

## Mémoire de fin d'études

Présenté pour l'obtention du Diplôme de Master international en Sciences et technologie de l'agriculture, de l'alimentation et de l'environnement

**Spécialité : Sélection et Évolution des Plantes Méditerranéennes et Tropicales (SEPMET)**

---

**Dynamique de la sélection du mil depuis sa domestication à nos jours**

---

par Abdou Rahmene WADE

Année de soutenance : 2018

# Mémoire de fin d'études

Présenté pour l'obtention du Diplôme de Master international en Sciences et technologie de l'agriculture, de l'alimentation et de l'environnement

**Spécialité : Sélection et Évolution des Plantes Méditerranéennes et Tropicales (SEPMET)**

---

**Dynamique de la sélection du mil depuis sa domestication à nos jours**

---

par Abdou Rahmane WADE

Mémoire préparé sous la direction de :

**Yves VIGOIROUX**  
**Anne-Céline THUILLET**  
**Philippe CUBRY**  
Tuteur pédagogique  
**Vincent RANWEZ**  
Membres du jury  
**Pierre BERTHOMIEU**  
**Nicolas BIERNE**  
**Jacques DAVID**  
Présenté le 13/09/2018

**Équipe DYNADIV**  
Organisme d'accueil : IRD UMR DIADE  
911 Avenue Agropolis, 34090 Montpellier

## **Remerciements**

Ce stage n'aurait pas été possible sans Yves VIGOUROUX, que je tiens à remercier pour m'avoir donné l'opportunité de travailler sur un sujet aussi intéressant. Merci à toi pour ta patience, ta disponibilité, pour m'avoir enseigné tellement en si peu de temps et aussi pour avoir eu le courage de corriger mon rapport, ce qui n'a pas dû être une mince affaire.

Un grand merci aussi à Anne-Céline THUILLET pour sa gentillesse, son aide précieuse et sa bonne humeur

Merci à Philippe CUBRY pour sa gentillesse, son soutien notamment sur les codes R

Je remercie toute l'équipe du projet CultiVar, pour m'avoir permis de réaliser ces deux années de master.

Merci à Vincent Ranwez pour m'avoir tutoré et répondu à toutes mes questions.

Je remercie les membres du jury de prendre le temps de lire ce mémoire et de se déplacer pour la soutenance.

Merci à tous les enseignants et responsables de la formation APIMET-SEPMET de Montpellier SupAgro.

Merci à tous mes Amis Mention spéciale à Mamadou MBAYE et Isidore A. Diouf.

**Je Dédie ce mémoire à mes parents et à Cheikh Abdoulahi Mbacké je vous aime !**

# Table des matières

<b>Introduction</b>	<b>1</b>
<b>Matériels et Méthodes</b>	<b>9</b>
<b>Matériels</b> . . . . .	9
<b>Méthodes</b> . . . . .	10
Déttection de la sélection avec une méthode haplotypique . . . . .	10
Détermination d'un seuil de significativité . . . . .	11
Identification des gènes candidats à la sélection et de leurs fonctions . . . . .	15
<b>Résultats</b>	<b>16</b>
Détermination d'un seuil de significativité . . . . .	16
Modélisation démographique neutre de la domestication du mil . . . . .	16
Estimation du taux de recombinaison ( $\rho_{cult}$ ) . . . . .	17
Simulation de séquences . . . . .	18
Détection des signatures de sélection forte et douce . . . . .	20
Analyse <i>GO terms</i> des gènes détectés . . . . .	22
<b>Discussion</b>	<b>23</b>
<b>Conclusion et Perspectives</b>	<b>24</b>
<b>Annexe A</b>	<b>A</b>
<b>Annexe B</b>	<b>B</b>
<b>Annexe C</b>	<b>C</b>
<b>Annexe D</b>	<b>D</b>
Résumé . . . . .	E
Abstract . . . . .	E

# Table des figures

1	Évolution des températures du globe de 1850 à 2012 (Source IPCC 2014) . . . . .	1
2	Variations de températures du globe de 1901 à 2012 (Source IPCC 2014) . . . . .	1
3	Tendances de précipitations annuelles 1901 et 2005 (IPCC, 2014) . . . . .	2
4	Évolution annuelle des précipitations au Sahel entre 1950 à 2010 (Salack et al. 2016)	2
5	Évolution du rendement mondial des céréales blé, riz et maïs entre 1960 et 2014 . . .	3
6	Types de balayage dur et doux (?hermissontsoft2017) . . . . .	6
7	Comparaison de généralogies générées par un <i>Bottleneck</i> et un palayage selectif (source [Pavlidis and Alachiotis 2017]) . . . . .	7
8	Présentation du mil <i>Pennisetum glaucum</i> . . . . .	8
9	<i>Distribution géographique des 190 RILs de mil cultivé utilisées dans ce stage</i> . . . . .	9
10	<i>Scénarios démographiques étudiés</i> . . . . .	12
11	Scan du génome avec la statistique H12 . . . . .	16
12	Résultat des modélisations démographiques faite en prenant que les mils du centre du Sahel . . . . .	17
13	Distribution des estimations du taux de recombinaison par génération . . . . .	19
14	Estimation du DL entre deux nucléotides en fonction de leur distance . . . . .	19
15	Distributions de H12 calculés à partir de nos séquences simulées et de nos données réelles . . . . .	21
16	Scan du génome avec la statistique H12 avec le seuil de significativité . . . . .	21
17	Modélisation démographique du Modèle 2 sans prendre en compte d'une erreur de polarisation et en utilisant tout notre jeu de données . . . . .	B
18	Modélisation démographique du Modèle 2 sans prendre en compte d'une erreur de polarisation et en utilisant que les mils du centre du Sahel . . . . .	B
19	Modélisation démographique du Modèle 2 avec prise en compte d'une erreur de polarisation et en utilisant tout notre jeu de données . . . . .	B
20	Modélisation démographique du Modèle 2 avec prise en compte d'une erreur de polarisation et en utilisant que les mils du centre du Sahel . . . . .	B
21	Résultat des modélisations démographiques . . . . .	C
22	Distribution des estimations de $\rho$ le long de chaque chromosome du mil . . . . .	D

# Liste des tableaux

1	Densité de marquage SNPs . . . . .	10
2	Distributions à priori des paramètres démographiques utilisées durant l'analyse $\delta aoi$	

### 3 Tableau des maximums de vraisemblance des paramètres démographiques obtenues après l'analyse *delta* . . . . . 18

# Introduction

Le système climatique mondial a subi plusieurs changements dans le passé. Cependant, les changements du climat auxquels nous assistons actuellement sont d'une rapidité exceptionnelle, comme en témoigne le réchauffement de la Terre de  $0,85^{\circ}\text{C}$  du *XIX<sup>e</sup>* siècle à nos jours (IPCC 2014, Figure 1 et 2).

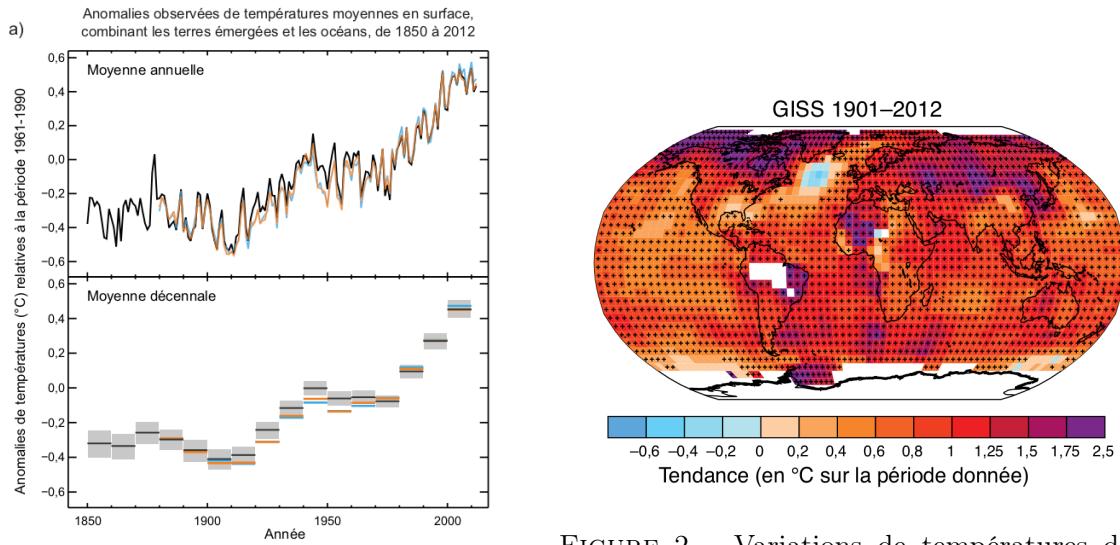


FIGURE 1 – Évolution des températures du globe de 1850 à 2012 (Source IPCC 2014).

Le changement de température est donné en  $^{\circ}\text{C}$  par année (figure du haut) et par décennie (figure du bas). Les variations de température sont mesurées par rapport à la normale climatologique de la période 1961-1990.

FIGURE 2 – Variations de températures du globe de 1901 à 2012 (Source IPCC 2014). Nous remarquons des spots de hausse de température en Afrique au niveau de la zone sahélienne.

Cette augmentation de température entraîne, par une plus intense évaporation des eaux du globe, des perturbations dans le cycle de l'eau et donc un dérèglement de la pluviométrie. Globalement les précipitations ont augmentées dans l'ensemble du globe depuis le début du XX<sup>e</sup> siècle notamment dans l'hémisphère nord (cite, Figure 3). Cependant pour certaines régions la tendance moyenne a été une réduction de la pluviométrie. En particulier le sahel qui a connu une période de forte sécheresse allant de la fin des années 1960 au début des années 1990 (Figure 4) conduisant à la dégradation complète de plus d'un quart du sol dans les zones pastorales sèches sahariennes (IRD). Depuis les années 1990 la pluviométrie a repris (GIEC, 2008). Cette reprise est cependant accompagnée de dérèglements dans le déroulement de la saison des pluies saharienne par rapport à ce qui a été habituellement observé dans cette région (Salack et al. (2016)).

Ce changement rapide et brutal du climat se traduit par des changements au niveau des populations de plantes et d'animaux.

En réponse à cette contrainte nouvelle, les espèces peuvent migrer vers des écosystèmes plus

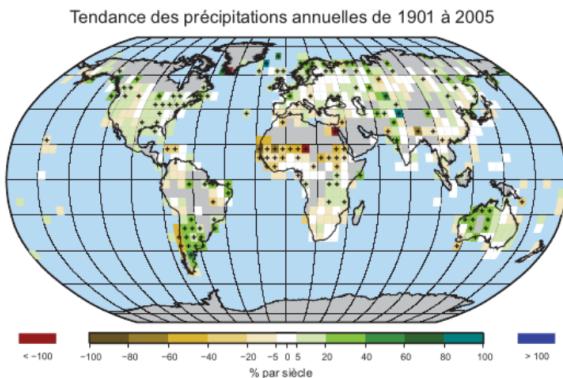


FIGURE 3 – Tendances de précipitations annuelles 1901 et 2005 (IPCC, 2014)  
Dans la zone sahélienne, les pluies ont été réduites de 40 et 80% par siècle.

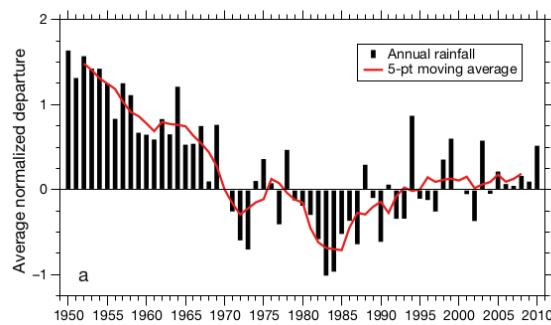


FIGURE 4 – Évolution annuelle des précipitations au Sahel entre 1950 à 2010 (Salack et al. 2016)  
Les variations de précipitations sont calculés par rapport aux données observées

propices à leur survie et reproduction.

Les individus capables de plasticité phénotypique peuvent modifier leur phénotype en réponse à un changement d'environnement. Cette réponse peut être insuffisante pour permettre aux individus de répondre à ces variations environnementales fortes. Par ailleurs elle peut parfois être maladaptative (Academie des sciences 2017).

Une autre stratégie est l'adaptation génétique. En effet, si un changement de conditions environnementales rapide entraîne une pression de sélection dans une population. Alors les individus qui ont un génotype plus favorables sont alors sélectionnés. Comme les mutations sont rares et les changements d'environnements rapides, une adaptation génétique à partir de la diversité fonctionnelle préexistante paraît plus probable (Messer and Petrov 2013).

Dans ce contexte, la diversité fonctionnelle existante au sein des populations pourrait être un paramètre important pour garantir la viabilité et le potentiel évolutif des populations face aux changements globaux.

Les populations de plantes cultivées subissent au même titre que les autres les effets du changement climatique. Ces effets se traduisent par une baisse des rendements (Figure 5) c'est ce qui a été observé pour les céréales les plus consommées au monde, blé, riz et maïs (de 1960 à nos jours). Cette baisse a conduit à plusieurs épisodes d'augmentation rapide des prix de ces céréales consécutifs d'où un problème de sécurité alimentaire (IPCC, 2014).

Les plantes cultivées, pendant la diffusion de l'agriculture, ont eu à s'adapter à différents environnements mais aussi au besoins de l'Homme. C'est pendant ces périodes que sont apparues les variétés traditionnelles (landraces). Ces variétés présentent sans doute une diversité fonctionnelle susceptible de répondre à de nouvelles pressions de sélection, notamment celles induites par les changements climatiques. De ce fait, identifier et comprendre comment la diversité fonctionnelle des

plantes cultivées a été façonnée durant leur histoire paraît aujourd’hui important.

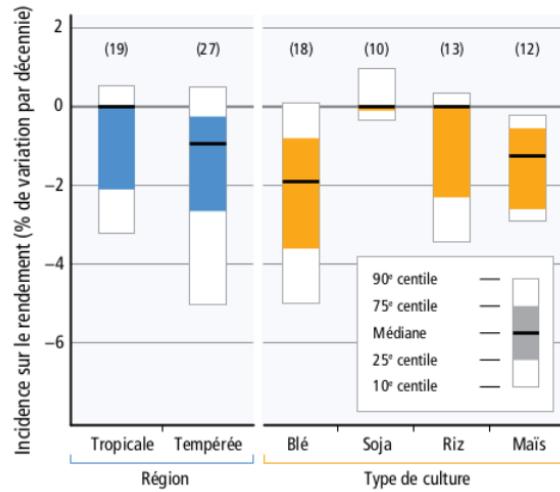


FIGURE 5 – Évolution du rendement mondial des céréales blé, riz et maïs entre 1960 et 2014 (IPCC,2014)

Identifier cette diversité fonctionnelle peut être réalisé par l'analyse de la sélection à l'échelle du génome. Lors qu'une mutation apparaît dans une population, conférant un avantage ou un désavantage aux individus qui la portent à survivre ou à ce reproduire. Alors l'évolution de la fréquence de cette mutation dans la population se fera sous l'effet de la sélection. Pour les mutation bénéfiques, la sélection à tendance à faire augmenter leurs fréquences dans la population au cours des générations (Haldane ; 1927). Inversement les mutations désavantageuses ont tendance à disparaître de la population sous l'effet de la sélection.

Plusieurs types de sélections ont été décrits (Hill and Robertson [1966], Smith and Haigh [1974], Glémin and Bataillon [2009], Charlesworth, 2006), nous allons nous concentrer dans ce travail sur la sélection positive (Smith and Haigh [1974]). On parle de sélection positive lorsque la sélection agit d'une manière directionnelle, en augmentant les fréquences d'une mutation bénéfique dans une population sous pression de sélection.

Face à un changement d'environnement, créant une nouvelle pression de sélection dans une population, l'adaptation génétique de cette dernière peut se faire de différentes manières via la sélection positive (Hermisson and Pennings [2005], Figure 6).

- L'adaptation génétique peut se faire à partir d'une nouvelle mutation bénéfique apparue après le changement d'environnement. La sélection positive agit en augmentant la fréquence de cette mutation bénéfique dans la population (Smith and Haigh [1974]). Cette augmentation de fréquence va entraîner les locus proches de ladite mutation créant ainsi un Déséquilibre de liaison (DL) autour de cette dernière (Kelly [1997]). Elle va aussi réduire la diversité au niveau de ce locus dans la population ainsi que son hétérozygotie (Smith and Haigh [1974], Kaplan, Hudson and Langley [1989], Stephan, Wiehe and Lenz [1992]). Ce phénomène est appelé balayage sélectif ou *selective sweep*. Ainsi chaque individu de la population présente le même haplotype (c'est-à-dire le même fond génétique sur lequel la mutation bénéfique s'est produite). En s'éloignant de la mutation bénéfique le long du chromosome, le polymorphisme se rétabli sous l'effet de la recombinaison (Kaplan, Hudson and Langley 1989). Ce processus produit également un excès d'allèles rares (Braverman et al. 1995, Fay and Wu [2000]). On appelle ce scénario *Hard selective sweep*.
- L'adaptation génétique peut se faire à partir de la diversité fonctionnelle préexistante. Dans ce modèle d'adaptation, ladite mutation a évolué d'abord sous l'effet de la dérive pendant un certain temps, jusqu'à ce qu'un changement dans l'environnement lui confère un avantage sélectif. De plus comme elle n'avait pas, durant ce temps, subie de sélection, ses loci voisins ont aussi subi les effets de la mutation, de la dérive et de la recombinaison. Donc plusieurs haplotypes peuvent être présents au niveau de cette région (Hermisson and Pennings [2005], Orr and Betancourt [2001]). Après le changement de l'environnement, les différents haplotypes présents ont tous la même valeur sélective et sont tous sélectionnés. La réduction de la diversité qui en résulte est donc moins prononcée que dans le *Hard selective sweep* (Hermisson and

---

(Pennings 2005). Ce scenario est appelé balayage doux *Soft selective sweep*. Les balayages doux ne produit pas un excès d'allèles rares de la même envergure que le balayage fort (Prezeworski, Coop and Wall 2005). Ce mode de sélection a également un impact différent sur le déséquilibre de liaison qui est en espérance plus faible dans cette région par rapport à celui attendu en *Hard selective sweep* (Schriener et al. 2015). De plus le *Soft selective sweep* peut conduire à une certaine structuration génétique au niveau la région chromosomique concernée due à la sélection des différents haplotypes ayant la mutation bénéfique.

Ces différents types de balayages sélectif participent au façonnement de la diversité. Les différentes méthodes qui permettent de détecter leur signatures sur les patrons de diversité sont basées sur leurs différents caractéristiques décrits plus haut de manière non exhaustive. La plupart des méthodes de détection de signature de sélection a été conçue pour détecter des traces de sélections positives forte ou *Hard selective sweeps* (Pavlidis and Alachiotis 2017). Ils ont donc en générale de faibles puissances pour détecter les *Soft selective sweep*. Cependant les méthodes basées sur la diversité haplotypique ou le déséquilibre de liaison ont en principe de bonnes capacité à détecter à la fois les balayages sélectifs forte comme douces (Garud et al. 2015, Hermisson and Pennings 2017). En effet ces méthodes se basent sur les principales différentes des conséquences entraînées par le balayage sélectif doux et balayage sélectif fort. Dans notre optique compréhension du façonnement de la diversité, nous avons choisi une méthode qui se base sur la diversité haplotypique afin de détecter à la fois les régions sous balayage sélectif fort comme doux. Cette méthode H12 a été appliquée par exemple sur la drosophile et le maïs (Garud et al. 2015, Lorant 2018).

Au delà de la sélection, la démographie peut induire des profils de diversité semblables à ceux engendrés par la sélection positive (Pavlidis and Alachiotis 2017). Par exemple, un goulet d'étranglement (*bottleneck*) conduit à une augmentation du taux de coalescent dans la population qui la subie (Figure 7). Elle entraîne de ce fait une réduction de la diversité au sein de la population, par conséquent une diminution de l'hétérozygotie et une augmentation du DL dans la population. Ces effets sont confondants avec ceux décrits pour un balayage sélectif (Jensen et al. 2005). Cependant on considère que les effets du *bottleneck* touchent la globalité du génome alors que les effets du balayage sélectif sont localisés. De ce fait nous pouvons utiliser un modèle démographique pour contrôler la signature de la sélection.

L'utilisation d'un modèle démographique nous permettra d'avoir une approximation du profil de diversité attendu en absence de sélection et de détecter avec plus de précision et de confiance les régions sous balayage sélectif. Cette approche a été réalisée dans beaucoup d'études concernant plusieurs espèces par exemple le maïs (Vigouroux et al. 2002). La plupart plantes cultivées en particulier ont une histoire démographique complexe avec une ou plusieurs phases de *bottleneck*, notamment durant leur(s) domestication(s) (Glémén and Bataillon 2009).

La recombinaison est aussi un paramètre important à considérer lors de la détection de sélection. En effet aussi bien la démographie que la sélection positive peuvent créer ou augmenter le DL qui

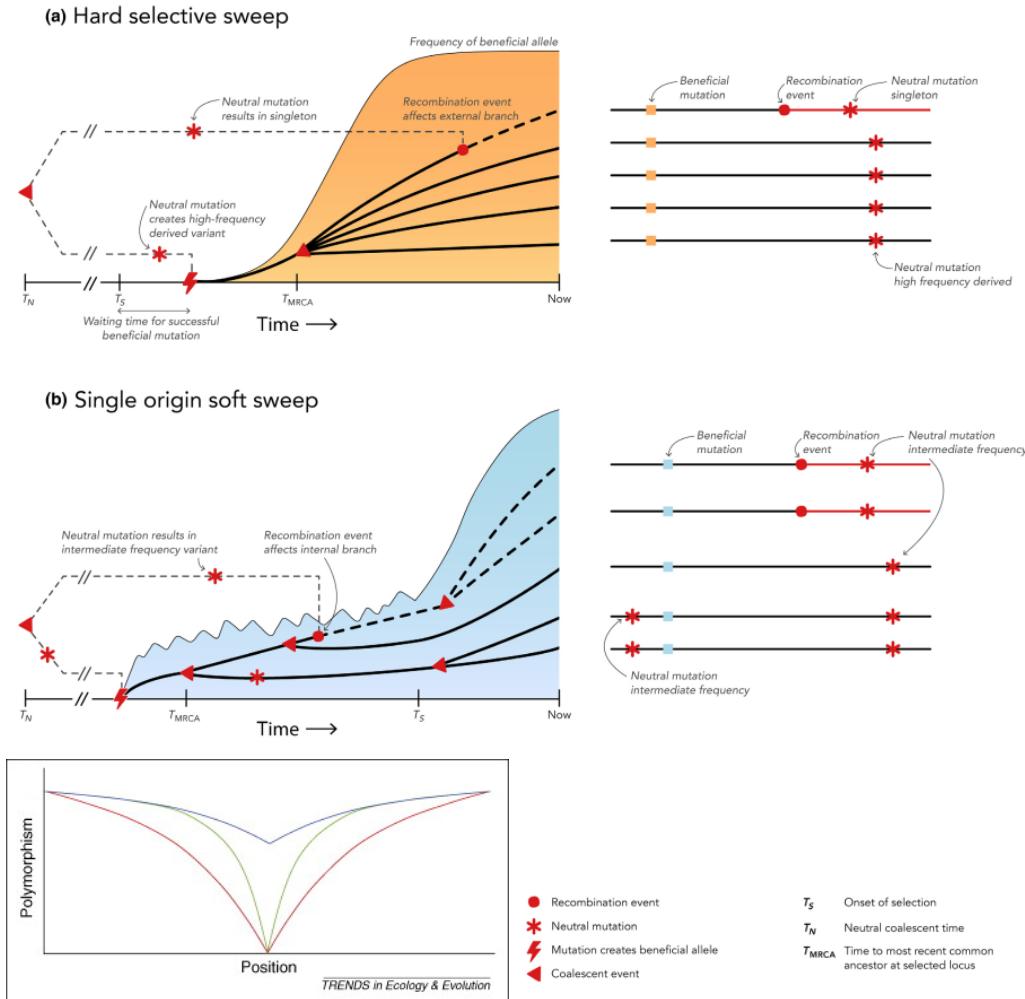


FIGURE 6 – Types de balayage dur et doux(Hermisson and Pennings 2017)

Les régions colorées (bleu et jaunes) représentent l'évolution de la fréquence des copies de l'allèle bénéfique. Les lignes noires peines et celle en pointillées retrace l'histoire de la coalescence des différents haplotypes contenant l'allèle bénéfique. les figures se trouvant à droite montre, les mutations et les événements de recombinaison sur les haplotypes des cinq individus échantillonnés. La partie (a) : On parle de balayage sélectif fort lorsque le temps jusqu'à l'ancêtre commun le plus récent de la mutation bénéfique(TMRCA) est plus court que le temps écoulé depuis le début de la sélection TS. De ce fait, toute variation ancestrale sur des sites étroitement liés à la mutation bénéfique est éliminée à cause de la sélection positive. La recombinaison agit en remettant de la diversité dans les régions adjacentes à la mutation bénéfique, ce qui crée un excès d'allèles rares dans la population. Partie (b) Pour un balayage doux à partir de la diversité préexistante, plusieurs haplotypes contiennent la mutation qui sera bénéfique après l'apparition de la pression de sélection. Là aussi la recombinaison agit de manière similaire au balayage sélectif dur. La Partie (c) compare la réduction de la diversité attendue pour un balayage sélectif fort (rouge) et celle attendu pour un balayage sélectif doux(vert).

subsistera tant que la recombinaison ne le cassera pas.

Pour comprendre d'avantage le fonctionnement de la diversité fonctionnelle, nous allons appliquer une méthode de détection de la sélection, qui se base sur la diversité haplotypique (Garud et al. 2015), sur des données génomiques de mil cultivé.

Le mil cultivé, *Pennisetum glaucum* subsp. *glaucum* est un céréale du genre *Pennisetum*, de la sous famille des *Panicoideae* (famille des *Poaceae*). Cette graminée est diploïde à reproduction

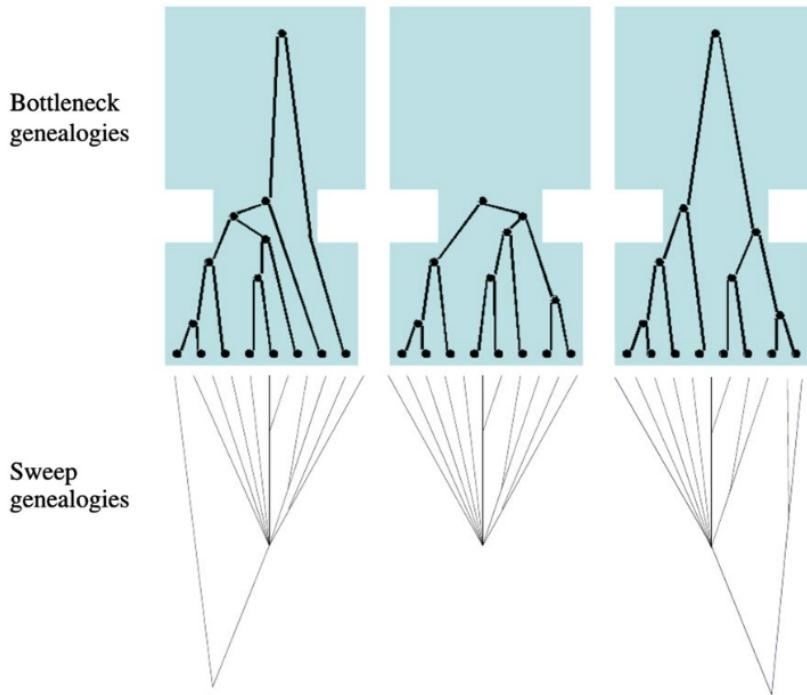


FIGURE 7 – Comparaison de généralogies générées par un *Bottleneck* et un palayage selectif (source (Pavlidis and Alachiotis 2017))

Les scénarii de goulot d'étranglement en bleu peuvent donner lieu à des généralogies similaires à un balayage sélectif (en bas).

Les deux modèles peuvent produire des arbres coalescents très courts.

allogame et possède 7 paires de chromosomes ( $2n = 2x = 14$ ).

Le mil cultivé (Figure 8a et 8b) sert d'aliment de base à près de 100 millions de personnes dans le monde, à travers le Sahel et les franges du désert de Thar en Inde (Gulia et al. 2007, USAID, 2014). Il est cultivé sur plus de 31 millions d'hectares dont 63% en Afrique. Il est aussi la plus tolérante à la sécheresse de toutes les céréales domestiques (Govindaraj et al. 2010). Il est couramment cultivé dans les régions les plus chaudes et les plus sèches du globe où d'autres céréales sont susceptibles d'échouer en raison de la sécheresse, du stress lié aux températures élevées et des mauvaises conditions du sol.

L'aire de répartition du mil couvre la région sahélienne, allant du Sénégal au Soudan, mais aussi dans une partie du sud-est du continent africain. En Asie, le mil est principalement cultivé en Inde, au Moyen-orient et en Chine (Figure 8a).

Le mil est la sixième culture céréalière la plus importante au monde, après le maïs (*Zea mays L.*), le riz (*Oryza sativa L.*), le blé (*Triticum aestivum L.*), l'orge (*Hordeum vulgare L.*) et le sorgho (*Sorghum bicolor (L.) Moench*) (FAOSTAT, 2016). Sa production mondiale était de 28.357.451 de tonnes en 2016. L'Inde étant le premier producteur de mil au monde avec environ 36% de cette production mondiale. Cependant parmi des 15 pays les plus grands producteurs de mil au monde 9 sont africains dont 8 pays sahéliens (Figure 8c).

Le grain du mil cultivé est plus nutritif que celui du blé, du riz, du maïs et du sorgho (Agte, Tarwadi and Chiplonkar [1999]; Muthamilarasan et al. [2016]). Cette qualité nutritionnelle provient de niveaux élevés de protéines, de vitamines, d'acides aminés essentiels, d'antioxydants et de micronutriments essentiels, comme le fer et le zinc (Agte, Tarwadi and Chiplonkar [1999])

La forte croissance démographique au Sahel<sup>a</sup>, et les contraintes liés aux changements climatiques constituent deux grand défis pour la production mil dans cette zone où une situation d'insécurité alimentaire prévaut. À ce problème s'ajoute la faible disponibilité en phosphore des sols sahéliens qui est une des principales contraintes à la culture du mil dans cette région (Khouma et al., 2005, Hash, Schaffert and Peacock [2002]). Ce qui pourrait conduire à une utilisation importante d'engrais importés coûteux économiquement.

Cependant les rendements de mil au Sahel et au Niger en particulier n'ont pas trop évolués depuis 1960 (FAOSTAT, 2016, Figure 8d). Tandis que la production de mil dans cette région a augmenté mais grâce à une l'extension des superficies de culture. La possibilité d'étendre les superficie agricoles étant limitée, la sécurisation de la production dans un contexte de changement climatique pourrait donc passer par des stratégies agricoles qui prendront en compte le potentiel adaptatif des variétés de mil<sup>b</sup>

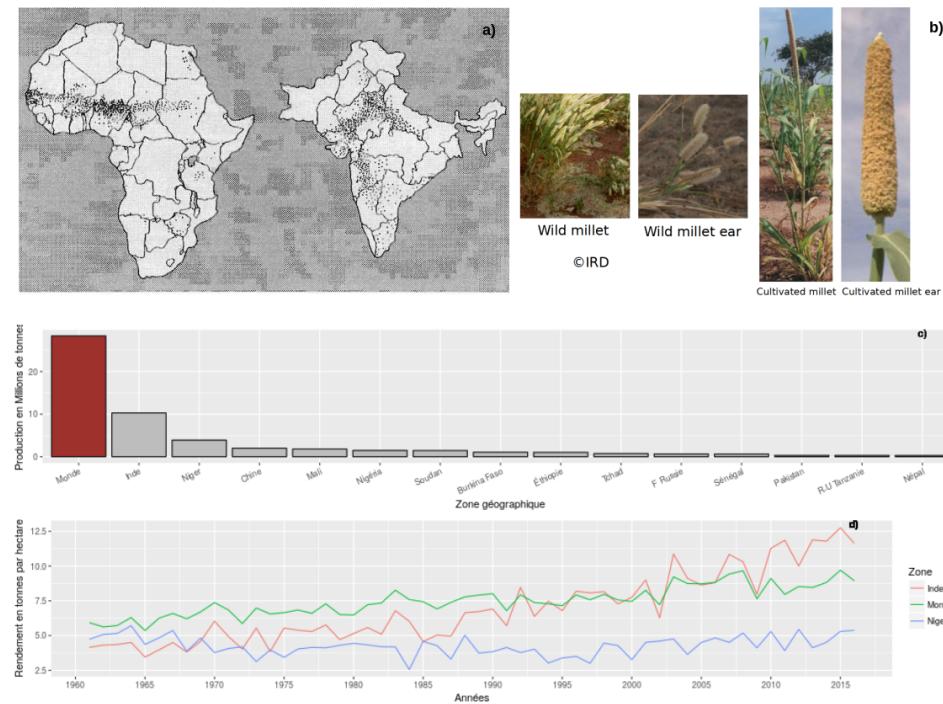


FIGURE 8 – Présentation du mil *Pennisetum glaucum*

La figure a) Zones de culture du mil , en Afrique et en Inde, d'après KUMAR (1989). Un point correspond à 20 000 hectares

(source:Bezanccon 1997). Figure b) les deux photos de droite montrent une plante entière de mil cultivé (gauche) et un épis (droite) ; les deux photos de gauche montrent comme celles de droite du mil sauvage. La Figure c)présente la production en millions de tonnes de mil des15 pays plus grands producteurs de mil. La Figure d) présente l'évolution du rendement mondial (en vert) en tonnes par hectare de mil, celui de l'Inde (en rouge) et en Niger (en bleu).

a. Selon les Nations Unies (N.U.), la population du Sahel a été multipliée par 5,2 entre 1950 et 2015. Les prévisions des N.U. prévoient multiplication de la population sahélienne de 6,1 entre 2015 et 2100

# Matériels et Méthodes

## Matériels

Un jeu de données issu d'un échantillonnage de 190 lignées de mil cultivé est utilisé. Ces mils cultivés proviennent de 23 pays d'Afrique ou d'Asie (Figure 9). Notre échantillonnage, comparé à la distribution des mils cultivés est assez représentatif des régions du monde où le mil est cultivé (Council, National Research, 2001, Figure 8a).

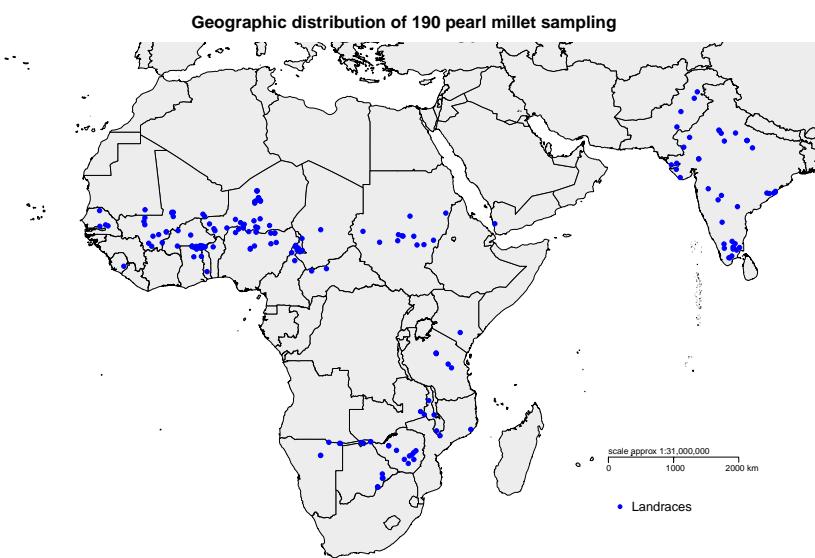


FIGURE 9 – *Distribution géographique des 190 RILs de mil cultivé utilisées dans ce stage*  
Nous avons utilisé 190 accessions de mil cultivé représentatives des mils cultivés dans le monde. En Afrique dans la presque totalité de la zone subsaharienne allant du Sénégal au Soudan. Aussi une partie du sud-est de l'Afrique. En Asie du sud, notamment en Inde et au Pakistan.

Ces 190 individus de mil cultivé sont des lignées obtenues par autofécondation (RILs) haute-ment homozygote mais il reste toujours une variabilité résiduelle. Cette variabilité résiduelle a été artificiellement fixée en tirant un allèle au hasard.

Ces individus entièrement séquencés, ont permis d'identifier 27.960.164 SNPs, soit une densité de 1,56 SNPs tous les 100 paires de bases. La distribution de ces SNPs est relativement homogène sur tout le génome du mil (*TABLE 1*)

TABLE 1 – Densité de marquage SNPs

Chromosome	Nombre de SNPs	Densité SNPs
Chr 1	4.482.530	
Chr 2	4.083.817	
Chr 3	4.894.224	
Chr 4	3.690.165	
Chr 5	2.558.102	
Chr 6	4.480.954	
Chr 7	3.770.372	

## Méthodes

### Détection de la sélection avec une méthode haplotypique

Pour un locus sous balayage sélectif fort à partir d'une nouvelle mutation (Hard selective sweep), on s'attend à ce que l'haplotype bénéfique soit le plus fréquent dans la population et donc il présentera une homozygotie haplotypique forte (proche de 1). Lors que l'adaptation se fait à partir de la variabilité existante (Soft selective sweeps), deux ou plusieurs haplotypes de grandes fréquences sont attendus dans la population ([Hermisson and Pennings 2005](#)).

Afin de détecter les traces récentes de sélections douces sur notre population de mil, nous utiliserons la méthode H12 développée par [Garud et al. 2015](#).

La méthode H12 définit et calcule 3 valeurs :

- H1 qui est une estimation de l'homozygotie haplotypique d'une région donnée

$$H1 = \sum_{i=1}^n pi^2$$

*pour une population contenant n haplotypes, pi est la fréquence du i<sup>eme</sup> haplotype le plus fréquent*

- H12 qui est une estimation de l'homozygotie haplotypique après avoir combiné les fréquences des deux haplotypes les plus fréquents de cette région considérés comme si ils étaient un seul et même haplotype.

$$H12 = (p1 + p2)^2 + \sum_{i>2} pi^2 = H1 + 2p1p2$$

*pour une population contenant n haplotypes, pi est la fréquence du i<sup>e</sup>me haplotype le plus fréquent*

- H2 qui est une estimation de l'homozygote haplotypique après avoir enlevé du calcul la fréquence de l'haplotype le plus fréquent de cette région considérée.

$$H2 = \sum_{i>1} pi^2$$

*pour une population contenant n haplotypes, pi est la fréquence du i<sup>e</sup>me haplotype le plus fréquent*

Pour un locus sous balayage sélectif fort à partir d'une nouvelle mutation (*Hard selective sweep*), on s'attend à ce que l'haplotype bénéfique soit le plus fréquent dans la population et donc il présentera une homozygote haplotypique forte (proche de 1). Lors que l'adaptation se fait à partir de la variabilité existante (*Soft selective sweeps*), deux ou plusieurs haplotypes de grandes fréquences sont attendus dans la population (Herisson and Pennings 2005).

Pour différencier *Soft* et *Hard selective sweep*, il est proposé de contraster deux statistiques  $H12$  et  $\frac{H2}{H1}$  (Garud et al. 2015) [4].

Nous avons calculé ces différentes statistiques utilisant un module de python Selectionhapstat. Nous avons réalisé ces calculs sur une fenêtre de 400 SNPs. Ce qui correspond environ à 25kb<sup>b</sup>. L'analyse est répétée sur des fenêtres glissantes de 50 SNPs (github.com/Tawfekh/Mes\_Codes\_Stage\_M2).

## Détermination d'un seuil de significativité

Pour déterminer, à partir de quelle valeur de ces statistiques nous sortons d'un attendu "neutre", nous avons réalisé une simulation d'une diversité haplotypique attendue. Pour cela, nous avons

- i réalisé une modélisation de scénarios démographique de la domestication du mil ;
- ii estimé un taux de recombinaison ;
- iii réalisé des simulations avec ces scénarios et les paramètres de ces modèles pour déterminer un seuil de significativité.

**Modélisation de scénarios démographique de la domestication du mil** Tout d'abord, 4 scénarios démographiques ont été considérés.

1. le premier modèle (modèle 1) renseigne sur l'histoire démographique de la domestication du mil avec une phase de *Bottleneck*. Cette phase de goulot d'étranglement est suivie d'une phase

b. Varshney et al ont observé une décroissance rapide du déséquilibre de liaison jusqu'à  $r^2 = 0,2$  dans une fenêtre de 0.5kb (Varshney et al. 2017)

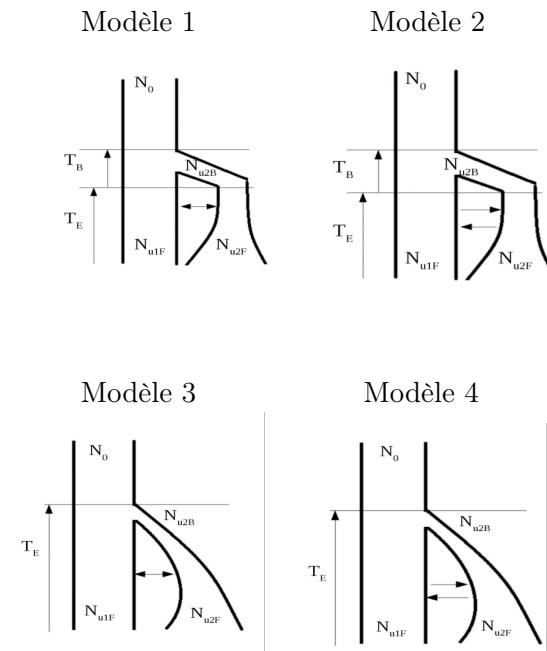


FIGURE 10 – Scénarios démographiques étudiés

Le modèle 1 renseigne sur l'histoire démographique de la domestication avec une phase de *Bottleneck*. Cette phase de goulot d'étranglement est suivie d'une phase de croissance démographique exponentielle de la population cultivée, avec des migrations par génération uniformes entre la population sauvage et celle cultivée. Le modèle 2 renseigne la même histoire que celle du modèle 1 à la différence que pendant la phase de croissance exponentielle les migrations entre populations sauvages et cultivées sont différentielles. Le modèle 3 décrit la même histoire que celle du modèle 1 à la différence qu'il y a une croissance exponentielle immédiate juste après le *bottleneck*. Le modèle 4 renseigne la même histoire que celle du modèle 3 à la différence que pendant la phase de croissance exponentielle les migrations entre la population sauvage vers celle cultivée d'une part et les migrations inverses d'autre part sont inégales.

de croissance démographique exponentielle de la population cultivée, avec des migrations uniformes entre la population sauvage et cultivée.

2. le deuxième modèle (modèle 2) renseigne la même histoire que le modèle 1 à la différence que pendant la phase de croissance exponentielle les migrations entre populations sauvages et cultivées sont inégales.
  3. le troisième modèle (modèle 3) décrit la même histoire que celle du modèle 1 à la différence qu'il y a une croissance exponentielle immédiate juste après le *bottleneck*. Ce modèle 3 a été utilisé par Cloutault et al 2012 (Cloutault et al. 2012).
  4. enfin le quatrième modèle (modèle 4) renseigne la même histoire que le modèle 3 à la différence que pendant la phase de croissance exponentielle les migrations entre la population sauvage vers celle cultivée d'une part et les migrations inverses d'autre part sont inégales(*Figure 10*).

Pour estimer les paramètres de ces modèles, nous avons utilisé une méthode basée sur une approche d'approximation en utilisant la théorie de la diffusion  $\delta a\delta i$  (Gutenkunst et al., 2009; Kimura, 1964). Voir [github.com/Tawfekh/Mes\\_Codes\\_Stag](https://github.com/Tawfekh/Mes_Codes_Stag) M2/).

Une distribution à priori de paramètres de chaque scénarios démographique est proposé (priors, TABLE 2). Nous avons donc les paramètres : taille efficace de la population ancestrale ( $N_0$ )

TABLE 2 – Distributions à priori des paramètres démographiques utilisées durant l’analyse  $\delta a\delta i$

Paramètres	Unité	Prior modèle 1	Prior modèle 2	Prior modèle 3	Prior modèle 4
$N_{u1}F$	$N_o$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$
$N_{u2}F$	$N_o$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$
$N_{u2}B$	$N_o$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$	$[1.10^{-2}; 100]$
$M$	$4N_o m$	$[0; 10]$	-	$[0; 10]$	-
$M_{12}$	$4N_o m_{12}$	-	$[0; 10]$	-	$[0; 10]$
$M_{21}$	$4N_o m_{21}$	$[0; 10]$	-	$[0; 10]$	-
$T_B$	$2N_o$	$[0; 3]$	$[0; 3]$	-	-
$T_E$	$2N_o$	$[0; 3]$	$[0; 3]$	$[0; 3]$	$[0; 3]$
pol error	-	$[0; 0, 1]$	$[0; 0, 1]$	$[0; 0, 1]$	$[0; 0, 1]$

Ces paramètres sont définies par rapport à la population ancestrale  $N_0$ .  $N_{u1}F$  désigne la taille efficace de la population de mil sauvage,  $N_{u2}F$  représente la taille efficace de la population de mil cultivé,  $N_{u2}B$  la taille efficace de la population de mil cultivé au moment du Bottleneck.  $M = 4N_o m$ ,  $m$  étant la proportion de chaque population (cultivée et sauvage) composée de nouveaux migrants à chaque génération (pour les modèles 1 et 3). Pour les modèles 2 et 4,  $M_{12} = 4N_o m_{12}$ , où  $m_{12}$  représente la fraction de la population sauvage qui est composée de migrants de la population cultivée à chaque génération alors que  $M_{21} = 4N_o m_{21}$ ,  $m_{21}$  désignant le taux de migration par génération dans le sens inverse.  $T_B$  désigne le temps en nombre de génération qu’a duré Bottleneck lors de la domestication du mil tandis que  $T_E$  est le temps en nombre de génération qu’a duré la phase de croissance exponentielle des mil cultivés après le goulot d’étranglement. Enfin pol error est le taux d’erreur de polarisation de nos données. (voir Figure 10)

la population ancestrale ( $N_0$ ) ; taille efficace de la population sauvage ( $N_1F$ ) ; taille efficace de la population cultivée ( $N_2F$ ) et taille efficace de la population cultivée durant le bottleneck ( $N_2B$ ) qui sont communs à l’ensemble des modèles. Et des paramètres spécifiques d’un modèle comme  $M$  qui est égale à  $M = 4N_o m$ , avec  $m$ , la fraction de chaque population (cultivée et sauvage) composée de nouveaux migrants à chaque génération, pour les modèle 1 et 3 . Pour les modèles 2 et 4, nous avons aussi  $M_{12} = 4N_o m_{12}$ ,  $m_{12}$  représentant la fraction de la population sauvage qui est composée de migrants de la population cultivée à chaque génération et  $M_{21} = 4N_o m_{21}$ , où  $m_{21}$  est la même portion de migration par génération dans le sens inverse.  $T_B$  désigne le temps en nombre de génération qu’a duré Bottleneck lors de la domestication du mil (modèle 1 et 2 uniquement) tandis que  $T_E$  est le temps en nombre de génération qu’a duré la phase de croissance exponentielle des mil cultivés après le goulot d’étranglement. Nous avons aussi inclus un paramètre d’erreur de polarisation des SNPs (voir partie Résultat). Une analyse par maximum de vraisemblance permet l’estimation de ces paramètres à partir des données réelles. Nous avons construits un spectre de fréquence en utilisant les lignées de mil cultivé et de 25 accessions de mil sauvages (*Pennisetum glaucum subsp. monodii*) provenant de tout le sahel (Sénégal, Mali, Mauritanie, Niger, Soudan et Tchad).

Nous avons utilisé les 4.486.566 SNPs obtenus après polarisation de nos 215 séquences (190 cultivé et 25 sauvages) en utilisant un "outgroup" *Pennisetum pedicellatum* [Chemisquy et al. 2010] [Veldkamp (2014)]. Nous avons cependant inclus dans notre modèle aussi un taux d’erreur de polarisation de nos SNPs.

Nous avons calculé les SFS à partir d’un code R(Philippe Cubry). Les priors ont été fixés pour chaque modèle et paramètres (Table2)

Pour comparer nos 4 modèles, nous avons un calcul du maximum de vraisemblance. Nous pourrons alors calculer des valeurs du critère d'information d'Akaike (*AIC*) (Akaike, Hirotugu. 2011, Wikipedia :Critère d'information d'Akaike. 2017) pour chaque scénario. Le modèle démographique avec la plus faible valeur d'*AIC* sera celui qui explique le mieux nos données observées sur notre population de mil.

$$AIC = 2k - 2 \log(L)$$

$k$  est le nombre de paramètre du scénario démographique.  $L$  étant la valeur du maximum de vraisemblance.

**Estimation du taux de recombinaison par génération dans la population de mil cultivé  $\rho$  ( $4N_e c$ )** Pour estimer  $\rho$ , nous avons utilisé "Estimation of Population Recombination rates" (EPRR) qui est une méthode basée sur la modélisation et le machine learning (Lin, Futschik and Li 2013). Cette méthode est efficace pour des données génomiques et est aussi précise que des méthodes basées sur la vraisemblance (Lin, Futschik and Li 2013).

La méthode EPRR est implémenté dans le package R (RStudio version 1.0.153) nommé FastE-PRR(version 1.0) (Gao et al. 2016). Nous avons appliqué cette méthode sur 15 fragments de 100kb pris au hasard sur chacun des 7 chromosomes. Pour un total de 105 fragments. Le  $\rho$  est estimé par fragment de 100kb et son unité est  $4Nec \times L$  (ici  $L = 100kb$ )

Pour avoir un  $\rho$  global pour tout le génome nous avons pris la moyenne de la valeur de  $\rho$  sur les 105 fragments (voir [github.com](#)). Ces données seront comparées à des données déjà publiés (Clotault et al. 2012)

**Simulation de séquences de mil cultivé pour estimer un seuil de détection des gènes sous sélection** La simulation de nos scénarios est faite en utilisant *ms* (Hudson 2002). Ce programme est implémenté sous *R* (RStudio version 1.0.153) dans le package *phyclust* (version 0.1-22). Nous avons simulé 10000 fois des séquences de 100kb de la population cultivée. Les paramètres démographiques et un  $\rho$  (soit le  $\rho$  estimé dans cette étude, soit le  $\rho$  estimé par Clotault et al 2012) sont utilisés dans ces simulations.

Le DL sur les données simulées est calculé en estimant un  $r^2$  (VanLiere and Rosenberg 2008) avec le programme *PLINK*(version 1.07) (Purcell et al. 2007). Nos différentes statistiques  $H12$  sont calculées à partir de chacune de nos 10000 simulations pour simuler un attendu neutre. Afin de construire un seuil de significativité, nous avons considérer que pour les 10000 simulations faites sous modèle neutre, 0,01% de nos haplotypes simulés de 400SNPs sortent de l'attendu neutre, ce qui nous donne un taux de fausse découverte (FDR) de  $1.10^{-4}$  sur nos simulations.

## **Identification des gènes candidats à la sélection et de leurs fonctions**

Nous avons identifier les gènes candidats à la sélection en comparant leurs positions avec celles des haplotypes qui sortent d'un attendu "neutre". Nous n'avons sélectionné que les gènes dont tout ou une partie est inclus dans un intervalle de 50kb autours des haplotypes candidats.

Il n'a été considéré que les gènes annotés du génome du mil. Nous avons ensuite cherché un enrichissement des termes de *Gene Ontology* par des tests de Fisher(Voir [github.com/Tawfekh/Mes\\_Codes\\_Stage\\_M2/](https://github.com/Tawfekh/Mes_Codes_Stage_M2/)). Pour cette analyse nous avons utilisé le Package *R topGo version 2.32.0* (Alexa, Rahnenführer and Lengauer 2006).

# Résultats

Les analyses qui suivent ont été effectuées de deux manières différents, d'une part en utilisant que sur les individus de mils sauvages et cultivés issus du centre du Sahel et d'autre part, en prenant tous les individus de mils cultivés et sauvages de notre jeux de données.

Un *scan* entier du génome a été effectué afin de détecter les traces récentes de sélections fortes et douces sur notre population de mil en utilisation la statistique H12 (Figure 11). Plusieurs pics de valeur extrême de la statistique sont observés (Figure 11). Des analyses supplémentaires ont été faites pour déterminer, à partir de quelle valeur de statistiques H12 nous pouvons supposer qu'une région est sous sélection.

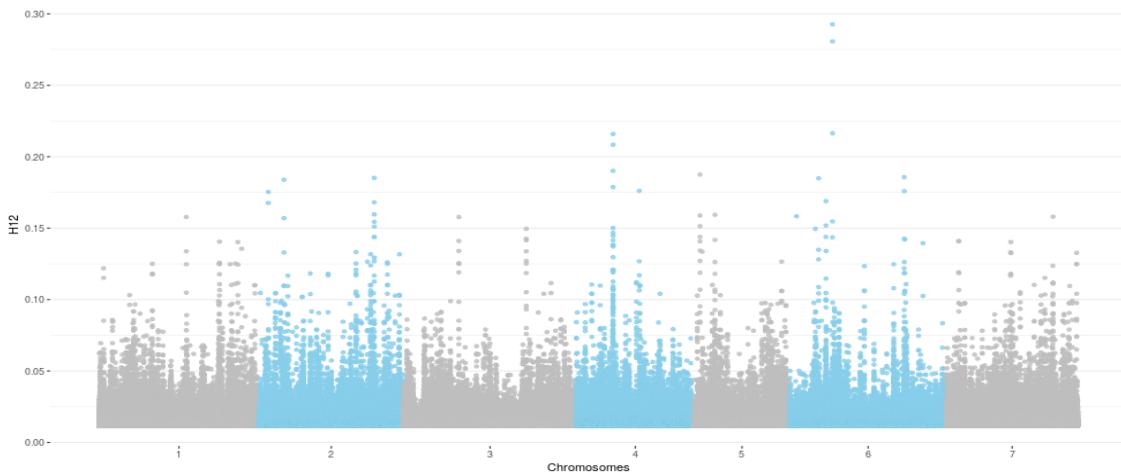


FIGURE 11 – Scan du génome avec la statistique H12

Chaque point représente un haplotype. En abscisse sont représentés les chromosomes du mil. Pour ce scan nous n'avons utilisé que les données de mils du centre du Sahel.

## Détermination d'un seuil de significativité

### Modélisation démographique neutre de la domestication du mil

La modélisation de l'histoire démographique du mil a été faite d'une part, avec l'intégralité les données de mils et d'autre part en ne prenant que les données de mils échantillonnes au centre du Sahel.

Comme nous avons utilisé des données polarisées pour modéliser nos scénarios démographiques, il est donc possible que nous ayons des erreurs de polarisation des données. Mis à part l'effet d'échantillonnage, nous avons donc implémenté la modélisation démographique de deux manières :

i sans prendre en compte une erreur de polarisation (Annexe A) ;

ii et en prenant en compte une erreur de polarisation (Table 3, Annexe A).

Pour chaque scénario démographique, la modélisation faite en échantillonnant que les individus issus du Centre du Sahel a un AIC plus faible que si cette modélisation est réalisée en prenant tous

les individus de mils cultivés (Figure 12, Annexe C). Dans chacun de ces deux cas, la prise en compte d'une erreur de polarisation a entraîné une amélioration de l'ajustement du modèle à nos données (Figure 12, Annexe C).

Parmi toutes nos implémentations de scénarios démographiques, celle prenant en compte une migration différentielle entre la population sauvage et cultivée, et qui atteste d'un temps de *bottleneck* suivi d'une expansion démographique de la population cultivée (Modèle 2, mils du centre du Sahel, prise en compte d'une erreur de polarisation) a eu la plus faible AIC. Donc s'ajuste le mieux à nos données (Figure 12, Annexe B).

Sur ce modèle, le taux de mutation par génération a été estimé à  $\theta_{anc} = 4N_o\mu = 503, 23 \cdot 10^{-5}$  par pb pour la population sauvage ancestrale, de  $\theta_{wild} = 4N_{u_1F}\mu = 86, 55 \cdot 10^{-5}$  par pb pour la population sauvage actuelle et de  $\theta_{cult} = 4N_{u_2F}\mu = 161, 53 \cdot 10^{-5}$  par pb pour la population cultivée actuelle. L'intensité du *bottleneck* lors de la domestication a été estimée à 3%. Le modèle a révélé aussi un taux de migration par génération ( $m = \frac{M}{2N_o}$ ) qui est différent entre la population sauvage et cultivée. La fraction de migrants vers la population sauvage est 2,5 fois plus élevée que celle allant dans le sens inverse (*i.e.* population sauvage vers population cultivée). Le temps écoulé depuis la domestication du mil a été estimé à  $0,612 \times 2N_o$  générations, avec une courte durée de *bottleneck* ( $0,009 \times 2N_o$  générations) et une phase de croissance exponentielle 67 fois plus long que celui du goulot d'étranglement (Table 3). Pour cette implémentation l'erreur de polarisation a été estimée à 7%.

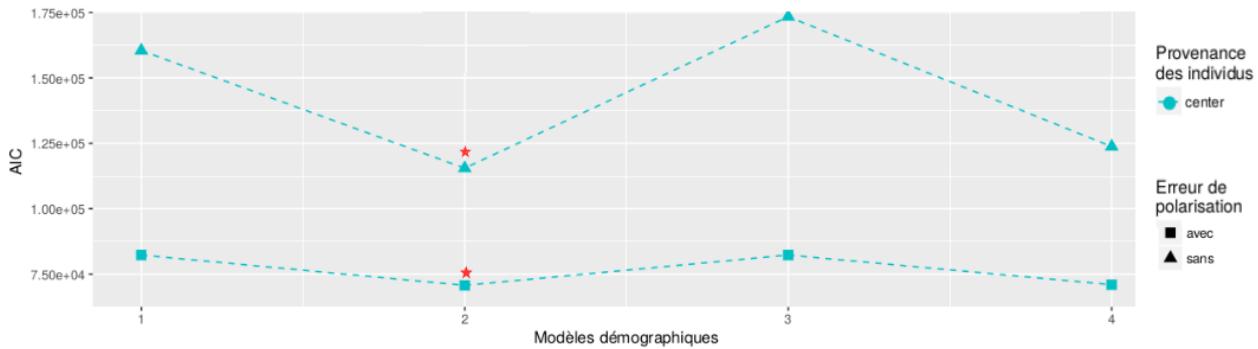


FIGURE 12 – Résultat des modélisations démographiques faite en prenant que les mils du centre du Sahel

En abscisse les chiffres désignent les numéros de modèle (voir figure 10). En ordonné les AICs de chaque scénario démographique inféré. Les carrés désignent les modélisations faites en prenant en compte une erreur de polarisation. Les triangles désignent les modélisations faites en ne prenant pas en compte une erreur de polarisation

### Estimation du taux de recombinaison ( $\rho_{cult}$ )

Les paramètres démographiques ont été estimés grâce à une méthode utilisant la théorie de la diffusion et posant comme hypothèse l'indépendance entre locus. Donc la méthode ne prends pas

TABLE 3 – Tableau des maximums de vraisemblance des paramètres démographiques obtenues après l’analyse  $\delta a\delta i$

Paramètres	Unité	Centre Sahel			
		scénario 1	scénario 2	scénario 3	scénario 4
$-1 * lik$	-	82268.02	<b>70671.60</b>	82276.14	<b>70923.47</b>
$AIC$	-	82282.02	<b>70687.60</b>	82288.14	<b>70937.47</b>
$\theta$	$4N_o\mu$	0.0025.6	0.00503	0.0016	0.0023
$N_u1F$	$N_o$	0.446	0.172	0.705	0.379
$N_u2F$	$N_o$	0.602	0.321	0.953	0.707
$N_u2B$	$N_o$	0.013	0.030	0.096	0.129
$M$	$2N_o m$	1.881	-	1.178	-
$M_{12}$	$2N_o m_{12}$	-	6.245	-	2.821
$M_{21}$	$2N_o m_{21}$	-	2.463	-	1.146
$T_B$	$2N_o$	0.0002	0.009	-	-
$T_E$	$2N_o$	1.425	0.603	1.356	0.876
<i>pol error</i>	-	0.0810	0.0709	0.0837	0.0719

Ces paramètres sont définis par rapport à la population ancestrale  $N_0$ .  $N_{u1}F$  désigne la taille efficace de la population de mil sauvage,  $N_{u2}F$  représente la taille efficace de la population de mil cultivé,  $N_{u2}B$  la taille efficace de la population de mil cultivé au moment du Bottleneck.  $M$  étant le taux de migration par génération entre la population cultivée et sauvage (pour les modèles 1 et 3). Pour les modèles 2 et 4,  $M_{12}$  représente le taux de migration par génération de la population sauvage vers la population cultivée alors que  $M_{21}$  représente le taux de migration par génération dans le sens inverse.  $T_B$  désigne le temps en nombre de génération qu'a duré Bottleneck lors de la domestication du mil tandis que  $T_E$  est le temps en nombre de génération qu'a duré la phase de croissance exponentielle des mil cultivés après le goulot d'étranglement. Enfin *pol error* est le taux d'erreur de polarisation de nos données. (voir Figure )

en compte le DL. L'effet du DL a été donc estimé. En utilisant une approche de *machine learning*, une distribution du taux de recombinaison par génération sur 100kb de la population cultivée de mil a été estimée pour chaque chromosomes (Figure 13, Annexe D). Pour l'ensemble des chromosomes letaux de recombinaison par génération estimé est compris entre 400 et 600 sur 100kb. La moyenne de la distribution du génome, qui prend en compte celle obtenue pour chaque chromosome, a été choisie comme une estimation du taux de recombinaison moyen ( $\rho_{A_{cult100kb}}$ ) du mil sur 100 fenêtres de 100kb.  $\rho_{A_{cult100kb}} = 484,42$  avec un écart-type de  $\sigma_{A_{cult100kb}} = 63.09$  (Figure). D'où un taux de recombinaison unitaire par génération de  $\rho_{A_{cult}} = 4,8442 \cdot 10^{-3}$

## Simulation de séquences

Les paramètres démographiques issu du modèle 2 fitté à partir des mils du centre du Sahel en prenant en compte une erreur de polarisation ( $N_{u1}F$ ,  $N_{u2}F$ ,  $N_{u2}B$ ,  $M_{12}$ ,  $M_{21}$ ,  $T_B$  et  $T_E$ ), le taux de mutation par génération ( $\theta_{cult}$ ) et d'une part le taux de recombinaison estimé ( $\rho_{A_{cult}}$ ) et d'autre part le taux de recombinaison estimé par Clotault et al. 2012 ( $\rho_C$ ) ont été utilisés pour simuler à l'aide de la coalescence :

- a) 88 séquences sous hypothèse neutre *ie* sans sélection ;
- b) 190 séquences sous hypothèse neutre *ie* sans sélection.

Le DL entre deux nucléotides en fonction de leur distance sur 100kb (*LD decay*) des séquences de mils a été comparé avec celui des séquences simulées (Figure 13).

Le *LD decay* calculé à partir de l'intégralité de notre échantillon de mils cultivées a décris de 0,366 (à 1pb) à 0,1 (à 100kb). Celui calculé à partir de nos mils issus du centre du Sahel maintient

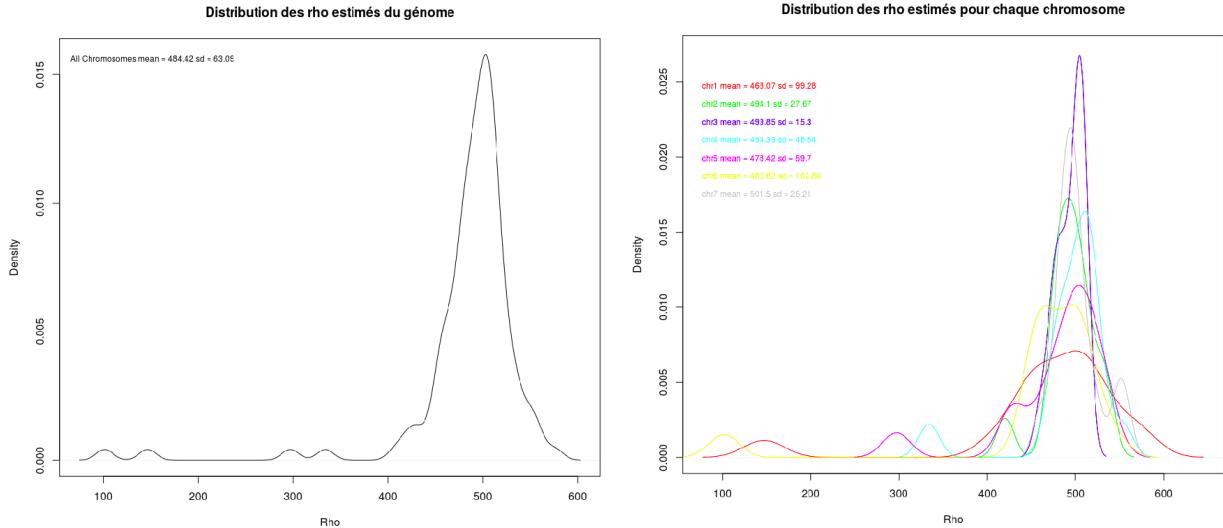


FIGURE 13 – Distribution des estimations du taux de recombinaison par génération  
 À droite nous avons les distributions des estimation de  $\rho_{A_{cult}100kb}$  pour chaque chromosome. La figure de gauche montre la distribution de l'estimation du taux de recombinaison pour tout le génome

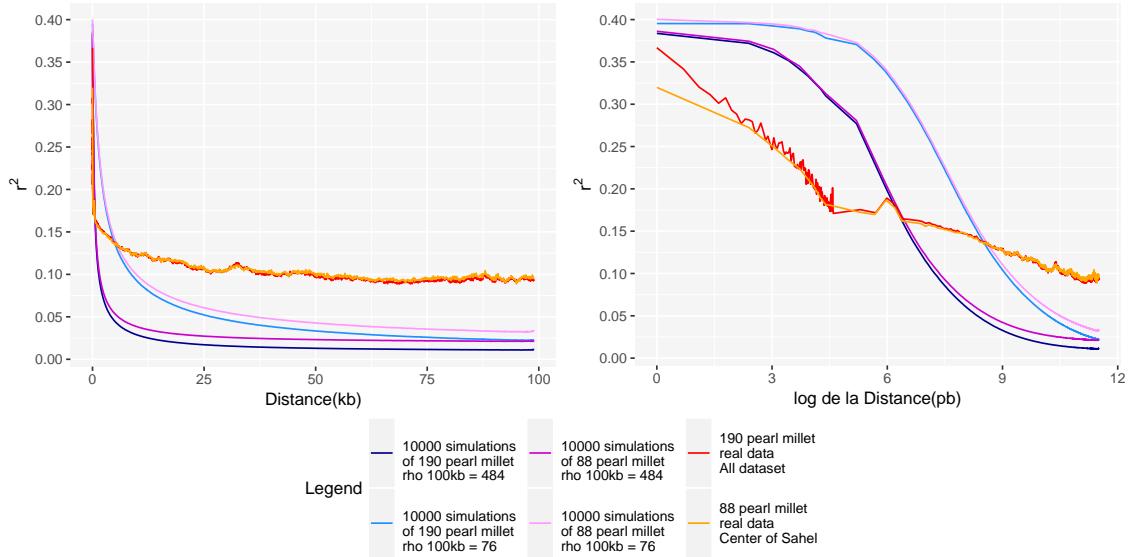


FIGURE 14 – Estimation du DL entre deux nucléotides en fonction de leur distance

la même tendance (0,319 à 1pb et 0,1 à 100kb). Alors que le *LD decay* des séquences simulées avec  $\rho_{A_{cult}100kb}$  sur 100kb ont des valeurs à 1pb de 0,38 pour les simulations de 190 séquences et de 0,386 pour celles de 88 séquences. Le LD decay à 1pb est respectivement de 0,01 et de 0,02. Les séquences simulées avec  $\rho_C$  sur 100kb ont des valeurs de *LD decay* à 1pb de 0,39 pour les simulations de 190 séquences et de 0,4 pour celles de 88 séquences. Le LD decay à 1pb est respectivement de 0,02 et de 0,03.

La décroissance du DL sur 100kb les données de mils cultivés est plus rapide que celles simulées sous modèle neutre. Cependant les *LD decay* de nos séquences simulées

La comparaison de la valeur du *LD decay* à 25kb, qui correspond à 400SNPs en espérance, par rapport à sa valeur sur une distance de 100kb montre :

- le *LD decay* calculé à partir de nos mils issus du centre du Sahel et celui calculé à partir de l'intégralité de notre échantillon de mils cultivées ont décrus de 93% .
- Les *LD decay* de nos séquences simulées ont la même tendance, celles modélisées avec  $\rho_{A_{cult}100kb}$  ont des *LD decay* qui ont décrus de 98%. Tandis que les séquences modélisées avec  $\rho_C$  ont des *LD decay* qui ont diminués de 91%

La figures 14 suggère que les fenêtres de 100kb sont suffisamment grandes pour que la démographie neutre ne génère pas par hasard des valeurs élevées de DL et augmente l'homozygotie de type haplotype, et devrait donc empêcher un taux élevé de faux positifs.

Nous avons aussi comparé la distribution des H12 calculés à partir de nos séquences simulées correspondant à un attendu neutre et ceux de nos données réelles.

La comparaison(Figure 15) montre que les valeurs de H12 à l'échelle du génome dans les données de mils cultivés sont considérablement plus élevées que ce à quoi on s'attendrait pour notre modèle démographique neutres. Les séquences modélisées avec  $\rho_C$  ont des distribution de H12 qui son respectivement plus hautes que celles simulé avec  $\rho_{A_{cult}100kb}$ .

Par ailleurs, nous remarquons aussi un effet d'échantillonnage sur l'homozygotie haplotypique car les simulations faites avec 190 individus ont des distributions de H12 respectivement plus basses que celle faite en modélisant 88 individus.

De plus, il y a une longue suite de valeurs aberrantes H12 dans nos données réelles de mils cultivés, ce qui suggère des balayages sélectifs récents(Figure 15).

## Détection des signatures de sélection forte et douce

Pour la détection de la sélection, nous avons choisi comme attendu neutre les simulations des 88 séquences faites avec les paramètres démographiques issus du modèle 2 fitté à partir des mils du centre du Sahel en prenant en compte une erreur de polarisation ( $N_u1F$ ,  $N_u2F$ ,  $N_u2B$ ,  $M_{12}$ ,  $M_{21}$ ,  $T_B$  et  $T_E$ ), le taux de mutation par génération ( $\theta_{cult}$ ) et le taux de recombinaison ( $\rho_{A_{cult}}$ )

Le seuil de significativité choisi est  $H12_o = 0.032$  correspondant à un taux de fausse découverte (FDR) de  $1.10^{-4}$  sur nos simulations (Figure 15 et 16).

La figure montre le scan génomique obtenu en appliquant le seuil de significativité  $H12_o$  à nos statistiques H12 calculés à partir des mils échantillonnés au centre du Sahel(Figure 15 et 16).

Nous avons détecté 13.838 fenêtres sur les 559.152 calculées, correspondant à 2,5% des Haplotypes à partir desquels les Homozygosités H12 ont été calculées. Ces haplotypes détectés correspondent à 6,9% (2606 gènes) des 37617 gènes annotés du génome du mil.

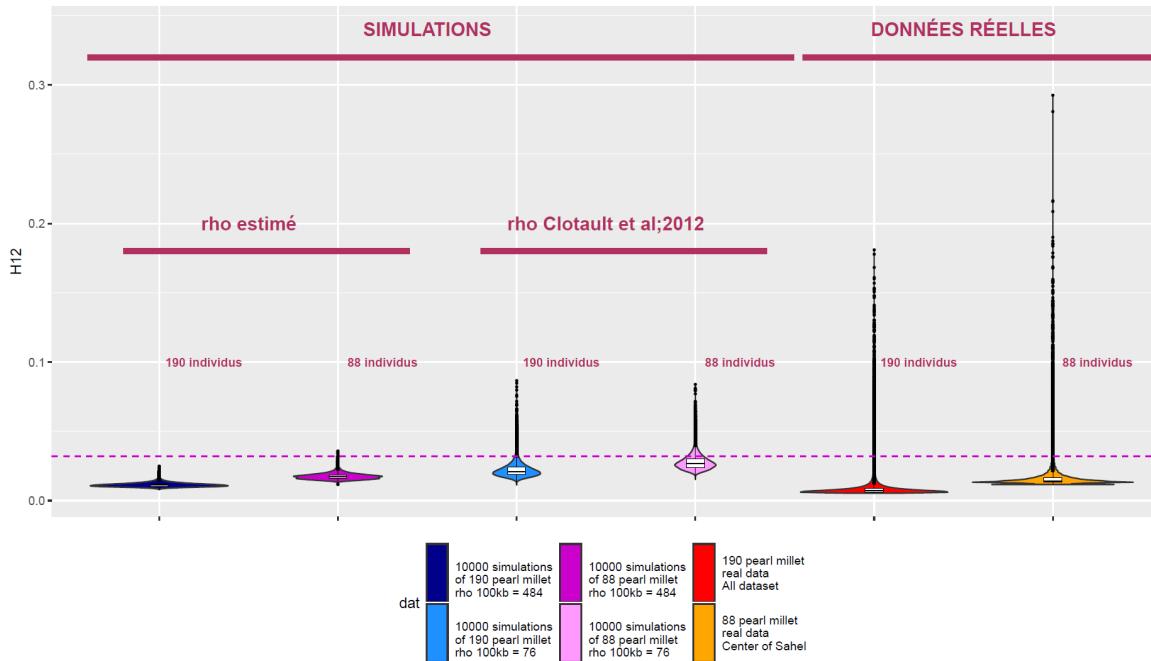


FIGURE 15 – Distributions de H12 calculés à partir de nos séquences simulées et de nos données réelles

Les deux distributions à droite représentent celles des H12 calculés à partir de nos données réelles : celui de 190 individus correspond à la totalité de nos données sur le mil tandis que celle de 88 individus représente nos données de mils du centre du Sahel.

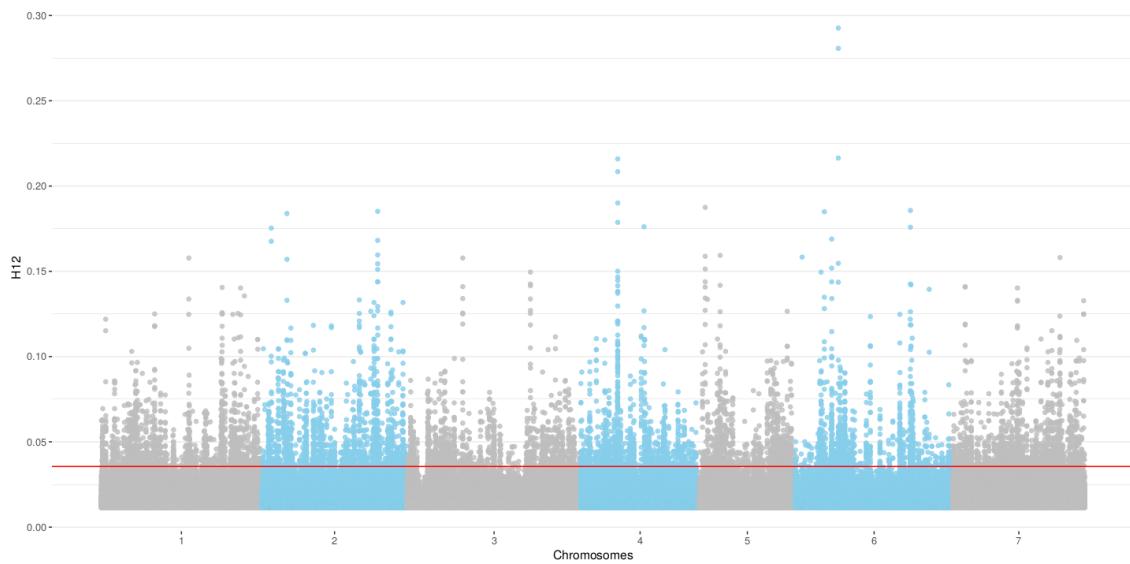


FIGURE 16 – Scan du génome avec la statistique H12 avec le seuil de significativité  
Chaque point représente un haplotype. En abscisse sont représentés les chromosomes du mil. Pour ce scan nous n'avons utilisé que les données de mils du centre du Sahel. Le trait rouge représente notre seuil de significativité

## Analyse *GO terms* des gènes détectés

L'analyse d'enrichissement en termes de GO a révélé pour les composants cellulaires, un enrichissement en gènes liés à la structure cellulaire impliquée dans le transport de molécules. Par exemple nous avons détecter un enrichissement en gènes liés aux vésicules *COP1* (GO :0030126, GO :0030137 et GO :0030663) et d'autres vésicules du cytoplasme et de l'empilement golgien (GO :0031410, GO :0005798, GO :0031982 et GO :0097708). Mais aussi à des structures cellulaires comme les plastides et le chloroplaste (GO :0009536 et GO :0009507).

Pour les processus biologiques, nos gènes détectés sont liés au métabolisme cellulaire et au transport de substances appelées *drugs en anglais*. Par exemple nos gènes candidats à la sélection participent significativement au métabolisme des acides aminés (GO :0006520 et GO :0043038), des macromolécules comme l'ADN ou l'ARN (GO :0033865, GO :0033875, GO :0034032, GO :0006259, GO :0034660, GO :0006418 et GO :0043039), au transport et réponse cellulaire aux *drugs* (GO :0006855, GO :0015893 et GO :0042493). Nos gènes détectés participent aussi aux processus régulations du cycle cellulaire (GO :0051726).

Pour les fonctions moléculaires, les gènes candidats à la sélection participent le plus significativement aux activités de phosphatase acide(GO :0003993). Mais aussi aux activités liées à l'amidon (GO :2001070), à la production d'énergie ATPasique(GO :0016887 et GO :0042623), au transport cellulaire de *drugs*(GO :0022804, GO :0015238 et GO :0090484) et de métabolisme d'ARN (GO :0034062, GO :0097747 et GO :0140098).

# Discussion

Nous avons choisi de modéliser des scenarii démographiques de deux manières différents, d'une part en nous basant que sur les individus de mils sauvages et cultivés issus du centre du Sahel et d'autre part modélisation est réalisée en prenant tous les individus de mils cultivés. Car il a été démontré que le mil cultivé a une origine monophylétique que sont les mils sauvage du centre du Sahel (Oumar et al. 2008, Burgarella et al. 2018). Et que les mils cultivés de l'est et de l'ouest du Sahel ont subi des flux de gènes avec leur apparentés sauvages de ces zones ce qui peut conduire à une différenciation plus prononcée entre ces population de mil cultivé et leurs ancêtres sauvage du Sahel (Burgarella et al. 2018). Donc à terme ce fait pourrait biaiser les estimations des paramètres démographiques de nos modèles. Les résultats de nos modélisations démographique vont dans le sens de cette hypothèse.

Parmi nos scenarii démographiques ceux qui assument une migration différentielle entre la population de mil sauvage et celle cultivée sont plus vraisemblables que ceux qui attestent d'une migration uniforme. Ce résultat suggère que le flux de gènes entre les populations de mils cultivés et sauvages a un impact non négligeable dans le façonnement de la diversité du mil cultivé. En effet (Mariacet al. 2006) ont estimé chez les populations de mil cultivé du Niger 2,8 fois plus d'introgression d'allèles cultivés dans la population de mil sauvage que d'allèles sauvages dans la population de mil cultivé. Ce qui est proche de notre estimation de taux de migration de la population cultivée vers la population sauvage qui est 2,5 fois plus élevé que celle allant dans le sens inverse (*i.e.* population sauvage vers population cultivée). Ce taux migration différentiel entre Mil cultivé et mil sauvage peut s'expliquer par les barrières reproductives qui existent entre ces deux populations comme en atteste l'étude de Amoukou and Marchais 1993 qui a montré que des croisements entre les pollens de mil sauvage et des ovules de mils cultivés donnaient des grains mal-formées associées à une faible capacité germinative et d'une diminution de leurs poids. Tandis que le croisement inverse donnaient des graines de poids plus faible mais avec des capacités de germination similaires à celle des mils cultivés. De surcroît la période de floraison plus longue des mils sauvages par rapport à celle des mils cultivés peut causer ce flux de gènes asymétrique.

Le modèle 2 (Modèle 2, figure10)<sup>c</sup> estime une intensité de bottleneck (IB de 3% de la taille efficace de la population sauvage au début de la domestication. Ce résultat est assez similaire à ceux qui a été obtenu chez le maïs (5% d'IB, Beissinger et al. 2016), le soja dont l'IB a été estimé à 3,6% de la taille de la population sauvage (Guo et al. 2010).

Nos analyses nous ont permis de détecter 2606 gènes supposés être sous sélection forte ou douce. Tandis que (Burgarella et al. 2018) ont détecté 215 gènes considérés comme étant sous *Hard selective sweep* soit 12 fois moins que dans cette étude. Chez la téosinte des analyses similaires ont été faites

c. qui prend en compte une migration différentielle entre la population sauvage et cultivée, et qui atteste d'un temps de bottleneck suivi d'une expansion démographique de la population cultivé

montrant 30 fois moins de gènes exclusivement sous *Hard selective sweep* que de gènes sous balayage fort ou doux (Lorant 2018). Cette proportion se situe entre 8 et 16.5 chez la drosophile. Ces résultats rejoignent l'hypothèse que les sélections sur la diversité existante (soft sweep) seraient le mode d'adaptation prépondérant des organismes (Hermisson and Pennings 2005, Hermisson J & Pennings PS ; 2017, Schrider and Kern 2017)

Nos gènes candidats à la sélection sont significativement enrichis en terme d'activité de phosphatase acide. Cette activité entre en jeu dans le processus d'absorption du phosphore chez la plante par hydrolyse du phosphore organique afin de le rendre biodisponible (Tarañdar and Jungk 1987, Nannipieri et al. 2011). Ce résultat suggère une adaptation des mils cultivés à la faible disponibilité en phosphore des sols du Sahel (Hash, Schaffert and Peacock 2002).

Parmi les gènes détectés, nous avons le gène *pif3* qui code pour un facteur suspecté d'interagir avec les protéines phytochrome (PHYA et PHyB en particulier), cette protéine PIF3 jouerai un rôle dans le contrôle du temps de floraison chez *Arabidopsis thaliana* (Oda et al. 2004) au même titre que le gène PHYC le jouera chez le mil (Saïdou et al. 2009).

## Conclusion et Perspectives

Durant ce travail, nous avons étudié la diversité fonctionnelle des plantes. En essayant de comprendre comment la diversité fonctionnelle des plantes cultivées a été façonnée durant leur histoire. En prenant comme modèle le mil (*Pennisetum glaucum*), nous avons identifié sa diversité fonctionnelle par l'analyse de la sélection à l'échelle du génome. Pour cela nous avons utilisé une méthodes basées sur la diversité haplotypique (H12). Nous avons modélisé l'histoire démographique de cette plante afin d'avoir une approximation du profil de diversité attendu en absence de sélection et de détecter avec plus de précision et de confiance les régions sous balayage sélectif. Cette étude nous a permis d'entrevoir l'importance des migrations entre la population sauvage et cultivée du mil dans le façonnement de sa diversité actuelle. Nous avons détecté chez le mil les gènes potentiellement sous sélection qui participent à des processus et fonctions biologiques importants notamment dans le transport des substances toxiques ou aux processus d'absorption du phosphore. Des analyses supplémentaires pourraient permettre de savoir si ces gènes ont permis une adaptation du mil. Par exemple on peut imaginer une analyse de génétique d'association qui aurait pour but d'étudier une corrélation entre des traits relatifs à un de ces processus et les gènes identifiés. Nous pouvons aussi explorer parmi ces gènes quelles sont ceux sous balayage sélectif fort et ceux sous balayage doux en utilisant nos modélisations et une approche de machine learning. Modéliser les Hard et des soft sweeps, et calculer des homozygotie attendu afin de les confrontés aux données réelles.

# Références

- Agte, V. V., K. V. Tarwadi and S. A. Chiplonkar. 1999. "Phytate Degradation During Traditional Cooking : Significance of the Phytic Acid Profile in Cereal-Based Vegetarian Meals." *Journal of Food Composition and Analysis* 12(3) :161–167.
- URL:** <http://www.sciencedirect.com/science/article/pii/S0889157599908268>
- Alexa, Adrian, Jörg Rahnenführer and Thomas Lengauer. 2006. "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." *Bioinformatics* 22(13) :1600–1607.
- URL:** <https://academic.oup.com/bioinformatics/article/22/13/1600/193669>
- Amoukou, A. I. and L. Marchais. 1993. "Evidence of a partial reproductive barrier between wild and cultivated pearl millets (*Pennisetum glaucum*)."*Euphytica* 67(1) :19–26.
- URL:** <https://doi.org/10.1007/BF00022720>
- Beissinger, Timothy M., Li Wang, Kate Crosby, Arun Durvasula, Matthew B. Hufford and Jeffrey Ross-Ibarra. 2016. "Recent demography drives changes in linked selection across the maize genome." *Nature Plants* 2(7) :16084.
- URL:** <https://www.nature.com/articles/nplants201684>
- Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan. 1995. "The hitchhiking effect on the site frequency spectrum of DNA polymorphisms." *Genetics* 140(2) :783–796.
- URL:** <http://www.genetics.org/content/140/2/783>
- Burgarella, Concetta, Philippe Cubry, Ndjido A. Kane, Rajeev K. Varshney, Cedric Mariac, Xin Liu, Chengcheng Shi, Mahendar Thudi, Marie Couderc, Xun Xu, Annapurna Chitikineni, Nora Scarcelli, Adeline Barnaud, Bénédicte Rhoné, Christian Dupuy, Olivier François, Cécile Berthouly-Salazar and Yves Vigouroux. 2018. "A western Sahara centre of domestication inferred from pearl millet genomes." *Nature Ecology & Evolution* 2(9) :1377–1380.
- URL:** <https://www.nature.com/articles/s41559-018-0643-y>
- Chemisquy, M. Amelia, Liliana M. Giussani, María A. Scataglini, Elizabeth A. Kellogg and Osvaldo Morrone. 2010. "Phylogenetic studies favour the unification of *Pennisetum*, *Cenchrus* and *Odon-telytrum* (Poaceae) : a combined nuclear, plastid and morphological analysis, and nomenclatural combinations in *Cenchrus*."*Annals of Botany* 106(1) :107–130.
- URL:** <https://academic.oup.com/aob/article/106/1/107/95610>
- Clotault, Jérémie, Anne-Céline Thuillet, Marylène Buiron, Stéphane De Mita, Marie Couderc, Bettina I. G. Haussmann, Cédric Mariac and Yves Vigouroux. 2012. "Evolutionary History of Pearl Millet (*Pennisetum glaucum* [L.] R. Br.) and Selection on Flowering Genes since Its Domestication."

*Molecular Biology and Evolution* 29(4) :1199–1212.

**URL:** <https://academic.oup.com/mbe/article/29/4/1199/1191863>

Fay, Justin C. and Chung-I. Wu. 2000. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155(3) :1405–1413.

**URL:** <http://www.genetics.org/content/155/3/1405>

Gao, Feng, Chen Ming, Wangjie Hu and Haipeng Li. 2016. “New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era.” *G3 : Genes, Genomes, Genetics* 6(6) :1563–1571.

**URL:** <http://www.g3journal.org/content/6/6/1563>

Garud, Nandita R., Philipp W. Messer, Erkan O. Buzbas and Dmitri A. Petrov. 2015. “Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps.” *PLOS Genetics* 11(2) :e1005004.

**URL:** <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005004>

Glémén, Sylvain and Thomas Bataillon. 2009. “A comparative view of the evolution of grasses under domestication.” *New Phytologist* 183(2) :273–290.

**URL:** <https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.2009.02884.x>

Govindaraj, M., P. Shanmugasundaram, P. Sumathi and A. R. Muthiah. 2010. “Simple, Rapid And Cost Effective Screening Method For Drought Resistant Breeding In Pearl Millet.” *Electronic Journal of Plant Breeding* 1 :590–599.

**URL:** <http://sites.google.com/site/ejplantbreeding/vol-1-4>

Gulia, S. K., J. P. Wilson, J. Carter and B. P. Singh. 2007. “Progress in grain pearl millet research and market development.” *Issues in new crops and new uses. ASHS Press, Alexandria, VA* pp. 196–203.

Guo, Juan, Yunsheng Wang, Chi Song, Jianfeng Zhou, Lijuan Qiu, Hongwen Huang and Ying Wang. 2010. “A single origin and moderate bottleneck during domestication of soybean (*Glycine max*) : implications from microsatellites and nucleotide sequences.” *Annals of Botany* 106(3) :505–514.

**URL:** <https://academic.oup.com/aob/article/106/3/505/183285>

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson and Carlos D. Bustamante. 2009. “Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data.” *PLOS Genetics* 5(10) :e1000695.

**URL:** <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000695>

Hash, C. T., R. E. Schaffert and J. M. Peacock. 2002. Prospects for using conventional techniques and molecular biological tools to enhance performance of ‘orphan’ crop plants on soils low in

available phosphorus. In *Food Security in Nutrient-Stressed Environments : Exploiting Plants' Genetic Capabilities*, ed. J. J. Adu-Gyamfi. Developments in Plant and Soil Sciences Dordrecht : Springer Netherlands pp. 25–36.

**URL:** [https://doi.org/10.1007/978-94-017-1570-6\\_4](https://doi.org/10.1007/978-94-017-1570-6_4)

Hermissen, Joachim and Pleuni S. Pennings. 2005. "Soft Sweeps : Molecular Population Genetics of Adaptation From Standing Genetic Variation." *Genetics* 169(4) :2335–2352.

**URL:** <http://www.genetics.org/content/169/4/2335>

Hill, W. G. and Alan Robertson. 1966. "The effect of linkage on limits to artificial selection." *Genetics Research* 8(3) :269–294.

**URL:** <https://www.cambridge.org/core/journals/genetics-research/article/effect-of-linkage-on-limits-to-artificial-selection/5CCFE11C1F8108242ED02AEC8BA5DD50>

Hudson, Richard R. 2002. "Generating samples under a Wright–Fisher neutral model of genetic variation." *Bioinformatics* 18(2) :337–338.

**URL:** <https://academic.oup.com/bioinformatics/article/18/2/337/225783>

Jensen, Jeffrey D., Yuseob Kim, Vanessa Bauer DuMont, Charles F. Aquadro and Carlos D. Bustamante. 2005. "Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data." *Genetics* 170(3) :1401–1410.

**URL:** <http://www.genetics.org/content/170/3/1401>

Kaplan, N. L., R. R. Hudson and C. H. Langley. 1989. "The "hitchhiking effect" revisited." *Genetics* 123(4) :887–899.

**URL:** <http://www.genetics.org/content/123/4/887>

Kelly, John K. 1997. "A Test of Neutrality Based on Interlocus Associations." *Genetics* 146(3) :1197–1206.

**URL:** <http://www.genetics.org/content/146/3/1197>

Kimura, Motoo. 1964. "Diffusion models in population genetics." *Journal of Applied Probability* 1(2) :177–232.

**URL:** <https://www.cambridge.org/core/journals/journal-of-applied-probability/article/diffusion-models-in-population-genetics/B0362BC3DE386579CBBCAF4663456717>

Lin, Kao, Andreas Futschik and Haipeng Li. 2013. "A Fast Estimate for the Population Recombination Rate Based on Regression." *Genetics* p. genetics.113.150201.

**URL:** <http://www.genetics.org/content/early/2013/04/08/genetics.113.150201>

Lorant, Anne. 2018. Plasticity and genetic adaptation as contributors to the evolutionary history of cultivated maize and its wild relatives PhD thesis université Paris-Saclay.  
**URL:** <http://www.theses.fr/225749440>

Mariac, Cedric, Viviane Luong, Issoufou Kapran, Aïssata Mamadou, Fabrice Sagnard, Monique Deu, Jacques Chantereau, Bruno Gerard, Jupiter Ndjeunga, Gilles Bezançon, Jean-Louis Pham and Yves Vigouroux. 2006. “Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers.” *Theoretical and Applied Genetics* 114(1) :49–58.

**URL:** <https://doi.org/10.1007/s00122-006-0409-9>

Messer, Philipp W. and Dmitri A. Petrov. 2013. “Population genomics of rapid adaptation by soft selective sweeps.” *Trends in Ecology & Evolution* 28(11) :659–669.

**URL:** <http://www.sciencedirect.com/science/article/pii/S0169534713002073>

Muthamilarasan, Mehanathan, Annni Dhaka, Rattan Yadav and Manoj Prasad. 2016. “Exploration of millet models for developing nutrient rich graminaceous crops.” *Plant Science* 242 :89–97.

**URL:** <http://www.sciencedirect.com/science/article/pii/S0168945215300558>

Nannipieri, Paolo, Laura Giagnoni, Loretta Landi and Giancarlo Renella. 2011. Role of phosphatase enzymes in soil. In *Phosphorus in action*. Springer pp. 215–243.

Oda, Atsushi, Sumire Fujiwara, Hiroshi Kamada, George Coupland and Tsuyoshi Mizoguchi. 2004. “Antisense suppression of the *Arabidopsis PIF3* gene does not affect circadian rhythms but causes early flowering and increases FT expression.” *FEBS letters* 557(1-3) :259–264.

Orr, H. Allen and Andrea J. Betancourt. 2001. “Haldane’s Sieve and Adaptation From the Standing Genetic Variation.” *Genetics* 157(2) :875–884.

**URL:** <http://www.genetics.org/content/157/2/875>

Oumar, Ibrahima, Cédric Mariac, Jean-Louis Pham and Yves Vigouroux. 2008. “Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci.” *Theoretical and Applied Genetics* 117(4) :489–497.

**URL:** <https://doi.org/10.1007/s00122-008-0793-4>

Pavlidis, Pavlos and Nikolaos Alachiotis. 2017. “A survey of methods and tools to detect recent and strong positive selection.” *Journal of Biological Research-Thessaloniki* 24(1) :7.

**URL:** <https://doi.org/10.1186/s40709-017-0064-0>

Prezeworski, Molly, Graham Coop and Jeffrey D. Wall. 2005. “The Signature of Positive Selection on Standing Genetic Variation.” *Evolution* 59(11) :2312–2323.

**URL:** <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0014-3820.2005.tb00941.x>

Salack, Seyni, Cornelia Klein, Alessandra Giannini, Benoit Sarr, Omonlola N. Worou, Nouhoun Belko, Jan Bliefernicht and Harald Kunstman. 2016. “Global warming induced hybrid rainy seasons in the Sahel.” *Environmental Research Letters* 11(10) :104008.

**URL:** <http://stacks.iop.org/1748-9326/11/i=10/a=104008>

Saïdou, Abdoul-Aziz, Cédric Mariac, Vivianne Luong, Jean-Louis Pham, Gilles Bezançon and Yves Vigouroux. 2009. “Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet.” *Genetics* .

Schrider, Daniel R. and Andrew D. Kern. 2017. “Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome.” *Molecular Biology and Evolution* 34(8) :1863–1877.

**URL:** <https://academic.oup.com/mbe/article/34/8/1863/3804550>

Schrider, Daniel R., Fábio K. Mendes, Matthew W. Hahn and Andrew D. Kern. 2015. “Soft Shoulders Ahead : Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps.” *Genetics* p. genetics.115.174912.

**URL:** <http://www.genetics.org/content/early/2015/02/24/genetics.115.174912>

Smith, John Maynard and John Haigh. 1974. “The hitch-hiking effect of a favourable gene.” *Genetics Research* 23(1) :23–35.

**URL:** <https://www.cambridge.org/core/journals/genetics-research/article/hitchhiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B>

Stephan, Wolfgang, Thomas H. E. Wiehe and Marcus W. Lenz. 1992. “The effect of strongly selected substitutions on neutral polymorphism : Analytical results based on diffusion theory.” *Theoretical Population Biology* 41(2) :237–254.

**URL:** <http://www.sciencedirect.com/science/article/pii/004058099290045U>

Tarafdar, J. C. and A. Jungk. 1987. “Phosphatase activity in the rhizosphere and its relation to the depletion of soil organic phosphorus.” *Biology and Fertility of Soils* 3(4) :199–204.

**URL:** <https://doi.org/10.1007/BF00640630>

VanLiere, Jenna M. and Noah A. Rosenberg. 2008. “Mathematical properties of the r<sup>2</sup> measure of linkage disequilibrium.” *Theoretical Population Biology* 74(1) :130–137.

**URL:** <http://www.sciencedirect.com/science/article/pii/S0040580908000609>

Varshney, Rajeev K., Chengcheng Shi, Mahendar Thudi, Cedric Mariac, Jason Wallace, Peng Qi, He Zhang, Yusheng Zhao, Xiyin Wang, Abhishek Rathore, Rakesh K. Srivastava, Annapurna Chitikineni, Guangyi Fan, Prasad Bajaj, Somashekhar Punnuri, S. K. Gupta, Hao Wang, Yong Jiang, Marie Couderc, Mohan A. V. S. K. Katta, Dev R. Paudel, K. D. Mungra, Wenbin Chen, Karen R. Harris-Shultz, Vanika Garg, Neetin Desai, Dadakhalandar Doddamani, Ndjido Ardo

Kane, Joann A. Conner, Arindam Ghatak, Palak Chaturvedi, Sabarinath Subramaniam, Om Parkash Yadav, Cécile Berthouly-Salazar, Falalou Hamidou, Jianping Wang, Xinming Liang, Jérémy Clotault, Hari D. Upadhyaya, Philippe Cubry, Bénédicte Rhoné, Mame Codou Gueye, Raman-julu Sunkar, Christian Dupuy, Francesca Sparvoli, Shifeng Cheng, R. S. Mahala, Bharat Singh, Rattan S. Yadav, Eric Lyons, Swapan K. Datta, C. Tom Hash, Katrien M. Devos, Edward Buckler, Jeffrey L. Bennetzen, Andrew H. Paterson, Peggy Ozias-Akins, Stefania Grando, Jun Wang, Trilochan Mohapatra, Wolfram Weckwerth, Jochen C. Reif, Xin Liu, Yves Vigouroux and Xun Xu. 2017. "Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments." *Nature Biotechnology* 35(10) :969–976.

**URL:** <https://www.nature.com/articles/nbt.3943>

Veldkamp, J. F. 2014. "A revision of Cenchrus incl. Pennisetum (Gramineae) in Malesia with some general nomenclatural notes.".

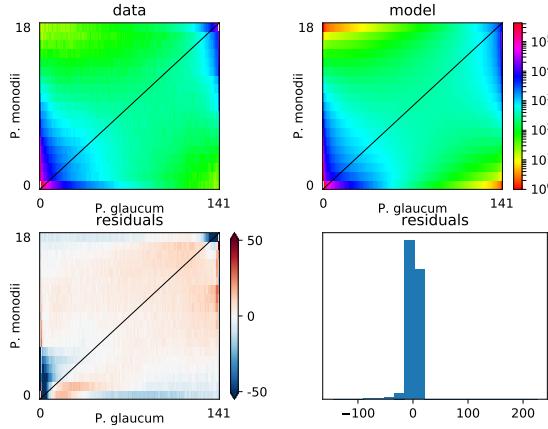
**URL:** <https://www.ingentaconnect.com/content/nhn/blumea/2014/00000059/00000001/art00012>

# Annexe A

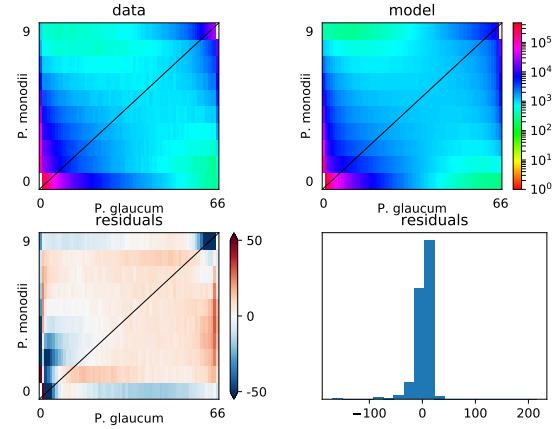
Paramètres	Unité	Centre Sahel				Tout Sahel			
		scénario 1	scénario 2	scénario 3	scénario 4	scénario 1	scénario 2	scénario 3	scénario 4
$-1 * lik$	-	82268.02	<b>70671.60</b>	82276.14	<b>70923.47</b>	230077.66	176339.64	178403.16	175929.02
$AIC$	-	82282.02	<b>70687.60</b>	82288.14	<b>70937.47</b>	230091.66	176355.64	178415.16	175943.02
$\theta$	$4N_o\mu$	0.0025.6	0.00503	0.0016	0.0023	0.0024	0.0024	0.0024	0.0029
$N_u1F$	$N_o$	0.446	0.172	0.705	0.379	0.832	0.764	0.854	0.661
$N_u2F$	$N_o$	0.602	0.321	0.953	0.707	0.631	0.602	0.608	0.514
$N_u2B$	$N_o$	0.013	0.030	0.096	0.129	0.010	0.062	0.032	0.023
$M$	$2N_o m$	1.881	-	1.178	-	1.962	-	2.017	-
$M_{12}$	$2N_o m$	-	6.245	-	2.821	-	2.505	-	2.817
$M_{21}$	$2N_o m$	-	2.463	-	1.146	-	1.818	-	2.194
$T_B$	$2N_o$	0.0002	0.009	-	-	0.185	0.088	-	-
$T_E$	$2N_o$	1.425	0.603	1.356	0.876	0.692	0.590	0.697	0.660
$pol\ error$	-	0.0810	0.0709	0.0837	0.0719	0.0075	0.0443	0.0469	0.0420

Ces paramètres sont définies par rapport à la population ancestrale  $N_0$ .  $N_{u1F}$  désigne la taille efficace de la population de mil sauvage,  $N_{u2F}$  représente la taille efficace de la population de mil cultivé,  $N_{u2B}$  la taille efficace de la population de mil cultivé au moment du Bottleneck.  $M$  étant le taux de migration par génération entre la population cultivée et sauvage (pour les modèles 1 et 3). Pour les modèles 2 et 4,  $M_{12}$  représente le taux de migration par génération de la population sauvage vers la population cultivée alors que  $M_{21}$  représente le taux de migration par génération dans le sens inverse.  $T_B$  désigne le temps en nombre de génération qu'a duré Bottleneck lors de la domestication du mil tandis que  $T_E$  est le temps en nombre de génération qu'a duré la phase de croissance exponentielle des mil cultivés après le goulot d'étranglement. Enfin  $pol\ error$  est le taux d'erreur de polarisation de nos données. (voir Figure )

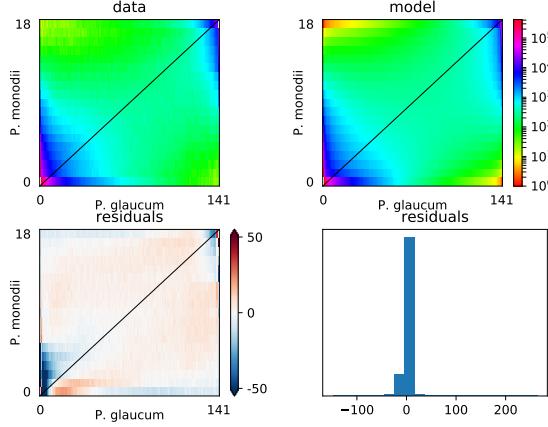
## Annexe B



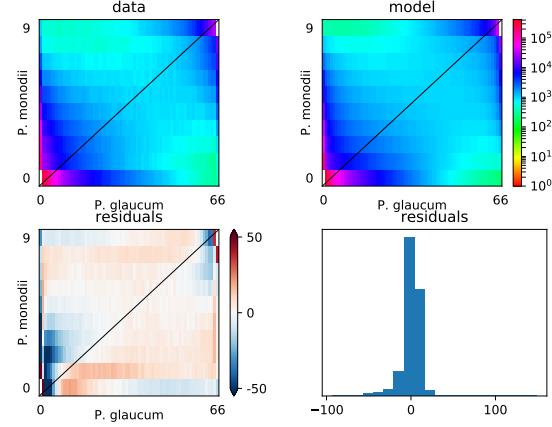
**FIGURE 17** – Modélisation démographique du Modèle2 sans prendre en compte d'une erreur de polarisation et en utilisant tout notre jeu de données



**FIGURE 18** – Modélisation démographique du Modèle2 sans prendre en compte d'une erreur de polarisation et en utilisant que les mils du centre du Sahel



**FIGURE 19** – Modélisation démographique du Modèle2 avec prise en compte d'une erreur de polarisation et en utilisant tout notre jeu de données



**FIGURE 20** – Modélisation démographique du Modèle2 avec prise en compte d'une erreur de polarisation et en utilisant que les mils du centre du Sahel

# Annexe C

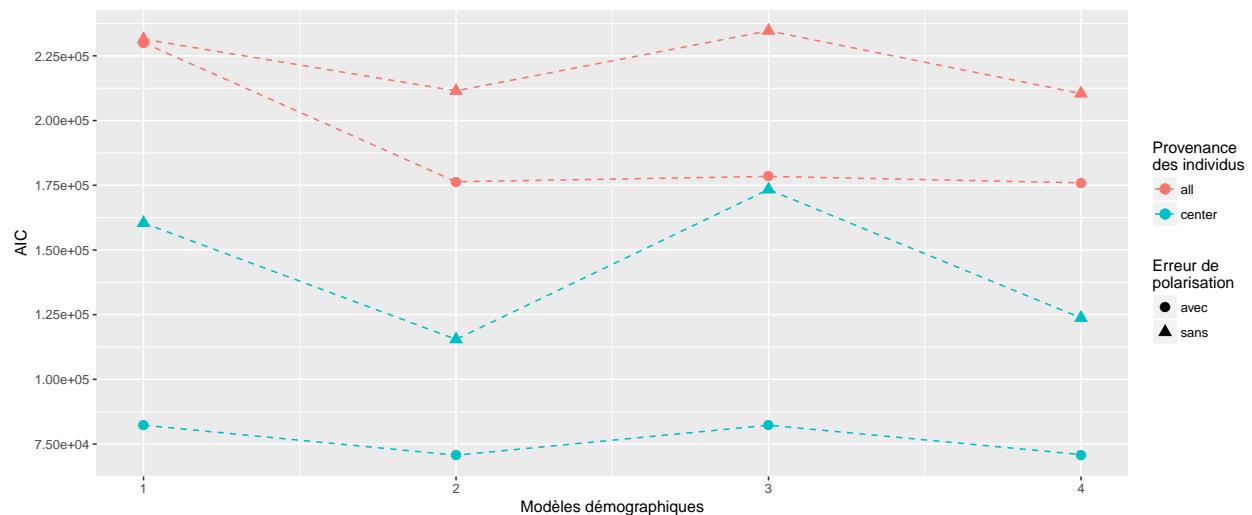


Figure 21– Résultat des modélisations démographiques.

En abscisse les chiffres désignent les numéros de modèle (voir figure 10). En ordonné les AICs de chaque scénario démographique inféré. Les carrés désignent les modélisations faites en prenant en compte une erreur de polarisation. Les triangles désignent les modélisations faites en ne prenant pas en compte une erreur de polarisation. En rouge sont représentés les AICs des scénarios démographiques inférés à partir de tous nos individus.

# Annexe D

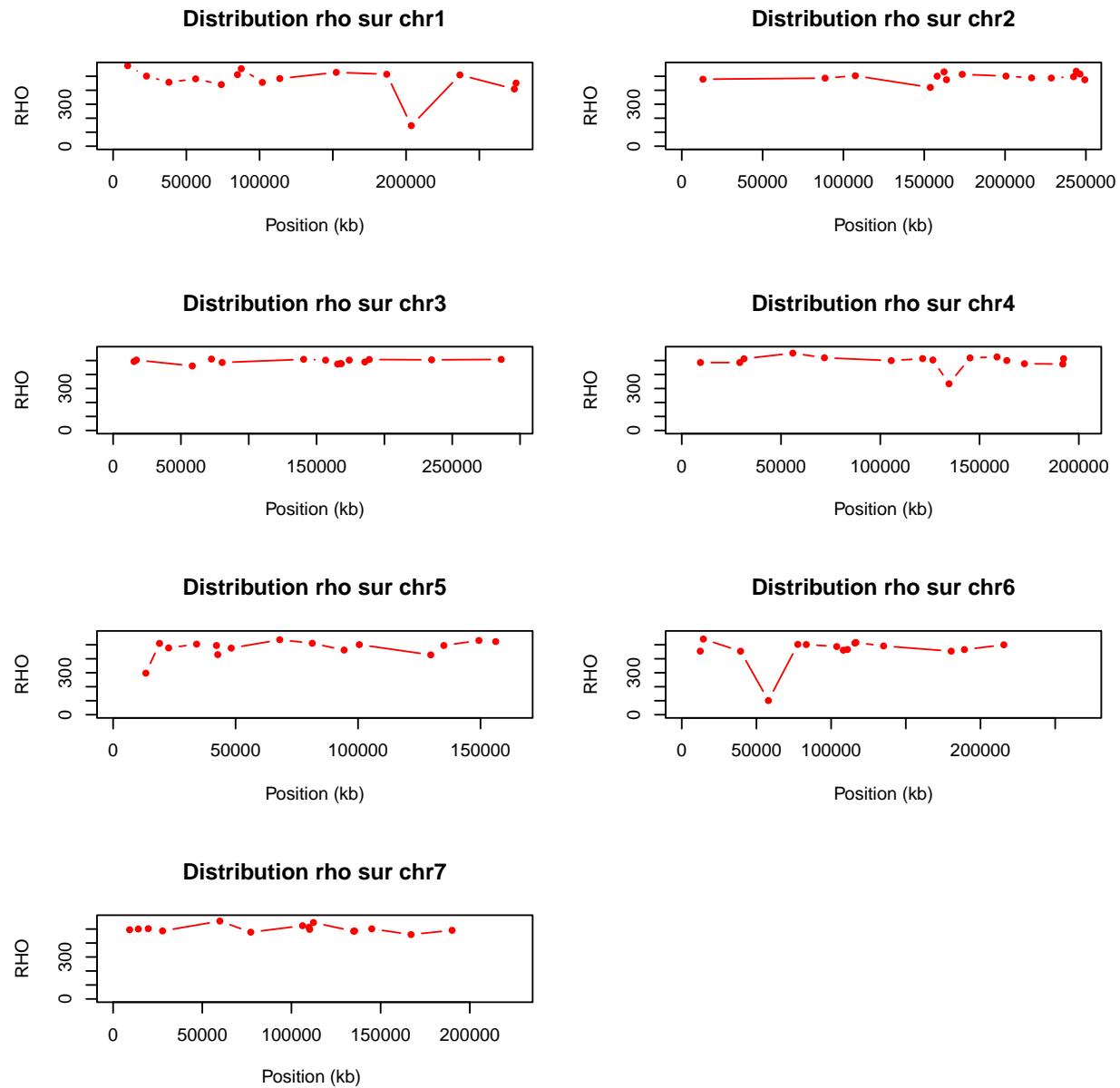


FIGURE 22 – Distribution des estimations du taux de recombinaison par génération le long de chaque chromosome de mil

## Résumé

Les changements climatiques auxquels nous assistons ces dernières décennies sont rapides et brutals. Ils se traduisent par des changements au niveau des populations de plantes et d'animaux. Pour faire face à ces contraintes, ces derniers développent plusieurs stratégies dont l'adaptation génétique. Dans ce contexte, la diversité fonctionnelle existante au sein des populations pourrait être un paramètre important pour garantir la viabilité et le potentiel évolutif des populations face aux changements globaux. De ce fait, identifier et comprendre comment la diversité fonctionnelle des plantes cultivées a été façonnée durant leur histoire paraît aujourd'hui important. En prenant comme modèle le mil (*Pennisetum glaucum*), nous avons identifié sa diversité fonctionnelle par l'analyse de la sélection à l'échelle du génome. Pour cela nous avons utilisé une méthode basée sur la diversité haplotypique (H12). Nous avons modélisé l'histoire démographique de cette plante afin d'avoir une approximation du profil de diversité attendu en absence de sélection et de détecter avec plus de précision et de confiance les régions sous balayage sélectif. Cette étude nous a permis d'entrevoir l'importance des migrations entre la population sauvage et cultivée du millet dans le façonnement de sa diversité actuelle. Les gènes potentiellement sous balayage sélectif que nous avons détectés participent aux activités de transport des substances toxiques, mais aussi aux processus d'absorption du phosphore. Sachant que le millet est principalement cultivé dans des zones avec des sols acides, donc à forte toxicité et pauvre en phosphore biodisponible. Nous avons aussi détecté un gène qui participerait au contrôle du temps de floraison chez les plantes. Des analyses supplémentaires seront nécessaires pour associer ces résultats à une adaptation du millet.

**Mots-Clés** Domestication, Balayage sélectif fort, Balayage sélectif doux, Sélection positive, Diversité fonctionnelle, *Pennisetum glaucum*

## Abstract

The climate changes in recent decades are rapid and brutal. They cause changes in plant and animal populations. To face these constraints, they develop several strategies including genetic adaptation. In this context, diversity based on standing genetic variation within the population could be an important parameter to ensure the viability and evolutionary potential of populations in the face of global change. Therefore, identifying and understanding how the functional diversity of cultivated plants has been shaped during their history seems important in the present day. Using millet (*Pennisetum glaucum*) as a model, we identified its functional diversity through genome-wide selection analysis. For this we used a method based on haplotypic diversity (H12). We modelled the demographic history of this plant in order to have an approximation of the expected diversity profile in the absence of selection and to detect with more precision and confidence the regions under selective scanning. This study has allowed us to see the importance of migration between the wild and cultivated millet population in shaping its current diversity. The genes potentially under selective scanning that we have detected participate in the transport activity of drugs, but also in the phosphorus absorption processes. Pearl millet is cultivated mainly in areas with acid soils, therefore highly toxic and low in bio-available phosphorus. We also detected a gene would participate in the control of flowering time in plants. Additional analyses will be necessary to associate results with an adaptation of millet.

**Keywords** Domestication, *Hard selective sweep*, *Soft selective sweep*, Positive selection, Functional diversity, *Pennisetum glaucum*