui la rend particulièrement efficace par rapport aux autres gestionnaires. Ce pipeline a été développé pour fonctionner à la fois sur des données de WBGS et de MC-Seq : des paramètres (par défaut) ont été fixés mais ils sont modifiables dans le fichier de configuration. De plus, nous avons fait en sorte d'avoir tous les fichiers de log et de statistiques possibles (intégrées dans MultiQC) pour voir si toutes les étapes se sont bien passées. Il sera donc aisé de relancer, au besoin, les analyses sur un individu en cas d'erreurs.

Mon pipeline comprend plusieurs étapes. La première est le nettoyage des reads. Généralement, la vérification de la qualité des séquences brutes est la première étape à effectuer pour déterminer si elles ont besoin d'être nettoyées. Cependant, dans notre cas, nous vérifions de façon automatisée uniquement la qualité des données avec FastQC après le nettoyage. En effet, compte-tenu du grand nombre d'individus qui seront analysés par la suite, il est impensable qu'un collaborateur vérifie, un à un, les résultats de FastQC des 500 jeux de données afin de déterminer lesquels ont besoin d'un nettoyage. Aussi, nous avons décidé d'effectuer un nettoyage systématique des données qui n'a que des avantages : si les séquences étaient déjà de bonne qualité, elles le resteront et si elles sont de mauvaise qualité, alors elles seront améliorées pour ne pas fausser les résultats par la suite (Tableau 1). Ainsi, avant le lancement des 500 échantillons, une vérification manuelle sur un nombre réduit d'entre eux pourra être effectuée et permettra de définir les paramètres à appliquer sur tous les individus.

Il existe, à l'heure actuelle, plusieurs outils permettant d'aligner des séquences traitées au bisulfite contre un génome de référence. Parmi eux, Bismark [60] et Bsmap [61] réalisent généralement les meilleurs alignements de données de méthylation [62, 63]. Les collaborateurs dans Epitree avaient constaté que Bismark était beaucoup moins performant que Bsmap pour l'alignement de données issues d'organismes végétaux [64, 65] lors de l'analyse des 30 premiers WGBS du projet. J'ai moi-même testé ces 2 outils sur 3 autres échantillons WGBS et les pourcentages d'alignements de Bismark sont toujours inférieurs à ceux de Bsmap. Néanmoins, nous avons aussi constaté que Bsmap n'est plus maintenu depuis 2013 et qu'il est désormais incompatible avec les versions plus récentes de SAMtools. Nous avons identifié une alternative à Bsmap : BsmapZ. Ce dernier a été développé à

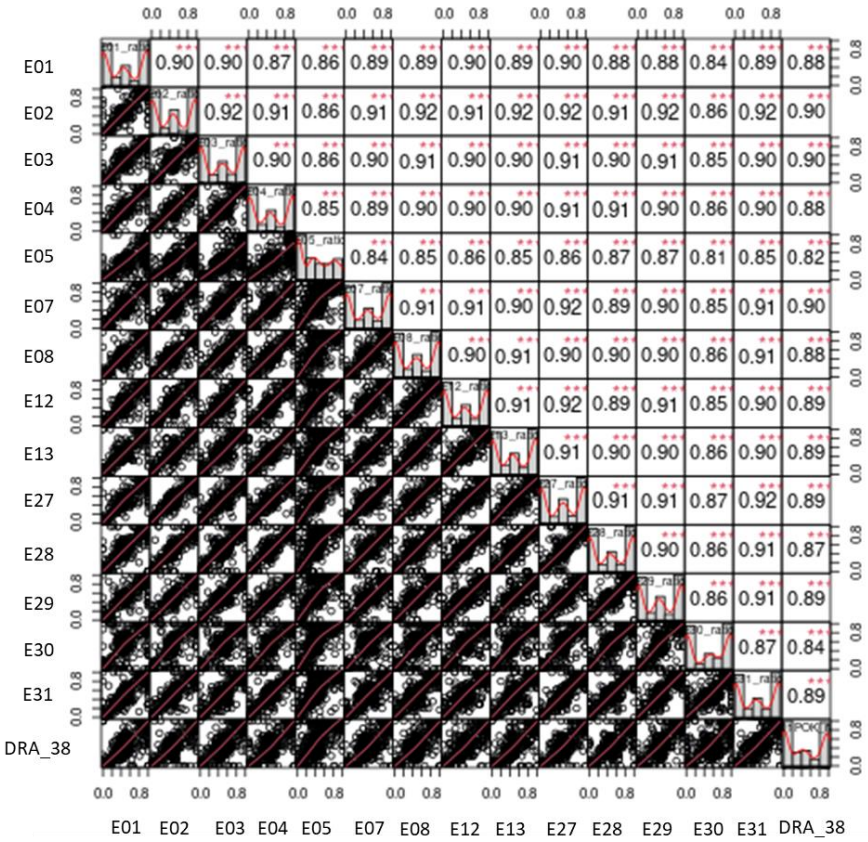


Figure 19: Représentation (Scatter plots) du niveau de corrélation entre les échantillons du projet visant à optimiser les conditions expérimentales de MC-Seq. Diagrammes de dispersion des valeurs de % de méthylation pour chaque paire de 18 échantillons MC-Seq et un échantillon WGBS (dernière ligne). Les nombres dans le coin supérieur droit indiquent les valeurs de corrélations de Spearman par paires d'échantillon. Les histogrammes sur la diagonale représentent le pourcentage de méthylation pour chaque échantillon.

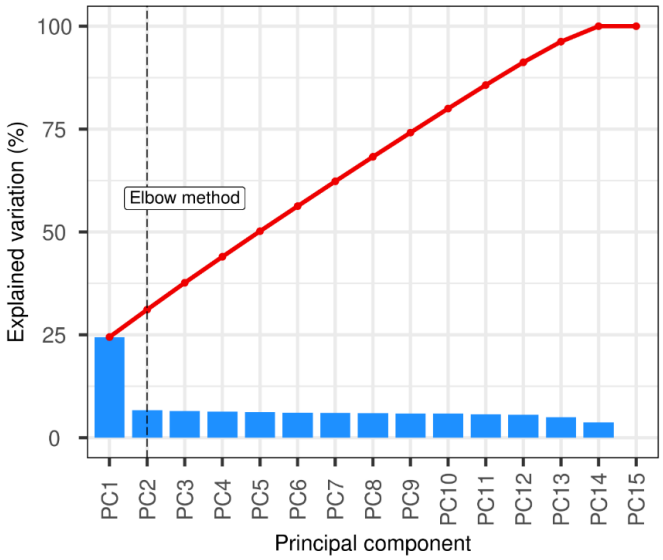


Figure 20: Analyse en Composantes Principales (ACP) montrant la part de la variabilité observée dans les niveaux de méthylation des échantillons MC-Seq qui est expliquée par chaque composante principale (CP) (Scree Plot) La courbe noir droit représente la trace de Elbow. Et la courbe rouge correspond au cumul de variabilité expliquée par les CP.

partir du code source (Fork) de Bsmap (v2.90). En comparant les résultats obtenus avec Bsmap et BsmapZ sur un jeu de données WGBS, nous avons constaté que les résultats étaient quasi-identiques (27.8% pour notre échantillon de WGBS DRA_38) (tableau 5). BsmapZ est aussi plus rapide, plus sensible et plus flexible par rapport aux autres aligneurs prenant en compte le traitement bisulfite. Après discussions avec les utilisateurs (qui lanceront les analyses en aval du pipeline), il a été décidé de s'arrêter à l'étape de Methylkit pour la partie automatisée.

Le deuxième objectif de mon stage était de vérifier si l'utilisation de la méthode MC-Seq, qui se limite au séquençage de certaines zones d'intérêt du génome, nous donne des résultats comparables en termes de méthylations à ceux obtenus sur ces mêmes zones du génome avec le séquençage génome entier (WGBS).

Les taux d'alignement trouvés pour tous les échantillons (aux environs de 45 %) sont des taux qui correspondent à ceux attendus pour des alignements avec un traitement bisulfite [61]. De plus, nous avons utilisé le génome de référence de *P. trichocarpa* malgré le fait que nos individus sont des *P. nigra*. Nous avons fait ce choix car le génome de *P. trichocarpa* est de très grande qualité, *P. trichocarpa* et *P. nigra* sont très proches phylogénétiquement et il n'existe pas à l'heure actuelle de génome de *P. nigra* publié. De plus, pour éliminer les duplications, nous avons choisi d'inclure samtools markdup dans notre pipeline plutôt que d'utiliser un paramètre dans methratio.py (-r) qui permet de les enlever. En effet, il n'est pas possible de connaître le nombre de duplications avec le paramêtre -r de methratio.py alors que c'est possible avec Samtools markdup qui produit un fichier de statistiques. Enfin, le pourcentage d'alignement de reads ciblant les zones d'intérêt (reads « On Target ») d'environ 60 % (tableau 2) figure parmis les meilleurs taux trouvés dans d'autres études utilisant le séquençage par capture [66].

Les distributions du pourcentage de méthylations, pour les contextes CG et CHG, sont bimodales (Figure 17). Ceci est en accord avec les pourcentages de méthylation attendus (avec des niveaux élevées). La distribution pour CHH n'est pas bimodale, ce qui reflète généralement le fait que la majorité des bases ont une faible méthylation. De plus, la distribution de la couverture des reads est également une métrique importante qui dépend en partie de la quantité de duplicats de PCR [52].

Nous nous sommes ensuite attachés à comparer les niveaux de méthylation en se limitant aux zones du génome commun aux 2 approches. Nous avons constaté que les échantillons MC-Seq sont tous fortement corrélés avec l'échantillon séquencé en WGBS, avec des valeurs de corrélation comprises entre 82% et 90% (figure 19). Notons que l'échantillon E05 est le moins bien corrélé avec l'échantillon WGBS. On peut remarquer, grâce à l'ACP, que cet échantillon se distingue des autres échantillons par sa distance sur la première composante (figure 22). Cet échantillon est le seul pour lequel la

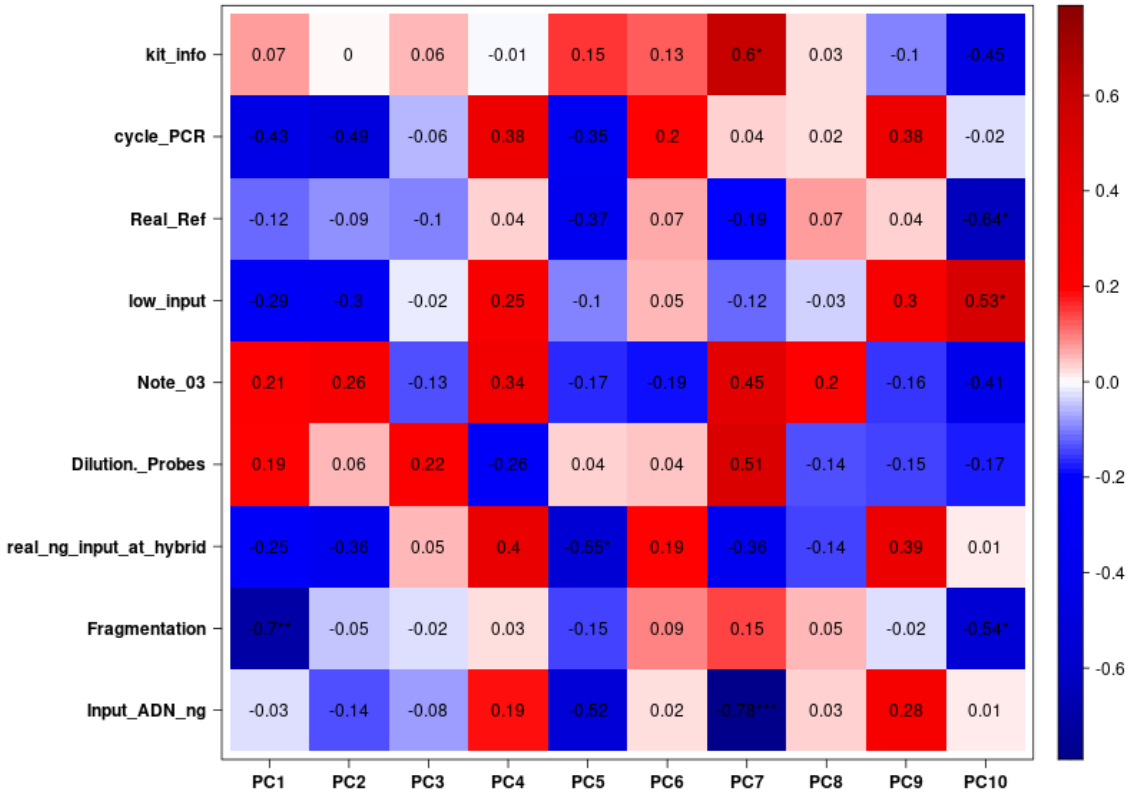


Figure 21 : Plot Eigoncor qui présente la corrélation de Pearson entre les composantes de l'ACP et les conditions expérimentales des 14 échantillons MC-Seq.

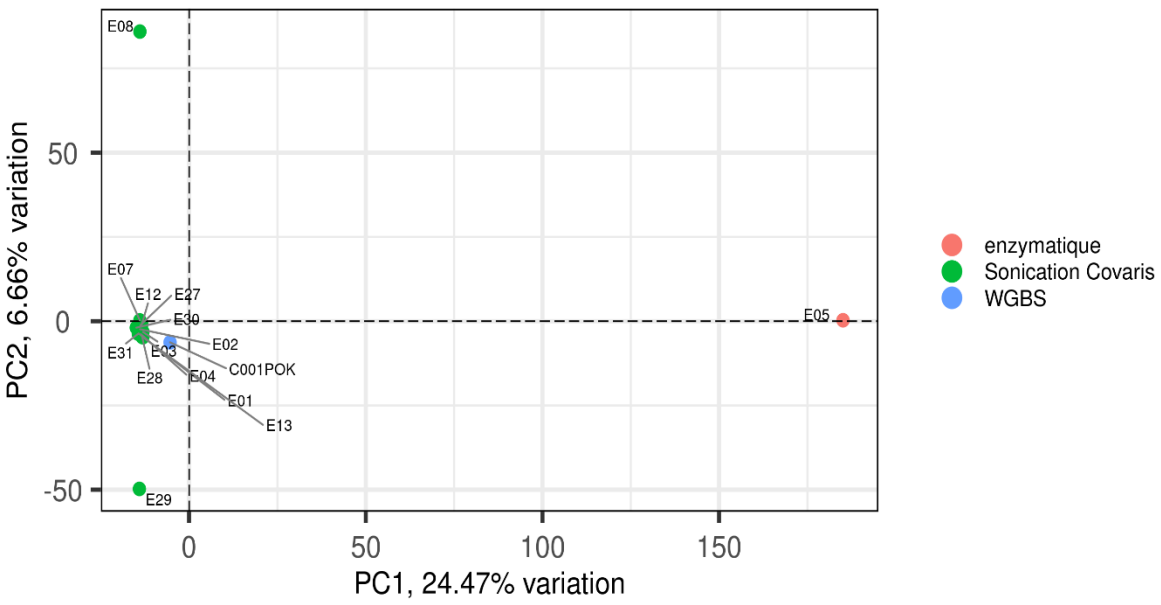


Figure 22: ACP des profils de méthylation d'ADN pour les 18 échantillons MC-Seq et l'échantillon WGBS. On considère ici les composantes principales 1 (PC1) et 2 (PC2) pour chaque échantillon.

fragmentation enzymatique a été utilisée, alors que les autres échantillons (WGBS et MC-Seq) ont subi une fragmentation Covaris par sonication. Ce résultat est en accord avec le fait que la seule corrélation significative avec la première composante est la méthode de fragmentation (figure 21). De plus, la distribution des valeurs de méthylation de l'échantillon E05 est la seule qui soit unimodale (figure 19). Selon la plateforme PGTB en charge du séquençage des échantillons du projet EpiTree [68], la méthode de fragmentation Covaris est plus reproductible par rapport à la fragmentation enzymatique, elle permet d'obtenir des fragments de taille très homogène. A titre d'exemple, sur la figure 24, les courbes orange, jaune, bleu turquoise et verte correspondent à cette méthode, montrant que la distribution de la taille des fragments est une gaussienne avec un pic à environ 200 pb [68]. La fragmentation enzymatique est, quant à elle, moins reproductible et conduit à des fragments d'ADN de tailles très variables (figure 23). Il semble donc logique que le type de fragmentation d'ADN utilisé ait un impact significatif sur les résultats. Ce constat a conduit à choisir la fragmentation Covaris pour l'étude des 500 individus.

Enfin, d'après les résultats de la comparaison des taux de méthylation obtenus dans les différentes conditions expérimentales pour un même individu. Nous avons trouvé que tous les échantillons étaient extrêmement bien corrélés, nous laissant penser que la majorité des conditions testées n'a pas d'influence significative sur les niveaux de méthylation détectés. En effet, de façon surprenante, la dilution des sondes ne semble pas avoir d'impact sur les résultats observés lorsqu'on compare les résultats obtenus avec les 4 taux de dilution (1, 1/8, 1/10, 1/16). On s'attendait à ce que le fait de diluer les sondes s'hybridant sur les zones d'intérêt conduise à une diminution du séquençage de ces zones et donc à une incapacité à déterminer leurs niveaux de méthylation. De même, le pourcentage de reads ciblant les zones d'intérêt (reads « On Target ») et le taux de duplication ne semblent pas impactés par la dilution des sondes. Une dilution de 1/8 permettant de diminuer considérablement le coût de l'expérience, elle a donc été choisie par les partenaires d'Epitree. L'impact de la quantité d'ADN initiale (Input_DNA_ng; 3 µG vs 1 µG) ne semble pas avoir d'effet sur les résultats obtenus. Il est donc possible de réaliser le séquençage MC-Seq avec 1µG d'ADN au lieu des 3 µG recommandés par Agilent. En revanche, le nombre de cycles de PCR semble impacter le nombre de duplications. En effet, un cycle de PCR supplémentaire conduit à un nombre plus élevé de duplicats de séquences.

Nos analyses nous ont permis de faire des hypothèses sur l'impact des conditions expérimentales (la dilution des sondes, la technique de fragmentation, la quantité initiale d'ADN, la qualité de l'ADN, la quantité d'ADN après fragmentation, etc.) du MC-Seq sur les taux de méthylation observés. De plus, étant donné le nombre de répétitions de certaines conditions, il faut rester prudent sur les résultats des tests statistiques. Néanmoins, le fait que des résultats identiques pour le contexte CHG aient été

	BsmapZ mapper			
	% Mapping		N° of reads	
genotype	Aligned pairs	Unique pairs	Aligned pairs	Unique pairs
1-A26	25.3%	20.3%	39490880	31686312
CTRL C	18.9%	15.2%	41125038	33033989
DRA_38	27.8%	22.8%	60245518	49243014
	Bsmap mapper			
1-A26	25.3%	20.3%	39481843	31680268
CTRL C	18.9%	15.2%	41114176	33027203
DRA_38	27.8%	22.8%	60229721	49232946

Tableau 5: Tableau comparatif des taux d'alignement obtenus avec les deux outils Bsmap et BsmapZ sur 3 échantillons WGBS.

trouvés, peut nous conforter dans nos conclusions. Enfin, compte-tenu du temps à notre disposition, nous n'avons fait que des analyses préliminaires qui mériteraient d'être approfondies et complétées.

Nous avons par exemple considéré une couverture minimale de 1X pour analyser les données de séquençage. Par la suite, en accord avec les collaborateurs du projet, un seuil de 10X sur la couverture sera appliqué ce qui conduira à l'élimination des sites de méthylation qui ont un taux de couverture très faible et des données de méthylation singulières par rapport à tous les sites de méthylation d'autres échantillons. En effet, cela semble beaucoup plus pertinent, car selon la littérature, pour les analyses WGBS, une profondeur minimale de 10X est recommandée, avec un nombre de réplicats biologiques au moins égal à 2 [67].

Perspectives

Ce travail a fourni des résultats intéressants qui contribuent à l'avancement du projet EpiTree car il va notamment permettre l'analyse rapide et automatique de plus de 500 échantillons de peupliers et de chênes dont les séquences seront disponibles cet été. Les résultats obtenus seront mis en lien avec les données transcriptomiques et phénotypiques déjà disponibles pour ces mêmes échantillons afin de comprendre l'impact de la méthylation sur l'adaptation aux variations environnementales.

En parallèle, nous allons explorer de nouvelles pistes pour améliorer le pipeline et pour compléter les analyses dans le but d'optimiser les conditions expérimentales de l'approche MC-Seq.

Au niveau du pipeline, étant donné le nombre prévu d'individus à analyser, il a été proposé d'essayer de réduire la taille des fichiers générés (en sortie du pipeline) tout en diminuant le temps de traitement. Il a donc été décidé de ne se concentrer que sur les zones d'intérêt du génome (fichier BED) pour le calcul des taux de méthylation (par le script « Methratio .py »).

Ensuite, il pourrait être intéressant d'ajouter un fichier de log contenant l'ensemble des paramètres de lancement et de configuration des outils permettra d'assurer une traçabilité des analyses tout en facilitant la détection d'erreurs d'exécution.

De même, au niveau des sorties, on pourrait ajouter des représentations complémentaires de la méthylation telles que des heatmaps et regrouper les figures produites au sein d'un unique rapport HTML qui faciliterait une première interprétation rapide des résultats par l'utilisateur.

Nous envisageons aussi de modifier la méthode actuelle de gestion de la « wildcard » (Expression régulière pour faire correspondre les fichiers d'entrée) destinée à collecter le nom des échantillons afin qu'elle ne nécessite plus d'intervention humaine pour renommer les échantillons.

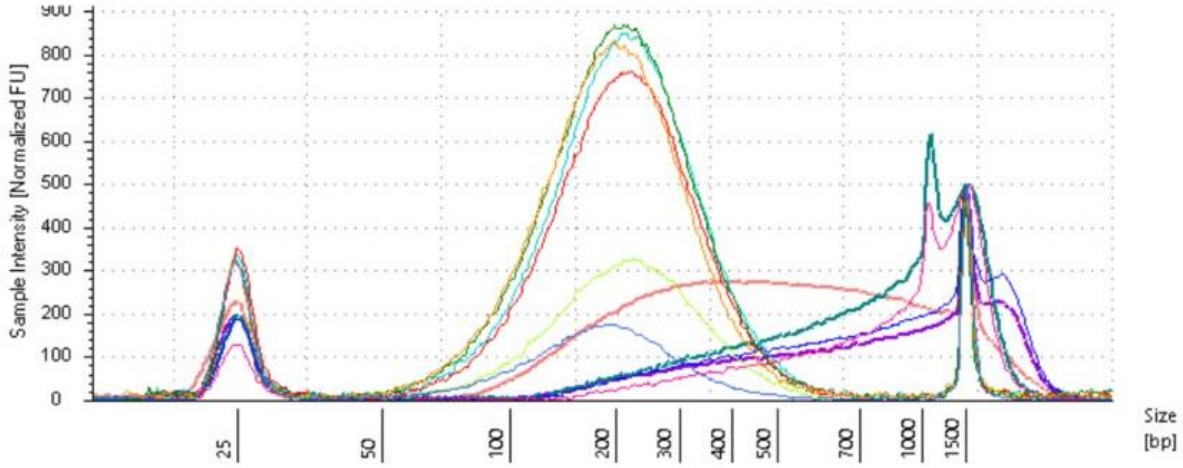


Figure 23: Effets de la fragmentation par sonication Covaris (courbes rouge, orange, bleu turquoise et verte) et de la frangmentation enzymatique (courbes rose, violette, turquoize foncé) sur la taille des fragments d'ADN [68].

Nous pourrions permettre aux utilisateurs de choisir l'aligneur (Bismark, BSSeeker2, BISCUIT et BWA-Meth…) à utiliser. Cela rendrait le pipeline utilisable dans des contextes où l'aligneur BsmapZ n'est pas le plus performant, par exemple lorsque les données d'entrée ne sont pas issues de plantes.

Enfin, transformer le pipeline en outil autosuffisant par son intégration au sein d'un environnement propre (CONDA, par exemple) incluant l'ensemble des packages et logiciels nécessaires à son fonctionnement assurerait une portabilité maximale sur toute plateformes autorisant leur utilisation. Le pipeline ne serait de fait plus que contraint par les ressources à sa disposition sur la machine ou le cluster où il est installé.

Pour ce qui est de l'analyse des taux de méthylation notamment dans l'approche MC-Seq, celle-ci devra être complétée par une étude approfondie du pourcentage d'alignement et du nombre de duplications en lien avec les conditions testées. Ceci pourrait se faire à l'aide de modèle de régression qui pourrait montrer l'impact de certaines conditions expérimentales sur d'autres mesures que les taux de méthylation.

De même, nous comptons ajouter un filtre sur la couverture de 10X (contre 1X actuellement) afin d'augmenter la confiance que l'on a envers les taux de méthylation détectés dans les zones d'intérêt.

Toutes les perspectives décrites ci-dessus sont des exemples de ce qui sera (ou pourrait être) fait d'ici à la fin de mon stage ou à plus long terme, pour compléter le travail que j'ai réalisé au cours de ces 4 mois et demi de stage.

Pour conclure, les résultats trouvés sont innovants : aucune étude comparative de la méthylation de l'ADN entre les techniques de séquençage MC-Seq et WBGS n'ayant été publiée à ce jour, nous comptons valoriser ce travail sous forme d'une publication scientifique.

Conclusion

Au cours de ce stage, j'ai développé un pipeline bioinformatique pour la détection de la méthylation de l'ADN à partir de données de séquençage WGBS et MC-Seq. Ce pipeline, adapté aux besoins du projet EpiTree, est totalement transposable sur d'autres jeux de données de méthylation chez les plantes. Simple d'utilisation et facilement adaptable, ce pipeline développé sous Snakemake est optimisé pour fonctionner sur un cluster de calcul. Comme demandé lors de sa conception, ce pipeline permet de quantifier les niveaux de méthylation par contexte (CG, CHG, CHH) chez le peuplier, données indispensables aux collaborateurs du projet Epitree.

De plus, nous avons validé l'utilisation de la méthode MC-Seq pour l'analyse du méthylome sur une portion du génome en la comparant au WGBS à l'aide d'analyses statistiques. Enfin, nous avons

identifié les conditions expérimentales pouvant être modifiées afin de réduire les coûts tout en permettant d'obtenir les mêmes taux de méthylation que ceux obtenus dans les conditions expérimentales optimales. Parmi celles-ci, notons la quantité initiale d'ADN qui peut être réduite et la concentration des sondes qui peuvent être diluées. Par contre, la fragmentation par sonication Covaris de l'ADN doit être choisie car elle est plus reproductible. En conclusion, les objectifs fixés au début du stage ont tous été atteints.

Références :

[1] IPCC SRCCL 2019, p. 7: Since the pre-industrial period, the land surface air temperature has risen nearly twice as much as the global average temperature (high confidence). Climate change... contributed to desertification and land degradation in many regions (high confidence).; IPCC SRCCL 2019, p. 45: Climate change is playing an increasing role in determining wildfire regimes alongside human activity (medium confidence), with future climate variability expected to enhance the risk and severity of wildfires in many biomes such as tropical rainforests (high confidence).

[2] IPCC SR15 Summary for Policymakers 2018, p. 7

[3] Knutson, T.; Kossin, J. P.; Mears, C.; Perlwitz, J.; Wehner, M. F. (2017). "Chapter 3: Detection and Attribution of Climate Change" (PDF). In USGCRP2017.

[4] "Responding to Climate Change". NASA. 21 December 2020. Archived from the original on 4 January 2021.*

[5] "The State of the Global Climate 2020". World Meteorological Organization. 14 January 2021. Retrieved 3 March 2021.

[6] Chapin III, F. Stuart; Zavaleta, Erika S.; Eviner, Valerie T.; Naylor, Rosamond L.; Vitousek, Peter M.; Reynolds, Heather L.; Hooper, David U.; Lavorel, Sandra; Sala, Osvaldo E. (May 2000). "Consequences of changing biodiversity". Nature. 405 (6783): 234–242. doi:10.1038/35012241. ISSN 0028-0836. PMID 10821284. S2CID 205006508.

[7] Anderegg, W.R.L., Klein, T., Bartlett, M., Sack, L., Pellegrini, A.F.A., Choat, B., Jansen, S., 2016. Meta-analysis reveals that hydraulic traits explain cross-species patterns of drought-induced tree mortality across the globe. Proc Natl Acad Sci USA 113, 5024–5029.

[8] Bruce Toby J.A., Michaela C. Matthes, Johnathan A. Napier, John A. Pickett, Stressful “memories” of plants: Evidence and possible mechanisms,

Plant Science, Volume 173, Issue 6, 2007, Pages 603-608, ISSN 0168-9452

[9] West-Eberhard MJ. 2003. Developmental Plasticity and Evolution. OUP USA.

[10] Nicotra AB, Atkin OK, Bonser SP, Davidson AM, Finnegan EJ, Mathesius U, Poot P, Purugganan MD, Richards CL, Valladares F, et al. 2010. Plant phenotypic plasticity in a changing climate. Trends in Plant Science 15: 684–692.

[13] Dickmann DI, Kuzovkina J. 2008. Poplars and willows of the world, with emphasis on sylviculturally important species (Chapter 2). Poplars and willows in the world: meeting the needs of society and the environment. Rome: FAO/IPC (Food and Agricultural Organization of the United Nations / International Poplar Commission),

[14] G. A. Tuskan, S. DiFazio, S. Jansson et J. Bohlmann, « The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray) », Science, vol. 313, no 5793, 15 septembre 2006, p. 1596–1604 (ISSN 0036-8075 et 1095-9203, PMID 16973872)

[15]https://data.jgi.doe.gov/refinedownload/phytozome?organism=Ptrichocarpa&expanded=444%2C533

[16] Antoine Kremer, Rémy-J. Petit et Alexis Ducousso, « Gene diversity in natural populations of oak species », Annales des Sciences forestières, vol. 54, 1993, p. 186-203 (DOI 10.1051/forest:19930717).

[17] https://www.oakgenome.fr/

[18] Kawakatsu, T., Ecker, J.R., 2019. Diversity and dynamics of DNA methylation: epigenomic resources and tools for crop breedin. Breed. Sci. 69, 191–204.

[19] Rey, O., Danchin, E., Mirouze, M., Loot, C., Blanchet, S., 2016. Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. Trends in Ecology & Evolution 31, 514–526.

[20] Xiao, S., Cao, X., Zhong, S., 2014. Comparative epigenomics: defining and utilizing epigenomic variations across species, time-course, and individuals. Wiley Interdiscip Rev Syst Biol Med 6, 345–352.

[21] Callis J, Vierstra RD. 2000. Protein degradation in signaling. Current opinion in plant biology 3: 381–386.

[22] Exner V, Hennig L. 2008. Chromatin rearrangements in development. Current Opinion in Plant Biology 11: 64–69.

[23] Qiu J. 2006. Epigenetics: unfinished symphony. Nature 441: 143–145

[24] Soria G, Polo SE, Almouzni G. 2012. Prime, repair, restore: the active role of chromatin in the DNA damage response. Molecular cell 46: 722–734.

[25] Andreas Vilcinskas, The role of epigenetics in host–parasite coevolution: lessons from the model host insects Galleria mellonella and Tribolium castaneum,Zoology, Volume 119, Issue 4, 2016, Pages 273-280, ISSN 0944-2006, https://doi.org/10.1016/j.zool.2016.05.004.

[27] Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., Ecker, J.R., 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126, 1189–1201.

[28] Henderson, I.R., Jacobsen, S.E., 2007. Epigenetic inheritance in plants. Nature 447, 418–424.

[29] Chen, Z.J., 2010. Molecular mechanisms of polyploidy and hybrid vigor. Trends in Plant Science 15, 57–71.

[30] Zemach, A., McDaniel, I.E., Silva, P., Zilberman, D., 2010. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science 328, 916–919

[31] Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219

[32] Yong, WS., Hsu, FM. & Chen, PY. Profiling genome-wide DNA methylation. Epigenetics & Chromatin 9, 26 (2016). https://doi.org/10.1186/s13072-016-0075-3

[32] Kernaleguen M. et al. (2018) Whole-Genome Bisulfite Sequencing for the Analysis of Genome-Wide DNA Methylation and Hydroxymethylation Patterns at Single-Nucleotide Resolution. In: Jeltsch A., Rots M. (eds) Epigenome Editing. Methods in Molecular Biology, vol 1767. Humana Press, New York, NY.

[33] Wilson GA, Dhami P, Feber A, Cortázar D, Suzuki Y, Schulz R, et al. Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers. Gigascience. 2012;1:3

[33] Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L.; Paul, C. L. (1992-03-01). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". Proceedings of the National Academy of Sciences of the United States of America. 89 (5): 1827–1831.

[34] Teh AL, Pan H, Lin X, et al. Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics. 2016;11(1):36-48. doi:10.1080/15592294.2015.1132136

[35] Bossdorf O, Richards CL, Pigliucci M. 2008. Epigenetics for ecologists. Ecology Letters 11: 106–115.

[36] Richards EJ. 2011. Natural epigenetic variation in plant species: a view from the field. Current Opinion in Plant Biology 14: 204–209.

[37] Ledón-Rettig CC. 2013. Ecological epigenetics: an introduction to the symposium. Integrative and Comparative Biology 53: 307–318.

[38] Bräutigam, K., Cronk, Q., 2018. DNA Methylation and the Evolution of Developmental Complexity in Plants. Front. Plant Sci. 9, 1447.

[39] Sow, M.D., Segura, V., Chamaillard, S., Jorge, V., Delaunay, A., Lafon-Placette, C., Fichot, R., Faivre-Rampant, P., Villar, M., Brignolas, F., Maury, S., 2018b. Narrow-sense heritability and PST estimates of DNA methylation in three Populus nigra L. populations under contrasting water availability. Tree Genetics & Genomes 14.

[40] Plomion, C., Bastien, C., Bogeat-Triboulot, M.-B., Bouffier, L., Déjardin, A., Duplessis, S., Fady, B., Heuertz, M., Le Gac, A.-L., Le Provost, G., Legué, V., Lelu-Walter, M.-A., Leplé, J.-C., Maury, S., Morel, A., Oddou-Muratorio, S., Pilate, G., Sanchez, L., Scotti, I., Scotti-Saintagne, C., Segura, V., Trontin, J.-F., Vacher, C., 2016. Forest tree genomics: 10 achievements from the past 10 years and future prospects. Annals of Forest Science 73, 77–103.

[41] Carbó, M., Iturra, C., Correia, B., Colina, F.J., Meijón, M., Álvarez, J.M., Cañal, M.J., Hasbún, R., Pinto, G., Valledor, L., 2019. Epigenetics in Forest Trees: Keep Calm and Carry On, in: Alvarez-Venegas, R., De-la-Peña, C., Casas-Mollano, J.A. (Eds.), Epigenetics in Plants of Agronomic Importance: Fundamentals and Applications. Springer International Publishing, Cham, pp. 381–403.

[42] Https://www6.inra.fr/epitree-project/

[43] J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics, 28(19) :2520–2522, Oct 2012.

[44] Mölder F, Jablonski KP, Letcher B et al. Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. F1000Research 2021, 10:33

[45] Krueger F: Trim Galore!. [http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/],

[46] Andrews S. FastQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (Accessed 14 September 2011).

[47] Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354

[48] Xi, Y., Li, W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10, 232 (2009). https://doi.org/10.1186/1471-2105-10-232

[49] Xi, Y., Li, W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10, 232 (2009). https://doi.org/10.1186/1471-2105-10-232

[50] G. J. Zynda, Bsmapz, https://github.com/zyndagj/bsmapz, 2019.

[51] Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

[52] Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012). “methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.” Genome Biology, 13(10), R87.

[52] Robert L. Thorndike (December 1953). "Who Belongs in the Family?". Psychometrika. 18 (4): 267–276. doi:10.1007/BF02289263.

[53] Prates ML, Machado JC, Silva LSD, Avelar PS, Prates LL, Mendonça ET, Costa GDD, Cotta RMM. Performance of primary health care according to PCATool instrument: a systematic review. Cien Saude Colet. 2017 Jun;22(6):1881-1893. Portuguese, English. doi: 10.1590/1413-81232017226.14282016. PMID: 28614508.

[54] Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020; 38(3):276–8.

[55] Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, Heath SC. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics. 2018; 35(5):737–42. https://doi.org/10.1093/bioinformatics/bty690.

[56] Graña O, López-Fernández H, Fdez-Riverola F, González Pisano D, Glez-Peña D. Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. Bioinformatics. 2017; 34(8):1414–5.https://doi.org/10.1093/bioinformatics/btx778.

[57] Huang KYY, Huang Y-J, Chen P-Y. Bs-Seeker3: ultrafast pipeline for bisulfite sequencing. BMC Bioinformatics. 2018; 19(1):111.

[58] Shahryary, Y., Hazarika, R.R. & Johannes, F. MethylStar: A fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. BMC Genomics 21, 479 (2020). https://doi.org/10.1186/s12864-020-06886-3

[59] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. doi:10.1038/nbt.3820

[59] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571-2. doi: 10.1093/bioinformatics/btr167. Epub 2011 Apr 14. PMID: 21493656; PMCID: PMC3102221.

[60] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009 Jul 27;10:232. doi: 10.1186/1471-2105-10-232. PMID: 19635165; PMCID: PMC2724425.

[61] Adam Nunn, Christian Otto, Peter F Stadler, David Langenberger, Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis, Briefings in Bioinformatics, 2021;, bbab021, https://doi.org/10.1093/bib/bbab021

[62] Grehl C, Wagner M, Lemnian I, Glaser B, Grosse I. Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. Front Plant Sci. 2020 Feb 28;11:176. doi: 10.3389/fpls.2020.00176. PMID: 32256504; PMCID: PMC7093021.

[63] Adam Nunn, Christian Otto, Peter F Stadler, David Langenberger, Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA

methylation analysis, Briefings in Bioinformatics, 2021;, bbab021, https://doi.org/10.1093/bib/bbab021

[64] Grehl C, Wagner M, Lemnian I, Glaser B, Grosse I. Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. Front Plant Sci. 2020 Feb 28;11:176. doi: 10.3389/fpls.2020.00176. PMID: 32256504; PMCID: PMC7093021.

[65] Thèse https://tel.archives-ouvertes.fr/tel-03142915/document

[66] [33333] Olohan L, Gardiner LJ, Lucaci A, et al. A modified sequence capture approach allowing standard and methylation analyses of the same enriched genomic DNA sample. BMC Genomics. 2018;19(1):250. Published 2018 Apr 13. doi:10.1186/s12864-018-4640-y

[67] M. J. Ziller, K. D. Hansen, A. Meissner, and M. J. Aryee, "Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing," Nature Methods, vol. 12, pp. 230–232, nov 2015.

[68] C. Boury, com. Pers.

Annexe 1 : Les conditions expérimentales de la technique MC-Seq.

Manip_ID	Sample ID	Input ADN ng	Qualité ADN	Fragmentation	real ng input hybrid	ng Hybrid OK	Dilution Probes	Note 01	Note 02	Note 03
E01	REF_populous	3000	Intègre	Sonication Covaris	408,8	**yes**	1	Ref modulo convertion	kit convertion périmé	bon volume
E02	REF_populous	1000	Intègre	Sonication Covaris	278,4	no	1	Ref modulo convertion	kit convertion périmé	bon volume
E03	REF_populous	1000	Intègre	Sonication Covaris	299	no	1/8	repet E04		erreur +Vol hyb block (9 au lieu de 5,6µL)
E04	REF_populous	1000	Intègre	Sonication Covaris	275,6	no	1/8	repet E03		erreur +Vol hyb block (9 au lieu de 5,6µL)
E05	REF_populous	1000	Intègre	*enzymatique*	184,6	**no**	1/8		+1 cycle PCR	erreur +Vol hyb block (9 au lieu de 5,6µL)
E06	populous_03	935	?	*enzymatique*	276	no	1/8			erreur +Vol hyb block (9 au lieu de 5,6µL)
E07	REF_populous	500	Intègre	Sonication Covaris	130	**no**	1/8	Low input	+1 cycle PCR	erreur +Vol hyb block (9 au lieu de 5,6µL)
E08	REF_populous	750	Intègre	Sonication Covaris	178,1	**no**	1/8	Low input	+1 cycle PCR	erreur +Vol hyb block (9 au lieu de 5,6µL)
E09	populous_02	1000	?	Sonication Covaris	304,2	no	1/8			erreur +Vol hyb block (9 au lieu de 5,6µL)
E10	populous_04	1000	?	Sonication Covaris	165,1	**no**	1/8		+1 cycle PCR	erreur +Vol hyb block (9 au lieu de 5,6µL)
E11	populous_08	1000	?	Sonication Covaris	267,8	no	1/8			erreur +Vol hyb block (9 au lieu de 5,6µL)
E12	REF_populous	1000	Intègre	Sonication Covaris	278,2	no	1/10			erreur +Vol hyb block (9 au lieu de 5,6µL)
E13	REF_populous	1000	Intègre	Sonication Covaris	239,2	no	1/16			erreur +Vol hyb block (9 au lieu de 5,6µL)
E27	REF_populous	1000	Intègre	Sonication Covaris	238	no	1	Real Ref		bon volume
E28	REF_populous	1000	Intègre	Sonication Covaris	301	no	1/8	repet E29		bon volume
E29	REF_populous	1000	Intègre	Sonication Covaris	301	no	1/8	repet E28		bon volume
E30	REF_populous	600	Intègre	Sonication Covaris	126	**no**	1/8	Low input	+1 cycle PCR	bon volume
E31	REF_populous	1000	Intègre	Sonication Covaris	273	no	1/10			bon volume

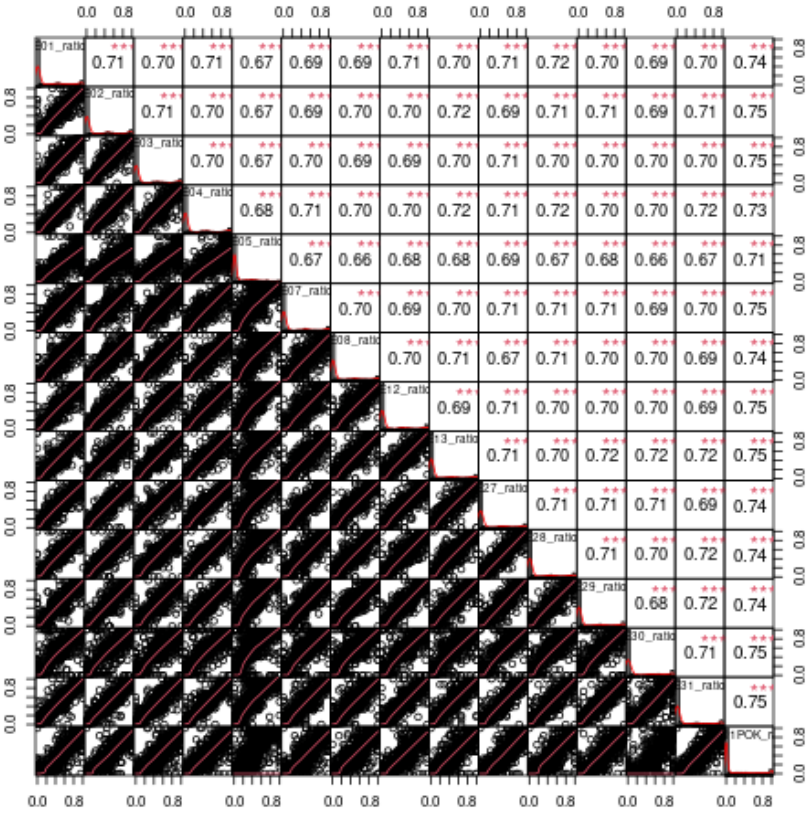
Annexe 2 : Les conditions expérimentales de la technique MC-Seq pour la partie post analyse statistique.

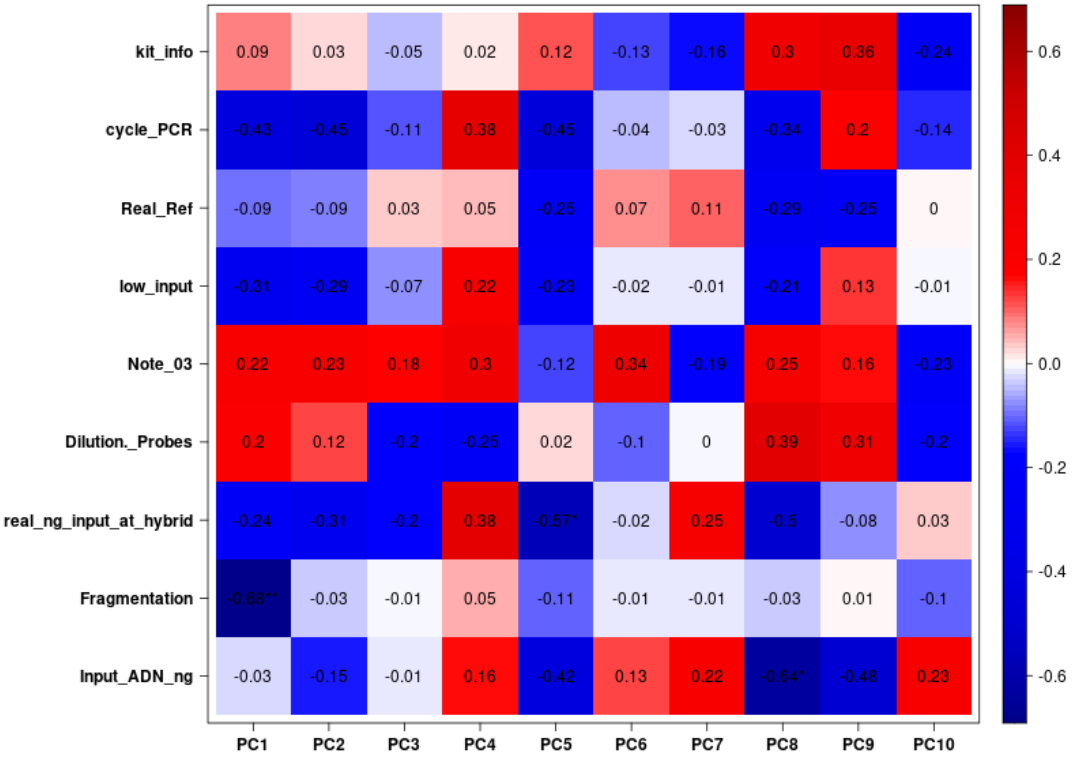
Manip_ID	Sample_ID	Input_ADN_ng	Qualite_ADN	Fragmentation	real_ng_input_at_hybrid	ng_Hybrid_OK	Dilution._Probes	Note_03	low_input	Real_Ref	cycle_PCR	kit_info
E01	REF_populous	3000	Integre	Sonication Co	408,8	yes	1	bon volume	normal	Ref modulo c	normal	kit convertion bisulfite expired
E02	REF_populous	1000	Integre	Sonication Co	278,4	no	1	bon volume	normal	Ref modulo c	normal	kit convertion bisulfite expired
E03	REF_populous	1000	Integre	Sonication Co	299	no	1/8	erreur +Vol hy	normal	normal	normal	normal
E04	REF_populous	1000	Integre	Sonication Co	275,6	no	1/8	erreur +Vol hy	normal	normal	normal	normal
E05	REF_populous	1000	Integre	enzymatique	184,6	no	1/8	erreur +Vol hy	Low input	normal	+1 cycle PCR	normal
E07	REF_populous	500	Integre	Sonication Co	130	no	1/8	erreur +Vol hy	Low input	normal	+1 cycle PCR	normal
E08	REF_populous	750	Integre	Sonication Co	178,1	no	1/8	erreur +Vol hy	Low input	normal	+1 cycle PCR	normal
E12	REF_populous	1000	Integre	Sonication Co	278,2	no	1/10	erreur +Vol hy	normal	normal	normal	normal
E13	REF_populous	1000	Integre	Sonication Co	239,2	no	1/16	erreur +Vol hy	normal	normal	normal	normal
E27	REF_populous	1000	Integre	Sonication Co	238	no	1	bon volume	normal	Real Ref	normal	normal
E28	REF_populous	1000	Integre	Sonication Co	301	no	1/8	bon volume	normal	normal	normal	normal
E29	REF_populous	1000	Integre	Sonication Co	301	no	1/8	bon volume	normal	normal	normal	normal
E30	REF_populous	600	Integre	Sonication Co	126	no	1/8	bon volume	Low input	normal	+1 cycle PCR	normal
E31	REF_populous	1000	Integre	Sonication Co	273	no	1/10	bon volume	normal	normal	normal	normal
C001POK	WGBS	NA	WGBS	WGBS	NA	WGBS	WGBS	WGBS	WGBS	WGBS	WGBS	WGBS

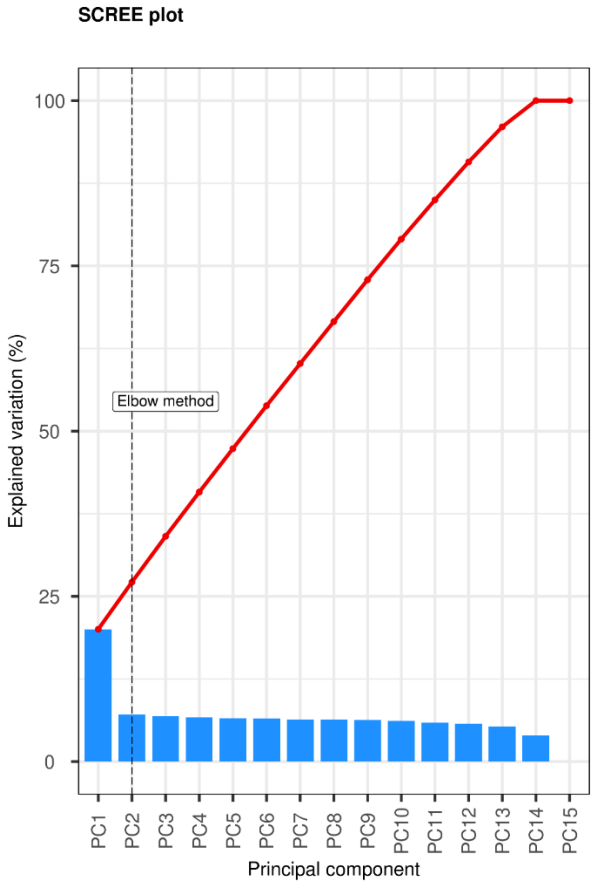
Annexe 3 : Les résultats obtenu après les calculs des On-Target

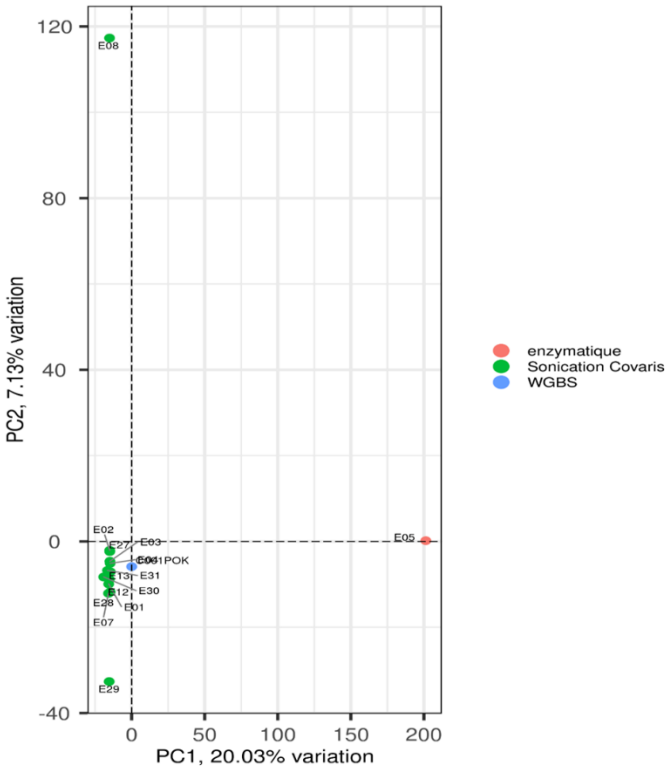
Techniques	Echantillons	Reads_OnTarget_Dedup	Total_Reads_Dedup	OnTargetPerc	Alignment_Perc	Duplicate_Reads	Total_Reads	Duplicate_Perc	CG	CHG	CHH	Total_mC
MC-Seq	E01	13133801	20546048	63,9	45,9	7765283	28311331	27,4	1651768	2460998	13402495	17515261
	E02	14017001	22711791	61,7	46,9	8685616	31397407	27,7	1762876	2609903	14119374	18492153
	E03	10651863	17633431	60,4	45,3	9119175	26752606	34,1	1691561	2513765	13517857	17723183
	E04	10073926	17044018	59,1	46,1	9067583	26111601	34,7	1736287	2585860	13673900	17996047
	E05	9094598	16030632	56,7	43	11315224	27345856	41,4	1731123	2544811	13331464	17607398
	E06	8662341	15129317	57,3	43,4	8896723	24026040	37	1901994	2859895	15142331	19904220
	E07	8268515	15118921	54,7	46,1	10912310	26031231	41,9	1668747	2444243	12789973	16902963
	E08	8920429	15081788	59,1	45,8	7464906	22546694	33,1	1631743	2430118	12886868	16948729
	E09	9793782	16523325	59,3	44,1	8372114	24895439	33,6	1757730	2622119	14200412	18580261
	E10	8772312	15720448	55,8	45,2	12629868	28350316	44,5	1831152	2760750	14678900	19270802
	E11	10058320	16830791	59,8	45,3	8898483	25729274	34,6	1706486	2541907	13688207	17936600
	E12	9918522	16427729	60,4	45,6	8418616	24846345	33,9	1696627	2535353	13567280	17799260
	E13	9925942	17202914	57,7	46,1	12483953	29686867	42,1	1941441	2975150	15437966	20354557
	E27	14767909	25565380	57,8	47,9	11301825	36867205	30,7	1918427	2822291	14809802	19550520
	E28	10707708	17464804	61,3	46,8	7286989	24751793	29,4	1774362	2676680	14225631	18676673
	E29	10270114	16603263	61,9	46,7	6580009	23183272	28,4	1711015	2575099	13724648	18010762
	E30	8421205	16516779	51	48,8	10988504	27505283	40	1708586	2461130	12013675	16183391
	E31	10180699	17283844	58,9	47,2	8591143	25874987	33,2	1763456	2613577	13715239	18092272
WGBS	DRA 38	_	_	_	44.0	30339122	223580450	_	8559982	14349389	78751719	101661090

Annexe 4 : Résultats de la corrélation et ACP pour le contexte CHG



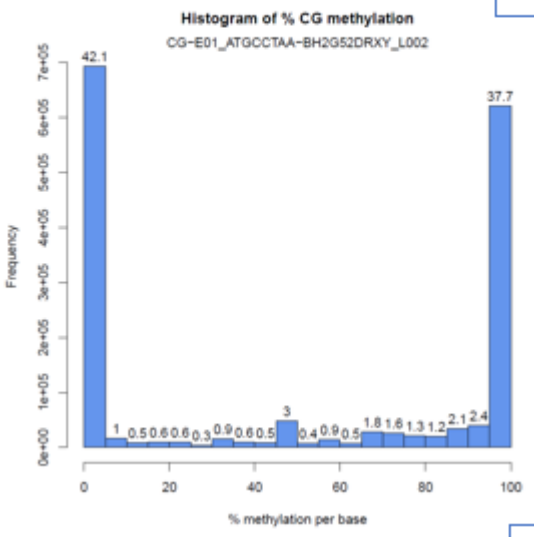


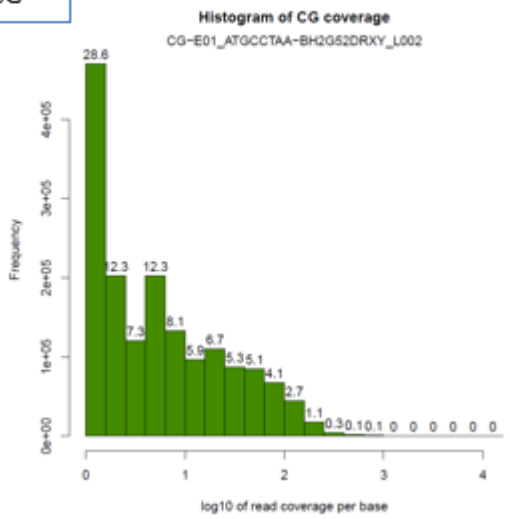

SCREE plot
Explained variation (%)
100
75
50
25
0
Elbow method
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
PC12
PC13
PC14
PC15
Principal component


PC2, 7.13% variation
120
80
40
0
-40
E08
E02
E27
E03
E05
E31
E30
E01
E07
E29
0
50
100
150
200
PC1, 20.03% variation
enzymatique
Sonication Covaris
WGBS

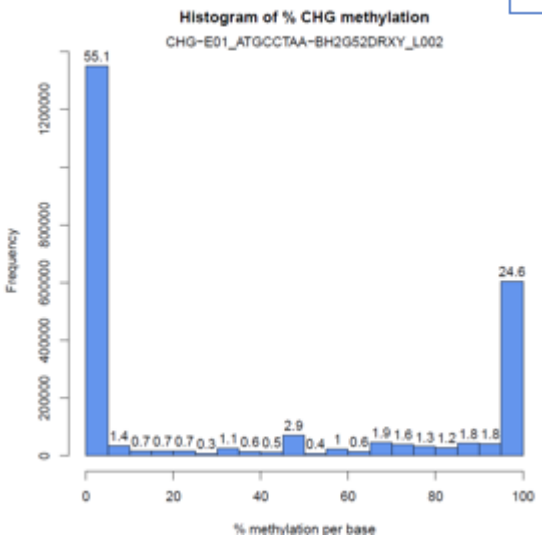
Annexe 5: Résultats de la distribution de pourcentage de méthylation et couverture d'un échantillon de MC-Seq

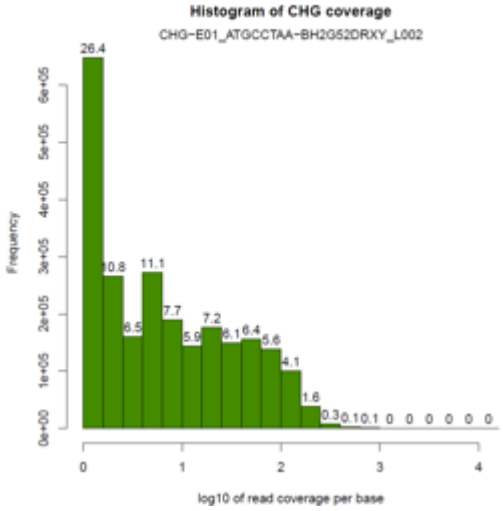
Contexte CG





Contexte CHG





Résumé :

Le peuplier et le chêne sont deux espèces d'arbres pérennes capables de s'adapter aux variations environnementales. Ils subissent le réchauffement climatique qui s'est amorcé il y a 180 ans. De nombreuses études ont montré que le phénotype d'un individu résulte à la fois de sa constitution génétique et de l'effet de son environnement. L'impact de l'environnement (température, sécheresse, etc.) sur l'expression des gènes se traduit notamment par des modifications des marques épigénétiques qui jouent ainsi un rôle dans la plasticité phénotypique des arbres. Aujourd'hui, grâce à des techniques de séquençage à haut débit devenues suffisamment accessibles, il est possible d'étudier les marques épigénétiques au niveau de génomes complets.

Les techniques d'étude de la méthylation par séquençage telles que le WGBS (Whole Genome Sequencing Bisulfite) et le MC-Seq (Methylation Capture Sequencing) sont des applications récentes du séquençage NGS qui permettent la détection de la méthylation à un très haut niveau de résolution. Le traitement au bisulfite de sodium qui consiste en la conversion des cytosines non méthylées en uracile rend possible la détection différentielle des deux types de cytosine dans l'ADN. Le développement de ces nouvelles technologies de biologie moléculaire rend indispensable l'adaptation des outils bio-informatiques à l'analyse de données de séquençage de méthylation.

Dans le cadre du projet ANR EpiTree portant sur l'étude de l'impact des variations environnementales sur le chêne et le peuplier, j'ai développé un pipeline bioinformatique de détection de méthylation de l'ADN. Ce pipeline comprend 6 étapes: (1) nettoyage des données brutes suivi d'un contrôle qualité, (2) alignement sur un génome de référence, (3) élimination des duplications, (4) détection des cytosines méthylées (mC) dans les trois contextes de méthylation, (5) extraction par contexte et (6) quelques analyses statistiques de base sur la détection des méthylations. Ce pipeline est implémenté sous snakemake et il est fonctionnel sur le cluster de calcul du CEA.

Compte tenu du coût du séquençage bisulfite du génome complet (WGBS), j'ai testé la possibilité de limiter l'étude des taux de méthylations de l'ADN à un sous-ensemble du génome (MC-Seq) correspondant à des zones d'intérêt définies par les collaborateurs du projet Epitree. Pour ce faire, j'ai comparé les résultats obtenus après séquençage de ces zones d'intérêt par les 2 approches : 1 WGBS et 18 MC-Seq. Cette comparaison m'a permis de contribuer à l'identification des conditions expérimentales de MC-Seq les moins onéreuses qui seront utilisées pour étudier 500 individus de chêne et de peuplier. De même, les niveaux de méthylationdétectés à partir de séquençage MC-Seq chez ces individus seront obtenus à l'aide du pipeline que j'ai développé.