Analyse bio-informatique de données issues de RNA-seq de populations naturelles de peupliers noirs

Alae Eddine Lekchiri

UNIVERSITÉ DE RENNES

MASTER 1 BIO-INFORMATIQUE

2022-2023

Analyse bio-informatique de données issues de RNA-seq de populations naturelles de peupliers noirs

Auteur : Alae-Eddine LEKCHIRI

Encadré par : Odile Rogier

Harold Duruflé

Alexandre Duplan

Equipe d'accueil

Centre INRA Val de Loire - Site d'Orléans

Équipe Génétique, Adaptation et Amélioration (GA2)

L'Unité Mixte de Recherche Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt (UMR 0588)

2163 Avenue de la Pomme de Pin - CS 40001 ARDON - 45075 ORLEANS Cedex 2

Université de Rennes

UFR Sciences de la vie & de l'environnement

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) Alae-Eddine LEKCHIRI

Etudiant (e) en Master 1 Bio-informatique

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature

INSERM U1242 OSS Equipe PROSAC

Centre Eugène Marquis Avenue de la Bataille Flandres Dunkerque
35042 Rennes

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr

TÉL. 33 (0)2 23 23 61 14

Remerciements :

Ce rapport de Master 1 marque la conclusion d'une expérience scientifique enrichissante menée au sein de l'INRAE d'Orléans avec l'équipe GA2. Je tiens à exprimer mes sincères remerciements à :

- Harold Duruflé : pour son soutien, sa confiance et son encadrement tout au long du stage.
- Odile Rogier : pour ses conseils, son encadrement et son expérience professionnelle précieuse.
- Alexandre Duplan : pour sa disponibilité, ses idées et sa sympathie.
- Leopoldo SANCHEZ RODRIGUEZ : le directeur de l'unité BioForA de m'avoir accueilli durant ce stage. Je tiens également à remercier toute l'équipe GA2 pour leur collaboration et leur contribution à cette expérience.

Leur contribution a été essentielle pour la réalisation de ce travail et j'apprécie grandement leur soutien tout au long de ce parcours.

Table des matières

I - INTRODUCTION

I-1 Contexte scientifique

Les forêts jouent un rôle majeur dans l'environnement, tels que dans le cycle de l'eau 1, la préservation des sols et le stockage du carbone et donc dans la régulation du climat [2]. Les arbres sont des organismes pérennes et ils doivent faire face et s'adapter aux divers changements environnementaux [3].

Mon unité d'accueil BioForA (INRAE, ONF) travaille à comprendre les facteurs moléculaires et génétiques qui influencent la production de biomasse et de bois chez les arbres forestiers, tant d'un point de vue qualitatif que quantitatif. Ainsi, les projet ANR SYBIOPOP et EPITREE, dans lequel s'inscrit mon stage, explore les interactions épistatiques (interactions gène-gène) et les modifications épigénétiques, qui pourraient jouer un rôle essentiel dans la détermination des caractères d'intérêt (biomasse, qualité du bois, résistance au stresses) [4]. De plus, en utilisant l'expression des gènes comme un phénotype intermédiaire, ce projet cherche à éclairer la manière dont les variations génétiques impactent les mécanismes moléculaires et les processus biologiques associés à la production de biomasse.

I-2 Modèle d'étude : le peuplier noir

I-2.1 Intérêt général

Le peuplier, du genre *Populus*, est un modèle d'étude scientifique similaire à *Arabidopsis thaliana,* modèle couramment utilisé pour les plantes annuelles, pour ces caractéristiques intéressantes telles que : sa facilité de culture pour une espèce pérenne, les possibilités de bouturage et de transformation génétique ainsi que sa capacité à produire du bois [5].

Il est important de noter que le peuplier noir constitue également un intérêt écologique [6]. En effet, cette espèce dite pionnière joue un rôle important pour la biodiversité, en hébergeant des écosystèmes complexes (insectes, oiseaux et chauves-souris) et possède un système racinaire très développé, lui permettant de piéger les sédiments et ainsi de préserver les berges [6]. Cette espèce est dite bio-indicatrice du dynamisme de son écosystème.

La répartition du peuplier noir dans le monde s'étend de l'Europe jusqu'à l'Asie centrale et aux côtes de l'Afrique du Nord. Ainsi, ces limites de croissance climatique vont de la Méditerranée,

au sud, jusqu'au 64º de latitude au nord, et des Iles Britanniques jusqu'à la Chine (figure 1A) [7]

.

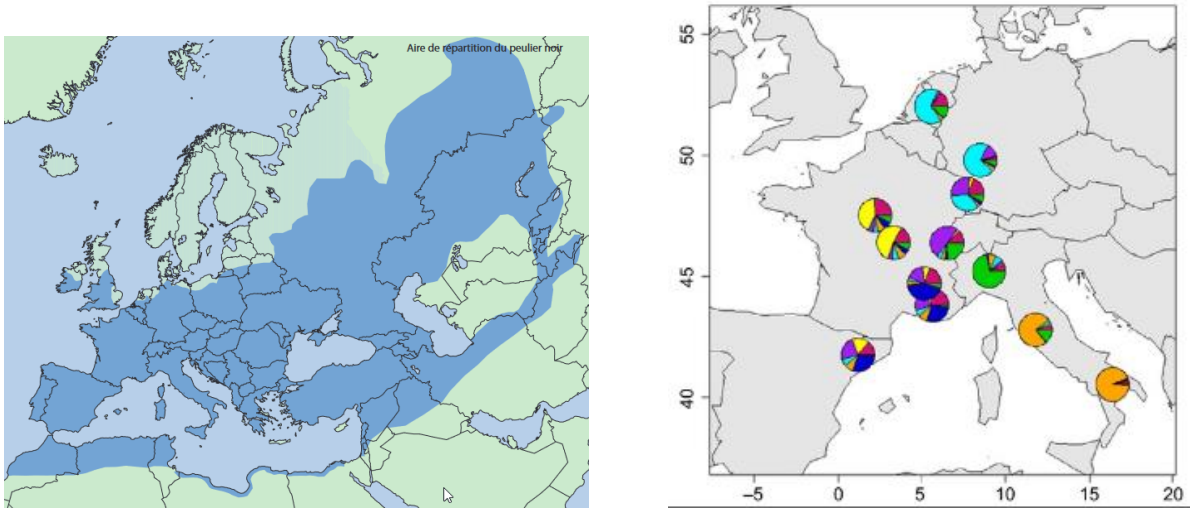

Figure 1 : A) Carte de l'aire de répartition de l'espèce[8] *Populus nigra*. B) Répartition géographique et structure génétique révélée par ADMIXTURE[9] des populations étudiées dans le cadre du projet ANR SYBIOPOP.

Depuis 2001, l'INRAE et le FCBA (Forêt Cellulose Bois construction Ameublement) travaillent conjointement dans le cadre d'une collaboration, le Groupement d'Intérêt Scientifique (GIS) peuplier, afin de produire des hybrides de peupliers performants et résistants à des stress biotique (maladie comme la rouille) et abiotique (sécheresse) [10].

I-2.2 Diversité génétique

Le peuplier noir est une espèce qui mérite une attention particulière en matière de conservation tant au niveau national qu'européen. Des études sur la variation génétique basées sur les microsatellites ont révélé une grande diversité génétique au sein des populations, ainsi qu'une différenciation génétique faible, mais significative entre les bassins fluviaux. Ces résultats suggèrent des niveaux élevés de flux de gènes entre les populations [11]. Une étude plus récente basée sur l'identification des SNP à l'échelle du génome entier à confirmer cela (figure 1B) [9].

I-2.3 Génome du peuplier

Le génome du peuplier (*Populus trichocarpa*) a été le premier génome d'arbre entièrement séquencé en 2006 [12]. Le choix du peuplier pour être le premier organisme ligneux séquencé s'explique par son génome relativement petit, estimé à 485 ± 10 Mb, soit environ 50 fois plus

petit que le génome du pin [13]. Le génome du peuplier est composé de 19 chromosomes et est diploïde, c'est-à-dire que ces cellules possèdent un jeu double de chromosomes semblables. L'analyse du génome a montré un événement de duplication récent, qui a conduit au dénombrement d'environ 8 000 paires de gènes dupliqués dans le génome du peuplier qui en compte un total de 34 700.[13]
Des assemblages du génome du peuplier noir (*Populus nigra* L.) sont en cours, mais ne sont pas disponibles à ce jour.

I-3 Transcriptomique

La transcriptomique désigne l'étude du transcriptome, permettant l'analyse qualitative ou quantitative des molécules d'ARN exprimées dans les cellules d'un organisme donné. [14]

Le RNA-Seq est une technique basée sur les technologies de séquençage de court fragment de nouvelle génération. Elle permet de quantifier en profondeur les échantillons d'ARN, qu'il s'agisse de l'ARN total ou d'une fraction spécifique [15]. Le processus général consiste à convertir ces ARN en une bibliothèque de fragments d'ADN complémentaires (ADNc) en y attachant des adaptateurs à l'une ou aux deux extrémités [15].

Les molécules d'ADNc, seront ensuite soumises à un séquençage à haut débit, générant ainsi des séquences courtes à partir d'une seule extrémité (single-end) ou des deux extrémités (paired-end)[15]. La longueur des lectures varie généralement de 100 à 300 paires de bases (bp), selon la technologie de séquençage d'ADN utilisée [16]. En principe, toutes les technologies de séquençage à haut débit peuvent être utilisées pour le RNA-Seq, notamment Illumina HiSeq 3000/HiSeq 4000, qui a déjà été appliquée à cette fin [15].

I-4 Présentation du sujet

L'objectif de ce stage est d'étudier les profils d'expression transcriptomique d'un panel de 241 peupliers noirs provenant de l'ouest de l'aire de répartition de cette espèce [17].
Mon étude porte principalement sur l'analyse de deux tissus spécifiques du bois : le cambium et le jeune xylème. Le cambium est une couche de cellules dans les racines et les tiges de certaines plantes qui se divisent pour produire de nouveaux tissus, comme le xylème et le phloème[18]. Le xylème est un tissu spécialisé présent chez les plantes vasculaires. Il assure le transport de l'eau et des nutriments depuis l'interface sol-plante jusqu'aux tiges et aux feuilles.

En plus de sa fonction de transport, le xylème offre également un soutien mécanique et permet le stockage de certaines substances [19].

Mon travail consiste à mettre en place un pipeline d'analyse bio-informatique de données de RNA-seq pour obtenir une quantification de l'abondance des transcrits (le nombre de lectures alignées par gène et par génotype) sur la version la plus récente du génome de *P.trichocarpa* (v4). Une première analyse avait déjà été réalisée précédemment [20], mais sur une version précédente du génome (v3) [21].

J'ai, dans un premier temps, testé sur un nombre réduit d'individus différents outils bio-informatiques préconisés dans l'analyse de données transcriptomiques, afin de comprendre leur rôle et me familiariser avec leur utilisation. Dans un second temps, j'ai comparé leurs résultats dans le but de sélectionner les outils pour le pipeline de type nf-core/Rnaseq . Enfin, j'ai lancé le pipeline sur l'ensemble des 461 échantillons puis lancé quelques analyses préliminaires pour vérifier la qualité des résultats obtenus.

II- MATERIELS ET METHODES

II-1 Matériel biologique et séquençage

Des échantillons de tissus de jeune xylème et de cambium provenant de deux arbres par génotype ont été prélevés [20]. Les tiges les plus vigoureuses ont été coupées, l'écorce a été retirée et les jeunes tissus de différenciation du jeune xylème et du cambium ont été grattés [20]. Les tissus ont été immédiatement congelés à l'azote liquide en vue de l'extraction de l'ARN [20].

Les échantillons de tissus ont été finement broyés et l'ARN total du xylème et du cambium a été isolé séparément à partir de chaque plante en utilisant un broyeur oscillant [20]. Une purification de l'ARN a été effectuée avec un traitement à la DNase I pour éliminer l'ADN génomique. L'ARN a été quantifié et stocké dans de l'eau sans ARNase-ADNase [20].

Les échantillons d'ARN du jeune xylème et du cambium de chaque plante ont été regroupés dans des extraits équimolaires avant d'être envoyés à la plateforme de séquençage RNA-Seq POPS (Plateforme transcriptOmic de l'Institut des Sciences du Végétal - Paris-Saclay) [20]. Les bibliothèques de séquençage d'ARN ont été préparées à partir de la sélection de l'ARN messager polyadénylé (polyARNA). Les échantillons ont été multiplexés et séquencés à une

seule extrémité (Single-End) sur un séquenceur Illumina HiSeq2000. Chaque échantillon a généré plus de 20 millions de lectures Single-End de 100 paires de bases (pb) [20].

Les échantillons ont été nommés en utilisant le code de nomenclature suivant : population_génotype_bloc (Tableau 1).

Nom de la population	Pays	Nombre de génotypes
Adour	France	36
Basento	Italie	5
Dranse	France	16
Kuhkopf	Allemagne	19
Loire	France	34
NL	Pays-Bas	4
Paglia	Italie	13
Ramières	France	26
Rhin	France	15
Ticino	Italie	54
ValAllier	France	19

Tableau 1 : Récapitulatif des échantillons utilisés

II-2 Cluster de calcul

En raison du volume important de données brutes (700 Go), nous avons opté pour l'utilisation d'un cluster de calcul.

Pour mener à bien nos travaux, nous avons choisi d'utiliser la plateforme GenoToul. Lancée en 2000, cette plateforme dispose d'une infrastructure robuste qui comprend les équipements suivants depuis 2009 :

- Une ferme informatique composée d'environ 3000 cœurs et 600 threads, avec une capacité mémoire totale de 36 To. Chaque machine dispose de 3 To de mémoire dans le cas d'une machine SMP (symmetric multiprocessing). L'interconnexion InfiniBand (QDR/FDR) assure des performances élevées en termes de communication entre les nœuds du système [22].

- Un système de fichiers parallèle (GPFS) est mis en place pour gérer efficacement les énormes volumes de données générées par les séquenceurs de deuxième et troisième génération [22].

- La plateforme dispose de plus de 6 Po (pétaoctet) d'espace disque ce qui représente une capacité de stockage considérable [22].

Cette plateforme utilise le système de gestion de travaux Slurm. Pour sélectionner les outils nécessaires, il est d'abord nécessaire de vérifier leur disponibilité, puis de créer et paramétrer le script d'exécution. Lorsque le script est soumis pour exécution, il sera placé en file d'attente si les ressources nécessaires ne sont pas disponibles.

II-3 Outils pour le pipeline d'analyse

II-3-1 Langage de programmation

GenoToul utilise le système d'exploitation CentLinux release 7.9.2009, ce qui a permis d'effectuer la manipulation de tous les outils de bio-informatique à l'aide de scripts bash. En ce qui concerne le traitement et la visualisation des données, j'ai utilisé principalement les langages de programmation :

- Python version 3 (via le conteneur Anaconda3, en utilisant spyder V5.4.1) avec les packages suivants : OS, Pandas, Matplotlib, Numpy.
- R (version 4.3.0) (via Rstudio) avec les librairies R suivantes : tidyverse, ggplot2, dplyr, tidyr, mixOmics, heatmaply.

II-3-2 Téléchargement des données brutes de séquençage

Les données sont disponibles en accès libre sur le dépôt NCBI (National Center for Biotechnology Information) sous le BioProject 527833. L'outil Sratoolkit-V.2.11.3[23] offre la possibilité de télécharger les données en les convertissant (entre autres) du format SRA au format fastQ. J'ai donc téléchargé les données brutes sur la plateforme GenoToul.

II-3-3 Contrôle qualité et nettoyage des données brutes

Le contrôle qualité des données de séquençage à haut débit a été effectué à l'aide de FastQC-v0.11.2[24]. Trim Galore-V.0.6.5[25] est un outil bioinformatique qui agit comme un wrapper

(enveloppe) autour des outils Cutadapt-V.1.14[26] et FastQC, facilitant ainsi l'application cohérente de ces outils : retirer les adaptateurs et évaluer la qualité des fichiers FastQ.

J'ai également effectué du contrôle qualité avec Multiqc-v.1.14 [27]. Car cet outil permet de générer des rapports d'analyse en regroupant les résultats et les statistiques produites par différents outils bio-informatiques, et cela pour un grand nombre d'échantillons simultanément.

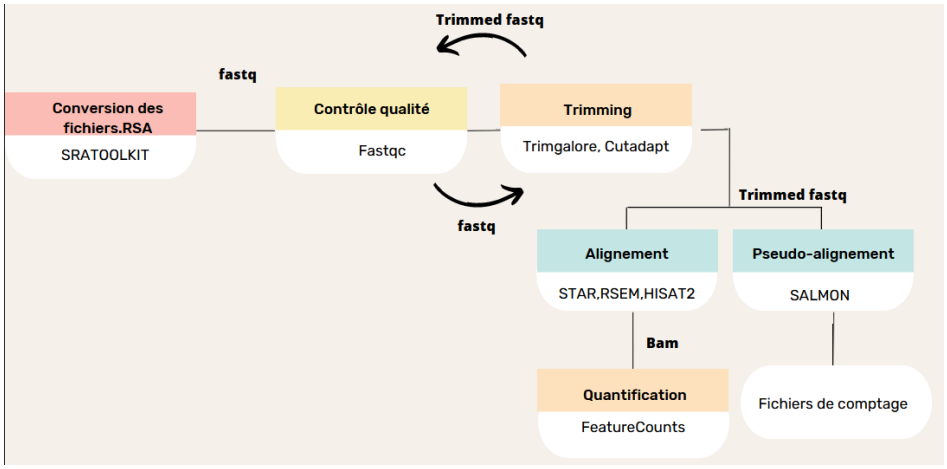

Figure 2 : Pipeline appliqué lors de ce stage.

II-3-4 Outils d'alignement

L'une des étapes fondamentales de l'analyse des données RNA-seq est l'alignement des lectures (reads) séquencées sur un génome de référence. Les outils qui vont suivre passent par deux étapes majeures.

1. L'indexation du génome de référence : Processus indépendant des lectures de séquençage, elle s'effectue une seule fois pour un génome de référence [28]. Cette étape permet de créer un index, ce qui permet d'accélérer l'étape ultérieure.
2. Alignement : Alignement des ensembles importants de lectures séquencées consiste à positionner ces lectures de manière précise et cohérente sur un génome de référence donné [29].

STAR-2.4.0i [30](Spliced Transcripts Alignment to a Reference) est un logiciel bioinformatique puissant qui offre un alignement précis et rapide des lectures de l'RNA-seq sur un génome de

référence. HISAT2-2.1.0[31] est un logiciel d'alignement conçu pour cartographier de manière rapide et précise les lectures issues de séquençage de nouvelle génération (NGS) sur un génome de référence spécifique.

Bowtie2-2.4.4[32] c'est un aligneur qui combine les points forts de l'index des minutes en texte intégral avec la flexibilité et la vitesse des algorithmes de programmation dynamique accélérés par le matériel pour obtenir une combinaison de vitesse élevée, de sensibilité et de précision.

II-3-5 Comptage/ quantification

FeatureCounts [33] est un outil bioinformatique couramment utilisé pour effectuer le comptage des lectures (reads) provenant de séquences d'ARN sur des régions génomiques spécifiques.

RSEM-1.3.3 (RNA-Seq by Expectation-Maximization)[34] est un logiciel qui permet d'estimer les niveaux d'expression des gènes et des isoformes à partir de données de séquençage d'ARN (RNA-Seq). Le package RSEM offre une interface conviviale et facilite l'estimation des niveaux d'expression en utilisant l'algorithme de l'espérance-maximisation (EM).

Salmon-1.9.0[28] est un outil spécialisé dans la quantification rapide et précise des transcriptions à partir de données de séquençage d'ARN (RNA-seq). Il est conçu pour estimer les niveaux d'expression des gènes et des transcrits de manière efficace. À la différence des outils précédents, Salmon ne nécessite pas l'utilisation d'un aligneur séparé. Il est capable de lire directement un fichier au format fastQ et d'effectuer la quantification simultanément. Salmon est considéré comme un pseudo-aligneur, car il utilise une approche différente pour estimer les niveaux d'expression sans effectuer un alignement complet des séquences sur un génome de référence.

II-4 Gestionnaires de workflow : nextflow et nf-core

La figure 3, accessible via le lien (https://nf-co.re/rnaseq), illustre le pipeline rnaseq basé sur nf-core. Ce pipeline est créé en utilisant Nextflow, un système de gestion de flux de travail en utilisant les conteneurs comme Docker et Singularity pour gérer le calcul à différentes échelles à l'aide de conteneurs. Il exploite les avantages des conteneurs pour garantir une reproductibilité des résultats obtenus. [35].

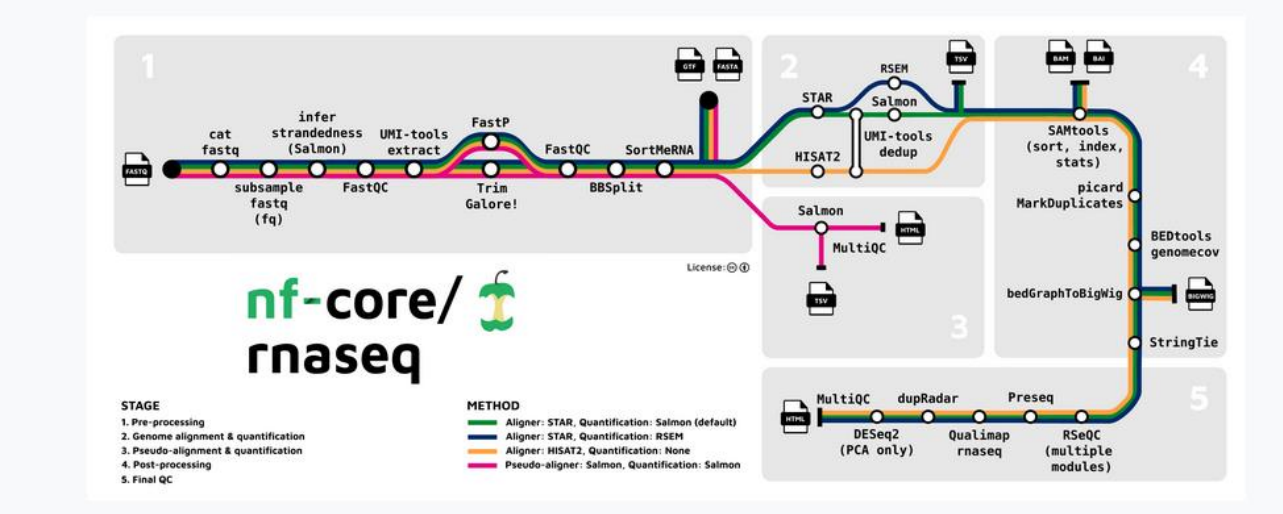

Figure 3 : Pipeline nf-core/rnaseq

III- Résultats

III-1 Test des outils

J'ai initialement sélectionné douze échantillons, représentant deux populations distinctes (Adour et Loire). Cette sélection était destinée à évaluer les outils bio-informatiques dans le cadre du pipeline. Par la suite, j'ai étendu mon échantillonnage (16 échantillons) avec un total de quatre populations (Adour, Loire, Rhin et Ticino) afin de diversifier les échantillons et d'avoir une plus grande variété de données à tester avant de procéder au traitement de l'ensemble des échantillons. Cette approche me permet de réaliser plusieurs tests préliminaires avant de lancer l'analyse complète.

III-1-1 Contrôle qualité des échantillons

Pour effectuer le contrôle qualité des échantillons, j'ai utilisé l'outil FastQC qui prend en entrée des fichiers au format. fastQ. FastQC génère ensuite des rapports au format HTML. À partir de ces fichiers, j'ai utilisé l'outil MultiQC afin de pouvoir regrouper les résultats dans un seul rapport.

III-1-2 Test des outils d'alignement et de quantification

Toujours à partir des seize échantillons que j'avais choisis, j'ai testé plusieurs outils d'alignement, afin de déterminer le plus performant (tableau 2 et figure 4).

Echantillion	Rsem	Salmon	Hisat	Star
Adour_AST-002_1	85.90	91.00	88.50	95.00
Adour_AST-002_3	84.90	90.60	88.90	95.50
Adour_AST-004_1	86.60	92.00	89.50	95.50
Adour_AST-004_3	86.70	91.60	86.30	95.70
Adour_AST-005_1	82.80	90.10	86.50	94.90
Adour_AST-005_3	83.20	90.80	88.30	95.30
Loire_92510-1_1	84.90	90.60	86.10	95.50
Loire_92510-1_3	82.30	89.20	89.30	94.70
Loire_92520-1_1	86.40	91.80	88.80	95.80
Loire_92520-1_3	85.80	91.30	87.70	95.70
Loire_92520-6_1	84.70	90.60	88.70	94.90
Loire_92520-6_3	85.30	90.80	88.05	95.50
moyenne	84.96	90.87	88.05	95.33

Tableau 2 : Moyennes des alignements de 16 échantillons testés sur 4 aligneurs.

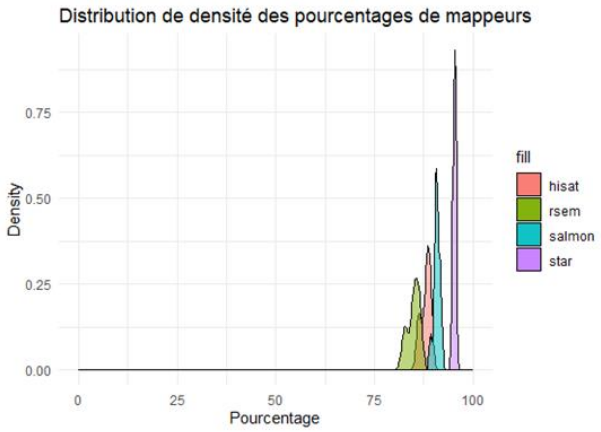

Figure 4 : Distribution des densités des pourcentages d'alignements en fonction des 4 aligneurs choisis.

Pour le comptage des lectures, j'ai opté pour l'utilisation de FeatureCounts, RSEM et Salmon comme outil pour les trois mappeurs, à savoir STAR, HISAT2 et RSEM via Bowtie2 [32].

Suite à son utilisation et après l'obtention des résultats d'alignement via STAR, puis une justification via FeatureCounts j'ai obtenu plusieurs fichiers .txt qui contiennent le nombre de lectures pour chaque échantillon (figure 5).

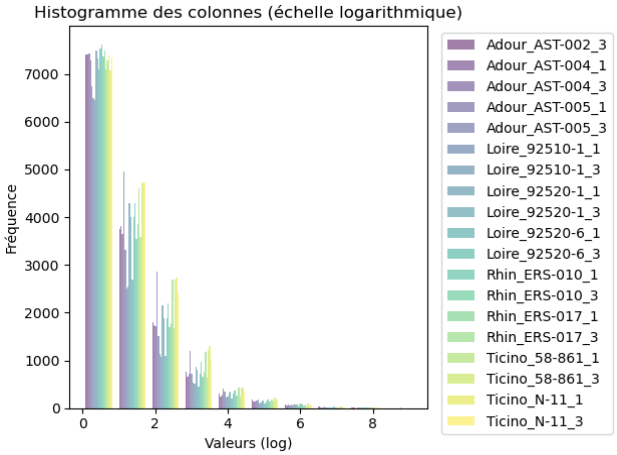

Figure 5 : Distribution des reads pour les 16 échantillons en utilisant STAR comme aligneur et FeatureCounts comme mappeur

III-2 Mise en place et lancement du pipeline nf-core

III-2-1 Choix des outils d'alignement et de quantification pour le pipeline nf-core rna-seq

Pour le pipeline nf-core, j'ai choisi les outils mentionnés dans la figure 6 (fastQC,Trimgalore,STAR,RSEM et Multiqc) disponible via le lien (https://nf-co.re/rnaseq). En ce qui concerne le choix de l'outil de comptage entre RSEM et Salmon dans le cadre du pipeline nf-core, j'ai effectué un test préliminaire sur mes 16 échantillons. Les résultats ont montré que les moyennes de pourcentages d'alignement obtenus étaient respectivement de 96,01% pour RSEM et de 89,90% pour Salmon. Ces résultats indiquent que l'utilisation de RSEM pour le comptage des reads était plus intéressante en termes de pourcentage de mapping.

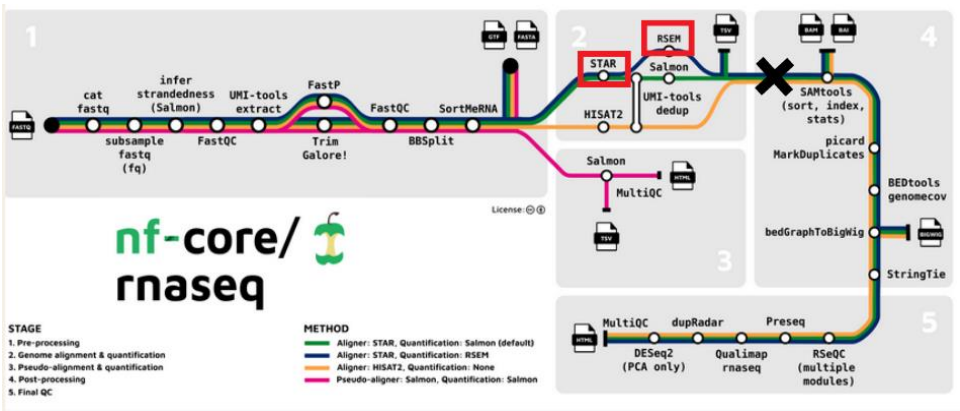

Figure 6 : Pipeline nf-core/rnaseq, les carrés rouges présentent les outils choisits et la croix représente la partie non inclue. (Issue du site Web https://nf-co.re/rnaseq)

III-2-2 Lancement du pipeline nf-core

Après avoir vérifié la disponibilité des modules nécessaires pour l'exécution du pipeline, j'ai créé trois fichiers qui m'ont permis de lancer le pipeline. Le premier fichier est au format JSON et contient tous les paramètres d'entrée, tels que les liens vers les différents répertoires nécessaires. Le deuxième fichier est un fichier de configuration spécifiant les options et les paramètres spécifiques au pipeline. Enfin, le dernier fichier est un script bash qui contient la commande d'exécution permettant de lancer le pipeline.

III-3 Bio-analyse des données RNAseq

III-3-1 Contrôle qualité des reads

Une fois que j'ai effectué un contrôle de qualité sur tous les fichiers bruts, j'ai utilisé l'outil Multiqc pour regrouper toutes les figures en une seule. Il est ainsi observé que les différents fichiers. fastQ présentent des scores de bonne qualité (un score de Phred supérieur à 30, représenté dans la figure 3 en vert). De plus, ils affichent des profils similaires dans l'ensemble.

Figure 7 : Score de qualité des 461 échantillons bruts.

III-3-2 Alignement

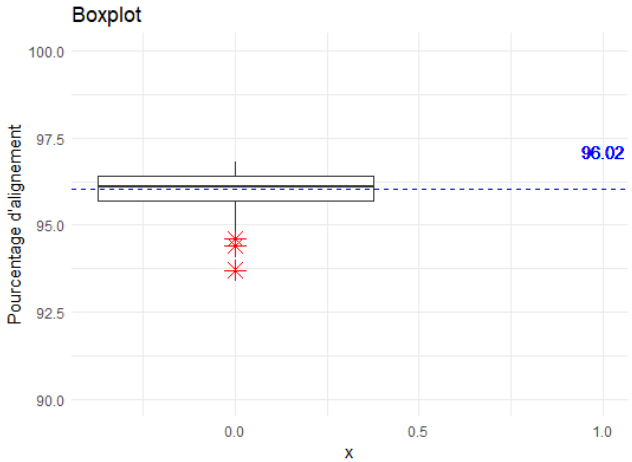

Figure 8 : Graphique en boîte à moustache des pourcentages
D'alignement des 461 génotypes avec la méthode STAR.

La figure montre clairement que la moyenne d'alignement est de 96,02%, à l'exception de sept échantillons. Le pourcentage le plus bas d'alignement est de 93,7%.

III-3-3 Quantification des reads

J'ai alors procédé à une analyse de regroupement (clustering) (Figure 9 A) et à la création d'une carte thermique (heatmap) (Figure 9 B) afin d'observer la répartition des groupes au sein des 16 échantillons.

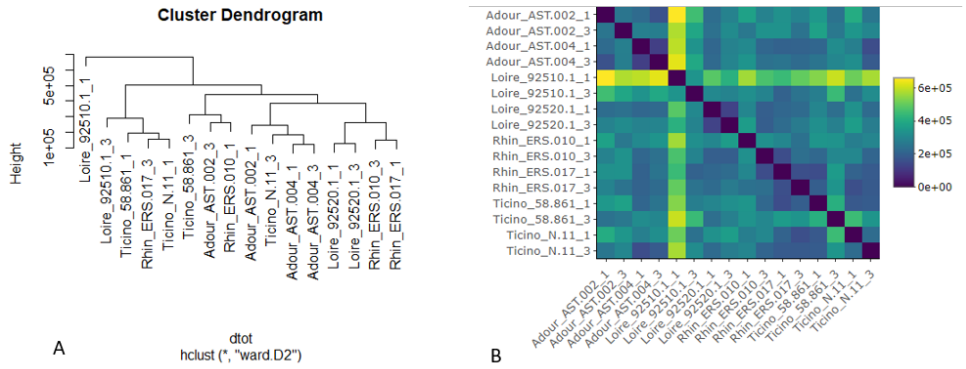

Figure 9 : A analyse de regroupement des 16 échantillons B carte thermique des 16 échantillons à partir des fichiers de comptage.

Afin d'obtenir une meilleure compréhension de la distribution de mes échantillons, j'ai réalisé une Analyse en Composantes Principales (ACP). L'objectif de l'ACP était de regrouper les échantillons en fonction de leurs populations respectives (figure 10).

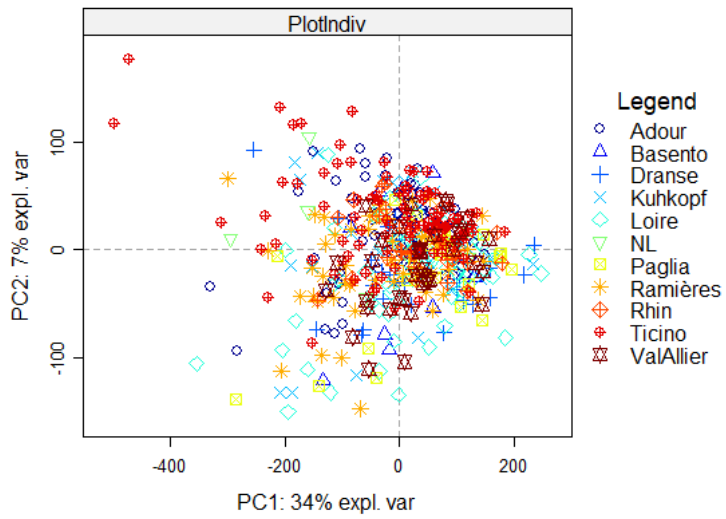

Figure 10 : Analyse en Composante Principale des transcriptomes des 461 échantillons

IV- Discussion

Afin d'approfondir notre compréhension de la diversité génétique au sein des différentes populations chez le peuplier noir, il est possible de mener une étude sur l'expression des gènes.

Pour sélectionner les outils les plus appropriés, j'ai d'abord pris en compte les outils les plus fréquemment mentionnés dans les articles scientifiques[36]. Ensuite, j'ai effectué des tests, tels que l'évaluation du pourcentage d'alignement, afin de choisir le meilleur outil d'alignement. Cette approche a été utilisée à la fois pour le pipeline de test et le pipeline nf-core/rnaseq.

Concernant le pipeline nf-core/rnaseq, le choix des outils est plus restreint, ce qui a limité ma capacité à construire moi-même un pipeline Nexflow en raison de contraintes de temps. Pour les échantillons de test, j'ai sélectionné les 16 échantillons de quatre populations différentes afin de représenter une diversité génétique maximisée. Ce nombre d'échantillons a été choisi dans le but d'économiser les ressources et de gagner du temps, tout en garantissant une diversité adéquate au sein des populations et la présence de réplicats pour chaque échantillon.

Après avoir effectué un contrôle qualité sur les données brutes, comprenant 461 échantillons, il est clair que tous les échantillons sont de bonne qualité (un score de Phred supérieur à 30). Cette observation peut probablement être attribuée au fait que les échantillons ont sûrement subi un prétraitement au sein de la plateforme de séquençage POPS.

Pour approfondir notre analyse du contrôle qualité des résultats, nous avons généré une carte thermique et un dendrogramme. Les résultats révèlent clairement que les réplicats ne sont pas regroupés avec les sous-groupes les plus proches. Cette observation peut être expliquée par le fait que les réplicats sont des clones biologiques, plantés dans un jardin commun, et au fil du temps, ils peuvent avoir évolué différemment en raison de facteurs tels que la position sur la parcelle, le sol, la période de récolte.

D'autre part, le génotype Loire_92510.1_1 se distingue et ne fait partie d'aucun groupe des 16 testés. Cette conclusion est également étayée par la carte thermique. Pour mieux comprendre ces observations, il serait intéressant de réaliser ces mêmes analyses sur l’ensemble des individus pour le comparer à une plus grande diversité.

V- Conclusion et perspectives

En conclusion, les résultats obtenus confirment que le choix du pipeline était judicieux, avec un pourcentage d'alignement significatif. L'analyse du comptage du nombre de lectures, visualisée à l'aide d'une ACP, montre la présence d'un seul bloc, ce qui est cohérent étant donné qu'il s'agit de la même espèce ayant été cultivée en jardin commun.

Après l'obtention de ces résultats il serait intéressant de faire une comparaison avec les résultats obtenus précédemment avec la version 3 du génome du *P.trichocarpa* afin de voir s'il existe des changements majeurs entre les deux versions.

La partie biostatistique jouera un rôle important après l'obtention de ces résultats. Des vérifications approfondies seront effectuées pour évaluer la qualité des données générées. Ensuite, la normalisation sera réalisée dans le but de limiter et réduire les biais introduits par les différentes étapes expérimentales (position des génotypes sur la parcelle, heure et jours de récoltes…,). Cette étape est essentielle pour rendre les échantillons comparables entre eux.

Une fois les données normalisées, elles seront utilisées dans le cadre de la thèse d'Alexandre Duplan. Elles seront alors intégrées avec d'autres données multi-omiques, comme l'épigénomique, pour permettre une étude de prédiction des phénotypes. Cette approche intégrée offrira de nouvelles perspectives et contribuera à une meilleure compréhension des relations entre les différentes variables.

Références

1. The nocturnal water cycle in an open-canopy forest - Berkelhammer - 2013 - Journal of Geophysical Research: Atmospheres - Wiley Online Library. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/jgrd.50701.
2. Forests and forestry. https://www.eea.europa.eu/en/topics/in-depth/forests-and-forestry (2023).
3. Anderson, J. T., Willis, J. H. & Mitchell-Olds, T. Evolutionary genetics of plant adaptation. *Trends Genet.* **27**, 258–266 (2011).
4. Une approche de biologie intégrative pour améliorer le peuplier en vue de sa valorisation en bio-raffinerie grâce à une meilleure compréhension de l'architecture génétique de la production et de la qualité de la biomasse lignocellulosique. *Agence nationale de la recherche* https://anr.fr/Projet-ANR-13-JSV6-0001.
5. Jansson, S. & Douglas, C. Populus: A Model System for Plant Biology. *Annu. Rev. Plant Biol.* **58**, 435–58 (2007).
6. Villar, M. & FORESTIER, O. Le Peuplier noir en France : pourquoi conserver ses ressources génétiques et comment les valoriser ? *Rev. For. Fr.* **61**, (2009).
7. Broeck, V. Fiche tech-nique d'EUFORGEN pour la con-servation des ressources géné-tiques et l'utilisation du peuplier noir (Populus nigra L.). (2003).
9. Faivre-Rampant, P. *et al.* New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. *Mol. Ecol. Resour.* **16**, 1023–1036 (2016).
10. GIS Peuplier - Accueil. https://www6.inrae.fr/gispeuplier/.
11. DeWoody, J., Trewin, H. & Taylor, G. Genetic and morphological differentiation in

Populus nigra L.: isolation by colonization or isolation by adaptation? *Mol. Ecol.* **24**, 2641–2655 (2015).

12. Tuskan, G. A. *et al.* The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).

13. Phytozome info: P.trichocarpa v4.1. https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v4_1.

14. Transcriptomics - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/transcriptomics.

15. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

16. RNA Sequencing | RNA-Seq methods & workflows. https://www.illumina.com/techniques/sequencing/rna-sequencing.html.

17. Chateigner, A. *et al.* Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC Genomics* **21**, 416 (2020).

18. Nieminen, K., Blomster, T., Helariutta, Y. & Mähönen, A. P. Vascular Cambium Development. *Arab. Book Am. Soc. Plant Biol.* **13**, e0177 (2015).

19. Myburg, A. A., Lev-Yadun, S. & Sederoff, R. R. Xylem Structure and Function. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2013). doi:10.1002/9780470015902.a0001302.pub2.

20. Rogier, O. *et al.* Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra. *BMC Genomics* **19**, 909 (2018).

21. Phytozome info: P.trichocarpa v3.1. https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v3_1.

22. Home. *genotoul-bioinfo* https://bioinfo.genotoul.fr/.

23. Heldenbrand, J., Ren, Y., Asmann, Y. & Mainzer, L. S. Step-by-Step guide for downloading very large datasets to a supercomputer using the SRA Toolkit. (2017).

24. Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**, 3137–3139 (2017).

25. Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

26. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

27. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

28. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

29. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **51**, 11.14.1-11.14.19 (2015).

30. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

31. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

33. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program

for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

34. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

35. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).

36. Ji, F. & Sadreyev, R. I. RNA-seq: Basic Bioinformatics Analysis. *Curr. Protoc. Mol. Biol.* **124**, e68 (2018).

Annexes

Annexe 1

L'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) est un institut de recherche publique composé de 18 centres de recherche au cœur de dynamiques régionales. Il est dirigé par Monsieur Philippe MAUGUIN.

Le centre Val de Loire, où s'est déroulé mon stage, mène des recherches autour de quatre axes : « Dynamique des sols et gestion de l'environnement », « Biologie intégrative des arbres et de la biodiversité associée pour la gestion durable des écosystèmes forestiers », « Biologie animale intégrative, durabilité des systèmes d'élevage » et « Infectiologie et "One Health" ». Ces axes de recherche sont répartis sur quatre sites : Orléans, Tours, Nogent-sur-Vernisson et Bourges.

Le centre INRAE qui se situe à Orléans est composé de plusieurs unités. :

• l'Unité Mixte de Recherche Biologie intégrée pour la valorisation de la diversité des arbres et de la forêt (INRAE - ONF, BioForA)

• l'Unité de Recherche Zoologie Forestière (URZF)

• l'Unité Expérimentale Génétique et Biomasse Forestière (GBFOr)

• L'Unité InfoSol est une unité de service (US) du Département AgroEcoSystem d'INRAE.

Durant les deux mois de mon stage, j'étais accueilli au sein de l'Unité Mixte de Recherche (UMR) INRAE-ONF BioForA, au sein de l'équipe "Génétique, Adaptation et Amélioration" (GA2). Cette unité mène des recherches visant à valoriser les ressources génétiques forestières en vue d'une production durable de bois d'œuvre et de biomasse, tout en prenant en compte l'impact écologique des populations domestiquées sur l'écosystème et un contexte climatique changeant.

Annexe 2

Au cours de mon stage, j'ai eu l'opportunité d'explorer le domaine de la transcriptomique qui m'intéressait et d'acquérir des compétences précieuses. J'ai été accueilli au sein de l'unité BioFora, qui m'a réservé un accueil chaleureux et où j'ai pu apprendre beaucoup sur le fonctionnement et l'organisation de la recherche. De plus, j'ai eu la chance d'être encadré par plusieurs mentors qui étaient toujours disponibles lorsque j'avais besoin d'aide. Ce stage m'a permis de me rendre compte de ce que pouvaient être les applications concrètes dans un environnement professionnel de ma formation académique et cela a renforcé ma détermination pour mon projet d'études.

J'ai également eu l'occasion de travailler de manière autonome, ce qui m'a permis de mettre à l'épreuve les compétences que j'ai acquises au cours de ma première année de Master. Une expérience qui m'a particulièrement plu a été l'apprentissage de l'utilisation de la librairie Python OS. Cette bibliothèque m'a permis d'interagir avec le système d'exploitation, ce qui était très utile pour la manipulation de fichiers et de répertoires, une tâche fréquente pendant mon stage.

Chaque jour, j'ai été confronté à des problématiques à résoudre, et c'est ce défi constant qui m'a le plus plu dans mon stage. Si j'avais la possibilité de refaire ce stage, je le referais avec grand plaisir.

Résumé

Le contexte scientifique de ce projet souligne l'importance des forêts dans l'environnement, particulièrement leur rôle dans le cycle de l'eau, la préservation des sols et le stockage du carbone pour la régulation du climat. Le peuplier noir est utilisé comme modèle d'étude de par sa facilité de culture, sa capacité à produire du bois et son intérêt écologique en tant qu'espèce bio-indicatrice. Le projet de recherche dont s'inscrit mon stage, vise à étudier les profils d'expression transcriptomique dans un tissu spécifique du bois. Un pipeline bio-informatique est mis en place pour analyser les données de RNA-seq et quantifier l'abondance des transcrits. L'étude sur l'expression des gènes dans le peuplier noir permet d'approfondir notre compréhension de sa diversité génétique au sein des différentes populations naturelles. Le choix des outils pour cette étude a été basé sur des critères tels que leur fréquence de mention dans la littérature scientifique et des tests d'évaluation, notamment en termes de pourcentage d'alignement. Pour le pipeline nf-core/rnaseq, le choix des outils était plus restreint en raison de contraintes de temps, ce qui a motivé la sélection des échantillons de test les plus représentatifs. Les contrôles qualité effectués sur les 461 échantillons ont montré une bonne qualité globale des données, avec des scores de Phred supérieurs à 30. Cependant, l'analyse du contrôle qualité a révélé des différences entre certains réplicats, probablement dues à des facteurs environnementaux tels que la position sur la parcelle, le sol, la période de récolte. Pour mieux comprendre ces variations, une étude biostatistique plus approfondie sera nécessaire afin d'obtenir des réponses.

Mots clés : Populus, Génotype, Expression transcriptomique, pipeline nf-core/rnaseq, Contrôle qualité.

Abstract : *Bioinformatics analysis of RNA-seq data from natural black poplar populations*

The scientific background to this project highlights the importance of forests in the environment, particularly their role in the water cycle, soil preservation and carbon storage for climate regulation. The black poplar is used as a study model because of its ease of cultivation, its capacity to produce wood and its ecological interest as a bio-indicator species. The EPITREE project aims to study transcriptomic expression profiles in two specific wood tissues. A bioinformatics pipeline is set up to analyse the RNA-seq data and quantify the abundance of transcripts. The study of gene expression in black poplar will enable us to gain a deeper understanding of its genetic diversity within different populations. The choice of tools for this study was based on criteria such as their frequency of mention in the scientific literature and evaluation tests, particularly in terms of percentage of alignment. For the nf-core/rnaseq pipeline, the choice of tools was more limited due to time constraints, which led to the selection of the most representative test samples. The quality checks carried out on the 461 samples showed good overall data quality, with Phred scores above 30. However, quality control analysis revealed differences between some replicates, probably due to environmental factors such as position on the plot, soil, harvesting period. To gain a better understanding of these variations, a more in-depth biostatistical study will be required.

Key words : Populus, Genotype, Transcriptomic expression, nf-core/rnaseq pipeline, Quality control.