



HAL
open science

Caractérisation de QTL contrôlant des traits racinaires chez le Mil

Princia Aurore Biawallys Nakombo-Gbassault

► **To cite this version:**

Princia Aurore Biawallys Nakombo-Gbassault. Caractérisation de QTL contrôlant des traits racinaires chez le Mil. Agronomie. 2020. dumas-03032947

HAL Id: dumas-03032947

<https://dumas.ccsd.cnrs.fr/dumas-03032947>

Submitted on 1 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mémoire de fin d'études

Présenté pour l'obtention du Master 2

**Mention : Sciences et technologies de l'agriculture, de l'alimentation et
de l'environnement**

Parcours : Sélection et évolution des plantes méditerranéennes et tropicales

Caractérisation de QTL contrôlant des traits racinaires chez le Mil



par Princia Aurore Biawallys NAKOMBO-GBASSAULT

Année de soutenance : 2020

**Organisme d'accueil : Institut de Recherche pour le Développement
UMR DIADE Diversité Adaptation et DEveloppement des plantes
Equipes CERES & DYNADIV**

Mémoire de fin d'études

Présenté pour l'obtention du Master 2

**Mention : Sciences et technologies de l'agriculture, de l'alimentation et
de l'environnement**

Parcours : Sélection et évolution des plantes méditerranéennes et tropicales

Caractérisation de QTL contrôlant des traits racinaires chez le Mil



par Princia Aurore Biawallys NAKOMBO-GBASSAULT

Année de soutenance : 2020

**Mémoire préparé sous la direction
de : Anna MEDICI**

Présenté le : 17/09/2020

Devant le jury :

Dominique THIS

Pierre BERTHOMIEU

David POT

**Organisme d'accueil : Institut de
Recherche pour le DEveloppement
(IRD)**

Maîtres de stage :

Laurent LAPLAZE

Yves VIGOUROUX

Résumé

TITRE : Caractérisation de QTL contrôlant des traits racinaires chez le mil

Le mil, *Pennisetum glaucum*, constitue l'aliment de base de plus de 90 millions de personnes en Afrique sub-saharienne et en Inde. La production mondiale de mil est limitée par la sécheresse et une faible fertilité des sols qui caractérisent sa zone de culture. L'amélioration de son système racinaire pourrait contribuer à optimiser l'acquisition de l'eau et des éléments minéraux. La connaissance de la base génétique qui contrôle l'établissement de ce système racinaire peut permettre l'utilisation de la sélection assistée par marqueurs, un puissant outil qui fixe la sélection sur les traits sélectionnés. L'étude de l'architecture racinaire et de la dynamique de croissance racinaire du mil a permis de mettre en évidence certains traits racinaires d'intérêt : la croissance précoce de la racine séminale et l'agrégation du sol rhizosphérique. Des études préliminaires ont abouti à la détection de plusieurs QTL contrôlant ces traits par génétique d'association. L'objectif de ce stage était de valider ces QTL d'intérêt en utilisant des populations obtenues à partir de deux parents présentant les caractères contrastés. Nous avons tout d'abord comparé l'efficacité de quatre méthodes de détection de QTL à partir d'une analyse de ségrégation en masse. Une étude de simulation a été réalisée pour évaluer l'effet de la variation de l'effet du QTL et du nombre d'individus sur la précision de détection de QTL pour ces différentes méthodes. Par ailleurs, nous avons initié la caractérisation d'une population de lignées recombinantes fixées (BC2F4) en évaluant l'homogénéité d'un échantillon de lignées par génotypage et phénotypage.

Ces travaux contribueront à la détection de QTL d'intérêt.

Mots clés : *Pennisetum glaucum*, Lignées BC2F4, Racines, Phénotypage, analyse de ségrégation en masse.

Abstract

TITLE: Characterization of QTLs controlling root traits in Pearl millet

Pearl millet, *Pennisetum glaucum*, is the staple food of more than 90 million people in sub-Saharan Africa and India, and global millet production is limited by the drought and low soil fertility that characterize its growing area. Improving its root system could help optimize the acquisition of water and mineral elements. Knowledge of the genetic basis that controls the establishment of this root system may allow the use of marker-assisted selection, a powerful tool that fixes the selection on the selected traits. The study of the root architecture and root growth dynamics of millet has made it possible to highlight certain root traits of interest: early growth of the seminal root and rhizospheric soil aggregation. Preliminary studies led to the detection of several QTLs controlling these traits by association genetics. The objective of this training course was to validate these QTLs using two types of population coming from crosses between two contrasted parents. We first tested 4 methods for QTL detection using bulk segregant analysis. Simulations were conducted to evaluate the impact of the QTL effect and number of individuals on QTL detection accuracy on these methods. Beside, we initiated the characterization of a population of backcross inbred lines (BC2F4). We evaluated the homogeneity of a sample of BC2F4 lines using genotypic and phenotypic methods. We observed inter-linear heterogeneity and intra-linear homogeneity. This work will help detect QTLs of interest.

Keywords : *Pennisetum glaucum*, *BIL lines*, *Roots*, *Phenotyping*, *Bulk segregant analysis*.

Remerciements

Mes sincères remerciements vont à mes maitres de stage Laurent LAPLAZE et Yves VIGOUROUX pour leur encadrement durant ce stage. Merci d'avoir cru en moi et conforté ma volonté de continuer en thèse.

A special attention for Carla without whom this internship would have been nonsense. A big thank you to you for your help at each step of this internship both in the setting up of the experiment and in the scientific approach to adopt.

Merci à chaque membre des équipes CERES et DYNADIV qui de près ou de loin m'ont apporté leur savoir-faire pour la réalisation de mes expérimentations.

Je remercie également ma tutrice SupAgro, Anna MEDICI pour ses suggestions dans la rédaction du mémoire.

Merci à la promotion APIMET-SEPMET 2019-2020 ainsi que toute l'équipe pédagogique pour le soutien apporté au cours de cette année.

Tout ceci n'aurait été possible sans l'appui du gouvernement Français qui a financé ce Master aboutissant à la rédaction de ce mémoire.

Merci !

Table des matières

Résumé	3
Abstract.....	4
Remerciements	5
Table des matières	6
Liste des figures.....	8
Listes des tableaux.....	9
Avant-propos	10
Glossaire	11
Sigles et acronymes	12
Introduction	13
1. Synthèse bibliographique	15
1.1. Le système racinaire des céréales	15
1.1.1. Architecture du système racinaire.....	15
1.1.2. Interaction racines - sol - microbiome : la rhizosphère.....	17
1.1.3. Le système racinaire des céréales : une cible pour l'amélioration variétale ?	18
1.2. Le mil.....	20
1.2.1. Biologie de la plante	20
1.2.2. Origine et domestication	22
1.2.3. Importance économique et sociale.....	23
1.2.4. Facteurs limitants la production du mil	24
1.3. Problématique du stage.....	25
2. Matériel et méthodes	31
2.1. Matériel végétal	31
2.2. Génotypage par séquençage	31
2.3. Modélisation de QTL : Bulk Segregant Analysis (BSA)	32
2.3.1. Simulation de la population F2	33
2.3.2. Analyses statistiques	33
2.4. Evaluation de l'homogénéité des lignées BC2F4	34
2.4.1. Diversité phénotypique de la croissance racinaire précoce des lignées BC2F4	34
2.4.2. Diversité génétique des lignées BC2F4	35
2.5. Analyse statistique des données	36
3. Résultat.....	36
3.1. Modélisation de QTLs : Analyse de ségrégation en masse (Bulk Segregant Analysis ou BSA)	36
3.1.1. Impact de la variation de qtleffect sur la précision d'estimation de la position de QTL	37
3.1.2. Impact de la variation du nombre d'individus d'une population en fonction de l'effet QTL	38
3.2. Analyses de lignées recombinantes fixées.....	40
3.2.1. Analyse de l'homogénéité génétique des lignées	41
3.2.2. Analyse de l'homogénéité phénotypique des lignées	42
4. Discussion	50
Conclusion.....	52

Travail de l'étudiant.....	53
Références bibliographiques.....	54
Annexes	57
Résumé	97

Liste des figures

Figure 1: Système racinaire des céréales (Maïs), Smith et De Smet, 2012.....	16
Figure 2: Phénotypes de maïs simulés ;	19
Figure 3: Cycle végétatif du mil (Maiti et Bidinger, 1981).....	21
Figure 4 : Dynamique du développement précoce des racines de mil au cours du temps, Passot 2016.....	27
Figure 5: Agrégation du sol autour des racines, Pang et al, Plant & Soil 2017.....	29
Figure 6 : Schéma de sélection des lignées BC2F4.....	31
Figure 7 : Impact de la variation de QTL effect de 0 à 2 avec un pas de 0, 5 sur la précision de détection de la position de QTL simulé.....	37
Figure 8 : Dispositif expérimental.....	42
Figure 9 : Distribution des phénotypes parentaux.....	43
Figure 10 : Distribution des lignées BC2F4 et leur parent en fonction de la longueur racinaire	44
Figure 11 : Analyse des résidus.....	46
Figure 12 : Analyse des résidus des données transformées.....	47
Figure 13 : Comparaison des moyennes des lignées dans la répétition B.....	48
Figure 14 : Comparaison des moyennes des lignées dans la répétition C.....	48
Figure 15 : Comparaison globale des moyennes de longueur racinaire des lignées	49

Listes des tableaux

Tableau 1 : Corrélation entre les positions de QTL simulé et estimé par les différentes méthodes pour différentes valeurs de qtl effect.....	38
Tableau 2 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 100 individus	39
Tableau 3 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 500 individus	39
Tableau 4 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 1000 individus	40
Tableau 5 : Longueur racinaire moyenne et écart à la moyenne des différentes lignées	44
Tableau 6 : Résultat de l'ANOVA sur les lignées	45
Tableau 7 : Résultat de l'ANOVA des données transformées	47
Tableau 8 : Variances.....	50

Avant-propos

Durant six mois, le stage a été réalisé au sein de l'Institut de Recherche pour le développement (IRD) associant les équipes CERES, qui travaille sur l'étude des déterminants génétiques régulant le développement des racines, et l'équipe DYNADIV qui travaille sur la dynamique de la biodiversité des plantes tropicales.

Le stage, intitulé « Caractérisation de QTL contrôlant des traits racinaires chez le mil », a été réalisé dans un contexte très particulier en raison de la pandémie liée à la COVID 19. Le programme initial du stage n'a donc pas pu être suivi normalement à cause du confinement instauré le 16 Mars 2020 soit une semaine après le début du stage. Le stage étant essentiellement expérimental une grande partie des activités ont été modifiées afin de réaliser des expériences dans la mesure du possible et en ajoutant une large part d'analyses théoriques qui n'étaient pas initialement prévues.

Ainsi le mémoire ici rédigé ne comporte que les étapes préliminaires afin de répondre à la problématique initiale du stage. Il s'agit des activités expérimentales qui ont pu être réalisées dans les trois mois qui ont suivi le déconfinement.

Glossaire

Analyse de ségrégation en masse (BSA) : technique de cartographie génétique qui permet d'identifier les régions génomiques associées à un trait d'intérêt

Epistasie : Interaction existant entre deux ou plusieurs gènes

Génétique d'association : Méthode d'association statistique des variations du phénotype aux variations du génotype

QTL effect : l'effet de dominance du locus QTL sur le phénotype

Sigles et acronymes

ANOVA : Analysis of variance (Analyse de variance)

BSA : Bulk segregant Analysis (Analyse de ségrégation en masse)

FAO: Food and Agriculture Organization of the United Nations (l'Organisation des Nations unies pour l'alimentation et l'agriculture)

GBS: Genotyping by sequencing (Génotypage par séquençage)

GWAS : Genome wide association study (étude de génétique d'association)

ICRISAT: Institut internationale de recherche sur les cultures des zones tropicales semi-arides

PCR : Polymerase Chain Reaction (réaction de polymérisation en chaîne par polymérase)

QTL : Quantitative trait loci (locus de caractère quantitatif)

SNP : Single nucleotid polymorphism (polymorphisme nucléotidique)

RIL: recombinant inbred line (lignée recombinante fixée)

Introduction

A l'échelle mondiale, les céréales constituent les principales sources d'alimentation humaine et animale. Le mil perlé (*Pennisetum glaucum* (L.) R. Br., Syn. *Cenchrus americanus* (L.) Morrone) est la sixième céréale en termes de production mondiale. En 2018, selon l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO), environ 28 millions de tonnes de millet dont 15 millions de *P. glaucum* (L.) R. Br. ont été produites dans le monde (FAOSTAT, 2018). Originaire de l'Ouest du Sahel, le mil constitue l'aliment de base de plus de 90 millions de personnes en Afrique sub-saharienne et en Inde (Burgarella *et al.*, 2018). Il est reconnu pour ses qualités nutritionnelles pour l'alimentation humaine (Shweta, 2015). Cette céréale joue ainsi un rôle majeur dans la sécurité alimentaire des régions caractérisées par un climat chaud et sec où il est principalement cultivé (Debieu *et al.*, 2017). Dans les milieux arides et semi-arides, qui occupent 30 % de la surface terrestre de la planète, la faible fertilité des sols et la sécheresse sont les facteurs les plus importants qui limitent la production du mil. Les plantes ont développé différentes stratégies d'adaptation en ajustant leurs systèmes métaboliques et physiologiques. La résistance à ces stress dépend du génotype et peut mobiliser des mécanismes développementaux, physiologiques et/ou biochimiques.

En Afrique subsaharienne, notamment dans l'Ouest du Sahel, les modèles climatiques prédisent un retard de la mousson et une diminution des précipitations au début de la mousson accompagné d'un réchauffement moyen de +2,8 ° C à l'horizon 2031-2060 par rapport à la période de référence 1961-1990 (Sultan *et al.*, 2014). Les épisodes de sécheresse devraient donc être plus fréquents pendant la saison des pluies ce qui entraînera une réduction du rendement du mil dans la région. Par conséquent, le changement climatique pourrait intensifier les conditions défavorables à la croissance des cultures, telles que la réduction des ressources en eau, l'augmentation de la salinité et l'épuisement des éléments nutritifs des sols (Sultan *et al.*, 2014). Ainsi, il risque de nuire à la sécurité alimentaire dans ces régions où une grande partie de la population est déjà confrontée à des difficultés d'accès à la nourriture (Sultan *et al.*, 2013). Afin de soutenir la demande alimentaire d'une population en forte croissance, la sélection variétale devrait donc s'intéresser à l'amélioration de l'efficacité des plantes à absorber l'eau et les nutriments du sol dans des conditions environnementales contraignantes. Le système racinaire étant l'organe qui remplit cette fonction, l'amélioration des rendements des futures variétés pourrait être basée sur une sélection de systèmes racinaires ayant des caractéristiques adaptées aux futures conditions agro-climatiques.

La partie souterraine des plantes a longtemps été ignorée en amélioration des céréales, mais elle bénéficie dorénavant de plus d'intérêt car elle représente une opportunité pour améliorer l'efficacité d'absorption des plantes et la tolérance aux stress abiotiques (Smith et De Smet, 2012). Outre sa fonction d'ancrage, le système racinaire détermine la capacité d'une plante à absorber les nutriments et l'eau du sol et d'autres fonctions importantes telles que la production d'hormones, le stockage d'assimilats ou les associations symbiotiques dans la rhizosphère. Une idée novatrice de sélection serait de cibler les traits racinaires qui favoriseraient la résilience des cultures. Cette possibilité est

encore plus intéressante avec des sols pauvres et sous un climat sec, étant donné que l'amélioration du rendement ne peut être obtenue ni par l'irrigation ni par une augmentation de la fertilisation des sols dans les zones de production du mil (Passot, 2016). L'avancée notable dans les techniques de phénotypage racinaire a ouvert la voie à l'identification des caractéristiques racinaires susceptibles d'être responsables de la tolérance des plantes à certaines contraintes environnementales (Fahlgren *et al.*, 2015 ; Koebernick *et al.*, 2017 ; Lynch, 2019). De récentes études ont permis de déterminer chez les céréales telles que le riz, le maïs ou le blé des régions chromosomiques impliquées dans le déterminisme génétique de certains traits racinaires d'intérêt (Bai *et al.*, 2013 ; Burton *et al.*, 2014). L'identification des mécanismes génétiques sous-jacents de ces caractères seraient une avancée notable pour la recherche agricole qui souhaite assurer une production alimentaire équilibrée dans un contexte de changements mondiaux qui font peser de nouvelles menaces sur l'agriculture.

1. Synthèse bibliographique

1.1. Le système racinaire des céréales

1.1.1. Architecture du système racinaire

Le système racinaire correspond à l'ensemble du réseau formé par les racines d'une plante. La distribution de ces racines dans le temps et l'espace à travers le sol est régie par trois processus développementaux : l'allongement des racines, leur taux et angles de ramification. Ces processus caractérisent l'architecture racinaire (Rich et Watt, 2013). Celle-ci varie selon les espèces, mais aussi au sein des espèces, en fonction du génotype et de l'environnement. Ces processus développementaux impactent la capacité d'explorer et exploiter différents horizons du sol. Le système racinaire des céréales est un système fasciculé (**Figure 1**) et s'oppose au système pivotant des dicotylédones par le fait que le pivot (racine primaire) cesse de croître pour laisser place à des racines adventives d'égale importance disposées à la base de la tige et qui peuvent devenir plus importantes que la racine primaire. C'est un système plus horizontal et superficiel que celui des dicotylédones. Il est composé de différents types de racines: la ou les racines primaires et des racines latérales, qui sont communes aux plantes monocotylédones et les plantes dicotylédones, et les racines coronaires, qui sont caractéristiques des monocotylédones (Passot *et al.*, 2018). La première racine à émerger de la graine est la racine primaire dont le nombre peut varier en fonction des céréales : un chez le maïs à six ou sept chez le blé (Passot, 2016). Aux stades ultérieurs du développement, les racines coronaires émergent de la base de la tige. En fonction des génotypes, elles ont un diamètre et un taux d'allongement relativement grand, un type de croissance indéterminé et un gravitropisme positif. Les racines latérales apparaissent ensuite sur les racines primaires et/ou coronaires par un processus de ramification acropète (Passot *et al.*, 2016). Les racines latérales peuvent elles-mêmes se ramifier, ces ramifications sont appelées racines latérales secondaires, tertiaires, etc... (Smith et De Smet, 2012). On distingue différents types de racines latérales ; chez le mil par exemple, on en a identifié trois types (Passot *et al.*, 2016 ; Passot *et al.*, 2018). Elles diffèrent en fonction de leur longueur, de leur diamètre, de leur anatomie, de leur cinétique de croissance mais aussi de l'angle de ramification et du tropisme ce qui suggère des rôles différents lors de l'exploration du sol pour l'acquisition de l'eau et des nutriments (Passot *et al.*, 2016 ; Passot *et al.*, 2018). Les avantages fonctionnels de la diversité des racines latérales pourraient être une meilleure efficacité et une plasticité dans un contexte pédologique variable et imprévisible dans le temps et dans l'espace (Muller *et al.*, 2019).

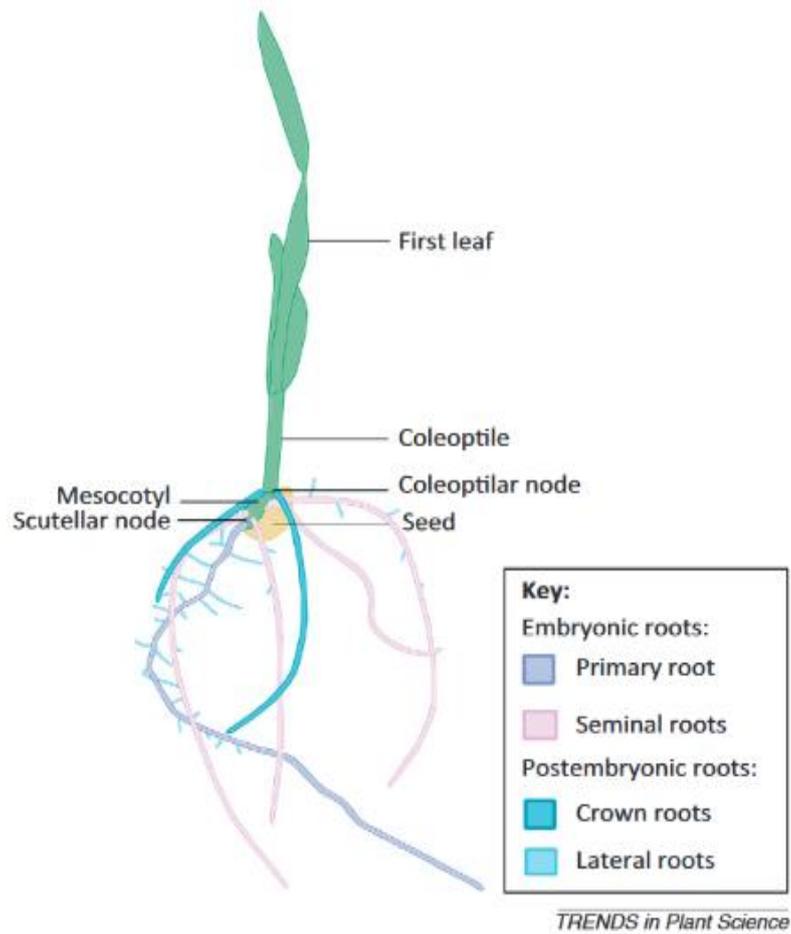


Figure 1: *Système racinaire des céréales (Maïs), Smith et De Smet, 2012*

Primary root: Racines primaires;

Seminal roots: Racines séminales ;

Crown roots: Racines coronaires ;

Lateral roots: Racines latérales.

1.1.2. Interaction racines - sol - microbiome : la rhizosphère

La rhizosphère, du grec “rhiza” qui signifie “racine” et “sphère” champ d'action ou d'influence, a été introduit en 1904 par Lorenz Hiltner pour définir la région du sol sous l'influence des racines des plantes (Hartmann *et al.*, 2008 ; Pang *et al.*, 2017 ; De la Fuente Cantó *et al.*, 2020). Cette région où s'effectuent les échanges entre le sol et la plante est caractérisée par des propriétés physico-chimiques particulières et une intense activité microbienne (Valencia, 2008).

Le sol est constitué de macroagrégats (diamètre > 250µm) et microagrégats (diamètre < 250µm) dont la stabilité est fortement corrélée à la nature et à la teneur en matière organique du sol (MOS) qui assure la cohésion des particules minérales (Ndour *et al.*, 2017). La structuration et la stabilité de ces agrégats déterminent les propriétés mécaniques et physiques du sol (rétention/mobilité de l'eau et des éléments minéraux, aération, régulation de la température, etc...) qui constituent des facteurs déterminants pour la croissance racinaire et la nutrition de la plante (Dickson *et al.*, 1991). La croissance des racines modifie la structure de la rhizosphère en augmentant la taille ainsi que la densité des pores du sol à proximité de la surface des racines (Koebernick *et al.*, 2017). Les racines ont la capacité de synthétiser, d'accumuler et de sécréter un éventail diversifié de composés biochimiques qui constituent l'ensemble des exsudats racinaires et peut représenter jusqu'à 17% des photo-assimilats végétaux (Ndour *et al.*, 2017 ; De la Fuente Cantó *et al.*, 2020). Les exsudats racinaires contribuent à l'allocation du carbone dans le sol et à l'approvisionnement en carbone et en énergie microbienne (Ndour *et al.*, 2017), ce qui a ensuite un impact sur l'agrégation du sol autour des racines. D'autres molécules complexes telles que le mucilage (sucres polymérisés) sont activement excrétées par les cellules de la coiffe des racines, formant une couche gélatineuse autour de l'extrémité de la racine. La quantité et la composition de ces substances varient considérablement en fonction de la diversité des communautés végétales, des espèces de plantes, des génotypes, de l'âge des plantes et des conditions de croissance (De la Fuente Cantó *et al.*, 2020). Les exsudats racinaires jouent donc un rôle majeur dans le maintien du contact racine-sol, la lubrification de la pointe des racines, la protection des racines contre la dessiccation, la stabilisation des micro-agrégats du sol, et l'adsorption sélective et le stockage des ions (Walker *et al.*, 2003). La rhizosphère forme ainsi une couche de sol fortement lié et compact qui adhère fermement la surface des racines.

L'agrégation du sol autour des racines dépend fortement de la lignée végétale, ce qui suggère un contrôle par le génome de la plante, et est liée à la teneur en carbone du sol (Ndour *et al.*, 2017). Par ailleurs, le génotype de la plante influence de façon remarquable la diversité et l'activité du microbiote du sol. La colonisation microbienne le long des racines en croissance est façonnée suivant le schéma d'exsudation des racines. Celui-ci n'est pas homogène le long de l'axe racinaire, la densité bactérienne diminue donc progressivement de la zone d'élongation vers la zone de la racine mature (De la Fuente Cantó *et al.*, 2020). Les microorganismes rhizosphériques : bactéries et champignons saprophytes, pathogènes et symbiotes, participent aussi à la structure du sol (Pang *et al.*, 2017). La

principale contribution des microorganismes rhizosphériques à l'agrégation est associée à la production de sécrétions extracellulaires (polysaccharides, protéines, acides nucléiques et lipides). Ils produisent des exopolysaccharides (EPS) qui constituent l'interface entre les cellules et leur environnement immédiat (Valencia, 2008 ; Ndour *et al.*, 2017). Les EPS ont une forte affinité pour les surfaces minérales et contribuent ainsi à la stabilité des micro agrégats ; ils jouent également un rôle positif dans les phénomènes de régulation de la teneur en eau, notamment dans des conditions de stress hydrique (Valencia, 2008).

On remarque donc que les racines et les microorganismes rhizosphériques vivent en étroite interaction et que les résultats de leur interaction modifient leur environnement le rendant propice à leur survie (Martinez, 2011). Cela illustre bien le concept de phénotype étendu énoncé par Richard Dawkins en 1982 qui se définit comme étant l'action des gènes d'un organisme au-delà même de celui-ci, c'est-à-dire sur son environnement (Dawkins, 1982). Ainsi, au cours de la croissance racinaire, les propriétés physico-chimiques de la rhizosphère sont modifiées par l'interaction entre les exsudats racinaires, l'activité microbienne et les variations du potentiel hydrique du sol.

1.1.3. Le système racinaire des céréales : une cible pour l'amélioration variétale ?

Face aux enjeux climatiques, l'agriculture moderne a pour objectif l'amélioration de la sécurité alimentaire tout en limitant les impacts environnementaux. La recherche scientifique s'investit de plus en plus dans la sélection de cultures présentant une efficacité d'absorption et d'utilisation de nutriment dans un environnement contraignant. De ce point de vue, les traits racinaires apparaissent comme des cibles intéressantes. Des phénotypes racinaires cibles ont ainsi été identifiés pour leur impact sur l'efficacité d'absorption de l'azote (N) et du phosphore (P) ; du potassium (K), du calcium (Ca) et du magnésium (Mg) dans les sols acides ; et l'efficacité d'absorption du fer (Fe) et du zinc (Zn) dans les sols alcalins (Lynch, 2019 ; Tracy *et al.*, 2020). La disponibilité contrastée de ces nutriments dans le temps et l'espace, et leur dépendance aux processus physico-chimiques et microbiologiques du sol, impliquent des compromis pour les stratégies d'amélioration des traits racinaires. Ainsi, différents idéotypes (**Figure 2**) peuvent être proposés (Lynch, 2019). Le phénotype « Topsoil foraging » qui favorise l'exploration de la couche arable est utile pour l'absorption du P et devrait également l'être pour celle du K, du Ca et du Mg dans les sols acides. L'idéotype « Steep, cheap, and deep » présente quant à lui un système racinaire profond qui favorise l'exploration de sous-sol ; il est efficace pour l'absorption de l'azote et de l'eau. La fixation du Fe et du Zn peut être améliorée en ciblant les mécanismes de mobilisation des métaux dans la rhizosphère. Les programmes de sélection devraient également s'intéresser davantage aux poils absorbants car ils réduisent le coût métabolique de l'exploration du sol.

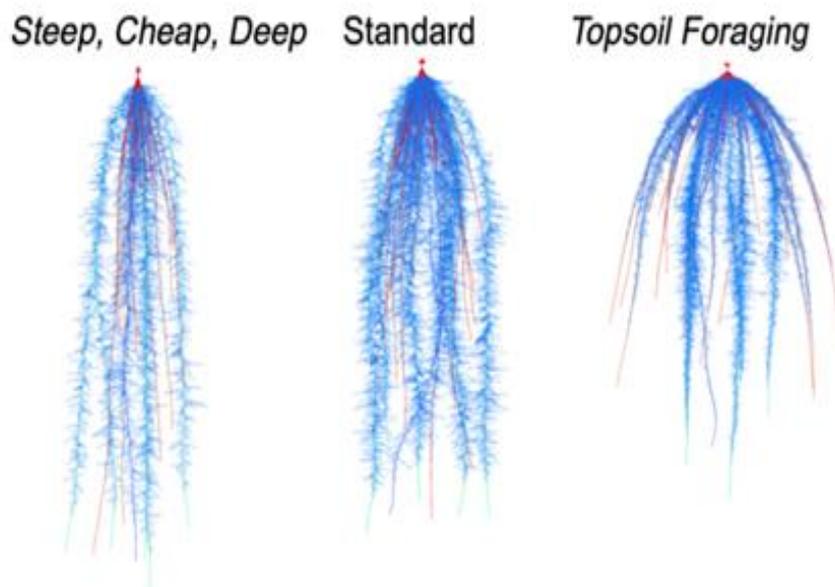


Figure 2: *Phénotypes de maïs simulés ;
Au centre le phénotype normal et à gauche et droite les différents idéotypes racinaires (Lynch, 2019)*

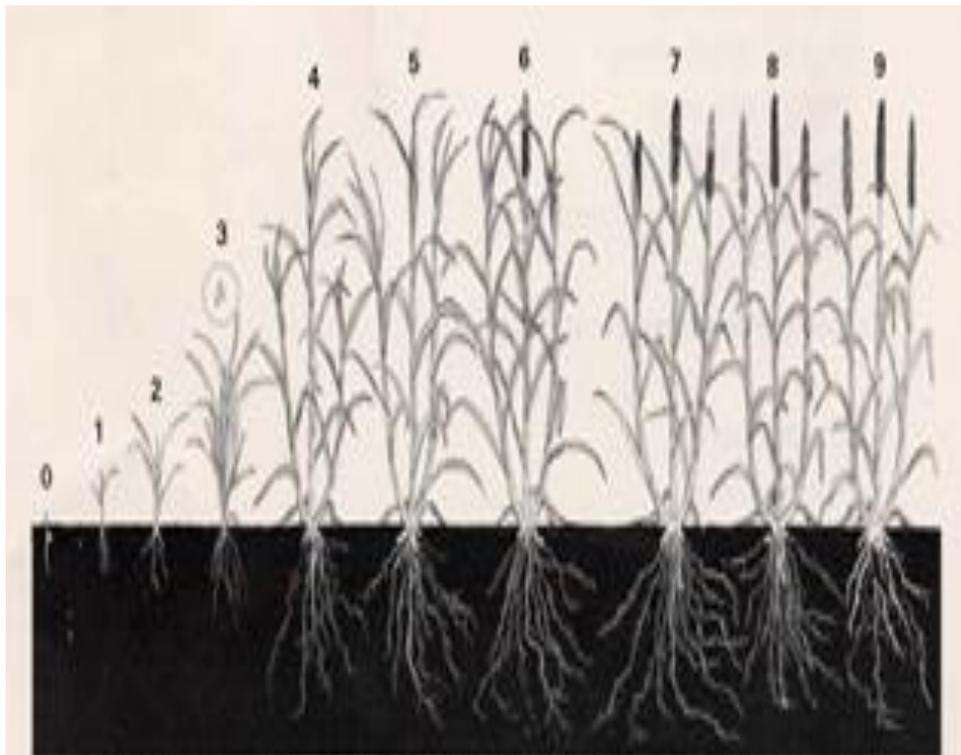
1.2. Le mil

1.2.1. Biologie de la plante

Le mil, *Pennisetum glaucum* [L.] R. Br syn. *Cenchrus americanus* (L.) Morrone est une espèce diploïde annuelle ($2n = 2x = 14$) de monocotylédone appartenant au clade PACMAD c'est-à-dire Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae et Danthoioideae. La taille de son génome est estimée à 1,76 Gb. Il a longtemps été classé au sein de différentes espèces en raison d'une grande diversité morphologique mais il fait désormais partie de la section *Penicillaria* du genre *Pennisetum*, de la famille des *Poaceae*, sous-famille *Panicoideae*, tribu des *Paniceae*. C'est une graminée avec un métabolisme en C4 comme le sorgho ou le maïs.

Le système végétatif aérien du mil est composé de talles qui se forment à partir du plateau de tallage située au niveau du sol (Maiti et Bidinger, 1981). Chaque talle, après un développement complet, est formée d'une tige, de feuilles et d'une inflorescence appelée panicule ou chandelle. En fonction des variétés et zones écologiques, la plante mature présente un port érigé de 0,5 à 3 mètres et des tiges épaisses de 10 à 40 mm de diamètre sur lesquelles s'alternent sur deux lignes opposées des feuilles de 20 à 100 cm de long et 5 à 50 mm de large (Passot, 2016). Les graines de mil sont maintenues sur la chandelle qui mesure 10 à plus de 150 cm de long (Passot, 2016). Le rachis cylindrique de la chandelle et la présence d'une touffe de poils sur le haut des étamines est une caractéristique de l'espèce au sein du genre. En fonction des variétés, il y a généralement entre 500 et 3000 épillets sur une panicule. Les graines de mil sont de diverses couleurs et pèsent environ 5 à 20 mg avec des tailles relativement petites (environ 2 à 5 mm) et des formes allant de globulaire à lancéolée (Passot, 2016). Le système racinaire quant à lui est de type fasciculé et présente des racines fibreuses avec une racine principale fine et petite qui est ensuite rapidement remplacée par des racines secondaires qui se répandent très largement dans le sol (Orman-Ligeza *et al.*, 2013). Le développement racinaire du mil est important et peut atteindre 300 cm de profondeur à la récolte.

Le mil est une plante sexuée, à fleurs hermaphrodites, allogame préférentielle (80 à 90%) grâce à l'anémogamie et une protogynie prononcée (Maiti et Bidinger, 1981). Son cycle végétatif varie entre 60 jours pour les variétés des zones de marge du désert dans le nord-ouest de l'Inde (précoces) et 180 jours dans la zone nord de la Guinée en Afrique de l'Ouest (tardives). A l'instar des autres plantes, le développement phénologique du mil est caractérisé par l'ordre et la vitesse d'émission des organes végétatifs et reproducteurs. On distingue ainsi une phase végétative (GS1) qui commence par une germination hypogée et fini par l'initiation des panicules, une phase de développement des panicules (GS2) et enfin une phase de maturation (GS3) qui comprend la floraison à la fin de la période de remplissage des grains (Maiti et Bidinger, 1981). Les durées relatives de ces phases dépendent de la variété, mais représentent généralement environ un tiers du cycle complet (**Figure 3**).



Phase végétative | Phase de développement de la panicule | Phase de maturation

Figure 3: Cycle végétatif du mil (Maiti et Bidinger, 1981)

1.2.2. Origine et domestication

Le mil (*Pennisetum glaucum* (L.) R. Br. syn. *Cenchrus americanus* (L.) Morrone) est la plus ancienne céréale africaine dans les archives archéologiques d'Afrique (Fuller et Macdonald, 2007 ; Manning *et al.*, 2011). L'ensemble des mils cultivés provient d'un ancêtre sauvage commun, *P. glaucum monodii* (L.) R.Br., que l'on retrouve de nos jours dans le Sahel central (Brunken, 1977). Cependant l'aire de distribution actuelle des mils sauvages est différente de celle des populations d'origine. En effet, un modèle spatial basé sur l'analyse des séquences du génome entier d'accessions de formes sauvages et de variétés traditionnelles représentatives de la diversité géographique du mil a permis d'identifier une origine à une latitude supérieure à l'aire de répartition actuelle des populations sauvages (Burgarella *et al.*, 2018). Il s'agit du bassin de Taoudeni au Sahara occidental (-6,61° E, 23,58° N), correspondant à un climat plus humide qui a prévalu au Sahara il y a 6 000 ans (Burgarella *et al.*, 2018). Autrefois, le Sahara était caractérisé par un système étendu de lacs; son assèchement a conduit il y a environ 3 200 ans les communautés végétales à se déplacer vers le sud pour atteindre leur répartition actuelle (Tostain, 1998 ; Burgarella *et al.*, 2018). La diffusion de l'agriculture du mil a débuté il y a environ 4 892 ans (95% IC 3 685-5 889) au Sahara occidental (Burgarella *et al.*, 2018). Au cours de celle-ci, le mil cultivé s'est hybridé avec des mils sauvages dans l'ouest et l'est du Sahel, ce qui a conduit à des points de diversité en dehors du centre d'origine (Oumar *et al.*, 2008). Le mil a donc été domestiqué en Afrique de l'Ouest avant que sa culture se diffuse vers les autres régions du monde. On peut supposer deux processus de migration : une vers l'Est de l'Afrique puis l'Asie notamment l'Inde et une vers le sud de l'Afrique en rapport avec le déplacement des populations de l'ethnie Bantou.

Le groupe des mils cultivés est très différencié des groupes sauvages et présente très peu de divergence génétique (Oumar *et al.*, 2008). La diversité génétique du mil sauvage est structurée en trois grands groupes en fonction de leur origine géographique. On distingue un groupe occidental qui comprend des individus du Sénégal et de la Mauritanie, un groupe central qui comprend des individus du Mali et du Niger et un groupe oriental comprenant des individus du Tchad et du Soudan (Burgarella *et al.*, 2018). Par ailleurs, on rencontre également dans la zone sahélienne une forme intermédiaire issue de l'hybridation des mils sauvages et cultivés appelé shibra (Brunken, 1977 ; Passot, 2016). Ils présentent des caractères morphologiques intermédiaires mais il est souvent difficile de les distinguer du mil cultivé avant la maturation, car ils imitent les caractères spécifiques de la plante cultivée à côté d'eux. Les shibras poussent dans la majeure partie de la zone de culture du mil en Afrique cependant ils n'ont pas été trouvés en Inde (Passot, 2016).

1.2.3.Importance économique et sociale

Le mil est l'une des céréales les plus cultivées dans les régions arides et semi-arides du monde en raison de sa grande capacité de résilience. Selon l'Institut internationale de recherche sur les cultures des zones tropicales semi-arides (ICRISAT), environ 30 millions d'hectares de la surface mondiale est dédiée à la culture du mil principalement dans des régions où l'agriculture traditionnelle est prédominante. En effet, le mil est cultivé principalement en Asie et Afrique où il constitue l'aliment de base pour plus de 90 millions d'agriculteurs. Ces deux continents ont produit à eux seuls respectivement 48,9% et 48,1% de la production mondiale de mil en 2018 (FAOSTAT, 2018). Le mil est la sixième céréale la plus importante au monde après le blé, le maïs, le riz et le sorgho. Le premier producteur de mil est l'Inde avec plus de 11 millions de tonnes produites en 2018 pour une surface cultivée de 9 millions d'hectares (FAOSTAT, 2018). Selon l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), près de 15 millions d'hectares sont consacrés à la culture du mil en Afrique. Le mil constitue 6% de la production céréalière totale en Afrique et 14% pour l'Afrique de l'Ouest seule (Passot, 2016). L'Afrique de l'Ouest est donc aussi une grande zone de production. Les principaux pays producteurs en Afrique sont le Niger avec 3,8 millions de tonnes produites, le Nigéria et le Mali avec 1,8 millions de tonnes produite chacun en 2018 (FAOSTAT, 2018).

Le mil est généralement cultivé dans le cadre d'un système de culture mixte en alternance avec des légumineuses (arachide, niébé, ...) ou d'autres céréales (maïs, sorgho, blé, ...) pour ses grains qui sont destinés à l'alimentation humaine mais aussi de plus en plus pour ses parties végétatives utilisées comme fertilisant ou aliment pour bétail (Andrews et Kumar, 1992). Au niveau mondial, les rendements en mil sont inférieurs aux autres céréales. En outre, la production mondiale est principalement consommée dans les zones de production notamment en Afrique où la récolte est exclusivement destinée à l'autosubsistance (Andrews et Kumar, 1992). Les échanges internationaux sont donc relativement faibles (0,5 millions de tonnes) avec des prix d'environ 250 \$/tonne ; le premier exportateur de mil est l'Inde (Passot, 2016).

Le mil est reconnu pour ses valeurs médicinales, énergétiques et nutritionnelles supérieures aux autres céréales dans des conditions de cultures identiques (Shweta, 2015). En effet, les grains de mil sont une importante source de micronutriments (fer, zinc, magnésium) et potentiellement l'une des sources alimentaires les moins chères d'oligo-éléments et de protéines. Ils contiennent 67% d'amidon, 11% de protéines et 5% de lipides ; sa teneur en acide aminé est supérieure à celle du maïs ou du sorgho (Andrews et Kumar, 1992 ; Shweta, 2015). Cette céréale sans gluten est très nutritive, c'est pourquoi elle est recommandée aux enfants, aux personnes âgées ou convalescentes. On la consomme généralement sous forme de bouillie, couscous, pain plat ou encore de boisson (Andrews et Kumar, 1992).

1.2.4. Facteurs limitants la production du mil

1.2.4.1. Contraintes biotiques

La faible production du mil est due en partie à des pressions parasitaires dont la plus importante est d'origine fongique (Mbaye, 1993). On peut citer le mildiou causé par *Sclerospora graminicola* et *Plasmopara penniseti* qui attaque la plante dès le stade plantule et réduit le rendement de l'ordre de 20 à 40% ; le charbon causé par *Moesziomyces penicillariae* responsable des pertes de rendement de près de 30%, l'ergot causé par *Claviceps fusiformis* qui réduit les rendements de 58 à 70% ; la pyriculariose causée par *Pyricularia grisea* et la rouille causée par *Puccinia substriata var. penicillariae* (Mbaye, 1993 ; Moumouni, 2014 ; Kadri *et al.*, 2019). Les maladies bactériennes comme la tache bactérienne (*Pseudomonas syringae*) et la strie foliaire bactérienne (*Xanthomonas campestris pv. pennamericanum*) occasionnent peu de dégâts (Passot, 2016). Les maladies virales qui ont quant à elles été signalées sur le mil sont peu importantes, il s'agit de Freckled Yellow Diseases et la mosaïque de la canne à sucre (Mbaye, 1993).

Le mil est attaqué par un nombre réduit d'insectes par rapport aux autres céréales. Ils appartiennent principalement aux ordres des lépidoptères, des diptères et parfois des coléoptères et peuvent être regroupés en cinq classes : les ravageurs de l'épi comme la chenille mineuse (*Heliocheilus albipunctella*) qui attaque près de 70% des épis et peut occasionner des pertes de rendement de 15%, les foreurs de tiges (*Acigona (Coniesta) ignefusalis* et *Sesamia calamistis*), les insectes de stocks (*Sitotroga cerealella* Oliv, *Tribolium castaneum*,...), les ravageurs des plantules et les insectes phyllophages. On dénombre également des ravageurs polyphages comme les acridiens (*Oedaleus senegalensis*, *Schistocerca gregaria*), les oiseaux granivores (*Quelea quelea*, *Q. erythrops*, ...) et les rongeurs nuisibles (*Arvicanthis niloticus*, *Mastomys sp*) (Mbaye, 1993 ; Kadri *et al.*, 2019).

Les adventices peuvent également réduire la productivité du mil. Ce dernier est souvent en concurrence avec des dicotylédones annuelles à larges feuilles (*Cassia obtusifolia*, *Commelina benghalensis*,...), les graminées annuelles (*Cenchrus biflorus*, *Digitaria spp*) et les cyperaceae (*Kyllingia squamulata*, *Cyperus amabilis*,...) qui occasionnent généralement des pertes de rendement d'au moins 30% (Mbaye, 1993). *Striga hermonthica* est un parasite obligatoire qui affecte environ 40% de la zone de production céréalière en Afrique subsaharienne et cause des pertes de rendement allant jusqu'à 100% (Kadri *et al.*, 2019). Son incidence augmente la carence en phosphore du sol et rend le mil plus vulnérable aux attaques parasitaires.

1.2.4.2. Contraintes abiotiques

Le mil est généralement cultivé dans des régions marginales où d'autres céréales peuvent difficilement être cultivées. L'une des principales contraintes de développement du mil est la sécheresse, elle survient généralement au stade précoce ou pendant les phases de remplissage des grains (Passot, 2016). Lorsque le mil est soumis à un stress hydrique en phase juvénile, cela a des

conséquences sur la maturation des graines même si le développement végétatif bénéficie à nouveau d'un apport en eau (Debieu *et al.*, 2018). Par conséquent, le stress hydrique précoce réduit significativement les composantes du rendement (hauteur des plantes, biomasse aérienne, nombre total de grains et leur poids). En Afrique sub-saharienne, le mil est généralement semé après ou juste avant la première pluie de la saison des pluies. Cette dernière est de courte durée et aléatoire selon les années ; les prédictions climatiques indiquent l'apparition d'évènements de plus en plus extrêmes tels que de fortes pluies, des sécheresses ou inondations et une augmentation d'année en année de la variabilité des précipitations ce qui ne feront qu'accentuer par moment les carences en eau dans ces zones (Sultan *et al.*, 2013). En outre, les températures au niveau du sol dans les zones arides peuvent atteindre facilement 45 °C favorisant ainsi une déshydratation immédiate du sol laissant la culture avec très peu de ressource hydrique (Passot, 2016). Par ailleurs, les sols sont également caractérisés par une faible teneur en matière organique et les cultivateurs les fertilisent peu. Le manque de nutriments est donc l'un des facteurs limitants pour la croissance du mil, notamment le phosphore dont une carence crée ou renforce le stress hydrique déjà existant. Après le phosphore, l'azote est le deuxième nutriment limitant la croissance des plantes. Cependant, cela ne compromet pas la mise en place de la culture. Le manque d'éléments nutritifs limite ainsi le développement du système racinaire en empêchant la plante d'utiliser de manière optimale les ressources en eau du sol. A cela s'ajoute d'autres contraintes telles que la structure et la texture (profondeur, faible capacité de rétention d'eau, salinité...) du sol, les températures élevées, les pratiques agronomiques, ...

1.3. Problématique du stage

La production mondiale de mil est limitée par les conditions environnementales qui caractérisent sa zone de culture. Les principaux facteurs limitant cette production sont la sécheresse et une faible fertilité des sols. L'amélioration de son système racinaire pourrait contribuer à optimiser l'acquisition de l'eau et des éléments minéraux grâce à l'exploration du sol qui dépend de l'architecture racinaire et à la mobilisation des nutriments, en particulier les moins mobiles comme le phosphore, via des interactions physicochimiques et biologiques dans la rhizosphère. Cela permettrait d'augmenter le rendement dans des conditions limitantes et sans avoir recours à des intrants plus importants. Une bonne connaissance de la structure et du fonctionnement du système racinaire est donc nécessaire pour identifier les traits pertinents à améliorer. La connaissance de la base génétique qui contrôle l'établissement de ce système racinaire peut permettre l'utilisation de la sélection assistée par marqueurs, un puissant outil qui fixe la sélection sur les traits sélectionnés. L'étude de l'architecture racinaire et de la dynamique de croissance racinaire du mil a permis de mettre en évidence certains traits racinaires d'intérêt (Passot *et al.*, 2016 ; Passot *et al.*, 2018 ; De la Fuente Cantó *et al.*, 2020). Il s'agit entre autres de la croissance précoce de la racine séminale qui est associée à une meilleure tolérance au stress hydrique précoce et l'agrégation du sol rhizosphérique qui serait associée à une

meilleure tolérance au stress hydrique et à la carence en phosphore (Rabbi *et al.*, 2018).

En effet, le développement précoce du système racinaire du mil est caractérisé par une racine primaire à croissance rapide qui colonise des horizons de sol plus profonds, tandis que les racines latérales et coronaires commencent à émerger seulement six jours après la date de semis (Passot, 2016). Cette stratégie de croissance peut être favorable à la colonisation rapide des horizons pédologiques profonds qui retiennent un peu d'eau dans un sol pauvre en matière organique et à capacité de rétention d'eau minimale. Il pourrait donc s'agir d'une stratégie importante d'adaptation pour faire face au stress de la sécheresse précoce sachant que cette dernière présente un grand impact sur les composantes du rendement; elle réduit significativement la hauteur des plantes, la biomasse aérienne, le nombre total de grains et leur poids mais affecte de façon modérée la longueur des panicules, la longueur des tiges et le nombre de feuilles (Debieu *et al.*, 2018). La vitesse de croissance de la racine primaire après la germination a une héritabilité au sens large relativement élevée (Passot, 2016). Une analyse de génétique d'association (GWAS) a permis de mettre en évidence 9 QTL contrôlant ce caractère chez le mil (**Figure 4**).

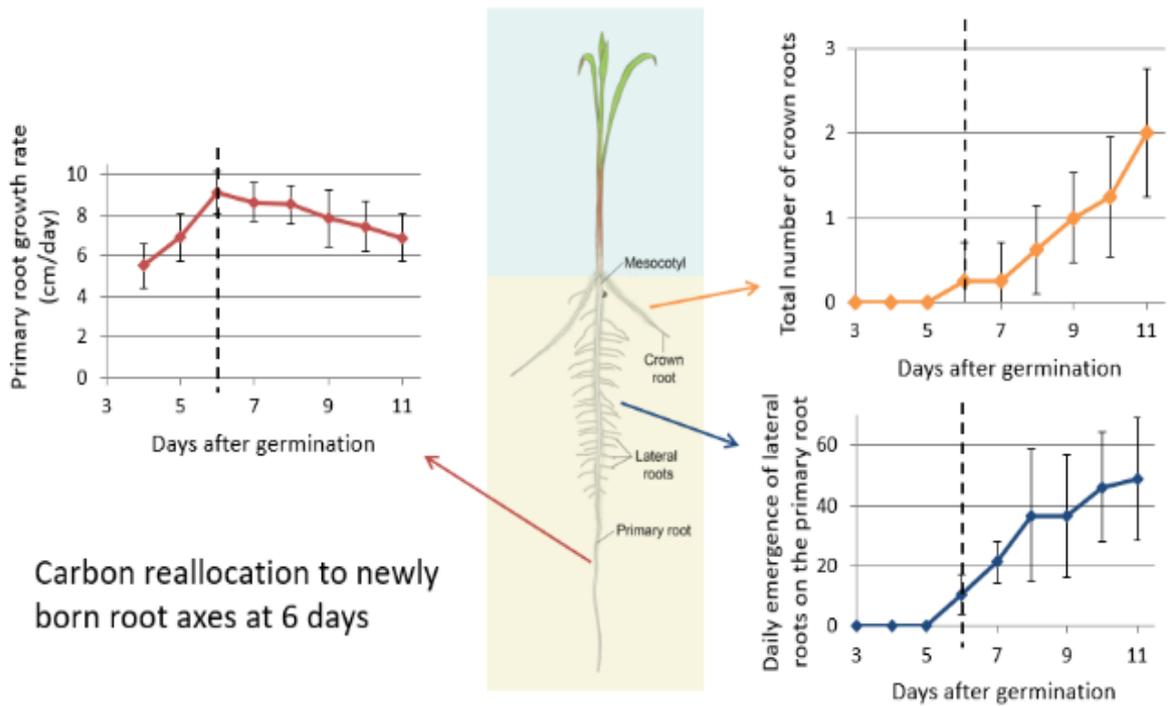
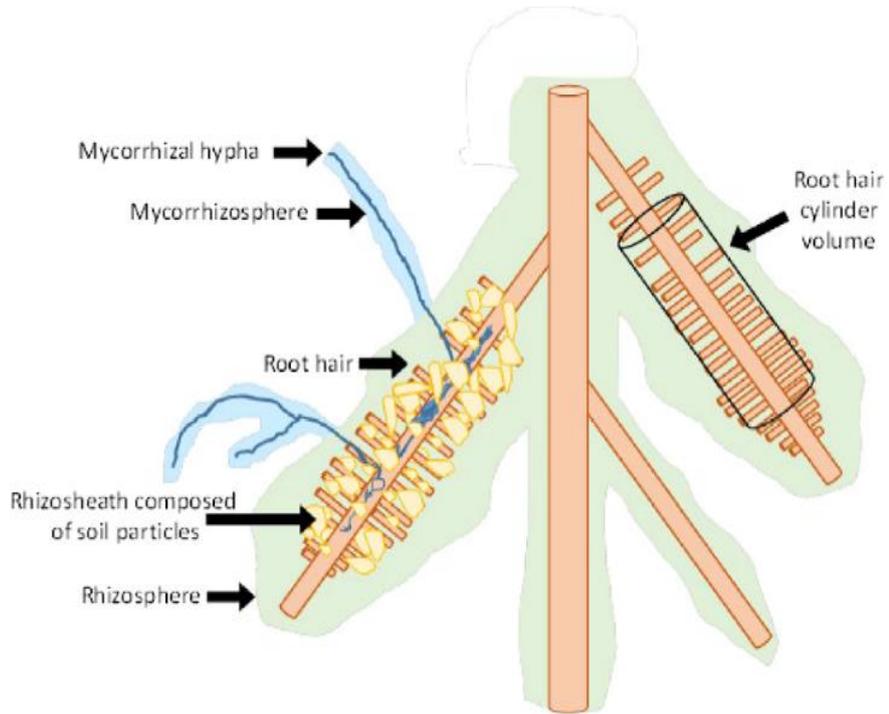


Figure 4 : Dynamique du développement précoce des racines de mil au cours du temps, Passot 2016.

- Vitesse de la croissance des racines primaires 11 jours après la germination
- Nombre de racines coronaires 11 jours après la germination
- Dynamique d'apparition des racines latérales sur les racines primaires 11 jours après la germination

L'agrégation du sol autour des racines est un caractère héritable qui est présent chez toutes les angiospermes et a été associé à une meilleure tolérance à la sécheresse et à la carence en P (Rabbi *et al.*, 2018). Elle favorise l'activité microbienne et l'augmentation des communautés microbiennes et a un impact sur la stabilité mécanique et les processus hydrauliques dans la rhizosphère (Walker *et al.*, 2003 ; Pang *et al.*, 2017 ; Rabbi *et al.*, 2018). La capacité de rétention d'eau et de diffusion des nutriments dans l'interface sol-racine est définie par cette interaction complexe et devient particulièrement importante dans les sols secs où le réseau de mucilages renforce le potentiel de la matrice du sol autour de la racine, ce qui contribue à maintenir la rhizosphère humide et à prévenir les chutes soudaines du flux d'eau, en particulier autour de l'extrémité de la racine (De la Fuente Cantó *et al.*, 2020). En ce qui concerne l'agrégation du sol rhizosphérique, 5 QTLs ont été identifiés à la suite d'une GWAS chez le mil (**Figure 5**).

1



2

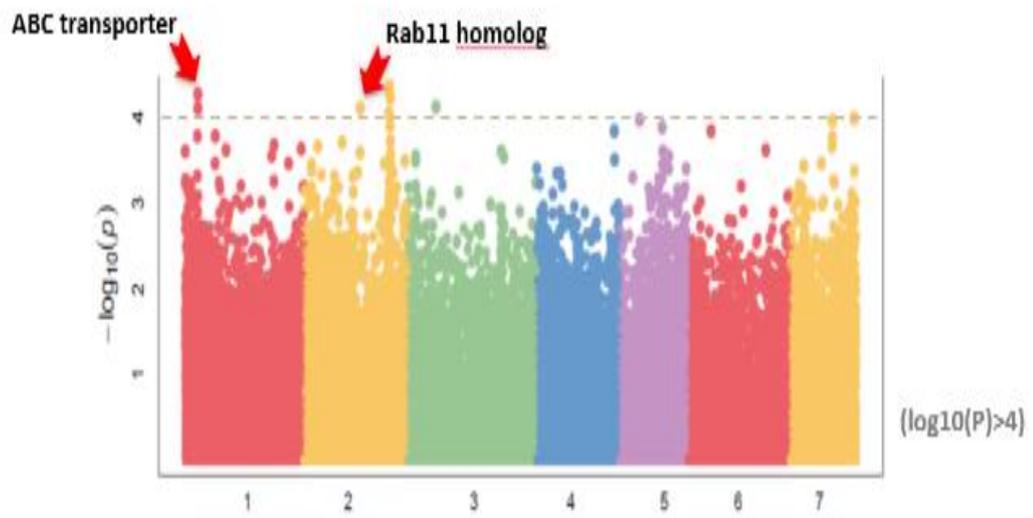


Figure 5: Agrégation du sol autour des racines, Pang et al, Plant & Soil 2017.

1-Association rhizosphérique.

2-Manhattan plot: 5 QTLs détectés par GWAS.

L'objectif initial du stage était de contribuer à la caractérisation de QTL de ces deux traits racinaires d'intérêt pour le mil en utilisant des lignées recombinantes (backcross inbred lines (BILs) correspondant à des BC2F4). Celles-ci sont issues de croisement entre des populations bi-parentales (2 par traits) à partir de lignées contrastées. Il s'agissait plus spécifiquement d'analyser les données de génotypage par séquençage des lignées BILs produites, ensuite de caractériser le phénotype des lignées quasi isogéniques pour valider l'impact des QTL sur les traits d'intérêt et enfin d'identifier de façon plus précise les QTL d'intérêt c'est-à-dire réduire la taille des QTL déjà obtenus, en identifier des nouveaux ou encore identifier des gènes directement impliqués. En raison de la situation sanitaire actuelle, des modifications ont été apportées au programme de recherche du Master initial.

Sa version initiale prévoyait des activités expérimentales importantes :

- Phénotypage de 200 lignées BILs pour la croissance racinaire à l'Université de Nottingham,
- Génotypage des BILs par séquençage (GBS : Genotyping by sequencing),
- Analyse QTL.

La version actuelle a été modifiée pour tenir compte de la réalisation des activités en télétravail et l'impossibilité de réaliser le phénotypage à l'Université de Nottingham :

- Développement de simulation pour l'analyse de l'apport des approches d'analyse de ségrégant en masse (BSA : Bulk Segregant Analysis) dans l'étude du lien génotype/phénotype sur F2 issues du même croisement bi-parental (activité en télétravail),
- Evaluation de l'homogénéité des lignées BILs par génotypage SSR et phénotypage,
- Génotypage des RILs par GBS,
- Analyses des données génotypiques pour trouver des individus recombinants dans des QTL déjà identifiés et (si faisable) caractérisation phénotypique de quelques individus recombinants.

La revue bibliographique a été réalisée sur la version initiale du Master.

2. Matériel et méthodes

2.1. Matériel végétal

Les expériences ont été menées sur des lignées recombinantes BC2F4 de mil *Pennisetum glaucum* (L.) R. Br. Celles-ci sont issues de croisement entre des lignées génétiquement fixées contrastées, deux par phénotype racinaire d'intérêt (**Figure 6**). Les lignées recombinantes ayant un phénotype caractérisé par une croissance précoce de la racine séminale sont issues du croisement ((SL2 x ICML IS 11155) x ICML IS 11155) noté 249 (faible croissance) x 337 (forte croissance). Celles étudiées pour leur capacité à agréger le sol rhizosphérique proviennent du croisement (line220 x line132) présentant une faible (220) et une forte agrégation (132) respectivement. Les graines utilisées proviennent de l'Institut international de recherche sur les cultures des zones tropicales semi-arides (ICRISAT) au Niger où ont été produites ces populations.

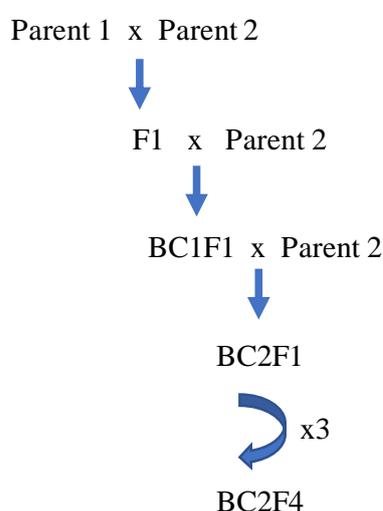


Figure 6 : Schéma de sélection des lignées BC2F4

- Parent 1 : Forte expression du caractère
- Parent 2 : Faible expression du caractère

2.2. Génotypage par séquençage

Les graines des lignées BC2F4 249 x 337 et 220 x 132 ont été mises à germer en serre dans des plaques de germination (108 alvéoles) contenant un substrat de culture (Terreau professionnel horticoles avec engrais, mélange de tourbe de sphaigne blonde, de tourbe de sphaigne noire et d'argile (6%), caractéristiques physico-chimiques : matière sèche % sur brut : 35; matière organique % sur sec : 65; pH : 5,8; conductivité en mS/cm : 100; engrais NFU 42001 : 1KG/m³ (NPK 17+10+14)). Au total, pour le croisement 249 x 337, 544 graines ont été semées (2 graines/lignée avec 272 lignées)

tandis que pour le croisement 220 x 132 seulement 448 graines ont été semées (2 graines/lignée avec 224 lignées). Après deux semaines de croissance, les échantillons de feuilles des deux groupes de lignées recombinantes (249 x 337 et 220 x 132) ont été collectés et stockés à -80°C. Les échantillons sont constitués de coupes de feuilles d'environ 10 à 12 cm de longueur (environ 100 mg de poids frais), 192 lignées par croisement ont été collectées. Ainsi, 8 plaques de 96 puits freemarkers (quatre plaques par croisement) ont été préparées. Les échantillons de feuilles prélevées ont ensuite subi une lyophilisation pendant trois jours à -47°C et 0,056 mbar. L'extraction de l'ADN ainsi que le génotypage sont réalisés par la société Freedom Markers (USA). Les conditions sanitaires ont allongé les délais d'obtention des marqueurs, qui ne sont pas encore disponibles.

2.3. Modélisation de QTL : Bulk Segregant Analysis (BSA)

L'analyse de ségréants en masse (BSA) est une technique de cartographie génétique qui permet d'identifier les régions génomiques associées à un trait d'intérêt. La méthode est basée sur le phénotypage d'une population en ségrégation et l'analyse génétique en mélange (bulk) des individus avec des phénotypes extrêmes (généralement les 10% les plus extrêmes). Nous nommerons dans la suite la méthode NGS-BSA dans le cas d'une approche de génotypage du mélange des phénotypes extrêmes par des méthodes de séquençage haut débit (NGS). Pour cette modélisation, nous avons utilisé le package QTLseqr, un package R pour NGS-BSA, qui permet d'identifier des QTLs. Ces approches calculent les fréquences alléliques dans les deux groupes d'individu de phénotype extrême et testent si ces fréquences sont statistiquement différentes. Nous attendons dans une population biparentale (utilisant une seule F1 autofécondée) en ségrégation une fréquence allélique de 50% pour les deux allèles. Si un variant n'est pas associé au phénotype, nous attendons donc : 1) une fréquence allélique de 50% dans les populations extrêmes et 2) pas de différence de fréquence allélique entre ces deux groupes.

Le package QTLseqr utilise un G' test pour évaluer la différence allélique entre groupe. Ce test peut être lissé sur plusieurs SNP proches, G' test lissée. QTLseqr permet aussi d'importer et filtrer les données génétique (SNP), calculer les fréquences alléliques relatives, les valeurs G' et sa significativité permettant ainsi l'identification de QTL.

Nous avons réalisé une simulation afin de modéliser une analyse BSA-NGS sur des populations F2 obtenues à partir d'un croisement biparental afin d'identifier et évaluer la significativité statistique de la position d'un QTL d'intérêt connu (car modélisé).

L'objet de cette simulation consistait à étudier premièrement l'impact d'une variation de la valeur de l'effet de QTL (qtleffect) et deuxièmement l'impact du nombre d'individus phénotypés dans la population F2 sur la précision d'estimation de la position d'un QTL d'intérêt.

2.3.1. Simulation de la population F2

Les paramètres des simulations sont : le nombre de simulation (Nbsim), le nombre de chromosome (Nbchr : 1000) ainsi que le nombre de SNP par chromosome (loci=10000).

Nous avons simulé un chromosome de 90 cM, qui pour chaque individu pouvait avoir q recombinaisons de manière uniforme. Le nombre q de recombinaison pour un chromosome est simulé suivant le modèle de Haldane et suit une distribution de Poisson avec pour paramètre ($\lambda=0.9$). Si $q=0$ aucune recombinaison n'est présente, si $q=1$ une seule et ainsi de suite. Pour chaque chromosome, on construit ensuite son haplotype associé à l'absence (0) ou présence (1) de l'allèle de référence en fonction du nombre de recombinaison. Un chromosome sans recombinaison sera exclusivement 0 pour ces 10000 SNP (50% de chance) ou 1 (50% de chance). Un chromosome avec une recombinaison alternera des génotypes 0 puis 1 (50% de chance) ou 1 puis 0 (50% de chance). La position de la recombinaison est tiré aléatoirement dans une loi uniforme. Une fois les 1000 chromosomes simulés, on obtient une matrice de génotype avec les chromosomes (en colonne) et loci/SNP (en ligne). Les chromosomes sont groupés par deux pour construire des génotypes diploïdes. La matrice de phénotype est basée sur le génotype d'un SNP particulier pour lequel le génotype sera 0, 1 ou 2 fois l'effet du QTL (QTL effect) en fonction des génotypes 0/0, 0/1, ou 1/1. Pour ajouter du bruit associé au phénotypage, on ajoute un effet phénotypique pour tous les individus suivant une loi normale $N(0; 1)$. La position du SNP causatif (aléatoire ou au centre du chromosome suivant les simulations) est nommée *qtlpos* et stocké dans la matrice *QTLpos* qui contient toutes les positions de QTL (simulé et estimé).

Une fois le phénotype obtenu, nous pouvons classer les individus suivant ce phénotype et créer deux groupes (10% avec le phénotype le plus faible, 10% avec le phénotype le plus grand). Les comptages des génotypes 0 et 1 de ces deux groupes sont formatés pour permettre leur analyse à l'aide du package *QTLseqr*. Ces comptages sont parfaits et ne prennent pas en compte la variabilité associée à un séquençage de type NGS. Du bruit est ajouté au comptage allélique parfait, en simulant une profondeur totale de $n=100$ et la fréquence allélique (p) de chaque SNP à l'aide d'une fonction binomiale $B(n, p)$. Les données ainsi obtenues sont ensuite analysées statistiquement.

2.3.2. Analyses statistiques

Les données simulées sont importées en utilisant la fonction *importFromTable* du package *QTLseqr*. Puis, la fonction *runQTLseq Analysis* est utilisée pour calculer (**Annexe v**) :

- la fréquence allélique de l'allèle alternatif pour chaque groupe ;
- la différence de fréquence allélique groupe haut – groupe bas (*deltaSNP*, Takagi *et al.*, 2013) ;
- la statistique distance euclidienne entre les deux groupes (*PEDm*, Hill *et al.*, 2013; Zhang *et al.*, 2019) .

Ces statistiques sont calculées SNP par SNP (deltaSNP, G, PEDm) ou lissées sur une fenêtre de plusieurs centaines de SNPs (PdSNP, PTdSNP statistique lissée dérivée de la statistique individuelle deltaSNP ; PED100_4 basé sur la distance euclidienne individuelle PEDm).

Pour la statistique DeltaSNP (PdSNP et PTdSNP) ou la statistique PEDm (et PED100_4), un seuil d'intervalle de confiance par 1000 bootstrap sont calculés pour tester la significativité (Takagi *et al.*, 2013).

Les positions de QTL simulées et celles estimées (valeur la plus grande de chaque statistique) sont ensuite représentées sur des graphiques. Des corrélations entre les positions de QTL simulées et QTL estimées ont également été calculées.

2.4. Evaluation de l'homogénéité des lignées BC2F4

2.4.1. Diversité phénotypique de la croissance racinaire précoce des lignées BC2F4

2.4.1.1. Condition de germination

Dix lignées recombinantes du croisement 249 x 337 ont été étudiées. Les graines de ce croisement (10 graines par lignée BC2F4) ont été désinfectées pendant deux minutes à l'éthanol (C₂H₅OH) 70% puis cinq minutes à l'hypochlorite de sodium (NaClO) 50%. Chaque lavage a été suivi d'un rinçage à l'eau distillée : une fois après l'éthanol et deux fois après l'hypochlorite de sodium. Les graines ont ensuite été mises à germer à l'obscurité dans des boîtes de Petri (96x16mm) tapissées de papier Whatman légèrement humide et enveloppées dans du papier aluminium. Les boîtes de Petri ont été placées pendant 24 heures à 27°C dans une chambre de culture. Le pourcentage de germination a été calculé (**Annexe i & Annexe ii**).

Pour chaque génotype, cinq graines ayant une radicule mesurant environ 2 cm ont été placées à 3,5 cm d'intervalles sur des papiers de germination (Anchor Paper Company, USA) qui ont ensuite été enroulés (Hetz *et al.*, 1996). Chaque rouleau de papier a été placé dans des erlenmeyer contenant 250 mL de solution nutritive (¼ strength Hoagland ; pH 6,8 ; Hoagland *et al.*, 1950). Les erlenmeyer ont été placés pendant une semaine dans une chambre de culture dont les conditions de croissance sont : 27°C, 57% d'hygrométrie et une photopériode de 12h (8h à 20h jour et 20h à 8h nuit). Le volume de solution nutritive a été maintenu à 250 mL durant toute la période de croissance.

L'expérimentation a été répétée trois fois (**Annexe iii**).

2.4.1.2. Phénotypage racinaire

Après une semaine, les rouleaux de papier portant les plantules ont été délicatement déroulés. Des photos ont été prises afin de mesurer la longueur des racines de chaque génotype. Le traitement d'image s'est fait à l'aide du logiciel ImageJ. Aussi, la partie aérienne des plantules a été collectée, pesée puis stockée à -80°C.

2.4.2. Diversité génétique des lignées BC2F4

2.4.2.1. Condition de croissance

Les graines des lignées recombinantes 249 x 337 et 220 x 132 ont été mises à germer en serre. Cinq graines de chaque génotype (10 génotypes par croisement) ont été semées dans des plaques de germination (54 alvéoles) contenant un substrat de culture (Terreau professionnel horticole avec engrais). Celles-ci ont poussé durant une semaine dans les conditions de température de 28°C, 70% d'hygrométrie et 12 heures de photopériode. Le phénotype aérien des plantes de chaque génotype a été relevé. Des échantillons des feuilles ont ensuite été collectés et stockés à -27°C.

2.4.2.2. Extraction et quantification de l'ADN

Dans des tubes de 0.2 mL, des pastilles de feuilles ont été écrasées dans le tampon d'extraction contenant 50 mM de Tris (pH 8), 150 mM de NaCl, 2% de PVP 40 et 1% de tween. Les tubes (100 tubes = 10 génotypes/croisement et 5 répétitions) ont été mis dans un thermocycleur (Applied Biosystems Veriti 96 well thermal cycler) pendant 15 minutes à 75°C puis la température a progressivement baissé jusqu'à 10°C. 50 µL du surnageant ont ensuite été récupérés dans un nouveau tube auquel on a ajouté 50 µL de solution de billes magnétiques (Ampure XP). Ces billes sont conçues pour lier l'ADN présent dans la solution. Après 10 minutes, les tubes ont été placés sur un aimant métallique puis le surnageant a été enlevé. Les billes ont ensuite été lavées deux fois avec de l'éthanol 70%. Enfin, le culot de billes est repris dans 15 µL de tampon TE.

Le spectrophotomètre Thermo Scientific™ NanoDrop a été utilisé pour quantifier et évaluer la pureté de l'ADN. 1 µL de chaque tube a été prélevé pour effectuer les mesures, chaque échantillon avait une concentration supérieure ou égale à 10 µg/µL.

2.4.2.3. Génotypage microsatellites

Les génotypages de contrôle des parents et des lignées BC2F4 ont été réalisés à l'aide de 12 marqueurs microsatellites (IRD 46, PsmP 2227, PsmP 2249, PsmP 2247, PsmP 2237, IRD 12, IRD 25, PsmP 2214, PsmP 2219, PsmP 2220, PsmP 2206 et PsmP 2202 ; Allouis *et al.*, 2000; Budak *et al.*, 2003 ; Qi *et al.*, 2004 ; Mariac *et al.*, 2006, **Annexe iv**).

L'amplification par réaction de polymérisation en chaîne par polymérase (PCR) a été effectuée à l'aide du kit Multiplex PCR (Qiagen ®) dans un thermocycleur selon le protocole suivant : pré-dénaturation pendant 15 min à 95°C, puis 40 cycles consistant en 30 secondes de dénaturation à 94°C, 1 minute d'appariement à 55°C et 1 minute d'élongation à 72°C, et enfin 30 min d'élongation finale à 60°C. Trois plaques de 96 puits (2 plaques pleine et une demi-plaque) ont été préparées à cet effet. Chaque plaque est organisée en deux multiplex (A et B) en fonction des 12 microsatellites couramment utilisés pour le génotypage du mil. Le mélange réactionnel de PCR est composé de Master Mix (1µM), Solution Q X (1 µM), Primer (0,2 µM) de chaque multiplex. Ainsi, chaque puits

des différentes plaques contient 12,3 μL de mélange réactionnel et 3,5 μL d'ADN. Chaque plaque contient les échantillons d'ADN et trois témoins : 2 témoins positifs (parents des différents croisements) et 1 témoin négatif (eau).

Le contrôle qualité des échantillons amplifiés est réalisé à l'aide d'un appareil d'électrophorèse capillaire (12 capillaires) sur gel haute résolution : QIAxcel (Qiagen®). Dans des tubes de 0,2 mL, 2 μL des produits de PCR ont été prélevés puis 8 μL de Q (X) DNA Dilution Buffer ont été ajoutés au mélange suivant le protocole recommandé par le fournisseur (Qiagen®). Les résultats ont ensuite été traités par des applications de l'appareil.

Les produits de PCR sont ensuite dilués au 1/100 et 2 μL de produit PCR dilué sont mélangés avec 10 μL de formamide + Liz 500 afin d'être séquencé par un séquenceur ABI 3500XL (Applied Biosystems®, Plateau de génotypage CIRAD). Les résultats de génotypage ont été lus avec le logiciel GeneMapper™ (version 3.7; Applied Biosystems®) qui permet d'identifier les allèles présents dans chaque échantillon.

2.5. Analyse statistique des données

Les données phénotypiques ont été traitées en faisant une analyse de variance et un test non paramétrique. Le test de Duncan a été utilisé pour comparer les moyennes. Le test de Levene a été utilisé pour évaluer l'homogénéité de la variance. R studio 3.6.1 a été utilisé pour le traitement des données.

3. Résultat

3.1. Modélisation de QTLs : Analyse de ségrégation en masse (Bulk Segregant Analysis ou BSA)

Une approche d'analyse de ségrégation en masse (BSA) sur une population F2 issue d'un croisement entre 2 lignées contrastées a été développée dans l'équipe d'accueil pour identifier des QTLs de croissance racinaire. 737 plantes F2 ont été phénotypées pour la croissance racinaire à 7 jours. Les plantes correspondant aux 10% de phénotype le plus fort et aux 10% de phénotype le plus faible ont été utilisées pour produire 2 mélanges d'ADNs groupés qui ont été génotypés par séquençage (NGS). Nous avons donc construit nos simulations pour être proche de ces paramètres.

Des scripts R de simulation ont été écrits dans le but d'optimiser les analyses de ces données de BSA pour l'étude du lien génotype/phénotype. Quatre méthodes d'analyse ont été utilisées : deux méthodes basées sur le calcul de la distance euclidienne (PEDm et PED100_4), le calcul des différences des fréquences alléliques entre les bulk : deltaSNP (PdSNP) et TricubeDeltaSNP (PTdNSP).

Dans un premier temps, nous avons étudié l'impact d'une variation de la valeur de l'effet de QTL (qtlexfect) sur la précision d'estimation de la position d'un QTL d'intérêt par BSA. Nous avons

ensuite analysé l'impact d'une variation du nombre d'individus d'une population en fonction de qtleffect sur la précision de cette estimation en utilisant les différentes méthodes statistiques.

3.1.1. Impact de la variation de qtleffect sur la précision d'estimation de la position de QTL

Au début de la simulation, des paramètres de base pour la simulation d'une analyse de ségrégation en masse dans un groupe de 500 individus d'une génération F2 ont été définis (**Annexe vi**) :

- Le nombre de chromosomes : 1000
- Le nombre de marqueurs SNP : 10000
- Le nombre de répétition de la simulation : 100
- La variation de QTL effect: de 0 à 2 avec un pas de 0,5 (2 correspond à 2 fois la standard déviation)
- Nous avons d'abord positionné le QTL au centre du chromosome (SNP position 5000)

Cette simulation retrouve bien le QTL au centre du chromosome et la précision d'estimation de la position du QTL simulé augmente avec la valeur de l'effet du QTL. Nos simulations indiquent que celle-ci est mieux estimée par la méthode PTdSNP. Aussi, à partir d'une valeur d'effet QTL de 0,5, la distribution des QTL simulés présente peu de variance c'est-à-dire que la précision de détection est plutôt bonne tandis qu'entre 0 et 0,5 l'estimation de la position du QTL simulé est encore assez variable (**Figure 7**).

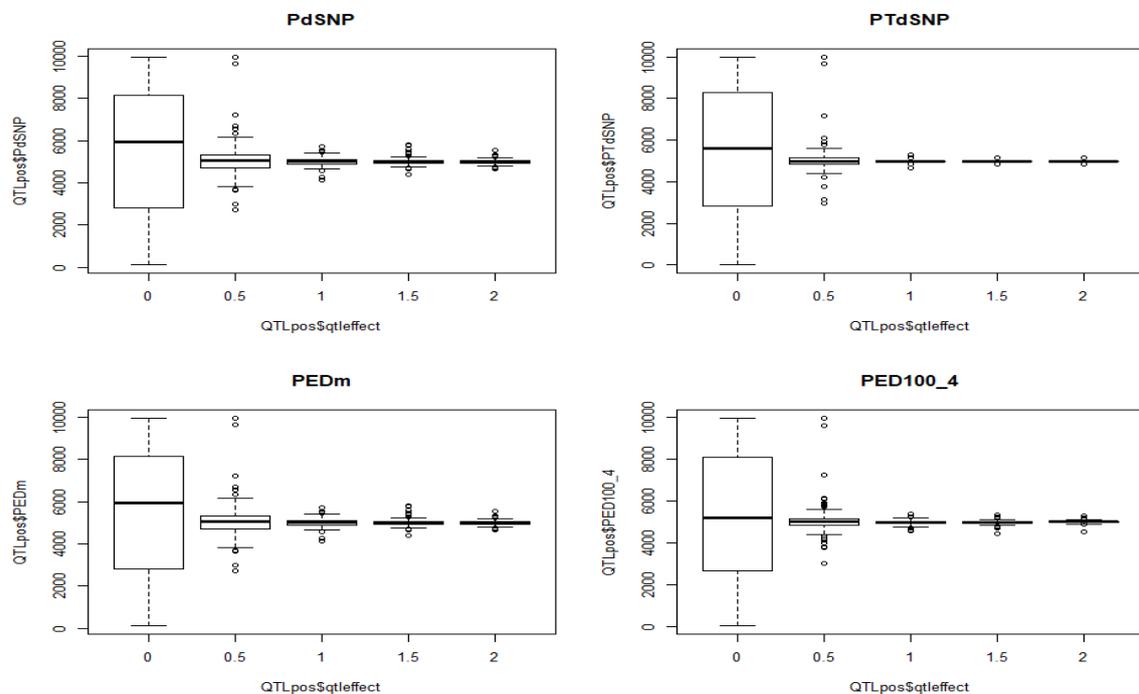


Figure 7 : Impact de la variation de QTL effect de 0 à 2 avec un pas de 0, 5 sur la précision de détection de la position de QTL simulé.

Nous avons ensuite utilisé une position aléatoire du QTL et analysé la corrélation entre la position du QTL simulé et la position estimée par le modèle. La simulation a été répétée pour un effet QTL (QTLeffect) variant entre 0 et 0,5 avec des pas de 0,1(Annexe vii). A partir de 0,1 il y a une corrélation positive (Tableau 1). Plus la valeur de l'effet QTL est grande plus la corrélation est forte statistiquement. PTdSNP (0,19) et PED100_4 (0.17) sont les statistiques qui montrent les corrélations les plus fortes dès 0,1. Ces deux méthodes sont les plus sensibles car elle permet une estimation de la position de QTL pour de faible valeur de QTL effect.

Tableau 1 : Corrélation entre les positions de QTL simulé et estimé par les différentes méthodes pour différentes valeurs de qtl effect

	Qtleffect	0	0,1	0,2	0,3	0,4	0,5
<i>PosQtl/méthode</i>							
PdSNP		-0.10	0.05	0.58	0.89	0.92	0.97
PTdSNP		0.04	0.19	0.71	0.89	0.97	0.98
PEDm		-0.10	0.05	0.58	0.89	0.92	0.97
PED100_4		0.06	0.17	0.72	0.92	0.96	0.98

3.1.2.Impact de la variation du nombre d'individus d'une population en fonction de l'effet QTL

Au début de la simulation, des paramètres de base pour la simulation d'une analyse de ségrégation en masse dans un groupe de plusieurs individus d'une génération F2 ont également été définis (Annexe viii) :

- Les différentes valeurs de QTL effect: celles-ci varient entre 0 et 5 avec un pas de 0,5
- Le nombre des individus : 100, 500,1000 et 2000
- Le nombre de répétition pour 1 QTL effect
- Le nombre de simulation
- Le nombre de marqueurs SNP : 10000

La précision d'estimation de la position de QTL simulé a été évaluée en calculant la corrélation (corrélations de Pearson) entre la position de QTL simulée aléatoirement et la position de QTL estimée par les différentes méthodes statistiques. On remarque que la précision d'estimation augmente avec le nombre d'individu et la valeur de QTL effect ; cette estimation est meilleure par la méthode PTdSNP.

Pour un petit nombre d'individus tel que 100, l'estimation de la position de QTL simulé croit avec la valeur de QTL effect (**Tableau 2**). A partir de QTL effect = 0,5 l'estimation de la position de QTL est de l'ordre de 0,6 ; elle est optimum a QTL effect = 2 pour une valeur de corrélation de l'ordre de 0,9. L'estimation est meilleure par la méthode PTdSNP.

Tableau 2 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 100 individus

	<i>qtleffect</i>	0	0,5	1	1,5	2
<i>PosQtl/méthode</i>						
PdSNP		-0.12	0.66	0.88	0.98	0.993
PTdSNP		-0.18	0.69	0.89	0.987	0.995
PEDm		-0.12	0.66	0.88	0.983	0.993
PED100_4		-0.12	0.66	0.89	0.986	0.992

Lorsque le nombre d'individus est de 500, la corrélation entre la position de QTL simulé et celle estimée est très forte à partir de QTL effect =0,5. L'augmentation est minime pour de plus grande valeur de QTL effect (**Tableau 3**).

Tableau 3 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 500 individus

	<i>qtleffect</i>	0	0,5	1	1,5	2
<i>PosQtl/méthode</i>						
PdSNP		-0.005	0.960	0.992	0.998	0.9994
PTdSNP		0.115	0.986	0.998	0.999	0.9996
PEDm		-0.005	0.960	0.992	0.998	0.9994
PED100_4		0.080	0.981	0.998	0.999	0.9998

Pour un grand nombre d'individus (>1000), dès 0.5, les corrélations sont presque parfaites (**Tableau 4**).

Tableau 4 : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour 1000 individus

	<i>qtleffect</i> 0	0,5	1	1,5	2
PosQtl/méthode					
PdSNP	0.06	0.991	0.997	0.999	0.9995
PTdSNP	0.014	0.997	0.999	0.999	0.9997
PEDm	0.061	0.991	0.997	0.999	0.9995
PED100_4	0.060	0.998	0.999	0.999	0.9999

Nous avons donc aussi analysé plus finement pour des effets QTL variant de 0 à 0,5 avec un pas de 0,1. La précision d'estimation de la position de QTL est clairement fonction du nombre d'individus et de la valeur de QTL effect. Plus le nombre d'individu est grand, meilleure est la prédiction (valeur de la corrélation). Une taille entre 500 et 1000 paraît un bon compromis (**Annexe ix- Annexe xiv**).

En conclusion, nos analyses indiquent que plus le QTL a un effet fort, meilleure sera la précision d'estimation de sa position. Le nombre d'individus utilisé pour la simulation est également un paramètre important. Un grand effectif d'individus améliore aussi la précision d'estimation de la position du QTL simulé. Nous avons comparé différentes méthodes statistiques d'estimation de la position de QTL simulé. Nos analyses suggèrent que la méthode statistique PTdSNP (et souvent PED100_4) est la plus performante pour estimer la position du QTL simulé sur une large gamme de paramètres testés et en particulier pour des QTLs à effet faible et pour des effectifs moins importants. Les scripts obtenus serviront pour l'analyse des données de phénotypage de la population.

3.2. Analyses de lignées recombinantes fixées

Des lignées « Backcross Inbred lines » (BILs correspondant à des individus BC2F4) ont été obtenues par l'équipe d'accueil à partir de croisement de parents très contrastés pour le phénotype longueur racinaire : le parent 249 caractérisé par une croissance racinaire lente et le parent 337 caractérisé par une croissance racinaire rapide. Nous avons cherché à caractériser l'homogénéité de ces lignées.

Nous avons des attendus concernant la variabilité du matériel végétal obtenu : un taux

d'hétérozygotie de l'ordre de 6% dans la descendance BC2F4, celle-ci doit également présenter uniquement les allèles parentaux. En raison du nombre d'autofécondations réalisées, nous nous attendons aussi à une homogénéité des phénotypes. Cependant, le mil étant une plante allogame, il y a des risques de croisements allogames lors des phases d'autofécondation. Deux analyses ont donc été réalisées pour vérifier l'homogénéité des lignées :

- 1) une analyse basée sur le génotypage des lignées BC2F4 avec des marqueurs microsatellites,
- 2) une analyse basée sur le phénotypage de ces lignées.

3.2.1. Analyse de l'homogénéité génétique des lignées

Nous avons tout d'abord lancé une analyse de génotypage par microsatellites (12 marqueurs) sur 10 lignées BILs. Nous avons planifié l'analyse sur 10 plantes par lignée BIL et avons inclus les parents du croisement. L'extraction et l'analyse par PCR microsatellite des échantillons a été entamée. Une des difficultés rencontrées lors de la mise en œuvre de cette expérience a été la présence de faux positifs dans certain des témoins négatifs (**Annexe xv**). Plusieurs mises au point ont été faites pour supprimer ce problème et une nouvelle amplification d'ADN n'a pas révélé d'amplification dans les témoins négatifs. Cependant, vu le temps de stage imparti et les problèmes d'accès à la plateforme d'analyse de nos PCR, nous n'avons pas pu finaliser ce génotypage.

3.2.2. Analyse de l'homogénéité phénotypique des lignées

Afin de phénotyper les lignées BC2F4 et s'assurer de leur homogénéité phénotypique, nous avons utilisé un test de croissance racinaire sur papier enroulé sur 10 lignées BILs et les parents du croisement (**Figure 8**). Le phénotypage a été réalisé en trois expériences indépendantes avec dix plantes phénotypées par génotype et par expérience.



Figure 8 : Dispositif expérimental

Nous avons tout d'abord voulu s'assurer que les deux parents présentaient dans les conditions expérimentales appliquées le phénotype racinaire attendu (croissance rapide/lente). Après 7 jours de croissance, les parents montrent une taille moyenne de racine de 9,59 cm pour la lignée 249 et de 16 cm pour la lignée 337 (**Tableau 5, Annexe xvi**). Le parent 337 présente un phénotype racinaire en moyenne supérieur à celui de la lignée 249 avec des médianes respectivement égales à 17 et 11 (**Figure 9**). La différence de moyenne des longueurs racinaires est de 6,4 cm et est très significative (Test de Wilcoxon, $W = 56$, $p\text{-value} = 5.6 \cdot 10^{-5}$) et leur variance ne sont pas significativement différentes (Test de Levene, $F\text{-value} = 2.84$, $p\text{-value} = 0.51$).

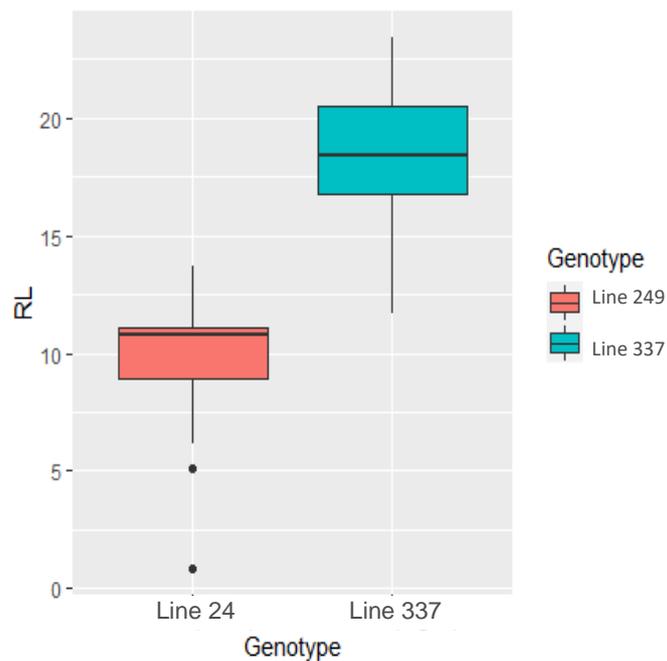


Figure 9 : Distribution des phénotypes parentaux
RL : Longueur racinaire en cm.

Nous avons ensuite étudié la distribution des lignées BC2F4 par rapport aux parents (**Figure 10**). On observe une certaine hétérogénéité entre les lignées : on remarque que les variances des lignées sont proches de celle du parent 337 bien que les distributions ne soient pas équi-variées. Les lignées 2, 124 et 132 ont des médianes approximativement égales ; celles-ci se trouvent entre les médianes des parents. Les autres lignées ont des médianes supérieures à celle des parents. Les lignées 125 et 130 ont également des médianes égales.

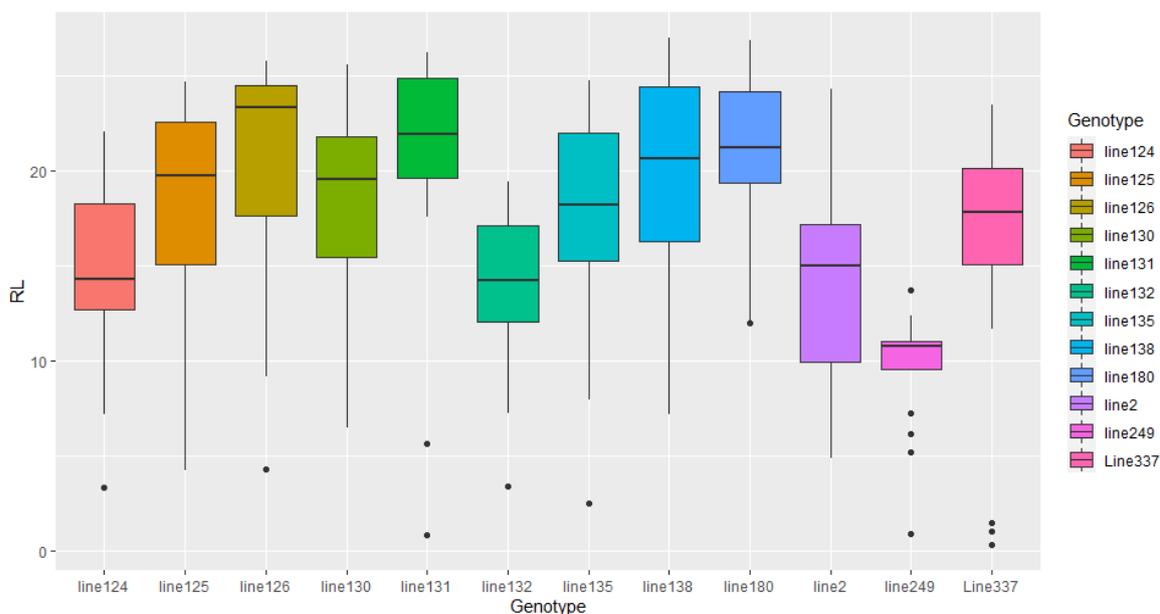


Figure 10 : Distribution des lignées BC2F4 et leur parent en fonction de la longueur racinaire

On remarque également qu'en moyenne la longueur racinaire de quelques lignées BC2F4 (line 125, 126, 130, 131, 135, 135, 138, 180) est supérieure à celle des parents (**Tableau 5**).

Tableau 5 : Longueur racinaire moyenne et écart à la moyenne des différentes lignées

Génotype	Longueur racinaire moyenne	Ecart type
line 2	14,1	5,48
line124	14.5	4.50
line125	18.6	5.28
line126	20.3	6.01
line130	18.1	4.68
line131	20.9	5.97
line132	14.0	3.82
line135	17.3	6.06
line138	19.7	6.08
line180	20.7	4.26
line249	9.59	3.01
line337	16.0	6.23

Nous avons ensuite procédé à l'analyse de la variance (ANOVA) des 10 lignées BC2F4 (Lignées 2, 124, 125, 126, 130, 131, 132, 135, 138 et 180). Le modèle statistique de l'ANOVA obtenu par la technique « Backfor selection » basé sur le test de Fisher inclut un effet génotype, répétition (3 répétitions réalisés) et d'interaction génotype x répétition (**Annexe xvii**) :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i * \beta_j + \epsilon_{ijk}$$

Où Y_{ijk} désigne la longueur racinaire k pour les lignées i , pour chaque répétition j et suit une loi normale de paramètre $(0, \sigma^2_{\text{Génotype}} + \sigma^2_{\text{répétition}} + \sigma^2_{\text{génotype*répétition}} + \sigma^2_{\text{résiduelle}})$. Cette longueur racinaire observée se décompose en 5 termes :

- μ est la moyenne générale
- α_i correspond à l'effet des lignées (ou génotype) sur la longueur racinaire k et suit une loi normale de paramètre $(0, \sigma^2_{\text{Génotype}})$
- β_j correspond à l'effet de la répétition sur la longueur racinaire k et suit une loi normale de paramètre $(0, \sigma^2_{\text{répétition}})$
- $\alpha_i * \beta_j$ correspond à l'effet de l'interaction entre le génotype i et la répétition j sur la longueur racinaire et suit une loi normale $(0, \sigma^2_{\text{génotype*répétition}})$
- ϵ_{ijk} suit une loi normale de paramètre $(0, \sigma^2_{\text{résiduelle}})$ et représente la variabilité résiduelle.

L'ANOVA montre un effet significatif du génotype ($F_{12,259}=11.35, p = 2 \cdot 10^{-16}$), de la répétition ($F_{2,259}=9.97, p=6,7 \cdot 10^{-5}$) ainsi que de l'interaction entre les deux facteurs ($F_{17,259} = 2.71, p=3 \cdot 10^{-4}$). Le modèle présenté explique 36% de la variabilité observée (Adjusted R-squared) avec une p -value significative (p -value $< 2,2 \cdot 10^{-16}$) ; la statistique de Fisher est égale à 6,36 (**Tableau 6**).

Tableau 6 : Résultat de l'ANOVA sur les lignées

Variation	Somme des carrés	ddl	Carré moyen	Valeur Fobs	p-value
Génotype	2859	11	259,89	11,35	$< 2.10^{-16}$
Répétition	457	2	228,50	9,97	$6.7.10^{-5}$
Génotype*Répétition	1056	17	62,09	2,71	$0,3.10^{-3}$
Erreur	5932	259	22.90		

L'analyse des résidus montrent qu'ils peuvent être considéré comme indépendants ($DW= 1.97$), mais ils s'éloignent d'une distribution normale (Test de Shapiro, $W = 0,92, p$ -value= $1,1 \cdot 10^{-9}$), et présentent une hétérogénéité des variances (Test de Bartlett, p -value $< 1,6 \cdot 10^{-8}$) (**Figure 11**).

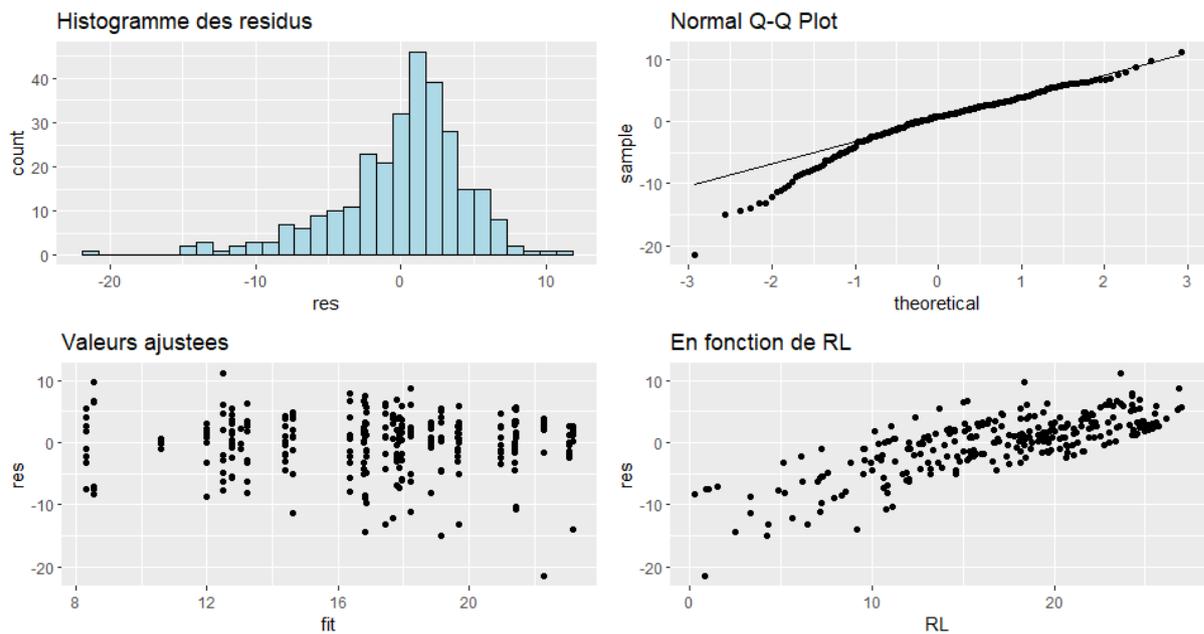


Figure 11 : Analyse des résidus

Pour corriger cela, nous avons donc procédé à une transformation de données en utilisant la transformation Box-Cox. Elle est définie par la formule suivante :

$$B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0 \end{cases}$$

Quand lambda est différent de zéro, la transformation BoxCox est très proche d'une transformation puissante puisqu'elle retranche 1 et divise par lambda, qui est une constante. Les sorties R de la fonction donne premièrement l'estimation du coefficient lambda. Celle-ci est faite par le maximum de vraisemblance. Ensuite un test statistique est réalisé pour évaluer si lambda peut être fixé à 0 c'est-à-dire pour évaluer si une simple transformation log est suffisante. Enfin, un autre test statistique est réalisé pour évaluer si la transformation BoxCox (avec lambda =1 ou lambda différent de 1) est réellement nécessaire. A l'issue de la transformation grâce à la fonction powerTransform, lambda a été estimé à 1,51 avec un intervalle de confiance à 95% = [1,27 ; 1,76]. Les données brutes ont donc été transformées en utilisation cette valeur de lambda. Une analyse de variance a été effectuée sur les données transformées, on obtient toujours un effet significatif du génotype ($F=13.257$, $p\text{-value} < 2$

10^{-16}), de la répétition ($F=12.12$, $p\text{-value} = 9.32 \cdot 10^{-6}$) et l'interaction ($F=2.71$, $p\text{-value} = 9.32 \cdot 10^{-6}$). Le modèle explique plus de variabilité soit 39% (Adjusted R-squared: 0.39) (**Tableau 7**).

Tableau 7 : Résultat de l'ANOVA des données transformées

Variation	Somme des carrés	ddl	Carré moyen	Valeur Fobs	p-value
Génotype	50533	11	4594	13.26	$< 2 \cdot 10^{-16}$
Répétition	8398	2	4199	12.12	$9.32 \cdot 10^{-6}$
Génotype*Répétition	15938	17	938	2.71	$4 \cdot 10^{-4}$
Erreur	89749	259	347		

Les résidus sont considérés indépendants ($DW= 1.96$), mais montrent toujours un écart à la loi normale (Test de Shapiro, $p\text{-value}= 9.45 \cdot 10^{-07}$) même s'ils sont 100 fois moins significatifs. La variance reste aussi hétérogène (Test de Bartlett, $p\text{-value} = 1.91 \cdot 10^{-07}$) (**Figure 12**).

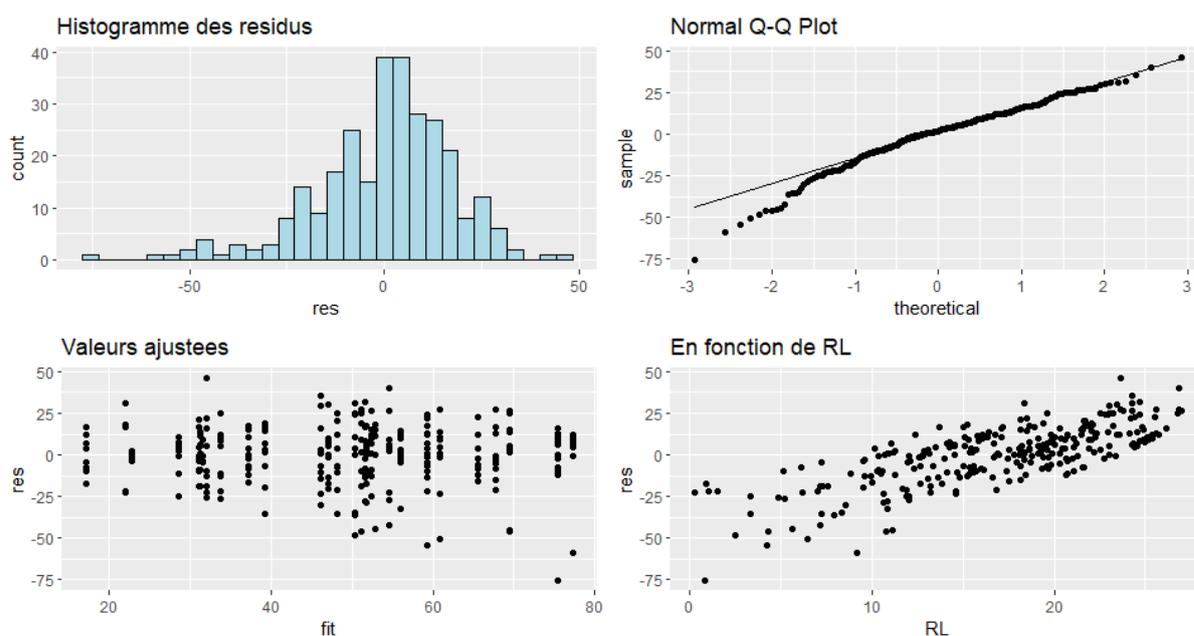


Figure 12 : Analyse des résidus des données transformées

Nous avons ensuite réalisé le test non paramétrique de Kruskal-Wallis. Du fait d'une évidence d'interaction entre génotype et répétition dans les modèles ANOVA, le test est réalisé par répétition. Une différence significative est observée entre les lignées dans les répétitions B ($p\text{-value} = 4.9 \cdot 10^{-08}$) et C ($p\text{-value} = 3.68 \cdot 10^{-08}$). Les résultats du test dans la première répétition montre qu'il n'y a pas de différence significative entre les lignées ($p\text{-value}=0.054$), cela pourrait s'expliquer également par le fait que dans la répétition A il n'y a pas toutes les lignées BC2F4. La comparaison par paire avec le test de Dunn confirme également les résultats obtenus.

On remarque que la répétition a un effet sur les distributions ; Il y a une certaine hétérogénéité. Dans la répétition B, la répartition en fonction des moyennes est présentée dans la **Figure 13**. Les moyennes sont bel et bien différentes. Cependant on retrouve quelque groupe qui se forme : les lignées 124 et 2 ont des moyennes qui ne sont pas significativement différentes, les lignées 135, 130 et 125, les lignées 131 et 180 ont la même moyenne du parent avec une croissance racinaire rapide.

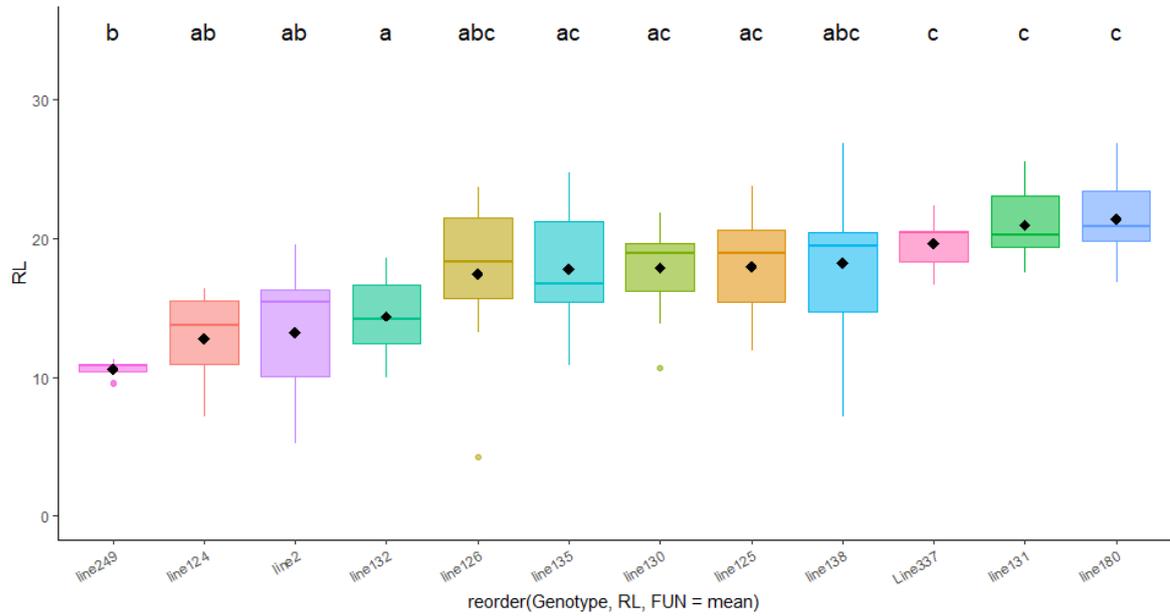


Figure 13 : Comparaison des moyennes des lignées dans la répétition B

Dans la répétition C, il se forme également des groupes (**Figure 14**) : les lignées 2 et 132, les lignées 125 et 131, les lignées 130 et 138. La plupart des lignées ont des moyennes élevées comparé aux parents.

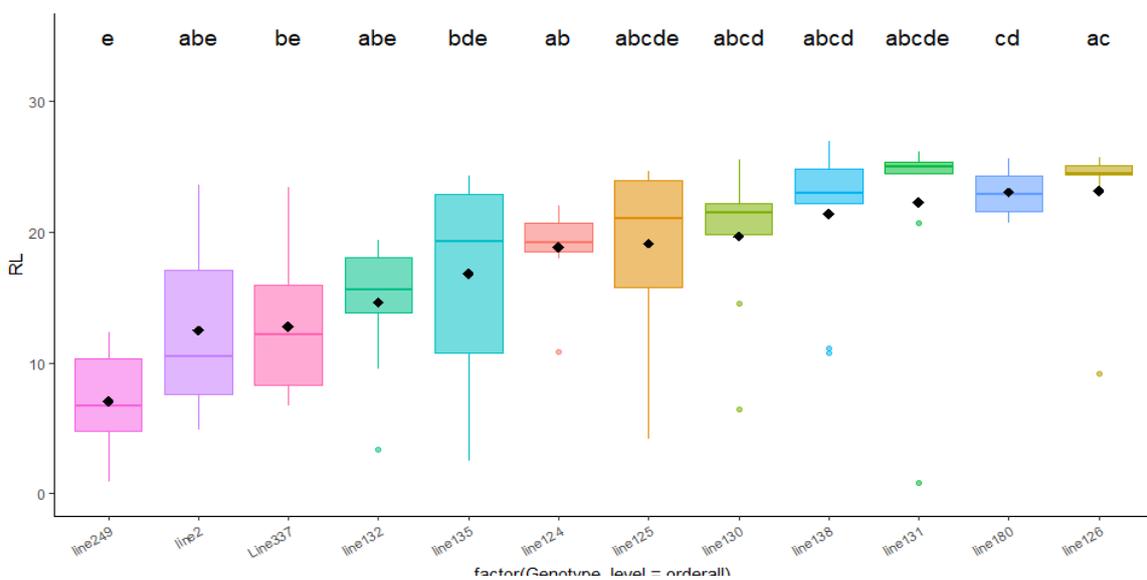


Figure 14 : Comparaison de moyenne des lignées dans la répétition C

Lorsqu'on fait une comparaison de moyenne globale des lignées, on observe que les lignées BC2F4 ont des longueurs racinaires significativement différentes de celles des parents. Cependant, il y a quelques similitudes : les moyennes des lignées 130 et 125 ne sont pas significativement différentes ; il en est de même pour les lignées 180 et 131 (**Figure 15**).

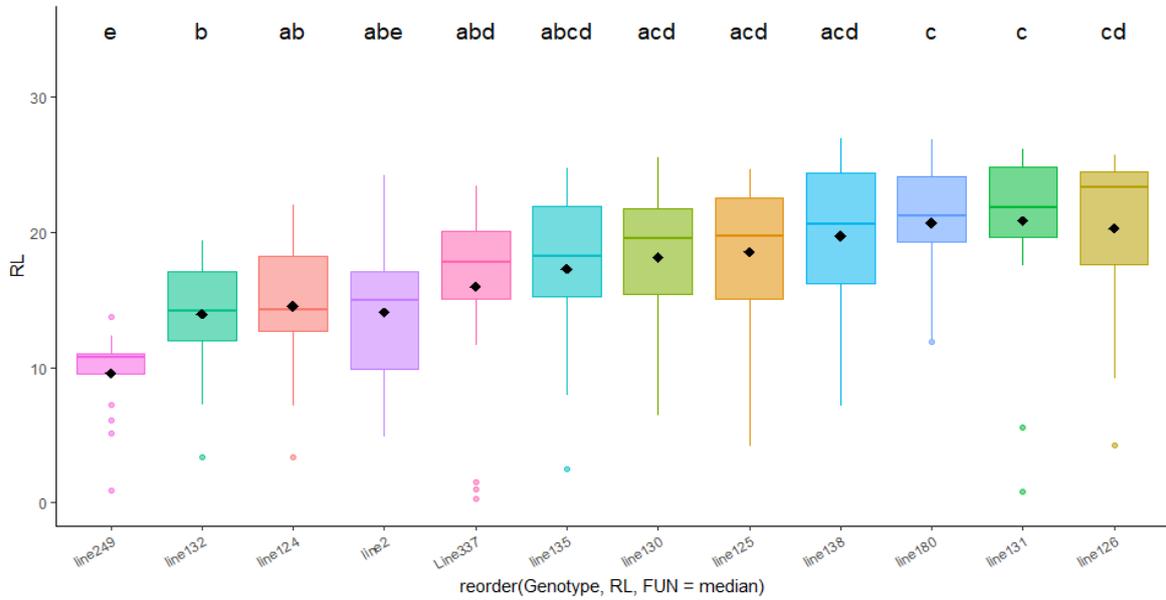


Figure 15 : Comparaison globale de moyenne de longueur racinaire des lignées

Enfin en prenant en compte les résultats de l'ANOVA où nous avons identifié un effet significatif de la répétition, l'analyse de l'homogénéité de la variance a donc été faite par répétition. Nous avons calculé la variance des lignées par répétition ainsi que la moyenne des variances (**Tableau 8**). On remarque que dans la répétition A la variance du parent 337 est supérieur à celle des BC2F4, dans les répétitions B et C la variance des lignées BC2F4 est plus grande que celle des parents. Nous avons donc cherché à savoir si cette différence est significative. Le test de Levene a révélé une différence de variance significative dans la première répétition avec une *p-value* de $4,9 \cdot 10^{-3}$, la différence de variance n'est pas significative dans les deux autres répétitions avec respectivement des *p-value* égale à 0.08 et 0.43. Cette significativité de la différence de variance dans la première répétition est due au fait que la lignée parentale 337 a une grande variance dans la répétition A. Les phénotypes des BC2F4 n'ayant pas une variance supérieure à celle du parent 337 qui a la plus grande variance dans cette répétition nous pouvons dire qu'ils ont une variance similaire aux lignées parentales. Si on considère les lignées parentales homogènes, ces lignées BC2F4 peuvent être aussi considérées homogènes.

Tableau 8 : Variances

Lignées	Variance Repetition A	Variance Répétition B	Variance Répétition C	Moyenne des variances
Lignée 2	26.287878	22.4501071	40.218109	29.652032
Lignée 124	12.125798	11.8020425	9.848503	11.258781
Lignée 125	NA	16.3063718	41.769290	29.037831
Lignée 126	NA	33.1360613	24.666963	28.901512
Lignée 130	23.846209	12.5985954	29.506254	21.983686
Lignée 131	47.661757	6.5379698	59.165148	37.788292
Lignée 132	11.701256	8.5778204	24.219323	14.832800
Lignée 135	NA	17.0536441	59.989932	38.521788
Lignée 138	NA	34.4406193	38.294002	36.367311
Lignée 180	2.468275	9.4356679	3.340757	5.081567
Lignée 337	71.587738	3.3368081	12.116824	29.013790
Lignée 249	NA	0.3606413	18.270703	9.315672

4. Discussion

Dans ce travail, nous avons participé à des travaux visant à identifier des QTLs contrôlant la croissance de la racine primaire chez le mil à partir de populations en ségrégation issues d'un croisement entre deux parents contrastés par deux approches : une approche de ségrégation en masse sur une population F2 et une approche d'analyse de lignées recombinantes BC2F4. Ce trait a été ciblé car il est associé dans des essais au champ à une meilleure tolérance à un stress hydrique précoce (Equipe CERES, non publié). Chez les plantes, les populations en ségrégation sont généralement utilisées pour localiser et comparer les loci contrôlant des caractères quantitatifs ; celles-ci sont généralement génotypées avec des marqueurs moléculaires. Cependant elles peuvent également être regroupées selon l'expression phénotypique du caractère d'intérêt et évaluer pour les différences de fréquences alléliques entre les deux groupes de population (Takagi *et al.*, 2013). L'analyse de ségrégation en masse est un outil précieux pour l'identification rapide de marqueurs dans une région chromosomique associée à un trait d'intérêt d'autant plus qu'on peut utiliser n'importe quel marqueur co-dominant y compris les marqueurs SNP. Cela a permis l'adaptation de cette technologie pour une utilisation avec des lectures de séquençage de nouvelle génération (NGS). Plusieurs méthodes d'analyse de ces résultats ont été développées (Hill *et al.*, 2013 ; Takagi *et al.*, 2013).

Pour essayer de déterminer la meilleure méthode d'analyse des données de BSA produites dans l'équipe d'accueil, quatre méthodes ont été testées. Dans ce travail, nous avons, par une étude de simulation, évalué l'impact de la variation de QTL effect, l'effet dominant du QTL sur le phénotype

d'intérêt, sur la capacité de détection de QTL de ces quatre méthodes. Nous avons montré que plus la valeur de ce paramètre est élevée plus la corrélation entre la position QTL simulé et celle de QTL estimé est grande pour l'ensemble des 4 méthodes c'est-à-dire il y a une meilleure estimation de la position du QTL. La méthode statistique PTdSNP s'est révélé être la méthode qui permet une meilleure estimation de la position de QTL simulé aléatoirement car elle est capable d'une meilleure détection des QTLs à faible effet comparé aux autres méthodes.

Aussi, en analyse de ségrégation en bulk, une question importante est la taille des échantillons d'individus qui doivent former les groupes d'individus contrastés (Guo *et al.*, 2015). On a remarqué lors de la simulation qu'avec un nombre d'individus élevé, l'estimation de la position de QTL est meilleure. Cela est sans doute dû au fait que lorsque les individus sont nombreux : 1) il y a plus de possibilité de crossing-over, 2) la taille des 10% extrêmes permet une meilleure estimation des fréquences. La méthode statistique PTdSNP s'est révélée être la méthode qui permet une meilleure estimation de la position de QTL simulé aléatoirement car elle est capable de détecter des QTLs pour de faibles effectifs de population.

La deuxième approche à laquelle nous avons contribué est une analyse de lignées recombinantes issues de rétrocroisements BC2F4. Il est intéressant de travailler avec des lignées génétiquement homogènes telles que les BC2F4 pour pouvoir effectuer des mesures phénotypiques sur un groupe d'individus isogéniques, permettant ainsi d'accroître la précision de ces mesures, d'évaluer l'influence des facteurs environnements et d'évaluer la variance intra-groupe. Ces lignées permettent une meilleure dissection génétique des caractères à déterminisme complexe et offrent une plus grande puissance statistique pour mettre en évidence des interactions épistatiques entre gènes du fait que les différences génétiques entre lignées concernent une plus petite fraction du génome du parent donneur. En effet, les lignées BC2F4 possèdent tous un fond génétique provenant en majorité du parent récurrent, dans notre cas le parent avec une croissance racinaire faible, et différent uniquement pour certaines régions du génome. D'un point de vue génétique, nous nous attendons à des individus homozygotes présentant que les deux allèles parentaux. Cependant le mil étant une plante fortement allogame, il y a des risques de croisement de nos lignées avec des pollens étrangers au cours de la sélection. Nous avons donc essayé d'évaluer l'homogénéité des lignées par génotypage microsatellite et par phénotypage de la croissance racinaire d'un échantillon de 10 lignées (sur les 282 lignées reçues). L'analyse de microsatellite n'a pu être menée à son terme faute de temps. L'évaluation de l'homogénéité phénotypique des lignées BC2F4 et de leurs parents a révélé, comme attendu, une hétérogénéité inter lignées. Le calcul des variances phénotypiques de chaque lignée a montré que la variance observée n'est pas significativement différente de celle des parents. Les parents étant homogènes, cela suggère que les lignées BC2F4 testées sont homogènes. La répétition a un effet significatif sur la longueur racinaire des lignées, ce qui suggère des voies d'amélioration pour avoir des phénotypes plus reproductibles. Nous avons quelques lignées avec des longueurs racinaires

supérieure à celle des parents. Cela pourrait s'expliquer par les phénomènes épistatiques et le fait que le caractère soit contrôlé par plusieurs locus. Les résultats de génétique d'association réalisée dans l'équipe d'accueil avaient mis en évidence au moins 9 loci contrôlant ce trait.

La méthode expérimentale utilisée pour le phénotypage semble appropriée car elle permet de mesurer facilement les racines et que celles-ci ne s'entremêlent pas. Mais l'effet significatif répétés suggèrent des optimisations à faire dans le phénotypage. L'utilisation du logiciel de traitement d'images de racines ImageJ permet également de réduire considérablement les erreurs dues à l'expérimentateur.

Conclusion

La croissance racinaire du mil ainsi que l'agrégation du sol rhizosphérique sont des caractères contrôlés par plusieurs régions chromosomiques. L'objectif de ce stage était de les identifier. Cependant, ce travail ne m'a permis que d'évaluer l'homogénéité des lignées BC2F4 obtenues à partir de deux parents dont l'un est caractérisé par une croissance racinaire rapide et l'autre une croissance racinaire lente. Nous avons produit de l'ADN à partir de l'ensemble des lignées BC2F4 qui a été envoyé à un prestataire de service pour réaliser un génotypage par séquençage (GBS). Les données de GBS permettront de procéder à la détection plus précise des QTL d'intérêt pour les deux traits étudiés.

Aussi, compte tenu des modifications apportées au programme lié à la fermeture du laboratoire pour raisons sanitaires, nous avons réalisé une simulation d'analyse de ségrégation en masse. Cette approche théorique nous a permis d'évaluer différentes méthodes statistiques et l'effet du nombre d'individus étudiés. Cela nous a permis de définir la méthode la plus efficace pour la détection de QTL par BSA. Cette méthode est actuellement utilisée pour détecter des QTLs de croissance racinaire chez le mil à partir de données de phénotypage et de génotypage de plantes F2 issues d'un croisement entre deux parents contrastés.

Travail de l'étudiant

Au cours de ces six mois de stage, j'ai contribué à la mise en place du phénotypage et du génotypage des lignées BC2F4. En effet, j'ai choisi les lignées BC2F4 (10 lignées par caractère d'intérêt) qui ont été utilisées au cours de l'expérimentation, celles-ci ont été choisies sur la base du nombre de graines disponibles pour chaque lignée. J'ai réalisé le phénotypage de 10 lignées BC2F4 et leur parent pour le caractère croissance racinaire par la méthode expérimentale expliquée dans la partie matériel et méthode, les données collectées ont été traitées avec le logiciel ImageJ. J'ai analysé statistiquement les données récoltées.

Aussi, j'ai semé et suivi la croissance de ces mêmes lignées afin d'évaluer leur diversité génotypique pour les deux caractères d'intérêt : croissance racinaire et agrégation rhizosphérique. J'ai collecté le matériel végétal (morceaux de feuilles), extrait l'ADN par la méthode Ampure, amplifié les échantillons par PCR, quantifié l'ADN et vérifié la qualité des échantillons avec Qiaxcel mais aussi sur gel d'agarose. J'ai fait un génotypage de 24 échantillons et j'ai analysé les données avec le logiciel GenMapper.

J'ai également produit les plantes qui ont servi pour le génotypage GBS.

Enfin, j'ai développé des Scripts R de simulation pour faire des analyses de ségrégation en masse.

Références bibliographiques

- Andrews D.J. et Kumar K.A. 1992. Pearl Millet for Food, Feed, and Forage. Dans : Sparks D.L. (éd.). Academic Press, p. 89-139.
- Bai C., Liang Y., et Hawkesford M.J. 2013. Identification of QTLs associated with seedling root traits and their correlation with plant height in wheat. *Journal of Experimental Botany*. 64(6), p. 1745-1753.
- Brunken J.N. 1977. A SYSTEMATIC STUDY OF PENNISETUM SECT. PENNISETUM (GRAMINEAE). *American Journal of Botany*. 64(2), p. 161-176.
- Burgarella C., Cubry P., Kane N.A., Varshney R.K., Mariac C., Liu X., Shi C., Thudi M., Couderc M., Xu X., Chitikineni A., Scarcelli N., Barnaud A., Rhoné B., Dupuy C., François O., Berthouly-Salazar C., et Vigouroux Y. 2018. A western Sahara centre of domestication inferred from pearl millet genomes. *Nature Ecology & Evolution*. 2(9), p. 1377-1380.
- Burton A.L., Johnson J.M., Foerster J.M., Hirsch C.N., Buell C.R., Hanlon M.T., Kaeppler S.M., Brown K.M., et Lynch J.P. 2014. QTL mapping and phenotypic variation for root architectural traits in maize (*Zea mays* L.). *Theoretical and Applied Genetics*. 127(11), p. 2293-2311.
- Dawkins R. 1982. *The Extended Phenotype*.
- De la Fuente Cantó C., Simonin M., King E., Moulin L., Bennett M.J., Castrillo G., et Laplaze L. 2020. An extended root phenotype: the rhizosphere, its formation and impacts on plant fitness. *The Plant Journal*. , p. tpj.14781.
- Debieu M., Kanfany G., et Laplaze L. 2017. Pearl Millet Genome: Lessons from a Tough Crop. *Trends in Plant Science*. 22(11), p. 911-913.
- Debieu M., Sine B., Passot S., Grondin A., Akata E., Gangashetty P., Vadez V., Gantet P., Foncéka D., Cournac L., Hash C.T., Kane N.A., Vigouroux Y., et Laplaze L. 2018. Response to early drought stress and identification of QTLs controlling biomass production under drought in pearl millet. *PLOS ONE*. 13(10), p. e0201635.
- Dickson E.L., Rasiah V., et Groenevelt P.H. 1991. Comparison of four prewetting techniques in wet aggregate stability determination. *Canadian journal of soil science*. 71(1), p. 67-72.
- FAOSTAT. 2018. FAOSTAT. Dans : *Organisation des Nations unies pour l'alimentation et l'agriculture* [En ligne]. Disponible sur : < <http://www.fao.org/faostat/fr/> > (Consulté le 29 avril 2020).
- Fuller D. et Macdonald K. 2007. Early domesticated pearl millet in Dhar N'ema (Mauritania): evidence. *Fields of change: progress in African archaeobotany*. 5, p. 71.
- Guo J., Fan J., Hauser B.A., et Rhee S.Y. 2015. Target Enrichment Improves Mapping of Complex Traits by Deep Sequencing. *G3: Genes/Genomes/Genetics*. 6(1), p. 67-77.
- Hartmann A., Rothballer M., et Schmid M. 2008. Lorenz Hiltner, a pioneer in rhizosphere microbial ecology and soil bacteriology research. *Plant and Soil*. 312(1-2), p. 7-14.
- Hill J.T., Demarest B.L., Bisgrove B.W., Gorski B., Su Y.-C., et Yost H.J. 2013. MMAPP: Mutation Mapping Analysis Pipeline for Pooled RNA-seq. *Genome Research*. 23(4),

p. 687-697.

- Kadri B.K., Halilou H., et Karimou I. 2019. Culture du mil [*Pennisetum glaucum* (L.) R. Br] et ses contraintes à la production: une revue. *International Journal of Biological and Chemical Sciences*. 13(1), p. 503.
- Koebnick N., Daly K.R., Keyes S.D., George T.S., Brown L.K., Raffan A., Cooper L.J., Naveed M., Bengough A.G., Sinclair I., Hallett P.D., et Roose T. 2017. High-resolution synchrotron imaging shows that root hairs influence rhizosphere soil structure formation. *New Phytologist*. 216(1), p. 124-135.
- Lynch J.P. 2019. Root phenotypes for improved nutrient capture: an underexploited opportunity for global agriculture. *New Phytologist*. 223(2), p. 548-564.
- Maiti R.K. et Bidinger F.R. 1981. Growth and development of the pearl millet plant. , p. 19.
- Manning K., Pelling R., Higham T., Schwenniger J.-L., et Fuller D.Q. 2011. 4500-Year old domesticated pearl millet (*Pennisetum glaucum*) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. *Journal of Archaeological Science*. 38(2), p. 312-322.
- Mansfeld B.N. et Grumet R. 2018. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. *The Plant Genome*. 11(2), p. 180006.
- Martinez J. 2011. *Expression et évolution du phénotype étendu dans une association parasitoïde-virus*. Lyon : Université Claude Bernard Lyon I
- Mbaye D.F. 1993. Contraintes phytosanitaires du mil dans le Sahel : état des connaissances et perspectives. Dans : *Contraintes phytosanitaires dans le Sahel*. , p. 14.
- Moumouni H. 2014. *Construction d'une carte génétique pour le mil, Pennisetum glaucum (L.) R. Br., par une approche de génotypage par séquençage (GBS)*. Québec, Canada : Université Laval
- Muller B., Guédon Y., Passot S., Lobet G., Nacry P., Pagès L., Wissuwa M., et Draye X. 2019. Lateral Roots: Random Diversity in Adversity. *Trends in Plant Science*. 24(9), p. 810-825.
- Ndour P.M.S., Gueye M., Barakat M., Ortet P., Bertrand-Huleux M., Pablo A.-L., Dezette D., Chapuis-Lardy L., Assigbetsé K., Kane N.A., Vigouroux Y., Achouak W., Ndoye I., Heulin T., et Cournac L. 2017. Pearl Millet Genetic Traits Shape Rhizobacterial Diversity and Modulate Rhizosphere Aggregation. *Frontiers in Plant Science*. 8, p. 1288.
- Orman-Ligeza B., Parizot B., Gantet P.P., Beeckman T., Bennett M.J., et Draye X. 2013. Post-embryonic root organogenesis in cereals: branching out from model plants. *Trends in Plant Science*. 18(8), p. 459-467.
- Oumar I., Mariac C., Pham J.-L., et Vigouroux Y. 2008. Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theoretical and Applied Genetics*. 117(4), p. 489-497.
- Pang J., Ryan M.H., Siddique K.H.M., et Simpson R.J. 2017. Unwrapping the rhizosheath. *Plant and Soil*. 418(1-2), p. 129-139.
- Passot S. 2016. *Exploration du système racinaire du mil et ses conséquences pour la tolérance à la sécheresse - Exploring pearl millet root system and its outcome for drought tolerance*.

- Passot S., Gnacko F., Moukouanga D., Lucas M., Guyomarc'h S., Ortega B.M., Atkinson J.A., Belko M.N., Bennett M.J., Gantet P., Wells D.M., Guédon Y., Vigouroux Y., Verdeil J.-L., Muller B., et Laplaze L. 2016. Characterization of Pearl Millet Root Architecture and Anatomy Reveals Three Types of Lateral Roots. *Frontiers in Plant Science*. 7.
- Passot S., Moreno-Ortega B., Moukouanga D., Balsera C., Guyomarc'h S., Lucas M., Lobet G., Laplaze L., Muller B., et Guédon Y. 2018. A New Phenotyping Pipeline Reveals Three Types of Lateral Roots and a Random Branching Pattern in Two Cereals. *Plant Physiology*. 177(3), p. 896-910.
- Rabbi S.M.F., Tighe M.K., Flavel R.J., Kaiser B.N., Guppy C.N., Zhang X., et Young I.M. 2018. Plant roots redesign the rhizosphere to alter the three-dimensional physical architecture and water dynamics. *New Phytologist*. 219(2), p. 542-550.
- Rich S.M. et Watt M. 2013. Soil conditions and cereal root system architecture: review and considerations for linking Darwin and Weaver. *Journal of Experimental Botany*. 64(5), p. 1193-1208.
- Shweta M. 2015. Pearl millet nutritional value and medicinal uses. *Antiquity*. (295), p. 414-418.
- Smith S. et De Smet I. 2012. Root system architecture: insights from *Arabidopsis* and cereal crops. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 367(1595), p. 1441-1452.
- Sultan B., Guan K., Kouressy M., Biasutti M., Piani C., Hammer G.L., McLean G., et Lobell D.B. 2014. Robust features of future climate change impacts on sorghum yields in West Africa. *Environmental Research Letters*. 9(10), p. 104006.
- Sultan B., Roudier P., Quirion P., Alhassane A., Muller B., Dingkuhn M., Ciais P., Guimberteau M., Traore S., et Baron C. 2013. Assessing climate change impacts on sorghum and millet yields in the Sudanian and Sahelian savannas of West Africa. *Environmental Research Letters*. 8(1), p. 014040.
- Takagi H., Abe A., Yoshida K., Kosugi S., Natsume S., Mitsuoka C., Uemura A., Utsushi H., Tamiru M., Takuno S., Innan H., Cano L.M., Kamoun S., et Terauchi R. 2013. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*. 74(1), p. 174-183.
- TOSTAIN S. 1998. LE MIL, UNE LONGUE HISTOIRE : HYPOTHÈSES SUR SA DOMESTICATION ET SES MIGRATIONS. Dans : *Plantes et paysages d'Afrique: une histoire à explorer*. KARTHALA et CRA, p. 461-490.
- Tracy S.R., Nagel K.A., Postma J.A., Fassbender H., Wasson A., et Watt M. 2020. Crop Improvement from Phenotyping Roots: Highlights Reveal Expanding Opportunities. *Trends in Plant Science*. 25(1), p. 105-118.
- Valencia L.J.H. 2008. *Etude des bases moléculaires de l'agrégation des sols par des exopolysaccharides bactériens*. Grenoble : Université Joseph FOURRIER, Grenoble 1
- Walker T.S., Bais H.P., Grotewold E., et Vivanco J.M. 2003. Root Exudation and Rhizosphere Biology: Fig. 1. *Plant Physiology*. 132(1), p. 44-51.

Annexes

Liste des annexes

Annexe i : Pourcentage de germination des graines de chaque lignée par répétition	59
Annexe ii : Pourcentage de germination globale de chaque lignée	60
Annexe iii : Mise en place du phénotypage.....	61
Annexe iv : Marqueurs microsatellites	62
Annexe v : Paramètres de la fonction R runQTLseq Analysis	63
Annexe vi : Script de simulation de l'impact de la variation de QTL effects sur la précision de détection de la position de QTL	64
Annexe vii : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect	69
Annexe viii : Script pour la variation du nombre d'individus d'une population en fonction de QTL effect.....	72
Annexe ix : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour individu =100.....	86
Annexe x : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 500	88
Annexe xi : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 100.....	88
Annexe xii : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 500	90
Annexe xiii : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 1000.....	91
Annexe xiv : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 2000.....	92
Annexe xv : Résultats du contrôle de la qualité des produits PCR.....	93
Annexe xvi : Phénotypes parentaux	94
Annexe xvii : Sélection du modèle ANOVA par « backfor selection »	95

Annexe i : Pourcentage de germination des graines de chaque lignée par répétition

Run	Génotype	Nbr graines	Graines germées	% germination
A	line 2	24	23	95.8
A	line 124	25	18	72
A	line125	25	2	8
A	line126	25	0	0
A	line130	25	25	100
A	line131	24	6	25
A	line132	25	9	36
A	line135	25	2	8
A	line138	25	0	0
A	line180	25	5	20
A	line249	25	0	0
A	Line337	25	4	16
B	line 2	25	22	88
B	line 124	25	23	92
B	line125	25	20	80
B	line126	25	20	80
B	line130	25	18	72
B	line131	25	25	100
B	line132	25	24	96
B	line135	24	20	83.3
B	line138	25	20	80
B	line180	25	23	92
B	line249	25	23	92
B	Line337	25	22	88
C	line 2	15	14	93.3
C	line 124	15	14	93.3
C	line125	15	14	93.3
C	line126	15	8	53.3
C	line130	15	14	93.3
C	line131	15	15	100
C	line132	15	15	100
C	line135	15	11	73.3
C	line138	15	7	46.6
C	line180	15	14	93.3
C	line249	15	3	20
C	Line337	15	8	53.3

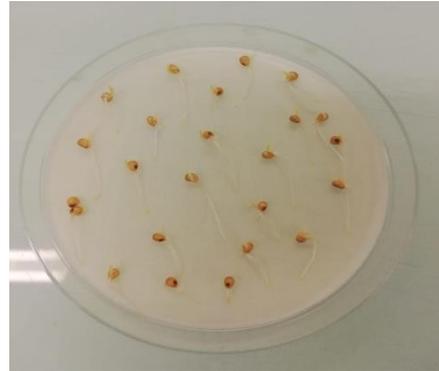
Annexe ii : Pourcentage de germination globale de chaque lignée

<i>Lignées</i>	<i>% Germination par lignée</i>
<i>Line2</i>	92.2
<i>Line124</i>	84.6
<i>Line125</i>	55.4
<i>Line126</i>	43.1
<i>Line130</i>	87.7
<i>Line131</i>	71.8
<i>Line132</i>	73.8
<i>Line135</i>	51.5
<i>Line138</i>	41.5
<i>Line180</i>	64.6
<i>Line249</i>	40
<i>Line337</i>	52.3

Annexe iii : Mise en place du phénotypage



Pré-germination des graines



Graines germées après 24H



Disposition des graines sur le rouleau de papier



Dispositif expérimental



Dispositif expérimental après
4 jours

Annexe iv : Marqueurs microsatellites

Multiplex A		Taille attendue	Multiplex B		Taille attendue
IRD 46	ned	80-110	IRD 25	pet	160-210
PsmP 2227	ned	190-210	PsmP 2214	pet	223-255
PsmP 2249	fam	130-155	PsmP 2219	ned	235-290
PsmP 2247	fam	175-215	PsmP 2220	ned	115-140
PsmP 2237	pet	220-270	PsmP 2206	fam	190-210
IRD 12	vic	115-150	PsmP 2202	vic	140-165

Annexe v : Paramètres de la fonction R *runQTLseq Analysis*

Les paramètres de la fonction R *runQTLseq Analysis* sont :

- les données importées grâce à la fonction *ImportFromGATK* ;
- la taille de la fenêtre (en paires de base) mettant entre parenthèse chaque SNP pour lequel il faut calculer la statistique: 300 ;
- la structure de la population : F2 ;
- le nombre d'individus dans chaque ensemble simulé: c(50,50) ;
- une profondeur de lecture pour laquelle il est possible de reproduire les appels de l'indice SNP : 70:120 ;
- le nombre de répliquions bootstrap: 1000 ;
- un filtre d'indice SNP de 0,1 qui permet de supprimer les SNP qui ne sont pas significatifs.

On obtient ainsi un dataframe contenant des colonnes CHROM (Chromosome) et POS (Position), la profondeur de lecture (DP) ainsi que les profondeurs des allèles de référence (AD_REF) et alternatifs (AD_ALT), le SNPindex pour chaque bulk. La fréquence allélique de référence (REF_FRQ) et le deltaSNP pour chaque marqueur SNP. Toutes les colonnes associées aux SNP possèdent l'extension .HIGH ou .LOW pour chaque bulk. Les seuils des index delta SNP correspondent aux intervalles de confiance demandés ainsi qu'à la profondeur lissée tricube à chaque SNP.

Les positions de QTL estimées en fonction du deltaSNP, PdSNP qui correspond à la moyenne des deltaSNP calculés et PTdSNP qui correspond à la moyenne de tricubeDeltaSNP sont ensuite calculées et stockées dans la matrice QTLPos.

Annexe vi : Script de simulation de l'impact de la variation de QTL effects sur la précision de détection de la position de QTL

Paramètres :

```
nbchr<-1000 #nombre de chromosomes dans un nombre d'individus diploïdes (nbchr/2)
```

```
loci<-10000 #nombre de marqueurs SNP
```

```
nrep<- 100 #nombre de répétition pour 1 effet qtl
```

```
qtleffectrange<- seq(0, 2, by=0.5) #différentes valeurs de qtleffectrange
```

```
nbsim<-length(qtleffectrange) # nbre de simulation
```

Création de la matrice QTLpos afin de stocker les positions de QTL estimées par les différentes méthodes de calcul ("P0","PdSNP","PTdSNP","PEDm","PED100_4") ainsi que les variations de qtleffect

```
test<-6
```

```
QTLpos<-matrix(ncol = test, nrow = nbsim*nrep)
```

```
colnames(QTLpos) = c("P0","PdSNP","PTdSNP","PEDm","PED100_4","qtleffect")
```

```
QTLpos<-as.data.frame(QTLpos)
```

```
for (n in 1:nbsim) { #début de la boucle qui fait la simulation pour chaque valeur de qtleffect
```

```
for (p in 1:nrep) { #début de la boucle qui fait la simulation pour un nombre de répétition pour une valeur de qtleffect
```

```
q<-rpois(nbchr,0.90) #simulation d'un crossing-over pour chacun des chromosomes en suivant une loi de Poisson
```

Création des matrices génotypique et phénotypique :

Matrice génotypique : les chromosomes sont en colonne et les loci en ligne. On définit une boucle pour le remplissage de la matrice. Pour chaque chromosome, la position où a lieu le crossing-over est défini aléatoirement dans l'intervalle 1-10000 défini par le nombre de SNP ; si $q > 0$ il y a eu crossing over sinon mettre 1 dans la matrice.

Matrice Phénotypique : le caractère étudié suit une loi normale (moyenne=0, sd=1) pour (nbchr/2) individus diploïdes.

```
geno<-matrix(ncol = nbchr,nrow = loci, dimnames =list(c(1:loci), c(1:nbchr)))
```

```
for (j in 1:nbchr) {
```

```
allele <- ifelse(runif(1) < 0.5, 0, 1)
```

```
geno[1:loci,j]<-allele
```

```
if (q[j]>0) {
```

```
vec<-c(sort(sample(loci, q[j])),loci)
```

```

for (k in 1 : (length(vec)-1)) {
  allele<-ifelse(allele==1,0,1)
  geno[vec[k]:vec[k+1],j]<-allele
}
}
}
pheno<-rnorm(nbchr/2,0,1)
phenoqtl<-pheno

```

Simulation des positions de QTL et remplissage de la matrice QTLPos

```

qtleffect<-qtleffectrange[n] #attribue les différentes valeurs de contenu de qtleffectrange à chaque
simulation
QTLpos$qtleffect[(n-1)*nrep+p]<-qtleffect #stocke une valeur de qtleffect pour nrep simulation dans
la matrice QTLpos colonne QTLeffect
qtlpos<-sample(loci,1) #choisi aléatoirement la position initiale du QTL
QTLpos$P0[(n-1)*nrep+p]<-qtlpos # stocke la valeur de la position de QTL simulée pour nrep
simulation dans la matrice QTLpos
print(QTLpos) # afficher la matrice QTLpos après chaque simulation

```

Remplissage de la matrice phénotypique avec les valeurs des effets additifs de QTL

```

for (i in 1: (nbchr/2)) {
  phenoqtl[i]<-pheno[i]+geno[qtlpos,2*i-1]*qtleffect +geno[qtlpos,2*i]*qtleffect
}

```

Tri des individus en fonction du phénotype puis sélection des 10% d'individus extrêmes pour la formation des bulks sur la base des données génotypiques

```

individuals<-c(1: (nbchr/2))
bsa<-data.frame(individuals, phenoqtl)
bsa<-bsa[order(bsa$phenoqtl),]
lowpheno<-bsa$individuals[1:(0.10*length(bsa$individuals))]
poslow<-sort(c(2*lowpheno,2*lowpheno-1))
lowbulk<-geno[,poslow]
highpheno<-bsa$individuals[((0.90*length(bsa$individuals))+1):length(bsa$individuals)]
poshigh<-sort(c(2*highpheno,2*highpheno-1))
highbulk<-geno[,poshigh]

```

Organisation les données et ajout des noms aux profondeurs de référence et d'allèles alternatifs par bulk (AD_REF et AD_ALT dans .LOW et .HIGH) et à la profondeur totale par bulk (DP.LOW et DP.HIGH)

```
AD_ALT.LOW_i<-apply(lowbulk,1,function(x)sum(x == 1))
AD_REF.LOW_i<-apply(lowbulk,1,function(x)sum(x == 0))
DP.LOW<-AD_ALT.LOW_i+AD_REF.LOW_i
AD_ALT.HIGH_i<-apply(highbulk,1,function(x)sum(x == 1))
AD_REF.HIGH_i<-apply(highbulk,1,function(x)sum(x == 0))
DP.HIGH<-AD_ALT.HIGH_i+AD_REF.HIGH_i
simdata<-data.frame("chr",1:loci,AD_ALT.LOW_i,AD_REF.LOW_i,DP.LOW,AD_ALT.HIGH_i,AD_
REF.HIGH_i,DP.HIGH)
colnames(simdata)[1]<-"CHROM"
colnames(simdata)[2]<-"POS"
```

Ajout de bruit de séquençage à la profondeur allélique des données

```
simdata$AD_ALT.LOW<-0
simdata$AD_REF.LOW<-0
simdata$AD_ALT.HIGH<-0
simdata$AD_REF.HIGH<-0
depth<-100 #bruit de séquençage
for (i in 1:nrow(simdata)){
  simdata$AD_ALT.LOW[i]<-
  rbinom(1,depth,(simdata$AD_ALT.LOW_i[i]/(simdata$AD_ALT.LOW_i[i]+simdata$AD_REF.LOW_
i[i])))
  simdata$AD_REF.LOW[i]<-
  rbinom(1,depth,(simdata$AD_REF.LOW_i[i]/(simdata$AD_ALT.LOW_i[i]+simdata$AD_REF.LOW_
i[i])))
  simdata$AD_ALT.HIGH[i]<-
  rbinom(1,depth,(simdata$AD_ALT.HIGH_i[i]/(simdata$AD_ALT.HIGH_i[i]+simdata$AD_REF.HIG
H_i[i])))
  simdata$AD_REF.HIGH[i]<-
  rbinom(1,depth,(simdata$AD_REF.HIGH_i[i]/(simdata$AD_ALT.HIGH_i[i]+simdata$AD_REF.HIG
H_i[i])))
}
```

Enlever les valeurs initiales AD (colonnes i) et calculer la nouvelle valeur de profondeur par volume (DP)

```
simdata<-simdata[,-c(3,4,6,7)]
simdata$DP.LOW<-simdata$AD_ALT.LOW+simdata$AD_REF.LOW
simdata$DP.HIGH<-simdata$AD_ALT.HIGH+simdata$AD_REF.HIGH
```

Analyse QTL en utilisant la méthode statistique basée sur le DeltaSNP

Cette méthode a été proposée par (Takagi *et al.*, 2013) et déployé dans le package R QTLseqr par (Mansfeld et Grumet, 2018). Il est nécessaire d'importer les données d'un fichier externe pour utiliser le package QTLseqr. L'importation des données ajoute des informations aux données originales : SNPindex pour chaque masse, REF_FRQ et deltaSNP pour chaque marqueur.

```
directory<-getwd()
write.csv(simdata, file = paste(directory, "/simdata.csv", sep=""), row.names = FALSE)
simdata<-importFromTable(file = paste(directory, "/simdata.csv", sep=""),
highBulk = "HIGH",
lowBulk = "LOW",
chromList = c("chr"))
head(simdata)
simdata<-runQTLseqAnalysis(simdata,
windowSize = 300,
popStruc = "F2",
bulkSize = c(50,50),
replications = 1000,
depth = 70:120,
filter = 0.1,
intervals = c(90, 95, 99))
QTLpos$PdSNP[(n-1)*nrep+p]<-
mean(simdata[(abs(simdata$deltaSNP))==max(abs(simdata$deltaSNP)),c(2)])
QTLpos$PTdSNP[(n-1)*nrep+p]<-
mean(simdata[(abs(simdata$tricubeDeltaSNP))==max(abs(simdata$tricubeDeltaSNP)),c(2)])
```

Analyse QTL en utilisant des statistiques basées sur la distance euclidienne (ED)

EDm : Distance euclidienne entre les fréquences des allèles dans les bulks pour chaque position de marqueur

ED100: somme des EDm de 100 marqueurs consécutifs (Zhang et al 2019)

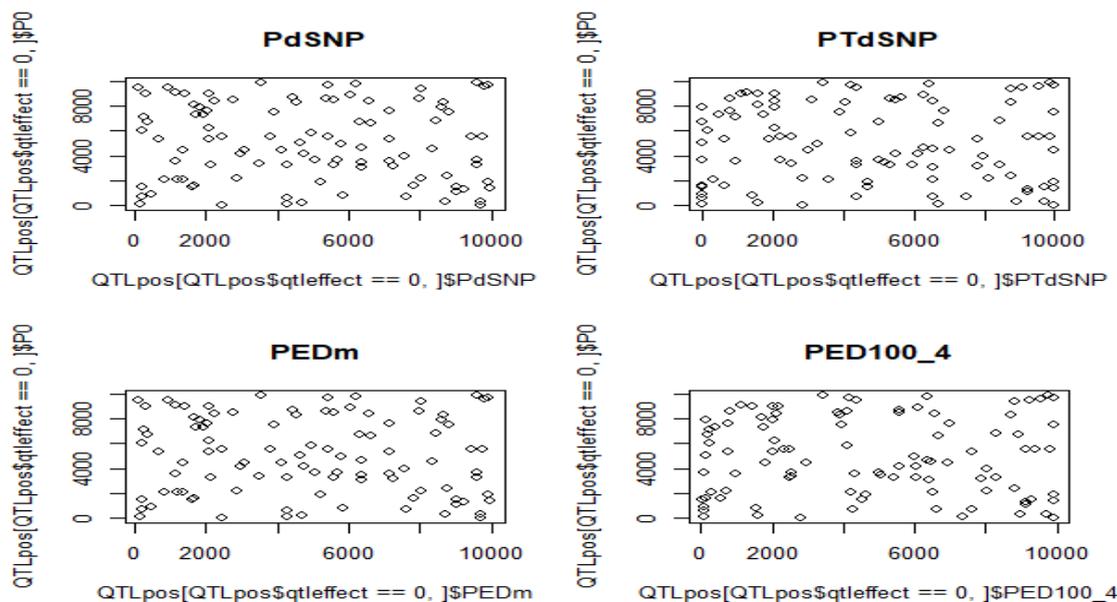
ED4 : ED100 (ou EDm) exposant 4 (pour augmenter le signe des valeurs ED et réduire le bruit des valeurs ED basses)

```
for (i in 1:nrow(simdata)){
  fqlow<-
  c((simdata$AD_ALT.LOW[[i]]/simdata$DP.LOW[[i]],(simdata$AD_REF.LOW[[i]]/simdata$DP.L
  OW[[i]]))
  fqhigh<-
  c((simdata$AD_ALT.HIGH[[i]]/simdata$DP.HIGH[[i]],(simdata$AD_REF.HIGH[[i]]/simdata$DP.
  HIGH[[i]]))
  simdata$EDm[i]<-dist(rbind(fqlow, fqhigh))
  simdata$ED4<-(simdata$EDm)^4
}
QTLpos$PEDm[(n-1)*nrep+p]<-mean(simdata[simdata$EDm==max(abs(simdata$EDm)),c(2)])
simdata$ED100<-0
range<-100
nsnp<-nrow(simdata)
for (i in 1:(range/2)){
  simdata$ED100[i]<-sum(simdata$EDm[1:(range)])
}
for (i in (range/2):((nsnp-(range/2)))) {
  simdata$ED100[i]<-sum(simdata$EDm[(i-(range/2)):(i+(range/2))])
}
for (i in (nsnp-(range/2)):nsnp){
  simdata$ED100[i]<-sum(simdata$EDm[(nsnp-(range/2)):nsnp])
}
simdata$ED100_4<-(simdata$ED100)^4
QTLpos$PED100_4[(n-1)*nrep+p]<-
mean(simdata[simdata$ED100_4==max(abs(simdata$ED100_4)),c(2)])

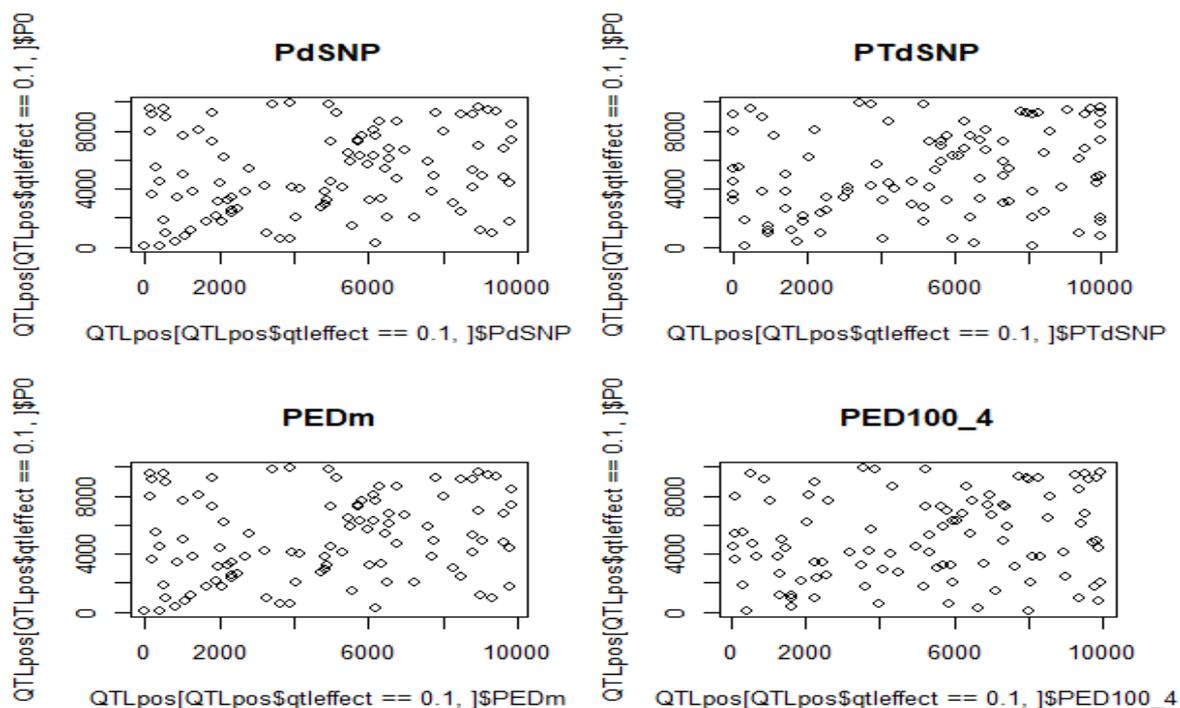
} #fin de la boucle qui fait la simulation pour un nombre de répétition pour une valeur de qtleffect } #
fin de la boucle pour toutes les valeurs de qtleffect
print(QTLpos)
```

Annexe vii : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect

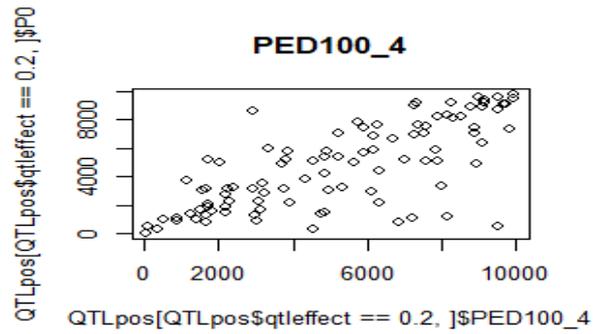
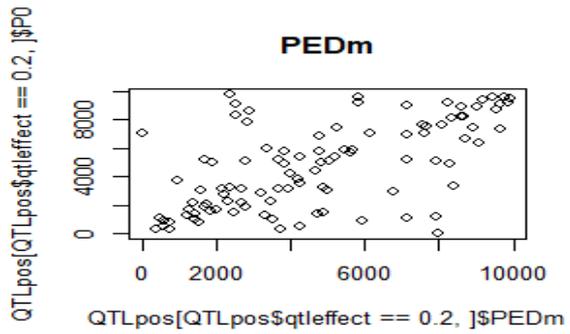
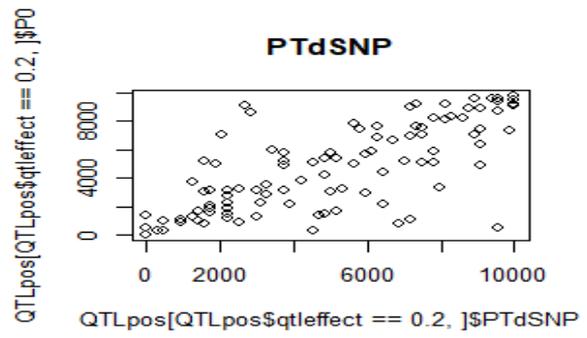
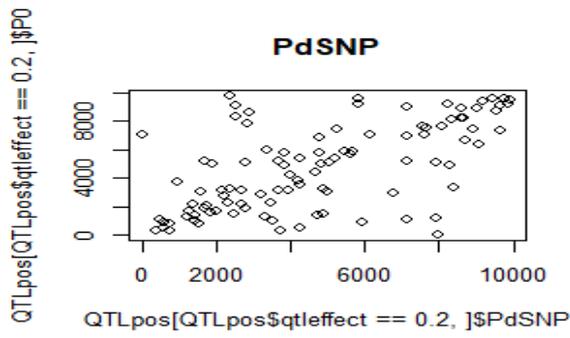
QTLeffect=0



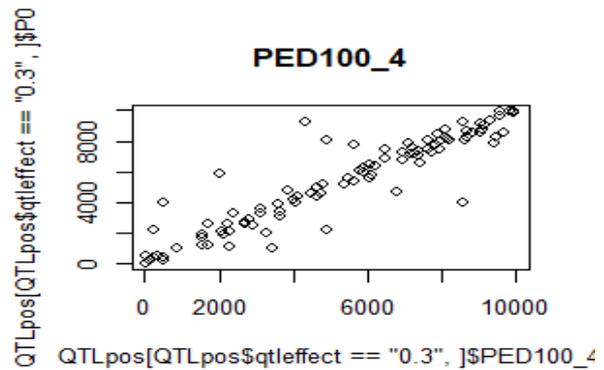
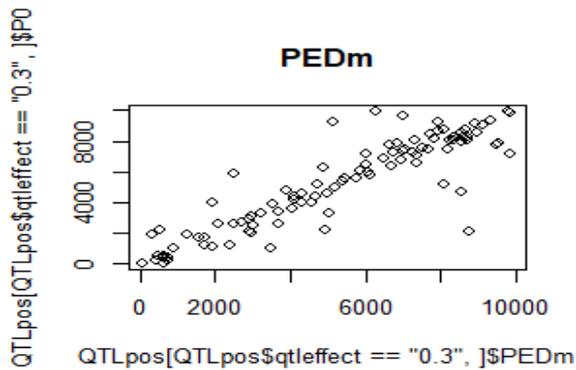
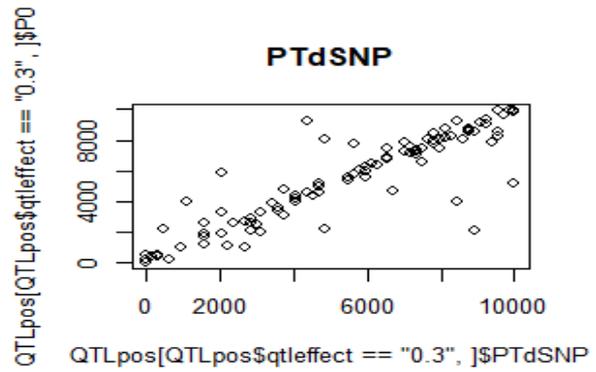
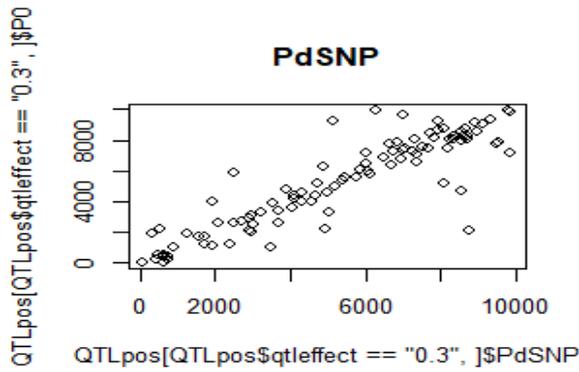
Qtl effect=0,1



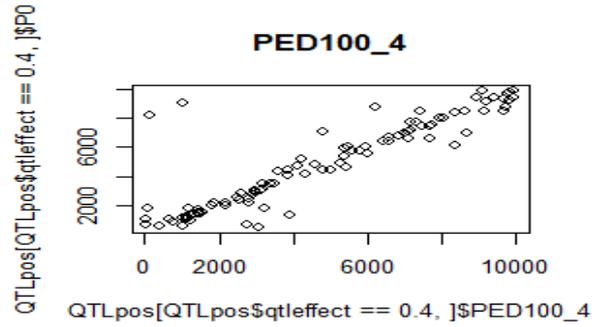
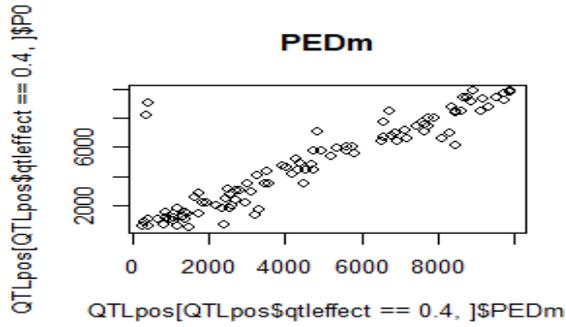
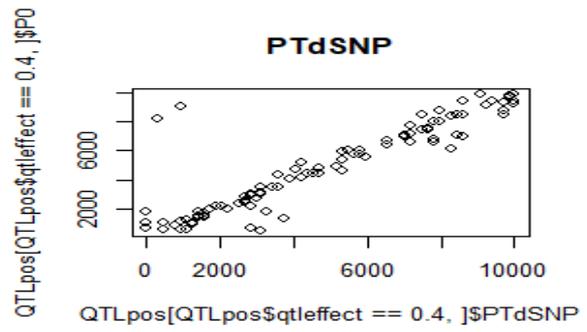
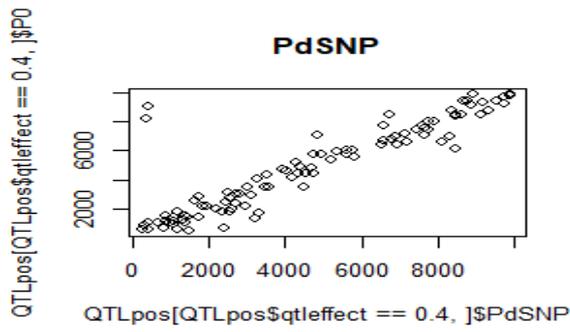
Qtl effect=0,2



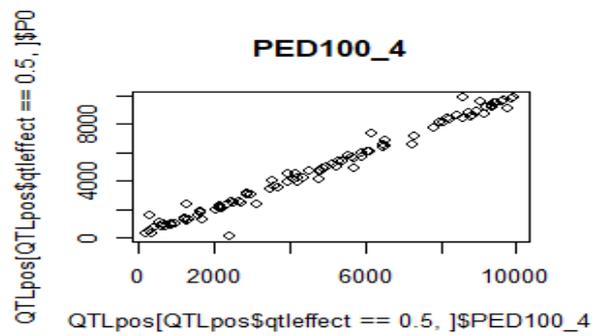
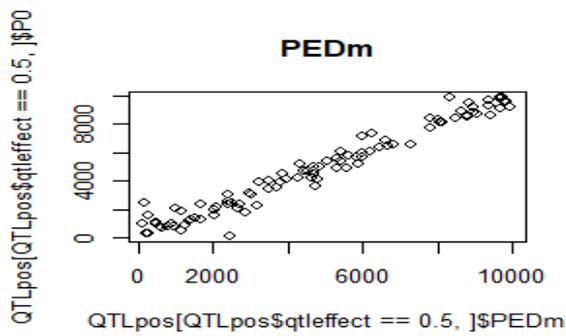
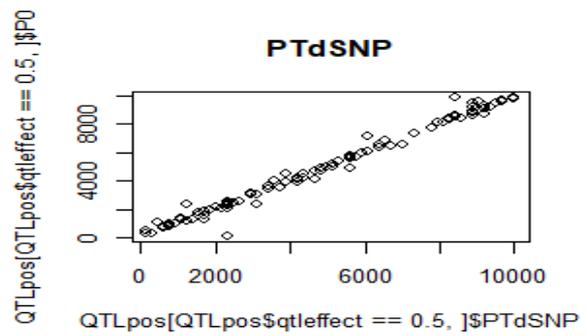
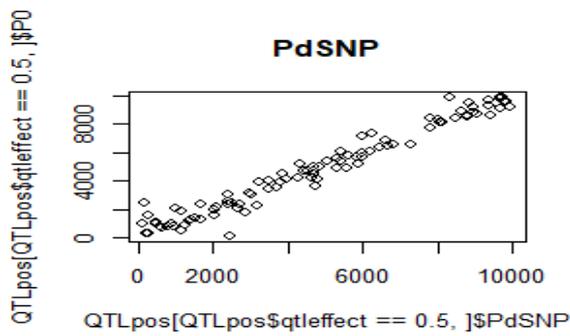
QTL effect= 0,3



Qtl effect=0,4



Qtl effect=0,5



Annexe viii : Script pour la variation du nombre d'individus d'une population en fonction de QTL effect

Paramètres :

```
qtlectrange<- seq(0, 2, by=0.5) #différentes valeurs de qtlectrange
individualsrange<-c(100,500,1000,2000) # taille de la population
nrep<- 100 #nbre de répétition pour 1 effet qtl
nbsim<-length(qtlectrange) # nbre de simulation
nbind<-length(individualsrange) #nbre d'individus
loci<-10000 #nombre de marqueurs SNP
```

Création de la matrice QTLpos afin de stocker les positions de QTL estimées par les différentes méthodes de calcul ("P0","PdSNP","PTdSNP","PEDm","PED100_4") ainsi que les variations de qtlect

```
test<-8
QTLpos<-matrix(ncol = test, nrow = nbsim*nrep*nbind)
colnames(QTLpos) = c("chr","ind","qtlect","P0","PdSNP","PTdSNP","PEDm","PED100_4")
QTLpos<-as.data.frame(QTLpos)
for (ind in 1:nbind){ #début de la boucle qui fait la simulation pour chaque effectif d'individus
  for (n in 1:nbsim) { #début de la boucle qui fait la simulation pour chaque valeur de qtlect
    for (p in 1:nrep) { #début de la boucle qui fait la simulation pour un nombre de répétition pour une
      valeur de qtlect
        ind_group <-individualsrange[ind] #ind_group (identifiant de groupe) prend une valeur dans
        individualsrange pour chaque simulation
        nbchr<-ind_group*2 #nombre de chromosomes
        q<-rpois(nbchr,0.90) #simulation d' un crossing-over pour chacun des chromosomes en suivant une
        loi de Poisson
```

Création des matrices génotypique et phénotypique :

Matrice génotypique : les chromosomes sont en colonne et les loci en ligne. On définit une boucle pour le remplissage de la matrice. Pour chaque chromosome, la position où a lieu le crossing-over est défini aléatoirement dans l'intervalle 1-10000 défini par le nombre de SNP ; si $q > 0$ il y a eu crossing over sinon mettre 1 dans la matrice.

Matrice Phénotypique : le caractère étudié suit une loi normale (moyenne=0, sd=1) pour (nbchr/2) individus diploïdes.

```
geno<-matrix(ncol = nbchr,nrow = loci, dimnames =list(c(1:loci), c(1:nbchr)))
for (j in 1:nbchr) {
```

```

allele <- ifelse(runif(1) < 0.5, 0, 1)
geno[1:loci,j]<-allele
if (q[j]>0) {
  vec<-c(sort(sample(loci, q[j])),loci)
  for (k in 1 : (length(vec)-1)) {
    allele<-ifelse(allele==1,0,1)
    geno[vec[k]:vec[k+1],j]<-allele
  }
}
pheno<-rnorm(nbchr/2,0,1)
phenoqtl<-pheno

```

Simulation des positions de QTL et remplissage de la matrice QTLPos

```

qtlexfect<-qtlexfectrange[n]
QTLpos$qtlexfect[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-qtlexfect #stocke une valeur de qtlexfect pour
nrep simulation dans la matrice QTLpos colonne QTLexfect
qtlpos<-sample(loci,1) #choisi aléatoirement la position initiale du QTL
QTLpos$P0[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-qtlpos # stocke la valeur de la position de QTL
simulée pour nrep simulation dans la matrice QTLpos
QTLpos$ind[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-ind_group # stocke la taille des individus pour nrep
simulation dans la matrice QTLpos
QTLpos$chr[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-nbchr # stocke la valeur du nombre de
chromosome pour nrep simulation dans la matrice QTLpos
print(QTLpos) # afficher la matrice QTLpos après chaque simulation

```

Remplissage de la matrice phénotypique avec les valeurs des effets additifs de QTL

```

for (i in 1: (nbchr/2)) {
  phenoqtl[i]<-pheno[i]+geno[qtlpos,2*i-1]*qtlexfect +geno[qtlpos,2*i]*qtlexfect
}

```

Tri des individus en fonction du phénotype puis sélection des 10% d'individus extrêmes pour la formation des bulks sur la base des données génotypiques

```

individuals<-c(1:ind_group)
bsa<-data.frame(individuals, phenoqtl)

```

```

bsa<-bsa[order(bsa$phenoqtl),]
lowpheno<-bsa$individuals[1:(0.10*length(bsa$individuals))]
poslow<-sort(c(2*lowpheno,2*lowpheno-1))
lowbulk<-geno[,poslow]
highpheno<-bsa$individuals[((0.90*length(bsa$individuals))+1):length(bsa$individuals)]
poshigh<-sort(c(2*highpheno,2*highpheno-1))
highbulk<-geno[,poshigh]

```

Organisation les données et ajout des noms aux profondeurs de référence et d'allèles alternatifs par bulk (AD_REF et AD_ALT dans .LOW et .HIGH) et à la profondeur totale par bulk (DP.LOW et DP.HIGH)

```

AD_ALT.LOW_i<-apply(lowbulk,1,function(x)sum(x == 1))
AD_REF.LOW_i<-apply(lowbulk,1,function(x)sum(x == 0))
DP.LOW<-AD_ALT.LOW_i+AD_REF.LOW_i
AD_ALT.HIGH_i<-apply(highbulk,1,function(x)sum(x == 1))
AD_REF.HIGH_i<-apply(highbulk,1,function(x)sum(x == 0))
DP.HIGH<-AD_ALT.HIGH_i+AD_REF.HIGH_i
simdata<-data.frame("chr",1:loci,AD_ALT.LOW_i,AD_REF.LOW_i,DP.LOW,AD_ALT.HIGH_i,AD_REF.HIGH_i,DP.HIGH)
colnames(simdata)[1]<-"CHROM"
colnames(simdata)[2]<-"POS"

```

Ajout de bruit de séquençage à la profondeur allélique des données

```

simdata$AD_ALT.LOW<-0
simdata$AD_REF.LOW<-0
simdata$AD_ALT.HIGH<-0
simdata$AD_REF.HIGH<-0
depth<-100 #bruit de séquençage
for (i in 1:nrow(simdata)){
  simdata$AD_ALT.LOW[i]<-
  rbinom(1,depth,(simdata$AD_ALT.LOW_i[i]/(simdata$AD_ALT.LOW_i[i]+simdata$AD_REF.LOW_i[i])))
  simdata$AD_REF.LOW[i]<-
  rbinom(1,depth,(simdata$AD_REF.LOW_i[i]/(simdata$AD_ALT.LOW_i[i]+simdata$AD_REF.LOW_i[i])))
  simdata$AD_ALT.HIGH[i]<-

```

```

rbinom(1,depth,(simdata$AD_ALT.HIGH_i[i]/(simdata$AD_ALT.HIGH_i[i]+simdata$AD_REF.HIGH_i[i])))
simdata$AD_REF.HIGH[i]<-
rbinom(1,depth,(simdata$AD_REF.HIGH_i[i]/(simdata$AD_ALT.HIGH_i[i]+simdata$AD_REF.HIGH_i[i])))
}

```

Enlever les valeurs initiales AD (colonnes i) et calculer la nouvelle valeur de profondeur par volume (DP)

```

simdata<-simdata[,-c(3,4,6,7)]
simdata$DP.LOW<-simdata$AD_ALT.LOW+simdata$AD_REF.LOW
simdata$DP.HIGH<-simdata$AD_ALT.HIGH+simdata$AD_REF.HIGH

```

Analyse QTL en utilisant la méthode statistique basée sur le DeltaSNP proposée par Takagi et al 2013 et déployé dans le package R QTLseqr par Mansfeld et Grumet 2018 pour l'analyse BSA en utilisant les données NGS. Il est nécessaire d'importer les données d'un fichier externe pour utiliser le package QTLseqr. L'importation des données ajoute des informations aux données originales : SNPindex pour chaque masse, REF_FRQ et deltaSNP pour chaque marqueur

```

directory<-getwd()
write.csv(simdata, file = paste(directory, "/simdata.csv", sep=""), row.names = FALSE)
simdata<-importFromTable(file = paste(directory, "/simdata.csv", sep=""),
highBulk = "HIGH",
lowBulk = "LOW",
chromList = c("chr"))
head(simdata)
simdata<-runQTLseqAnalysis(simdata,
windowSize = 300,
popStruc = "F2",
bulkSize = c(50,50),
replications = 1000,
depth = 70:120,
filter = 0.1,
intervals = c(90, 95, 99))
QTLpos$PdSNP[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-
mean(simdata[(abs(simdata$deltaSNP))==max(abs(simdata$deltaSNP)),c(2)])
QTLpos$PTdSNP[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-
mean(simdata[(abs(simdata$tricubeDeltaSNP))==max(abs(simdata$tricubeDeltaSNP)),c(2)])

```

Analyse QTL en utilisant des statistiques basées sur la distance euclidienne (ED)

ED_m : Distance euclidienne entre les fréquences des allèles dans les bulks pour chaque position de marqueur

ED100:somme des ED_m de 100 marqueurs consécutifs

ED4 : ED100 (ou ED_m) exposant 4 (pour augmenter le signe des valeurs ED et réduire le bruit des valeurs ED basses)

```
for (i in 1:nrow(simdata)){
  fqlow<-
  c((simdata$AD_ALT.LOW[[i]]/simdata$DP.LOW[[i]]),(simdata$AD_REF.LOW[[i]]/simdata$DP.L
  OW[[i]]))
  fqhigh<-
  c((simdata$AD_ALT.HIGH[[i]]/simdata$DP.HIGH[[i]]),(simdata$AD_REF.HIGH[[i]]/simdata$DP.
  HIGH[[i]]))
  simdata$EDm[i]<-dist(rbind(fqlow, fqhigh))
  simdata$ED4<-(simdata$EDm)^4
}
QTLpos$PEDm[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-
mean(simdata[simdata$EDm==max(abs(simdata$EDm)),c(2)])
simdata$ED100<-0
range<-100
nsnp<-nrow(simdata)
for (i in 1:(range/2)){
  simdata$ED100[i]<-sum(simdata$EDm[1:(range)])
}
for (i in (range/2):((nsnp-(range/2)))) {
  simdata$ED100[i]<-sum(simdata$EDm[(i-(range/2)):(i+(range/2))])
}
for (i in (nsnp-(range/2)):nsnp){
  simdata$ED100[i]<-sum(simdata$EDm[(nsnp-(range/2)):nsnp])
}
simdata$ED100_4<-(simdata$ED100)^4
QTLpos$PED100_4[(ind-1)*nbsim*nrep+(n-1)*nrep+p]<-
mean(simdata[simdata$ED100_4==max(abs(simdata$ED100_4)),c(2)])
} #fin de la boucle qui fait la simul pour un nbr de repeton pour une valeur de qtleffect
} # fin de la boucle pour toutes les valeur de qtleffect
```

```
} #fin de la boucle qui fait la simul pour tous les effectifs d'ind pour chaque qtleffect par nrep  
print(QTLpos)
```

Sélection des individus pour ind=100

```
QTLpos_ind100_0<-QTLpos[QTLpos$ind=="100" & QTLpos$qtleffect=="0" ,]  
QTLpos_ind100_0.5<-QTLpos[QTLpos$ind=="100" & QTLpos$qtleffect=="0.5" ,]  
QTLpos_ind100_1<-QTLpos[QTLpos$ind=="100" & QTLpos$qtleffect=="1" ,]  
QTLpos_ind100_1.5<-QTLpos[QTLpos$ind=="100" & QTLpos$qtleffect=="1.5" ,]  
QTLpos_ind100_2<-QTLpos[QTLpos$ind=="100" & QTLpos$qtleffect=="2" ,]
```

Plot

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0

par (mfrow=c(2,2))

```
plot(QTLpos_ind100_0$P0~QTLpos_ind100_0$PdSNP,main="PdSNP")  
plot(QTLpos_ind100_0$P0~QTLpos_ind100_0$PTdSNP,main="PTdSNP")  
plot(QTLpos_ind100_0$P0~QTLpos_ind100_0$PEDm,main="PEDm")  
plot(QTLpos_ind100_0$P0~QTLpos_ind100_0$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

par (mfrow=c(2,2))

```
plot(QTLpos_ind100_0.5$P0~QTLpos_ind100_0.5$PdSNP,main="PdSNP")  
plot(QTLpos_ind100_0.5$P0~QTLpos_ind100_0.5$PTdSNP,main="PTdSNP")  
plot(QTLpos_ind100_0.5$P0~QTLpos_ind100_0.5$PEDm,main="PEDm")  
plot(QTLpos_ind100_0.5$P0~QTLpos_ind100_0.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

par (mfrow=c(2,2))

```
plot(QTLpos_ind100_1$P0~QTLpos_ind100_1$PdSNP,main="PdSNP")  
plot(QTLpos_ind100_1$P0~QTLpos_ind100_1$PTdSNP,main="PTdSNP")  
plot(QTLpos_ind100_1$P0~QTLpos_ind100_1$PEDm,main="PEDm")  
plot(QTLpos_ind100_1$P0~QTLpos_ind100_1$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

par (mfrow=c(2,2))

```
plot(QTLpos_ind100_1.5$P0~QTLpos_ind100_1.5$PdSNP,main="PdSNP")
```

```
plot(QTLpos_ind100_1.5$P0~QTLpos_ind100_1.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind100_1.5$P0~QTLpos_ind100_1.5$PEDm,main="PEDm")
plot(QTLpos_ind100_1.5$P0~QTLpos_ind100_1.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

par (mfrow=c(2,2))

```
plot(QTLpos_ind100_2$P0~QTLpos_ind100_2$PdSNP,main="PdSNP")
plot(QTLpos_ind100_2$P0~QTLpos_ind100_2$PTdSNP,main="PTdSNP")
plot(QTLpos_ind100_2$P0~QTLpos_ind100_2$PEDm,main="PEDm")
plot(QTLpos_ind100_2$P0~QTLpos_ind100_2$PED100_4,main="PED100_4")
```

Sélection des individus pour ind=500

```
QTLpos_ind500_0<-QTLpos[QTLpos$ind=="500" & QTLpos$qtlexact=="0" ,]
QTLpos_ind500_0.5<-QTLpos[QTLpos$ind=="500" & QTLpos$qtlexact=="0.5" ,]
QTLpos_ind500_1<-QTLpos[QTLpos$ind=="500" & QTLpos$qtlexact=="1" ,]
QTLpos_ind500_1.5<-QTLpos[QTLpos$ind=="500" & QTLpos$qtlexact=="1.5" ,]
QTLpos_ind500_2<-QTLpos[QTLpos$ind=="500" & QTLpos$qtlexact=="2" ,]
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0

par (mfrow=c(2,2))

```
plot(QTLpos_ind500_0$P0~QTLpos_ind500_0$PdSNP,main="PdSNP")
plot(QTLpos_ind500_0$P0~QTLpos_ind500_0$PTdSNP,main="PTdSNP")
plot(QTLpos_ind500_0$P0~QTLpos_ind500_0$PEDm,main="PEDm")
plot(QTLpos_ind500_0$P0~QTLpos_ind500_0$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

par (mfrow=c(2,2))

```
plot(QTLpos_ind500_0.5$P0~QTLpos_ind500_0.5$PdSNP,main="PdSNP")
plot(QTLpos_ind500_0.5$P0~QTLpos_ind500_0.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind500_0.5$P0~QTLpos_ind500_0.5$PEDm,main="PEDm")
plot(QTLpos_ind500_0.5$P0~QTLpos_ind500_0.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

par (mfrow=c(2,2))

```
plot(QTLpos_ind500_1$P0~QTLpos_ind500_1$PdSNP,main="PdSNP")
```

```
plot(QTLpos_ind500_1$P0~QTLpos_ind500_1$PTdSNP,main="PTdSNP")
plot(QTLpos_ind500_1$P0~QTLpos_ind500_1$PEDm,main="PEDm")
plot(QTLpos_ind500_1$P0~QTLpos_ind500_1$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

```
par (mfrow=c(2,2))
plot(QTLpos_ind500_1.5$P0~QTLpos_ind500_1.5$PdSNP,main="PdSNP")
plot(QTLpos_ind500_1.5$P0~QTLpos_ind500_1.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind500_1.5$P0~QTLpos_ind500_1.5$PEDm,main="PEDm")
plot(QTLpos_ind500_1.5$P0~QTLpos_ind500_1.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

```
par (mfrow=c(2,2))
plot(QTLpos_ind500_2$P0~QTLpos_ind500_2$PdSNP,main="PdSNP")
plot(QTLpos_ind500_2$P0~QTLpos_ind500_2$PTdSNP,main="PTdSNP")
plot(QTLpos_ind500_2$P0~QTLpos_ind500_2$PEDm,main="PEDm")
plot(QTLpos_ind500_2$P0~QTLpos_ind500_2$PED100_4,main="PED100_4")
```

Sélection des individus pour ind=1000

```
QTLpos_ind1000_0<-QTLpos[QTLpos$ind=="1000" & QTLpos$qtleffect=="0" ,]
QTLpos_ind1000_0.5<-QTLpos[QTLpos$ind=="1000" & QTLpos$qtleffect=="0.5" ,]
QTLpos_ind1000_1<-QTLpos[QTLpos$ind=="1000" & QTLpos$qtleffect=="1" ,]
QTLpos_ind1000_1.5<-QTLpos[QTLpos$ind=="1000" & QTLpos$qtleffect=="1.5" ,]
QTLpos_ind1000_2<-QTLpos[QTLpos$ind=="1000" & QTLpos$qtleffect=="2" ,]
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0

```
par (mfrow=c(2,2))
plot(QTLpos_ind1000_0$P0~QTLpos_ind1000_0$PdSNP,main="PdSNP")
plot(QTLpos_ind1000_0$P0~QTLpos_ind1000_0$PTdSNP,main="PTdSNP")
plot(QTLpos_ind1000_0$P0~QTLpos_ind1000_0$PEDm,main="PEDm")
plot(QTLpos_ind1000_0$P0~QTLpos_ind1000_0$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

```
par (mfrow=c(2,2))
plot(QTLpos_ind1000_0.5$P0~QTLpos_ind1000_0.5$PdSNP,main="PdSNP")
```

```
plot(QTLpos_ind1000_0.5$P0~QTLpos_ind1000_0.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind1000_0.5$P0~QTLpos_ind1000_0.5$PEDm,main="PEDm")
plot(QTLpos_ind1000_0.5$P0~QTLpos_ind1000_0.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

```
par (mfrow=c(2,2))
plot(QTLpos_ind1000_1$P0~QTLpos_ind1000_1$PdSNP,main="PdSNP")
plot(QTLpos_ind1000_1$P0~QTLpos_ind1000_1$PTdSNP,main="PTdSNP")
plot(QTLpos_ind1000_1$P0~QTLpos_ind1000_1$PEDm,main="PEDm")
plot(QTLpos_ind1000_1$P0~QTLpos_ind1000_1$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

```
par (mfrow=c(2,2))
plot(QTLpos_ind1000_1.5$P0~QTLpos_ind1000_1.5$PdSNP,main="PdSNP")
plot(QTLpos_ind1000_1.5$P0~QTLpos_ind1000_1.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind1000_1.5$P0~QTLpos_ind1000_1.5$PEDm,main="PEDm")
plot(QTLpos_ind1000_1.5$P0~QTLpos_ind1000_1.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

```
par (mfrow=c(2,2))
plot(QTLpos_ind1000_2$P0~QTLpos_ind1000_2$PdSNP,main="PdSNP")
plot(QTLpos_ind1000_2$P0~QTLpos_ind1000_2$PTdSNP,main="PTdSNP")
plot(QTLpos_ind1000_2$P0~QTLpos_ind1000_2$PEDm,main="PEDm")
plot(QTLpos_ind1000_2$P0~QTLpos_ind1000_2$PED100_4,main="PED100_4")
```

Sélection des individus pour ind=2000

```
QTLpos_ind2000_0<-QTLpos[QTLpos$ind=="2000" & QTLpos$qtlexact=="0" ,]
QTLpos_ind2000_0.5<-QTLpos[QTLpos$ind=="2000" & QTLpos$qtlexact=="0.5" ,]
QTLpos_ind2000_1<-QTLpos[QTLpos$ind=="2000" & QTLpos$qtlexact=="1" ,]
QTLpos_ind2000_1.5<-QTLpos[QTLpos$ind=="2000" & QTLpos$qtlexact=="1.5" ,]
QTLpos_ind2000_2<-QTLpos[QTLpos$ind=="2000" & QTLpos$qtlexact=="2" ,]
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0

```
par (mfrow=c(2,2))
plot(QTLpos_ind2000_0$P0~QTLpos_ind2000_0$PdSNP,main="PdSNP")
```

```
plot(QTLpos_ind2000_0$P0~QTLpos_ind2000_0$PTdSNP,main="PTdSNP")
plot(QTLpos_ind2000_0$P0~QTLpos_ind2000_0$PEDm,main="PEDm")
plot(QTLpos_ind2000_0$P0~QTLpos_ind2000_0$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

```
par (mfrow=c(2,2))
plot(QTLpos_ind2000_0.5$P0~QTLpos_ind2000_0.5$PdSNP,main="PdSNP")
plot(QTLpos_ind2000_0.5$P0~QTLpos_ind2000_0.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind2000_0.5$P0~QTLpos_ind2000_0.5$PEDm,main="PEDm")
plot(QTLpos_ind2000_0.5$P0~QTLpos_ind2000_0.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

```
par (mfrow=c(2,2))
plot(QTLpos_ind2000_1$P0~QTLpos_ind2000_1$PdSNP,main="PdSNP")
plot(QTLpos_ind2000_1$P0~QTLpos_ind2000_1$PTdSNP,main="PTdSNP")
plot(QTLpos_ind2000_1$P0~QTLpos_ind2000_1$PEDm,main="PEDm")
plot(QTLpos_ind2000_1$P0~QTLpos_ind2000_1$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

```
par (mfrow=c(2,2))
plot(QTLpos_ind2000_1.5$P0~QTLpos_ind2000_1.5$PdSNP,main="PdSNP")
plot(QTLpos_ind2000_1.5$P0~QTLpos_ind2000_1.5$PTdSNP,main="PTdSNP")
plot(QTLpos_ind2000_1.5$P0~QTLpos_ind2000_1.5$PEDm,main="PEDm")
plot(QTLpos_ind2000_1.5$P0~QTLpos_ind2000_1.5$PED100_4,main="PED100_4")
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

```
par (mfrow=c(2,2))
plot(QTLpos_ind2000_2$P0~QTLpos_ind2000_2$PdSNP,main="PdSNP")
plot(QTLpos_ind2000_2$P0~QTLpos_ind2000_2$PTdSNP,main="PTdSNP")
plot(QTLpos_ind2000_2$P0~QTLpos_ind2000_2$PEDm,main="PEDm")
plot(QTLpos_ind2000_2$P0~QTLpos_ind2000_2$PED100_4,main="PED100_4")
```

Calcul des corrélations pour ind=100

pour qtleffect=0

```
cor(QTLpos_ind100_0$P0,QTLpos_ind100_0$PdSNP)
```

$cor(QTLpos_ind100_0\$P0, QTLpos_ind100_0\$PTdSNP)$
 $cor(QTLpos_ind100_0\$P0, QTLpos_ind100_0\$PEDm)$
 $cor(QTLpos_ind100_0\$P0, QTLpos_ind100_0\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

$cor(QTLpos_ind100_0.5\$P0, QTLpos_ind100_0.5\$PdSNP)$
 $cor(QTLpos_ind100_0.5\$P0, QTLpos_ind100_0.5\$PTdSNP)$
 $cor(QTLpos_ind100_0.5\$P0, QTLpos_ind100_0.5\$PEDm)$
 $cor(QTLpos_ind100_0.5\$P0, QTLpos_ind100_0.5\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

$cor(QTLpos_ind100_1\$P0, QTLpos_ind100_1\$PdSNP)$
 $cor(QTLpos_ind100_1\$P0, QTLpos_ind100_1\$PTdSNP)$
 $cor(QTLpos_ind100_1\$P0, QTLpos_ind100_1\$PEDm)$
 $cor(QTLpos_ind100_1\$P0, QTLpos_ind100_1\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

$cor(QTLpos_ind100_1.5\$P0, QTLpos_ind100_1.5\$PdSNP)$
 $cor(QTLpos_ind100_1.5\$P0, QTLpos_ind100_1.5\$PTdSNP)$
 $cor(QTLpos_ind100_1.5\$P0, QTLpos_ind100_1.5\$PEDm)$
 $cor(QTLpos_ind100_1.5\$P0, QTLpos_ind100_1.5\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

$cor(QTLpos_ind100_2\$P0, QTLpos_ind100_2\$PdSNP)$
 $cor(QTLpos_ind100_2\$P0, QTLpos_ind100_2\$PTdSNP)$
 $cor(QTLpos_ind100_2\$P0, QTLpos_ind100_2\$PEDm)$
 $cor(QTLpos_ind100_2\$P0, QTLpos_ind100_2\$PED100_4)$

Calcul des corrélations pour ind=500

pour qtleffect=0

$cor(QTLpos_ind500_0\$P0, QTLpos_ind500_0\$PdSNP)$
 $cor(QTLpos_ind500_0\$P0, QTLpos_ind500_0\$PTdSNP)$
 $cor(QTLpos_ind500_0\$P0, QTLpos_ind500_0\$PEDm)$
 $cor(QTLpos_ind500_0\$P0, QTLpos_ind500_0\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

$cor(QTLpos_ind500_0.5\$P0, QTLpos_ind500_0.5\$PdSNP)$
 $cor(QTLpos_ind500_0.5\$P0, QTLpos_ind500_0.5\$PTdSNP)$
 $cor(QTLpos_ind500_0.5\$P0, QTLpos_ind500_0.5\$PEDm)$
 $cor(QTLpos_ind500_0.5\$P0, QTLpos_ind500_0.5\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

$cor(QTLpos_ind500_1\$P0, QTLpos_ind500_1\$PdSNP)$
 $cor(QTLpos_ind500_1\$P0, QTLpos_ind500_1\$PTdSNP)$
 $cor(QTLpos_ind500_1\$P0, QTLpos_ind500_1\$PEDm)$
 $cor(QTLpos_ind500_1\$P0, QTLpos_ind500_1\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

$cor(QTLpos_ind500_1.5\$P0, QTLpos_ind500_1.5\$PdSNP)$
 $cor(QTLpos_ind500_1.5\$P0, QTLpos_ind500_1.5\$PTdSNP)$
 $cor(QTLpos_ind500_1.5\$P0, QTLpos_ind500_1.5\$PEDm)$
 $cor(QTLpos_ind500_1.5\$P0, QTLpos_ind500_1.5\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

$cor(QTLpos_ind500_2\$P0, QTLpos_ind500_2\$PdSNP)$
 $cor(QTLpos_ind500_2\$P0, QTLpos_ind500_2\$PTdSNP)$
 $cor(QTLpos_ind500_2\$P0, QTLpos_ind500_2\$PEDm)$
 $cor(QTLpos_ind500_2\$P0, QTLpos_ind500_2\$PED100_4)$

Calcul des corrélations pour ind=1000

pour qtleffect=0

$cor(QTLpos_ind1000_0\$P0, QTLpos_ind1000_0\$PdSNP)$
 $cor(QTLpos_ind1000_0\$P0, QTLpos_ind1000_0\$PTdSNP)$
 $cor(QTLpos_ind1000_0\$P0, QTLpos_ind1000_0\$PEDm)$
 $cor(QTLpos_ind1000_0\$P0, QTLpos_ind1000_0\$PED100_4)$

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

$cor(QTLpos_ind1000_0.5\$P0, QTLpos_ind1000_0.5\$PdSNP)$
 $cor(QTLpos_ind1000_0.5\$P0, QTLpos_ind1000_0.5\$PTdSNP)$

cor(QTLpos_ind1000_0.5\$P0,QTLpos_ind1000_0.5\$PEDm)
cor(QTLpos_ind1000_0.5\$P0,QTLpos_ind1000_0.5\$PED100_4)

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

cor(QTLpos_ind1000_1\$P0,QTLpos_ind1000_1\$PdSNP)
cor(QTLpos_ind1000_1\$P0,QTLpos_ind1000_1\$PTdSNP)
cor(QTLpos_ind1000_1\$P0,QTLpos_ind1000_1\$PEDm)
cor(QTLpos_ind1000_1\$P0,QTLpos_ind1000_1\$PED100_4)

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

cor(QTLpos_ind1000_1.5\$P0,QTLpos_ind1000_1.5\$PdSNP)
cor(QTLpos_ind1000_1.5\$P0,QTLpos_ind1000_1.5\$PTdSNP)
cor(QTLpos_ind1000_1.5\$P0,QTLpos_ind1000_1.5\$PEDm)
cor(QTLpos_ind1000_1.5\$P0,QTLpos_ind1000_1.5\$PED100_4)

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

cor(QTLpos_ind1000_2\$P0,QTLpos_ind1000_2\$PdSNP)
cor(QTLpos_ind1000_2\$P0,QTLpos_ind1000_2\$PTdSNP)
cor(QTLpos_ind1000_2\$P0,QTLpos_ind1000_2\$PEDm)
cor(QTLpos_ind1000_2\$P0,QTLpos_ind1000_2\$PED100_4)

Calcul des corrélations pour ind=2000

##pour qtleffect=0

cor(QTLpos_ind2000_0\$P0,QTLpos_ind2000_0\$PdSNP)
cor(QTLpos_ind2000_0\$P0,QTLpos_ind2000_0\$PTdSNP)
cor(QTLpos_ind2000_0\$P0,QTLpos_ind2000_0\$PEDm)
cor(QTLpos_ind2000_0\$P0,QTLpos_ind2000_0\$PED100_4)

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =0.5

cor(QTLpos_ind2000_0.5\$P0,QTLpos_ind2000_0.5\$PdSNP)
cor(QTLpos_ind2000_0.5\$P0,QTLpos_ind2000_0.5\$PTdSNP)
cor(QTLpos_ind2000_0.5\$P0,QTLpos_ind2000_0.5\$PEDm)
cor(QTLpos_ind2000_0.5\$P0,QTLpos_ind2000_0.5\$PED100_4)

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1

```
cor(QTLpos_ind2000_1$P0,QTLpos_ind2000_1$PdSNP)
cor(QTLpos_ind2000_1$P0,QTLpos_ind2000_1$PTdSNP)
cor(QTLpos_ind2000_1$P0,QTLpos_ind2000_1$PEDm)
cor(QTLpos_ind2000_1$P0,QTLpos_ind2000_1$PED100_4)
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =1.5

```
cor(QTLpos_ind2000_1.5$P0,QTLpos_ind2000_1.5$PdSNP)
cor(QTLpos_ind2000_1.5$P0,QTLpos_ind2000_1.5$PTdSNP)
cor(QTLpos_ind2000_1.5$P0,QTLpos_ind2000_1.5$PEDm)
cor(QTLpos_ind2000_1.5$P0,QTLpos_ind2000_1.5$PED100_4)
```

Corrélation entre QTL simulé et QTL observé pour chaque méthode pour QTL effect =2

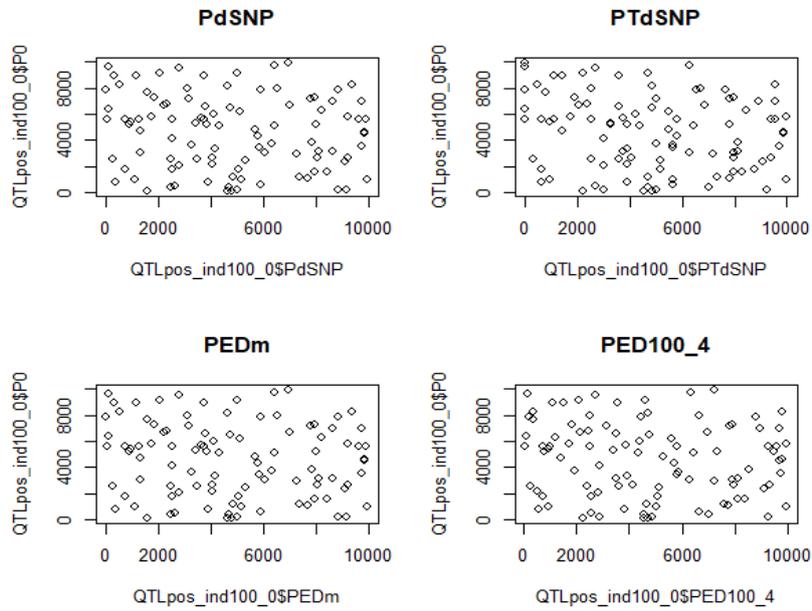
```
cor(QTLpos_ind2000_2$P0,QTLpos_ind2000_2$PdSNP)
cor(QTLpos_ind2000_2$P0,QTLpos_ind2000_2$PTdSNP)
cor(QTLpos_ind2000_2$P0,QTLpos_ind2000_2$PEDm)
cor(QTLpos_ind2000_2$P0,QTLpos_ind2000_2$PED100_4)
```

Le même script a été également écrit pour des valeurs de qtl effect variant de 0 à 0,5 avec un pas de 0,5 :

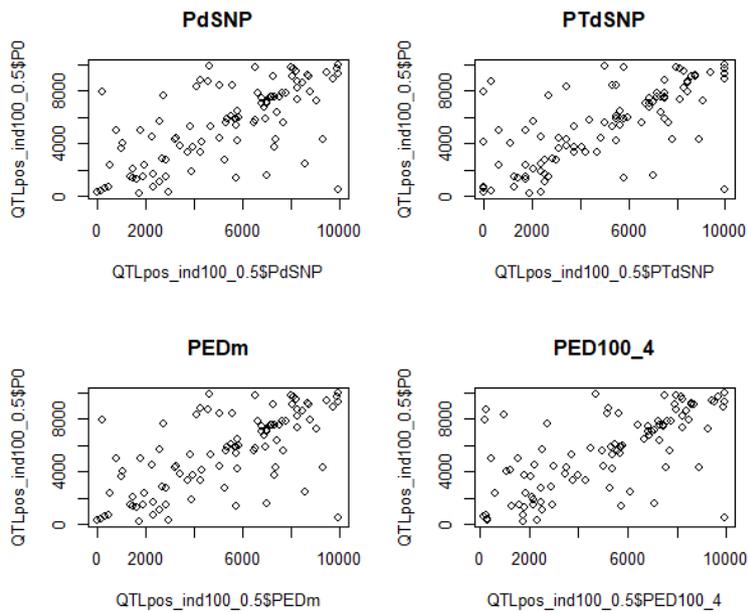
```
qtlectrange<- seq(0, 0.5, by=0.1)
```

Annexe ix : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de QTL effect pour individu =100

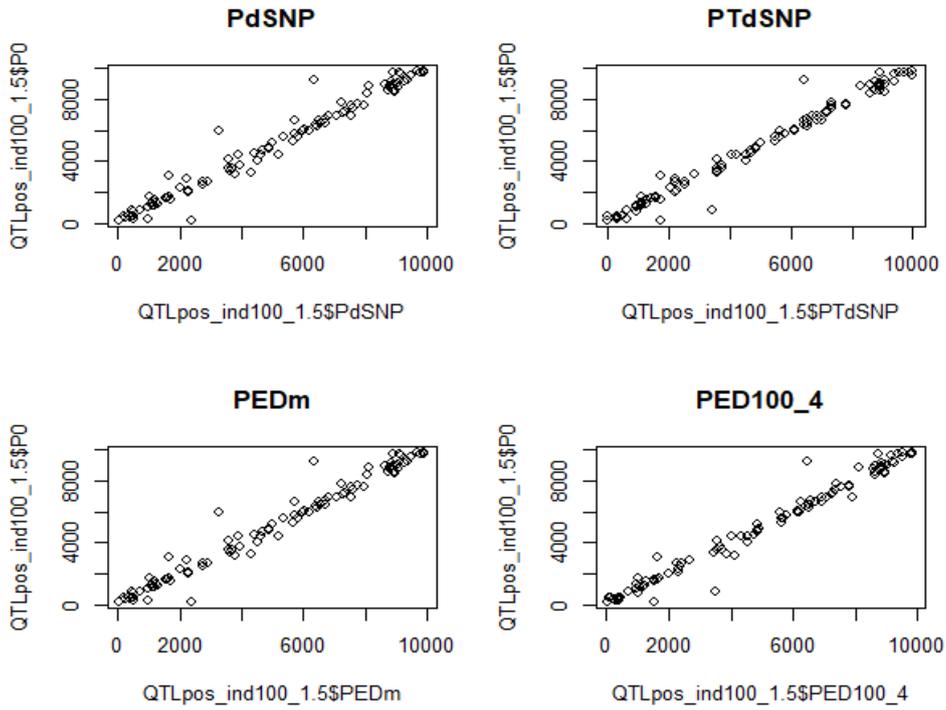
Qtl effect =0



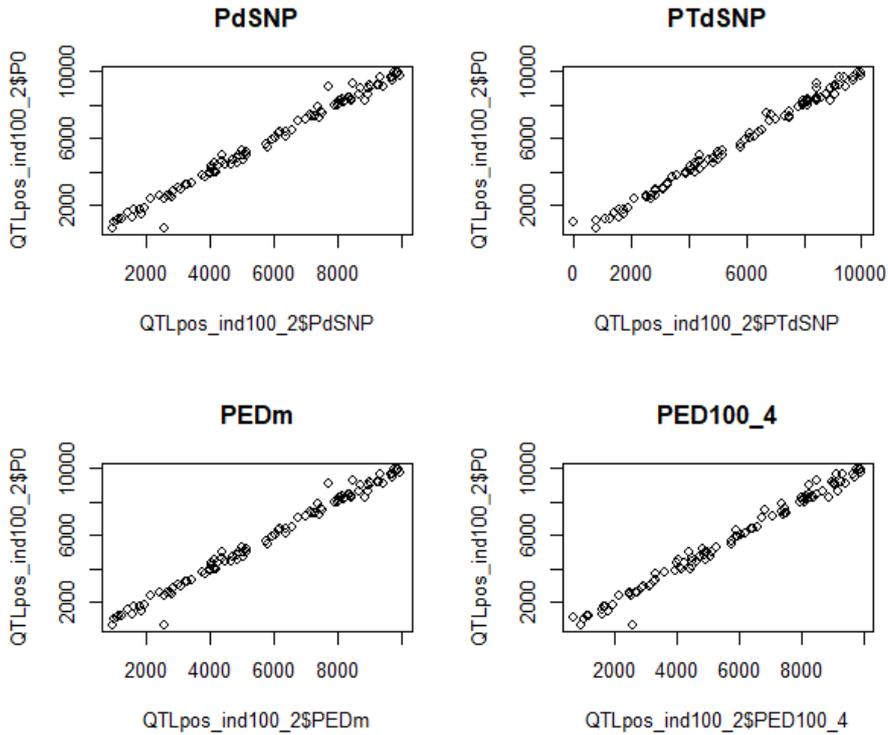
Qtl effect =0,5



Qtl effect =1,5

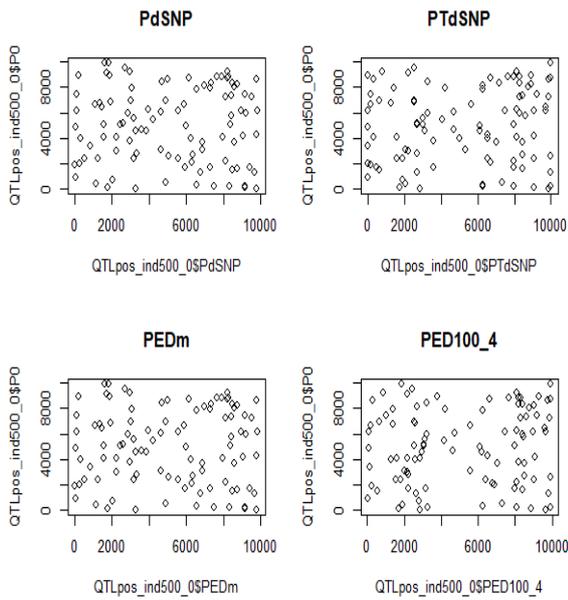


Qtl effect=2

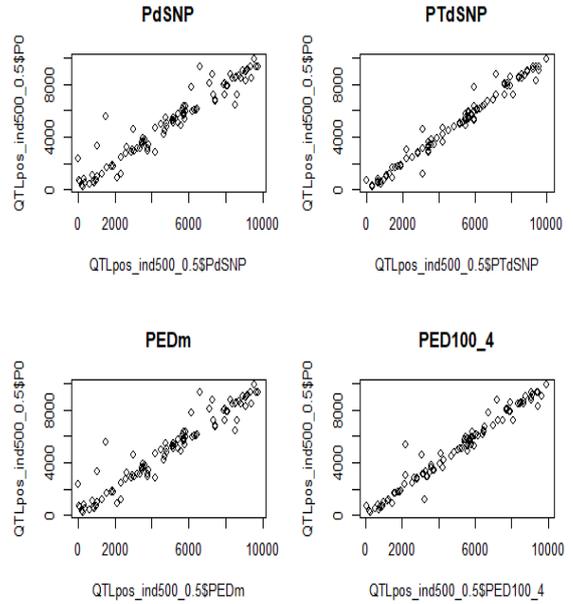


Annexe x : Représentation graphique de la corrélation entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 500

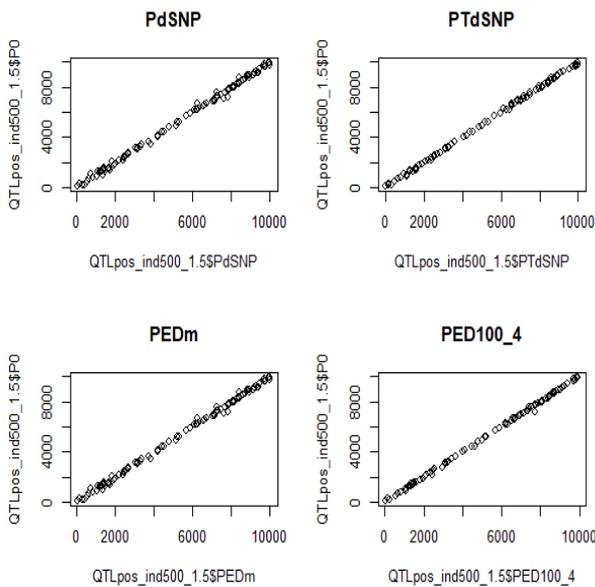
Qtl effect=0



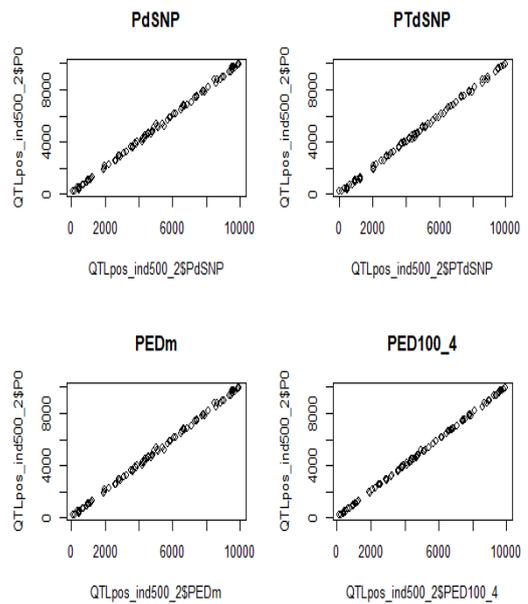
Qtl effect=0,5



Qtl effect=1,5



Qtl effect=2



Annexe xi : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 100

➤ **QTL effect varie de 0 à 2**

	<i>qtleffect</i> 0	0,5	1	1,5	2
<i>PosQtl/méthode</i>					
PdSNP	-0.125	0.666	0.888	0.98	0.993
PTdSNP	-0.187	0.691	0.895	0.987	0.995
PEDm	-0.125	0.666	0.888	0.983	0.993
PED100_4	-0.125	0.661	0.897	0.986	0.992

➤ **QTL effect varie de 0 à 0,5**

<i>Qtleffect</i>	0	0.1	0.2	0.3	0.4	0.5
<i>PosQtl/méthode</i>						
<i>PdSNP</i>	-0.054	-0.004	0.671	0.316	0.931	0.621
<i>PTdSNP</i>	-0.100	0.015	0.137	0.357	0.941	0.661
<i>PEDm</i>	-0.054	-0.004	0.067	0.316	0.931	0.621
<i>PED100_4</i>	-0.013	0.018	0.096	0.342	0.944	0.666

Annexe xii : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 500

➤ QTL effect varie de 0 à 2

	<i>qtleffect</i> 0	0,5	1	1,5	2
<i>PosQtl/méthode</i>					
PdSNP	-0.005	0.960	0.992	0.998	0.9994
PTdSNP	0.115	0.986	0.998	0.999	0.9996
PEDm	-0.005	0.960	0.992	0.998	0.9994
PED100_4	0.081	0.981	0.998	0.999	0.9998

➤ QTL effect varie de 0 à 0,5

	<i>Qtleffect</i> 0	0.1	0.2	0.3	0.4	0.5
<i>PosQtl/méthode</i>						
PDSNP	-0.068	0.177	0.643	0.842	0.932	0.976
PTDSNP	-0.106	0.200	0.650	0.892	0.942	0.986
PEDM	-0.068	0.176	0.643	0.842	0.932	0.976
PED100_4	-0.101	0.213	0.665	0.874	0.944	0.982

Annexe xiii : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 1000

➤ QTL effect varie de 0 à 2

	<i>qtleffect</i>	0	0,5	1	1,5	2
<i>PosQtl/méthode</i>						
PdSNP		0.06	0.991	0.997	0.9991	0.9995
PTdSNP		0.001	0.998	0.999	0.9997	0.9997
PEDm		-0.061	0.991	0.997	0.9981	0.9995
PED100_4		0.060	0.998	0.999	0.9998	0.9999

➤ QTL effect varie de 0 à 0,5

	<i>Qtleffect</i>	0	0.1	0.2	0.3	0.4	0.5
<i>PosQtl/méthode</i>							
PDSNP		-0.080	0.420	0.926	0.957	0.974	0.992
PTDSNP		-0.007	0.435	0.922	0.963	0.985	0.998
PEDM		-0.080	0.420	0.926	0.957	0.974	0.992
PED100_4		-0.014	0.406	0.916	0.965	0.987	0.997

Annexe xiv : Valeur numérique des corrélations entre QTL simulé et estimé pour chaque méthode en fonction de qtl effect pour individu = 2000

➤ **QTL effect varie de 0 à 2**

<i>PosQtl/méthode</i>	<i>qtleffect</i>	0	0.5	1	1.5
PdSNP		0.091	0.992	0.997	0.998
PdSNP		0.001	0.9985	0.999	0.9998
PEDm		0.091	0.992	0.997	0.998
PED100_4		0.093	0.9983	0.999	0.999

➤ **QTL effect varie de 0 à 0,5**

<i>PosQtl/méthode</i>	<i>qtleffect</i>	0	0.1	0.2	0.3	0.4	0.5
PdSNP		0.031	0.491	0.894	0.975	0.984	0.991
PdSNP		0.001	0.660	0.965	0.988	0.998	0.999
PEDm		0.031	0.491	0.895	0.975	0.984	0.991
PED100_4		-0.026	0.668	0.954	0.988	0.998	0.999

Annexe xv : Résultats du contrôle de la qualité des produits PCR

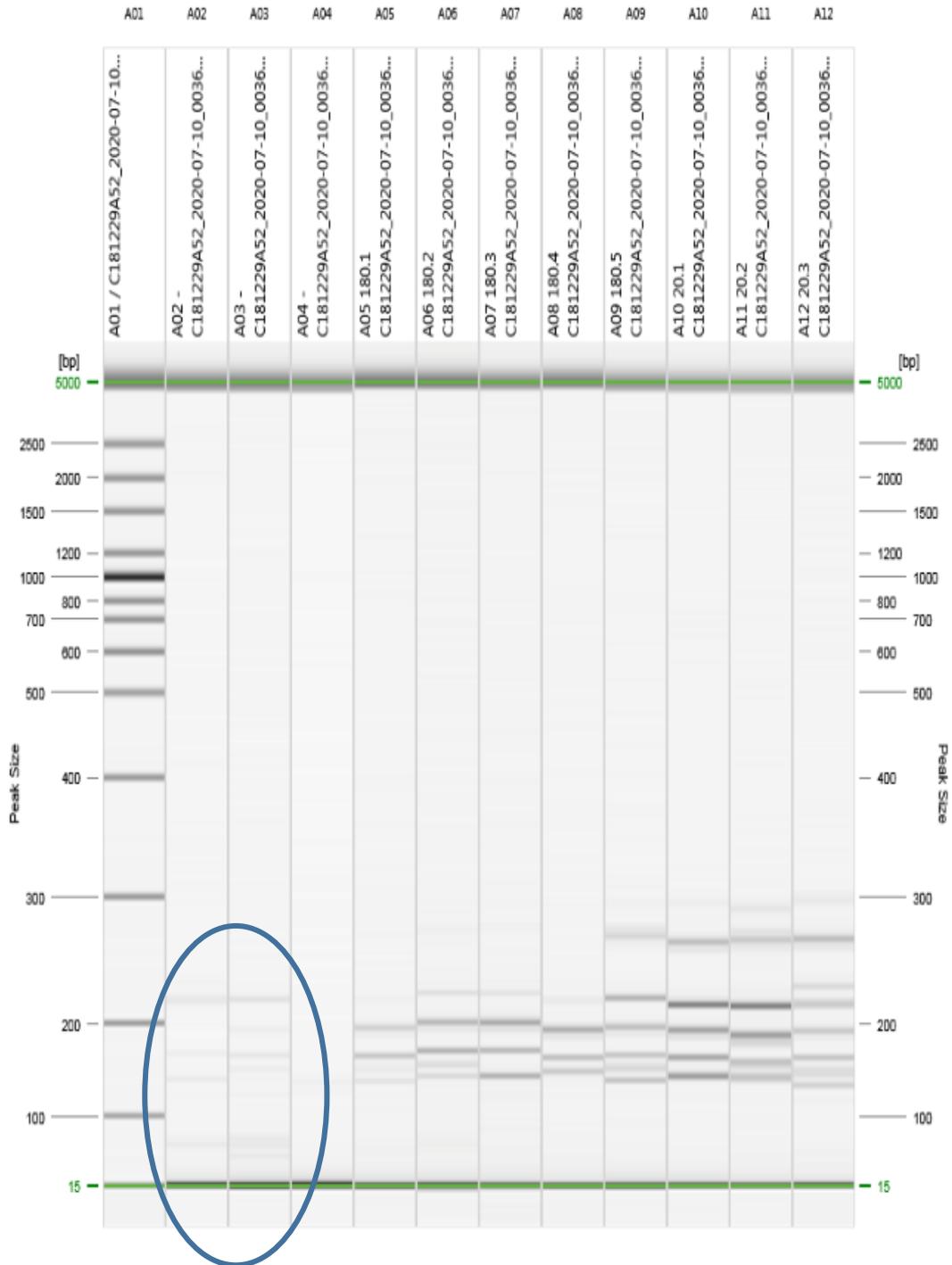


Figure: 1

Amplification dans les
témoins négatifs

Annexe xvi : Phénotypes parentaux

Line 249



Line 337



Annexe xvii : Sélection du modèle ANOVA par « backfor selection »

```
mod1 <- lm(RL ~ Genotype + Repetition + Erlenmeyer + Position +Genotype*Re
petition +Genotype*Erlenmeyer + Genotype*Position + Repetition* Erlenmeyer
+ Repetition*Position + Erlenmeyer*Position
, data= dataRL )

drop1(mod1, .~., test="F")

## Single term deletions
##
## Model:
## RL ~ Genotype + Repetition + Erlenmeyer + Position + Genotype *
## Repetition + Genotype * Erlenmeyer + Genotype * Position +
## Repetition * Erlenmeyer + Repetition * Position + Erlenmeyer *
## Position
## Df Sum of Sq RSS AIC F value Pr(>F)
## <none> 4257.5 989.10
## Genotype 6 535.80 4793.3 1011.48 3.8803 0.0011311 **
## Repetition 2 223.98 4481.5 999.97 4.8663 0.0087158 **
## Erlenmeyer 1 0.19 4257.7 987.12 0.0082 0.9277277
## Position 4 10.78 4268.3 981.84 0.1171 0.9763303
## Genotype:Repetition 17 1134.87 5392.4 1023.63 2.9008 0.0002094 ***
## Genotype:Erlenmeyer 11 249.51 4507.0 983.62 0.9856 0.4610359
## Genotype:Position 44 805.25 5062.8 951.34 0.7952 0.8137426
## Repetition:Erlenmeyer 2 123.38 4380.9 993.39 2.6807 0.0711782 .
## Repetition:Position 8 137.43 4394.9 982.32 0.7464 0.6503893
## Erlenmeyer:Position 4 111.32 4368.8 988.59 1.2093 0.3083699
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod2 <- lm(RL ~ Genotype + Repetition + Erlenmeyer +Genotype*Repetition +G
enotype*Erlenmeyer + Repetition* Erlenmeyer
, data= dataRL )
drop1(mod2, .~., test="F")

## Single term deletions
##
## Model:
## RL ~ Genotype + Repetition + Erlenmeyer + Genotype * Repetition +
## Genotype * Erlenmeyer + Repetition * Erlenmeyer
## Df Sum of Sq RSS AIC F value Pr(>F)
## <none> 5527.3 944.79
## Genotype 6 473.35 6000.6 956.62 3.4969 0.0024544 **
## Repetition 2 219.98 5747.2 952.11 4.8755 0.0083879 **
## Erlenmeyer 1 10.26 5537.5 943.33 0.4548 0.5006947
## Genotype:Repetition 17 1131.83 6659.1 964.82 2.9511 0.0001216 ***
## Genotype:Erlenmeyer 11 241.25 5768.5 935.18 0.9721 0.4724028
## Repetition:Erlenmeyer 2 134.37 5661.6 947.76 2.9780 0.0527396 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

mod3 <- lm(RL ~ Genotype + Repetition +Genotype*Repetition
, data= dataRL )
drop1(mod3,~,.,test="F")

## Single term deletions
##
## Model:
## RL ~ Genotype + Repetition + Genotype * Repetition
## Df Sum of Sq RSS AIC F value Pr(>F)
## <none> 5931.9 937.28
## Genotype 6 440.08 6371.9 946.04 3.2025 0.0047447 **
## Repetition 2 283.51 6215.4 946.82 6.1894 0.0023672 **
## Genotype:Repetition 17 1055.54 6987.4 950.78 2.7110 0.0003894 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Résumé

TITRE : Caractérisation de QTL contrôlant des traits racinaires chez le mil

Le mil, *Pennisetum glaucum*, constitue l'aliment de base de plus de 90 millions de personnes en Afrique sub-saharienne et en Inde. La production mondiale de mil est limitée par la sécheresse et une faible fertilité des sols qui caractérisent sa zone de culture. L'amélioration de son système racinaire pourrait contribuer à optimiser l'acquisition de l'eau et des éléments minéraux. La connaissance de la base génétique qui contrôle l'établissement de ce système racinaire peut permettre l'utilisation de la sélection assistée par marqueurs, un puissant outil qui fixe la sélection sur les traits sélectionnés. L'étude de l'architecture racinaire et de la dynamique de croissance racinaire du mil a permis de mettre en évidence certains traits racinaires d'intérêt : la croissance précoce de la racine séminale et l'agrégation du sol rhizosphérique. Des études préliminaires ont abouti à la détection de plusieurs QTLs contrôlant ces traits par génétique d'association. L'objectif de ce stage était de valider ces QTLs d'intérêt en utilisant des populations obtenues à partir de deux parents présentant les caractères contrastés. Nous avons tout d'abord comparé l'efficacité de quatre méthodes de détection de QTL à partir d'une analyse de ségrégation en masse. Une étude de simulation a été réalisée pour évaluer l'effet de la variation de l'effet du QTL et du nombre d'individus sur la précision de détection de QTL pour ces différentes méthodes. Par ailleurs, nous avons initié la caractérisation d'une population de lignées recombinantes fixées (BC2F4) en évaluant l'homogénéité d'un échantillon de lignées par génotypage et phénotypage. Ces travaux contribueront à la détection de QTLs d'intérêt.

Mots clés : *Pennisetum glaucum*, Lignées BC2F4, Racines, Phénotypage, analyse de ségrégation en masse.

- Pour citer ce document : NAKOMBO-GBASSAULT, Princia, 2020. Caractérisation de QTL contrôlant des traits racinaires chez le mil. Mémoire de Master 2, Sciences et Technologies de l'agriculture, de l'alimentation et de l'environnement, Sélection et évolution des plantes méditerranéennes et tropicales, Montpellier SupAgro. 98p.
- L'institut Agro/Montpellier SupAgro, 2 place Pierre Viala, 34060 Montpellier.
www.supagro.fr

