SOMMAIRE

LISTE DES ANNEXES :

Annexe 1. Genetic basis of pearl millet population adaptation along an environmental gradient investigated by a combination of genome scan and association mapping.

Annexe 2. Selection for earlier flowering crop associated to climatic variations in the Sahel.

CHAPITRE 1. INTRODUCTION

Contexte général de l'étude

Le changement global du climat : quels effets sur l'agriculture ?

L'évolution du climat est marquée par des changements importants au niveau planétaire. Parmi ces changements, on note la variation de la température et de la pluviométrie. La fin du siècle passé a connu une augmentation de température de l'ordre de 0.2° C par décade (Hansen *et al.* 2006). Ces changements ont été liés en grande partie aux activités humaines dont l'émission de gaz à effet de serres (Groupe d'experts intergouvernemental sur l'évolution du climat, GIEC 2007). La pluviométrie a subi des changements divers selon la région du globe. Ces changements ne sont pas toujours de même ampleur, ni de même direction, lorsqu'ils sont déclinés aux échelles régionales (Figure G1). De façon générale, on note à la fin du 20^{e} siècle une augmentation des précipitations dans les régions Nord du globe (les latitudes hautes), une réduction des précipitations en Chine, en Australie et dans le pacifique, et une augmentation de la variabilité des pluies dans les régions équatoriales (Dore 2005). Les augmentations de précipitation dans certaines régions, au niveau notamment l'hémisphère Nord, pourraient être associées à l'augmentation de la fréquence d'évènements pluvieux intenses et extrêmes, même si ces évènements extrêmes sont aussi observés dans des régions où le cumul pluviométrique est en baisse (Dore 2005).

Les conséquences observées ou prévisibles du changement climatique sont multiples. Elles impliquent à la fois des perturbations météorologiques et écologiques, des impacts sur la biodiversité et l'agriculture, et des impacts sur les activités humaines induits par la modification du milieu naturel. Parmi les impacts biologiques, on note la variation de la phénologie et la perturbation de la dynamique de populations animales et végétales dans différents contextes écologiques (Walther *et al.* 2002). Chez les espèces agricoles, on a pu observer, au niveau européen par exemple, l'occurrence plus tardive des stades phénologiques chez des espèces comme la vigne et le pommier (Seguin 2010). Ces changements ont, entre autres, entrainé la modification de l'organisation du travail agricole selon les régions (calendrier cultural et période de vendange), mais aussi ils interrogent sur la qualité futurs des produits agricoles (Seguin 2010).

Par ailleurs, des phénomènes climatiques naturels comme l’oscillation de l’*El Niño* australe expliqueraient de nos jours entre 15% et 35% de la variation global du rendement chez le blé, les oléagineux et les céréales secondaires; cela laisse clairement présager qu’une perturbation futur du climat pourrait avoir des effets bouleversants sur l’agriculture (Howden *et al.* 2007).

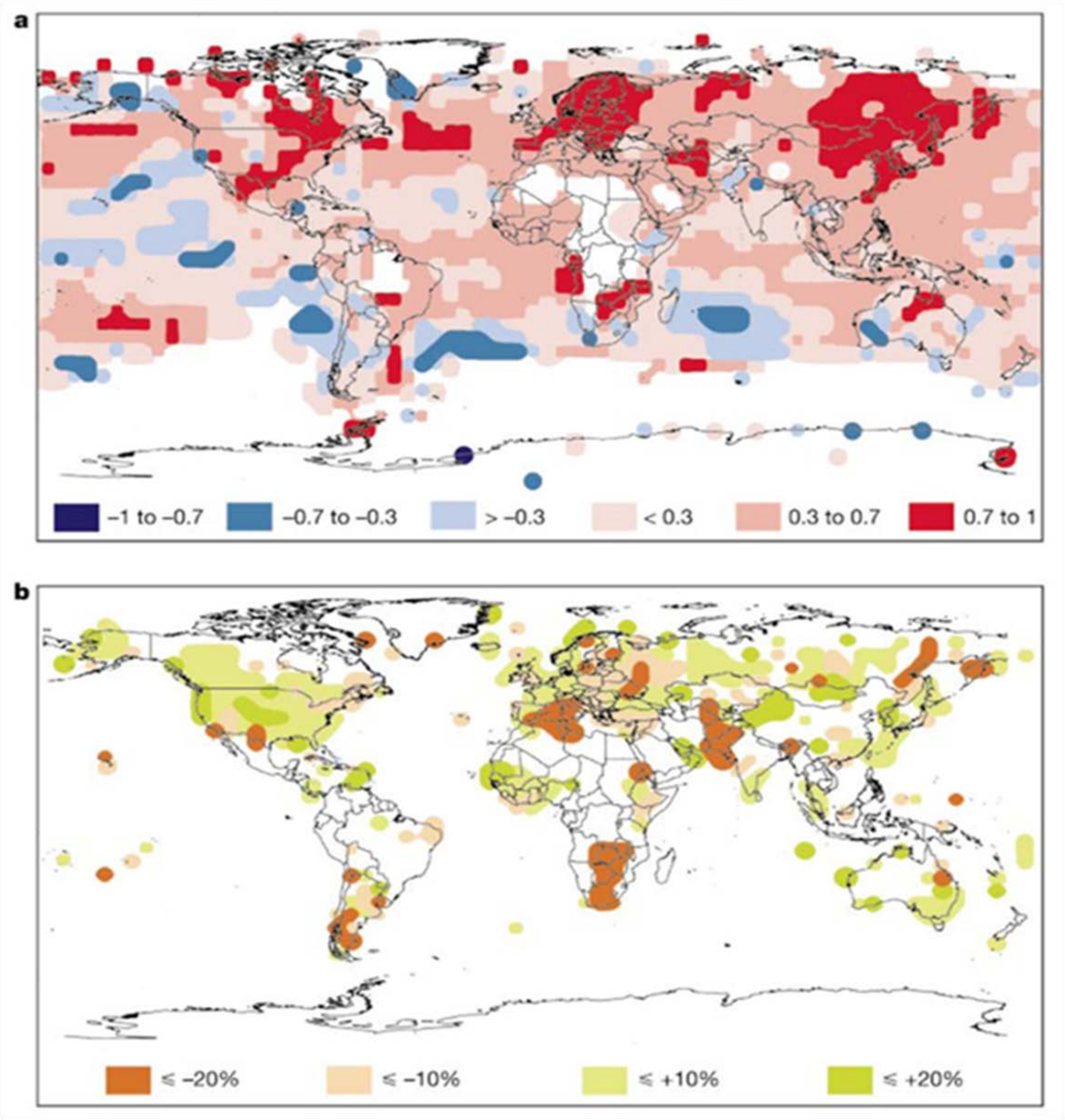


Figure G1. Variabilité spatiale des changements de température et de pluviométrie observés à la fin du 20^{e} siècle (d’après Walther *et al.* 2002). Le changement de température est donné en °C par décade (a). Le changement de pluviométrie est indiqué en pourcentage par décade (b). Les tendances de changement sont mesurées pour la période 1976-2002 par rapport à la normale climatologique de la période 1961-1990.

Le rapport du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) fait état de différentes hypothèses à propos de l’impact futur du réchauffement sur les rendements agricoles (GIEC, 2007). Pour les régions situées en moyennes et hautes latitudes, ces projections prédisent un effet positif sur les rendements pour une augmentation de température de l’ordre de 1 à 3°C selon la culture, et une baisse de rendement pour des hausses de température supérieures à ces seuils. Dans les régions sèches et dans les régions tropicales, une augmentation de température, même plus faible (1 à 2°C), pourrait conduire à des baisses significatives de rendement.

Les projections sur l’évolution du climat tentent de prévoir à la fois la direction et l’intensité des changements pour les prochaines décennies, de même que son impact potentiel. Les régions du globe qui concentrent aujourd’hui 95% des populations vulnérables à l’insécurité alimentaire (soit environ 825 millions de personnes) connaîtraient des augmentations de température de l’ordre de 0.5 à 1°C à l’horizon 2030 (Figure G2, Lobell *et al.* 2008). Ces régions sont situées principalement en Afrique, en Asie et en Amérique du Sud, et connaîtraient aussi des variations pluviométriques de différentes directions (Figure G2). La projection des impacts du changement climatique sur l’agriculture devra se faire, pour une région donnée, en prenant en compte i) l’évolution climatique attendue, ii) les espèces cultivées significatives pour la sécurité alimentaire, et iii) la dépendance de ces cultures vis-à-vis des principales variables climatiques. Pour les régions les plus vulnérables du globe en termes de sécurité alimentaire, trois grandes tendances se déclinent à l’horizon 2030, si des actions efficaces ne sont pas faites pour contrecarrer les impacts du changement climatique (Lobell *et al.* 2008) :

i) certaines cultures comme le sorgho au Sahel, ou le maïs et le blé en Afrique du Sud, connaîtraient une baisse de production significative. Il s’agit en général de cultures qui ont été jusqu’ici fortement sensibles à la température et pour lesquelles le réchauffement prévu aurait des effets potentiels négatifs. Même dans le cas où une augmentation de la pluviométrie serait probable, celle-ci est assez incertaine et donc son effet est dominé (dans les modèles prédictifs) par l’effet de la température;

ii) la deuxième tendance concerne des cultures pour lesquelles la prévision des impacts est très incertaine et peu prévisible. C’est le cas par exemple de l’arachide en Asie du Sud ou du sorgho en Afrique du Sud. Pour ces cultures, les prévisions d’impact couvrent à la fois l’éventualité de scénarios de baisse de production et de scénarios d’augmentation de production. Il s’agit en général de cultures qui par le passé ont été largement dépendantes de la pluviométrie, et pour lesquels les projections futures de pluviométrie restent très incertaines et ne permettent pas de prédiction fiable;

iii) enfin, certaines cultures ne seraient soumises qu’à un impact faible, voire négligeable. C’est le cas par exemple du blé en Asie occidentale, du riz au Sahel, ou du manioc en Afrique de l’Ouest. Dans certains cas, il s’agit de cultures (manioc ouest-africain par exemple) dont la dépendance vis-à-vis du climat saisonnier n’a pas été trop forte dans le passé.

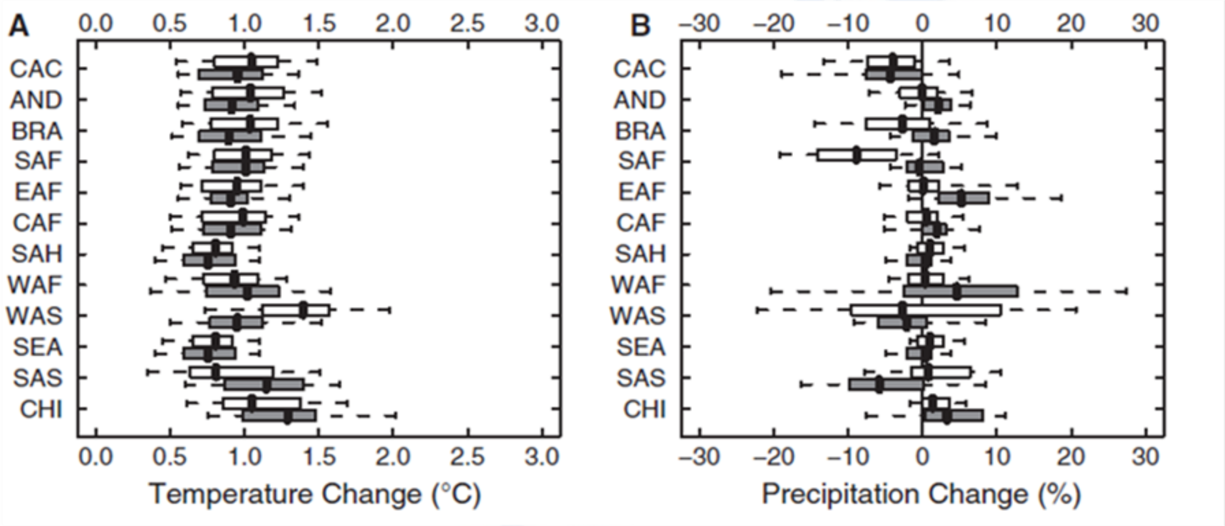


Figure G2. Projection des changements de température et de pluviométrie à l’horizon 2030 pour différentes régions vulnérables sur le plan alimentaires (d’après Lobell *et al.* 2008). Les variations de température (A) et de pluviométrie sont simulées sur la base de modèles. Les changements sont évalués relativement à la normale de 1980-1999. Les rectangles gris montrent la variation pour les mois de décembre à février, et les rectangles blancs montrent des variations pour les mois de juin à août. Les pointillés correspondent aux percentiles 5% et 95%, et les rectangles s’étendent aux percentiles 25% et 75%. Les valeurs médianes sont indiquées en noir pour chaque région. SAS : Asie du Sud. CHI : Chine. SEA : Asie du Sud Est. EAF : Afrique de l’Est. CAF : Afrique Centrale. SAF : Afrique du Sud. WAF : Afrique de l’Ouest. CAC : Amérique Centrale et Caraïbes.SAH : Sahel. WAS : Asie occidentale. AND : Andes. BRA : Brésil.

Si ces scénarios se confirmaient, les bouleversements qu’ils vont engendrer pourraient compromettre ou aggraver la situation agricole et alimentaire de beaucoup de populations. Pour répondre à ces défis, les systèmes de culture devraient évoluer pour faire face aux changements. Des modifications mineures dans les systèmes de culture actuels pourraient limiter les impacts d'un changement climatique modéré, mais des changements forts nécessiteraient des solutions plus systémiques comme la diversification des systèmes de culture (Howden *et al.* 2007). L'adaptation de l'agriculture doit également se penser en intégrant la variabilité climatique dans un cadre de gestion de risque qui engage les agriculteurs, les industriels et les politiques, tout en prenant en compte les critères de marchés et de développement durable (Howden *et al.* 2007). De même que la manifestation physique du changement climatique est hétérogène suivant l'espace, son impact et sa gestion en relation avec l'agriculture vont tout aussi dépendre de la région considérée et de ses spécificités multiples aux plans environnemental, agricole mais aussi socio-économique.

Changement climatique dans le contexte sahélien

La région sahélienne (ou le *Sahel*) est l'une des régions du globe qui ont connu les sécheresses les plus importantes à partir des années 1970[1] (Dai 2011). La baisse de pluviométrie dans cette région s'est maintenue depuis 40 ans (Figure G3). Cela s'est traduit par un déficit pluviométrique global d'environ 30% sur toute la région sahélienne (AGRHYMET et CIRAD 2005).

La sécheresse a été caractérisée par une saison de pluies plus courte qui restreint davantage la période agricole pour les cultures pluviales (AGRHYMET et CIRAD 2005). Sur plusieurs siècles, les périodes humides ont été souvent alternées, dans cette région, avec des périodes sèches (AGRHYMET et CIRAD 2005). Des causes naturelles sont donc associées au dessèchement, historique ou récent, du climat sahélien (Dai 2011). Toutefois, les changements récents, observés au Sahel à partir des années 1970, pourraient être aussi associés à l'effet des activités humaines, notamment l'émission des gaz à effet de serres (Biasutti et Sobel 2009). Ces changements seraient donc, au moins en partie, l'expression régionale du changement global du climat à l'échelle planétaire. Les projections pour le 21^{e} siècle prédisent un démarrage plus tardif et un raccourcissement de la saison hivernale (Biasutti et Sobel 2009). La région sahélienne pourra aussi être éprouvée, à l'horizon 2030, par une variation forte de pluviométrie et une augmentation de température pouvant atteindre 1°C (Lobell *et al.* 2008). Une des conséquences importantes de ces variations du climat est leur effet sur la production agricole et sur la sécurité alimentaire.

[1] La zone sahélienne constitue une bande géographique à climat semi aride traversant plusieurs pays de l'Afrique au Sud du Sahara, notamment des pays de l'Afrique de l'Ouest; elle est située en latitude entre 20-10N.

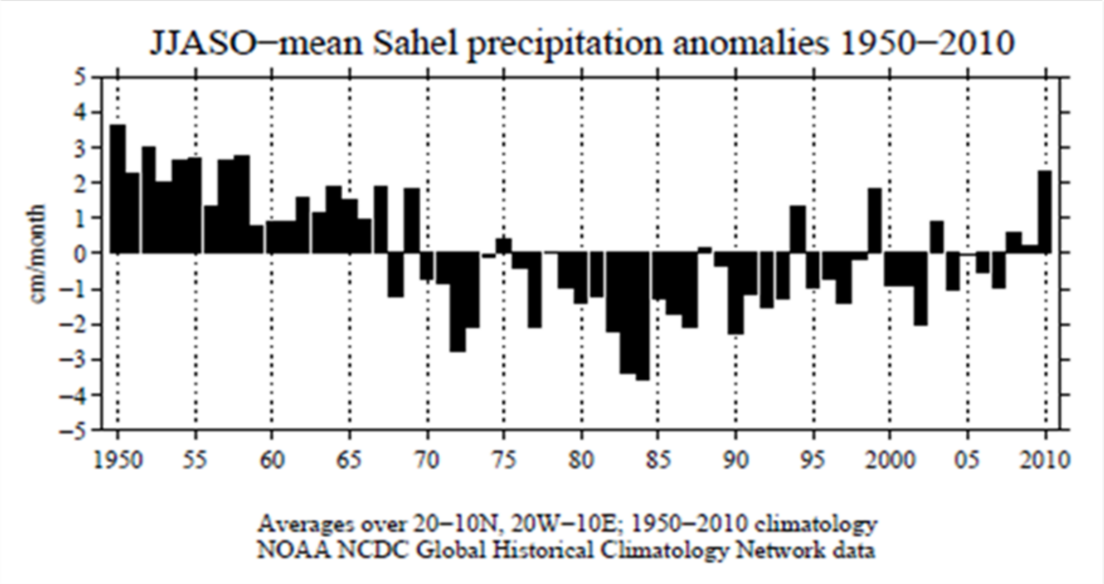


Figure G3. Déviation annuelle de la pluviométrie au Sahel : la sécheresse se maintient depuis 1970 (d'après *The Joint Institute for the Study of the Atmosphere and Ocean, University of Washington*, 2010). L'écart des moyennes annuelles à la normale est montré pour chaque année de 1950 à 2010. La normale a été calculée sur la base de la période 1950-2010. On peut observer la baisse persistante de la pluviométrie depuis 40 ans.

La sécurité alimentaire des populations sahéliennes dépend en très grande partie de céréales dont principalement le mil et le sorgho (Bezançon et Pham 2004). La culture du mil est conduite sans irrigation, et dépend essentiellement de la pluviométrie naturelle. La culture s'étend dans des zones pluviométriques dont les cumuls annuels descendent jusqu'à 300 mm par an, ce qui correspond à une disponibilité en eau faible, comparativement au sorgho et au maïs par exemple, pour lesquels le minimum d'exigence serait respectivement de l'ordre de 400 mm et 500-600 mm (ICRISAT and FAO 1996). La production du mil au Sahel est cependant très fortement tributaire des variations climatiques locales et interannuelles (AGRHYMET et CIRAD 2005). Cette production varie fortement avec la durée de la saison de pluies (Eldin 1993). La durée de la saison dépend essentiellement de la date d'installation des pluies (début de saison), la fin de saison étant plus stable d'une année à l'autre (AGRHYMET et CIRAD 2005, Sivakumar 1988). Le choix de cultivars en fonction de la durée de leur cycle est un élément clef de la stratégie de sécurisation de la production, au regard de la variabilité spatiale et interannuelle de la durée de la saison (Eldin 1993).

Au vu des prévisions et des incertitudes sur l'évolution future du climat, la compréhension des caractères biologiques complexes (tels que la durée du cycle) qui déterminent l'adaptation du mil au climat, est une perspective intéressante, qui pourrait contribuer à la gestion de l'impact du changement climatique.

Le mil : aspects agro-biologiques, défis de la production

Le nom commun *mil* désigne au sens large plusieurs espèces de graminées, dont entre autres *Pennisetum glaucum*, *Eleusine coracana*, *Panicum miliaceum*, *Setaria italica*, *Echinochloa crusgalli* (ICRISAT and FAO 1996, Larousse Agricole 2002). L'espèce objet de ce travail est *Pennisetum glaucaum* (Tableau G1). Cette espèce regroupe trois sous-espèces (Bezançon *et al.* 1994) : *P. glaucaum* ssp. glaucum (forme cultivée), *P. glaucum* ssp. violaceum (forme sauvage) et *P. glaucaum* ssp. sieberianum (forme hybride entre le cultivé et le sauvage). Notre travail porte principalement sur le type cultivé, *P. glaucaum* ssp. glaucum[2].

Tableau G1. Classification botanique du mil.

Critère botanique	Classification
Règne	Plantae
Division	Magnoliophyta
Classe	Liliopsida
Ordre	Cyperales
Famille	Poaceae
Sous-famille	Panicoideae
Tribu	Paniceae
Genre	Pennisetum
Nom binomial	*Pennisetum glaucum* (L.) R. Br.

Cette espèce est diploïde, avec un génome composé de 2n=14 chromosomes. Son mode de reproduction est sexué, avec une fécondation principalement allogame et anémophile. Elle a été très probablement domestiquée en Afrique de l'Ouest, puis diffusée en Afrique et en Asie (Oumar *et al.* 2008). L'aire de répartition actuelle couvre différentes régions du monde; les surfaces cultivées les plus importantes se trouvent au Sahel (Figure G4).

Les pays sahéliens comme le Niger, le Nigéria et le Mali constituent, après l'Inde, les plus grands producteurs de mil à l'échelle mondiale (Figure G5).

[2] Sauf précision contraire, nous utiliserons le terme *mil* pour désigner exclusivement cette espèce (*Pennisetum glaucum*).

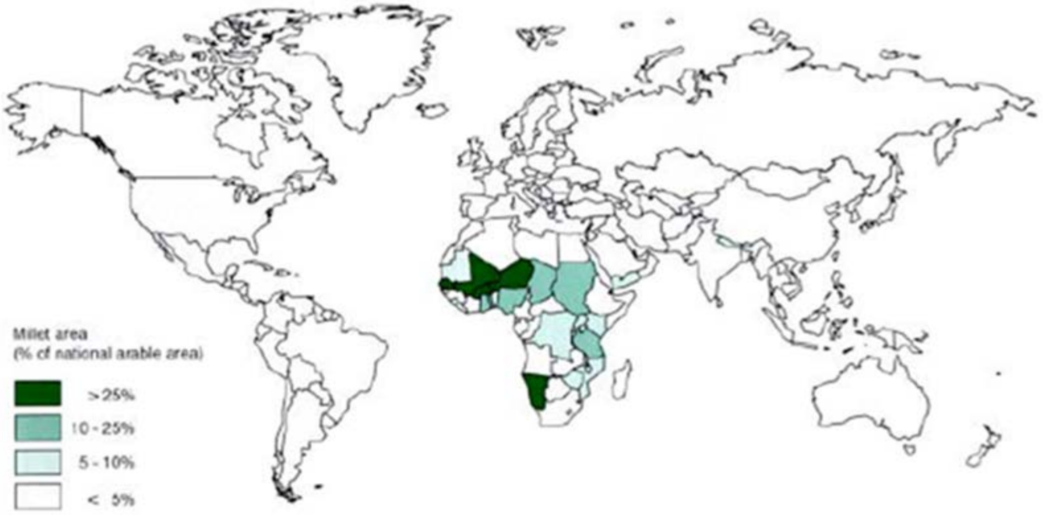


Figure G4. Répartition des aires de culture relatives du mil à l'échelle mondiale (d'après FAO and ICRISAT 1996). Les aires de cultures sont représentées en proportion de la surface agricole totale de chaque pays.

En Inde, la culture est importante principalement dans les zones semi-arides; la production totale du pays est estimée à plus de 11 millions de tonnes en 2008 (FAO, 2011). Au Nigéria et au Niger (2^{e} et 3^{e} producteurs mondiaux), cette production était estimée respectivement à 9,1 millions de tonnes et 3,5 millions de tonnes (FAO, 2011).

Le mil est une culture privilégiée par les paysans Sahéliens, du fait de son adaptation aux conditions particulières de production dans cette région[3]. Ces conditions incluent une moindre disponibilité en eau, des températures élevées, des terres à dominance plus ou moins sableuse, et des systèmes de cultures traditionnels avec des itinéraires extensifs (ICRISAT and FAO 1996, Serpentier et Milleville 1993). Les variétés cultivées sont généralement des variétés populations locales à base génétique large (Bezançon *et al.* 2009).

La production du mil doit faire face au triple défi de la croissance démographique (augmentant la demande vivrière), de la disponibilité réduite en terres dédiées à l'agriculture et de la contrainte climatique. Une des principales stratégies qui ont permis l'augmentation de la production céréalière au Sahel, dans la deuxième moitié du 20^{e} siècle, est l'extension des superficies de culture. Aujourd'hui, les possibilités d'extension de ces superficies s'amenuisent. L'amélioration

[3] Au Niger par exemple, le mil occupe jusqu'à 2/3 de la surface agricole (dont une partie en association avec d'autres cultures); il représente à lui seul 73% de la production céréalière totale (Bezançon et Pham, 2004).

des rendements constituera alors une nécessité réelle pour augmenter la production et répondre à une demande croissante[4]. Pour sécuriser cette production dans un contexte de changement climatique, les stratégies agricoles devront prendre en compte le potentiel adaptatif des variétés. La variation de la date de floraison est l'une des composantes majeures de l'adaptation aux différenciations climatiques chez le mil (Haussmann *et al.* 2006). Ce caractère peut donc être exploité pour la gestion future de la production dans un contexte de changement climatique.

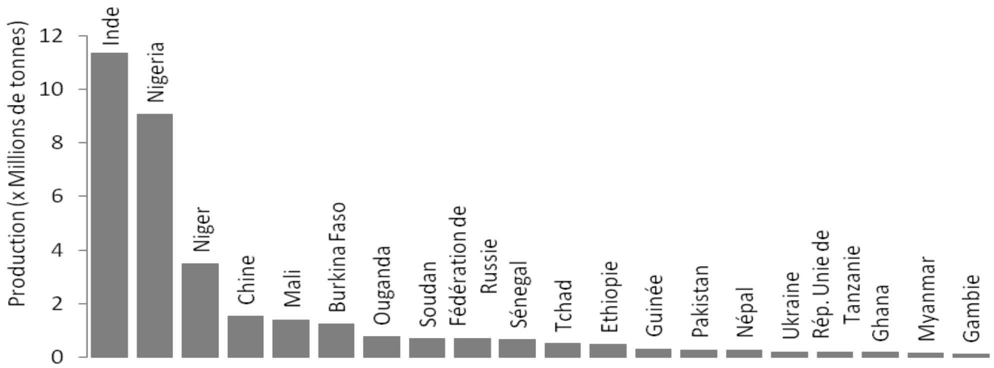


Figure G5. Production mondiale du mil (année 2008). Les vingt premiers pays producteurs sont représentés. Source de données : FAO, 2011.

La date de floraison : un caractère clef pour l'adaptation au climat

La variation de la date de floraison est un caractère clef dans l'adaptation des populations de plantes cultivées ou des populations naturelles aux conditions environnementales (Roux *et al.* 2006). Dans les conditions environnementales optimales, une date de floraison tardive est intéressante et permet d'allonger la phase de croissance et de favoriser ainsi l'accumulation de ressources pour la production des graines. Dans des environnements moins favorables, marqués par exemple par une saison pluvieuse courte, la précocité de la date de floraison est une stratégie pour éviter le stress tout en réalisant un minimum de rendement. La diffusion des plantes en dehors de leurs aires de domestication a été possible grâce, notamment, au potentiel adaptatif permettant d'ajuster le cycle de vie des variétés aux conditions environnementales. Chez le maïs, la sélection sur les gènes de floraison a joué un rôle dans l'adaptation des variétés, d'origine tropicale, aux conditions des climats tempérés (Camus-Kulandeivelu *et al.* 2008, Camus-Kulandeivelu *et al.* 2006). Chez le mil, il existe deux stratégies adaptatives associées à la date de floraison. La première est la précocité, qui permet un raccourcissement du cycle de la plante.

[4] Le rendement est une mesure de la production par unité de surface (notée en kg par hectare, par exemple). Cette mesure relative permet donc d'apprécier le niveau de productivité par rapport à une même superficie.

Dans ce premier cas, les génotypes ne sont pas forcément sensibles à la durée du jour. La seconde stratégie est la sensibilité à la photopériode, qui permet une floraison synchronisée avec la fin de la saison, grâce à la perception de la durée du jour (Li *et al.* 2010, Clerget *et al.* 2007). Les deux types de génotypes existent chez le mil (Clerget *et al.* 2007). Les variétés locales africaines présentent une date de floraison qui varie de 40 jours à 160 jours (entre le semis et la floraison) et qui suit le gradient climatique de l'aire de distribution (Haussmann *et al.* 2006). Les variétés à cycle de floraison précoce sont prédominantes dans les localités les plus sèches, alors que les variétés à cycle tardif prédominent dans les zones humides. La floraison est donc une des stratégies évolutives associées à l'adaptation des cultures à différentes conditions climatiques.

Plusieurs voies de contrôle de la floraison ont été identifiées chez les plantes, associées à la perception par la plante de la température (vernalisation), de la lumière, ou de signaux hormonaux (Putterill *et al.* 2004). Chez Arabidopsis, plus de 80 gènes impliqués dans la régulation de la voie de la floraison on été identifiés (Blázquez *et al.* 2001). Des études comparatives indiquent la conservation de plusieurs gènes de cette voie entre différentes espèces (Hayama *et al.* 2003, Hecht *et al.* 2005). La date de la floraison, mesurée souvent en nombre de jours entre le semis et l'apparition des organes floraux, est un des caractères phénotypiques dépendant de la voie de la floraison[5].

Chez le maïs par exemple, un grand nombre de *QTLs* à effets faibles sont associés à la date de floraison et permettent de prédire la variation de ce trait (Buckler *et al.* 2009) [6]. Des polymorphismes liés au gène candidat *Dwarf8* (Thornsberry *et al.* 2001), ou au locus et *Vgt1* (Ducroq *et al.* 2008), on également été identifiés comme de bons candidats associés à la variation de la date de floraison. Chez Arabidopsis, les études d'association phénotype-génotype sur l'ensemble du génome ont mis en évidence un grand nombre de marqueurs moléculaires associés à ce trait (Atwell *et al.* 2010, Brachi *et al.* 2010). Comparativement à d'autres céréales d'importance agricole, comme le riz, le maïs et le sorgho, le mil a fait l'objet de très peu de travaux dans ce domaine. Chez cette espèce, quelques QTLs associés à la date de floraison ont

[5] La date de floraison peut être exprimée aussi en unités, comme les degré-jour, qui prennent en compte des variables environnementales.

[6] Un *QTL*, ou *quantitative trait locus*, désigne un fragment chromosomique contenant des polymorphismes associés à la variation d'un trait quantitatif. Un QTL peut recouvrir plusieurs gènes liés physiquement.

été identifiés dans les groupes de liaison 2, 4, 5, 6 (2 QTLs par groupe) et le groupe de liaison 7 (Poncet 1998, Yadav *et al.* 2002, Yadav *et al.* 2003). Cependant, la localisation des régions impliquées est peu précise et le nombre d'études reste faible. La connaissance des bases génétique de la variation de la date de floraison chez le mil nécessite alors un effort de recherche important. L'état de l'art des études sur la base génétique des traits quantitatifs offre aujourd'hui, à cette fin, des méthodologies performantes qui permettront d'étudier les traits d'intérêt avec assez de *puissance* et de *résolution*[7] (Mackay *et al.* 2009, Mackay 2001, Myles *et al.* 2009). Nous présenterons les méthodes d'association génotype-phénotype que nous avons utilisées au cours de ce travail de thèse.

Méthodologie d'association génotype-phénotype

Introduction

Les études d'association permettent d'identifier les bases génétiques de la variation phénotypique. Ces approches analysent l'association statistique entre la variation phénotypique et le polymorphisme génétique. Chez les plantes, deux approches générales sont utilisées pour identifier les variants naturels liés au phénotype[8]. La première est la cartographie de liaison ou *linkage mapping* (Mackay *et al.* 2009). Cette approche utilise des populations expérimentales (*familles*) constituées à partir de croisements entre un nombre limité de parents (2 généralement). La seconde approche, plus récente, est la cartographie d'association basée sur des populations (ou *population-based association mapping*). Cette dernière utilise des collections de populations diverses, constituées d'individus dont les relations ne sont pas souvent connues *a priori*[9].

Ces deux approches ont en commun le fait qu'elles exploitent la possibilité de recombinaison pour séparer différents fragments du génome et mettre en évidence les fragments associés à la variation phénotypique (Myles *et al.* 2009). Dans le cas de la cartographie de liaison, les

[7] La *puissance* des méthodes détermine leur capacité à détecter des associations. La *résolution* détermine la précision dans la localisation des polymorphismes causaux.

[8] Les études de mutagénèse sont également utilisées pour identifier des variants d'intérêt en appliquant des mutations artificielles. Ces mutations ne sont pas toujours fonctionnelles dans les populations naturelles. Cette méthode n'est pas traitée dans le cas de la présente thèse.

[9] Nous utiliserons plus souvent dans le texte le terme *familles* (plutôt que *populations*) en référence aux groupes composés d'individus partageant un lien d'ascendance commune assez récente; c'est généralement le cas des populations classiques de cartographie de liaison. Le terme *populations* sera réservé préférentiellement (mais pas exclusivement) pour désigner les collections diverses utilisées dans le contexte de la seconde approche, dite *cartographie d'association*.

évènements de recombinaison exploitables sont ceux qui se sont produits dans les quelques générations faisant suite aux croisements effectués. Différents types de populations expérimentales ont été développées dans le cadre de la cartographie de liaison : F2, *near isogenic lines* (NILs), *recombinant inbred lines* (RILs), etc. Ces populations sont généralement basées sur le croisement entre deux parents. Plus récemment, une production de RILs à partir de croisements entre plusieurs parents a été proposée; ces lignées sont désignées sous le nom *multiparent advanced generation inter-cross lines*, ou lignées MAGIC (Kover *et al.* 2009). Les lignées *MAGIC* présentent l'avantage, comparées aux populations traditionnelles de cartographie de liaison, de capturer une plus grande part de diversité génétique et phénotypique de l'espèce, du fait du nombre de lignées parentales plus élevé. Le nombre plus élevé de croisements et de générations dans le cas des ces lignées permet aussi d'augmenter le nombre de recombinaisons dans la descendance, améliorant ainsi la résolution de la cartographie (Kover *et al.* 2009).

La cartographie d'association basée sur des populations exploite un nombre d'événements de recombinaison davantage plus grand que ce qui peut être obtenu avec n'importe laquelle des populations de cartographie de liaison citées. Ces recombinaisons sont le résultat d'une accumulation au cours de l'histoire évolutive des collections de populations utilisée par la cartographie d'association. Elles conduisent à une résolution supérieure dans la localisation des régions génomiques associées au phénotype. Le nombre d'allèles disponible au sein de ces populations est également plus large que ce qui peut être représenté au sein de familles biparentales (Figure G6, Yu and Buckler 2006). Ces avantages ont fortement motivés, depuis une décennie, le développement de cette méthodologie chez les plantes.

Nous passerons en revue différentes questions méthodologiques dans le cadre de la cartographie d'association basée sur des populations. Premièrement, nous discutons le problème de la structure génétique des populations et son effet confondant dans les études d'associations. Cette question est cruciale car le biais statistique lié à la méconnaissance des relations entre les individus peut se traduire en une hausse significative du taux de faux positifs des modèles statistiques. L'effet confondant de la structure a été l'obstacle majeur à l'introduction de cette approche de cartographie chez les plantes. Pour lever cet écueil, la démarche a consisté à introduire des méthodes pour inférer la structure des populations (Pritchard *et al.* 2000), puis à utiliser des modèles permettant de prendre en compte la structure inférée pour l'analyse d'association (Thornsberry *et al.* 2001). Nous présenterons, dans un deuxième temps, un ensemble de méthodes développées pour inférer et analyser la structure génétique à partir de données moléculaires. Enfin, nous présenterons les modèles statistiques communément utilisés dans les études d'association chez les plantes. Une démarche permettant d'évaluer la performance de différents modèles et de les comparer est présentée, autour de quelques critères fondamentaux, en l'occurrence la limitation du taux de faux positifs, la puissance et l'ajustement du modèle aux données.

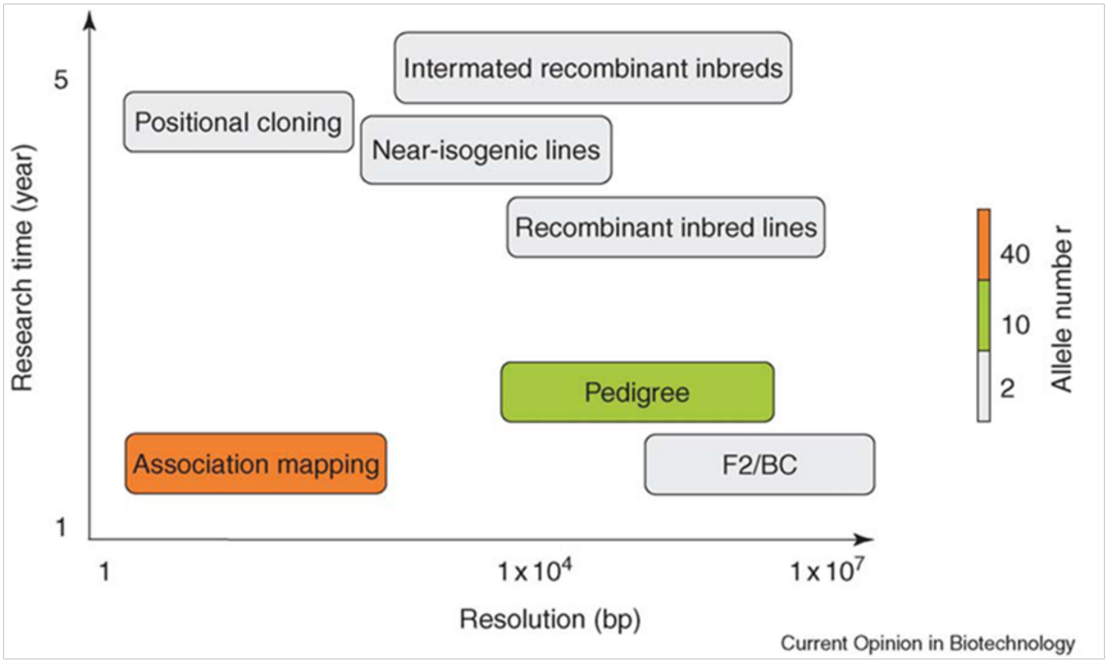


Figure G6. Comparaison de différentes méthodes d'association phénotype-génotype (d'après Yu and Buckler 2006). La différence entre les méthodes est illustrée en fonction de : i) la résolution (les distances les plus petites sur l'axe correspondent à une localisation plus précise du polymoprhisme causal); ii) le temps de recherche; et iii) le nombre d'allèles par locus que les échantillons peuvent contenir. La cartographie d'association (*association mapping*) est la méthode la plus performante au vu de ces critères.

Par commodité de langage, nous utiliserons plus souvent le terme de *cartographie d'association* ou *études d'association* pour désigner la cartographie d'association basée sur des collections de populations. Le sens ici est bien entendu restrictif, du moment où la *cartographie de liaison*, basée sur des familles issues d'un croisement contrôlé, demeure également une cartographie d'association[10].

Effet confondant de la structure génétique dans les études d'association

Les populations évoluent sous la pression des forces évolutives comme la dérive génétique et la sélection. La différenciation d'une population, dans le cas d'une adaptation locale par exemple, se décline à deux niveaux : la différenciation génétique, consistant à la sélection à chaque locus d'allèles particuliers; et la différenciation phénotypique, traduite par la convergence du phénotype des individus de cette population vers des valeurs particulières. Lorsque l'on mélange deux populations différentes, on crée artificiellement une corrélation entre les fréquences des allèles, fixées à des niveaux différents entre les deux populations, et les moyennes des traits phénotypiques divergeant entre ces deux populations.

Les études d'association utilisent des collections avec des populations diverses. Au sein de ces collections, les individus donc à des populations différentes (on parle d'échantillon structuré). Cette structuration, lorsqu'elle est ignorée dans les modèles statistiques, produit l'effet décrit plus haut : la distribution des allèles différenciant les populations va être virtuellement corrélée avec le phénotype, pour tous les traits ayant des moyennes différentes selon la population. Il devient donc difficile de faire la distinction entre les vrais allèles responsables d'un phénotype et les allèles associés au phénotype par le seul effet du biais statistique. Le partage de fond génétique commun entre individus sur la base de liens familiaux induit également de la covariance phénotypique entre les individus les plus proches, et cela produit un effet de confusion similaire à celui de la subdivision des populations.

Les modèles statistiques élémentaires sont généralement construits sur l'hypothèse d'indépendance entre individus. En général, ces modèles sont donc inappropriées pour les études d'associations avec des populations structurées, dans lesquels les individus ne sont pas

[10] Myles *et al.* (2009) ont proposé les termes anglais respectifs de *population mapping* et *family mapping* pour tenter de lever cette équivoque terminologique. Cependant, le terme *association mapping* demeure le plus couramment utilisé pour désigner de façon restrictive les études d'association basées sur des populations. Ce déficit de précision terminologique ne semble pas poser d'ambiguïté, au moins aux spécialistes. Nous garderons donc comme terminologie la traduction française directe de *association mapping*, c-à-d. *cartographie d'association*, tout en faisant l'effort, au besoin, de rajouter des précisions permettant de ne pas confondre les deux méthodes. Pour les études basées sur des croisements contrôlés, le terme *linkage mapping* (traduit à la lettre par *cartographie de liaison*) reste un des plus utilisés : nous l'adoptons.

indépendants. L'application de ces modèles élémentaires conduit à des associations virtuelles, pour lesquelles il n'y a pas de vrai lien fonctionnel entre le phénotype et les allèles détectés (*faux positifs*). L'utilisation de modèles prenant en compte les relations entre individus est nécessaire afin de réduire le taux de faux positifs. Ces relations sont souvent *a priori* inconnues dans les échantillons, et leur inférence se pose donc comme un préalable aux études d'association (Pritchard *et al.* 2000b). Nous discutons dans ce qui suit des concepts et des outils analytiques utilisables pour analyser et inférer la structure d'un panel d'association[11].

Méthodes d'inférence de la structure génétique

Deux dimensions analytiques pour inférer la structure génétique

Les relations entre individus au sein des panels sont l'aboutissement de leurs histoires évolutives (partage d'une population d'origine commune, partage d'un pédigrée commun). Selon cette histoire évolutive, divers patterns de structuration peuvent exister dans ces échantillons. Deux dimensions ont été proposées pour appréhender cette structuration (Yu *et al.* 2006). La première dimension est l'appartenance des individus à différentes populations. Sous l'effet de l'adaptation locale ou de la sélection diversifiante, chaque population tend à fixer ou à éliminer des allèles particuliers; les individus issus de cette population se distinguent ainsi des autres individus par leurs fréquences alléliques à différents loci. Ils partagent de ce fait un fond génétique plus similaire entre eux, comparés à des individus provenant d'une autre population. Cette dimension de structuration est désignée par le concept de *structure des populations*.

La seconde dimension de structuration au niveau des panels d'association est le lien familial entre individus. Ce lien provient du partage d'un ancêtre commun à une échelle de temps plus ou moins courte (pédigrée). Il se traduit génétiquement par la présence, au sein des individus liés, d'allèles identiques hérités de cet ancêtre commun (*Identity by descent*). Cette deuxième dimension permettant de décrire les relations entre individus est désignée par le concept d'apparentement (ou *kinship*) [12].

[11] Nous utiliserons souvent le terme *panel* pour désigner les échantillons utilisés en cartographie d'association. Cet usage est très courant dans la littérature.

[12] La traduction française *apparentement* a été proposée, même si l'anglicisme *kinship* reste d'usage commun même en milieu francophone.

La structure des populations et l'apparentement rendent décrivent des liens qui ne sont pas toujours indépendants. Selon l'histoire évolutive des populations, il est possible que les liens familiaux soient plus forts au sein d'une même population, et dans ce cas les deux dimensions ne sont pas indépendantes (appartenir à une population peut présager des liens familiaux). Il est aussi possible que les liens d'apparentement transcendent les populations, rendant dans ce cas indépendantes les deux dimensions.

Dans le cadre des études d'associations, deux ensembles de méthodes sont proposés pour analyser et inférer la structuration des panels. Le premier ensemble comprend des méthodes bayésiennes et des méthodes de type *analyse en composantes principales* (ACP), qui permettent d'analyser la subdivision du panel en populations. Le second ensemble comprend des méthodes permettant d'analyser l'apparentement entre tous les individus pris deux-à-deux.

Toutes les méthodes qui seront présentées ici infèrent la structure essentiellement sur la base de données moléculaires multilocus. En général, des marqueurs choisis au hasard, sans lien *a priori* avec le phénotype, sont utilisés pour ces analyses (*random markers* ou *background markers*).

Méthodes bayésiennes et méthodes de type ACP

Une des méthodes les plus utilisées pour analyser la structure est sans doute la méthode bayésienne STRUCTURE (Pritchard *et al.* 2000a). STRUCTURE considère un modèle génétique avec K populations en équilibre de Hardy-Weinberg, définie chacune par une série de fréquences alléliques. Partant des données moléculaires, la reconstitution des populations est faite suivant un processus bayésien à travers lequel sont inférées itérativement les fréquences alléliques des populations, la population d'origine de chaque allèle et l'appartenance des individus aux populations respectives. Le modèle le plus simple assigne chaque individu à une seule population et considère donc l'absence de mélange (*admixture)* entre populations (*No-admixture model*)[13]. L'extension de ce modèle a permis d'implémenter l'inférence de populations avec des individus présentant une origine (ascendance, *ancestry* en anglais) partagée entre plusieurs populations (*Admixture model*). Dans ce cas, le lien d'un individu avec chacune des K populations est défini de façon relative par un coefficient d'ascendance (ou *ancestry coefficient*).

[13] L'anglicisme *admixture* désigne le fait que deux populations partagent en commun des individus suite à des flux de gènes. Nous y ferons allusion en utilisant les expressions *mélange de populations* ou *populations mélangées*. Dans la littérature, l'adjectif est également rapporté à un individu, pour signifier que cet individu a une origine *hybride*, partagée entre plusieurs populations (*admixed individual*).

Ce coefficient estime la proportion d'allèles que l'individu aurait hérité de chacune des K populations. Des extensions du modèle de base de STRUCTURE ont été proposées pour prendre en compte divers paramètres, notamment : le déséquilibre de liaison (DL) entre les marqueurs (Falush *et al.* 2003); l'ambiguïté génotypique inhérente à l'utilisation de marqueurs dominants ou à la polyploïdie (Falush *et al.* 2007); et les informations préalables tel que l'origine géographique des accessions (Hubisz *et al.* 2009). L'utilisation d'informations comme l'origine géographique permet de mieux identifier des populations à effectif réduits, des populations génotypées avec peu de marqueurs, ou encore de populations présentant un faible niveau de divergence.

Par ailleurs, notons que le système de reproduction peut conduire à une déviation des fréquences alléliques au sein des populations, par rapport à ce qui est attendu sous les hypothèses de Hardy-Weinberg. Dans le cas d'une consanguinité ou d'une autofécondation assez fortes, l'hypothèse d'équilibre de Hardy-Weinberg assumée par STRUCTURE est mise à mal et cela peut conduire à des signaux erronés de mélange entre populations (Gao *et al.* 2007). Une extension de STRUCTURE, implémentée sous le programme INSTRUCT, a été développée afin d'améliorer l'inférence en présence de consanguinité ou d'autofécondation (Gao *et al.* 2007).

La détermination du nombre de populations (K) reste un élément délicat dans les analyses bayésiennes de structure. STRUCTURE et INSTRUCT utilisent l'instruction de l'utilisateur pour fixer K. Il peut sembler trivial de fixer la valeur de K sur la base d'informations comme l'origine géographique des individus ou la structure phénotypique du panel. Mais le plus souvent, les regroupements géographiques ou la structuration morphologique ne recoupent pas forcément la structure génétique. Des groupes morphologiquement homogènes peuvent cacher une différenciation génétique fine (structure cryptique). La démarche souvent adoptée consiste alors à tester un intervalle de valeurs possibles de K et de comparer les résultats des simulations respectifs pour identifier la valeur de K optimale. Sous STRUCTURE, la courbe d'évolution du log-likelihood en fonction de K peut montrer dans certains cas un maximum ou un plateau à un point donné, que l'on considère alors comme la valeur de K optimale (Pritchard *et al.* 2000a). Cette méthode n'est pas toujours efficace, ce qui a motivé le développement d'une autre méthode pour le choix de K (Evanno *et al.* 2005). Cette deuxième méthode est basée sur la variation de second ordre du log-likelihood, qui semble plus efficace pour certaines données (Camus-

Kulandaivelu *et al.* 2007). Il reste cependant que ces deux méthodes *ad hoc* ne donnent pas une garantie d'efficacité dans toutes les situations. Sous INSTRUCT, un choix du nombre optimal de populations est suggéré à l'aide d'un critère d'information, le *DIC* (*Deviance information criterion*). Enfin, une autre approche bayésienne (STRUCTURAMA) a été proposée (Huelsenbeck et Andolfatto 2006). Avec cette approche, le nombre de populations K de même que l'assignation des individus aux populations sont considérés comme des variables aléatoires. Cette méthode permet d'associer une probabilité à chaque espérance de valeur de K, sachant les données.

Ces méthodes bayésiennes ont fait preuve d'efficacité pour inférer la structure des populations. Néanmoins, elles requièrent un temps de calcul relativement important et sont dépendantes du modèle génétique assumé. Les méthodes plus classiques de type ACP offrent le double avantage de requérir un temps de calcul plus court et de ne nécessiter aucun modèle génétique. L'ACP regroupe les individus en fonction de la similarité des fréquences alléliques. Les axes principaux les plus structurants peuvent être déterminés en testant statistiquement la significativité des valeurs propres correspondantes (Patterson *et al.* 2006). Le nombre d'axes significatifs pourrait être interprété comme le nombre de populations supporté par les données. Des méthodes plus ou moins similaires à l'ACP, notamment le *nonmetric multidimensional scaling*, ont également été proposés (Zhu and Yu 2009).

L'analyse bayésienne permet d'obtenir des matrices de populations (ou *matrices Q*) au sein desquelles sont définis les coefficients d'ascendance (*ancestry*) pour chaque individu. Dans le cas de l'ACP, l'appartenance aux groupes est déterminée par les coordonnées individuelles sur les axes principaux retenus (*matrices P*).

Méthodes d'inférence de l'apparentement

Le deuxième niveau de structuration au sein des panels est la présence de relations familiales entre les individus. Ces relations peuvent peut être estimées par le coefficient relatif d'apparentement. L'estimation de l'apparentement à partir des marqueurs moléculaires serait plus précise que le calcul à partir des informations de pédigrée, car elle prend en compte l'écart des contributions parentales par rapport à l'attendu (Bernardo *et al.* 1996). A la base, l'estimation de l'apparentement entre deux individus devrait s'appuyer sur la fréquence d'allèles identiques hérités d'un ancêtre commun (*identity by descent*). Le coefficient d'apparentement estimerait dans ce cas la probabilité d'identité par descendance (Hardy O and X Vekemans 2007). Cependant, lorsque l'on compare deux individus, les allèles communs sur leurs génotype actuels (identité d'état ou *identity by state*) ne sont pas tous hérités d'un ancêtre commun (*identity by descent*). Les données moléculaires brutes renseignent sur l'identité d'état, mais ne disent pas directement si les allèles identiques sont issus d'un même ancêtre commun. Les méthodes d'estimation de l'apparentement à partir des marqueurs moléculaires diffèrent ainsi sur l'approximation utilisée pour statistiquement estimer (ou non) la fraction d'allèles hérités d'un ancêtre commun et la fraction d'allèles dont l'identité n'induit pas de lien familial.

Le programme SPAGEDI (Hardy et Vekemens 2002) a implémenté différentes méthodes de calcul de l'apparentement à partir de marqueurs dominants (notamment Hardy 2003) ou de marqueurs codominants (notamment Loiselle *et al.* 1995 et Ritland 1996). (Ritland 1996) définit l'apparentement entre deux individus i et j comme Kij= (Qij – Qm)/(1-Qm); où Qij est la probabilité d'identité d'état pour deux loci pris au hasard chez les individus i et j, et Qm est probabilité moyenne d'identité d'état pour des loci comparés chez deux individus pris au hasard dans l'échantillon. On voit que le coefficient d'apparentement est ici un coefficient relatif, comparant la probabilité d'identité d'état entre deux individus i et j avec la probabilité d'identité d'état obtenue par hasard. L'approximation ici est de considérer la probabilité d'identité d'état entre des individus pris au hasard. L'estimation dépend de l'échantillon considéré. Cette définition de l'apparentement conduit parfois à des valeurs négatives, ce qui suggérerait que les deux individus comparés seraient moins liés que deux individus pris au hasard. Ces valeurs négatives sont souvent remplacer par zéro dans les études d'association.

Zhao *et al.* (2007) ont définit la proportion d'allèles communs entre deux individus (identité d'état) comme une estimation alternative de l'apparentement. Ils font l'hypothèse qu'il n'y a pas en fait d'individus non liés dans le contexte des panels d'associations, donc tous les allèles seraient hérités d'un ancêtre commun. Aussi, pour des marqueurs avec un taux de mutation faible, l'identité d'état impliquerait l'identité par descendance (Zhao *et al.* 2007). L'hypothèse de lien systématique entre tous les individus ne serait valide que si l'on remonte assez loin dans l'histoire évolutive. Pour cette méthode, on ne considèrera donc sans lien que des individus qui ne partagent aucun allèle commun. Cela marque une différence avec le calcul présenté plus haut (Riltand 1996), où l'importance du lien est plutôt appréciée de façon relative en comparaison à un échantillon aléatoire.

Stich *et al.* (2008) suggèrent de définir l'apparentement comme Kij=1+(Sij-1)/(1-T), où T est la probabilité conditionnelle que deux allèles soit identiques (identité d'état), étant donné qu'ils ne sont pas identiques par descendance. T est défini comme une variable dont la valeur est inconnue et distribuée entre 0 et 1. Une approche basée sur le *restricted maximum likelihood (REML*) a été utilisée pour estimer la valeur optimale de T dans le contexte des études d'association. Cet optimum correspond à la valeur de T pour laquelle l'incorporation de la matrice d'apparentement dans le modèle d'association donne le meilleur ajustement avec les données (*REML* le plus élevé). L'avantage conceptuel de cette approche vient du fait qu'elle s'appuie sur un critère statistique (l'ajustement), et la définition d'individus non liés (qui sert de repère pour le calcul de l'apparentement) n'est pas fixé arbitrairement mais suggérée par les données.

Modèles statistiques d'association

Présentation des modèles statistiques

L'effet confondant de la structure a été le principal écueil pour le développement des approches de génétique d'association chez les plantes. La solution méthodologique pour lever cet écueil a consisté à inférer la structure génétique puis à intégrer les matrices inférées dans les modèles statistiques (Pritchard *et al.* 2000b). Chez les plantes, un modèle pionnier basé sur la régression logistique avait été proposé (Thornsberry *et al.* 2001). Ce modèle teste l'effet d'un gène (en utilisant un SNP par exemple) en prenant en compte l'effet des populations modélisé à travers les matrices Q[14]. D'autres types de modèles statistiques, notamment le modèle linéaire généralisé (GLM), peuvent également incorporer ce type de matrices (Bradbury *et al.* 2007, Buckler *et al.* 2007).

[14] *SNP*, ou *single nucleotide polymorphism :* correspond à la différence, sur une seule paire de base, de la séquence d'ADN de deux individus (ou de deux haplotypes).

Pour traiter des panels avec des niveaux de relations plus complexes entre individus, la prise en compte des matrices d'apparentement (matrices K) a été suggérée, à travers le modèle linéaire mixte (Yu *et al.* 2006). Ce modèle a été formulé comme suit :

$y=X\beta+S\alpha+Qv+Zu+e$,

où y est le vecteur des phénotypes, β est un effet fixe autre que l'effet du SNP ou du background des populations, α est le vecteur d'effets fixes du SNP, v est le vecteur d'effets fixes du fons génétique des populations, u est le vecteur d'effets aléatoires du fond génétique des lignées, e est le vecteur des erreurs résiduelles. Q représente les matrices de populations. X, S, et Z sont des matrices binaires d'occurrence (0/1) qui relient y aux vecteurs β, α, et u. La variance des effets aléatoires du fond génétique est donnée par $Var(u) = 2KV_{g;}$ K est une matrice de dimension n x n qui évalue l'apparentement pour les n individus et V_g est la variance génétique. La variance des effets résiduels est $Var(e)=RV_{R;}$ R est une matrice de dimension n x n avec en dehors de la diagonale des valeurs de 0 et en diagonale des valeurs réciproques au nombre d'observations sous-tendant chaque point de données; V_R est la variance résiduelle. Le modèle mixte permet ainsi de considérer à la fois les subdivisions du panel (Q) et les relations familiales entre individus (K).

L'utilisation des matrices P de composantes principales (ou de composantes issues du *nonmetric multidimensionnal scaling*) a également été proposé comme alternative aux matrices de populations Q (Price *et al.* 2006, Zhao *et al.* 2007, Zhu and Yu 2009, Patterson *et al*. 2006).

Le Tableau G1 résume la façon dont les différents modèles prennent en compte la structuration du panel.

Tableau G1 : Modèles statistiques prenant en compte l'effet de la structure

Type de modèle	Contrôler de la structure	Exemple de modèle
Q	L'effet de la structure est ajusté avec une matrice de populations Q	Régression logistique, GLM
P	L'effet de la structure est ajusté avec une matrice de composantes principales P	Régression logistique, GLM
K	L'effet de la structure est ajusté avec une matrice d'apparentement K	Modèle linéaire mixte
Q+K	L'effet de la structure est ajusté avec une matrice Q et une matrice K	Modèle linéaire mixte
P+K	L'effet de la structure est ajusté avec une matrice P et une matrice K	Modèle linéaire mixte

Avant de discuter la performance relative de ces modèles, nous présentons trois critères statistiques par rapport auxquels on peut comparer la performance des différents modèles : le taux de faux positifs, la puissance, et l'ajustement aux données.

Taux de faux positifs

Le taux de faux positifs correspond dans le langage statistique à l'erreur de type I. Le taux d'erreur de type I est la fréquence à laquelle on rejette l'hypothèse nulle (H_0) alors qu'elle est vraie. Dans le contexte des études d'associations, l'hypothèse H_0 peut être formulée dans le cas général comme suit : *l'allèle testé n'a pas d'effet significatif sur le phénotype*. En rejetant cette hypothèse alors qu'elle est vraie, on déclare comme significatif un allèle qui ne l'est pas : cet allèle est donc un faux positif, un allèle déclaré significatif par erreur, alors qu'il n'a pas de lien fonctionnel avec le phénotype.

L'estimation du taux de faux positifs associé à un modèle donné est souvent réalisée empiriquement dans les études d'associations, sur la base de N marqueurs pris au hasard le long du génome (*random markers*). L'association entre le trait et chacun des marqueurs aléatoires est analysé avec ce modèle pour le trait d'intérêt en question. La distribution nulle des p-values attendue en considérant que les marqueurs ont été choisis aléatoirement sans lien avec le phénotype (pour N suffisamment grand) est une distribution uniforme entre 0 et 1. Dans ce cas, le taux de tests significatifs attendu est égal au seuil de p-value fixé, α. Pour α=0.05, par exemple, on attend seulement 5% de tests significatifs. On compare alors le taux empirique d'allèles aléatoires significatifs avec ce taux attendu. Lorsque le taux empirique est supérieur à ce qui est attendu, cela indique que le modèle est biaisé vers un excès de faux positifs. Le niveau de différence avec le taux attendu indique la gravité du biais. L'écart entre la distribution empirique des p-values obtenues et la distribution nulle peut être appréciée en représentant graphiquement le nuage de points entre ces deux distributions. Un modèle parfait produirait une courbe correspondant à la diagonale y=x.

L'absence de lien entre les allèles aléatoires et le phénotype constitue dans certains cas une hypothèse conservative. Il est attendu en réalité qu'une partie des allèles choisis aléatoirement le long du génome soient associés au trait phénotypique. Dans ce cas, un écart minimal à la distribution nulle serait attendu. Cet écart est confondu avec le biais éventuel du modèle, ce qui conduit à une surestimation du biais. Cependant, lorsque l'on compare plusieurs modèles, l'appréciation du biais est surtout relative, et l'on choisit les modèles dont la distribution empirique de p-values s'approche le plus de la distribution nulle.

Puissance du modèle

Dans les études d'association, la puissance du modèle détermine la capacité à identifier les allèles réellement liés au phénotype (ou *vrais positifs*). Cette puissance est mesurée statistiquement par 1−β, β étant le taux d'erreur de type II. L'erreur de type II consiste à accepter l'hypothèse H_0 alors qu'elle est fausse. Dans le contexte des études d'association, le taux d'erreur de type II correspond à la fréquence de *faux négatifs*, c.-à-d. des allèles déclarés non significatifs alors qu'ils sont réellement associés au phénotype.

L'approche communément utilisée pour évaluer la puissance des tests dans les études d'association consiste à simuler des effets et à déterminer la proportion détectée avec le modèle. Les effets sont simulés en faisant varier plusieurs paramètres, notamment la fréquence allélique, la taille de l'effet phénotypique, l'effectif de l'échantillon et l'héritabilité du trait. Ces simulations permettent d'évaluer la puissance du modèle en fonction des paramètres. En général, ces résultats sont visualisés avec des courbes montrant la variation de la puissance selon les paramètres (*abaques de puissance*)[15]. Ces abaques permettent de mettre en évidence les gammes de conditions dans lesquelles les associations seraient le plus fortement détectables pour un modèle donné. Cette approche permet de comprendre les limites du modèle, c.-à-d. les conditions dans lesquelles le modèle n'est pas puissant pour identifier les effets même s'ils existent[16].

Ajustement du modèle

Les mesures d'ajustement permettent d'évaluer le niveau d'écart entre le modèle et les données. Dans le cadre des études d'association, le recours aux mesures d'ajustement permet de comparer différents modèles par rapport à un jeu de données particulier. La comparaison de modèles a aussi un avantage méthodologique car elle aide à définir les paramètres les plus pertinents du modèle sur la base du principe de parcimonie. La parcimonie implique de ne garder dans le modèle que les paramètres strictement nécessaires pour ajuster convenablement les données. La définition de cette série de paramètres peut poser problème dans le cas de dispositifs complexes où les facteurs descriptifs à la base sont très nombreux et ne sont pas tous très informatifs (modèle multi-essai par exemple). Dans le cas du modèle linéaire mixte, divers critères de mesure d'ajustement sont disponibles selon les effets comparés (effets fixes et/ou effets aléatoires) et selon la méthode utilisée pour estimer les paramètres de variance (*maximum likelihood*, *ML* ou *restricted maximum likelihood*, *REML*).

Sous *ML*, le *likelihood ratio test* est l'un des tests les plus utilisés pour comparer des modèles imbriqués qui diffèrent pour des effets fixes et/ou aléatoires[17]. La statistique du test (D) est le double de la différence de log-likelihood entre les deux modèles. La distribution de probabilité de cette statistique peut être approximée par la distribution de Khi-deux. Le degré de liberté du test est égal à la différence de nombre de paramètres entre les deux modèles. Des p-values plus précises peuvent être calculées par simulation lorsque la distribution de Khi-deux semble inappropriée (Faraway 2006). Les critères d'information tels que AIC (Akaike 1974) et BIC

[15] Les abaques sont définies en général comme des outils permettant de faciliter un calcul. Les abaques de puissance permettent de déterminer la puissance du modèle pour différentes combinaisons de paramètres (MILLOT 2009, p. 222).

[16] Il est à préciser que différents tests de significativité peuvent être appliqués après avoir ajusté les données avec un modèle donné. Ces tests peuvent eux aussi influer sur la puissance. La puissance définie est donc en toute rigueur associée conjointement au modèle statistique proprement dit et au test de significativité appplıqué, et non au modèle tout seul.

[17] Deux modèles sont dits imbriqués si le plus large des deux modèles est défini en rajoutant des paramètres supplémentaires au modèle le plus simple. L'anglicisme *modèles nestés* (de *nested*) revient aussi souvent dans la terminologie francophone.

(Schwarz 1978) permettent de comparer les modèles même si ces modèles ne sont pas imbriqués[18]. AIC et BIC appliquent des pénalités pour chaque paramètre supplémentaire dans le modèle. BIC prend aussi en compte l'effectif de l'échantillon.

REML donne une estimation non biaisée des paramètres de variance du modèle. L'utilisation de REML peut donc être préférée à ML, surtout lorsque l'intérêt porte sur les composantes de la variance (Verbeke et Molenberghs 2000). REML prend en compte l'ajustement des effets fixes (et donc la perte de degrés de liberté associée) pour déterminer les paramètres de variance. Lorsque deux modèles diffèrent dans les effets fixes, le log-likelihood obtenu est sous-tendu par des séries d'observations différentes; il n'est donc pas considéré comme directement comparable. Pour cela, il est en général considéré que seuls des modèles ayant des effets fixes strictement identiques peuvent être comparés sous REML, en utilisant le *likelihood ratio test*, ou en utilisant des critères d'informations basés sur le log-likelihood (Gilmour *et al.* 2006). Mais cette considération a fait l'objet de discussion et pourrait nécessiter plus d'examen méthodologique (Gurka 2006a).

Comment choisir le meilleur modèle statistique ?

Le choix du modèle statistique pour une analyse d'association devra répondre notamment à cette triple exigence : limiter le taux de faux positifs, offrir une puissance acceptable, et s'ajuster le mieux aux données. Les modèles présentés plus hauts (Tableau G1) prennent tous en compte les relations entre individus. *A priori*, ces modèles sont tous supérieurs aux modèles pour lesquels aucun contrôle de la structure génétique n'est défini, notamment en termes de taux de faux positifs. Toutefois, les niveaux de performance obtenus sont variables, à cause des différences quant à la façon dont chaque modèle intègre la structure génétique. Plusieurs études ont proposé une comparaison plus ou moins exhaustive de ces modèles (Yu *et al.* 2006, Stich *et al.* 2008, Zhao *et al.* 2007). Globalement, les modèles associant les matrices de structure de populations avec les matrices d'apparentement tendent à produire une meilleure performance globale en termes d'optimum combinant la limitation du taux de faux positifs, la puissance et l'ajustement. Les modèles basés sur les matrices d'apparentement seules sont généralement supérieurs aux modèles basés sur les matrices de populations seules; en général, ces modèles basés sur l'apparentement ont une performance similaire ou légèrement différente à celle des modèles de

[18] AIC : Akaike information criterion. BIC : Bayesian information criterion.

type Q+K. Différentes méthodes existent pour calculer les matrices de populations et les matrices d'apparentement. Chaque méthode peut tient sur des postulats et des approximations spécifiques (voir plus haut). La comparaison des types de modèles peut donc être nuancée en prenant en compte les méthodes d'inférence utilisées. Par ailleurs, ces tendances générales restent aussi à nuancer selon la spécificité des jeux de données analysées.

On peut penser que des estimations plus précises de l'apparentement pourraient, à terme, suffire pour modéliser l'effet de la structure sans nécessairement utiliser des matrices de populations (Myles *et al.* 2009, Zhao *et al.* 2007). Les mesures d'ascendance mettent en évidence des subdivisions assez grossières du panel. Les matrices d'apparentement rendent compte des niveaux de relations plus complexes entre tous les individus deux à deux (Yu *et al.* 2006). Toutefois, la nécessité de combiner ces matrices avec les matrices de populations dépend du jeu de données (Yu *et al.* 2006, Stich *et al.* 2008).

En conclusion, il n'y a pas de modèle absolument supérieur. La démarche adoptée dans plusieurs études d'associations consiste à faire une comparaison assez exhaustive et contextualisée pour éclairer le choix final de méthodes d'inférence et du modèle statistique. Méthodologiquement, cette démarche semble la plus rigoureuse, et les critères développés ici pourraient permettre une telle démarche de choix de modèles.

Plan de la thèse

Les travaux de cette thèse seront présentés en trois chapitres.

Premièrement, une étude des bases génétiques de la variation de la floraison chez le mil sera présentée (Chapitre 2). L'étude sera structurée en 3 trois grandes parties : i) développement d'un approche statistique pour des études d'association chez le mil; ii) étude du pattern de déséquilibre de liaison et du pattern d'association génotype-phénotype sur une large zone génomique autour du gène *PHYC*; et iii) étude de la valeur sélective de deux gènes candidats de la floraison chez le mil en condition de stress hydrique simulé.

Ensuite, nous présentons une étude sur le développement de méthodes de cartographie d'association permettant l'étude des interactions génotype x environnement avec des populations structurées (Chapitre 3).

Enfin, une discussion finale est proposée, en deux étapes (Chapitre 4). D'abord, les avantages et les limites des différentes méthodologies d'association phénotype-génotype chez les plantes sont discutés, en faisant une ouverture sur les approches émergentes. Ensuite, les résultats de cette thèse sont discutés par rapport à la perspective de gestion des impacts du changement climatique sur la culture de mil.

CHAPITRE 2. ETUDE DES BASES GENETIQUES DE LA VARIATION DE LA FLORAISON CHEZ LE MIL

Introduction

Ce chapitre présente une étude des bases génétiques de la variation du cycle de floraison chez le mil. La date de floraison est un des traits principaux qui permettent l'adaptation du mil à différentes conditions climatiques. La connaissance des bases génétiques de ce caractère complexe permettra de mieux comprendre l'évolution adaptative de l'espèce, mais également de mieux gérer les impacts du climat sur la culture. Notre étude repose essentiellement sur des approches d'association génotype-phénotype.

Dans la première partie, nous présentons une étude de cartographie d'association basée sur un panel de 90 lignées de mil originaires d'Afrique et d'Inde, et sur un échantillon de 598 accessions de variétés locales nigériennes[19]. Cette étude comporte deux parties principales. La première partie est méthodologique et a permis de choisir des méthodes appropriées pour les études d'association phénotype-génotype chez le mil. D'abord, nous avons comparé les méthodes bayésiennes STRUCTURE et INSTRUCT pour inférer la structure génétique du panel à partir de données moléculaires (25 marqueurs microsatellites). Les coefficients d'apparentement ont été calculés sur la base de 306 marqueurs AFLP. Ensuite, nous avons comparés différents modèles statistiques d'association pour identifier les paramètres à prendre en compte afin de limiter le taux de faux positifs et obtenir une meilleure confiance dans les tests d'association. Enfin, des simulations ont été effectuées pour évaluer la puissance du modèle linéaire mixte en fonction de la fréquence allélique, de la taille des effets sur la date de floraison, et de la méthode d'inférence de la structure. Ces simulations ont permis d'évaluer la capacité du modèle, en fonction des paramètres, à détecter des allèles liés au phénotype.

Dans la seconde partie, nous avons illustré l'application de la méthodologie pour tester des gènes candidats et identifier ceux qui sont significativement liés au phénotype. Huit gènes candidats ont été analysés vis-à-vis de sept traits phénotypiques liés à la floraison, la morphologie de la plante et la morphologie de l'épi. Cette première étude a permis de mettre en évidence une association

[19] Cette étude est déjà publiée chez la revue de la *Genetics Society of America*. L'article est présenté dans sa forme publiée (*Genetics* 182 : 899–910, 2009).

significative entre des SNPs du gène *PHYC* et la variation de la date de floraison chez le mil. Cette méthode d'association a aussi été utilisée pour valider statistiquement l'association d'un nouveau gène candidat, *MADS11*, avec le phénotype chez le mil (Annexe 1). *MADS11* montre une signature moléculaire de sélection, et l'étude d'association a permis de monter son effet significatif sur la date de floraison.

Dans la deuxième partie de ce chapitre, nous présentons une analyse du pattern de déséquilibre de liaison et du pattern d'association génotype-phénotype dans une région génomique plus large autour du gène *PHYC*[20]. L'objectif principal de cette deuxième étude était d'avoir une analyse plus fine de la zone chromosomique autour de *PHYC* et de mettre en évidence les polymorphismes expliquant le mieux l'association observée. Six loci homologues de différents gènes de céréales (dont *PHYC*) ont été séquencés dans la région pour l'ensemble des 90 lignées du panel de mil. Pour cette étude, nous avons également développé la séquence complète du gène *PHYC* (~6kb). Un total de 75 marqueurs polymorphiques (SNP/INDEL) avec une fréquence allélique supérieure à 2.5% ont été identifiés sur l'ensemble des six loci[21]. L'analyse standard avec le modèle linéaire mixte a révélé des associations significatives entre plusieurs de ces marqueurs et la variation phénotypique. Un déséquilibre de liaison très fort a été mis en évidence sur une partie de cette zone chromosomique, suggérant que l'association observée avec plusieurs loci est induite par la liaison. Nous avons développé une méthode de comparaison basée sur une chaîne de Markov (MCMC) afin d'identifier parmi ces marqueurs ceux qui présentent la plus forte probabilité d'association avec le phénotype. Nos résultats suggèrent que *PHYC* est le meilleur candidat pour expliquer l'association avec le phénotype. Nous avons également réalisé une analyse de QTL sur trois familles F2 respectives. Cette étude confirme la présence, dans la région de *PHYC*, de QTL pour les traits analysés.

[20] Cette deuxième partie constitue un projet d'article en préparation.

[21] Les INDELs sont des variations de séquences caractérisées par l'insertion ou la délétion chez certains individus d'un fragment d'ADN dont la taille peut varier d'une paire de bases à plusieurs paires de bases.

La troisième partie de ce chapitre analyse le lien de deux gènes candidats de floraison (*PHYC* et *MADS11*) avec les composantes de fitness. L'étude est effectuée en conditions expérimentales de gradient hydrique, avec 159 lignées recombinantes (RILs)[22].

[22] Les *RILs*, ou *recombinant inbred lines*, sont des lignées obtenues en autofécondant sur des générations successives des individus echantillonnés dans la descendance d'un même croisement initial.

DOI: 10.1534/genetics.109.102756

Association Studies Identify Natural Variation at *PHYC* Linked to Flowering Time and Morphological Variation in Pearl Millet

Abdoul-Aziz Saïdou,[*,†,‡] **Cédric Mariac,**[*,†] **Vivianne Luong,**[*] **Jean-Louis Pham,**[*]
Gilles Bezançon[†] **and Yves Vigouroux**[*,†,1]

[*] *Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/Sup-Agro, BP64501, 34394 Montpellier, Cedex 5, France,* [†] *Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/Sup-Agro, BP11416, Niamey, Niger and* [‡] *University Abdou Moumouni, BP 11040, Niamey, Niger*

Manuscript received March 13, 2009
Accepted for publication May 4, 2009

ABSTRACT

The identification of genes selected during and after plant domestication is an important research topic to enhance knowledge on adaptative evolution. Adaptation to different climates was a key factor in the spread of domesticated crops. We conducted a study to identify genes responsible for these adaptations in pearl millet and developed an association framework to identify genetic variations associated with the phenotype in this species. A set of 90 inbred lines genotyped using microsatellite loci and AFLP markers was used. The population structure was assessed using two different Bayesian approaches that allow inbreeding or not. Association studies were performed using a linear mixed model considering both the population structure and familial relationships between inbred lines. We assessed the ability of the method to limit the number of false positive associations on the basis of the two different Bayesian methods, the number of populations considered and different morphological traits while also assessing the power of the methodology to detect given additive effects. Finally, we applied this methodology to a set of eight pearl millet genes homologous to cereal flowering pathway genes. We found significant associations between several polymorphisms of the pearl millet *PHYC* gene and flowering time, spike length, and stem diameter in the inbred line panel. To validate this association, we performed a second association analysis in a different set of pearl millet individuals from Niger. We confirmed a significant association between genetic variation in this gene and these characters.

DOMESTICATION and dispersion of cultivated plants were associated with their adaptation to the agricultural environment. These adaptations led to genetic changes shared by all individuals of a cultivated species (domestication genes) or to variations between varieties within a cultivated species (genes controlling varietal differences). Domestication genes like *tb1* (Doebley *et al.* 1997; Wang *et al.* 1999) in maize (*Zea mays*) were selected very early by human populations (Jaenicke-Després *et al.* 2003). After the first early selection, adaptation of the flowering phenotype to different climatic conditions was certainly a key innovation that enabled colonization of new environments. One of the most well-known examples was the adaptation of maize—a tropical plant—to northern climates. Maize cultivation spread late to northeastern America. By 1000 YBP, only maize was an established staple crop (Fritz 1995). A genetic variant of the *Dwarf8* gene led to an earlier flowering phenotype (Thornsberry *et al.* 2001). This early allele was present at a high frequency in North America and was certainly selected after the domestication of maize under northern climatic conditions (Camus-Kulandaivelu *et al.* 2006).

Pearl millet (*Pennisetum glaucum* [(L.) R. Br.]), one of the most important West African cereals, was most likely domesticated once in the Sahelian zone of West Africa (Oumar *et al.* 2008). By 3500 YBP, it was already being cultivated throughout Sahelian and tropical West African countries (D'Andrea *et al.* 2001; D'Andrea and Casey 2002). The adaptation of pearl millet in West Africa was also associated with an environmental gradient (Haussmann *et al.* 2006). Pearl millet varieties from tropical coastal West Africa flower very late (up to 160 days from planting to female flowering) as compared to varieties from Sahelian West Africa, which may have a flowering time as short as 45 days (Haussmann *et al.* 2006). The genetic factors underlying the differences between these varieties are still unknown.

Association studies offer new opportunities for assessing the role of a particular gene on a phenotype. Contrary to QTL analysis, association studies have the challenging task of taking an unknown evolutionary history of studied individuals into account. For exam-

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.102756/DC1.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. FN376885–FN377564.

[1] *Corresponding author:* Institut de Recherche pour le Développement, 911 Ave. Agropolis, BP 64501, 34394 Montpellier, Cedex 5, France. E-mail: yves.vigouroux@mpl.ird.fr

ple, population structure is a common confounding effect in association studies (Pritchard *et al.* 2000a). Allele frequencies evolve between divergent structured populations via drift, mutation, and selection. Differences in allele frequencies may be correlated with any morphological traits that differentiate two populations. Then a statistical correlation between a gene and a trait is not necessarily associated with a "causative" relationship between the gene and the morphology, which can lead to a high number of false positives. The use of population structure to correct the number of false positives was a significant breakthrough in plant studies (Thornsberry *et al.* 2001). This approach was recently further refined by also using a matrix of kinship coefficients, which proves efficient when there is a complex structure and familial relationship between individuals (Yu *et al.* 2006; Kang *et al.* 2008; Stich *et al.* 2008). Complex structures and familial relationships are common in inbred cultivated crop material. In the current association study framework (Thornsberry *et al.* 2001; Yu *et al.* 2006; Casa *et al.* 2008; Kang *et al.* 2008; Stich *et al.* 2008), population structure was assessed using STRUCTURE software (Pritchard *et al.* 2000b). This tool is not implemented to deal with selfed inbred materials or inbred species (Pritchard *et al.* 2000b). Through new methodological developments, population structure analysis can now be performed using Bayesian methods in these particular cases (Gao *et al.* 2007). The extent to which the power of association studies will differ when dealing with inbred material or selfing species using either Bayesian method has yet to be evaluated.

In this study, we developed an association framework for pearl millet to assess the role of flowering pathway genes. We assessed the ability of the method to control the number of false positives, while taking different methodological inferences of population structure that allow inbreeding or not into account. We also assessed the power of the association framework to detect given additive genetic effects. Finally, we applied this method to a set of eight flowering time gene homologs sequenced in pearl millet. We assessed sequence variation in light perception genes (*PHYA*, *PHYB*, *PHYC*, and *CRY2*) and downstream regulators of flowering (*GI*, *Hd6*, *Hd1*, and *FLORICAULA*). Variation was detected in the *PHYC* gene associated with variations in flowering time and morphological traits. This association was noted in two different data sets.

MATERIALS AND METHODS

Field experiments: For the association framework, a set of 90 pearl millet inbred lines was used (supporting information, Table S1). These inbred lines had diverse origins: India and West and East Africa. They were obtained from T. Hash [International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hiderabad, India], J. Chantereau (Centre de Cooperation Internationale en Recherche Agronomique pour le Développement, Montpellier, France), and T. Robert and A. Sarr (University Paris XI, Paris).

These inbred lines were characterized in three experimental field trials during the rainy season. Planting dates were July 9, 2005, June 16, 2006, and July 13, 2006. Hereafter, we refer to these three different field trials as 2005, 2006a, and 2006b, respectively. The experiments were performed at the ICRISAT field station in Sadore, Niger. The plant spacing was 0.7 × 0.7 m. Inbred individuals from given inbred lines were sown in a row and the locations of inbred lines were randomized. For each pearl millet inbred line, data from 6–10 individuals were separately scored for days from planting to the female flowering stage (FT), the number of basal tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), basal primary spike diameter (BSpD), primary spike length (SpL), and primary spike diameter (SpD). Average values of each inbred line were calculated for each field trial and each morphological and phenological trait. To obtain an inbred line average trait effect for the total field trials, we fitted the mixed model $y_{ijkl} = \mu + x_i + z_j + v_{jk} + \varepsilon_{ijk}$, where y_{ijkl} was the phenotype of individual l of the i inbred line, in the j field trial, in the k subplot. The value ε_{ijk} was the residual error and μ the grand mean. Inbred lines (x_i) were considered as fixed effects and field trial (z_j) and subplot (v_{jk}) were considered as random effects. For each trait, the best linear unbiased effect (BLUE) was estimated for each inbred line i as $\hat{\mu} + \hat{x}_i$. The model was fitted using R (http://cran.r-project.org/) and the lmer() function. The BLUE of each trait was then used for association studies.

We also used a set of 598 different pearl millet varieties originating from Niger. These landraces were sampled throughout the country from 0°E to 13.3°E latitude and from 12°N to 15°N longitude (Table S1). Each landrace was sown in 2004 and 2005 during the rainy season at the ICRISAT field station in Sadore, Niger. The plant spacing was 0.7 × 0.7 m. For each accession, data from five individuals were recorded on flowering time from planting to female flowering stage, the number of tillers at head emergence, plant height, stem diameter, primary spike length, primary spike diameter, and thousand seed weight (TSW). The averages for each trait were calculated per accession for each field trial and used for association studies. We also used a BLUE estimate of each trait, using the procedure previously described for inbred lines.

SSR and AFLP genotyping: DNA was extracted from inbred lines and pearl millet varieties as previously described (Mariac *et al.* 2006b). Pearl millet inbred lines were genotyped three to four times, using a set of 27 microsatellite loci (Oumar *et al.* 2008) on plants from 2005 and 2006 field trials. The PCR conditions and methods were previously described (Oumar *et al.* 2008). Consensus genotypes were obtained as follows. If one of the four multilocus genotypes was markedly different from the three others, this genotype was excluded and classified as erroneous. If, for an inbred line, the multilocus was identical at most of the loci but a variation was observed at a given locus, the most frequent genotype was conserved. This variation was attributed to genotyping errors or to residual diversity segregating in the inbred lines. The same inbred lines were also genotyped with AFLP markers (Vos *et al.* 1995), using the method previously described in pearl millet (Allinne *et al.* 2008). A total of six primer pair combinations with three specific bases were used (E-AAC/M-CTT, E-ACA/M-CTT, E-AGC/M-CTT, E-ACA/M-CTG, E-AGC/M-CTG, and E-AAC/M-CTG). The letters E and M represent the sequences ACTGCGTACCAATTCAG and GAT GACTCCTGAGTAA corresponding to *Eco*RI and *Tru*I adapters, respectively. AFLP-Quantar (Keygen) software was used to identify and count the number of polymorphic bands. Two independent readings were performed per gel and only reliable loci were used. A total of 306 locus markers were identified.

For the second association population, an individual of each variety was genotyped with 25 microsatellite loci. A total of 598 different plants were genotyped. All varieties were genotyped according to a previously published (Mariac *et al.* 2006a) protocol and this data set has already been partially published (Mariac *et al.* 2006a).

Sequencing: Primers for partial amplification of eight flowering genes (Table S3) were designed or obtained from previously published studies (Mathews *et al.* 2000). Fragments ranging from 200 to 1175 bp in size were amplified by PCR with 0.2–0.4 μm of each primer, 0.5 units of Taq polymerase, 1× GoTaq Buffer (Promega, Madison, WI), 0.200 mm dNTP, and 20 ng genomic DNA in a 30-μl final volume. Amplifications were performed as follows: 35 cycles of 30 sec at 94°, 90 sec at 50°–64° (depending on the primer Tm), and 60 sec at 72°, ending with 10 min at 72°. PCR products were purified using Ampure kits (Agencourt Bioscience) and sequence reactions were performed using the BigDye v3.1 Terminator kit (Applied Biosystems, Foster City, CA). Sequence reactions were purified with CleanSeq kits (Agencourt Bioscience) and read on an ABI 3130 XL automated sequencer (Applied Biosystems). Forward and reverse sequences were obtained for inbred lines.

Sequence data analysis: To confirm amplification of the targeted gene, all gene sequence data obtained in pearl millet were confirmed using Blastn (MegaBlast) analysis. We calculated the percentage of polymorphic sites, the pairwise nucleotide diversity (π), Watterson's estimator (θ) of diversity, Tajima's D (Tajima 1989), and Fu and Li's D^* and F^* (Fu and Li 1993) using DNAsp version 4.10.3 (Rozas *et al.* 2003). All SNP and indel polymorphic sites were used for this analysis. The linkage disequilibrium and its significance were estimated on the basis of r^2, using TASSEL software version 2.0.1 (Buckler *et al.* 2007).

SNP genotyping: To genotype pearl millet varieties, a restriction assay using *Pvu*II was performed to recognize an SNP C/G at position 697 on the amplified *PHYC* fragment. The *PHYC* gene was amplified by PCR with 0.2 μm of each forward and reverse primer (Table S3), 0.5 units of Taq polymerase, 1× GoTaq buffer (Promega), 0.200 mm dNTP, and 20 ng genomic DNA in a 30-μl final volume. Amplifications were performed as follows: 35 cycles of 30 sec at 94°, 30 sec at 55°, and 60 sec at 72°, ending with 10 min at 72°. PCR reactions were digested with *Pvu*II as recommended by the supplier (Fermentas) immediately following amplification. About 10 μl of the digestion were loaded on a 2% (w/v) agarose gel for genotyping. Genotypes were scored as C/C, G/G, and C/G according to the digestion pattern.

Population structure analysis: *Bayesian methods:* For inbred lines, we analyzed the population structure using STRUCTURE (Pritchard *et al.* 2000b; Falush *et al.* 2003) and INSTRUCT (Gao *et al.* 2007) software. The number of populations tested ranged from $K = 1$ to $K = 10$. STRUCTURE runs were performed with 10^6 iterations and a burn-in period of 30,000. Ten independent simulations were performed. INSTRUCT parameters involved 200,000 iterations, including a burn-in of 100,000. INSTRUCT allows a different selfing rate for each individual plant and seems more appropriate for inbred line materials. For landraces, we used only the STRUCTURE method as pearl millet is an outcrossing cereal species. We varied the number of populations from $K = 1$ to $K = 5$ and 10 independent simulations were performed.

Choice of K and comparison of methods: For STRUCTURE, we used the method of Evanno *et al.* (2005) based on the second-order rate of change of likelihood. For INSTRUCT, we used the deviance information criterion (DIC) to infer optimal K (Gao *et al.* 2007). The results obtained by both methodologies were compared for each K value. To measure differences between INSTRUCT and STRUCTURE results, we compared the ancestry values for each population obtained with each method. For an individual i, let q_{ik} and q'_{ik} be the ancestry of individual i from STRUCTURE and INSTRUCT, respectively, where k is the population. The two methods gave relatively similar results and it was easy in the present case to associate the q_k and q'_{ik} values to "the same population," *i.e.*, a population that pooled a common set of individuals in the STRUCTURE and INSTRUCT results. We calculated a similarity index of ancestry per individual: $\mathrm{SI}_i = 1 - \sqrt{\sum_{k=1}^{K}(q_{ik} - q'_{ik})^2/K}$. We then calculated the average similarity index for all inbred lines: $\mathrm{SI} = (1/n)\sum_{i=1}^{n}\mathrm{SI}_i$. This index ranged from 0 if individuals were associated with different groups to 1 if the results obtained by both methods were identical. To compare the STRUCTURE and INSTRUCT results, we used the ancestry Q matrix obtained with the highest likelihood run.

Association studies: *Model:* We used a linear mixed model to determine associations between morphological traits and genetic variations (Yu *et al.* 2006). This model took into account (1) the population structure of the inbred lines based on the ancestry Q matrix of each individual inbred line in $K - 1$ populations and (2) the family relationship between individuals through the kinship coefficient matrix.

The association model was $y = X\beta + S\alpha + Qv + Zu + e$, where y was the phenotype vector, β was a fixed effect other than SNP or population structure, α was the vector of a given SNP fixed effect, v was the vector of population structure fixed effects, u was the vector of background genetic effects, and e was the residual error vector (Yu *et al.* 2006). Q was the population ancestry matrix. X, S, and Z were 0/1 matrices relating y to β, α, and u vectors. The variance of the random effect u was expected to be $\mathrm{Var}(u) = \mathrm{KM}\ V$, where KM is the kinship matrix and V the variance (Yu *et al.* 2006).

We used the kinship package (Atkinson and Therneau 2008; Ingvarsson *et al.* 2008) to implement the mixed-model approach. The mixed model was fitted using a maximum-likelihood method. Different nested models were assessed: the most complete model that included population structure ancestry and kinship matrix, models without kinship matrix or population structure, and a null model that disregards the population structure and kinship matrix. The different models were compared to the complete model by calculating a likelihood ratio Λ, and the $-2 \ln \Lambda$ value was statistically assessed for significance using a χ^2-distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models.

For the association analysis of an SNP with a trait, we used either the kinship package or the mixed-model method implemented in TASSEL (Buckler *et al.* 2007). The two methods gave similar results but the method implemented in TASSEL was particularly user friendly with respect to managing SNP, trait, matrix, and population structure data sets. For inbred lines, we used microsatellite loci to infer population structures and AFLP markers to calculate the kinship matrix. Kinship coefficients were calculated using SPAGeDI (Hardy and Vekemans 2002). Kinship coefficients lower than zero were set at zero. For pearl millet varieties, the kinship coefficient was calculated using the method of Loiselle *et al.* (1995) implemented in SPAGeDI (Hardy and Vekemans 2002). This method is adapted to heterozygote diploid individuals in the case of multiallele and multilocus data sets.

For the mixed-model analysis of the kinship package, the kinship matrix needs to be positive definite (Atkinson and

Therneau 2008); *i.e.*, all the matrix's eigenvalues need to be positive. However, kinship matrix estimations might lead to non-positive-definite matrices (Atkinson and Therneau 2008). To obtain a positive-definite matrix, we adapted an *ad hoc* procedure from Hayes and Hill (1981). With M being the non-positive-definite matrix, we need to find M', *i.e.*, a matrix highly correlated to M but positive definite with diagonal elements of 1 and only positive values for all elements. To obtain such a matrix, we decomposed the M matrix into its eigenvectors and eigenvalues. Eigenvalues lower than an arbitrary threshold of 10^{-4} times the higher eigenvalue were set to this threshold. There is at least one such element since a non-positive-definite matrix is defined as having a least one negative eigenvalue. A new matrix M' could then be rebuilt using the new eigenvalues and eigenvectors. The problem of this method is that the new matrix M' might have small negative values. To avoid this problem, we did not apply the procedure to M but rather to $M - \varepsilon$, with ε being a square matrix of the same size as M with all elements equal to a small negative value ε. A possible value for ε is the minimum value of M' (if negative) or 0. Using the previously described procedure, we obtained the matrix $(M - \varepsilon)'$. Each row of this new positive-definite matrix $(M - \varepsilon)'$ was then standardized, so the diagonal was 1. To measure the extent of the modification, a Spearman correlation between the initial matrix M and standardized $(M - \varepsilon)'$ matrix was calculated and compared using the Mantel test (Sokal and Rohlf 1991).

Assessment of type I error: We performed an analysis using microsatellite and AFLP alleles to assess the ability of the linear mixed-model (LMM) method to reduce type I errors for the inbred lines data set. We used all microsatellites of AFLP alleles having a frequency >2.5% to perform association studies. For each allele, the association between the presence or the absence of the allele and a trait was assessed. When the allele occurrence and phenotype are strictly independent, 5% of the alleles could be expected to have a significant association at the 5% level. This analysis was performed independently for three different phenotypic traits: flowering time, primary spike length, and primary spike diameter. We wanted to assess the extent to which taking the population structure and family relationship into account reduced the type I error. We thus considered a population number ranging from $K = 1$ (no structure) to $K = 7$ using the Q matrix obtained by the STRUCTURE and INSTRUCT methodological approaches. The kinship matrix (KM) obtained from AFLP data or a noninformative kinship matrix was also used. The uninformative matrix (UKM) was built by setting the relationship between two different individuals at 0 (no relatedness). The analysis output is a percentage of false positives for different inferences of population structure and family relationship (STRUCTURE + UKM, INSTRUCT + UKM, STRUCTURE + KM, and INSTRUCT + KM), for a different number of accepted populations, K ($K = 1$ to $K = 7$), for different phenotypic traits (FT, SpD, and SpL), and for the three field trials. The number of false positives was compared using the Kruskal–Wallis test. Paired data from AFLP-based false positive rates and SSR-based false positive rates were compared using Wilcoxon's paired tests.

Empirical P-value threshold: Taking the population structure and family relationship into account could, however, lead to a higher type I error rate than the commonly used 5% threshold. We therefore also calculated an empirical threshold. To do so, we used AFLP and microsatellite allele data to perform association studies taking the population structure ($K = 7$) and kinship matrix into account. To calculate a corrected threshold, the *P*-values associated with AFLP and microsatellite alleles were ordered from the lowest to the highest value. The corrected *P*-value threshold corresponded to the *P*-value associated with microsatellite or AFLP alleles having a rank of 5%. This value was specific to each phenotype/field trial, and we calculated a separate threshold on the basis of the AFLP and microsatellite data sets.

Power analysis: We performed a simulation analysis to assess the power of this methodology for detecting an additive effect in pearl millet. A set of inbred lines was used to create an SNP data set having a given flowering time effect. We first randomly attributed the causative SNP to an inbred line. Then, for each inbred line having the causative SNP, the flowering time value was increased by adding a certain amount of flowering time (in days). We used the best linear estimates of flowering times for all field trials. This additive effect ranged from 0 to 22 days. We also calculated this additive effect in terms of genetic effect ratio (Yu *et al.* 2006), *i.e.*, as a percentage of the flowering time standard deviation. The genetic effect ratio ranged from 0 to 2.9. We varied the frequency of the causative SNP allele in the inbred lines: frequencies of 50, 25, 12.5, 6.25, and 3.12%. One hundred random data sets were created for each given set of parameters (SNP frequency, a given additive effect). Association analyses using these data sets were performed to detect the SNP effect, using the mixed linear model with the INSTRUCT or STRUCTURE Q matrix for $K = 7$ and the kinship matrix. The percentage of tests that were significant (out of 100 data sets) at the 5% level was used as a measurement of the probability of detecting the SNP effect on the phenotype. This value was obtained for each SNP frequency (5 different values) and additive effect (21 values).

RESULTS

Pearl millet diversity and structure: Of the 27 microsatellite loci, 25 were polymorphic enough on the 90 inbred lines to be used for subsequent analyses. The total number of alleles detected was 188. An average number of 7.5 alleles per locus were found with an average gene diversity of 0.56. The observed heterozygosity was low (0.059) as expected for inbred materials. The data set structure was first estimated using STRUCTURE. The log-likelihood increased as K increased and did not show evidence of a maximum (Figure 1A). We calculated the second-order change in log-likelihood (Figure 1B) and found a strong signal for $K = 6$. On the basis of this result, we considered $K = 6$ as being the supported number of populations. INSTRUCT uses a deviance criterion to infer K. The DIC value was lowest for $K = 7$ ($\mathrm{DIC}_{K=7} = 6116.08$). We calculated a similarity index to assess the difference between the results of the two methods (Figure 1C). The average similarity index for all individuals was >82.5% regardless of the number of K populations. The highest value was obtained for $K = 4$ at 92% but then the similarity index tended to decrease to 82.5% for $K = 7$. Visual comparison of the output of STRUCTURE and INSTRUCT (Figure 1, D and E) showed an apparent similarity. However, numerous differences were noted and some individuals were grouped with different clusters.

A total of 306 AFLP markers were identified. The average gene diversity was 0.29. Kinship coefficients between 0 and 0.35 represented 99% of the data points of the distribution (Figure S1). A total of 67.5% of the

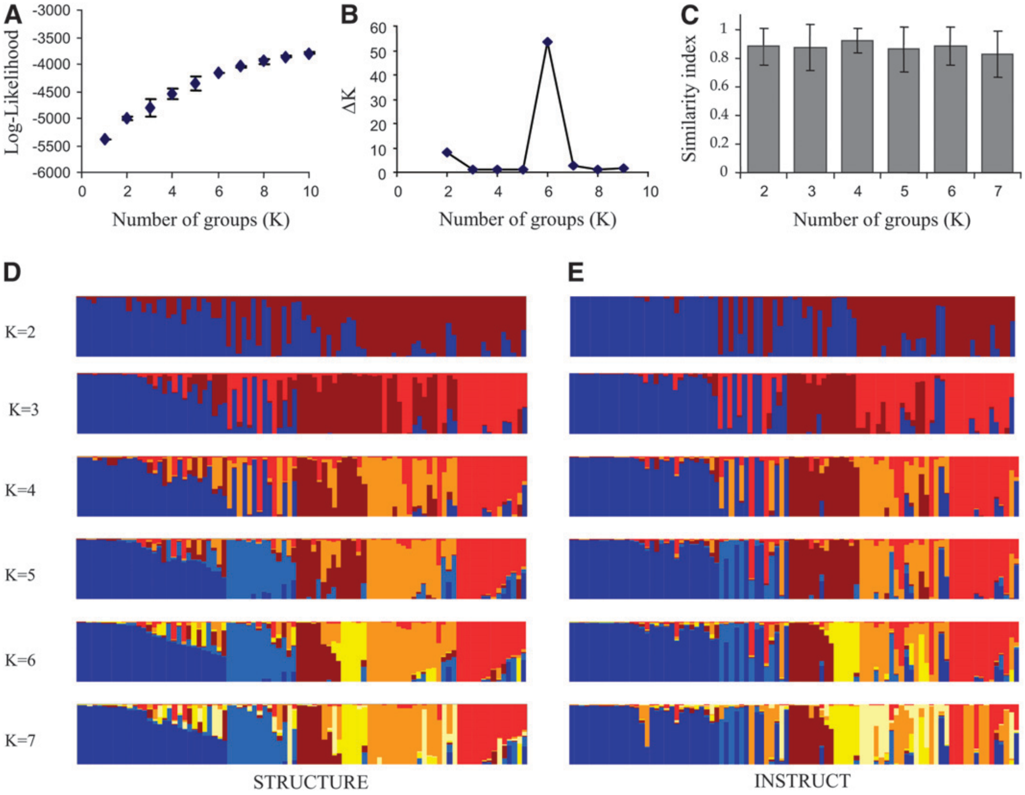


FIGURE 1.—Analysis of the population structure of pearl millet inbred lines. The analysis of population structure in inbred lines was performed using STRUCTURE (PRITCHARD *et al.* 2000b; FALUSH *et al.* 2003) and INSTRUCT (GAO *et al.* 2007). (A) The average log-likelihood and the standard error of 10 different runs of STRUCTURE were calculated. The log-likelihood showed a steady increase as the number of groups (*K*) increased, and no clear maxima were detected. (B) To assess the number of groups (*K*) supported by the analysis, we also calculated the second-order change in the log-likelihood Δ*K* (EVANNO *et al.* 2005). A clear change was detected for *K* = 6, suggesting six was the number of groups supported by the STRUCTURE analysis. To allow a comparison of the two Bayesian structure inference methods of STRUCTURE and INSTRUCT, we calculated a similarity index (see text for details). (C) The average similarity index and standard error values for each individual were reported. The average similarity index was >80% in most cases, suggesting similar inference of ancestry results for each plant. We finally represented the results of a run of STRUCTURE (D) and INSTRUCT (E) to enable direct visual comparison of the two methods. The graph (D) represents the run with the highest likelihood of STRUCTURE for a number of populations (*K*), and E represents the run of INSTRUCT that showed the lowest deviance information criterion (DIC). The ancestry (*q*) of each of the inbred lines in a population is represented by a different color. The different colors correspond to the different populations identified by STRUCTURE and INSTRUCT. The global visual comparison highlighted a global similarity, but some differences were clearly observed between the two analyses.

kinship coefficients suggested that there was no or low relatedness between inbred lines with kinship values ranging from 0 to 0.05. A significant fraction (31.5%) showed various degrees of relatedness, with kinship ranging from 0.05 to 0.35. Finally, only 1% showed relatedness >0.35. This relatedness was illustrated in a phylogenetic relationship between inbred lines (Figure S2). Few inbred lines showed weak genetic dissimilarity (and so high kinship), but a large share of the inbred lines were weakly related.

Morphological traits: The days to female flowering of inbred lines ranged from 35.0 to 98.0 days, with a mean of 58.8 (SE ±0.54) days for all field trials. The mean spike morphological values were 0.46 (SE ±0.54), 2.20 (SE ±0.034), and 25.6 (SE ±0.79) for basal primary spike diameter, primary spike diameter, and spike length, respectively. The number of basal tillers at heading date was 8.50 (SE ±0.20). Finally, the mean stem diameter was 1.04 (SE ±0.016) and the mean plant height was 83.9 (SE ±0.016).

Association study: We reported the likelihood for the different models considered, using STRUCTURE or INSTRUCT (Table 1). The complete model, including population structure and kinship matrix, is generally better than models with structure only or kinship only and always better than a null model (Table 1). Comparisons of nonnested models are generally based on the Akaike information criterion (AIC), with AIC = −2 log-likelihood + 2*k*, with *k* being the number of parameters. For our purposes, we wanted to compare models with STRUCTURE or INSTRUCT considering the same number of populations. So the highest likelihood would lead to the lowest AIC for the same number of *k* parameters. We noted that for *K* = 7, STRUCTURE always gave a better fit. However, when comparing the likelihood for different *K* values (Figure S3), the INSTRUCT and STRUCTURE results were similar, or sometimes better with INSTRUCT (for flowering time), for *K* < 4. However, as *K* increased, STRUCTURE always performed better for each of the considered traits (Figure S3). In summary, STRUCTURE led to a better likelihood for the highest number of assumed populations (*K* = 7).

Using microsatellite allele data, the population inference method (INSTRUCT/STRUCTURE) and an informative or noninformative kinship matrix did not have a significant effect on the rate of false positives for

TABLE 1

−2 log-likelihood of the different statistical models

	INSTRUCT			STRUCTURE		
Model	FT	SpL	SpD	FT	SpL	SpD
Null	576.86***	678.20***	84.96***	576.86***	678.20***	84.96***
KM	551.76**	638.26***	60.46 (NS)	551.76***	638.26***	60.46***
Q_7	533.76 (NS)	625.54***	57.76*	526.78(ns)	597.56***	40.80***
KM + Q_7	531.26	607.98	52.06	526.78	586.18	32.76

The models tested include the null model, where neither population structure nor family relatedness are considered, the model where only family relatedness is considered (KM), structure only (Q_7), and both KM + Q_7. Q_7 corresponds to ancestry obtained with STRUCTURE or INSTRUCT with seven populations. Comparison of the most complete model (KM + Q_7) to other models is based on a χ^2-test. Significance is noted as follows: NS, nonsignificant; *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

flowering time (Kruskal–Wallis test, $P = 0.97$), spike diameter (Kruskal–Wallis test, $P = 0.95$), or spike length (Kruskal–Wallis test, $P = 0.76$). The effect of the population number (Figure S4) was highly significant regardless of the character considered: flowering time (Kruskal–Wallis test, $P < 0.001$), spike length (Kruskal–Wallis test, $P < 0.001$), or spike diameter (Kruskal–Wallis test, $P < 0.001$). For flowering time, the type I error rate ranged from 18.1% ($K = 1$, no structure) to 5.6% for $K = 3$ (Figure S4). The type I error rate increased as K increased from $K = 4$ to $K = 7$, while for $K = 7$ the type I error rate was 7.2%. For spike diameter (Figure S4), the type I error rate decreased from 16.5% ($K = 1$) to 7.8% ($K = 7$). Finally, the spike length showed the highest rate of false positives (Figure S4), with values of 27.2% at $K = 1$ and 11.9% at $K = 7$. We observed similar results when we used AFLP alleles rather than SSR alleles (Figure S4, statistical analysis not shown). However, although no overall difference in false positive rate was observed between AFLP and SSR allele-based distributions for spike diameter ($P = 0.28$), the AFLP data showed a significantly higher global false positive rate for spike length (Wilcoxon's test, $P < 0.004$) and flowering time (Wilcoxon's test, $P < 10^{-6}$).

We analyzed how these three characters (spike length, spike diameter, and flowering time) were associated with the population structure. We thus considered only $K = 3$ to have enough individual plants in the different groups and set the ancestry threshold at 0.70 to determine whether the plants belong to one of the three groups. We then performed a Kruskal–Wallis test for each field experiment and used a Fisher combining probability to obtain a statistical test pooling the results of the three field experiments. All characters covaried with the population structure. Spike length showed the strongest covariation signal with respect to the population structure ($\chi^2 = 92.3$, $P < 10^{-17}$), then flowering time ($\chi^2 = 74.4$, $P < 6 \times 10^{-14}$), and finally spike diameter ($\chi^2 = 28.5$, $P < 8 \times 10^{-5}$).

The power of the method for detecting a given additive effect on the flowering time character was assessed with different allele frequencies (Figure 2). The given additive effect was a number of days or a genetic effect ratio (Yu *et al.* 2006; Stich *et al.* 2008). The genetic effect ratio was the number of days divided by the standard deviation. Modest effects of <2 days (a genetic effect ratio of 0.22) could not be easily detected regardless of the allele frequency of the SNP. An effect of 6 days was easily detected even for alleles with a frequency of 12.5%. Alleles present at low or very low frequency (1/16 or 1/32) were detected only if they had a strong effect on the phenotype (12–16 days). Some

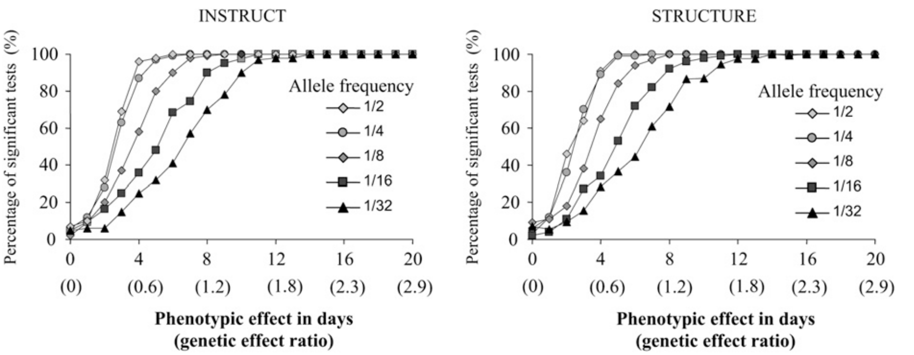


Figure 2.—Power to detect a given flowering phenotypic effect as a function of the allele frequencies. We calculated the probability of finding a significant association at $P < 0.05$ for a simulated allele having a given flowering phenotypic effect. The allele frequency ranged from 50% (1/2) to 3% (1/32). The additive effect was number of days to flowering from 0 to 22 days. Weak phenotypic effects of <2 days were difficult to identify regardless of the allele frequency. For an additive effect of 6 days, the probability of detection of the effect was high when alleles had a frequency of 12.5–50%. It was, however, only 60% for alleles with a frequency 6% and 40% for alleles with a frequency of 3%. Low-frequency alleles were detected only when they had a large phenotypic effect (≥16 days).

TABLE 2

Diversity of pearl millet genes

Name	Size (bp)	Polymorphic site (%)	π (10^{-3})	⊖ (10^{-3})	Tajima's *D*	Fu and Li's *D**	Fu and Li's *F**
Floricaula	819	0.24	0.82	0.51	1.00	0.72	0.94
CRY2	848	0.24	0.60	0.49	0.38	−0.99	−0.67
GI	1417	0.92	1.64	2.07	−0.59	1.51*	0.94
Hd3a	917	0.76	2.00	0.16	0.65	1.21	1.21
Hd6	652	0.92	2.27	1.78	0.61	1.06	1.07
PHYA	1051	0.29	1.24	0.58	2.12*	0.84	1.45
PHYB	1175	0.51	0.49	1.06	−1.28	−0.67	−1.02
PHYC	866	0.69	3.00	1.47	2.38*	1.13*	1.82*
Mean		0.57	1.51	1.02	0.66	0.60	0.72

For each gene, the size of the amplified fragment (SNP and indels), the percentage of the polymorphic site, the value of π, the value of ⊖, Tajima's *D* value, and Fu and Li's *D** and *F** are reported. *$P < 0.05$.

authors have presented this effect as a percentage of the explained variance, which depends both on the standard deviation of the studied trait and on the allele frequency (Yu *et al.* 2006; Stich *et al.* 2008). For comparison, with an SNP frequency of 20% in our simulation, the percentages of explained variance for differences of 2 days, 6 days, and 10 days were 1.4, 11.1, and 25.8%, respectively. The analysis performed using ancestry, as estimated with INSTRUCT, did not show a marked difference with respect to the STRUCTURE findings (Figure 2).

Gene sequence diversity: All primers designed in this study led to sequences with high Blast values with respect to the targeted gene (Table S4). The average percentage of polymorphic sites was 0.64% (Table 2). Polymorphic site indels and SNPs were considered in the present analysis. The average θ-value was 1.1×10^{-3} and the average π-value was 1.6×10^{-3}. The average Tajima's *D* value for all eight loci was 0.66, with a slight bias toward positive values. Two loci exhibited significant Tajima's *D* values: *PHYC* (Tajima's $D = 2.38$, $P < 0.05$) and *PHYA* (Tajima's $D = 2.16$, $P < 0.05$). The *PHYC* gene also showed significant Fu and Li's D^* ($D^* = 1.81$, $P < 0.05$) and F^* ($F^* = 1.13$, $P < 0.05$) test values. The linkage disequilibrium (LD) was calculated on the basis of r^2 (Figure 3). LD varied according to the SNP considered. Strong or weak LDs were observed for the short sequence considered here (<1000 bp between two polymorphisms). Some SNPs separated by only a few hundred base pairs presented no LD. LD was particularly high for *PHYC*, while all polymorphisms except one were strongly linked.

Association with candidate genes: Association analyses were performed for all polymorphic sites of the eight genes (Table S2). We present results obtained with a complete mixed model including the kinship matrix and ancestry inferred for seven populations, using STRUCTURE for SNP 101 of the *PHYC* gene (Table 3). Analyses were performed for each field trial and on BLUE for all field trials. Significant associations (Table 3) were found for flowering time, plant, and spike morphology. Spike length and basal spike diameter were the strongest associated morphological traits. Stem diameter was associated only when the best linear unbiased effects were used. Some morphological associations were significantly detected only in one field trial (NTHE). As most of the *PHYC* SNPs were tightly linked, the same association was observed for the entire *PHYC* amplified fragment (Table S2). Estimation of the SNP effect using BLUE values was 5.2 days for flowering time, 8.3 cm for spike length, 0.070 cm for basal spike diameter, and 0.10 cm for stem diameter (Figure 4).

Associations were also noted for *PHYA* polymorphism and spike length in all field trials for SNPs of the

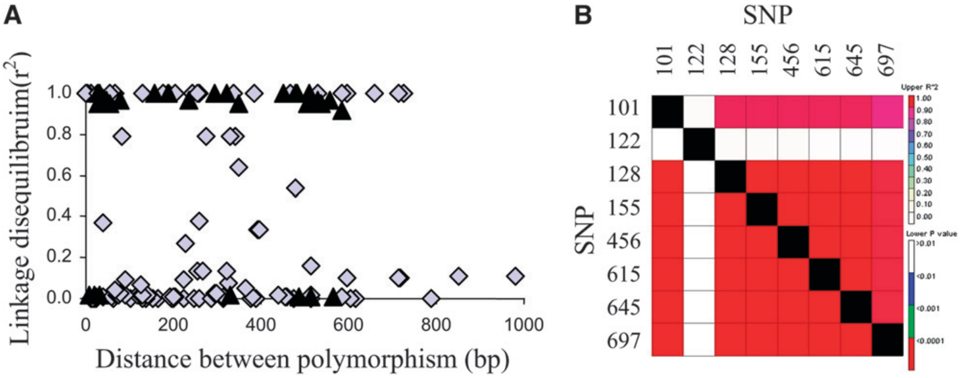


Figure 3.—Linkage disequilibrium in pearl millet. The linkage disequilibrium (LD) was estimated using r^2 (A) between each polymorphism (SNP or indel) for each gene except *PHYC* (gray diamonds) and *PHYC* (black triangles). For *PHYC*, LD values are also presented as a square matrix (B) with r^2 values (top matrix) and LD significance *P*-values (bottom matrix).

TABLE 3

Association of morphological character and *PHYC* polymorphism

	Field trials			
	2005	2006a	2006b	BLUE
	No. of inbred lines			
Traits	79	80	76	88
Flowering time				
FT	$P < 0.002$, $R^2 = 6.6\%$	$P < 0.01$, $R^2 = 6.5\%$	$P < 4 \times 10^{-5}$, $R^2 = 15.5\%$	$P < 3 \times 10^{-4}$, $R^2 = 8.9\%$
Plant morphology				
PH	$P = 0.59$	$P = 0.54$	$P = 0.33$	$P = 0.62$
SD	$P = 0.07$	$P = 0.08$	$P = 0.08$	$P < 0.004$, $R^2 = 5.7\%$
NTHE	$P = 0.22$	$P = 0.36$	$P < 0.04$, $R^2 = 5.2\%$	$P = 0.09$
Spike morphology				
SpD	$P = 0.81$	$P = 0.21$	$P = 0.74$	$P = 0.84$
BSpD	$P < 0.01$, $R^2 = 4.2\%$	$P < 0.03$, $R^2 = 2.9\%$	$P < 0.007$, $R^2 = 5.2\%$	$P < 0.004$, $R^2 = 3.8\%$
SpL	$P < 3 \times 10^{-4}$, $R^2 = 7.0\%$	$P < 0.003$, $R^2 = 5.4\%$	$P < 3 \times 10^{-4}$, $R^2 = 7.9\%$	$P < 3 \times 10^{-4}$, $R^2 = 6.6\%$

For each field trial the number of inbred lines having sequence data, morphological data, and phenological data is given. The mixed model used included a kinship matrix and STRUCTURE-inferred ancestry for seven populations. The *P*-value and percentage of variance explained (R^2) are presented for the SNP at position 101 of the amplified *PHYC* fragment and flowering time (FT), plant morphology (PH, SD, NTHE), and spike morphology (SpD, BSpD, SpL). The probability is presented for each field trial (2005, 2006a, and 2006b) and on best linear unbiased estimates for all field trials. The strongest significant association with *PHYC* was observed for flowering time, basal spike diameter, and spike length for the three field trials.

amplified fragment (Table S2). The SNP 146 of *PHYA*, for example, explained >4% of SpL variation in all field trials (2005, $P < 0.0005$, $R^2 = 7.8\%$; 2006a, $P < 0.02$, $R^2 = 4.0\%$; 2006b, $P < 0.005$, $R^2 = 5.9\%$; BLUE, $P < 0.0007$, $R^2 = 6.6\%$). The other two SNPs of this gene had similar association probability values.

To validate the association of SNP in the *PHYC* gene, we analyzed a new set of 598 pearl millet individuals from Niger. The structure analysis of this sample did not reveal a marked population structure (Figure S5). A kinship matrix was calculated and was not positive definite. We bent this matrix using an *ad hoc* method, using $\varepsilon = -10^{-2}$. The new positive-definite matrix was almost identical to the initial matrix (Spearman's correlation coefficient $R = 0.9999$, $P < 0.001$), showing the adjustment only very slightly modified the original matrix. The different individuals showed only weak relatedness (Figure S6). However, the model with the kinship matrix was significantly better than a null model for most traits (Table S5). We used the model with the kinship matrix for the association between the *PHYC* SNP and traits.

Genotypes for the presence of the C or G alleles of *PHYC* were obtained for 560 of these pearl millet individuals. We found 27 individual homozygotes G/G, 120 individual C/G, and 413 C/C. We assessed associations in this data set with a mixed model considering the kinship matrix and the three genotypes (C/C, C/G, and G/G), using BLUE of trait value for all field experiments. The analysis highlighted a significant effect of the genotype on flowering time [Wald test of fixed effects, WT = 12.1, degree of freedom (dof) = 2, $P < 0.003$], spike length (Wald test of fixed effects, WT = 11.9, dof = 2, $P < 0.003$), and stem diameter (Wald test of fixed effects, WT = 13.9, dof = 2, $P < 0.001$). The number of tillers, plant height, spike diameter, and thousand seed weight were not significantly associated with the SNP polymorphism (Figure 5). The Bonferroni-

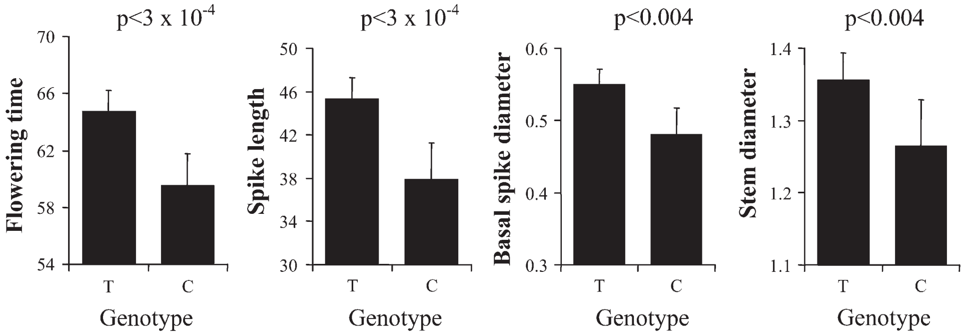


FIGURE 4.—Trait effect of *PHYC* SNP 101 in pearl millet inbred lines. The mean value and standard errors for each genotype of the SNP 101 in the *PHYC* gene (C or T) are presented for flowering time (in days), spike length (in centimeters), basal spike diameter (in centimeters), and stem diameter (in centimeters). The *P*-value was obtained using the mixed-model method implemented in TASSEL.

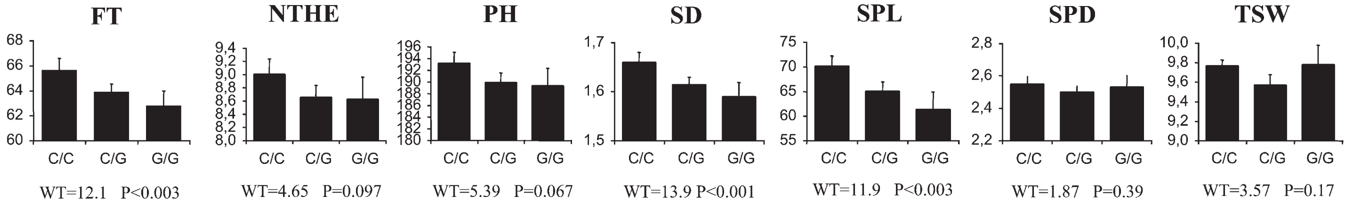


FIGURE 5.—Variation at *PHYC* and variation in phenology and morphology in pearl millet varieties. The mean value and standard errors for each genotype of the SNP 697 in the *PHYC* gene (C/C, C/G, and G/G) are presented for each morphological and phenological trait: flowering time in days (FT), number of tillers at head emergence (NTHE), plant height in centimeters (PH), stem diameter in centimeters (SD), spike length in centimeters (SpL), spike diameter in centimeters (SpD), and thousand seed weight in grams (TWS) are presented. The analysis was performed on the best linear unbiased effect of traits assessed in two different field trails (2004 and 2005). The analysis was performed using a mixed model incorporating a kinship matrix only (see text for details). We reported, for each trait and each field trial, the value of the Wald test statistics of fixed effects and the associated *P*-values with 2 dof. Three traits (FT, SPL, and SD) showed a significant genotypic effect even though we considered a Bonferroni-corrected significant threshold of 0.007.

corrected *P*-value for seven different tests was 0.007, so the association of flowering time, spike length, and stem diameter was significant with this corrected threshold. The association was also performed on individual field trials and led to a similar conclusion (Table S6). On the basis of BLUE, flowering time was on average 62.8 days for the G/G genotype and 65.7 days for C/C. A difference of 2.8 days was thus noted. The difference in stem diameter was 0.07 cm. The average stem diameter was 1.59 cm for G/G and 1.66 cm for C/C. The spike length difference was 8.7 cm. The average spike length was 61.4 cm for G/G and 70.1 cm for C/C.

DISCUSSION

Inference of population structure and association study: The STRUCTURE Bayesian method is frequently used to infer population structures in an association framework. However, this method is not yet tailored for studies with inbred materials or selfing species. New methods like INSTRUCT have been developed very recently for this specific purpose (GAO *et al.* 2007). Our comparison obviously showed some differences between INSTRUCT and STRUCTURE results, as also previously noted (GAO *et al.* 2007). The similarity between the two methods was high (generally >90%). However, for the mixed model, STRUCTURE tends to have higher likelihood for a number of assumed populations >5, whereas INSTRUCT tends to have higher likelihood for a lower number of assumed populations. Comparative analyses of the two population structure inference methods on a type I control in association studies did not show a significant difference. Although the INSTRUCT model seems to be the most appropriate method for inbred material, our results obtained on our current data set using only 27 microsatellite loci showed that STRUCTURE led to better control of population structure. For population structure inference, we assumed a *K* population number ranging from 1 to 10. For STRUCTURE, the optimal *K* was $K = 6$, while for INSTRUCT it was $K = 7$. A question that might be addressed is, What population number gave the best control of type I error? As expected, taking the population structure into account (assuming $K > 1$) led to a lower number of false positives. However, using the optimal number of populations ($K = 6$ or $K = 7$) did not necessarily lead to better control of the false positive rate than $K = 3$, for example. The number of false positives for a given *K* value is certainly dependent on the relationship between the genetic structure and the phenotype differentiation between populations (CASA *et al.* 2008).

Traits covarying with the population structure are the most problematic for effective control of the false positive rate (REMINGTON *et al.* 2001). Our results showed that spike length was most strongly linked to the population structure. For this trait, the false positive rate was never <10% regardless of the number of *K* populations considered. It could thus be hard to detect associations with this particular trait even if a particular SNP has an effect on the phenotype. We calculated a corrected *P* threshold based on microsatellite loci for each trait/field trial. This new significance threshold should partially overcome the gap between the expected 5% nominal ratio of false positive rate and the observed ≥10% for spike length.

In the present study, we used model-based Bayesian approaches to infer the population structure. Other methods like the principal components analysis (PCA)-based approach are not based on a particular model and can also be applied to detect population structures (PATTERSON *et al.* 2006). STICH *et al.* (2008) found that the PCA-based approach did not have better control of the false positive rate on a wheat data set. We observed—like previous studies—that taking the kinship matrix into account give a fitter model (YU *et al.* 2006; STICH *et al.* 2008). However, considering the control of spurious associations, we actually did not detect a significant difference when using a kinship matrix or not. However, we noted that the type I error rate was slightly lower when taking the kinship matrix into account. In this study, we used SPAGeDI to infer a kinship matrix between inbred lines. A very recent study (Stich *et al.* 2008) showed that a restricted maximum-likelihood-based method could be used to infer this kinship matrix. This approach leads to a slight improvement in false positive control (Stich *et al.* 2008) over the initial approach of Yu *et al.* (2006).

We considered the best model using kinship and a STRUCTURE population structure. Seven clusters were used to perform all subsequent analyses. We analyzed the power of the methodology for identifying a given additive effect. From this analysis, it is clear that frequent variants (>1/8) are easily spotted for even a modest effect of 7 days (genetic effect ratio of ~1.0). But a slight effect (<2 days, *i.e.*, a genetic effect of ~0.3) would be difficult to identify. The ability to detect slight effects may have been linked to the number of inbred lines considered in this study, whereas a higher number of inbred lines might be more effective for identifying such a slight effect. However, a previous power analysis study also detected a low power for a similar genetic effect, even though it considered threefold more inbred lines (Yu *et al.* 2006; Stich *et al.* 2008). Such a low effect may thus be difficult to detect, even though we used larger data set. Identification of variants using this approach in the present framework would likely be useful for flowering genes having an effect of at least 4 days; lower flowering differences were observed for allele frequencies of at least 1/8. In terms of explained variance, an effect of ≥10% is often easily detected. For the study of flowering time differences, some authors suggest that crop mutations might be more likely associated with large phenotypic effects (Roux *et al.* 2006). Although such alleles are relatively frequent, they could be easily spotted using this association framework.

Association of *PHYC* polymorphism with flowering time and morphological character: We identified some polymorphism in the *PHYC* gene correlated with the flowering phenotype and other morphological traits in a pearl millet inbred data set. Using pearl millet varieties from Niger, we validated the association between *PHYC* polymorphism and phenological variation that we first detected in our inbred line data set. This analysis was based on an estimation of the average morphological/phenological character of each variety. The association was based on a single individual per variety associated with the average morphological/phenological value of the variety. Detection of an effect based on this design could not be very powerful since we attributed the average value of a variety to a single individual and within-variety polymorphism is expected to be very high (Allinne *et al.* 2008). We nevertheless detected a significant effect of *PHYC* polymorphism on a similar set of characters: flowering time, spike length, and stem diameter. However, the design did not allow us to draw any conclusions on the recessivity or dominance of the C and G alleles based on the mixed model results (Figure 5).

The association we detected with *PHYC* polymorphism was thus validated in two independent samples. The extent to which the phenotype is controlled by the *PHYC* gene or a neighboring gene has yet to be determined. Several studies suggest that polymorphism at *PHYC* is related to flowering differences in rice (Takano *et al.* 2005) and Arabidopsis (Balasubramanian *et al.* 2006). The direct causative role of *PHYC* (although not yet fully demonstrated) is a very likely scenario. A sequence analysis is underway to identify potential functional polymorphism within the entire *PHYC* gene. However, the phenotype might also be associated with differences in expression pattern. *PHYA* and *GI* genes also showed a significant association with spike length. However, the character associated with these genes is one for which false positive control was the least effective. These results should be considered with caution until they are further validated.

We found evidence based on Tajima's *D* statistics of two *PHYA* and *PHYC* genes, suggesting that polymorphism was balanced at these loci. These statistics were accurate if there was no population structure within the study sample. We found a significant population structure signal in the inbred lines. The average Tajima's *D* value for all loci was 0.66 (0.13 when *PHYA* and *PHYB* were excluded), suggesting a slight positive bias. This effect certainly inflated the Tajima's *D* values of the two genes. However, when considered with the *PHYC* association, these values might indicate a real selection signal. Wild pearl millet populations are spread in a dry area at the southern limit of the Sahara desert (Oumar *et al.* 2008). In West Africa, pearl millet is cultivated throughout three agro-ecological zones: the Sahel zone (200–500 mm annual rainfall), the Sudano-Sahelian zone (500–900 mm), and the Sudanian zone (900–1100 mm). The adaptation of pearl millet to a wetter climate is associated with later flowering (Haussmann *et al.* 2006). A likely hypothesis is this adaptation to a wetter climate was associated with selection at the two genes: different alleles of these genes are maintained in different environments, leading to genetic diversity exhibiting balanced polymorphism. A study should be carried out on a regional scale to validate this hypothesis.

The LD study highlighted a fast decrease in pearl millet inbred material, with low r^2 values, as we observed here for SNPs separated by a few base pairs. The LD in Arabidopsis has a genomewide decrease to $r^2 < 0.20$ at a distance of 10 kb (Kim *et al.* 2007). In inbred maize lines, a decrease has been observed at a shorter range of a few hundred base pairs (Remington *et al.* 2001). The results obtained here were closer to maize results. However, as expected, we also found strong locus-specific variability, which was certainly linked to each particular gene history, gene location in the genome, selection, local diversity, and recombination rate. The LD for *PHYC* was particularly high, as expected for a selected genomic region. As we investigated a low number of genes, it was hard to pinpoint the factor controlling this high LD in *PHYC*. However, a better assessment of LD in pearl millet would require an analysis of a larger number of loci and a larger chunk of DNA.

Altogether, the positive association results, significant selection test results, and high LD at *PHYC* suggested that this locus is under diversifying selection in pearl millet.

Five phytochrome *PHYA-E* genes have been found in *Arabidopsis thaliana*, and only three *PHYA-C* genes are described in monocotyledon species like Oryza or Sorghum (Mathews 2006b). The *PHYC* gene seems to have a relatively minor functional role in Arabidopsis development (Franklin *et al.* 2003; Monte *et al.* 2003; Mathews 2006a). However, natural variation at *PHYC* is associated with a latitudinal gradient (Balasubramanian *et al.* 2006), and there is empirical evidence that *PHYC* mediates photoperiod sensitivity in natural populations of *A. thaliana* (Samis *et al.* 2008). *PHYC* in Arabidopsis thus has an important role for the adaptation of natural populations to different climates. A recent study has also revealed natural variation at the *PHYB* gene in Arabidopsis accessions causes differential responses to light (Filiault *et al.* 2008). In *Populus tremula*, *PHYB2* natural variations are also associated with variations in the timing of bud set (Ingvarsson *et al.* 2008). In rice, PHYC protein is required to delay flower initiation during long days (Takano *et al.* 2005). The *phyB* mutants have an earlier flowering phenotype similar to *phyC* mutants under long day conditions, but *phyB* and not *phyC* hastens the flowering time during short days (Takano *et al.* 2005). In sorghum, the *phyB* natural mutant is associated with a photoperiod-insensitive flowering time phenotype (Foster *et al.* 1994; Childs *et al.* 1997). Moreover, the *PHYC* in sorghum shows unusual nonsynonymous polymorphisms (White *et al.* 2004), which might be associated with functional effects. Overall, these results and the present study findings suggest that phytochromes might be preferential targets of selection for flowering time variation in plants (Balasubramanian *et al.* 2006). The upstream position of the photoreceptor gene in the flowering development network might partially explain why, in different species, variations may occur in the same set of genes associated with flowering time variation. Variations in the most upstream gene of a pathway might be associated with a lower pleiotropic effect (Roux *et al.* 2006).

To date, 3000 genomic DNA sequences are available for pearl millet in GenBank. This species is not a genomic research priority and is best described as an orphan crop. Pearl millet is adapted to marginal agricultural areas with low rainfall and plays a crucial role in feeding the poorest of the poorest, particularly in the Sahel. In Niger, pearl millet is grown on 65% of the total cultivated area. Conducting association studies in pearl millet provides an opportunity to rapidly validate important agronomic genes identified in other plant models and cereals for their role in the pearl millet phenotype. We hope that the identification of such key genes will favor the development of improved varieties using marker-assisted selection.

We thank T. Hash (International Crops Research Institute for the Semi-Arid Tropics), J. Chantereau (Centre de Cooperation Internationale en Recherche Agronomique pour le Développement), T. Robert, and A. Sarr (University of Orsay, Paris) for supplying us with pearl millet inbred line seeds. We thank Y. Orieux, C. Allinne, P. Sire, M. Couderc, D. Moussa, and M. Tidjani for help during the field studies and laboratory experiments. We thank Pär K. Ingvarsson for advice on R code development. This project was funded by an Institut de Recherche pour le Developpement (IRD) core grant and by a grant from the Agence Nationale de la Recherche to Y.V. (ANR-07-JCJC-0116-01). A.-A.S. is funded by an IRD Ph.D. fellowship.

LITERATURE CITED

Allinne, C., C. Mariac, Y. Vigouroux, G. Bezançon, E. Couturon *et al.*, 2008 Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. Genetica **133:** 167–178.

Atkinson, B., and T. Therneau, 2008 Kinship: mixed kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package, Versions 1.1.0–21. http://cran.r-project.org.

Balasubramanian, S., S. Sureshkumar, M. Agrawal, T. P. Michael, C. Wessinger *et al.*, 2006 The phytochrome C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. Nat. Genet. **38:** 711–715.

Buckler, E., P. Bradbury, D. Kroon, Y. Ramdoss, T. Casstevens *et al.*, 2007 Trait Analysis by Association, Evolution and Linkage (TASSEL). Version 2.0.1. http://www.maizegenetics.net/tassel.

Camus-Kulandaivelu, L., J.-B. Veyrieras, D. Madur, V. Combes, M. Fourmann *et al.*, 2006 Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. Genetics **172:** 2459–2463.

Casa, A. M., G. Pressoir, P. J. Brown, S. E. Mitchell, W. L. Rooney *et al.*, 2008 Community resources and strategies for association mapping in Sorghum. Crop Sci. **48:** 30–40.

Childs, K. L., F. R. Miller, M. M. Cordonnier-Pratt, L. H. Pratt, P. W. Morgan *et al.*, 1997 The sorghum photoperiod sensitivity gene, *Ma*$_3$, encodes a phytochrome B. Plant Physiol. **113:** 611–619.

D'Andrea, A. C., and J. Casey, 2002 Pearl millet and Kintampo subsistence. Afr. Archaeol. Rev. **19:** 147–173.

D'Andrea, A. C., M. Klee and J. Casey, 2001 Archaeological evidence for pearl millet (*Pennisetum glaucum*) in sub-Saharan West Africa. Antiquity **75:** 341–348.

Doebley, J., A. Stec and L. Hubbard, 1997 The evolution of apical dominance in maize. Nature **386:** 485–488.

Evanno, G., S. Regnaut and J. Goudet, 2005 Detecting the number of clusters of individuals using the software Structure: a simulation study. Mol. Ecol. **14:** 2611–2620.

Falush, D., M. Stephens and J. K. Pritchard, 2003 Inference of population structure using multilocus genoype data: linked loci and correlated allele frequencies. Genetics **164:** 1567–1587.

Filiault, D. L., C. A. Wessinger, J. R. Dinneny, J. Lutes, J. O. Borevitz *et al.*, 2008 Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. Proc. Natl. Acad. Sci. USA **105:** 3157–3162.

Foster, K. R., F. R. Miller, K. L. Childs and P. W. Morgan, 1994 Genetic regulation of development in *Sorghum bicolor*. Plant Physiol. **105:** 941–948.

Franklin, K. A., S. J. Davis, W. M. Stoddart, R. D. Vierstra and G. C. Whitelam, 2003 Mutant analyses define multiple roles for phytochrome C in Arabidopsis photomorphogenesis. Plant Cell **15:** 1981–1989.

Fritz, G. L., 1995 New dates and data on early agriculture: the legacy of complex hunter-gatherers. Ann. Mo. Bot. Gard. **82:** 3–15.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gao, H., S. Williamson and C. D. Bustamante, 2007 An MCMC approach for the joint inference of population structure and inbreeding rate from multi-locus genotype data. Genetics **176:** 1635–1651.

Hardy, O. J., and X. Vekemans, 2002 SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes **2:** 618–620.

Haussmann, B. I. G., A. Boubacar, S. S. Boureima and Y. Vigouroux, 2006 Multiplication and preliminary characterization of West and Central African pearl millet landraces. Int. Sorghum Millet Newsl. **47:** 110–112.

Hayes, J. F., and W. G. Hill, 1981 Modification of estimates of parameters in the construction of genetic selection indices ("bending"). Biometrics **37:** 483–493.

Ingvarsson, P. K., M. V. Garcia, V. Luquez, D. Hall and S. Jansson, 2008 Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, *Salicaceae*). Genetics **178:** 2217–2226.

Jaenicke-Després, V., E. S. Buckler, B. D. Smith, M. T. Gilbert, A. Cooper *et al.*, 2003 Early allelic selection in maize as revealed by ancient DNA. Science **302:** 1206–1208.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics **178:** 1709–1723.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clarck *et al.*, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **39:** 1151–1155.

Loiselle, B. A., V. L. Sork, J. Nason and C. Graham, 1995 Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (*Rubiaceae*). Am. J. Bot. **82:** 1420–1425.

Mariac, C., V. Luong, I. Kapran, A. Mamadou, F. Sagnard *et al.*, 2006a Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theor. Appl. Genet. **114:** 49–58.

Mariac, C., T. Robert, C. Allinne, M. S. Remigereau, A. Luxereau *et al.*, 2006b Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. Theor. Appl. Genet. **113:** 1003–1014.

Mathews, S., 2006a Seeing the light. Nat. Genet. **38:** 606–608.

Mathews, S., 2006b Phytochrome-mediated development in land plants: red light sensing evolves to meet the challenges of changing light environments. Mol. Ecol. **15:** 3483–3503.

Mathews, S., R. C. Tsai and E. A. Kellogg, 2000 Phylogenetic structure in the grass family (*Poaceae*): evidence from the nuclear gene phytochrome B. Am. J. Bot. **87:** 96–107.

Monte, E., J. M. Alonso, J. R. Ecker, Y. Zhang, X. Li *et al.*, 2003 Isolation and characterization of *phyC* mutants in Arabidopsis reveals complex crosstalk between phytochrome signaling pathways. Plant Cell **15:** 1962–1980.

Oumar, I., C. Mariac, J.-L. Pham and Y. Vigouroux, 2008 Phylogeny and origin of Pearl Millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. Theor. Appl. Genet. **117:** 489–497.

Patterson, N., A. L. Price and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. **2:** e190.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly, 2000a Association mapping in structured populations. Am. J. Hum. Genet. **67:** 170–181.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000b Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Roux, F., P. Touzet, J. Cuguen and V. Le Corre, 2006 How to be early flowering: an evolutionary perspective. Trends Plant Sci. **11:** 375–381.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Samis, K. E., K. D. Heath and J. R. Stinchcombe, 2008 Discordant longitudinal clines in flowering time and phytochrome C in Arabidopsis thaliana. Evolution **62:** 2971–2983.

Sokal, R. R., and F. J. Rohlf, 1991 *Biometry*, Ed. 3. W. H. Freeman, New York.

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E. S. Buckler *et al.*, 2008 Comparison of mixed-model approaches for association mapping. Genetics **178:** 1745–1754.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Takano, M., N. Inagaki, X. Xie, N. Yuzurihara, F. Hihara *et al.*, 2005 Distinct and cooperative functions of phytochromes A, B, and C in the control of deetiolation and flowering in rice. Plant Cell **17:** 3311–3325.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen *et al.*, 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28:** 286–289.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. Nature **398:** 236–239.

White, G. M., M. T. Hamblin and S. Kresovich, 2004 Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution. Mol. Biol. Evol. **21:** 716–723.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** 203–208.

Communicating editor: A. H. Paterson

GENETICS

Supporting Information

http://www.genetics.org/cgi/content/full/genetics.109.102756/DC1

Association Studies Identify Natural Variation at *PHYC* Linked to Flowering Time and Morphological Variation in Pearl Millet

Abdoul-Aziz Saïdou, Cédric Mariac, Vivianne Luong, Jean-Louis Pham, Gilles Bezançon and Yves Vigouroux

DOI: 10.1534/genetics.109.102756

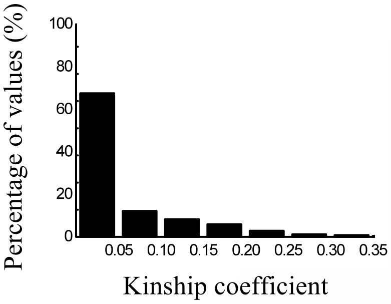


FIGURE S1.—Inbred line kinship coefficient distribution. The kinship coefficient between each inbred lines was calculated using 306 AFLP markers using SPAGeDi (HARDY AND VEKEMANS 2002). Kinship coefficients lower than 0 were set to zero.

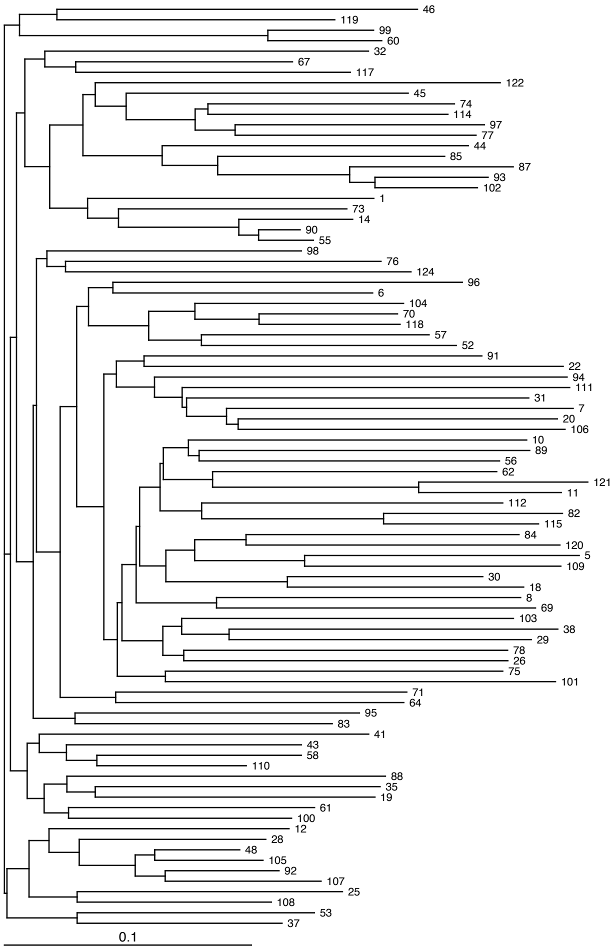


FIGURE S2. —Neighbor joining tree of inbred lines AFLP and SSR datasets were used to build a neighbor joining tree using a shared-alleles distance. The neighbor joining tree illustrates relatedness of some inbred lines: for example 48, 105 and 92.

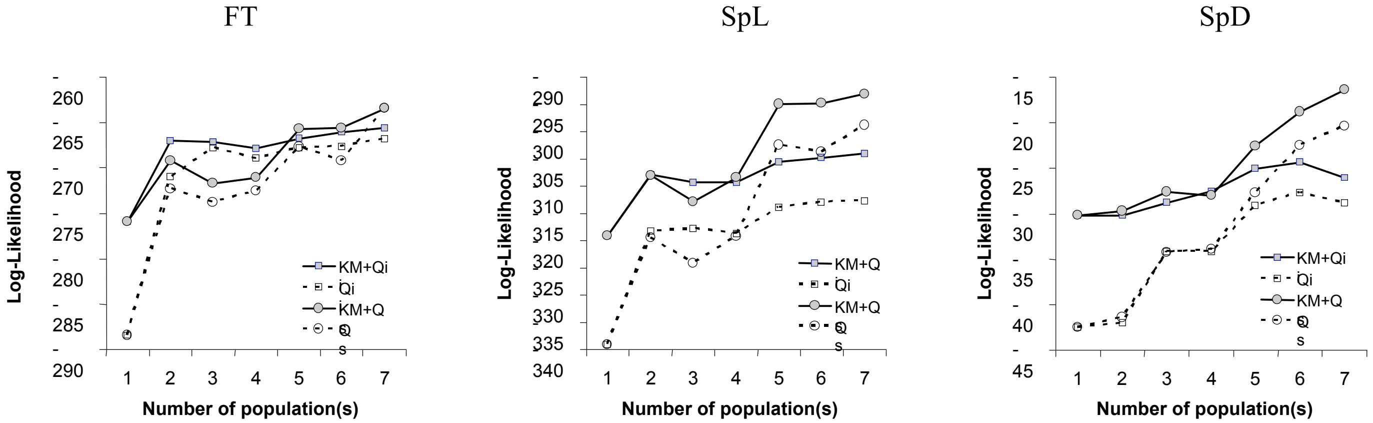


FIGURE S3.—Model log-likelihood considering population structure and kinship matrix. Log-likelihood for models considering population structure (K=1 no structure, to K=7) inferred using INSTRUCT (Qi) or STRUCTURE (Qs), and considering or not a kinship matrix (KM) are presented. Models considering a kinship matrix are represented with grey squares (for INSTRUCT inferred ancestry) and grey circles (for STRUCTURE inferred ancestry). Models without a kinship matrix are represented with white squares (for INSTRUCT inferred ancestry) and white circles (for STRUCTURE inferred ancestry). The models were assessed for three traits: flowering time (FT), spike length (SpL) and spike diameter (SpD).

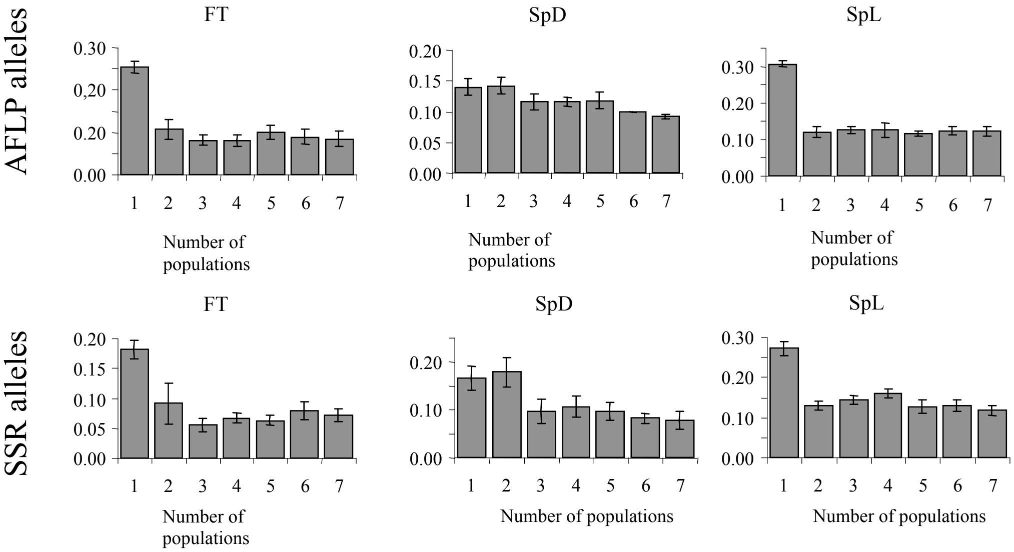


FIGURE S4.—Type I error control in function of the number of assumed populations (K). The percentage of type I error (false positive rate) was estimated using association studies performed on SSR and AFLP alleles using a mixed linear model (YU *et al.*, 2006). The average percentage of false positives was estimated for different assumed populations (K from 1 to 7) for the days from sowing to female flowering (FT), the diameter of the spike (SpD) and the length of the spike (SpL). The average false positive rate using population structure inferred using STRUCTURE and INSTRUCT was calculated. Standard errors were calculated using the false positive rate estimated on three different field trials. Using AFLP and SSR markers, taking into account population structure (K=2 to K=7) reduces false positive rate, compared to models without population structure (K=1). For example, with SSR markers, taking into account population structure significantly reduces the false positive rate for FT, from 18.1% at K=1 to 5.6% at K=3. The effect of considering a high number of populations (K higher than 3) do not lead to a better control of the type I error rate. For SpD, the type I error rate decrease from 16.5% at K=1 to 7.8% at K=7. For SpL, the average type I error rate is 27.2% at K=1 and decreases to 11.8% at K=7. Noted that AFLP markers showed a significantly higher level of false positives than SSR alleles for spike length (Wilcoxon test, P<0.05) and flowering time (Wilcoxon test, P<0.05).

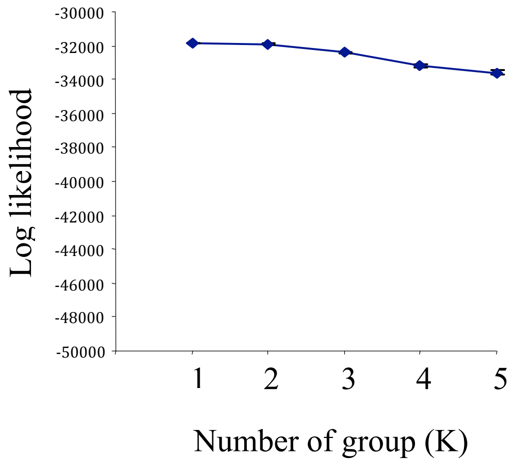


FIGURE S5.—Structure analysis of varieties from Niger. The graph represents the highest log-likelihood for pearl millet varieties from Niger considering from 1 to 5 populations. The analysis reveals no major signal of population structure, the log-likelihood slightly decreases from K=1 to K=2, then K=5. Ten different runs for each K value were performed. The highest log-likelihood for K=1 was −31839.5, and for K=2 was −31885.9.

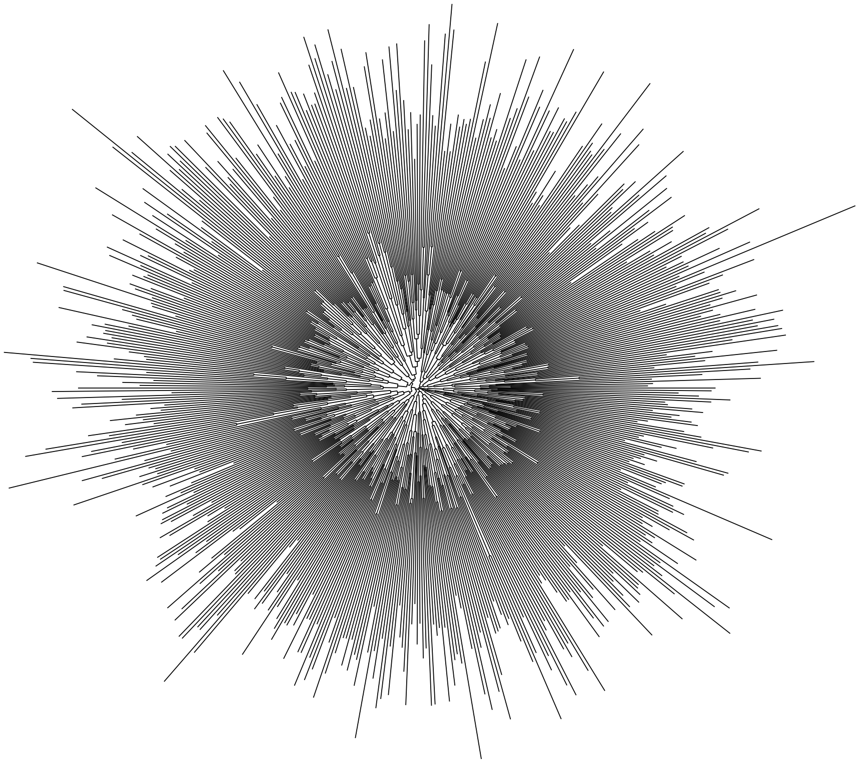


FIGURE S6.—Neighbor joining tree of the 598 pearl millet individuals from Niger. The neighbour joining tree was built using the shared allele distance using 2 microsatellite loci.

TABLE S1

Table S1 is available as an Excel file at http://www.genetics.org/cgi/content/full/genetics.109.102756/DC1.

TABLE S2

Table S2 is available as an Excel file at http://www.genetics.org/cgi/content/full/genetics.109.102756/DC1.

Table S3

Primers of pearl millet genes

Name	Forward	Reverse	Origin
FLORICAULA	GAGCTGGAGGACCTGGTG	CTCGGAGCTCGGGTTCAC	This study
CRY2	GAGCTGCACCTTGTTTCTCC	TCATGGTAGGCACCATCTGA	This study
GI	GCTGCCTATGGTTTGCTACC	GCCAGAGCAATGAGACAACA	This study
Hd3a	GGCAGGGACAGGGASC	TTGTAGAGCTCGGCGAAGT	This study
Hd6	GATTACTGCCATTCACAAGG	GAAGCTCAGGWCCCTTGAAGTA	This study
PHYA	ATTGCCTTCTGGCTTTCAGA	TACAAAGCACACCCCAACAA	This study
PHYB	GCRTCCATYTCKGCATTYTCCCA	GAGCCIGCYMGHACSGARGAYCC	MATHEWS *et al.* 2000
*PHY*C	CAGATTGCTCATYTRGAGTTCA	CGTGCCRCTCATCGTYTTC	This study

TABLE S4

Gene sequenced in pearl millet

Name	E value	Accession	Species	Gene
FLORICAULA	2e^{-103}	AY789048.1	*Zea mays*	*FLORICAULA/LEAFY-LIKE* 2
CRY2	0	EF601540.1	*Triticum aestivum*	*CRYPTOCHROME* 1a
GI	0	AY679115.1	*Triticum aestivum*	*GIGANTEA* 3
Hd3a	9e^{-44}	DQ157462.1	*Oryza sativa*	*Hd3a*
Hd6	4e^{-39}	EF114229.2	*Zea mays*	*Hd6*
PHYA	0	AY466082.1	*Sorghum bicolor*	*PHYTOCHROME* A
PHYB	0	AB109892.1	*Oryza sativa*	*PHYTOCHROME* B
*PHY*C	1e^{-34}	AY234829.1	*Zea mays*	*PHYTOCHROME* C1

For each gene sequenced in pearl millet the highest value with Blastn (MegaBlast) analysis is presented: E value, accession name, accession species name and gene name.

TABLE S5

Loglikelihood of mixed model for varieties using or not a kinship matrix

Traits	LogLikelihood without KM	LogLikelihood with KM
FT	-1813.0***	-1761.5
NTHE	-1100.9***	-1069.0
PH	-2265.2***	-2256.7
SD	258.6***	277.0
SpL	-2356.0***	-2343.1
SpD	-225.3***	-215.7
TSW	-765.8 (ns)	-765.8

For each trait : flowering time (FT), number of tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), spike length (SpL), spike diameter (SpD) and thousand seed weight (TWS), the log-likelihood is reported for a mixed model including a kinship matrix KM or not. The two models were compared using a likelihood ratio tests and using a χ^2 distribution to assess significance: (ns) non significant, *** P<0.001. The model including the kinship matrix gave a better fit for most of the traits except thousand seed weight. Note that log-likelihood is positive for stem diameter, a not surprising feature for quantitative continuous variable.

TABLE S6

Association of varieties with *PHYC* SNP for annual field data

Traits	**2004 field trial**		**2005 field trial**	
FT	WT=8.8	P=0.012 *	WT=14.5	P<0.001 ***
NTHE	WT=4.2	p=0.12	WT=3.1	p=0.22
PH	WT=3.4	p=0.18	WT=6.8	p=0.034 *
SD	WT=10.4	P<0.006 **	WT=13.7	P<0.002 **
SpL	WT=13.6	P<0.002 **	WT=9.3	P=0.0095 **
SpD	WT=2.2	p=0.34	WT=1.8	p=0.40
TSW	WT=2.0	p=0.36	WT=2.1	p=0.35

For each trait : flowering time (FT), number of tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), spike length (SpL), spike diameter (SpD) and thousand seed weight (TSW), the value of the Wald test of fixed effect was given with its associated p value for 2 degrees of freedom. * P<0.05, ** P<0.01, *** P<0.001. Note that the p values for these tests are not corrected by a Bonferroni adjustment.

Conclusions et perspectives

Nous avons développé une approche de génétique d'association exploitant diverses populations de mil. L'étude montre que le modèle linéaire mixte constitue une option efficace pour limiter le taux de faux positifs. Ce modèle a aussi révélé une puissance convenable, surtout pour des allèles ayant un effet phénotypique relativement large et une fréquence assez élevée. Cette méthodologie pourrait permettre l'analyse d'un plus grand nombre de traits et de gènes d'intérêt chez le mil.

Nous avons testé huit gènes candidats et avons identifié une association entre la région génomique de *PHYC* et la date de floraison (Saïdou *et al.* 2009).

Parmi les principales limites de cette première étude, se trouve le fait que, malgré une correction relativement efficace du taux de faux positifs avec le modèle mixte, cette correction est rarement parfaite dans les études d'association (Yu *et al.* 2006, Stich *et al.* 2008). Cela peut être dû, par exemple, à la limite des modèles utilisés pour inférer la structure, ou encore à la sensibilité de ces modèles à la taille de l'échantillon ou au nombre de marqueurs neutres utilisés pour inférer la structure. De façon intéressante, nous avons aussi montré que l'efficacité de cette correction du taux de faux positifs dépend de la relation du trait phénotypique avec la structure des populations. Les traits qui covarient le plus avec la structure, comme la taille des épis, ont ainsi montré plus de biais vers l'excès de faux positifs. Il reste cependant que, malgré cette relative imperfection, la correction du taux de faux positifs avec le modèle mixte, reste assez efficace. Pour la date de floraison par exemple, nous avons pu passer de 18.1 % de faux positifs avec le modèle naïf, à seulement 6-8% avec le modèle mixte. Le modèle mixte a donc considérablement réduit le biais, ne laissant qu'un plus faible écart au taux d'erreur souhaitable (5%). Nous avons ensuite fait une correction empirique du seuil de significativité des p-values (α), pour combler l'écart entre ce taux estimé de 6-8% de faux positifs pour le modèle mixte, et le taux standard de 5% communément admis. On peut donc estimer que les associations significatives relevées dans cette étude ont une très haute probabilité d'être de vraies associations. Une discussion autour du rôle connu de *PHYC* chez d'autres espèces comme le riz (Takano *et al.* 2005) et Arabidopsis (Balasubramanian *et al.* 2006).

Une approche pour renforcer la certitude sur l'association de la région chromosomique de *PHYC* avec le phénotype est d'effectuer une validation de cette association en utilisant une cartographie de liaison classique basée sur des croisements (Bergelson and Roux 2010). La cartographie de liaison n'est pas sujette à l'excès de faux positifs relevé dans le cas de la cartographie d'association, d'où l'intérêt de cette démarche comparative. Cela est l'un des objets de notre seconde étude, qui sera présentée dans ce qui suit. Cette étude présentera donc une analyse QTL permettant de valider la colocalisation entre *PHYC* et des QTLs détectés sur des familles indépendantes du panel de lignées (F2).

Il est également connu qu'un déséquilibre de liaison (DL) étendu dans une région génomique peut entrainer des associations virtuelles (Camus-Kulandeivelu *et al.* 2008). Afin de vérifier si *PHYC* n'est pas associé au trait du fait d'un éventuel DL avec un locus voisin, nous avons analysé la structure du DL dans la région chromosomique couvrant environ 80 kb au voisinage physique de *PHYC*. Cette étude a permis d'analyser 75 SNP et INDEL repartis sur six gènes trouvés dans la région (dont *PHYC*). Nous avons également développé un algorithme MCMC pour comparer ces marqueurs et identifier ceux qui sont les candidats les plus probablement responsables du phénotype.

Patterns of Linkage Disequilibrium and Association Mapping in the Vicinity of Phytochrome C Gene in Pearl Millet.

AUTHORS:

Abdoul-Aziz SAÏDOU[1,2,3,4], Jérémy CLOTAULT[1], Marie COUDERC[1], Cédric MARIAC[1], Katrien M. DEVOS[5], Anne-Céline THUILLET[1], Ibrahim ADAMOU AMOUKOU[2] and Yves VIGOUROUX[1]*.

[1] UMR DIADE, Institut de Recherche pour le Développement, Université Montpellier 2, BP 64501, 34394 Montpellier, France;

[2] Université Abdou Moumouni, BP 11040, Niamey, Niger;

[3] Montpellier SupAgro, 2 place Pierre Viala 34060 Montpellier, France;

[4] Institut de Recherche pour le Développement, BP 11416, Niamey, Niger; and

[5] Institute of Plant Breeding, Genetics and Genomics & Department of Plant Biology, University of Georgia, 3111 Plant Sciences, Athens, GA 30602, USA.

* Corresponding author:

Y Vigouroux
Institut de Recherche pour le Développement,
911, avenue Agropolis,
34394 Montpellier, France.
Phone: 33 (0) 467416165
Fax: 33 (0) 467416222
Email: yves.vigouroux@ird.fr

ABSTRACT

Great research efforts have been made to dissect phenotype-genotype relationship in living organisms. Association mapping has been introduced in plants recently to use historical recombination accumulated in diverse existing populations. This recombination breaks linkage disequilibrium between adjacent single nucleotide polymorphisms, providing a high resolution for the localization of genomic regions associated with phenotype. However, persisting linkage disequilibrium extends sometimes to a relatively long distance, notably in genomic regions containing polymorphisms that have been targets of selection. In such a case, numerous markers in the region would statistically associate to the trait shaped by the selected polymorphism. This makes the depicting of causal polymorphism less obvious. *PHYC* has been suggested through a candidate gene-based association study to be associated to flowering time and morphological variation in pearl millet. In the current study, we analyzed a larger genomic region of ~80 kb around *PHYC.* Using a panel of 90 pearl millet inbred lines, we showed that linkage disequilibrium decreases sharply around *PHYC*. Finally, the inbred panel was used to assess association between seven phenotypic traits and polymorphisms spanning the 80kb region around *PHYC*. Significant associations were reported with flowering time, spike length, number of tillers and basal spike diameter. We confirmed these results in a genotype-phenotype association at *PHYC* locus using three pearl millet F2 families. Finally, to pinpoint markers showing the strongest association, we developed a *Markov Chain Monte Carlo* (MCMC) approach to compare 75 markers distributed along the 80 kb region. Polymorphisms located in *PHYC* are the most strongly associated with the different phenotypes. Altogether, these results allow to validate *PHYC* as a strong candidate for the genetic basis of the different phenotypes and to narrow down the potential functional polymorphism associated with these phenotypes.

KEY WORDS: Association mapping, MCMC, linkage disequilibrium, *PHYC*, pearl millet.

SHORT RUNNING TITLE: Linkage disequilibrium and association mapping in pearl millet

INTRODUCTION

The study of the link between genotype and phenotype is a central theme of biology. In recent years, association mapping methods have been developed to exploit collections of existing populations for the analysis of genotype–phenotype relationship (Rafalski 2010, Bergelson and Roux 2010, Myles *et al.* 2009, Zhu *et al.* 2008). One advantage of these diverse collections is the presence of recombination events which accumulated inside populations through long periods of time. This historical recombination provides a mapping resolution that could be sometimes expected up to few kb (Yu and Buckler 2006). Moreover, association mapping could be used to validate the phenotypic role of polymorphisms associated with selection, bridging the gap between the identification of molecular selection signature and the identification of the traits shaped by selected polymorphisms (Mariac *et al.* 2011).

Two types of association studies are now coexisting: candidate gene association studies (CGA, *e.g.* Camus-Kulandaivelu *et al.* 2008, Brown *et al.* 2008, Saïdou *et al.* 2009) and genome-wide association studies (GWA, *e.g.* Atwell *et al.* 2010). The density of markers needed for an efficient GWA depends on genome size and on the extent of linkage disequilibrium at the genomic scale (LD). When LD is relatively extensive, a lower density of markers is sufficient to survey the whole genome. A number of 250,000 SNPs were used for GWA in Arabidopsis (Atwell *et al.* 2010) while 1.6 millions SNPs was used in maize (Tian *et al.* 2011). Candidate gene approach is a more flexible alternative when LD decreases rapidly (making the demand of markers density too high), or for orphan crops with low available genomic information (*e.g.* pearl millet).

The control of false positive rate (type I error) in plant association mapping requests a great attention. The adherence of simplest statistical models to nominal type I error rate is often biased due to the confounding effect of structured background sharing between individuals (Pritchard *et al.* 2000). Most sophisticated statistical models accounting for population structure were proposed to limit this bias (Thornsberry *et al.* 2001, Yu *et al.* 2006). However, the performance of these models to completely deal with the bias depends on data characteristics. For example, false positive limitation with mixed model (Yu *et al.* 2006) has been less efficient for pearl millet traits highly linked to population differentiation (Saïdou *et al.* 2009). The use of linkage mapping as a complementary method for association confirmation is a way to overcome this shortcoming. Linkage mapping is based on controlled crosses and is no longer biased by population structure effect. Marker-trait associations highlighted by association mapping could be consider as true positive with a markedly highest confidence when the markers colocalizes with QTL revealed by linkage mapping (Bergelson and Roux 2010).

Association mapping finds genomic regions associated with phenotype by exploiting usually the LD between genotyped markers and the unknown real causal polymorphisms. Causal polymorphisms shaped by selection often produce extensive LD in their surrounding region. This high LD leads to significant statistical associations with a large number of markers in the region (*e.g.* Ducroq *et al.* 2008). The genome could thus be surveyed using a good coverage of markers (GWA) or using markers linked to particular genomic regions (CGA). Regions corresponding to markers with significant association need to be finely examined to search for the likely causal polymorphisms. Causal polymorphisms could be located far away from genotyped markers, depending on LD (Camus-Kulandaivelu *et al.* 2008). To deal with this, larger fragments of genome around significantly associated markers could be considered to analyze extent of LD and association pattern in order to finally pinpoint the best candidate polymorphisms (Ducroq *et al.* 2008, Brown *et al.* 2008). This approach allows discarding non-causal markers that associate to phenotypic traits due to the simple effect of LD. This allows to pinpoint polymorphisms which could then be assess using functional studies.

A previous candidate gene study in pearl millet (*Pennisetum glaucum* [L.] R. Br.) examined a shorter fragment of *PHYC* (866 bp) and reported significant association with flowering time and morphological variation (Saïdou *et al.* 2009). In the current study, we used linkage analysis into three pearl millet F2 families to confirm the association between *PHYC* locus and several quantitative traits. We then used an association mapping panel of 90 pearl millet inbreds to examine the pattern of linkage disequilibrium around phytochrome C gene (*PHYC*) and to assess marker-trait association in this genomic region using mixed model. This study took into account 75 SNPs and INDELs from a larger genomic region (~80 kb) in the vicinity of *PHYC*. Full *PHYC* locus (~6 kb) and fragments of five putative genes surrounding *PHYC* were sequenced. We finally developed a Markov Chain Monte Carlo approach (MCMC) based on probability and mixed model fit to identify the best candidate loci associated with phenotype.

MATERIAL AND METHOD

Association mapping

Plant material and phenotypic scores. Association analysis was performed using a panel of 90 inbred lines derived from diverse pearl millet material (Saïdou *et al.* 2009). Seven traits were considered: basal primary spike diameter (BSpD), number of days from sowing to the female flowering stage (FT), number of basal tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), primary spike diameter (SpD) and primary spike length (SpL). Field trials and phenotype measurements have been described (Saïdou *et al.* 2009, Saïdou *et al.* submitted). A total of 9 field trials were performed between 2005 and 2008 at Sadoré, Niger. In each trial, approximately 7 to 10 plants per inbred line were measured for each trait to calculate the average phenotype for each inbred line. The complete design consisted thus of 810 entries per trait, which represents the average inbred scores of the 90 inbreds into the 9 trials (except BSpD that was scored only in 5 trials).

Sequencing. Protocols for DNA extraction and PCR have been described previously (Mariac *et al.* 2006b, Oumar *et al.* 2008). Primers were designed to sequence different fragments of *PHYC* (Table S1). We sequenced a single BAC clone from pearl millet (Allouis *et al.* 2001). Primers were designed based on this BAC sequence to complete *PHYC* sequencing and to sequence loci in the neighboring region of this gene (Table S1). The primers were designed on loci that blasted predicted or known genes of *Sorghum bicolor* (*BLASTN*, near-exact matches, GRAMENE 32). Fragments of the genes identified by this analysis were sequenced on the whole panel of inbreds. Independently from the BAC, we also sequenced two additional genes on this panel (named *Pg7840* and *Pg7880* hereafter). These two genes are found very close to *PHYC* in rice and sorghum (name in sorghum database: Sb01g007840 and Sb01g007880). These genes are also known as closely linked to *PHYC* based on pearl millet genetic map (Y. Vigouroux and C. Mariac, *ongoing study*). The BAC clone was sequenced using the method *454* (Roche Applied Science). All the sequences on the 90 inbreds were developed using a classical sequencing protocol previously described (Saïdou *et al.* 2009).

Polymorphism scoring. SNP/INDEL polymorphisms were extracted from sequence data (Geneious 4.8.5). Entries with ambiguity in base scoring were checked and if ambiguity remains, the entries were set as missing data. A few polymorphic sites presented a third allele hold by less than 2 individuals; these entries were also set as missing. The finally checked data matrices were then filtered for association mapping based on a minor allele frequency (MAF) threshold of 2.5%.

Population background. The genetic background of the inbred panel has been assessed in a previous study (Saïdou *et al.* 2009). The panel was structured into different populations and inbreds presented a globally moderate level of kinship relatedness. We used in this study ancestry matrix based on seven populations inferred using the bayesian model STRUCTURE (Pritchard *et al.* 2000). We also used a kinship matrix inferred using SPAGEDI (Hardy and Vekemens 2002). The kinship matrix has been transformed to be positive-definite to fit mixed model requirements (Saïdou *et al.* submitted).

Association analysis. Marker-trait association was based on 9 field trials available on the inbred panel. Single markers were respectively fitted to each phenotype using mixed linear model (see Yu *et al.* 2006). Fixed effect of ancestry (Q) was included using population structure matrices. The matrix of kinship (K) was used to set the relationship between individuals to estimate random effect of background. Field trial was added as random factor. The model was fitted using restricted maximum likelihood method (ASReml-R 2.0/32). The significance of fixed effects was assessed using incremental Wald test (Gilmour *et al.* 2006). We ordered terms in the model so that marker effect was adjusted for ancestry fixed effects which were fitted prior.

Correlation of p-values across traits. For each pair of traits, p-values for marker-trait association provided by the mixed model analysis were plotted across markers to check graphically markers that are significantly associated to both traits. We also assessed the correlation between these p-values distributions using Pearson's coefficient and their significance assess using Student's test. P-values (P) were transformed as $-\log_{10}(P)$ before these analyses.

QTL analysis

F2 plant families. F1 plants were derived from biparental crosses between cultivated and wild pearl millet individuals (Table S4). The progenies of these crosses were named respectively A7, C1 and D1. Fifteen microsatellite markers were genotyped on parents and F1 progeny to control each cross (data not shown). One single F1 individual was selfed to derive each F2 family used in this study. Two of the 15 microsatellites were genotyped into the F2 to validate self-pollination.

Field trial. Experiment was conducted in the field at Sadoré, Niger. Seeds of the F2 plants were sown into pots on 1 June 2009 and transplanted on the ground when the young plants reached 4-6 cm. For each family, plants were grown into plots of 8 columns and 12 rows with a distance of 1 m between rows and 0.7 m between columns. The trial was conducted under rainfall with supplemental irrigation when necessary. Flowering time was measured as the number of days from sowing to heading date (FTHE). The traits BSpD, NTHE, PH, SD, SpD and SpL were recorded as described for the inbred panel. Each plant phenotype was individually score. The total number of F2 plants was 270 for A1, 272 for D1 and 1182 for C1.

Genotyping of PHYC SNP and statistical analysis. DNA was extracted from leaf fragments of each individual plant. A restriction enzyme targeting The SNP at position 5525 in *PHYC* was used to genotype this SNP. We used a protocol described in (Saïdou *et al.* 2009). The genotype was scored as CC, GG and CG with respect to the digestion pattern. QTL analysis was performed based on this single marker. We fitted genotype with each trait using a generalized linear model (R 2.7.2). The significance of marker effect was assessed using Fischer's test.

LD analysis and MCMC approach to identify the best candidate SNPs

Measure of linkage disequilibrium. LD between all pairs of polymorphic sites inside and between all sequenced genes was calculated as squared correlation coefficient r^2. This measure was assessed for significance using two sided Fischer exact test. The R package LDtests (Lewin 2008) was used to calculate and test LD and the R package LDheatmap (Shin *et al.* 2010) was used to plot heatmap.

MCMC method. To highlight the best markers associated with trait, we implemented an iterative process of pairwise comparison of markers through the space defined by all of the available 75 markers. The algorithm was defined as follows:

1) Pick an initial SNP (or INDEL) at random; this SNP define the current position and is referred to as SNPcp.
2) Repeat the following step several times (N iterations):
    a. Pick a second SNP at random (noted SNPnp);
    b. Perform association analysis with SNPcp and SNPnp respectively by fitting mixed model to the same panel subset (inbreds with missing entries in either SNP are discarded to obtain a common sample with the same set of observations for the two markers);
    c. Compare the log-likelihood associated with SNPcp and SNPnp (L_{SNPcp} and L_{SNPnp}):
        i. If $L_{SNPnp} > L_{SNPcp}$, SNPnp is selected;
        ii. If $L_{SNPcp} > L_{SNPnp}$, the probability to select SNPnp is *P* (SNPnp) = $10^{-\Delta L}$ and the probability to select SNPcp is *P* (SNPcp) = $1 - 10^{-\Delta L}$; where $\Delta L = L_{SNPcp} - L_{SNPnp}$;
    d. Record the original ID of the selected SNP;
    e. Reset the new current position at the selected SNP and start next iteration.

3) Analyze the record of selection over iterations and calculate the frequency corresponding to the choice of each SNP (selection score).

Computationally, the probability $10^{-\Delta L}$ defined in the step 2 was set by comparing ΔL to a random value β sampled from a uniform random distribution of the range 0–1. SNPcp was selected for each iteration with $\Delta L >= -\log_{10} (\beta)$; otherwise, SNPnp was selected. So the probability of selecting SNPcp increased with ΔL which measures the improvement of fit provided by this marker compared to a randomly selected SNP.

We aimed to check the effect of the total number of iterations and burn-in length on MCMC result. First, we set the total number of iterations (including burn-in) to 200,000 iterations and we varied the length of burn-in from 10,000 to 100, 000. Second, we set the burn-in period at 10,000 and we varied the value of total chain length from 20,000 to 200,000 iterations. The frequencies of selection obtained for the different markers (selection scores) were compared with respect to MCMC parameters (Pearson's correlation and Student's test). Markers with a null score for all parameter values were discarded from correlation test. Parameters were set using flowering time trait. The analysis was then performed on each of the following traits: flowering time (FT), spike length (SpL), number of tiller at head emergence (NTHE) and basal spike diameter (BSpD).

RESULTS

Homology with sorghum. A total of 31 contigs were obtained through the sequencing of pearl millet BAC. The total size of contigs was 95,212 bp initially and 87,469 bp after vector inserts trimming (GENEIOUS 4.8.5). Individual contig size varied between 109 bp and 21,379 bp based on the filtered sequences. Four loci from pearl millet BAC provided significant BLAST with sorghum genes: *PHYC* and three genes that we named respectively *Pg7830*, *Pg7870* and *Pg7878* (Table S2). The prefix *Pg* refers to the species (*Pennisetum glaucum*) and the number refers to the index of the homolog gene in sorghum. The two supplemental genes (*Pg7840* and *Pg7880*) also produced significant BLAST with sorghum genes (Table S2). So we obtained a total of six genes with known or predicted protein coding in sorghum. Due to incomplete assembly of contigs to date, the physical position of these genes in pearl millet is not available (BAC assembly in progress). All the different genes (*Pg7830*, *Pg7840*, *Pg7870*, *Pg7878*, *Pg7880*) are found in the vicinity of *PHYC* in sorghum, and are in the same position around *PHYC* in rice suggesting little changes in gene order between rice and sorghum. All these genes were located on a unique chromosome in sorghum (chromosome 1). So we used the physical location of the homolog genes in sorghum to suggest a putative ordering (Figure

1). However, *Pg7840* was not found in the 80 kb BAC of pearl millet and could be located before *Pg7830* in this species.

Polymorphism and LD pattern. Complete *PHYC* locus (6115 bp) and fragments of 807 to 923 bp were sequenced for the 5 others genes on the 90 inbred lines. The sum of sequence size was 10,423 bp. A total of 75 SNP/INDEL with minor allele frequency superior to 2.5% was found (Supplemental file S1). Across the examined genomic region, LD value (r^2) ranged from approximately null value to 1 and presented an average pairwise measure of r^2=0.35 ±0.01 (Figure 2). The range and the average value of intra-gene LD were relatively different across genes. Polymorphisms inside *PHYC* were the most tightly linked and showed average r^2 of 0.74 ±0.01 (r^2 ranged from 0.15 to 1 for this gene). *Pg7830* (and *Pg7840* in a lesser extent) showed the highest LD with *PHYC*. The markers in *Pg7870*, *Pg7878* and *Pg7880* presented no LD (or only a weak LD) with *PHYC*.

Marker-trait association. Each of the 75 markers was fitted to phenotypic traits using mixed model. Numerous markers were significantly associated with flowering time (FT), primary spike length (SpL), number tillers at head emergence (NTHE) and basal spike diameter (BSpD) [Figure 3 and Supplementary file S2]. For plant height (PH), stem diameter (SD) and spike diameter (SpD), no association with high significance was found. The observed associations were linked to INDELs and SNPs at *PHYC*, *Pg7830* and *Pg7840*. Globally, no associations were detected with the other genes (*Pg7870*, *Pg7878* and *Pg7880*). Markers associated with BSpD were located at *PHYC* and *Pg7880*.

QTL analysis. The F2-based study revealed significant associations between the genotype at *PHYC* and different traits (Figure 5). Associations with BSpD, SpD and SpL were noticed with a highly significant p-value in all families. The association of the marker with flowering time (FTHE) was significant into families D1 and C1 ($p=7.93 \times 10^{-3}$ and $p=6.09 \times 10^{-20}$ respectively), but not significant in A7 family (p>0.05). Significant associations with stem diameter were found for A7 and C1 crosses. Finally, significant association was found for PH and NTHE for the cross C1 only.

Analysis of shared associations across traits. A part of the polymorphic sites statistically associated with phenotype were shared between two or more traits (FT, SpL, NTHE and BSpD). For each pair of traits, we performed a graphical comparison and a correlation between p-values to highlight markers that are associated to both traits at the same time (Figure 6). FT and SpL appeared to be associated with the same set of markers. Markers providing the highest p-values for FT also provided the highest p-values for SpL. The correlation of paired p-values across markers was positive and significant for these two traits, based on the whole set of markers (R=0.77, df=73, p=4.44 x 10^{-16}) and/or based on *PHYC* markers (R=0.81, df=33, p=2.79 x 10^{-9}). The remaining combinations of traits globally produced either null or weak correlation of p-values, or negative correlations of p-values.

Comparison of markers using MCMC analysis. We tested different parameter values for the MCMC analysis (Supplemental files S3-S4). All values in the considered range led to very similar results (R^2>0.99 for all pairwise comparisons of burn-in length and R^2>0.97 for all pairwise comparisons of total chain length). So we set the total number of iterations to 30,000 including a burn-in period of 10,000. The MCMC process was implemented for FT, SpL, NTHE and BSpD traits which showed significant association with markers into the initial mixed model analysis.

We reported the frequency of selection of each single marker with respect to the trait (Figure 4). The most frequently selected polymorphic sites were: *PHYC*$_{5004C>A}$ for FT and SpL traits (respectively 13.0% and 18.0%), *PHYC*$_{3219A>C}$ for NTHE (23.2%) and *PHYC*$_{4967G>A}$ for BSpD (15.8%). With respect to these scores, *PHYC* appeared as the best candidate for all of different traits FT, SpL, NTHE and BSpD.

DISCUSSION

Consistency between association mapping and QTL analysis. Unlike association mapping, the family-based QTL study does not suffer from false positive inflation due to population structure. We used F2 families derived from 3 different crosses and a marker in *PHYC* to check for any QTL linked to the traits of interest. The analysis supported the presence of QTL in this genomic region for different traits. All the 4 traits for which we reported significant associations based on association mapping (FT, SpL, NTHE and BSpD) were associated also in the QTL analysis. This result suggests that these detected effects are true positive. The association for flowering time (FTHD) was detected into two families (D1 and C1), but not in the third family (A7). Based on the results observed into D1 and C1, it could be suggested that a QTL does exists but the linkage between the causal loci underlying this QTL and the genotyped marker at *PHYC* has been broken by recombination into A7 family. The observed associations were higher into C1. This was actually expected, as the power for QTL detection depends on sample size (n=1182 for C1). Additional genotype-phenotype associations not identified into the inbred panel were noticed into the F2 families (Figure 5). These associations could be due to close linkage of multiple QTLs with the analyzed marker. Finally, based on the associations highlighted in the inbred panel and in the F2 families, the existence of true positive loci associated with trait could be assumed with a strong confidence for this chromosomal region.

Deciphering of the best statistically associated polymorphisms. Based on the mixed model analysis (Figure 3), several markers at *PHYC*, *Pg7830* and *Pg7840* showed significant association with FT, SpL, BSpD and NTHE. These three genes are tightly linked (r^2 up to 1 between *PHYC* and *Pg7830* on the inbred panel). In the presence of such LD, association at one polymorphism is expected to cause virtual association at the majority of linked loci. A comparison across markers was thus carried out to decipher the markers supporting the best association.

Inbred with missing genotype for a given marker were discarded from the analysis with this marker. The final subset of inbred was thus relatively different across markers (n=77 to n=87); and so was the number of observations (Supplemental file S2). This difference could cause a bias when comparing markers directly based on the p-values. For example, the rank of markers *Pg7830*$_{767T>G}$ and *PHYC*$_{5004C>A}$ in terms of significance level could be inversed by the change of data subset (Table S3). We implemented the MCMC method to overcome this shortcoming and to assess a statistically formal comparison between markers. All markers were compared iteratively based on pairwise common samples. The probability of selecting a marker as the best was based on the log-likelihood for the model with this marker. No correction for sample size or for the number of parameters was made, as these two properties were strictly common to both markers in each step of pairwise comparison. The frequency of selection of a marker through this process was expected to increase proportionally to the improvement of fit supported by this marker compared to the rest of markers. Across the genes examined in this region, *PHYC* appeared to be the best candidate for all of FT, SpL, BSpD and NTHE traits.

We examined the region around *PHYC* by sequencing homologs of grass genes. Highest associations could be validly found apart from the sequences we examined. The density of markers is reasonable (75 markers for 80 kb) only if these genes are approximately well distributed along the region with respect to the extent of LD. The pattern of LD variation with physical distance might help discussing these results. However, the incomplete assembly of the BAC did not allow a complete physical positioning of the genes. Furthermore, it could not be totally discarded that the causative polymorphisms are more or less far away from this 80 kb region, even though this eventuality is not actually expected with high probability for an outcrossing species. For comparison, note for example that LD around *Vgt1* locus in maize presented magnitudes of r^2>0.7 over 1 kb and r^2 up to 0.4 at 70 kb apart (Ducroq *et al.* 2008).

Pleiotropy versus close linkage. The MCMC analysis localized the best statistically associated polymorphisms for FT, SpL, NTHE and BSpD at *PHYC* gene. These results suggest pleiotropy at the gene scale. But the deciphering of association at the nucleotide level would lead to discern three situations: i) the association between different traits is driven by a same set of pleiotropic molecular polymorphisms (same SNP hypothesis); ii) association is driven by different sets of polymorphisms but appeared pleitropic because of linkage disequilibrium (linkage disequilibrium hypothesis), iii) different unlinked polymorphisms drive the association between different traits (unlinked SNP hypothesis). The analysis of correlation of p-values obtained for two different traits could assess the difference between hypothesis i or ii versus hypothesis iii (Figure 6). This analysis showed a cutting difference between the six tested combinations of traits, suggesting that FT and SpL shared the same set of associated markers. For the remaining combinations of traits, the observed null or weak correlation of p-values could be interpreted as independence between the p-values, so being associated to one trait does not imply any information about the association with the second trait (FT−BSpD for example). Negative correlations mean that polymorphisms linked to one trait are not linked to the other, and inversely (NTHE−BSpD for example). The hypotheses of pleiotropy or linkage could be tested into a more sophisticated QTL mapping using multivariate analysis for example (Stich *et al.* 2008; Szyda *et al.* 2003).

In the F2 family A7, association with flowering time was not observed while the association with spike length was still observed. This result suggests that the functional polymorphism(s) associated with each trait are probably different. The significant correlation found between flowering time and spike length association p-values might be driven by linkage disequilibrium (linkage disequilibrium hypothesis). The hypothesis is that selection for one of the traits drives by linkage functional polymorphism associated with another trait. So there is an apparent pleiotropy at the SNP if we analyze correlation of polymorphisms p-values but this does not means necessarily that the same functional SNPs drive the association with these two phenotypes. On a short time scale the effect of hypothesis i and ii are similar for the evolution of populations: selection for one trait would be associated with changes in the other trait. However, in the long term, the linkage case (hypothesis ii) would allow the sets of SNP associated with different traits to be disjunct thanks to recombination.

CONCLUSION

This study used association mapping jointly with linkage analysis to provide evidence for the presence of very likely true positive polymorphisms in the genomic region around *PHYC*. The MCMC analysis suggests that *PHYC* is probably the best candidate associated to trait among the set of genes sequenced. This reinforces the hypothesis of a possible role of *PHYC* in the phenotypic variation of pearl millet. Further studies might refine LD assessment using physical distance and potentially a larger genomic region. Moreover, functional studies could be designed to validate the best candidate polymorphisms highlighted by this study.

REFERENCES

Allouis S, X Qi, S Lindup, MD Gale and KM Devos (2001). Construction of a BAC library of pearl millet, *Pennisetum glaucum. Theor Appl Genet* 102: 120–125

ASReml package for R (ASReml-R), version 20/32. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Atwell S, Y S Huang, BJ Vilhjalmsson, G Willems, M Horton et al (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465: 627-631

Bergelson J and F Roux (2010). Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nature Reviews* 11: 867-879

Brown P J, W L Rooney, C Franks and S Kresovich (2008). Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. *Genetics* 180: 629–637

Camus-Kulandaivelu L, LM Chevin, C Tollon-Cordet, A Charcosset, D Manicacci and M I Tenaillon (2008). Patterns of molecular evolution associated with two selective sweeps in the Tb1-Dwarf8 region in maize. *Genetics* 180: 1107-1121

Drummond AJ, B Ashton, S Buxton, M Cheung, A Cooper *et al.* (2010). GENEIOUS v4.8.5. http://www.geneious.com

Ducrocq S, D Madur, JB Veyrieras, L Camus-Kulandaivelu, M Kloiber-Maitz *et al.* (2008). Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. *Genetics* 178: 2433-2437

Gilmour AR, BJ Gogel , BR Cullis, and R Thompson (2006). ASReml User Guide Release 20. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

GRAMENE release 32. http://www.gramene.org/ (access online: November 2010).

Hardy OJ and X Vekemans (2002). SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2: 618-620

Lewin A (2008). R package: Exact tests for linkage disequilibrium and Hardy-Weinberg equilibrium. http://www.r-project.org

Mariac C, V Luong, I Kapran, A Mamadou, F Sagnard *et al.* (2006). Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L] R. Br.) in Niger assessed by microsatellite markers. *Theor Appl Genet* 114:49-58

Mariac C, Jehin L, AA Saïdou, AC Thuillet, M Couderc, S Sire, H Jugdé, H Adam, G Bezançon, JL Pham, Y Vigouroux (2011). Genetic basis of pearl millet population adaptation along an environmental gradient investigated by a combination of genome scan and association mapping. *Mol Ecol* 20:81-91

Myles S, J Peiffer, P J Brown, E S Ersoz, Z Zhang, D E Costich, and E S Buckler (2009). Association mapping: critical considerations shift from genotyping to experimental design. *The Plant Cell* 21: 2194-2202

Oumar I, C Mariac, JL Pham and Y Vigouroux (2008). Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L] R. Br.) as revealed by microsatellite loci. *Theor Appl Genet* 117: 489-497

Pritchard JK, M Stephens, NA Rosenberg and P Donnelly (2000). Association mapping in structured populations. *Am J Hum Genet* 67:170-181

R Development Core Team (2008). R: A language and environment for statistical computing R. Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org

Rafalski JA (2010). Association genetics in crop improvement. *Curr Opin Plant Biol* 13:1-7

Roche Applied Science. 454 Sequencing. http://454com/

Saïdou AA, C Mariac, V Luong, JL Pham, G Bezancon, and Y Vigouroux (2009). Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics* 182: 899-910

Shin JH, S Blay, N Lewin-Koh, B McNeney and J Graham (2010). R package: Graphical display of pairwise linkage disequilibria between SNPs. http://www.R-project.org

Stich B, J Mohring, HP Piepho, M Heckenberger, ES Buckler and A E Melchinger (2008). Comparison of mixed-model approaches for association mapping. *Genetics* 178:1745-1754

Szyda J, E Grindflek, Z Liu and S Lien (2003). Multivariate mixed inheritance models for QTL detection on porcine chromosome 6. *Genet Res Camb* 81:65-73

Thornsberry JM, MM Goodman, J Doebley, S Kresovich, D Nielsen and E S Buckler (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 28:286-289

Tian F, P J Bradbury, P J Brown, H Hung, Q Sun *et al.* (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics. doi:10.1038/ng.746*

Yu J and E S Buckler (2006). Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17: 155-160

Yu J, G Pressoir, WH Briggs, I V BI, M Yamasaki *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genetics* 38: 233-208

Zhu C and J Yu (2009). Nonmetric multidimensional scaling corrects for population structure in whole genome association studies. *Genetics* 182: 875-888

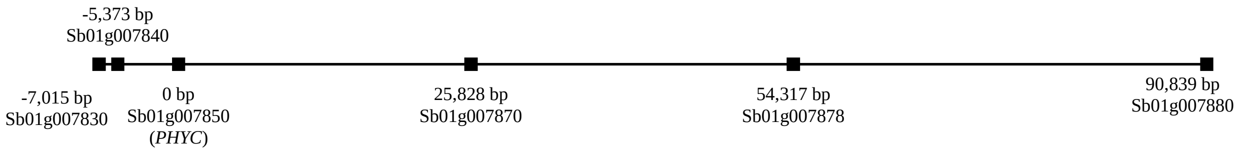
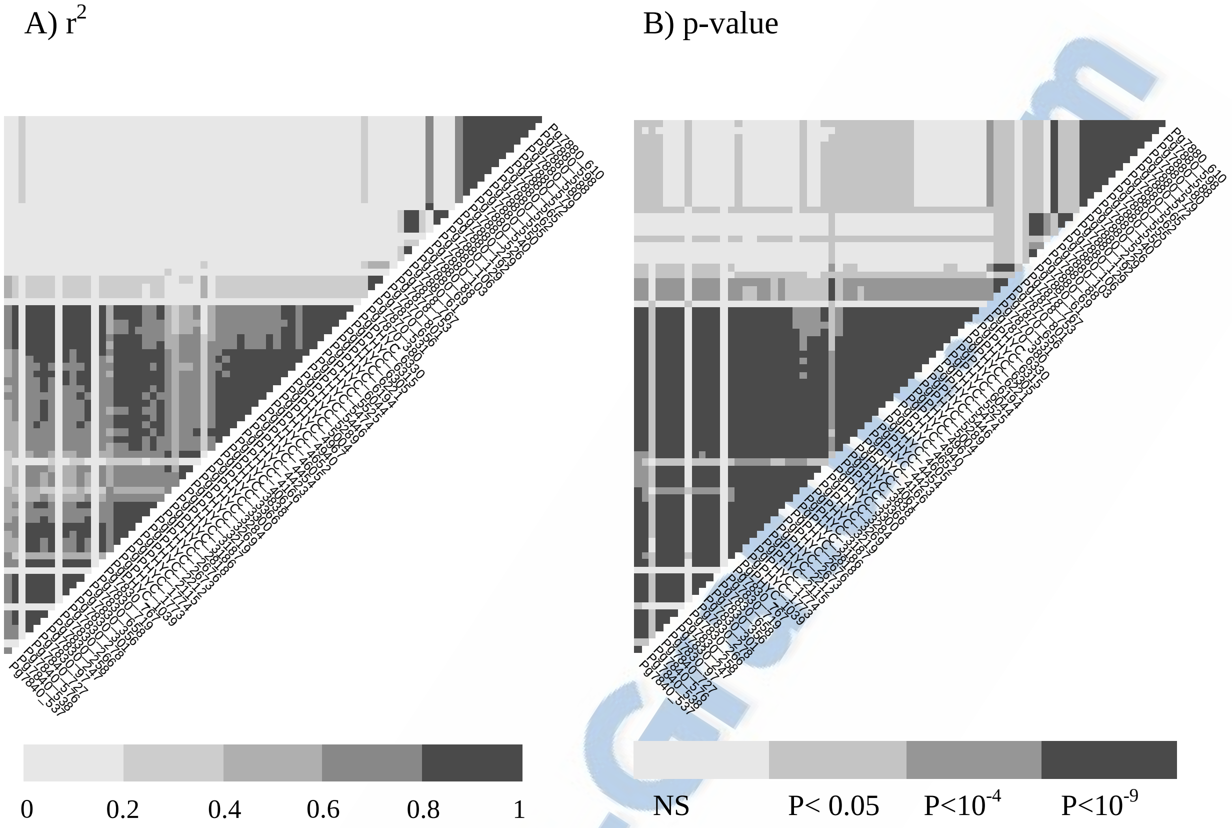


Figure 1. Physical position of pearl millet genes homologs in *Sorghum bicolor*.

Sorghum genes are presented according to the position and the nomenclature of GRAMENE. We rescaled the position with respect to *PHYC* position which was set to zero. All genes are located on chromosome 1 of sorghum.



C) Summary statistics of r^2

Locus	Mean ±SE	Min.	Max.	N
Pg7840	0.41 ±0.14	0.08	0.77	4
Pg7830	0.60 ±0.07	0.02	1	10
Pg*PHYC*	0.74 ±0.01	0.15	1	35
Pg7870	0.61 ±0.17	0.21	1	4
Pg7878	0.17	-	-	2
Pg7880	0.53 ±0.03	0.03	1	20
All genes	0.35 ±0.01	6 x 10^{-7}	1	75

Figure 2. Linkage disequilibrium in the vicinity of *PHYC* in pearl millet.

(A) Pairwise squared correlation coefficient (r^2) is plotted across 75 sites (SNPs/INDELs). (B) P-value of Fischer's test. (C) Summary statistics of r^2 across each respective gene and across all genes. Mean r^2 value, standard error (SE), minimum r^2 value (Min.), maximum r^2 value (Max.) and number of sites (N) underlying each summary statistic are given.

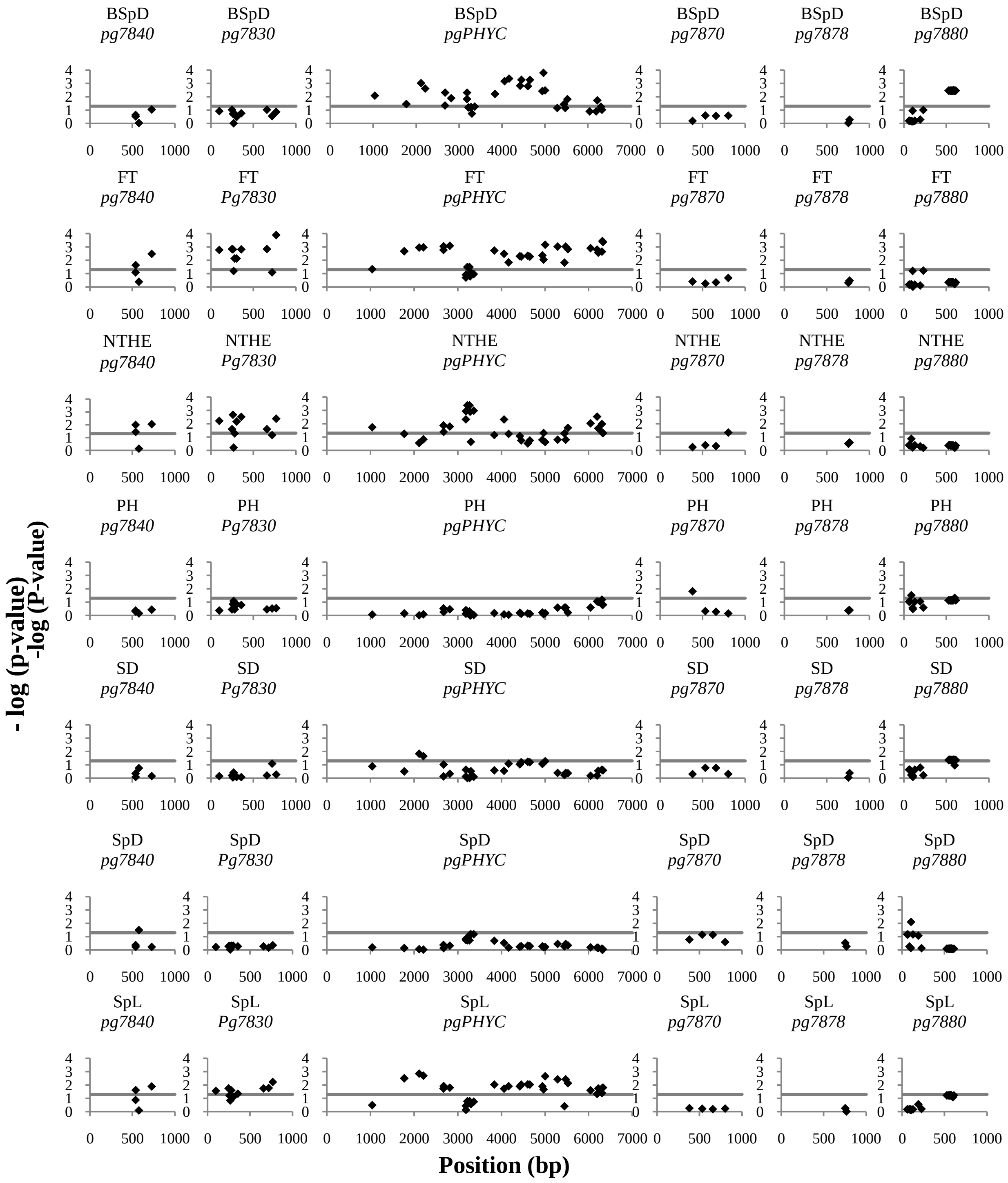


Figure 3. Marker-trait association across genes in the vicinity of *PHYC.*

Association was fitted with the inbred panel using mixed model. The –log(p-value) of marker effect is given at each position (Wald test). The analysis was performed for SNPs and INDELs having a minor allele frequency of at least 2.5%. The markers are distributed across six genes fragments : *Pg7840*, *Pg7830*, *PHYC*, *Pg7870*, *Pg7878*, *Pg7880*. Seven traits were considered : basal spike diameter (BSpD), flowering time (FT), number of tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), spike diameter (SpD) and spike length (SpL). Trait and gène codes are indicated on the top of each graph. The horizontal line represents the non-corrected standard p-value threshold (0.05).

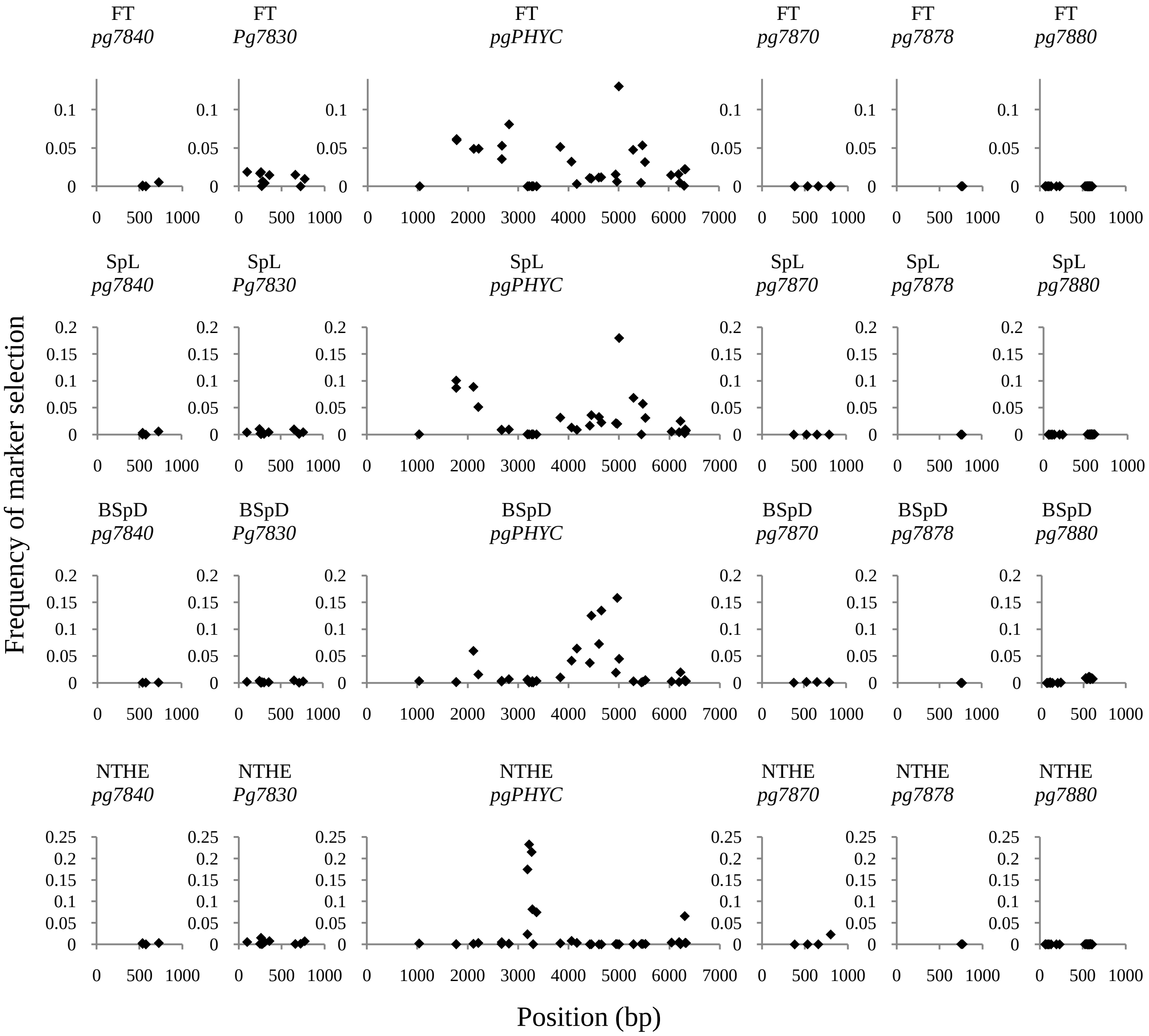


Figure 4. Comparison of markers using MCMC analysis.

MCMC algorithm was used to explore iteratively the whole set of 75 markers distributed on the six genes. In each step, two markers were respectively fitted to the trait by mixed model using a common inbreds subset. The log-likelihood obtained with the two respective markers was used to compare the markers using a criterion of probability (see text for details). Markers presenting best fit with the trait had the highest probability to be selected. The frequency of selection of each marker (selection score) is presented for flowering time (FT), spike length (SpL), basal spike diameter (BSpD) and number of tillers at head emergence (NTHE). The run consisted in 30,000 iterations per trait. *PHYC* gene shows the most frequently selected markers for all the traits.

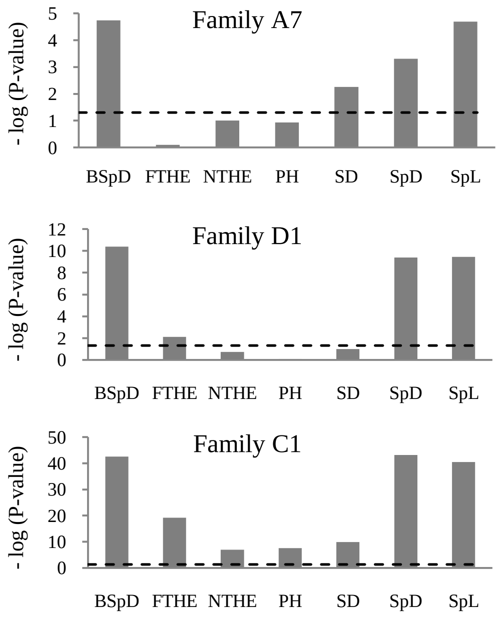
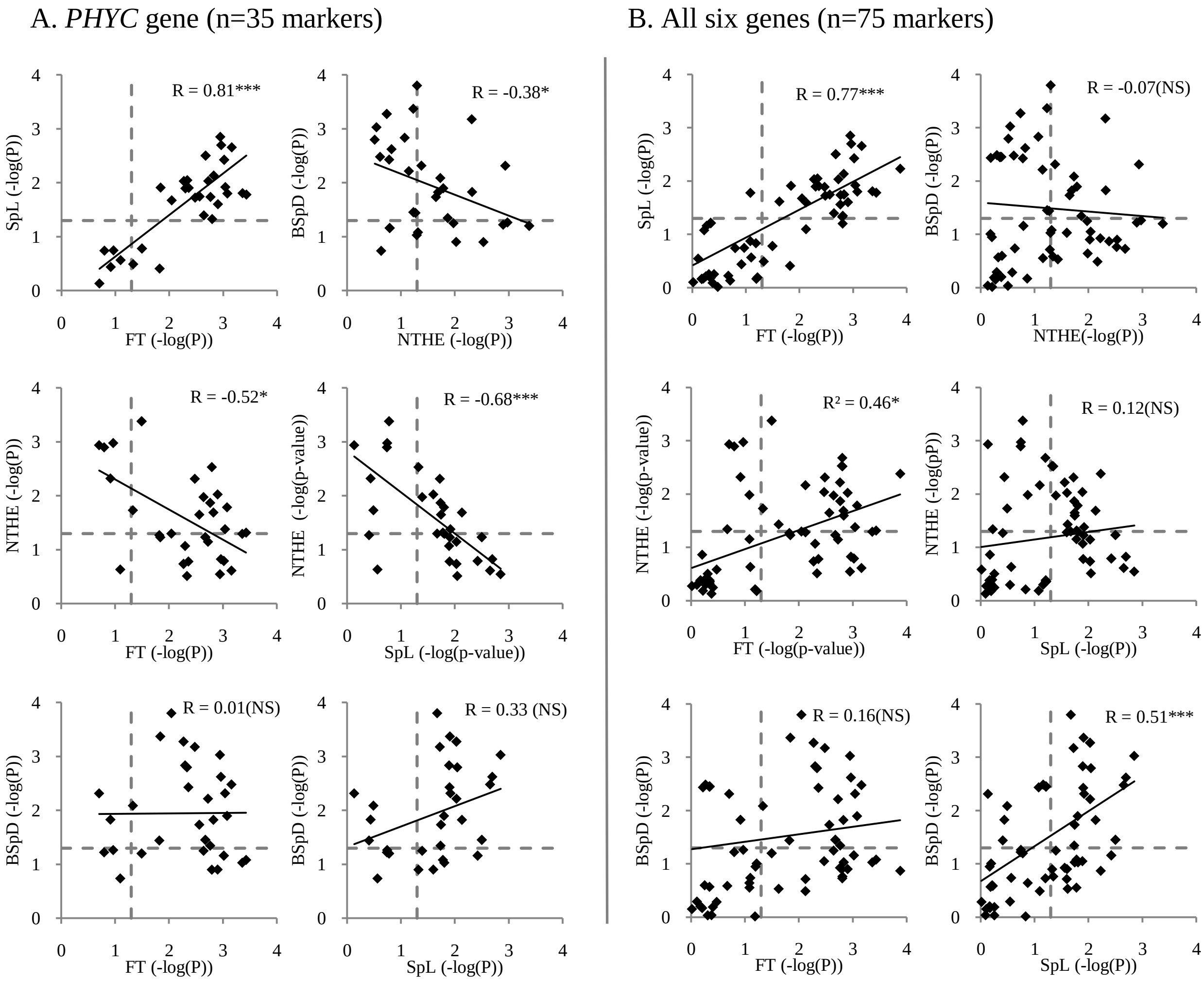


Figure 5. Single marker QTL analysis into F2 families.

Three F2 families were used for the analysis: A7 (n=270), D1 (n=272) and C1 (n=1182). The effect of *PHYC* genotype was assessed respectively on each of the seven traits indicated on the first axis (see text for details). Data were fitted using generalized linear model. The -$\log_{10}$(p-value) is given (Fischer's test). Dashed line represents the standard p-value threshold (0.05). Note the change in scale across graphs.



A. *PHYC* gene (n=35 markers)

B. All six genes (n=75 markers)

Figure 6. Sharing of associated markers across traits.

Marker-trait association was assessed on the inbred panel using mixed model for each trait: flowering time (FT), primary spike length (SpL), number of tillers at head emergence (NTHE) and basal spike diameter (BSpD). P-values of marker effect (Wald test) are compared for each pair of traits. P-values were transformed with decimal logarithm and were paired with respect to the marker. Horizontal dashed lines represent the p-value threshold 0.05 for the trait on the first axis, and vertical line represents the same threshold for the trait on the second axis. Dots located over both thresholds indicate markers associated to two traits. Linear regression of the p-values is also plotted (solid line) for each pair of traits. Pearson's coefficient of correlation is given (R). This coefficient was assessed for significance using Student's test (*P<0.05, ***P<0.001 and NS: none significant). The analysis was performed respectively for the 35 markers at *PHYC* gene (A), and for all the 75 markers across the six genes sequenced in this study (B). Significant positive correlation indicates potential pleiotropy. Pleiotropy is highly suggested for FT and SpL based on these results.

Table S1. Primers for pearl millet genes

Gene	Fragment code	Forward primer	Reverse primer	Amplified size (bp)
Pg7878	*Contig6.2 (BAC)*	TCGCACTGACCGCATGTGAAA	TCTCAAGCGGCTAAAGCAAGCA	923
Pg7870	*Contig20 (BAC)*	CACAGGTGGCTGGCTGGCTT	ACCGGTGCTGGTGGGGGAAG	858
Pg7830	*Contig5 (BAC)*	TAAGCTTTCACCGTCCGTTC	GGCCATCAGATTTGTTGCTT	807
PgPHYC	*Contig7.2 (BAC)*	CCGGGTGTCCCAGATGCCCT	TGACGTGGCTGGTGGGACCT	845
	PHYC6b	AGCACCGTCTCCGCCTACC	CTTGCAGGGAGCCGCGTAA	1600
	PHYC6b1R	CGAGGTTATCTCCGAGTGTAGG	TCCACTCTAAGCATGAACCAA	1600
	PHYC7	TCGGGACCTTTCAGATAACG	GATTTGGACAGTTTGGCACTTA	750
	PHYC8	AGGTGCAGAAAATGTCTGAG	ATATACCCCCTTCGTACCAT	785
	PHYC9	GGTATCCCCATAGTCACCAG	TCTACGGACCTAGAGGGAGT	825
	PHYC10	TCCGTGTGACTAAGGGTATC	CGTGGGTTACAGTTACAGG	882
	PHYC4	CAGATTGCTCATYTRGAGTTCA	CGTGCCRCTCATCGTYTTC	800
Pg7840	-	TTTGCTCCTCTTCCTGATATGG	CCCTCTTCACAGGCATAAACC	807
Pg7880	-	CTCCTCATCGTCCTCTCGAC	CTGCTTCGCTTTTAGTGCAA	913

Primers were designed on pearl millet to amplify gene sequences. Fragment code refers to the ID used in our laboratory to recognize different fragments of sequence. Primers directly based on the pearl millet BAC are indicated.

Table S2. BLAST between pearl millet and sorghum sequences

Pearl millet locus	Size (bp)	Sorghum gene	E-value
Pg7830	807	Sb01g007830	0.00044
Pg7840	807	Sb01g007840	4.4e-114
Pg7870	858	Sb01g007870	0.023
Pg7878	923	Sb01g007878	8.3e-94
Pg7880	913	Sb01g007880	4.0e-13
PgPHYC	6115	Sb01g007850 (*PHYC*)	0

E-value of BLAST between pearl millet sequenced fragments and sorghum genes is given. The size of sequenced fragments for pearl millet genes is also given. We named the novel sequences of pearl millet using the corresponding gene index of the sorghum homolog preceded by the prefix *Pg* that refers to *Pennisetum glaucum*. Sorghum genes are named according to GRAMENE nomenclature (Species name, chromosome index, gene index).

Table S3. Comparison of mixed model results at *PHYC* and *Pg7830*

Sampling	Site	N	Effect estimate	P-value	LK
Analysis with different subsets	*PHYC*$_{5004C>A}$	87	4.22 ±1.24	6.93 x 10^{-04}	-1403.14
	Pg7830$_{767T>G}$	77	4.32 ±1.13	1.32 x 10^{-04}	-1278.95
Analysis with common subset	*PHYC*$_{5004C>A}$	75	5.54 ±1.3	2.07 x 10^{-05}	-1227.78
	Pg7830$_{767T>G}$	75	4.24 ±1.15	2.36 x 10^{-04}	-1229.80

Mixed model was fitted to associate polymorphism with flowering phenotype scored in 9 trials. The results of two respective markers at *PHYC* and *Pg7830* are compared with respect to the method used to handle missing sequence data. In the available dataset, entries with missing data were different between these markers. In the first case, the analysis is performed after filtering entries with missing data for each marker considered solely. In the second case, the dataset was filtered to choose only entries with valid sequence data for both markers (common subset). The number N of inbreds in the final subset, the effect estimate (coefficient ±standard error, in days), the p-value for marker effect significance (Wald test) and the log-likelihood (LK) of the model are given. Based on different subset, *Pg7830* marker presents the highest level of significance and a higher effect estimate. Based on the common inbreds subset, the situation is inversed and *PHYC* marker presents the highest level of significance and a higher effect estimate.

Table S4. Linkage mapping families

Family	Parent 1	Parent 2
A7	PE8151-AF2-P2 (Wild, Niger)	PE5887-AF2-P2 (Cultivated, Burkina Faso)
C1	PE8151-AF2-P2 (Wild, Niger)	PE1205-AF2-P3 (Cultivated, Burkina Faso)
D1	PE8504-AF3-P1 (Wild, Senegal)	PE1205-AF2-P3 (Cultivated, Burkina Faso)

For each family, the code of the two parents is given, as well of the type (Wild/Cultivated) and the country of origin.

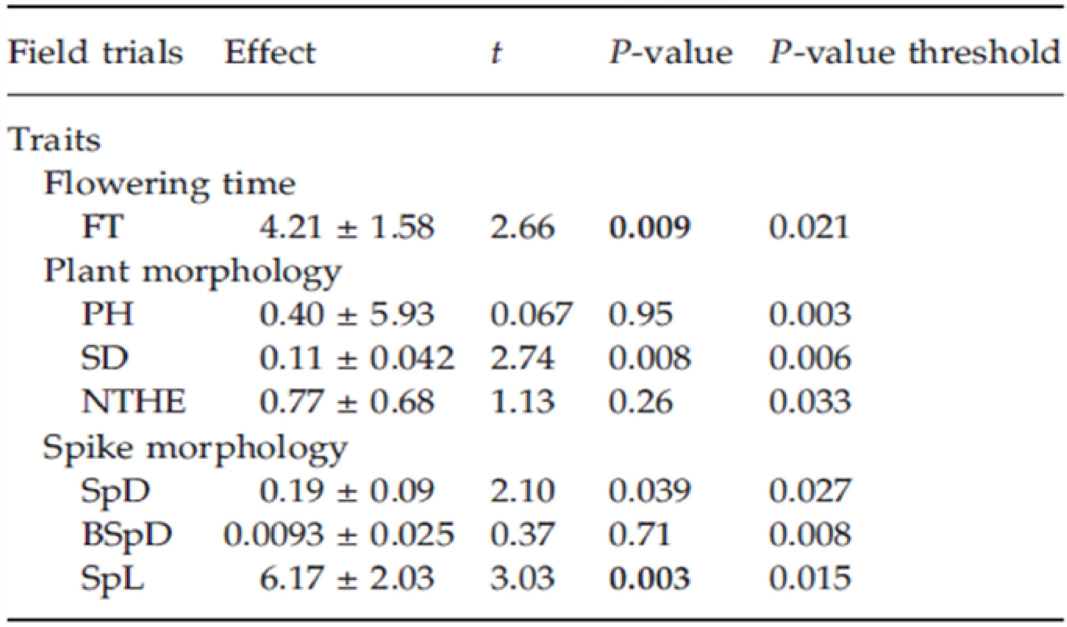
Final note: *The supplemental tables cited for this article (S1-S4) are available on demand as Excel files.*

Résultats complémentaires

Nous avons également appliqué la méthodologie d'association développée dans cette thèse pour valider l'association du gène *MADS11* avec la date de floraison (Annexe 1).

Le gène *MADS11* a été identifié en se basant sur des tests de sélection, qui suggèrent que ce gène a été cible de sélection au cours de l'histoire évolutive du mil (Annexe 1). Ce gène code un facteur de transcription qui est associé à la date de floraison chez d'autres espèces (Becker and Theissen 2003, Atwell *et al.* 2010). Nous avons utilisé le panel de 90 lignées (décrit plus haut) pour tester l'association entre ce gène et différents phénotypes chez le mil (Tableau G2). Cette analyse valide l'association entre ce gène-candidat et deux phénotypes, la date de floraison et la taille de l'épi, suggérant qu'un effet de précocité serait aussi associé à des épis plus courts.

Tableau G2. Association entre *MADS11* et les traits phénotypiques



Field trials	Effect	*t*	*P*-value	*P*-value threshold
Traits				
Flowering time				
FT	4.21 ± 1.58	2.66	**0.009**	0.021
Plant morphology				
PH	0.40 ± 5.93	0.067	0.95	0.003
SD	0.11 ± 0.042	2.74	0.008	0.006
NTHE	0.77 ± 0.68	1.13	0.26	0.033
Spike morphology				
SpD	0.19 ± 0.09	2.10	0.039	0.027
BSpD	0.0093 ± 0.025	0.37	0.71	0.008
SpL	6.17 ± 2.03	3.03	**0.003**	0.015

Association (modèle mixte) entre un marqueur situé dans le gène *MADS11* et 7 traits phénotypique : date de floraison (FT), hauteur de la plante (PH), diamètre de la tige principale (SD), nombre de talles au stade d'épiaison (NTHE), diamètre de l'épi principal (SpD), diamètre du rachis, et taille de l'épi principal (SpL). Sont donnés respectivement : l'estimation de l'effet moyen du marqueur (± erreur standard); la valeur du test statistique (t de Student pour cette analyse), la p-value, et le seuil empirique de significativité. Les tests significatifs en considérant le seuil empirique sont surlignés en gras.

La fréquence allélique du gène *MADS11* varie sur 21 variétés-populations cultivées couvrant l'aire de culture au Niger (Annexe 1). La pluviométrie est différente selon les localités. Elle correspond à un cumul annuel moyen allant de 250 mm à 650 mm sur la zone d'étude.

L’allèle associé sur ce gène à un effet de floraison précoce (ou *allèle de précocité*) est prédominant dans les localités à climat plus sec (fréquence atteignant q=0.7); sa fréquence diminue progressivement (jusqu’à q<0.4) lorsque l’on descend vers les zones les plus humides.

Conclusions et perspectives

L'étude d'association présentée en début de ce chapitre avait révélée un gène candidat associé au phénotype du mil, *PHYC*. Nous avons donc ensuite développé une seconde étude, qui a adressé trois questions importantes relevant des perspectives de la première étude, à savoir :

i) Une analyse QTL indépendante permet-elle d'avoir plus de certitude dans le fait que l'association observée n'est pas une fausse association ?

ii) Quelle est la structure du déséquilibre de liaison (DL) autour du gène *PHYC* ?

iii) *PHYC* est-il le meilleur gène candidat dans la région, ou alors, un gène voisin serait responsable de l'association observé à *PHYC* ?

L'étude que nous venons de présenter complète substantiellement la première en apportant des éléments de réponse à ces questions :

i) Nous avons trouvé un QTL colocalisant avec *PHYC* et renforçant l'hypothèse que la région génomique étudiée (région de *PHYC*) est réellement associée au phénotype.

ii) Le déséquilibre autour de *PHYC* s'émousse sur l'échelle de distance considérée (80 kb), comme le montre le DL très faible (r^2<0.2) entre *PHYC* et les gènes *Pg7870*, *Pg7878* et *Pg7880*. Ce pattern suggère que l'on ne peut espérer avec une forte probabilité de trouver un locus fortement lié à *PHYC* plus loin, sans cependant exclure totalement cette éventualité.

iii) Nous avons utilisé une approche de Monte Carlo (MCMC) pour comparer la force de l'association entre tous les marqueurs repartis sur les six gènes. Cette approche a été basée sur la probabilité de chaque marqueur d'expliquer mieux l'association que les autres marqueurs (ajustement aux données plus fort). Parmi tous les gènes trouvés dans la région, *PHYC* ressort comme le meilleur candidat.

Ces deux études portent sur différents phénotypes d'intérêt (7 phénotypes). Cela nous a permis de discuter les effets pléiotropiques éventuels associés à *PHYC* en même temps que les effets sur la date de floraison. Le pattern observé suggère une pléiotropie à l'échelle du gène, mais pas à l'échelle des nucléotides (SNPs). On montre que la date de floraison (FT), la taille de l'épis (SpL), le nombre de talles au stade d'épiaison (NTHE) et le diamètre du rachis (BSpD) sont tous liés à *PHYC*, mais les SNPs au sein de *PHYC* qui sont les plus probables pour expliquer ces phénotypes ne sont pas les mêmes. Chez *D. Melanogaster*, il a été observé que les effets pléiotropiques au niveau d'un gène quelconque sont dans plusieurs cas portés par des SNPs différents à l'intérieur de ce gène (Flint and Mackay 2009). Ainsi, un gène peut être pléiotropique, mais ce n'est pas le cas des polymorphismes individuels à l'intérieur de ce gène (Mackay *et al.* 2009). Nos résultats sont concordants avec ces observations et constituent donc une preuve supplémentaire, alimentant la réflexion sur cette théorie. Il est à noter cependant que chez l'humain, des SNPs individuels ont été associés à plusieurs phénotypes à la fois (Mackay *et al.* 2009).

La date de floraison est un trait déterminant dans le cycle de vie d'une plante. Ce trait est une des composantes essentielles de la fitness. La *fitness*, ou *valeur sélective*, est le nombre moyen de descendants viables et fertiles produits par un individu ayant un génotype particulier, comparé au nombre de descendants produits par des individus de génotype différent (Roux 2006). Elle détermine la contribution d'un génotype donné aux générations suivantes. Nous avons testé si les gènes de floraison identifiés dans notre étude (*PHYC* et *MADS11*) étaient associés à la fitness. En particulier, il est intéressant de savoir si un génotype particulier donne un avantage aux individus qui le portent, en termes de survie de la population ou de contribution à la génération suivante, dans un contexte de variation environnementale (stress hydrique par exemple). Ce travail a fait l'objet de l'étude qui suit.

Etude de la valeur sélective de deux gènes candidats de la floraison chez le mil en condition de stress hydrique simulé

(Résultats préliminaires)

OBJECTIFS

Nous avons réalisé des essais sur une famille de 159 RILs qui ségrégent pour chacun des gènes candidats de floraison *PHYC* et MADS1, afin d'étudier la relation du génotype au niveau de ces gènes avec la date de floraison et des phénotypes associés à la fitness.

MATERIEL ET METHODES

Matériel végétal. Le matériel est composé d'une population de 159 lignées recombinantes (LR72). Ces lignées (F7) ont été obtenues de l'ICRISAT Hyderabad (Inde) et sont issues du croisement entre les individus de mil (81B)-P6 et (ICMP 451)-P8 (ICRISAT).

Génotypage. Une enzyme de restriction (PvII) a été utilisée pour reconnaître le site associé à un SNP (C/G) en position 5525 dans le gène *PHYC*. Le pattern de digestion permet de reconnaître les individus CC, CG et GG (Saïdou et al. 2009). Um marqueur de même type a été défini pour *MADS11* (Mariac et al. 2010). Nous avons utilisé ces deux marqueurs pour génotyper individuellement chacune des 159 lignées recombinantes.

Dispositif expérimental. L'essai mis en place comporte 4 parcelles contiguës (axe horizontal) avec un traitement hydrique spécifique par parcelle. L'essai a été implanté en plein champ (50 m x 50 m) à Sadoré, Niger. Pour chaque traitement, on a mis en place 3 blocs (A, B, C) disposés l'un à la suite de l'autre de façon contigüe (axe vertical). Pour chaque lignée, on a placé aléatoirement un poquet par bloc.les poquets ont été démarié à trois plantes après la levée. Les plantes ont été semées en ligne (sur billons), avec 0.7 m d'espacement entre deux plantes. Les billons étaient également espacés de 0.7 m. Une distance de 5 m a été laissée entre les différentes parcelles (traitements), et une bordure végétal (variété de mil cultivé) de 3 colonnes a été insérée des deux côtés de chaque parcelle. Le semis a été fait en fin de saison pluvieuse (15/09/2009), afin de contrôler l'apport d'eau en utilisant un plan d'irrigation défini. Dans le premier mois suivant le semis, toutes les parcelles ont reçues une irrigation homogène (en moyenne 2 x 20 mm par semaine). Pour le premier traitement (T1), les parcelles ont été irriguées jusqu'à maturité des plantes. Pour les autres traitements, l'irrigation a été poursuivie respectivement jusqu'à 73 jours après le semis (T2), 55 jours après le semis (T3) et 30 jours après le semis (T4). Les dates d'arrêt de l'irrigation correspondantes étaient donc respectivement le 27/11/2009 (T2), le 9/11/2009 (T3) et le 15/10/2009 (T4).

Notation phénotypique. Chaque plante individuelle a été notée au champ pour le nombre de jours entre le semis et la date de floraison femelle (FLO), la hauteur totale de la plante à la fin du cycle (HTP) et la longueur de l'épi principal (LEP). Après récolte, on a mesuré pour chaque individu le poids brut de la récolte par plante (épis et involucres avant battage, PBR), le diamètre de l'épi principal (DEP), le nombre d'épis par plante (NEP), et le poids de 100 graines (PCG). Le taux d'attaque des épis par les oiseaux ravageurs (TDO) et le niveau de remplissage des épis (NRE) ont été également notés, sur la base d'une appréciation visuelle. Pour le dégât des oiseaux (TDO), la note est une estimation en pourcentage de la proportion de graines perdues suite au dégât des ravageurs. Pour le remplissage des épis (NRE), une échelle discrète variant de 0 à 5 a été définie ; 0 représente un épi vide, et 5 représente un épi complètement rempli. Pour 50 individus choisis aléatoirement sur l'ensemble de l'essai, le poids total des graines récoltées par plante (PGR) a été mesuré, après battage de l'ensemble des épis de chaque plante respectivement. Cet échantillon a été utilisé pour évaluer la régression entre le poids brut récolté par plante (PBR) et le poids de graines correspondant, après battage (PGR). Nous avons ensuite utilisé les coefficients de cette droite de régression pour estimer le poids de graines récoltées pour chaque individu sur l'ensemble de l'essai. Le nombre de graines par plante (NGP) a été enfin estimé en divisant le poids total de graines récoltées par le poids de 100 graines de cette plante, soit NGP = 100 x (PGR / PCG).

Analyses statistiques. Nous avons effectué l'analyse statistique en deux étapes.

Premièrement, nous avons ajusté les données de chaque traitement séparément pour calculer le phénotype ajusté de chaque lignée au sein de chaque traitement. Le modèle linéaire généralisé (GLM) suivant a été utilisé, séparément pour chaque traitement, pour ajuster les données:

$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$,

Où y_{ij} est le phénotype de la lignée j au sein du bloc i, μ est la moyenne générale (constante) ; α_i est l'effet du bloc i ; β_j est l'effet de la lignée j ; et ε_i est l'effet résiduel aléatoire. La moyenne ajustée pour chaque lignée j a été extraite comme $y_j = \mu + \beta_j$.

Ensuite, nous avons utilisé les phénotypes ajustés pour effectuer l'analyse d'association entre le phénotype et les marqueurs aux gènes candidats *PHYC* et *MADS11*. Ce modèle a été utilisé pour ajuster les données des 4 traitements ensembles :

$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_i j + \varepsilon_{ijk}$, où y_{ijk} est la $k^{ième}$ l'observation phénotypique, α_i est l'effet du traitement i, β_j est l'effet de l'allèle j du gène candidat, γ_{ij} est l'effet de l'interaction entre le traitement i et l'allèle j, et ε_{ijk} est le résidu pour la $k^{ième}$ observation. Enfin, pour estimer le phénotype moyen par traitement, nous avons utilisé la forme réduite de ce même modèle, en enlevant du modèle l'effet du gène et l'interaction gène-traitement. Nous avons ainsi ajusté un modèle avec uniquement l'effet principal du traitement, et les coefficients obtenus avec ce modèle ont été utilisés pour calculer le phénotype moyen par traitement, défini comme $y_i = \mu + \alpha_i$. La significativité de chaque effet a été testée sur la base du test de Ficher. L'analyse statistique a été conduite sous R (R 2.7.2).

RESULTATS PRELIMINAIRES

Effet du stress hydrique sur le phénotype

Nous avons simulé des scénarios de stress hydrique se produisant avec une précocité différente (arrêt systématique de l'apport en eau à des stades différents de la culture). Dans cet essai, le stress le plus précoce est T4 (arrêt d'irrigation 30 jours après le semis), suivi de T3 (55 jours après le semis), et T2 (73 jours après le semis). L'irrigation a été faite jusqu'à maturité pour le traitement T1. Nous avons analysé l'effet de ce stress graduel sur le phénotype, en ajustant les observations des 4 traitements avec le modèle linéaire généralisé. Pour tous les phénotypes, l'effet du traitement a été significatif (test de Fischer, $P<0.05$), à l'exception de la longueur de l'épi qui est restée en moyenne identique d'un traitement à l'autre (P=0.72). Nous avons extrait du modèle la valeur moyenne de chaque phénotype, estimée pour les 159 lignées recombinantes en fonction du traitement (Figure 1). Le stress a induit une réduction significative de la croissance des plantes, d'autant plus prononcée que l'arrêt de l'irrigation est précoce (Figure 1). On observe cet effet pour la hauteur de la plante (P=1.19 x 10^{-36}) et pour le diamètre de l'épi (P=6.5 x 10^{-9}). Un retard générale de la floraison est aussi observé sous l'effet du stress (P=5.27 x 10^{-5}). Enfin, les composantes de rendement sont également en baisse: nombre d'épis par plante (P=1.36 x 10^{-71}), poids brut de la récolte par plante (P=4.8 x 10^{-55}), poids de 100 graines (P=8.78 x 10^{-7}), et nombre de graines estimé par plante (P=1.6 x 10^{-44}).

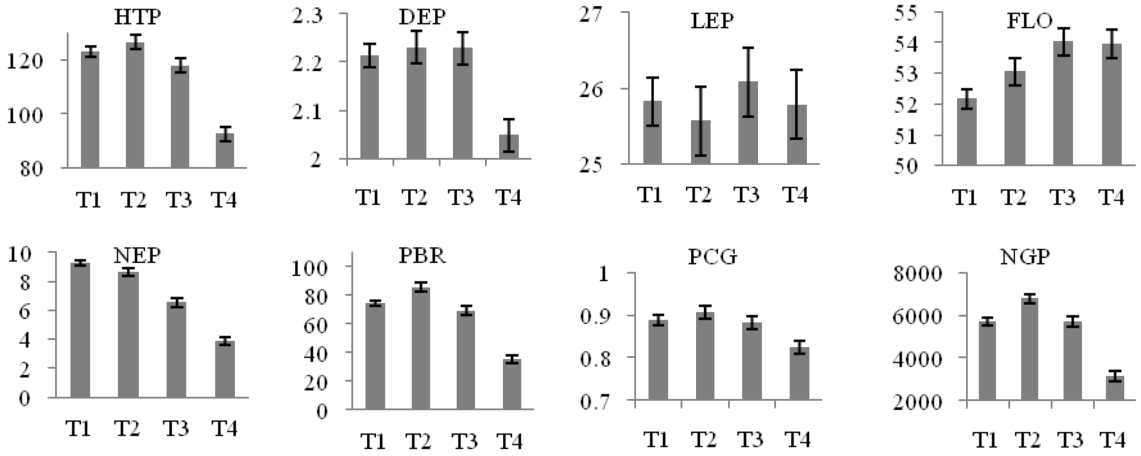


Figure 1. Effet du traitement sur le phénotype moyen des lignées. La valeur moyenne du phénotype, pour les 159 lignées recombinantes, est représentée pour chaque traitement. Ces valeurs ont été extraites du modèle linéaire généralisé, après ajustement des observations sur 4 traitements (T1 à T4). Les barres verticales indiquent l'erreur standard. Le stress le plus précoce est T4 (30 jours après le semis), suivi de T3 (55 jours après le semis) et T2 (73 jours après le semis). Huit caractères phénotypiques ont été considérés : hauteur de la plante (HTP), diamètre de l'épi (DEP), longueur de l'épi (LEP), floraison (FLO), nombre d'épis par plante (NEP), poids brut de la récolte par plante (PBR), poids de 100 graines (PCG), et nombre de graines estimé par plante (NGP). On observe une réduction de croissance (HTP, DEP) et une baisse de la valeur des composantes de rendement (NEP, PBR, PCG et NGP). La date de floraison est tardive lorsque les plantes sont stressées précocement.

Analyse de l'effet des gènes candidats sur la variation phénotypique

Nous avons testé l'effet sur le phénotype du génotype au niveau de deux gènes candidats de floraison (*PHYC* et *MADS11*). Huit traits ont été analysés. Le gène *PHYC* a montré un effet significatif sur la date de floraison et la morphologie de l'épi (Tableau 1). La taille de ces effets est respectivement estimée, à 1.2 ±0.66 jours sur la date de floraison, -1.73 ±0.68 cm sur la longueur de l'épi, et 0.17 ±0.05 cm sur le diamètre de l'épi. Le gène *MADS11* a également montré un effet significatif sur ces mêmes caractères, puis sur le nombre d'épis par plante (Tableau 1). Les effets de *MADS11* sont estimés respectivement à -0.47 ±0.77 jours sur la date de floraison, -0.92 ±0.8 cm sur la longueur de l'épi, 0.27 ±0.06 cm sur le diamètre de l'épi, et -0.82 ±0.52 sur le nombre d'épis par plante. En cohérence avec le modèle d'analyse phénotypique (sans effet gène), ces modèles avec *PHYC* et *MADS11* montrent aussi un effet significatif du traitement. Aucun effet d'interaction entre les gènes et le traitement n'est détecté par ces modèles.

Tableau 1. Association gènes candidats-phénotype

Phénotype	Effet	*PHYC*	*MADS11*
Date de floraison (FLO)	Gène	4.76 x 10^{-03}	1.99 x 10^{-02}
	Traitement	8.99 x 10^{-05}	9.10 x 10^{-04}
	Traitement x Gène	0.54	0.54
Poids de 100 graines (PCG)	Gène	0.28	0.99
	Traitement	1.55 x 10^{-06}	4.20 x 10^{-06}
	Traitement x Gène	0.93	0.8
Nombre de grains par plante (NGP)	Gène	0.91	0.12
	Traitement	2.22 x 10^{-40}	3.95 x 10^{-34}
	Traitement x Gène	0.57	0.85
Hauteur de la plante (HTP)	Gène	0.48	0.1
	Traitement	1.50 x 10^{-34}	2.53 x 10^{-32}
	Traitement x Gène	0.82	0.98
Diamètre de l'épi principal (DEP)	Gène	4.65 x 10^{-04}	1.58 x 10^{-04}
	Traitement	1.49 x 10^{-08}	9.57 x 10^{-07}
	Traitement x Gène	0.35	3.99 x 10^{-02}
Longueur de l'épi principal (LEP)	Gène	5.80 x 10^{-05}	4.56 x 10^{-02}
	Traitement	0.71	0.93
	Traitement x Gène	0.9	0.99
Poids brut de la récolte par plante (PBR)	Gène	0.81	0.14
	Traitement	1.63 x 10^{-51}	5.46 x 10^{-44}
	Traitement x Gène	0.73	0.97
Nombre d'épis par plante (NEP)	Gène	0.6	2.36 x 10^{-2}
	Traitement	1.11 x 10^{-62}	3.36 x 10^{-54}
	Traitement x Gène	0.99	0.34

L'analyse a été faite avec modèle linéaire généralisé. Elle porte sur les données phénotypiques des 159 lignées recombinantes, évaluées dans 4 différents traitements. Pour chaque lignée recombinante, la moyenne ajustée par traitement a été utilisé dans ce modèle. Le génotype au gène *PHYC* et au gène *MADS11* ont respectivement été associés au trait. On montre les tests de significativité de l'effet principal de chaque gène, du traitement, puis l'effet d'interaction entre chaque gène et le traitement. Les p-values du test pour la significativité de chaque effet sont indiquées (test de Fischer). Les valeurs significatives (P<0.05) sont surlignées en gris.

DISCUSSION PRELIMINAIRE ET PERSPECTIVES

Nous avons conduit un essai en conditions de stress hydrique afin d'étudier l'association de deux gènes candidats sur des traits phénotypiques impliqués dans la fitness. Les traitements appliqués ont conduit à un effet global sur ces traits, notamment sur les composantes de rendement (Figure 1). Cet effet a montré un gradient, suggérant une importance relative du stress selon le stade de croissance auquel il est appliqué. Les stress les plus précoces ont entraînés un retard dans la floraison. Cet effet pourrait être dû au déficit de croissance induit (comme on l'observe sur les variables morphologiques), qui n'aurait pas permis la mise en place des organes floraux assez tôt. On observe aussi, sous stress, un nombre d'épis plus faible, dénotant potentiellement une stratégie de la plante pour assurer le maintien et la croissance d’un minimum d’organes, les ressources disponibles ne permettant pas une croissance optimale. La variation du poids total récolté par plante semble s’expliquer à la fois par un nombre limité de graines produites et un remplissage de graines sensiblement limité. Le nombre de graines est vraisemblablement lié à un nombre d’épis plus bas, mais il pourrait aussi dépendre du taux de remplissage des épis.

Nous avons analysé l’association entre ces caractères et deux gènes candidats. Les deux gènes sont significativement associés à la date de floraison, mais on ne trouve pas d’interaction significative entre ces gènes et le statut du stress hydrique appliqué. Cela semble indiqué que dans les conditions testées, ces gènes n’interviennent pas significativement dans les changements phénotypiques observés d’un traitement à un autre. Il reste que la variation à ces gènes induit, indépendamment des conditions de stress, un changement faible mais significatif sur la date de floraison et la morphologie des épis. *MADS11* est associé également à la variation du nombre d’épis par plante. L’absence d’interaction signifie autrement dit que les effets des gènes sont stables, et la différence entre les génotypes demeure similaire dans différentes conditions. Dans ce cas, les lignées porteuses des allèles précoces auront tendance à fleurir plus tôt quelque soit le stress. On aurait pu s’attendre, dans ce cas, que ces génotypes précoces bénéficient d’un avantage pour les composantes de rendement arrivant en fin de cycle, en présence notamment de stress appliqué tardivement. Car dans ce cas, ces génotypes précoces ont plus de chance d’avoir fini leur cycle ou d’avoir atteint un stade de remplissage et de maturation avancé, leur permettant de ne pas se confronter à l’effet du stress (évitement de stress). Cet effet aurait pu s’observer sur des composantes comme le remplissage des graines (poids de 100 graines), mais il n’en est rien. La précocité de floraison chez *MADS11* est associée à un effet pléiotropique négatif sur le nombre d’épis par plante. Il est donc possible que ce type d’effet sur les composantes de rendement initiales contrebalance un éventuel avantage en fin de cycle. Nos études précédentes ont généralement montré un effet plus fort de *PHYC* et *MADS11* sur la date de floraison (>4 ou 5 jours). Sur ce croisement spécifique, l’effet s’est révélé très faible (~0.5 à 1 jour) ; donc dans ce fond génétique, ces deux gènes ont des effets très faibles. Il parait difficile avec de tel effets d’espérer détecter un effet sur les composantes des rendements. Une étude sur des variétés nigériennes où l’effet estimé de *PHYC* et *MADS11* est plus élevé a pu détecter des effets associés de ces gènes avec les composantes de rendement (Y Vigouroux et C Mariac, *com. pers*.). La présente population des RILs, à cause du faible effet du gène sur les phénotypes, n’était sans doute pas la meilleure pour réaliser cette expérience.

Le modèle linéaire général a été utilisé pour cette analyse. Ce modèle est plus robuste que le modèle linéaire standard à la présence de manquants dans l’analyse, par exemple. La façon de prendre en compte l’hétérogénéité environnementale dans ce cas a été d’utiliser le facteur bloc comme effet fixe dans l’analyse. L’utilisation d’un effet bloc comme facteur qualitatif peut être préjudiciable en cas de gradients insoupçonnés. Aussi, cette approche ne permet pas de prendre en compte la covariance entre individus spatialement proches. Vu le dispositif utilisé (parcelles placées horizontalement avec les blocs en vertical), il est possible d'utiliser la structure spatiale des points d'observations pour mieux contrôler l'hétérogénéité et la covariance entre observations. Nous disposons de la position spatiale de chaque plante dans l'essai. Il sera donc possible d'utiliser le modèle linéaire mixte pour affiner ces analyses, avec la performance qu'il offre en termes de modélisation de l'hétérogénéité et de la variance-covariance des données. Cela peut permettre d'avoir plus de puissance pour éventuellement détecter des effets faibles cachés par l'hétérogénéité spatiale ou la structure des données. Cela peut aussi permettre une estimation moins biaisée de l'effet des gènes sur le phénotype. Par ailleurs, le taux de couverture en graines des épis et les attaques de ravageurs ont été notées sur cet essai. Il conviendrait de prendre en compte ces paramètres pour d'une part discuter davantage le lien entre les composantes de rendement, et d'autre part ajuster le poids des récoltes et le nombre de graines par plante en tenant compte des dégâts estimés. Ces analyses seront effectuées pour compléter le présent travail.

Chapitre 3. Développement de méthodes de génétique d'association pour l'étude des interactions génotype -environnement dans des populations structurées

Contexte et intérêt méthodologique de l'étude

Les études d'association avec des populations structurées ont été introduites récemment chez les plantes (Thornsberry *et al.* 2001). Très vite, cette approche a fait l'objet d'examens méthodologiques, portant notamment sur le contrôle du taux de faux positifs et la puissance de détection de gènes d'intérêt. Le modèle linéaire mixte (Yu *et al.* 2006) prenant en compte la subdivision des populations et l'apparentement entre individus a montré dans ce contexte un niveau de performance généralement supérieur. Cependant, les intérêts ont été principalement portés sur l'analyse des effets principaux du génotype, sans prendre en compte les interactions éventuelles entre gènes ou les interactions de ces gènes avec l'environnement. Jannink (2006) a proposé un modèle linéaire mixte permettant de détecter des interactions entre gènes (épistasie) dans le cadre des études d'associations basées sur des populations. Cette étude a utilisé l'apparentement entre les individus comme variable proximale pour détecter une interaction entre un marqueur analysé et des allèles plus ou moins fixés dans des familles particulières[1]. Cette approche améliore la prédiction de l'effet des gènes dans des fonds génétiques différents (Jannink 2007).

Les effets d'interaction entre le génotype et l'environnement (GxE) sont une composante importante de la variation des traits phénotypiques. Ces effets peuvent confondre le résultat des études d'association. De même, la connaissance des interactions existantes est importante pour pouvoir prédire l'effet d'un gène dans une gamme plus large de fonds génétiques et/ou d'environnements. Dans le contexte de populations structurées, les interactions GxE peuvent se décliner en interaction entre le fond génétique des populations et l'environnement. Dans ce cas, la réponse d'un individu à la variation environnementale serait différentielle selon son appartenance à une population.

[1] On parle de variable proximale lorsque la variable utilisée n'est pas directement (ou n'est pas forcément) responsable de l'effet analysé, mais que cette variable est au moins corrélée à la variable causale et permet donc de détecter l'effet, en l'absence de la vraie variable causale. Ici par exemple, le lien d'apparentement est utilisé pour rendre compte de l'effet des allèles présents dans les familles. Des individus apparentés ont une probabilité forte de porter des allèles communs.

Nous avons proposé d'intégrer différents dans le modèle linéaire mixte des termes d'interactions, y compris des interactions entre gène et environnement et des interactions entre le fond génétique (mesurée par le coefficient d'ascendance ou *ancestry*) et l'environnement[2].

Cette étude comporte quatre points principaux : i) la simulation de jeux de données populationnels contenant divers effets d'interaction; ii) l'évaluation de la puissance du modèle linéaire mixte à détecter les effets simulés; iii) l'évaluation de la performance des critères d'information pour identifier les termes adéquats du modèle linéaire selon le jeu de données; et iv) l'illustration des méthodes proposées sur des jeux de données réels. Nous avons utilisés pour cette étude un panel de 90 lignées de mil et un panel trois fois plus large, de 277 lignées de maïs. L'étude a permis de discuter, autour des 4 points précités, la perspective d'utilisation du modèle linéaire mixte pour étudier les interactions dans le cadre des études d'association chez les plantes.

En l'occurrence, nous avons montré l'impact de différents paramètres sur la puissance du modèle. L'ajustement statistique de données avec divers effets d'interaction peut nécessiter des modèles statistiques plus complexes et cela pose le problème de choix de modèles statistiques. Nous avons discuté la performance des critères d'information pour le choix de modèle sous REML. Enfin, nous avons illustré l'application de notre démarche sur avec deux jeux de données (maïs et mil). Sur le maïs, nous avons mis en évidence une interaction entre le locus Vgt1 et l'environnement, et aussi des interactions entre le fond génétique des populations et l'environnement.

[2] Cette étude a fait l'objet d'un article soumis. Cet article est présenté à partir de la page suivante.

Association studies including genotype by environment interactions: prospects and limits.

AUTHORS:

Abdoul-Aziz Saïdou,[1, 2, 3, 4, *], Anne-Céline Thuillet, [1] , Marie Couderc, [1] , Cédric Mariac,[1, 2], and Yves Vigouroux[1,*]

[1] Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/ Montpellier SupAgro, BP64501, 34394 Montpellier, France ;

[2] Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/ Montpellier SupAgro, BP11416, Niamey, Niger ;

[3] Abdou Moumouni University, BP 11040, Niamey, Niger; and

[4] Montpellier SupAgro, 2, place Pierre Viala 34060 Montpellier, France.

[*] Corresponding authors:

A Saïdou & Y Vigouroux
Institut de Recherche pour le Développement,
911, avenue Agropolis,
34394 Montpellier, France.
Phone: 33 (0) 467416165
Fax: 33 (0) 467416222
Email: saidou_aa@yahoo.fr & yves.vigouroux@ird.fr

RUNNING HEAD: Genotype by environment interactions in association study

KEY WORDS: association study, G x E, power simulation, model selection, REML, *PHYC*, *Vgt1*

ABSTRACT

Association mapping studies are increasingly used in plant genetics to identify polymorphisms associated with phenotype. These studies lead great promise for the understanding of the genetic basis of quantitative trait variation. The major part of association mapping studies assesses the main effects of genetic polymorphisms while accounting for background effects. Mixed linear model is one of the most efficient frameworks used for such studies. The extension of this framework to deal with genotype by environment interaction or interaction between genetic factors will be useful to complete the dissection of phenotypic variation. In this study, we proposed a methodological prospect of mixed linear model to analyze interaction effects using association panels. First, we modelled effects of interaction involving environment, genetic polymorphism and/or population background (ancestry). We simulated data based on these models to assess the power of linear mixed model to detect interaction effects. The simulation was based on two respective association panels of 90 inbreds (pearl millet) and 277 inbreds (maize). Variation of power was assessed depending on effect size, allele frequency, heritability and sample size. Second, the performance of information criteria to select the best statistical model under REML was assessed for association mapping data involving interactions. This framework proved powerful for simulated traits with relatively high heritability and/or shaped by common alleles with relatively large effects. Limitations of the framework were discussed regarding for instance rare polymorphisms or polymorphisms with small effects. Information criteria showed different levels of performance for model selection. Depending on data characteristics, these criteria could help defining the most adequate model fitting the data . The importance of larger sample size to improve the power or to facilitate model selection was highlighted by the study. Applications on real data revealed significant interactions in maize between *Vgt1* gene and environment, and between population ancestry and environment.

INTRODUCTION

Deciphering the genetic basis of quantitative trait variation is a great challenge of biology. Linkage mapping and association mapping are two complementary methods that are commonly used to study the relationship between genotype and phenotype. Linkage mapping (or *family mapping* as suggested by Myles *et al.* 2010) is in general based on the progeny of controlled crosses. Association mapping (or *population mapping*) takes benefit of large populations which have inter-crossed for a large number of generations, allowing a high number of recombination events to occur. This strong historical recombination between loci generally leads to a very fine scale for genotype-phenotype association analysis (Nordborg and Weigel 2008). Association mapping is especially powerful for common alleles and for moderate-to-large effects (Saïdou *et al.* 2009, Nordborg and Weigel 2008). A main pitfall of this method is that the genetic background of the populations could produce confounding effects which bias the statistical analysis and inflates false positive rate (Pritchard *et al.* 2000a). Statistical models accounting for the genetic background were thus proposed to limit spurious associations and enhance power to detect true associations (Thornsberry *et al.* 2001, Yu *et al.* 2006). As the evolutive history and the genetic background of the populations used in these studies are often *a priori* unknown, different methods of inference have been developed to provide background matrixes for use in statistical models (Pritchard *et al.* 2000b, Gao *et al.* 2007). These methods rely on multilocus molecular data, for instance randomly distributed background markers such as AFLPs, SSRs or SNPs. A first approach consists in Bayesian algorithms that model the data based on assumed genetic models in order to infer clusters and to assign individuals to these clusters (e.g. Pritchard *et al.* 2000b). Linkage disequilibrium between markers, admixture between populations, property of the background markers (dominant/null alleles), low level of divergence, and/or limited sample size could be handled through these Bayesian analyses (Falush *et al.* 2003, Falush *et al.* 2007, Hubisz et al. 2009). A method that handles selfing or inbreeding has also been proposed (Gao *et al.* 2007). The resulting population matrixes set an ancestry coefficient (or membership probability) for each individual into each considered cluster. As well as an adequate set of background markers is used, the use of these matrixes into adequate statistical models lead to a good fit with phenotypic data (Yu *et al.* 2009). This proved efficient for limiting false positive rate (e.g. Thornsberry *et al.* 2001), but the algorithms are generally seen as time-consuming, especially for large datasets. A second approach that reduces running time consists in the use of traditional statistical methods rather than assuming any genetic model. For instance, principal components analysis (Patterson *et al.* 2006) or nonmetric multidimensional scaling

(Zhu and Yu 2009) have been suggested for such purpose. Another level of relatedness that often occurs in population panels is kinship relationship between individuals (Yu *et al.* 2006). Methodologies were also developed to infer this relatedness (Loiselle *et al.* 1995, Hardy and Vekemans, 2002, Stich *et al.* 2008). The resulting matrixes are used in association analyses to set the variance-covariance structure of the model and correct for polygenic background effect (Yu *et al.* 2006). All these methodologies contribute to the limitation of false positives to different extents depending on the data (e.g. Stich *et al.* 2008, Saïdou *et al.* 2009). False positive control remains most difficult when the trait of interest covaries highly with population structure (Remington *et al.* 2001, Saïdou *et al.* 2009). Such traits are generally traits with particular interest, for instance, the control of local adaptation of populations. In the other hand, the correction of background confounding effect might be too conservative when the polymorphisms involved in trait expression covary with population structure. This prevents the detection of association between the traits and these polymorphisms. These latter issues constitute methodological limits that still have to be solved. Despite this, association studies lead great promise to accelerate the study of genotype-phenotype relationships, especially with the continuously developing high throughput genomic technologies, and with the advances in statistical methodology and computational resources (Zhang *et al.* 2010a). In particular, genome-wide association studies (GWAS) are challenging to take the maximum benefits from these advances to exhaustively identify polymorphisms linked to the traits of interest (Zhang *et al.* 2010a, Myles *et al.* 2009). GWAS already proved very useful in plants (Atwell *et al.* 2010, Brachi *et al.* 2010).

The number of association studies performed in plant is currently increasing (see reviews of Zhu *et al.* 2008, Hall *et al.* 2010, Rafalski 2010). However, most of these studies focus on the analysis of the main effect of molecular polymorphism on the phenotype. To date, only few association mapping studies reported tests of interaction effects (Stracke *et al.* 2009, Zhang *et al.* 2010b, Li *et al.* 2010). Interaction effects include genotype by environment interactions (G x E) and epistatic interactions between genetic factors themselves. Genotype by environment interaction occurs when there is variation among genotypes in the rank order or relative magnitude of effects in different environments (Falconer and Mackay, 1996). Epistasis could be seen as the statistical deviation from the additive combination of two loci in their effects on a phenotype (Phillips, 2008). This could be detected as interaction between different loci. In association mapping populations, the distribution of certain loci is correlated to population background, due for instance to the predominance of particular alleles in specific populations or families. Thus, interaction between a given marker and background could be indicative of epistasis involving alleles linked to this background (Jannink, 2007). The relative part of phenotypic variation explained either by G x E or by epistatic interactions varies with the trait and the biological species. In a study of *Drosophila melanogaster* populations, about 50% of phenotypic variation in adult olfactory behavior has been attributed to G x E (Sambandan *et al.* 2008). In maize Nested Association Mapping populations (McMullen *et al.* 2009), the part of variance explained by G x E or epistatic interactions in flowering traits was low, compared to the genetic variance due to QTL main effects (Buckler *et al.* 2009). But studies in rice (Uwatoko *et al.* 2008) and Arabidopsis (El-Lithy *et al.* 2006, Caicedo et al. 2004) highlighted that epistatic effects importantly contribute to shape flowering time. Furthermore, exhaustive association studies reveal that, in general, QTLs explain only a part of trait heritability, even when a large number of loci is considered (Kover *et al.* 2009; Buckler et al. 2009). This is known in human genetics as the problem of missing heritability (Manolio *et al.* 2009). G x E and epistatic interactions are part of the factors that are expected to explain the missing heritability (Hall et *al.* 2009, Myles *et al.* 2010). The extension of the plant association mapping framework to deal with G x E or epistatic interactions is thus interesting in order to refine the dissection of quantitative traits.

The mixed linear model (MLM) framework (Yu *et al.* 2006) is the most commonly used in plant association mapping. Two methods exist for the estimation of variance parameters in MLM: the method based on the maximum likelihood (ML) and the method based on restricted (or *residual*) maximum likelihood (REML). REML gives unbiased estimates of the variance components of mixed linear model and might be preferred to ML, especially when interest lies on the variance components (Verbeke and Molenberghs, 2000). Both ML and REML mixed linear models are increasingly used in association mapping to assess the main effect of genetic factors on phenotype. A first methodological examination of the MLM framework for the purpose of interaction testing using structured populations was recently proposed (Jannink 2007, Jannink 2008). A model was specified to test for interactions between family background (pedigree) and QTL marker. Data were simulated with a background statistically linked to a set of loci that interact with the QTL. These data were analyzed to assess the power of MLM given different parameters, notably genetic variance explained by the epistatic interaction, sample size, and linkage disequilibrium between the marker and the causal QTL. With such patterns of data, epistatic interactions involving a QTL could be assessed through the test of interaction between the QTL and the background, even when the loci interacting with the QTL are not identified individually (Jannink 2007). However, further methodological examination of MLM is needed to prospect other patterns of interactions, for instance genotype by environment interactions or interaction between gene and ancestry. The latter could occur when a set of loci correlated to population structure are involved in epistatic interaction with the gene.

On the other hand, the fitting of models including multiple terms or multiple levels of interactions raises the problem of parsimony and model simplification. Parsimony implies the choice of a minimum number of parameters without significant loss of fit (Crawley 2007). The selection of such a *minimum adequate model* is not always obvious for association mapping data generated from complex designs. This is in general the case of data designed for the analysis of genotype by environment interaction. These data consist of measurements in different environments, with different factors describing the structure of the design. In such data, the full model with higher order interactions contains a too large number of parameters, and the number of competing reduced models is also large. To perform model simplification, a rigorous method of model comparison is necessary. There exist different information criteria traditionally used for model comparison under linear models or under mixed linear models based on ML, namely the log likelihood, the Akaike Information Criterion (AIC, Akaike 1974), the Corrected Akaike Information Criterion (AICC, Hurvich and Tsai 1989), the Bayesian Information Criterion (BIC, Schwarz, 1978) and the Consistent Akaike Information Criterion (CAIC, Bozdogan 1987). Adjusted versions of the squared correlation coefficient R^2 (R^2_{adj}; Vonesh *et al.* 1996) has also been used for model selection. Note that R^2 could be calculated using marginal or conditional formulation (Schabenberger, 2004). The suitability of these criteria for model selection under REML is not yet a clearly solved issue (Gurka 2006a, Leng 2008). For instance, the use of log likelihood ratio test (LRT) which is commonplace under ML is no longer appropriate under REML when the models to be compared have different fixed effects parameterization (Gilmour *et al.* 2006). A new criterion called the Residual Information Criterion (RIC) was derived based on REML (Shi and Tsai, 2002), but the derivation of this criterion was considered as erroneous and the criterion is seen as systematically selecting overfitted models (Leng, 2008). Gurka (2006a) stated it is not apparent as to why information criteria such as the AIC or BIC are seen commonly as inappropriate for model comparison under REML. A simulation study was proposed to examine the performance of AIC, AICC, BIC and CAIC in model selection, when interest lies on the choice of mean structure (set by *p* fixed parameters), variance structure (set by *k* random parameters), or both (Gurka, 2006a). This study provided numerical evidence for the fact these information criteria might be adequate for model selection under REML. However, the performance of the criteria varied with data characteristics, notably total variance of the trait, covariance structure and sample size. A similar study (Wang 2007) examined the performance of AIC, BIC and predictive criteria (R^2 and R^2_{adj}). The conclusions were consistent with those of Gurka (2006a). Note the existence of variant formulations for these information criteria (Gurka, 2006a), with respect notably to the way of defining the number of parameters or the sample size (number of individuals in the sample *m* versus total number of observations *N*). These studies (Gurka, 2006a; Wang 2007) were based on basic simulation models. So the feature with more specific or more complex patterns of data has yet to be assessed (Gurka, 2006a).

The aim of this study is to investigate the use of mixed linear model in association mapping framework to deal with G x E or epistatic interactions involving population background. First, we specified models with different terms of interaction, including gene by environment interaction, interaction between gene and ancestry, and three way interaction between ancestry, gene and environment. Second, we conceptualized such patterns of interaction in the context of association mapping populations, and we simulated the data given different parameters (allele frequency, effect size and heritability) to assess the power of mixed linear model. The method was implemented on two distinct association panels (pearl millet and maize). Model components were estimated based on REML (ASReml-R, version 2.0/32). Next, we performed numerical simulations to assess the performance of information criteria (AIC, AICC, BIC, CAIC and R^2_{adj}) to select the best model in the case of competing models with different fixed effects. Finally, a strategy for interaction analysis based on the considered mixed models and on the assessed methods of model selection was illustrated using two real datasets (pearl millet and maize). Results of association with flowering phenotype are presented and discussed. More globally, perspectives about interaction analysis into the association mapping framework are discussed in light of the presented work.

MATERIAL AND METHOD

Linear mixed model specification

The commonly used mixed model in plant association mapping studies is:

$y=X\beta+Qv+S\alpha+Zu+e$,

where y is the vector of phenotype, β is a fixed effect other than SNP or population structure, α is the vector of a given SNP fixed effect, v is the vector of population structure fixed effects, u is the vector of polygenic background effects, and e is the residual error vector (Yu et al. 2006). Q is the population ancestry matrix. X, S, Z were 0/1 incidence matrices relating y to

β, α and u vectors respectively. The variance of the random effect u is expected to be Var(u)=K V, where K is the kinship matrix and V the variance. In the rest of this paper, we will use a simplified style of notation. So the previous model could be rewritten as follows:

$Y = \mu + E + Q + S + K + e$; (1)

Y is the phenotypic trait, μ is the intercept, Q is the fixed effect of population structure, S is the fixed effect of a SNP (or any gene polymorphism), E is an additional fixed effect (e.g. environmental effect). Q is commonly set by matrixes of population membership (e.g. ancestry matrixes or principal components). K is the polygenic background random effect and e is the random residual of the model. K is set by a matrix of kinship relationship between individuals.

In this canonical form, linear mixed model was mainly used so far to assess the main effect of genetic polymorphism (and other covariates) on the phenotype. We proposed a first extension of this model to fit gene by environment interaction (S x E). A term for S x E could be added in the model as follows:

$Y = \mu + E + Q + S + S \times E + K + e$; (2)

If the environmental variable (E) is set as random effect, the S x E term has to be set as random too. But we consider in this study the case where both S and E are fixed effects, so the interaction could be set as fixed effect and will contribute to the mean structure.

Next, we proposed a full extension of the model to fit two and three way interactions between factors:

$Y = \mu + E + Q + S + Q \times E + S \times E + Q \times S + Q \times S \times E + K + e$; (3)

Q x E is the effect of interaction between ancestry and environment; Q x S is the interaction between SNP and ancestry; and Q x S x E is the three way interaction between ancestry, SNP and environment. All interactions were considered as fixed effects.

Basic scheme for the simulation of association mapping data

A basic modelling of genotype-phenotype association consists in the simulation of datasets characterized by the main effect of a single locus on the trait. Such effect could be incremented on real phenotypic scores from a structured panel, so that the generated phenotype is linked to real genetic background information. With this scheme, there is no longer need to simulate background matrixes. So let’s consider a panel of n individuals. We note p_i the initial (real) phenotype of individual i. Considering a binary causative polymorphism (for instance presence/absence of a SNP), we can attribute to each individual i a simulated phenotype y_i, as follows:

$yi = p_i + S_i\ r\ \sigma_G + \varepsilon_i$ (Simulation scheme 1)

where S is a random variable with possible values “1” and “0”, standing respectively for the presence or the absence of the causative allele; σ_G is the standard deviation of the initial trait across the whole panel; r is the genetic effect ratio (or *effect ratio*, i.e. a numeric variable that modulates the size of the effect as a function of σ_G); and ε_i is a random variable that adds noise to the trait. The random variable S follows a Bernoulli distribution with P(S=1)= q, q being the expected frequency of the causative allele. The random variable ε follows a normal distribution with a mean of 0 and a variance σ_ε^2. This variance σ_ε^2 is set to satisfy a given trait heritability $h^2 = \sigma_G^2/ (\sigma_G^2 + \sigma_\varepsilon^2)$. Note that the simulated SNP effect is independent of population structure and kinship relationship between individuals.

Hereafter, this basic simulation scheme was complexified in order to generate data patterns with genotype by environment interaction and/or background interaction.

Simulation of gene by environment interaction

A phenotypic variation caused by a single SNP was simulated into two virtual environments E_1 and E_2 (Fig. 1, A). The trait simulated in E_1 follows the basic simulation scheme described in the previous paragraph. A new term was added in the simulation model for E_2, to make the effect of the SNP vary between environments proportionally to a coefficient λ. This simulation model was specified as follows:

In E1, $yi = p_i + S_i\ r\ \sigma_G + \varepsilon_i$

(Simulation scheme 2)

In E2, $y_i = p_i + S_i\ r\ \sigma_G + (S_i – q)\ \lambda\ r\ \sigma_G + \varepsilon_i$

λ is a numerical variable. The frequency of the causative allele (q) is taken into account to obtain an equal mean phenotype (y) between the two environments E_1 and E_2. In other words, this model simulates a null expected main effect of environment, but generates specifically a gene by environment interaction.

The simulation was performed based on two inbred panels described below (pearl millet and maize inbred). Field experiment-based phenotypic scores (flowering time) from each respective panel were used to set the initial phenotypic values p_i. The effect ratio r varied across iterations from 0 to 1.5 (18 values); λ varied from 0.05 to 1 (4 values); q was set to 0.05, 0.25 and 0.5 respectively; and h^2 was set to 0.25 and 0.75 respectively. The parameter σ_G was fixed at the value of the standard deviation of the initial trait in each panel (σ_G= 6.83 days for pearl millet and σ_G= 8.72 days for maize). For each combination of parameters, thousand samples were simulated for pearl millet and a hundred for maize. The lower number of iterations for maize allowed simulation to be performed in a reasonable timescale. So in this step, around 12 millions individual genotype-phenotype datasets were simulated for maize and around 40 millions for pearl millet. A summary of simulation parameters is given (Table 1).

Simulation of interaction between population background, gene and environment

We considered two loci in a structured panel of n individuals: a SNP randomly distributed with a frequency expectation q and a background marker (Bk) linked to the ancestry in a population P_0. P_0 is one of the populations constituting the panel. The SNP and the background marker were not linked. Defining two environments E_1 and E_2, we modelled the effect of the SNP on each individual phenotype so that this effect increases in the second environment only for individuals carrying simultaneously the causative allele at the SNP and the background marker Bk (Fig. 1B). For each run, Bk was attributed to an individual if the ancestry coefficient of this individual in P_0 (ranging from 0 to 1) is superior to a value randomly taken from a uniform distribution with the range 0-1. Given this approach, the expected frequency of the marker Bk in the sample was q_0, the ancestry coefficient averaged from all individuals. Also, into each iteration, the probability of an individual to have the marker equals the ancestry of this individual in P_0. The final phenotype was simulated using the following model:

In E1, $yi = p_i + S_i\, r\, \sigma_G + \varepsilon_i$ (Simulation scheme 3)

In E2, $y_i = p_i + S_i\, r\, \sigma_G + Bk\, (S_i - q)\, \lambda\, r\, \sigma_G + \varepsilon_i$

The current simulation scheme is therefore basically identical to the previous (simulation scheme 2), except the presence of Bk that makes the effect variation dependent on ancestry. Bk was set in the model as a binary numerical variable scoring the presence or the absence of the background-dependent marker (1 or 0, respectively). The parameter r varied from 0 to 1.5 (18 values) and λ varied from 0.05 to 2 (9 values); q was set respectively at 0.05, 0.25 and 0.5; h^2 was set at 0.25 and 0.75 respectively. For each combination of parameters, thousand samples were generated for pearl millet panel and hundred samples for maize, as in simulation scheme 2. Among the populations constituting each panel, the population with the highest average ancestry (q_0=0.28 for pearl millet and q_0=0.49 for maize, respectively) was defined as P_0, to maximize sample size. We developed scripts with the software R (version 2.7.2) to implement all the described simulation schemes.

Validation of the simulation step

A regression was performed to assess the link between expected effects and effect estimates actually obtained after simulation. Effect estimates were computed using the procedure of effect prediction based on the fitted mixed model (Gilmour et al. 2006). To eliminate stochastic deviations, one hundred iterations were performed and averaged for each considered combination of parameters. For each effect, the regression was assessed using 50 combinations of parameters (10 values of the parameter r and 5 values of λ*;* q was set at 0.5 and h^2 was set at 0.75). All regression lines followed approximately the diagonal (y = x), and the squared correlation coefficient approached R^2=1 (Table S1). So the simulated data were consistent with the expectations, and this validated the model implemented with R.

Estimation of model components and assessment of power

The data with gene by environment interaction (simulation scheme 2) were fitted using the adequate mixed model (2). The model with two and three way interactions (3) was used to fit the data including the corresponding interactions (simulation scheme 3). The background was set for maize and pearl millet respectively using the real data-based kinship and ancestry matrixes described hereafter. Model components were estimated using the REML method (ASReml-R package, version 2.0/32). A significance test of fixed effects was performed using the incremental Wald test implemented in the ASReml package (Gilmour et al. 2006). The power of each model to detect a given simulated effect was estimated for each parameter combination as the proportion of significant tests over the total number of carried out iterations ($p<0.05$).

Assessment of model selection criteria

The information criteria AIC, AICC, CAIC, BIC and the predictive criteria R^2_{adj} were assessed. Two variant formulations of each criterion (but AIC) were considered (Table 2). Datasets were generated based on the three respective simulation schemes described above: (i) the pattern with no interaction (simulation scheme 1), (ii) gene by environment interaction (simulation scheme 2) and (iii) two to three way interactions (simulation scheme 3). We fit each simulated dataset, whatever its pattern, by a set of competing models. For simplicity, this set of models was composed only of the 3 mixed linear models described: the model without interaction (1) , the model with gene by environment interaction (2) and the model with two and three way interactions (3) . These models were fitted under REML and presented different mean structures. For each single dataset, model selection criteria were used alternately to compare the competing models. The joint process of data simulation and model selection was iterated 1000 times for each combination of parameters. We then assess, for each criterion, the frequency by which each competing model was selected. For this analysis, the parameters h^2, q and λ did not vary and were fixed respectively at 0.75, 0.5 and 1. Only the effect ratio r varied from 0.1 to 1.5 (7 values) to set a gradual effect size. This analysis was implemented on both maize and pearl millet panels.

The larger sample (maize) was also used to assess properly the impact of sample size on model selection in the particular case of three way interaction (simulation scheme 3). The sample size n was set at six equally spaced values between n=90 and n=240 (90, 120, 150, 180, 210 and 240). For each value, 10 subsets were randomly sampled from the whole sample of 277 lines. Simulation was incremented on these subsets to assess the variation of criteria performance with sample size (r was set to 1 for this step).

The maize and the pearl millet panels

Independent field experiments were available on two respective panels (maize and pearl millet). Maize data originated from *Cornell university* (Ithaca, New York). Pearl millet data were from the *Institut de Recherche pour le Développement* (IRD, France and Niger).

For pearl millet, the number of days from sowing to female flowering stage was scored on the panel of 90 inbred lines. Nine field trials were carried out at different dates of sowing through the rainy season, over 4 years (2005 to 2008). For this dataset, we used the term *trial* to refer to the replicate of the field experiment at a given date of sowing. All trials were conducted at the same location (Sadoré, Niger). Three of these trials (2005-2006) and the experimental design are exhaustively described in Saïdou *et al*. (2009). The six additional trials (2007-2008) were conducted with the same experimental design, to allow a study of the trait over the whole period of sowing in Niger. In 2007, two dates of sowing were considered (16 June and 9 July), and 4 dates were considered in 2008 (26 June and 1, 7, 18 July). The averaged inbred flowering score in each trial was used in this study. The set of background markers for this panel (27 SSRs, 306 AFLPs) is described in Mariac *et al*. (2006) and Alline *et al*. (2007). Population structure was previously analyzed by Saïdou *et al*. (2009) and we got from that study the ancestry matrixes on the set of 7 pearl millet populations inferred using STRUCTURE algorithm (Pritchard *et al*. 2000b). But we recomputed the kinship matrix using the joint set of AFLP and SSR markers (Supplementary file 1). As candidate loci, we used 7 SNPs with minor allele frequency q > 5% at *PHYC* locus (sequence described in Saïdou *et al*. 2009).

The maize panel contains 277 inbred lines. Flowering phenotype (days to silk) was recorded in 8 locations across the USA. Hereafter, we used the term *trial* to refer to the maize experiment in a given location. In this study, only 7 trials were used, due to missing data in the eighth location. For each inbred, best linear unbiased predictor (BLUP) from each trial was used. Ancestry estimates were provided from previous study (Flint-Garcia *et al.* 2005). Three populations were considered: *NonStiffStalk* (NS), *StiffStalk* (SS) and *Tropical* (TS) populations. Kinship coefficients were calculated using EMMA method (Kang *et al.* 2008) based on a set of background markers previously described (Flint-Garcia *et al.* 2005). Genotype score for a MITE insertion at *Vgt1* locus was available from previous study (Buckler *et al.* 2009). One inbred was dropped from the sample due to data missingness. So the final sample size used in all part of this study was actually 276.

The maize kinship matrix was positive definite, but the pearl millet matrix was not. We used an *ad hoc* procedure (Saïdou *et al.* 2009) to make a transformed matrix which is positive definite and highly correlated to the original matrix (Supplementary file 1).

Statistical analyses

A common set of 5 competing models were defined to fit each of the maize and pearl millet datasets (Table 3). All the five models are nested and differ from each other by one or more terms. For each panel, one efficient criterion (AICC) and one consistent criterion (BIC) were used to make a comparative selection of model. Wald test was performed to assess the significance of each fixed effect retained in the models selected by either criterion. The Wald test was incremental and adjusted each term for those above (Gilmour *et al.* 2006). To be more conservative, we set systematically main effects and interactions involving environment or population background prior to effects involving candidate genes. Prior to statistical analyses, maize data were transformed using square root Box-Cox transformation (power of transformation 0.5) to minimize the departure of these data from model assumptions. A scaled variant of Box-Cox transformation adapted for mixed models (Gurka, 2006b) was used.

RESULTS

Power of mixed model with gene by environment interaction

We reported the power of MLM to detect gene main effect and gene by environment interaction (Fig. 2, Fig. S1). In both maize and pearl millet samples, the power increased with the genetic effect ratio r and the allele frequency q. The power was highest with allele frequencies of 0.5 and 0.25 and with high heritability. Also, the power to detect the interaction was more sensitive to the coefficient λ; this was expected because λ measures the increase of the effect of the SNP from the first environment to the second. As a consequence, the power is challengingly low for the detection of gene by environment interaction when the low heritability is combined with low allele frequency and/or low effect size. The global pattern of power variation with the parameters (r, λ, q) is consistent for both datasets (maize and pearl millet), but maize sample performed globally better than pearl millet sample in terms of absolute power. So in maize sample, environmental interaction underlined by common alleles with relatively strong effects were still detectable at h^2=0.25 and even the detection of rare alleles (q=0.05) was improved, compared to pearl millet (Fig. S1). Finally, we calculated the simulated effects resulting from the combination of parameters. The range of effects highly detected in both panels is reported (Table S2). For instance, SNP by environment interaction effects of 3.05 days and 4.44 days were detected for flowering trait with a power of about 95% in maize and pearl millet respectively (h^2=0.75, q=0.5).

Power to detect two and three order interactions

We also assessed the power of MLM to detect complex interactions (Fig. 3, Fig. S2 and Fig. S3). Two and three way interactions between SNP (S), ancestry (Q) and/or environment (E) were examined. The impact of the parameters (r, q, h^2) on the power was globally similar to the impact described in the case of gene by environment interaction. However, in the current situation, the power to detect the interactions appeared more sensitive to the value of λ. Thereby, even in the best condition of allele frequency and heritability (Fig. 3), the three way interaction (i.e. Q x S x E) was not highly detected unless a critical value of λ is reached (roughly λ >1 in the present samples). A stronger decrease of power was noticed with h^2 (Fig. S2 and Fig. S3). With h^2=0.75 and q=0.5, three way interaction effects of about 3.5 days or more were detected with a high power (>95%) for both panels (Table S3). Note that the results for complex interactions were not displayed at the lower allele frequency 0.05 in the pearl millet sample. A part of ASReml-R runs aborted in this case due to the low number of individuals carrying the SNP and the background marker. Indeed, with respect to the model of simulation, the expectation for this number of individuals is q x q_0 x n (q and q_0 are respectively the expected frequencies of S and Bk, and n the sample size). So for the present case (q=0.05, q_0=0.28 and n=90), only 1.26 individuals in average were expected to carry the two markers. With stochastic deviations, this number falls across iterations to higher values (>2 individuals) allowing model fit, but also to lower values (1 or 0 individuals) producing singularity.

Performance of information criteria for model selection

We tested the ability of different information criteria to select the right mean structure of MLM fitted by REML. We set the same variance parameters for all the competing models, so that the selection targets specifically the set of fixed parameters (mean structure). We reported, given each criterion, the frequency of selection of each of the competing models for pearl millet and maize datasets respectively (Fig. 4 and Table S4). The appropriate structure of each simulated dataset is known *a priori*. Given that, the model with no interaction (1) is the best model for the data generated under simulation scheme 1; the gene by environment interaction model (2) is the best model for data generated under simulation scheme 2 and the model with three way interaction (3) is the best model for data from simulation scheme 3. Over all schemes, the frequency of success of each information criterion (i.e. frequency of selection of the correct model) tends to increase with r (Table S4). In some cases, higher values of r leaded to a plateau at the maximum frequency of success. Among the tested values, r=1 could represent an intermediate-to-high relative effect size. We displayed the performance of information criteria at r=1 (Fig. 4). For each criterion, the result for only one variant was plotted, as the feature was globally similar between two variants of the same criterion (Table S4). For pearl millet sample (Fig. 4A), three categories of behaviour were noticeable: (i) AIC criterion tended to systematically choose the model with the highest number of parameters (full model); (ii) AICC (and R^2_{adj} to a certain extent) showed moderate-to-good performance (frequency of success around 50% to 85%) in all the 3 simulation schemes, but with a relative bias to overselect the full model; (iii) BIC and CAIC showed a very high performance (frequency of success around 95%) in the scenario where the simplest model was right (simulation scheme 1), but tended to often miss the selection of the full model when this model is true (frequency of success around 33% only). The simulation on maize sample showed a relative improvement of the performance of all criteria (Fig. 4B). The biased preference of AIC and AICC for the full model was reduced and the frequency of success of BIC and CAIC for the full model became acceptable (roughly 90%).

The main difference between pearl millet and maize samples relies on sample size, maize sample being 3 fold larger. The hypothesis about the impact of sample size was more properly tested using inbred subsets of gradual size in maize (Fig. S4). We therefore established that the frequency of success of all selection criteria (but AIC) is significantly linked to sample size (Table S5; $p<10^{-10}$). The impact of sample size was stronger with BIC and CAIC. Note that besides sample size, the parameters σ_G and q_0 varied with sampling. σ_G (involved in effect size setting) had also significant impact on the frequency of success of all the criteria; the frequency of background marker (q_0) showed significant impact on selection success only for the consistent criteria (BIC, CAIC) and for R^2_{adj} (Table S5). Note that the impact of σ_G and q_0 was lower than the impact of sample size.

Analysis of interactions in the pearl millet association panel

We used the extended mixed linear models to analyze association with pearl millet flowering trait scored into a design including 9 trials. Prior to this analysis, pearl millet kinship matrix were transformed to obtain a positive definite matrix. The obtained matrix was highly correlated to the original matrix (Spearman's correlation coefficient *R*= 0.9998; Mantel test *p*<0.01). We then fit the data using a set of 5 competing models (Table 3). As different behaviors were highlighted among the information criteria through our empirical assessment, we chose to make a comparative model selection using two different criteria (AICC and BIC). Figure 5 presents the result of this procedure for one of the SNPs of *PHYC* gene (SNP in position 101 in Saidou *et al.* 2009). AICC led to the selection of the full model (Fit1). This model includes, besides main effects, two and three way interactions between the SNP, environment and ancestry. BIC selects the simplest linear mixed model (Fit5) which specifies main effects and no interaction.

Wald test of fixed effects in each of the two selected models provided consistent p-values for the effects that are common between the two selected models (Table 4). Among these, we noted a significant main effect of *PHYC* SNP in position 101 (Table 4). In the contrary, none of the interaction terms of the full model (Fit1) was significant at 5% threshold, even though a few probabilities were close to this threshold (Table 4). The results of model selection procedure and Wald test for the other SNPs found at the *PHYC* locus were consistent with the results presented for SNP in position 101 (Fig. S5 and Table S6). This was actually expected, as all these SNPs are tightly linked (Saïdou *et al.*. 2009).

Analysis of interactions in maize association panel

The preliminary analysis of maize data using the standard framework (MLM with no interaction effects) showed different results across trials (Table S7). *Vgt1* effect was detected in some environments, but not in others ($p<0.05$). However, the combined p-value obtained from these 7 independent trials using Fischer's method (Table S7) supported a globally significant effect of *Vgt1* ($p=7.52 \times 10^{-5}$). We analyzed these data with the extended mixed linear model, to fit the whole multitrial information at once and to test for possible background or environmental interactions. The Box-Cox transformation (Gurka, 2006b) slightly reduced departure of the residuals from model assumptions (data not shown). We performed model selection with AICC and BIC in parallel (Fig. 6), based on the set of 5 candidate models listed in Table 3. AICC chose a model including gene by environment interaction and ancestry by environment interaction (Fit3). BIC chose a close nested model (Fit4), in which the gene by environment interaction is dropped.

We assessed the significance of the terms fitted by each of the two models respectively (Table 5). For terms that are common between the two models, the p-values obtained by either model were unambiguously consistent, and this result supported a significant effect of interaction between environment (trial) and ancestry ($p=2.52 \times 10^{-10}$ for the NS population, and $p=5.06 \times 10^{-18}$ for the TS population). A highly significant environment main effect was also highlighted ($p<10^{-26}$), as expected with contrasting environments. Finally, the part of *Vgt1* effects set as main effect was not significant ($p=0.37$); instead, the term of interaction between *Vgt1* and environment (Fit4) was significant ($p=2 \times 10^{-6}$).

DISCUSSION

How data parameters have impacts on the power of mixed linear model?

Previous association mapping studies in plant investigated the power of mixed linear model mainly in the case of patterns involving main effects (Yu *et al.* 2006; Stich *et al.* 2008; Saïdou *et al.* 2009; Zhu et Yu 2009). In this study, we addressed background interaction and genotype by environment interactions in structured populations. We simulated patterns with interaction to investigate the power of MLM framework to detect such effects on the phenotype. The simulation was incremented on real plant panels presenting a structured background (maize and pearl millet). Diverse patterns of G x E interactions do occur in real dataset (Mackay and Anholt, 2007). Our first example of interaction modelling (i.e. simulation scheme 2) was relatively simplified to provide a learning case of gene by environment interaction, and to address the problem from a basic level. The last presented scheme (two and three way interactions) was more complexified and allowed the examination of higher order interactions.

We analyzed how different parameters have repercussions on the ability to detect interaction effects in the assessed situations. Among these parameters, trait heritability and allele frequency showed strong effects on the power. We relatively simplified the setting of heritability. For instance, we did not explicitly modelled the components of variance linked to interaction for example, and only two levels of h^2 were considered. Nonetheless, the analysis does support a general trend about heritability effect. It highlighted to which extent the lack of heritability reduces the power to detect main effects of genetic factors, as well as the interactions of these factors with environment. Besides, effects caused by common alleles were more easily spotted, compared to effects of rare alleles. Such limitation about allele frequency has been already reported for MLM framework dealing with main effects solely, but it appeared more crucial in the current situation with interactions, especially when higher order interactions are considered. One obvious reason is that the estimation of higher order interaction (Q x S x E, for example) relies not on the allele frequency of one independent variable (q or q_0 in this example), but on the combination of the frequency of all variables involved in the interaction (q x q_0 in this example). As allele frequency is a proportion (i.e. a numeric value between 0 and 1), this combination is necessarily lower than any of the initial variable frequencies. Given these remarks, considerations on sample size are to be discussed more specifically when interest lies on higher order interaction.

Furthermore, we noted that the absolute level of power was improved in the simulations incremented on maize, compared to pearl millet. Although several initial properties of these two panels are relatively different (e.g. population structure, trait variance), it could be suggested that the improvement of power reflects mainly the effect of the larger sample size in maize (3 times the pearl millet sample size). We properly confirmed the effect of sample size on power using gradual sampling in maize panel. But despite the difference about the absolute value of power estimates between maize and pearl millet samples, it is important to note that the pattern of power variation (increase or decrease), as a function of allele frequency, heritability or effect size, was similar for these two panels.

The importance of a large sample size is reflected in results from the simulation scheme 3 with 5% allele frequency on the pearl millet data. We reported the occurrence of singularity and crash in ASReml-R runs with this parameter set. The size of subsamples representing the combinations of markers underlying the interaction was too low into the corresponding runs. This numerical exemple underlies a problem which will certainly be limitative for higher order interactions fitting in real data, especially when rare alleles and/or small samples are processed. The partition of the phenotypic variation between all the components specified in the model became no longer possible in such cases. Even if no singularity occurs, the reliability of model components defined on the basis of rare combinations is statistically questionable. The problem can be handled in simulation studies by limiting the range of deviation of allele frequency to avoid singularity and secondly by averaging results across a high number of iterations to establish the robustness of the result. It might be interesting for future studies: i) to discuss appropriate methods to handle such singularities in the analysis of real data and ii) to assess the robustness of interactions underlined by a small number of individuals.

Integrating all the simulation results together, it appears that no single parameter determines alone the power of the model and that the effects of all parameters might balance each other. For instance, a good allele frequency does not necessarily determine a good power when heritability is not sufficient as well, and reciprocally. Thereby, the power of MLM framework on a particular set of data has to be discussed with respect to the combination of data characteristics in place, instead of focusing on one or two single parameters. When planning a plant association study, they are some parameters that could be more easily handled. For instance, the size of a panel can be extended with an appropriate sampling effort to reach a desired sample size, while being aware that the increase of global sample size does not necessary improve as well the frequency of particular combinations of factors. Adequate sampling strategies are needed to tackle this problem without sampling bias. In the same way, small phenotypic effect or low allele frequency also limit the performance of MLM framework. To be able to detect rare alleles, the use of family mapping (linkage analysis) is recommended (Myles *et al.* 2009). A joint approach combining linkage and association mapping was developed in maize, with the Nested Association Mapping population (McMallen *et al.* 2009). This framework proved very efficient for the study of the genetic architecture of flowering time (Buckler *et al.* 2009).

Model selection for multitrial data

Model selection procedures are used to select, from different sets of parameters, those that fit better the data. The principle of parsimony is commonly associated with model selection. This principle implies to limit, as well as possible, the number of parameters in the model. The performance of information criteria was assessed to select the mean structure of mixed model. The behaviour of the considered criteria was partly antagonistic. AIC and AICC biased the selection by wrongly preferring larger models. On the contrary, BIC and CAIC tended to wrongly reject the full model. This behaviour has been previously reported for particular simulated pattern of data and seems to occur particularly when the total variance of the data is large (Gurka, 2006a). With lower variance, the bias in all these criteria would be the overselection of the full model (Gurka, 2006a). Nevertheless, we showed that the selection criteria, notably BIC and CAIC, perform better with larger samples, and this is in agreement with previous studies (Hurvish and Tsai 1989, Gurka 2006a, Wang 2007).

The overall performance of the selection criteria, as examined for structured population data in this study, suggests it is possible to use these criteria in model selection under REML, while being aware of their limits. In particular, we should keep in mind that, depending on the data, efficient criteria (AIC, AICC) may violate parsimony while consistent criteria (BIC and CAIC) may lead to the irrelevant removal of informative parameters from the model. There is no a clear objective consideration that might lead to prefer one of the categories of criteria over another (see also Gurka 2006a, Bozdogan 1987). Moreover it is not straightforward to choose which bias is more acceptable, even though certain authors consider more severe the fact of removing relevant parameters from a model (Gurka 2006a). Interestingly, we showed that both biases are reduced with larger samples. As information criteria do not lead to an unambiguous model selection in small samples, additional strategies should be adopted. Methods assessing uncertainty of model selection exist (Chatfield 1995, Posada and Buckley

2004, Spiegelhalter 2002). Uncertainty assessment could conciliate, in some situations, small differences in model selection, by clustering different models in the same confidence interval. Also, tests of fixed effects such as Wald test could be used as *ad hoc* strategies to appreciate, *a posteriori*, the significance of the parameters selected in a model (Gurka 2006a).

Statistical evidence suggesting genotype by environment interactions in maize

To select parameters for maize and pearl millet real data, we adopted an *ad hoc* strategy which consisted of the comparative use of both efficient (AICC) and consistent (BIC) criteria. For maize data, AICC and BIC led to similar results, as the model selected by AICC had only one additional term (*Vgt1* by environment interaction), compared to the model selected by BIC. It is very likely that a procedure of model selection uncertainty could cluster these two close models in the same confidence interval, so that one cannot discard either of the models to be a good approximation of reality. We did not implemented model selection uncertainty procedures in the current study.

We relied on Wald test to appreciate the informativeness of the terms retained by each of the criteria. The p-values were consistent for the terms common to both models. Significant ancestry by environment interactions were reported with both models. This might be probably indicative of environmental adaptation linked specifically to the considered populations. The information captured by ancestry inference in population structure analyses relies on multilocus data and are shaped by alleles predominant or fixed in a given population. One obvious assumption to explain ancestry by environment interaction could be that specific alleles (identified or not) are present in the considered population and shape the environmental plasticity of individuals from that population.

Flowering time data was scored in contrasted environments. The main effect term for *Vgt1* was not significant in the multitrial model, but the *Vgt1* by environment interaction was significant ($p<10^{-6}$). This could make sense if the difference between the two *Vgt1* alleles is modulated by environmental cues. This hypothesis is actually sustained by the trial-by-trial analysis (Table S7), which showed that *Vgt1* effect was significantly detected in some but not all environments. Such results are traditionally known as indicative of gene by environment interactions. Given all these statistical results (single trial model, model selection and multitrial model) our study statistically supports interactions between *Vgt1* and environment in the one hand, and interaction between ancestry and environment in the other hand. Nonetheless, these statistical evidences need further assessment to be biologically interpreted with confidence.

Difficulties in selecting a model to analyze the pearl millet flowering in multitrial design

The comparative model selection using AICC and BIC led to a contrasting result that pinpoints two extreme models for pearl millet (Fig. 5). AICC selected the full model with two and three way interactions, whereas BIC selected the simplest model with main effects solely. Given the considerable difference in the number of parameters between the two models, it is unlikely that considering model uncertainty could conciliate these results. Also, confronting the two results is not straightforward, particularly for the current sample size for which both criteria are shown to produce specific biases. Considering these biases, the probability is not nil that the full model is the best as pointed out by AICC, but BIC still prefers the simple model because of applying too much penalty to the full model; and inversely, it cannot be discarded that the simple model is the best as pointed out by BIC, but AICC failed the selection due to its biased preference for full models. One singular fact is the brutal change in AICC around the full model. This might be explained if, for example, the three order level of interactions was informative but a part in the entire set of two way interactions was not. If this is the case, it could explain why the full model is highly penalized (using BIC) by the presence of too much none informative parameters. Anyway, it is generally a good practice to do not remove interaction of small order until higher order interaction are set, so we kept this way of setting the full model. So finally, no obvious element conciliated the difference in model selection by AICC and BIC in this particular data of pearl millet.

To appreciate the terms with regard to their significance, we applied Wald test to the effects fitted by two models respectively. Terms that were retained by the two models produced highly consistent p-values. This included notably the main effect of *PHYC* polymorphism, suggesting an association already reported in a study that used the classical MLM framework (Saïdou *et al.* 2009). None of the interaction terms added in the full model was significant, even though some values were found at the limit of the threshold ($p<0.05$). This Wald test result could reinforce the choice of the simple model, even though it could not arbitrate definitely when keeping in mind other methodological limits (notably the limit of power). For instance, although *PHYC* allele frequency is at favourable level for this sample, the structure of pearl millet populations is characterized by a low average ancestry in each population (from 0.08 to 0.28). Such structure lead to a low number of individuals for each combination of the interacting factors and this is actually expected to limit the power for the detection of any three way interaction. In the same way, there is a partial covariance of *PHYC* with the ancestry in one of the seven populations (data not shown). This covariance is expected to confound *PHYC* by environment interaction even if this interaction was true, as the incremental Wald test adjust systematically population effects prior to gene effects. Therefore, it is not clear which model reflects better the real pattern of these pearl millet data. Only the parameters for main effects (notably *PHYC* main effect) could be retained with no statistical contradiction. More generally, this case illustrates possible contradictions that could occur between statistical methods. Statistical perspectives to efficiently handle such contradictions in the case of plant association studies are yet to be addressed. However, further biological experiments about the interactions may allow to the get round the statistical problem.

CONCLUSION

In this study the perspective of using structured association mapping panels to detect interaction effects on phenotype was assessed. Interactions were simulated to investigate methodological issues including model power and model selection. We discussed, with respect to diverse parameters, the variation of the power for the detection of interactions involving genetic factors and/or environment. The methods were implemented on two plant panels with different size (maize and pearl millet). These simulations suggest that the current mixed linear model framework will be efficient to detect mainly effects of interaction involving common alleles with moderate-to-large effects on phenotype. In particular, interactions with environment would be detected more frequently only if the difference in effect magnitude due to environment is relatively strong. High trait heritability would also facilitate the detection of associations. Besides, we showed that larger samples perform better, so increasing the size of association panels could improve the efficiency of the framework. Sample size is also important to achieve model selection under REML. Particularly, we showed that mean structures could be compared under REML with better performance when sample size is larger. For samples of small size, biases that could often occur were highlighted, notably the overselection of the full model by efficient criteria (AIC, AICC) and underfitting by consistent criteria (BIC, CAIC). We also illustrated the use of the presented methods with real datasets based on multitrial field experiments in maize and pearl millet. Pearl millet dataset did not allow a straight statistical assessment of genotype by environment interactions. But the results were interesting and allowed an open discussion about methodological difficulties. We hope that this discussion will instigate further examination by the scientific community. With the maize dataset, we found statistical evidence of genotype by environment interaction characterized by interaction between ancestry and environment in the one hand, and interaction between *Vgt1* gene and environment in the other hand. For both datasets, additional biological examination of the current results has to be performed, as this paper focused mainly on statistical aspects. Finally, it appears that mixed linear model framework could be a suitable tool to address genotype by environment interactions in structured population, even if some limitations are to be resolved. This could be promising for the study of a large set of genotype by environment interactions or adaptive population differentiation.

ACKNOWLEDGEMENT

This study is supported by a grant from the Agropolis Fondation to A-A.S., Y.V. and A-C.T., and by a grant from the National Research Agency of France (ANR-07-JCJC-0116 to Y.V). A.-A.S. was supported by IRD Ph.D. fellowship and performed part of this research in Cornell University (February-April 2009). We thank Edward S. Buckler, Peter Bradbury, Elhan Ersoz and Zhiwu Zhang for discussion and help during modelling and statistical analyses. We acknowledge Alain Charcosset, Jacques David and Laurène Gay for discussion about the preliminary versions of this work. We thank Jean-Louis Pham for advice during project design.

REFERENCES

Akaike H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control. AC 19: 716-723.

Allinne C., C. Mariac, Y. Vigouroux, G. Bezançon, E. Couturon *et al.*, 2008. Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. Genetica 133:167-178.

ASReml package for R (ASReml-R), version 2.0/32. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Atwell S., Y. S. Huang, B.J. Vilhjalmsson, G. Willems, M. Horton *et al.*, 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465: 627-631.

Bozdogan H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52: 345-370.

Brachi B., N. Faure, M. Horton, E. Flahauw, A. Vazquez, M. Nordborg, J. Bergelson, J. Cuguen, and F. Roux, 2010. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genet 6(5): e1000940. doi:10.1371/journal.pgen.1000940

Buckler E. S., James B. H., P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009.The genetic architecture of maize flowering time. Science 325: 714-718.

Caicedo A. L., J.R. Stinchcombe, K.M. Olsen, J. Schmitt and M. D. Purugganan, 2004. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. PNAS 101: 15670-15675.

Chatfield C., 1995. Model uncertainty, data mining and statistical inference. J. R. Stat. Soc. A 158:419-466.

Crawley M. J., 2007. The R book. John Wiley & Sons, Ltd; The Atrium, Southern Gate, Chichester,West Sussex PO19 8SQ, England. 942 p.

El-Lithy M., L. Bentsink, C.J. Hanhart, G.J. Ruys, D. Rovito *et al.,* 2006. New Arabidopsis recombinant inbred line populations genotyped using SNPwave and their use for mapping flowering-time quantitative trait loci. Genetics 172: 1867- 1876.

Falconer D.S. and T.F.C. Mackay, 1996. Introduction to Quantitative Genetics. Ed. 4, Addison-Wesley Longman, Harlow, UK.

Falush D., M. Stephens and J. K. Pritchard, 2003. Inference of population structure using multilocus genoype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Falush D., M. Stephens and J. K. Pritchard, 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes 7(4) : 574-578.

Gao H., S. Williamson and C.D. Bustamante, 2007. An MCMC approach for the joint inference of population structure and inbreeding rate from multi-locus genotype data. Genetics: 176: 1635-1651.

Gilmour A.R., B.J. Gogel , B.R. Cullis, and R. Thompson, 2006. ASReml User Guide Release 2.0. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Gurka M. J., 2006a. Selecting the best linear mixed model under REML. The American Statistician, 60(1): 19-26.

Gurka M.J., 2006b. Extending the Box-Cox transformation to the linear mixed model. J. R. Statist. Soc. A 169 (2): 273-288.

Hall D., C. Tegström and P. K. Ingvarsson, 2010. Using association mapping to dissect the genetic basis of complex traits in plants. Briefings in Functional Genomics 9 (2): 157-165.

Hardy O.J., and X. Vekemans, 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618-620.

Hubisz M.J., D. Falush, M. Stephens and J.K. Pritchard, 2009. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9, 1322-1332.

Hurvich C. M. and C. L. Tsai, 1989. Regression and time series model selection in small samples. Biometrika, 76: 297-307.

Jannink J.-L., 2007. Identifying quantitative trait locus by genetic background interactions in association studies. Genetics 176: 553-561.

Jannink J.-L., 2008. QTL by genetic background interaction: predicting inbred progeny value. Euphytica 161:61-69.

Flint-Garcia S.A., A.-C. Thuillet, J.M. Yu, G. Pressoir, S.M. Romero *et al.*. 2005. Maize association population: a high resolution platform for QTL dissection. Plant Journal 44: 1054-1064.

Kang H.M., Zaitlen N.A., Wade C.M., Kirby A., Heckerman D. *et al.*, 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Kover P.X., Valdar W., Trakalo J., Scarcelli N., Ehrenreich I.M. *et al.*, 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet 5(7): e1000551.

Leng C, 2008. The Residual information criterion, corrected. ARXIV, Bibliographic code: 2007arXiv0711.1918.

Li L., M.J. Paulo and F. van Eeuwijk, 2010. Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theor Appl Genet, in press. DOI: 10.1007/s00122-010-1389.

Loiselle B. A., V. L. Sork, J. Nason and C. Graham, 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. J. Bot. 82: 1420-1425.

Manolio T.A., F.S. Collins, N.J. Cox *et al.*, 2009. Finding the missing heritability of complex diseases. Nature 461:747-53.

Mariac C., V. Luong, I. Kapran, A. Mamadou, F. Sagnard *et al.*, 2006. Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theor Appl Genet 114:49-58.

McMullen M. D., S. Kresovich, H. S. Villeda, P. Bradbury, Huihui Li *et al.*, 2009. Genetic properties of the maize nested association mapping population. Science 325: 737: 740

Myles S, J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang, D. E. Costich, and E. S. Buckler, 2009. Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell 21: 2194-2202.

Nordborg M. and D. Weigel, 2008. Next-generation genetics in plants. Nature 456:720-723.

Patterson N., A. L. Price and D.Reich, 2006. Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190.

Phillips P.C., 2008. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855-867.

Posada D. and T. R. Buckley, 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53 (5): 793-808.

Pritchard J.K. , M. Stephens, N. A. Rosenberg, and P. Donnelly, 2000a. Association mapping in structured populations. Am. J. Hum. Genet. 67:170-181.

Pritchard, J.K., M. Stephens and P. Donnelly, 2000b. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rafalski J.A., 2010. Association genetics in crop improvement. Curr Opin Plant Biol 13:1-7.

Remington D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt *et al.*, 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. PNAS 98: 11479-11484

Saïdou A.-A., C. Mariac, V. Luong, J.-L. Pham, G. Bezancon, and Y. Vigouroux, 2009. Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. Genetics 182: 899-910.

Sambandan D., M.A. Carbone, R.R.H. Anholt and T.F.C. Mackay, 2008. Genetics 179(2): 1079-1088.

Schabenberger O., 2004. Mixed model influence diagnostics. SUGI 29 Proceedings: Paper 189-29.

Schwarz G., 1978. Estimating the dimension of a model. The Annals of Statistics, 6: 461-464.

Shi P. and C.-L. Tsai, 2002. Regression model selection - A residual likelihood approach. J. R. Stat. Soc. B 64: 237-252.

Spiegelhalter D. J., N. G. Best, B. P. Carlin and A. van der Linde, 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B 64 (4): 583-639.

Stich B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler *et al.*, 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745-1754.

Stracke S, Haseneyer G, Veyrieras J-B,Geiger H H, Sauer S, Graner A, and H-P Piepho, 2009. Association studies reveals gene action and interactions in the determination of flowering time in barley. Theor Appl Genet 118:259-273.

Thornsberry J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen *et al.*, 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28: 286-289.

Uwatoko N., A. Onishi, Y. Ikeda, M. Kontani, A. Sasaki, K. Matsubara, Y. Itoh, and Y. Sano, 2008. Epistasis among the three major flowering time genes in rice: coordinate changes of photoperiod sensitivity, basic vegetative growth and optimum photoperiod. Euphytica 163:167-175.

Verbeke G. and G. Molenberghs, 2000. Linear mixed models for longitudinal data, New York: Springer-Verlag.

Vonesh E.F., V. M. Chinchilli and K. Pu, 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics 52:572-587.

Wang J., 2007. Selecting the best linear mixed model using predictive approaches. Master of Science. Brigham Young University.

Yu J., G. Pressoir, W.H. Briggs, I. V. Bi, M.Yamasaki *et al.*, 2006. A unified mixed-model method for association studies that accounts for multiple levels of relatedness. Nat. Genet. 38: 233-208.

Yu J., Z. Zhang, C. Zhu, D. A. Tabanao, G. Pressoir, M. R. Tuinstra, S. Kresovich, R. J. Todhunter, and E. S. Buckler, 2009. Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Gen. 2:63-77.

Zhang N., A. Gur, Y. Gibon, R. Sulpice, S. Flint-Garcia, M. D. McMullen, M. Stitt and E. S. Buckler, 2010b. Genetic analysis of central carbon metabolism unveils an amino acid substitution that alters maize NAD-dependent isocitrate dehydrogenase activity. PLoS ONE 5(4): e9991. doi:10.1371/journal.pone.0009991.

Zhang Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas and E. S. Buckler, 2010a. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics, 42: 355-360.

Zhu C. and J. Yu, 2009. Nonmetric multidimensional scaling corrects for population structure in whole genome association studies. Genetics 182: 875-888.

Zhu C., M. Gore, E.S. Buckler and J. Yu, 2008. Status and prospects of association mapping in plants. The Plant Genome 1:5-20.

FIGURES

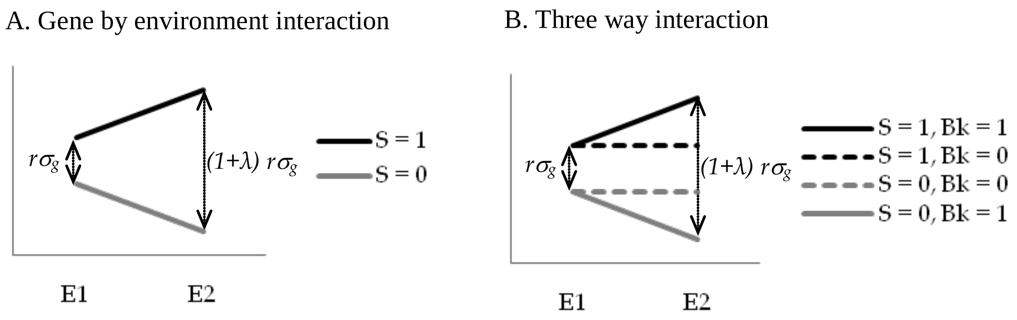


Figure 1. Modelling of genotype by environment interactions. Trait values (y axis) are presented into two environments E1 and E2. A) The value of individual phenotype is given with respect to the presence (1) or the absence (0) of a SNP (*S*); $r\sigma_g$ is the effect of the SNP into E1; $(1+\lambda)\ r\sigma_g$ is the effect of the SNP into E2. The coefficient λ is a numeric value that quantifies the change of SNP effect from one environment to another. B) Besides the marker *S*, a marker correlated to population background is considered. The variable *Bk* denotes the presence (1) or the absence (0) of this marker. The SNP effect ($r\sigma_g$) is stable from environment E1 to E2 when *Bk* is absent. The size of this SNP effect varies with environment only in the presence of *Bk*.

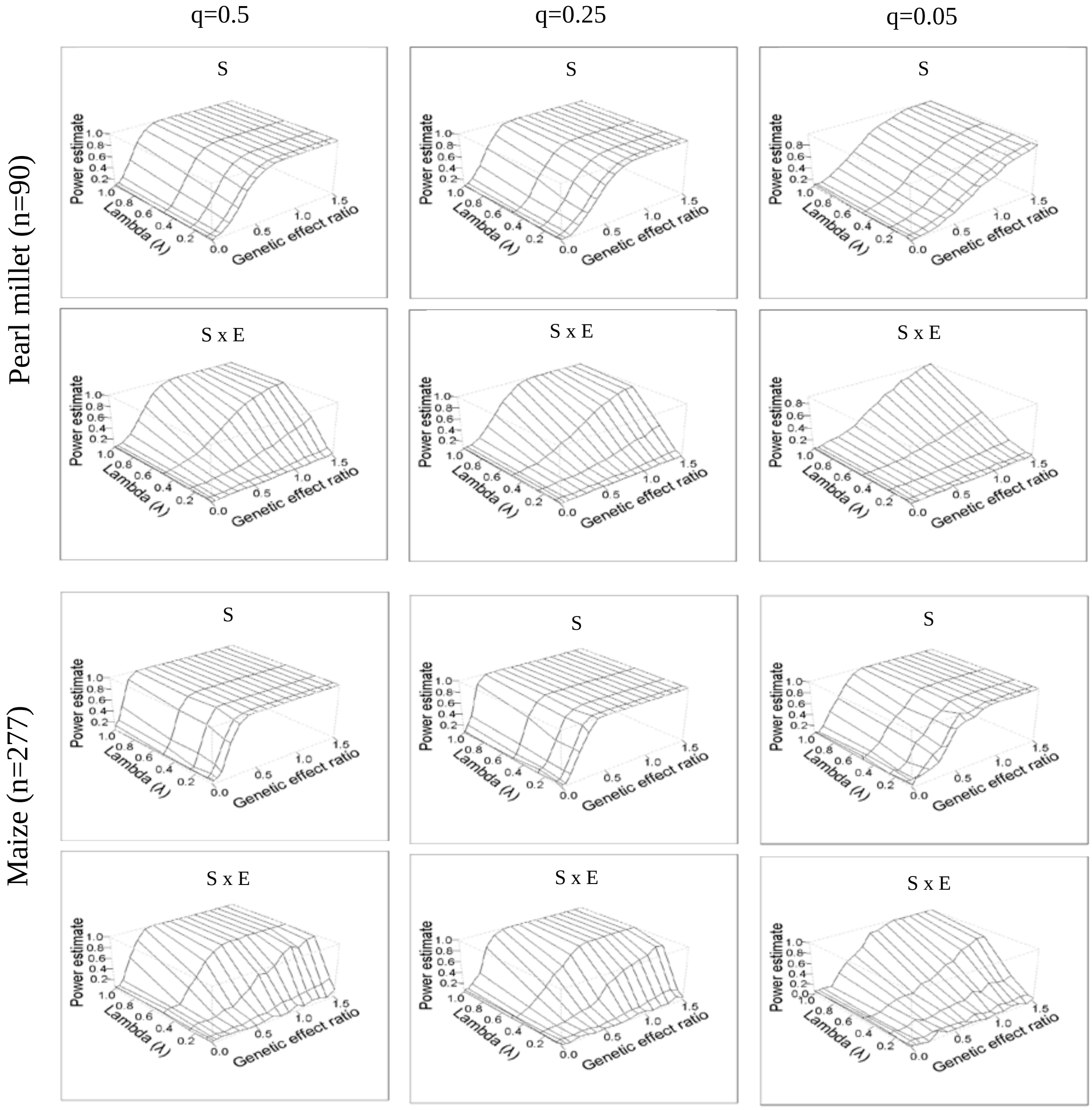


Figure 2. Power of mixed linear model to detect phenotypic effects including gene by environment interaction. A SNP with differential effect across environment was simulated based respectively on pearl millet and maize association mapping populations (see text for details). The main effect of the SNP (*S*) and the interaction between this SNP and environment (*SxE*) were fitted into a mixed linear model. The power of the model is calculated as the ratio of the number of runs where a given effect is significantly detected and the overall number of runs. The power is plotted for h^2=0.75 and according to allele frequency (*q*), genetic effect ratio (*r*) and λ. The parameter λ measures the variation of SNP effect magnitude with environment. The largest panel (maize, n=277 individuals) performed globally better. However, the relative variation of power as a function of parameters shows

similar feature into both panels. Power increased with r and was higher with common allele frequencies (q=0.5, q=0.25). The ability to detect the interaction was more particularly sensitive to λ. The highest range of power (say power > 80%) corresponded overall to relatively large parameters' values. This indicates that these current mapping frameworks might be limitative for traits that are fundamentally shaped by loci with too small individual effects.

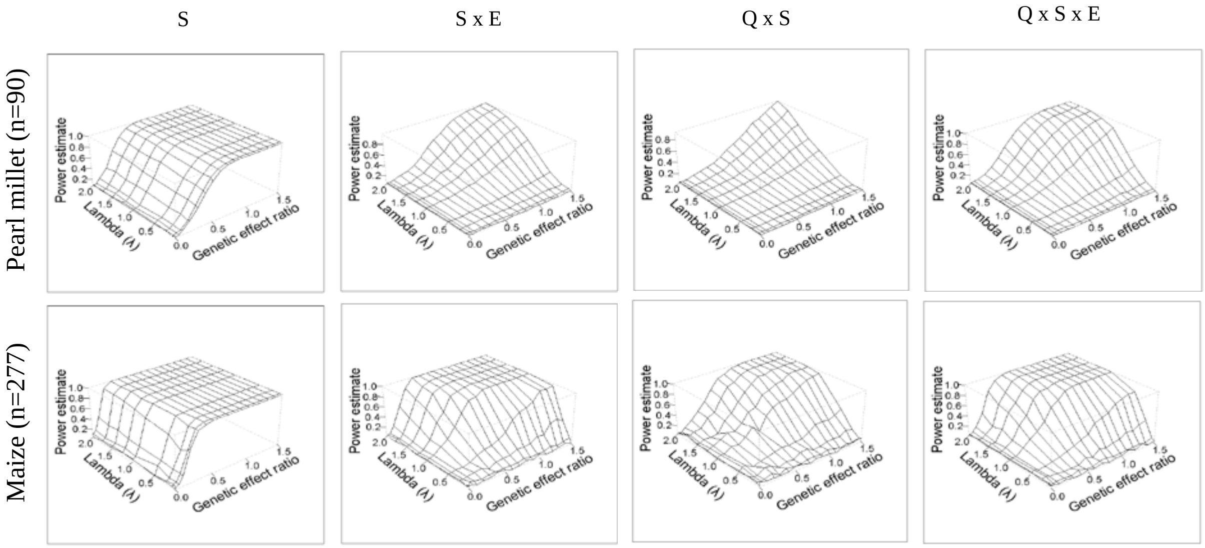


Figure 3. Power of mixed linear model to detect phenotypic effects including three way interaction. The data were simulated using pearl millet and maize panels respectively (see text for details). The set of simulated effects consists of SNP main effect (*S*), SNP by environment interaction (*S x E*), ancestry by SNP interaction (*Q x S*) and three way interaction between ancestry, SNP and environment (*Q x S x E*). The power to detect each effect is plotted for the heritability h^2=0.75 and the allele frequency *q*=0.5, according to the effect ratio *r* and the parameter λ that modulates the variation of effect with environment. Power increases with *r* for all the effects, and a strong effect of *λ* is observed for the interactions. The highest range of power (say power > 80%) is reached only with relatively large effect size. Otherwise, note the relative improvement of power for the maize panel (3 times larger than pearl millet panel).

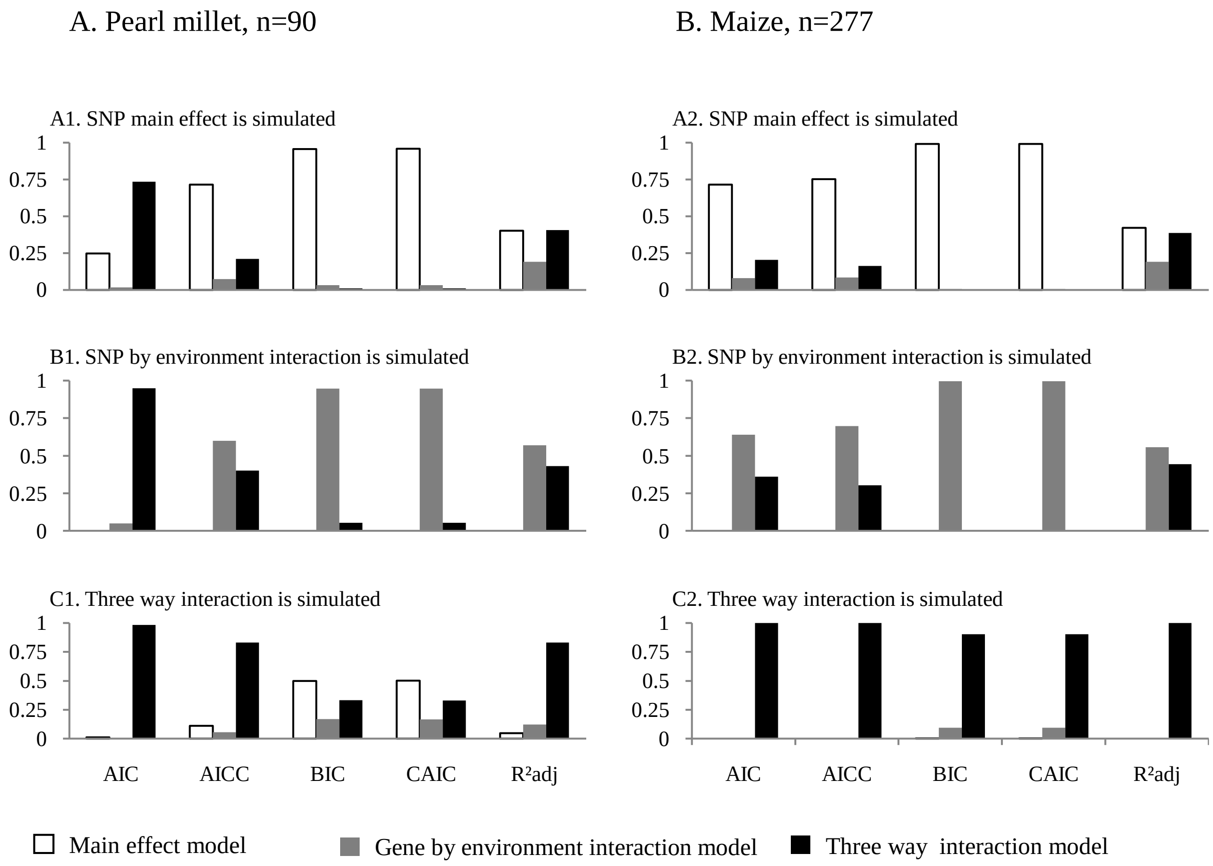


Figure 4. Selection of the mean structure of mixed linear model under REML. Data were simulated respectively with SNP effect (A1- A2), SNP by environment interaction (B1-B2) and up to three way interaction between SNP, ancestry and environment (C1-C2). Simulation was based on initial datasets from pearl millet (left) and maize (right). Each dataset was fitted by three competing models with different fixed parameters, and models were compared using respectively AIC, AICC, BIC, CAIC and R^2_{adj}. All models were set with the same variance parameters. The frequency of selection of each competing model was assessed on the basis of 1000 iterations per point. The result is displayed here for one combination of parameters' values (q=0.5, h^2=0.75, r=1, λ=1; see text for details). With respect to the simulation schemes used to generate the data, the expected adequate model is respectively the main effect model (A), the gene by environment interaction model (B), and the three way interaction model (C).

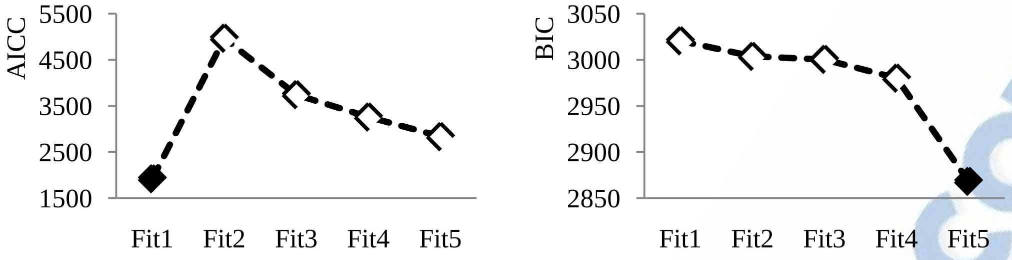


Figure 5. Model selection for the analysis of associations in pearl millet multitrial design. A set of five nested models (named Fit1 to Fit5) were specified to fit flowering time scored on 90 pearl millet inbreds in 9 field trials at Sadoré (Niger). Fit1 is the full model specifying up to three way interaction; Fit5 is the simple model with only main effects; Fit2 to Fit4 are intermediate models between the two (see Table 3). The main fixed effects in each model are environment (trial), ancestry and gene polymorphism gene (SNP in position 101 at *PHYC* locus is displayed). We compared the models using AICC (left) and BIC (right). AICC criterion pinpoints the full model (highlighted by dark point), while BIC criterion prefers the reduced model (highlighted by dark point). Note that for both criteria, lower value is considered as better.

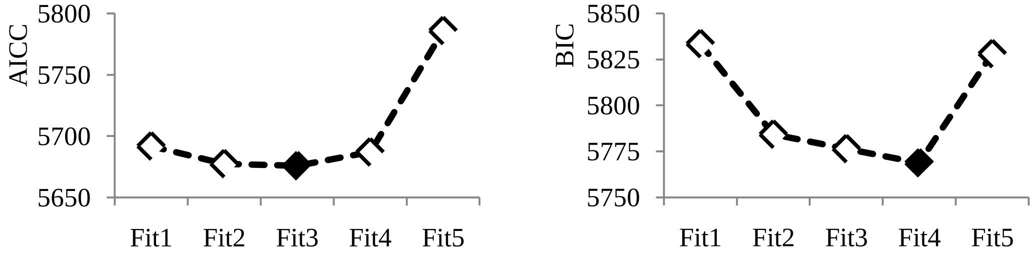


Figure 6. Model selection for the analysis of associations in maize multitrial design. Flowering time trait was scored on 7 environments. We fitted the data using a set of five nested mixed linear models with differences in mean structure. Fit1 is the full model with up to three way interaction and Fit5 is the main effects model; Fit2 to Fit4 are intermediate nested models (see Table 3). Genetic factors (ancestry, polymorphism at *Vgt1* gene) as well as environmental factor (trials) were fitted as fixed main effects in all the models; two and three way interactions were added (or not) given the model. AICC criterion (lower value is better) selects Fit3 as the best model (highlighted by dark point); while BIC criterion selects Fit4 (highlighted by dark point). Fit3 includes ancestry by environment interactions and *Vgt1* by environment interaction. Fit4 is the reduced model with *Vgt1* by environment interaction dropped.

TABLES

Table 1. Summary of parameters used in simulation

Symbol	Definition	Range examined in this study
n	Number of individuals in the panel (sample size)	90 [a] to 277 [b]
p_i	Initial phenotype of individual i	Field-based flowering score [a, b]
r	Effect ratio	0 to 1.5
σ_G	Standard deviation of the initial trait	6.83[a]; 8.72[b]
λ	Numeric value measuring the change in SNP effect between environments	0.05 to 2
q	Expected SNP allele frequency	0.05; 0.25; 5
q_0	Mean ancestry (also expected frequency of the marker correlated to ancestry)	0.28[a]; 0.49[b]
h^2	heritability	0.25; 0.75

[a] based on pearl millet real data

[b] based on maize real data

Table 2. Formulation of information criteria

Criterion	Formula	Notation
AIC	−2l + 2s	AIC
AICC	−2l + 2s m / (m−s−1)	AICC
	−2l + 2s (N−p) / (N−p−s−1)	AICC*
BIC	−2l + s log(m)	BIC
	−2l + s log(N−p)	BIC*
CAIC	−2l + s log(m+1)	CAIC
	−2l + s log(N−p+1)	CAIC*
R^2_{adj}	$1 - (1-R^2)\ N / (N-p)$	$R^2_{\ adj}$
	$1 - (1-R^2)\ N/(N-p-k)$	R^2_{adj}*

The maximum log likelihood is noted l; p is the number of fixed parameters in the model, k is the number of variance parameters, s is the total number of parameters (i.e. s=p+k), m is the number of individuals (or *statistical units*), N is the total number of observations (product of the number of individuals by the number of repeats). Log function is natural logarithm. R^2 is the squared correlation coefficient calculated as ; where r_i is the conditional residual for the ith observation, calculated as the difference between the observed value and the predicted value. For all criteria but AIC, two variants are defined, with respect to the way of accounting for sample size. A star (*) is used in this paper to distinguish these two variants.

Table 3. Set of competing models specified for pearl millet and maize data

Model	Specification	Number of parameters	
		Maize	Pearl millet
Fit1	$Y = E + Q_i + S + Q_i \times E + S \times E + Q_i \times S + Q_i \times S \times E + (K + e)$	44	128
Fit2	$Y = E + Q_i + S + Q_i \times E + S \times E + Q_i \times S + (K + e)$	32	80
Fit3	$Y = E + Q_i + S + Q_i \times E + S \times E + (K + e)$	30	74
Fit4	$Y = E + Q_i + S + Q_i \times E + (K + e)$	24	66
Fit5	$Y = E + Q_i + S + (K + e)$	12	18

Y is the trait, E is environmental effect (trial effect), Q_i is population structure effect set by ancestry in the the ith population (from k available populations, k-1 are used in the model), S is SNP effect. K is polygenic background random effect set by kinship matrix and e is the residual. A cross between terms represents interaction effect. For each dataset, the total number of model parameters is given.

Table 4. Wald test of fixed effects for pearl millet dataset.

Effect	df	Fit1	Fit5
Intercept	1	$< 10^{-26}$	$< 10^{-26}$
Trial	8	$< 10^{-26}$	$< 10^{-26}$
Q_1	1	0.1917	0.1954
Q_2	1	0.0016	0.0017
Q_3	1	0.1610	0.1663
Q_4	1	0.0231	0.0244
Q_5	1	0.1103	0.1128
Q_6	1	0.0511	0.0519
PHYC	1	0.0044	0.0044
Trial x Q_1	8	0.2942	-
Trial x Q_2	8	0.3272	-
Trial x Q_3	8	0.1102	-
Trial x Q_4	8	0.4017	-
Trial x Q_5	8	0.6050	-
Trial x Q_6	8	0.8051	-
Trial x PHYC	8	0.0658	-
Q_1 x PHYC	1	0.5669	-
Q_2 x PHYC	1	0.2556	-
Q_3 x PHYC	1	0.1405	-
Q_4 x PHYC	1	0.6343	-
Q_5 x PHYC	1	0.6349	-
Q_6 x PHYC	1	0.5684	-
Trial x Q_1 x PHYC	8	0.1705	-
Trial x Q_2 x PHYC	8	0.3126	-
Trial x Q_3 x PHYC	8	0.8426	-
Trial x Q_4 x PHYC	8	0.0537	-
Trial x Q_5 x PHYC	8	0.0821	-
Trial x Q_6 x PHYC	8	0.6632	-

Fit1 is the model selected by AICC, and Fit5 is selected by BIC. P-values of Wald test are given for each term with respect to the fitted model. Significant p-values ($p<0.05$) are highlighted. For *PHYC* gene, SNP in position 101 is displayed. Q_i: ancestry in population i; df: degree of freedom.

Table 5. Wald test of fixed effects for maize dataset

Effect	df	Fit3	Fit4
Intercept	1	$< 10^{-26}$	$< 10^{-26}$
Trial	6	$< 10^{-26}$	$< 10^{-26}$
Q_{NS}	1	6.27×10^{-05}	5.86×10^{-05}
Q_{TS}	1	4.43×10^{-04}	4.21×10^{-04}
Vgt1	1	0.379	0.373
Trial x Q_{NS}	6	2.52×10^{-10}	4.16×10^{-10}
Trial x Q_{TS}	6	5.06×10^{-18}	1.25×10^{-17}
Trial x Vgt1	6	2.00×10^{-06}	-

Fit3 is the model selected by AICC, and Fit4 is the model selected by BIC. P-value of Wald test is given for each term with respect to the fitted model. Significant p-values ($p<0.05$) are highlighted. Q_i: ancestry in population i (see text); df: degree of freedom.

Saïdou *et al.*

SUPPLEMENTARY FIGURES

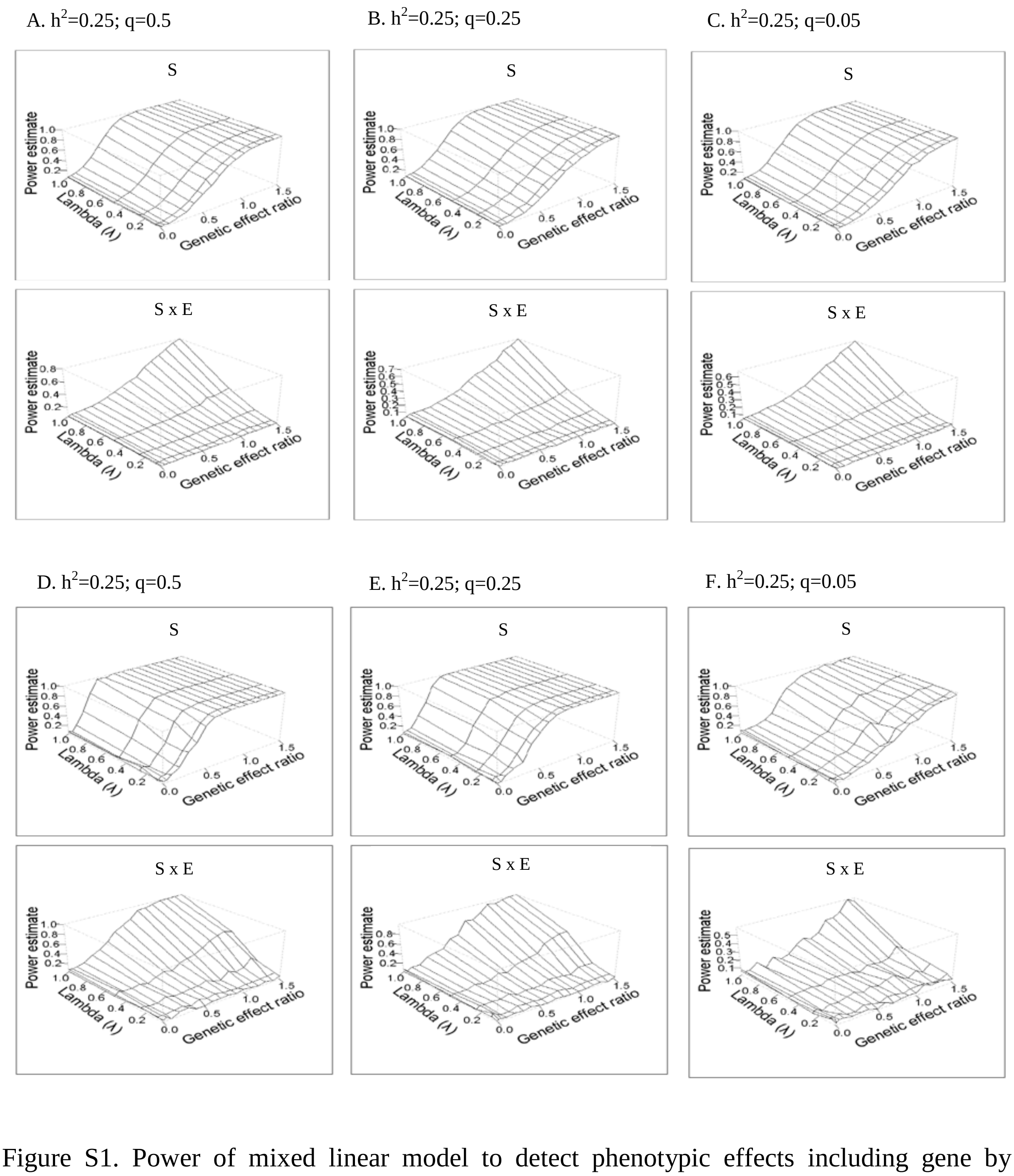


Figure S1. Power of mixed linear model to detect phenotypic effects including gene by environment interaction. –Pearl millet and maize samples, additionnal parameters levels. Data were simulated (see main text for details) considering the main effect (S) of a SNP, as well as the interaction of this SNP with environment (S x E). The simulated datasets were analyzed using a mixed linear model. The power was estimated as the ratio between simulated datasets for which a significant effect is detected and the total number of simulations. The

results are presented in function of the heritability (h^2), allele frequency (q), genetic effect ratio (r) and the parameter modulating difference of SNP effects between environment (λ). Simulation was carried out for pearl millet (A, B, C) and maize (D, E, F).

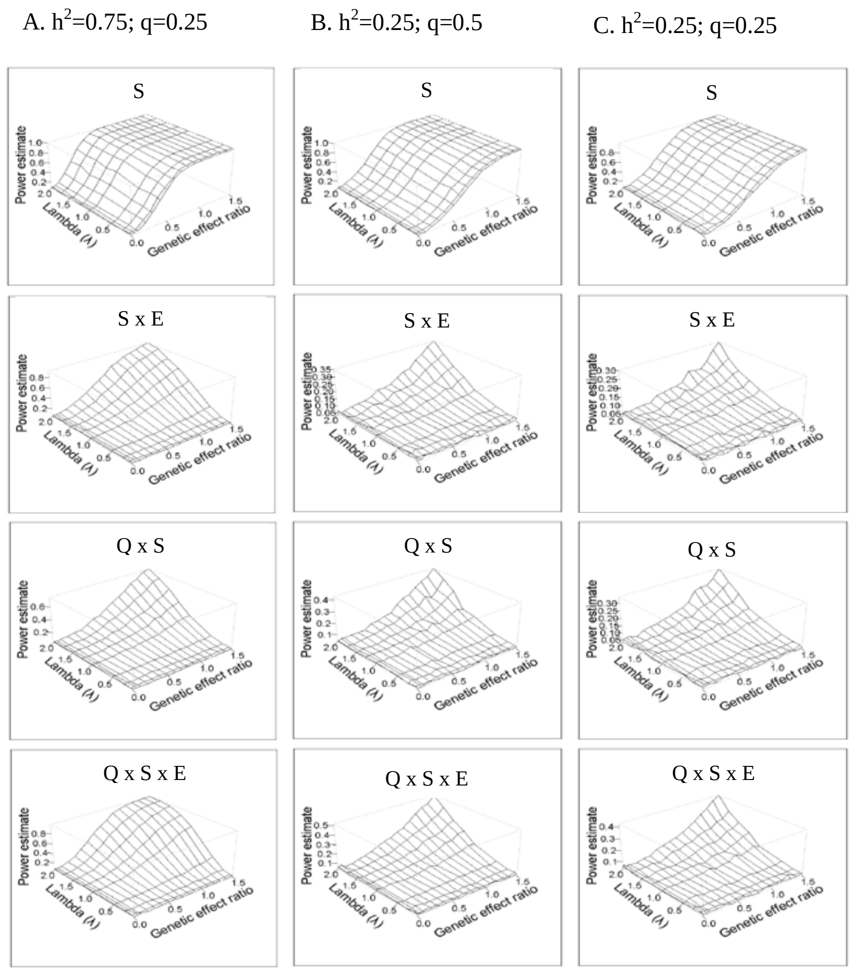
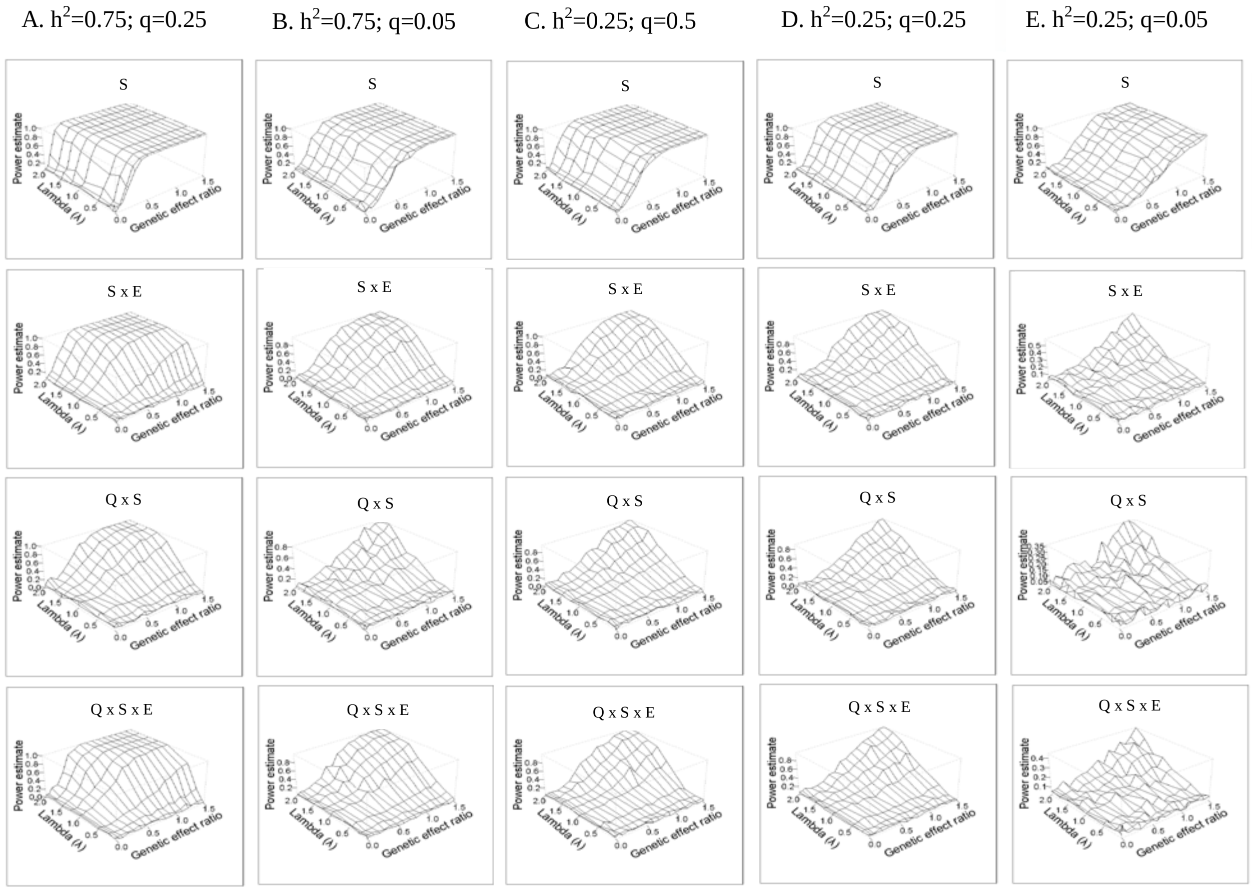


Figure S2. Power of mixed linear model to detect phenotypic effects including three way interaction. –Pearl millet sample, additional parameters levels. The data were simulated with two and three way interactions (see text, simulation scheme 3). The set of simulated effects includes SNP main effect (S), SNP by environment interaction (S x E), ancestry by SNP interaction (Q x S) and three way interaction between ancestry, SNP, and environment (Q x S x E). The proportion of simulations for which the effects were significantly detected by the statistical model is given, according to heritability (h^2), allele frequency (q), genetic effect ratio (r) and a parameter modulating effect across environment (λ).



A. h^2=0.75; q=0.25 B. h^2=0.75; q=0.05 C. h^2=0.25; q=0.5 D. h^2=0.25; q=0.25 E. h^2=0.25; q=0.05

Figure S3. Power of mixed linear model to detect phenotypic effects including three way interaction. –Maize sample, additional parameters levels. The data were simulated considering SNP main effect (S), SNP by environment interaction (S x E), ancestry by SNP interaction (Q x S) and three way interaction between ancestry, SNP, and environment (Q x S x E). The proportion of simulations in which each effect was detected by the model is given, according to heritability (h^2), allele frequency (q), genetic effect ratio (r) and parameter modulating the effects across environment (λ).

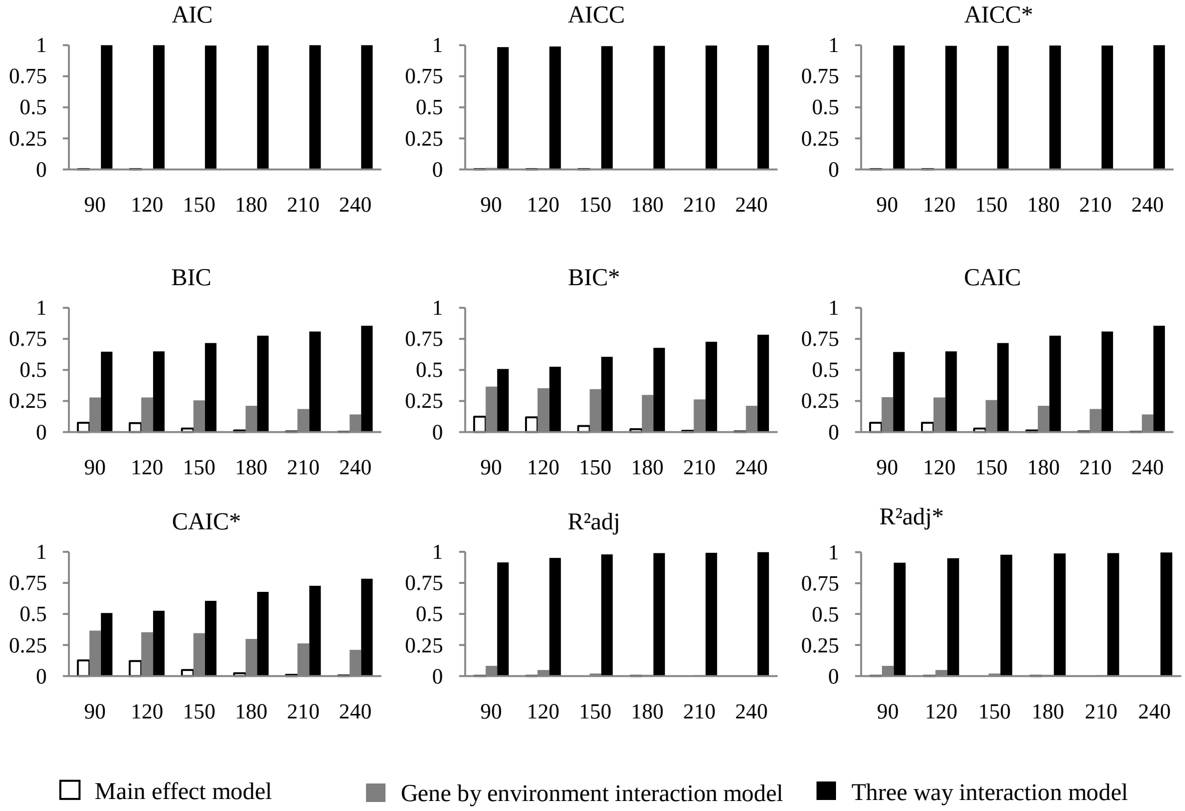


Figure S4. Frequency of selection of competing models as a function of sample size. Datasets with three way interaction were generated (see text). In each run, three competing models were fitted to the simulated dataset: the simple main effect model, the gene by environment interaction model and the three way interaction model. Five criteria for model selection were respectively used to compare the competing models: AIC, AICC, BIC, CAIC and R^2_{adj} (two variants are considered for all criteria but AIC, see text for details). We presented in this figure, for each criterion, the frequency of selection of each model over the total number of simulations. The data were based on the the maize inbred lines panel and used different sampling size (from n=90 to 240). Inbred lines were randomly sampled among the 277 inbred constituting the whole panel. The detection of the right model increased with sample size, particularly for BIC, BIC*, CAIC and CAIC*. This underlined the importance of sample size for the performance of model selection under REML.

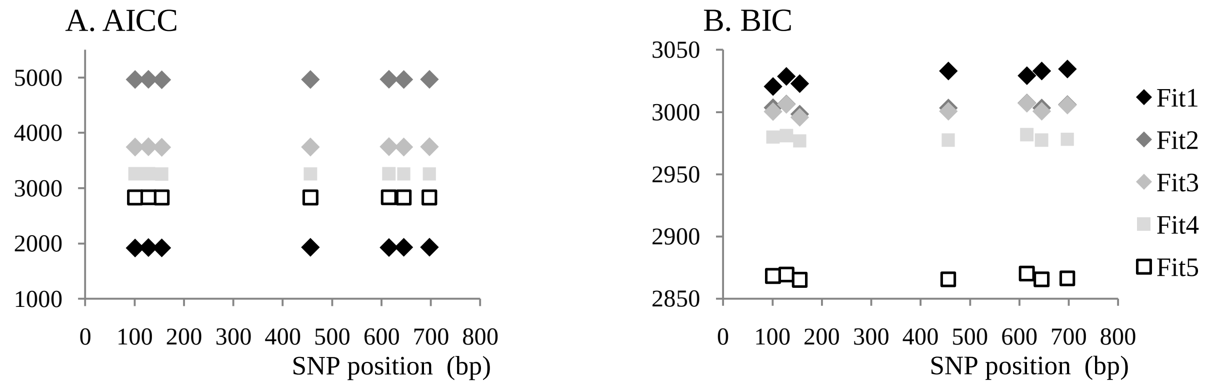


Figure S5. Model selection for the analysis of associations in pearl millet multitrial design. A set of five nested models (named Fit1 to Fit5) were specified to fit flowering time scored on 90 pearl millet inbreds in 9 field trials. Fit1 is the full model with all terms up to the three way interaction and Fit5 is the simple model with only main effects; the remaining models (Fit2 to Fit4) are intermediate specifications derived by removing terms from the full model (see text). The fixed main effects in all models are the effects of environment (trials), ancestry and SNP. Each of the seven SNPs considered at *PHYC* gene was analyzed independently. We compared the models using AICC and BIC (for both criteria, lower value is better).

SUPPLEMENTARY TABLES

Table S1. Numerical validation of the simulation code

		Linear regression between expected and estimated effects	
Effect	Expected effect size	Pearl millet sample	Maize sample
Simulation scheme 2			
S	$(1 + ½\ \lambda)\ r\ \sigma_G$	y = 0.9995x; R^2 = 0.9990	y = 0.9993x; R^2 = 0.9998
SxE	$\lambda\ r\ \sigma_G$	y = 0.9997x; R^2 = 0.9982	y = 1.0046x; R^2 = 0.9993
Simulation scheme 3			
S	$(1 + ½\ q_0\ \lambda)\ r\ \sigma_G$	y = 1.0029x; R^2 = 0.9992	y = 1.0011x; R^2 = 0.9999
SxE	$q_0\ \lambda\ r\ \sigma_G$	y = 0.9856x; R^2 = 0.9950	y = 1.0016x; R^2 = 0.9995
QxS	$½\ q_0\ \lambda\ r\ \sigma_G$	y = 1.0029x; R^2 = 0.9992	y = 1.0011x; R^2 = 0.9999
QxSxE	$q_0\ \lambda\ r\ \sigma_G$	y = 1.0088x; R^2 = 0.9951	y = 1.0016x; R^2 = 0.9995

The expected size of each effect was assessed with respect to simulation parameters. The considered effects in simulation scheme 2 were SNP main effect (S) and SNP by environment interaction (SxE). In simulation scheme 3, interaction between ancestry and SNP (QxS), and three way interaction between ancestry, SNP and environment (QxSxE) were also considered. For each effect, an average effect estimate (y) was calculated from fitted data over 100 iterations. These estimates were compared to the expectation (x). The results are presented for pearl millet and maize panels respectively. The regression line and the squared correlation coefficient (R^2) are given, based on 50 data points for each effect (see text for details). The result is consistent and validated the coding step.

Table S2. Minimum effect detected with a power of 95% in the model with gene by environment interaction.

Parameters			Effects	
Sample	h^2	q	S	S x E
Pearl millet (n=90)	0.75	0.5	4.61	4.44
		0.25	5.12	5.46
		0.05	12.29	>10.25a
	0.25	0.5	7.68	>10.25b
		0.25	9.10	>10.25c
		0.05	>15.37d	>10.25e
Maize (n=277)	0.75	0.5	3.27	3.05
		0.25	3.58	3.92
		0.05	6.87	9.6
	0.25	0.5	5.23	10.46
		0.25	6.54	11.34
		0.05	12.75	>13.08 f

a: 0.904 b: 0.803 c: 0.705 d: 0.832 e: 0.274 f:0.55

Effects on the number of days to flowering were simulated. For each effect, the minimum effect size detected in 95% of iterations is given. When the maximum effect simulated for a given term was detected in less than 95% of iterations, a footnote gives the actual power reached for the considered effect size.

Table S3. Expectation of minimum effect detected with a power of 95% in the model with three way interaction.

Parameters			Effects			
Sample	h^2	q	S	S x E	Q x S	Q x S x E
Pearl millet (n=90)	0.75	0.5	4.67	4.97	>2.87aa	3.45
		0.25	5.1	>5.74bb	>2.87cc	4.59
	0.25	0.5	7.65	>5.74dd	>2.87ee	>5.74ff
		0.25	8.94	>5.74gg	>2.87hh	>5.74ii
Maize (n=277)	0.75	0.5	2.94	3.2	3.2	3.85
		0.25	3.53	3.74	3.85	4.27
		0.05	6.83	9.4	>5.61jj	11.11
	0.25	0.5	5.36	9.61	6.41	11.22
		0.25	6.48	10.25	>6.41kk	12.82
		0.05	13.03	>12.82ll	>6.41mm	>12.82nn

aa: 0.921 bb: 0.924 cc: 0.744 dd: 0.353 ee: 0.429 ff: 0.52 gg: 0.302 hh: 0.344 ii: 0.428 jj: 0.82 kk: 0.9 ll: 0.59 mm: 0.32 nn: 0.45

Effects on the number of days to flowering were simulated. For each effect, the minimum size detected in 95% of iterations is given. When the maximum effect simulated for a given term was detected in less than 95% of iterations, a footnote gives the actual power reached for the considered effect size.

Table S4. Rate of selection of competing models by information criteria, at different values of effect size parameter r.

Table S4 is available as Excel file.

Table S5. Significance of the effect of data parameters on the rate of success of information criteria

Criterion	Intercept	Sample size	σ_G	q_0
AIC	0.073	0.75	4.42×10^{-08}	0.74
AICC	3.52×10^{-07}	2.4×10^{-46}	8.92×10^{-20}	0.17
AICC*	8.5×10^{-4}	1.38×10^{-10}	7.28×10^{-13}	0.45
BIC	1.59×10^{-38}	$<10^{-324}$	3.81×10^{-44}	2.70×10^{-08}
BIC*	8.68×10^{-50}	$<10^{-324}$	5.28×10^{-42}	1.39×10^{-07}
CAIC	2.86×10^{-38}	$<10^{-324}$	9.71×10^{-44}	4.79×10^{-08}
CAIC*	1.45×10^{-49}	$<10^{-324}$	9.67×10^{-42}	1.89×10^{-07}
R^2_{adj}	3.68×10^{-3}	3.07×10^{-246}	2.25×10^{-3}	9.45×10^{-3}
R^2_{adj}*	2.28×10^{-3}	1.03×10^{-249}	1.36×10^{-3}	8.4×10^{-3}

Data were simulated with three way interaction (simulation scheme 3) on maize sample and fitted by 3 competing models. Subsets of different sizes were sampled (from n=90 to n=240) and 10 random sampling were performed for each sample size. The rate of success (frequency of selection of the three way interaction model) was recorded in each single run (one run consisted of 1000 iterations of data simulation and model selection using the same subset). We then analyzed the distribution of the resulting rate of success using the generalized linear model (GLM) $y \sim \mu+\alpha_i+\beta_j+\lambda_l$, were μ is the intercept, α_i is the effect of sample size level i, β_j is the effect of trait standard deviation (σ_G) in the run j, and λ_l is the effect of average ancestry (q_0) in the sampled subset l. As the rate of success y is a proportion, GLM was set with a logit link function (R, version 2.7.2). The p-value of significance of each term is given in the table, with respect to the criterion used for model selection. Significant p-values are highlighted ($p<0.05$).

Table S6. Wald test of fixed effects for pearl millet data.

Table S6 is available as Excel file. Data coding for this supplementary table is already described in Table 4.

Table S7. Wald test for *Vgt1* effect analyzed independently into each trial.

Environment code	Df	Sum of squares	Wald statistic	P-value (Chisq)
11	1	74.50	5.18	0.0229
12	1	48.69	5.38	0.0203
13	1	38.58	2.80	0.0940
14	1	37.01	5.15	0.0232
15	1	69.04	4.19	0.0406
17	1	10.79	2.28	0.1310
18	1	11.44	3.25	0.0713

The association between *Vgt1* and flowering time were assessed independently in each field trial using mixed linear model as Y_i~Q_p+S_i +(K+ e), where Q_p is the fixed effect of the ancestry in population P, S_i is fixed effect of *Vgt1* allele hold by individual i, and K modulates the variance covariance structure based on kinship matrix; e is the residual of the model. Environment code refers to the location in which each trial was performed. Combining these probabilities using Fischer's method over the 7 trials led to a Chi square probability of 7.52 x 10^{-5}, meaning the effect of *Vgt1* is significant when the independent results of all trials are combined. Df: degree of freedom for *Vgt1* effect.

Final Note: *Supplemental tables S4 and S6 of this article are available on demand as Excel files.*

Discussion finale

Méthodes d'association génotype-phénotype

De la détection des QTLs à la localisation des gènes…

Il convient de distinguer deux étapes importantes dans les études d'association phénotype-génotype :

i) *la détection* consiste à montrer qu'une zone chromosomique donnée est associée au phénotype. Cette détection est basée généralement sur un échantillon de marqueurs qui sont rarement situés au sein du gène causal responsable du phénotype. En général, les marqueurs détectés le sont car ils ont une liaison physique et un déséquilibre de liaison (DL) avec le locus causal. Ainsi, cette étape permet d'identifier les zones chromosomiques contenant un locus causal, mais permet rarement d'en dire beaucoup plus sur l'identité précise du polymorphisme causal. Dans la cartographie de liaison par exemple, les QTLs détectés sont des zones qui couvrent souvent un nombre élevé de gènes, parmi lesquels peut se trouver seulement un ou quelques gènes causaux.

ii) *la localisation* constitue une deuxième étape, qui consiste à localiser avec plus de précision les locus causaux. La finesse de cette localisation dépendra de la résolution de la méthodologie. La résolution définit la capacité à cartographier avec précision l'emplacement physique des locus causaux. L'échelle de localisation (la résolution) comprend plusieurs niveaux, depuis les QTLs étalés sur des zones très larges à la localisation du SNP (voir Mackay *et al.* 2009).

La cartographie de liaison a été longtemps la méthode de choix pour l'étude d'association phénotype-génotype chez les plantes. Malgré le succès éclatant de cette méthode dans la détection de zones chromosomiques associées au phénotype (QTLs), les espoirs pour localiser finement les gènes ou les changements nucléotidiques (SNPs) agissant directement sur le phénotype sont restés sur leur soif (Mackay *et al.* 2009). Généralement, les QTLs recouvrent un grand nombre de gènes que la résolution de cette méthodologie ne permet pas toujours de séparer. Les détections de QTLs basées sur des F2 ou des RILs sont souvent suivies d'études plus fines, avec des dispositifs comme les lignées quasi isogéniques (*NILs*), qui permettent de localiser plus finement les QTLs. Les approches de cartographie fine sont cependant très lentes (Weigel and Nordborg 2005).

La recombinaison est l'élément le plus important qui détermine la résolution dans toutes les approches de génétique d'association, qu'il s'agisse de la cartographie de liaison ou de la cartographie d'association : « *It's all about recombination*» (Myles *et al.* 2009). La recombinaison casse les blocs haplotypiques en fragments plus courts, ce qui permet de dissocier l'effet de différents loci sur la variation phénotypique. Dans le cas de la cartographie de liaison, les évènements de recombinaison utiles pour la cartographie sont uniquement ceux qui se produisent après le croisement utilisé pour développer la famille de cartographie (Mackay *et al.* 2009). Or le nombre d'haplotypes recombinants (haplotypes non parentaux) que l'on peut obtenir sur les quelques générations souvent réalisées reste relativement faible. La cartographie d'association, au contraire, bénéficie d'un très large nombre de recombinaisons accumulés au cours de l'histoire évolutive des populations. Cela vient du fait que les échantillons d'étude, dans ce cas, ne viennent pas d'un croisement récent entre deux ou quelques parents, mais il s'agit de collections d'accessions très diverses avec une histoire évolutive longue et un nombre de recombinaisons plus conséquent. Cette propriété est à l'origine de la résolution plus élevée de la cartographie d'association. Cet avantage ouvre une nouvelle perspective en termes d'identification des bases génétiques de la variation phénotypique, et permet potentiellement de remonter à l'échelle, non pas d'un fragment chromosomique large couvrant plusieurs gènes (QTL), mais à l'échelle du gène voire du nucléotide (Yu and Buckler 2006). Selon le cas, il serait donc possible de localiser directement le gène responsable du caractère, voire de localiser précisément le polymorphisme nucléotidique qui, au sein de ce gène, explique l'effet phénotypique. Un des intérêts forts de la cartographie d'association réside clairement dans ce potentiel de résolution.

Toutefois, il faut noter qu'il est difficile que le déséquilibre de liaison entre des gènes proches soit totalement éliminé, même dans le cas des populations de cartographie d'association. Le déséquilibre de liaison (DL) au sein d'une population est défini par l'équilibre des forces évolutives qui jouent sur ce DL. La recombinaison casse la liaison physique, et donc tend à réduite le DL. A l'opposé, certaines forces comme la sélection tendent à augmenter ce DL. C'est ce que l'on observe dans le cas de balayage sélectif, où un grand nombre de loci le long d'un bloc chromosomique sont maintenus en DL plus ou moins forts, sous l'effet de la sélection à un locus présent dans la zone (Olsen *et al.* 2006). La tentative de localisation précise des gènes dans le cadre de la cartographie d'association devra donc se faire en prenant en compte le déséquilibre de liaison potentiel au sein de la région génomique candidate (Brown *et al.* 2008, Ducroq *et al.* 2008, Camus-Kulandeiveilu *et al.* 2008). Sans cette attention méthodologique, des locus pourront être considérés, à tort, comme responsables d'un phénotype.

Cartographie de liaison et cartographie d'association : des différences vers la complémentarité…

Nous avons discuté la résolution de la cartographie d'association par rapport à la cartographie de liaison classique. Cette résolution est sans doute l'un des apports les plus importants de la cartographie d'association, par rapport à la cartographie de liaison.

La différence des deux méthodes, loin de les opposer, en fait finalement des outils complémentaires. Par exemple, la détection des allèles rares pose problème dans la cartographie d'association car la puissance des modèles ne permet de les détecter qu'avec des effectifs très larges (Mackay *et al.* 2009). A ce niveau, la cartographie de liaison peut être la solution, car le recours aux croisements expérimentaux permet d'obtenir artificiellement une descendance où les allèles sont en fréquence plus élevée (McMallen *et al.* 2009, Myles *et al.* 2009). Il est connu chez l'humain, par exemple, que ces allèles rares contribuent significativement à expliquer la variation de certains traits quantitatifs (Mackay *et al.* 2009). L'identification de ces allèles permettra donc de mieux expliquer la variation des traits.

La cartographie de liaison peut être aussi utilisée pour aider aussi à discerner les vrais positifs dans les résultats de la cartographie d'association (Bergelson and Roux 2010)[3]. Cette dernière méthode est sujette à l'effet de la structure génétique des populations, qui malgré les corrections peut produire un biais dans le contrôle du taux de faux positifs (voir Chapitre 1). Diverses publications ont illustré la complémentarité méthodologique entre ces approches (Brachi *et al.* 2010, Ducroq *et. al* 2008).

Des approches émergentes, comme cartographie de liaison basée sur les lignées de type MAGIC (voir Chapitre 1), sont en cours développement, et procurent des avantages et des opportunités nouvelles. Toutes ces méthodes pourront avoir des apports complémentaires, en termes par exemple d'études de propriétés sous-tendant l'effet des gènes (dominance, récessivité) ou de propriétés génétiques du trait comme l'héritabilité. Parmi les approches émergentes, on peut noter en particulier le dispositif expérimental de type *Nested Association Mapping ou* NAM, qui permet de combiner de façon originale les avantages respectifs de la cartographie de liaison et de la cartographie d'association (MacMullen *et al.* 2009).

[3] Les concepts de vrais et faux positifs, les facteurs qui les déterminent et la façon de les prendre en compte dans la cartographie d'association, sont expliqués plus loin dans ce document.

Nested Association Mapping : une approche émergente

Une des limites de la cartographie de liaison classique c'est qu'elle était souvent basée sur des croisements biparentaux, ce qui limite la diversité exploitée. Aussi, un faible nombre de recombinaisons limitait la résolution de la cartographie. La cartographie d'association basée sur des populations apporte des réponses à ses limites, en offrant notamment du matériel avec plus de diversité, et une résolution beaucoup plus fine. Le développement de la cartographie d'association à l'échelle du génome entier est hautement prometteur mais requière un nombre de marqueurs important et pouvant exploser selon la taille du génome et la structure de déséquilibre de liaison de l'organisme étudié. Pour le maïs par exemple, plus de 1.6 millions de SNP ont été nécessaires pour une étude d'association à l'échelle du génome entier. Par ailleurs, l'effet confondant de la structure pose problème pour la limitation du taux de faux positifs, même si les méthodologies développées ont permis d'atténuer sa portée. La cartographie d'association nécessite aussi un effectif très large pour détecter des allèles à faible fréquence (Myles *et al.* 2010).

Le développement chez le maïs de la population NAM est fondé sur la volonté de combiner les avantages spécifiques de la cartographie d'association et la cartographie de liaison (McMallen *et al.* 2009). Pour créer cette première population NAM du maïs, 25 lignées parentales représentant une part importante de la diversité du maïs ont été choisies. Chacune de ces 25 lignées a été ensuite croisée avec la lignée de référence B73. Vingt cinq familles de lignées recombinantes ont découlées de ces croisements, totalisant environ 5000 RILs (en moyenne 200 RILs par famille). Une telle population permet d'exploiter à la fois les recombinaisons expérimentales obtenues suite aux croisements et les recombinaisons ancestrales capturées en utilisant plusieurs lignées parentales. Cette ressource maximise aussi le nombre d'allèles par rapport aux approches QTL classiques.

Ce dispositif a permis d'analyser, avec une puissance de détection de QTL améliorée, l'architecture génétique de la floraison chez le maïs (Buckler *et al.* 2009). La cartographie d'association à l'échelle du génome entier (1.6 millions de SNP génotypés) a aussi permis, avec cette population, d'étudier l'architecture génétique de phénotypes foliaires (Tian *et al.* 2011). Dans les deux cas, l'architecture des traits quantitatifs est dominée par un grand nombre de loci à effet individuel faible, peu d'interactions épistatiques et environnementales, et peu de pléiotropie.

Des simulations ont démontré la puissance élevée et la haute résolution du dispositif de type NAM, pour détecter et localiser les polymorphismes sous-tendant la variation phénotypique (Guo *et al.* 2010, Stich 2009). Ces études suggèrent que la puissance de ce type de dispositif serait accrue en augmentant le nombre de lignées parentales lors de la mise en place de la population. Un plus grand nombre de lignées parentales maximise le nombre de QTL dans les descendances. Cependant, pour un même nombre de parents, les schémas de croisement peuvent aussi influer sur la puissance du dispositif final (Stich *et al.* 2009). L'architecture des traits étudiés devrait aussi être prise en compte dans l'optimisation des dispositifs (Myles *et al.* 2009). Dans ce sens, la puissance semble plus forte pour des traits sous-tendus par quelques QTLs, comparés à des traits dont l'architecture génétique est définie par un grand nombre de QTLs (Stich 2009).

Le développement de ce type d'approches est très intéressant pour combler le déficit des approches d'association existantes. Par exemple, les NAM peuvent être une alternative pour détecter des allèles d'intérêt qui sont plutôt rares dans les populations naturelles ou les collections d'accessions existantes. Les fréquences alléliques peuvent être artificiellement augmentées dans ces populations de type NAM (McMallen *et al.* 2009). Les allèles à faible fréquence pourraient expliquer la part encore inexpliquée de l'héritabilité, connue en génétique humaine par le terme de *missing heritability* (Hall *et al.* 2010, Manolio *et al.* 2009). La notion de *missing heritability* renvoie au fait que les variants identifiés dans les études les plus exhaustives ne suffisent à expliquer qu'une part limitée, voire faible de l'héritabilité des traits. Des facteurs génétiques difficiles à identifier avec les approches actuelles (comme les allèles rares et les interactions) pourraient être associés à cette part inexpliquée de l'héritabilité.

Les effets non stables, ou effets *contexte-dépendants*, pourraient aussi être responsable d'une part importante de l'héritabilité et de la variation phénotypique encore inexpliquée (Mackay 2001). Ces effets incluent les interactions entre gènes (épistasie), et les interactions entre génotype et environnement (Mackay and Anholt 2007). Les études en cartographie d'associations ont essentiellement mis l'accent sur les effets principaux de gènes. Dans l'avenir, l'identification d'allèles impliqués dans l'interaction avec l'environnement ou dans des interactions épistatiques permettra d'une part de mieux caractériser la variation des traits quantitatifs, et d'autre part de pouvoir mieux prédire les phénotypes pour des environnements et des fonds génétiques différents (Jannink 2007).

Les méthodologies de cartographie deviennent de plus en plus sophistiquées. En parallèle, le développement rapide des technologies de séquençage accélère aujourd'hui la possibilité de disposer d'une densité de marqueurs assez large pour couvrir le génome. Cela est prometteur pour l'identification des allèles responsables du phénotype, à l'échelle du génome entier. Ces progrès sont une opportunité pour une connaissance plus poussée de l'architecture génétique des caractères quantitatifs (Mackay 2001, Buckler *et al.* 2009, Tian *et al.* 2011). L'étude de l'architecture des traits pose plusieurs questions, notamment : quelles sont les allèles impliqués dans la variation ? Quelle est la taille des effets phénotypiques de ces allèles ? Quelle est la part relative de variation expliquée par l'effet additif des allèles, par leur interaction épistasique, et par l'interaction de ces allèles avec l'environnement ? Quels sont les traits liés à un même gène et comment la pléiotropie se décline-t-elle à l'échelle fine du gène? Ces questions sont importantes pour comprendre la variation des traits complexes et leur évolution.

Quelles perspectives pour la gestion de l'impact du changement climatique chez le mil ?

Le mil est une céréale d'intérêt agricole majeur, notamment dans les zones semi-arides de l'Afrique et de l'Inde. Le fort potentiel adaptatif de cette espèce a permis sa culture dans des conditions extensives, avec peu d'intrants, et une pluviosité erratique et limitée. La sécurité alimentaire des pays sahéliens repose essentiellement sur cette espèce, à coté d'autre céréales comme le sorgho. Le climat sahélien, sec et difficilement prévisible, rend les conditions de culture très incertaines. Cette incertitude est grandissante au regard des tendances climatiques à l'échelle globale de la planète, et à l'échelle régionale sahélienne en particulier. Les conditions de production futures méritent dans ce contexte une attention importante.

La compréhension des facteurs biologiques qui déterminent l'adaptation du mil au climat permettra de mieux exploiter sa diversité naturelle pour gérer efficacement les effets du changement climatique. L'ajustement de la date de floraison est un des traits adaptatifs clefs permettant l'adaptation à différentes conditions climatiques, chez le mil comme chez d'autres espèces. Nous nous sommes donc intéressés à l'étude des bases génétique de ce trait.

Les résultats cumulés au cours de cette thèse identifient le gène *PHYC* (ou, selon les limites de l'étude, un locus très fortement lié à ce gène) comme un gène prometteur, expliquant une partie de la variation de la date de floraison. Ce gène semble avoir été, au cours de l'histoire évolutive du mil, une des cibles de la sélection. Le polymorphisme de ce gène est fortement associé à la précocité de la floraison. Par ailleurs, il a été montré une augmentation de fréquence de l'haplotype précoce de *PHYC* au sein des variétés locales nigériennes, au cours des 30 dernières années du 20^{e} siècle (Annexe 2, Vigouroux *et al. Soumis*). Ces années ont été caractérisées par une sécheresse persistante au Sahel, et ont entrainé un changement adaptatif rapide des variétés. Ce changement adaptatif est faible, mais significatif. Il a été marqué notamment par le raccourcissement global du cycle de floraison (Annexe 2). Le changement de fréquence du gène *PHYC* semble indiquer que ce gène a très probablement été impliqué dans les changements adaptatifs qui se sont produits au sein de ces populations de mil en réponse au changement climatique. L'hypothèse d'un rôle adaptatif a également été développée pour le gène *MADS11*(Annexe 1). Ce gène, également associé à la précocité de floraison, est corrélé avec le gradient climatique le long de l'aire de distribution du mil cultivé au Niger. La distribution allélique de *MADS11* suivant un gradient pluviométrique, et son association avec la date de floraison, suggèrent que ce gène contribue à l'adaptation des variétés aux conditions locales à travers l'ajustement de la précocité.

La plupart des études de diversité chez le mil avaient, jusque-là, examinées la diversité neutre, non sujette à la sélection (Mariac *et al.* 2006a, Mariac *et al.* 2006b, Allinne *et al.* 2008, Stich *et al.* 2010, etc). Nos résultats suggèrent que la caractérisation, dans les populations cultivées, de la diversité des gènes associés au phénotype et de leur dynamique, pourrait aider à la gestion de la diversité *in situ*. Ce type de connaissance pourrait, en pratique, aider à la gestion et au choix variétal à l'échelle spatiale et temporelle, en rapport avec les conditions climatiques.

Par ailleurs, la sélection assistée par marqueurs (SAM) pourrait valoriser de tels résultats, en exploitant les marqueurs moléculaires liés aux gènes identifiés. L'accumulation de plusieurs allèles favorables (ou *gene pyramiding*) sera possible une fois qu'un nombre critique de gènes seraient identifiés, pour renforcer la sélection sur la date de la floraison. D'autres gènes seront certainement identifiés dans l'avenir, grâce notamment aux efforts de développement de nouvelles données de séquence sur le mil (Y Vigouroux *Com. pers.*).

Il est important de notifier que les gènes identifiés dans cette étude sont associés à des effets pléiotropiques. La pléiotropie est déterminante dans le rôle que peut jouer un gène au cours de l'évolution adaptative (Roux *et al.* 2006). Chez le mil, on considère en général que la précocité de floraison est associée à une limitation de rendement. Cette limitation de rendement peut être vue comme un compromis nécessaire en conditions sèches, car elle permet de boucler le cycle de la plante dans le délai court autorisé par la saison pluvieuse, et d'éviter ainsi le stress de fin de cycle (Do et Winkel 1993). Le stress de fin de cycle a un effet drastique sur les rendements et peut conduire à la perte totale des plantes à cycle plus tardif (Eldin 1993). La précocité permet dans ce cas de sécuriser un minimum de production.

Le développement chez le mil de la méthodologie d'association présenté dans cette thèse est une démarche aujourd'hui assez originale. Les outils développés et rendus disponibles pourraient permettre d'étudier davantage de gènes et de caractères d'intérêt chez le mil, espèce restée en marge des travaux pointus sur les bases génétiques des caractères quantitatifs. Un des avantages du panel est qu'il est constitué d'un matériel végétal fixé (lignées), pour lequel les données génétiques développées sont donc réutilisables pour d'autres études ou d'autres caractères phénotypiques. Les travaux futurs dans le cadre de la cartographie d’association, de cartographie de liaison ou d’autres approches émergentes aideront à disséquer de façon plus exhaustive l’architecture génétique de la date de floraison chez le mil. Ces développements nécessiteront un effort pour la production de ressources génomiques plus larges chez cette espèce.

Références bibliographiques

AGRHYMET et CIRAD (2005). Centre régional AGRHYMET (CILSS) & Centre de coopération internationale en recherche agronomique pour le développement. Après la famine au Niger…Quelles actions de lutte et de recherche contre l'insécurité alimentaire au Sahel ? Dossier de presse, décembre 2005, 41 pp.

Akaike H, 1974. A new look at the statistical model identification. IEEE transactions on automatic control. AC 19: 716-723.

Allinne C, C Mariac, Y Vigouroux, G Bezançon, E Couturon, *et al.* (2008). Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. Genetica 133: 167-178. .

Allouis S, X Qi, S Lindup, MD Gale and KM Devos (2001). Construction of a BAC library of pearl millet, *Pennisetum glaucum*. Theor Appl Genet 102: 120–125.

ASReml package for R (ASReml-R), version 20/32. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Atkinson B, and T Therneau (2008). Kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package, versions 1.1.0-21. http: //cran.r-project.org.

Atwell S, YS Huang, BJ Vilhjalmsson, G Willems, M Horton, *et al.* (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465: 627-631.

BALASUBRAMANIAN, S, S. SURESHKUMAR, M. AGRAWAL, T.P. MICHAEL, C. WESSINGER, *et al.* (2006. The phytochrome C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. Nat Genetics 38: 711-715. .

Becker A, and Theissen G (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Molecular Phylogenetics and Evolution 29: 464-489.

Bergelson J and F Roux (2010). Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nature Reviews 11: 867-879.

Bernardo R, A Murigneux, and Z Karaman (1996). Marker-based estimates of identity by descent and alikeness in state among maize inbreds. Theor Appl Genet 93: 262–267.

Bezançon G, JF Renno, and KA Kumar (1994). Le mil. In: L'amélioration des plantes tropicales. Edition CIRAD & OSRTOM. La librairie du CIRAD, Montpellier France, 1994; 457-481. ISSN 1251-7224.

Bezançon G et JL Pham (2004). Ressources génétiques des mils en Afrique de l'Ouest. Diversité, conservation et valorisation. Editions de l'IRD (2004).

Bezançon G, JL Pham, M Deu, Y Vigouroux, F Sagnard, *et al.* (2009). Changes in the diversity and geographic distribution of cultivated millet (*Pennisetum glaucum* (L.) R. Br.) and sorghum (*Sorghum bicolor* (L.) Moench) varieties in Niger between 1976 and 2003. Genet Resour Crop Evol 56: 223–236.

Biasutti M, and AH Sobel (2009). Delayed Sahel rainfall and global seasonal cycle in a warmer climate. Geophysical Research Letters, doi: 10.1029/2009GL041303.

Blázquez M, M Koornneef, and J Putterill (2001). Flowering on time: genes that regulate the floral transition. EMBO reports 2 (12): 1078-1082.

Bozdogan H, 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52: 345-370.

Brachi B, N Faure, M Horton, E Flahauw, A Vazquez, M Nordborg, J Bergelson, J Cuguen, and F Roux (2010). Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. PLoS Genet 6(5): e1000940. doi: 10.1371/journal.pgen.1000940.

Bradbury PJ, Z Zhang, DE Kroon, TM Casstevens, Y Ramdoss, and ES Buckler (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 23: 2633–2635.

Brown PJ, WL Rooney, C Franks, and S Kresovich (2008). Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics 180: 629–637.

Buckler, E, P. Bradbury, D. Kroon, Y. Ramdoss, T. Casstevens, *et al.* (2007. Trait Analysis by Association, Evolution and Linkage (TASSEL). Version 2.0.1. www.maizegenetics.net/tassel.

Buckler ES, James BH, PJ Bradbury, CB Acharya, PJ Brown, *et al.* (2009).The genetic architecture of maize flowering time. Science 325: 714-718.

Caicedo AL, JR Stinchcombe, KM Olsen, J Schmitt and MD Purugganan (2004). Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. PNAS 101: 15670-15675.

Camus-kulandaivelu L, JB Veyrieras, D Madur, V Combes, M Fourmann, *et al.* (2006). Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172: 2459-2463. .

Camus-Kulandaivelu L, LM Chevin, C Tollon-Cordet, A Charcosset, D Manicacci and MI Tenaillon (2008). Patterns of molecular evolution associated with two selective sweeps in the Tb1-Dwarf8 region in maize. Genetics 180: 1107-1121.

Camus-Kulandaivelu (2007). Evolution génomique du maïs durant son adaptation aux conditions européennes. Thèse de doctorat en génétique végétale. UMR 8120 (Gif-sur-Yvette, France).

Casa,AM, G Pressoir, PJ Brown, SE Mitchell, WL Rooney, *et al.* (2008). Community ressources and strategies for association mapping in sorghum. Crop Sci 48: 30-40.

CBSU, Cornell university. http: //cbsuapps.tc.cornell.edu/index.aspx.

Chatfield C, 1995. Model uncertainty, data mining and statistical inference. J. R. Stat Soc A 158: 419-466.

Childs KL, FR Miller, MM Cordonnier-Pratt, LH Pratt, PW Morgan, *et al.* 1997 The sorghum photoperiod sensitivity gene, Ma3, encodes a phytochrome B. Plant Physiol 113: 611-619.

Clerget B, BIG Haussmann, SS Boureima, and E Weltzien (2007). Surprising flowering response to photoperiod: Preliminary characterization of West and Central African pearl millet germplasms. SAT eJournal 5 (1).

Crawley MJ (2007). The R book. John Wiley & Sons, Ltd; The Atrium, Southern Gate, Chichester,West Sussex PO19 8SQ. England. 942 p.

Dai A (2011). Drought under global warming: a review. Wiley Interdisciplinary Reviews: Climate Change 2 (1): 45-65.

D'andrea AC and J. Casey (2002). Pearl millet and Kintampo subsistence. Afr Archaeol Rev 19: 147-173.

D'andrea AC, M. Klee and J. Casey (2001). Archaeological evidence for pearl millet (*Pennisetum glaucum*) in sub-saharan West Africa. Antiquity 75: 341-348.

Do F et T Winkel (1993). Mécanismes morpho-physiologiques de résistance du mil a la sécheresse. Intérêt d'une approche agrophysiologique et résultats expérimentaux. *In*: Le mil en Afrique. Diversité génétique et agro-physiologique: potentialités et contraintes pour l'amélioration génétique et l'agriculture. Editeur scientifique: S Hamon. Editions de l'ORSTOM, Paris 1993.

Doebley, J, A. Stec and L. Hubbard (1997). The evolution of apical dominance in maize. Nature 386: 485-488.

Dore MHI (2005). Climate change and changes in global precipitation patterns: What do we know? Environment International 31(8): 1167-1181.

Drummond AJ, B Ashton, S Buxton, M Cheung, A Cooper, *et al.* (2010). GENEIOUS v4.8.5. http: //www.geneious.com.

Ducrocq S, D Madur, JB Veyrieras, L Camus-Kulandaivelu, M Kloiber-Maitz, *et al.* (2008). Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178: 2433-2437.

Eldin M (1993). Analyse de l'effet des déficits hydriques sur la récolte du mil au Niger: Conséquences agronomiques. *In*: Le mil en Afrique. Diversité génétique et agro-physiologique: potentialités et contraintes pour l'amélioration génétique et l'agriculture. Editeur scientifique: S Hamon. Editions de l'ORSTOM, Paris 1993.

El-Lithy M, L Bentsink, CJ Hanhart, GJ Ruys, D Rovito, *et al.* (2006). New Arabidopsis recombinant inbred line populations genotyped using SNPwave and their use for mapping flowering-time quantitative trait loci. Genetics 172: 1867- 1876.

Evanno, G, S. Regnaut and J. Goudet (2005). Detecting the number of clusters of individuals using the software Structure: a simulation study. Mol. Ecol. 14: 2611-2620.

Falconer D.S. and T.F.C. Mackay, 1996. Introduction to Quantitative Genetics. Ed. 4, Addison-Wesley Longman, Harlow, UK.

Falush D, M Stephens and JK Pritchard (2003). Inference of population structure using multilocus genoype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Falush D, M Stephens, and JK Pritchard (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7(4): 574-578.

FAO (2011). Food and Agriculture Organization of the United Nations. FAOSTAT: http: //faostat.fao.org/ (accès en ligne: 27 janvier 2011).

Faraway J (2006). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Chapman & Hall, Boca Raton, FL.

Filiault DL, CA Wessinger, JR Dinneny, J Lutes, JO Borevitz, *et al.* (2008). Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. Proc. Natl. Acad. Sci. USA 105: 3157-3162.

Flint J and TFC Mackay (2009). Genetic architecture of quantitative traits in flies, mice and humans. Genome Res 19: 723-733.

Flint-Garcia SA, AC Thuillet, JM Yu, G Pressoir, SM Romero, *et al.* (2005). Maize association population: a high resolution platform for QTL dissection. Plant Journal 44: 1054-1064.

Foster KR, FR Miller, KL Childs, and PW Morgan (1994). Genetic regulation of development in *Sorghum bicolor*. Plant Physiol. 105: 941-948.

Franklin, KA, SJ DaviS, WM SToddart, RD Vierstra, and GC Whitelam (2003). Mutant analyses definemultiple roles for phytochrome C in arabidopsis photomorphogenesis. Plant Cell 15: 1981-1989.

Fritz GL (1995). New dates and data on early agriculture: the legacy of complex hunter-gatherers. Ann. Missouri Bot. Gard. 82: 3-15.

Fu YX and WH Li (1993). Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Millot G (2009). Comprendre et réaliser les tests statistiques à l'aide de R. Manuel pours les débutants. Editions De Boeck. 704 pages

Gao H, S Williamson, and CD Bustamante (2007). An MCMC approach for the joint inference of population structure and inbreeding rate from multi-locus genotype data. Genetics: 176: 1635-1651. .

GIEC (2007). Climate change 2007: impacts, adaptation and vulnerability. Summary for policy makers. Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change. www.ipcc.ch.

Gilmour AR, BJ Gogel, BR Cullis, and R Thompson (2006). ASReml User Guide Release 20. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

GRAMENE release 32. http: //www.gramene.org/ (access online: November 2010).

Guo B, DA Sleper, and WD Beavis (2010). Nested association mapping for identification of functional markers. Genetics 186: 373–383.

Gurka M J (2006a). Selecting the best linear mixed model under REML. The American Statistician, 60(1): 19-26.

Gurka MJ (2006b). Extending the Box-Cox transformation to the linear mixed model. J. R. Statist. Soc. A 169 (2): 273-288.

Hall D, C Tegström, and P K Ingvarsson (2010). Using association mapping to dissect the genetic basis of complex traits in plants. Briefings in Functional Genomics 9 (2): 157-165. .

Hansen J, MS Makiko, R Reto, L Ken, WL David, and ME Martin (2006). Global temperature change. PNAS 103: 14288-14293.

Hardy OJ and X Vekemans (2002). SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Mol Ecol Notes 2: 618-620.

Hardy O.J (2003). Estimation of pairwise relatedness between individuals and characterisation of isolation by distance processes using dominant genetic markers. Molecular Ecology 12: 1577-1588.

Hardy O and X Vekemans (2007). SPAGeDi 1.2: a program for spatial pattern analysis of genetic diversity. User's manual. Université Libre de Bruxelles. Bruxelles, Belgium.

HAUSSMANN, BIG, A BOUBACAR A, SS BOUREIMA SS, and Y VIGOUROUX (2006). Multiplication and preliminary characterization of West and Central African pearl millet landraces. Int. Sorghum and Millet Newsletter 47: 110-112.

Hayama R, S Yokoi, S Tamaki, M Yano, and K Shimamoto (2003). Adaptation of photoperiodic control pathways produces short-day flowering in rice. Nature 422: 719 - 722.

Hayes JF, and WG Hill (1981). Modification of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics 37: 483-493.

Hecht V, F Foucher, C Ferrándiz, R Macknight, C Navarro, *et al.* (2005). Conservation of Arabidopsis flowering genes in model legumes. Plant Physiology 137: 1420-1434.

Howden SM, JF Soussana, FN Tubiello, N Chhetri, M Dunlop and H Meinke (2007). Climate change and food security special feature: adapting agriculture to climate change. PNAS 104 (50): 19691-19696.

Hubisz MJ, D Falush, M Stephens and JK Pritchard (2009). Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9, 1322-1332.

Huelsenbeck JP and P Andolfatto (2007). Inference of population structure under a Dirichlet process model. Genetics. 175:1787-1802.

Hurvich CM and CL Tsai (1989). Regression and time series model selection in small samples. Biometrika, 76: 297-307.

ICRISAT and FAO (1996). International Crops Research Institute for the Semi-Arid Tropics; Food and Agriculture Organization of the United Nations. The world sorghum and millet economies: facts, trends and outlook.

Ingvarsson PK, MV Garcia, V Luquez, D Hall, and S Jansson (2008). Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula,* Salicaceae). Genetics 178: 2217-2226.

Jaenicke-Després V, ES Buckler, BD Smith, MT Gilbert, A Cooper, *et al.* (2003). Early allelic selection in maize as revealed by ancient DNA. Science 302: 1206-1208.

Jannink JL (2007). Identifying quantitative trait locus by genetic background interactions in association studies. Genetics 176: 553-561.

Jannink JL. (2008). QTL by genetic background interaction: predicting inbred progeny value. Euphytica 161: 61-69.

Kang HM, NA Zaitlen, CM Wade, A Kirby, D Heckerman, *et al.* (2008). Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Kim S, V Plagnol, TT Hu, C Toomajian, RM Clarck, *et al.* (2007). Recombinaison and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151-1155. .

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, *et al.* (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5(7): e1000551.

Larousse Agricole (2002). Direction: Marcel Mazoyer. Edition Larousse, Montréal, Québec, 2002; p. 419-420.

Leng C (2008). The Residual information criterion, corrected. ARXIV, Bibliographic code: 2007arXiv0711.1918.

Lewin A (2008). R package: Exact tests for linkage disequilibrium and Hardy-Weinberg equilibrium. http: //www.r-project.org.

Li L, MJ Paulo and F van Eeuwijk (2010). Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theor Appl Genet: DOI: 10.1007/s00122-010-1389.

Li Y, S Bhosale, BIG Haussmann, B Stich, AE Melchinger, and HK Parzies (2010). Genetic diversity and linkage disequilibrium of two homologous genes to maize D8: sorghum SbD8 and pearl millet PgD8.

Lobell DB, MB Burke, C Tebaldi, MD Mastrandrea, WP Falcon and RL Naylor (2008). Prioritizing climate change adaptation needs for food security in 2030. Science 319: 607-610.

Loiselle BA, VL Sork, J Nason, and C Graham (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. Journal of Botany 82: 1420-1425.

Mackay TFC (2001). The genetic architecture of quantitative traits. Annu Rev Genet 35: 303-339.

Mackay TFC and RRH Anholt (2007). Ain't misbehavin'? Genotype-environment interactions and the genetics of behavior. Trends Genet: 23: 311-314.

MacKay TFC, E Stone, and J Ayroles (2009). The genetics of quantitative traits: Challenges and prospects. Nature Reviews Genetics 10: 565-577.

Manolio TA, FS Collins, NJ Cox, *et al.* (2009). Finding the missing heritability of complex diseases. Nature 461: 747-53.

Mariac C, V Luong, I Kapran, A Mamadou, F Sagnard, *et al.* (2006). Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L] R. Br.) in Niger assessed by microsatellite markers. Theor Appl Genet 114: 49-58.

Mariac C, Jehin L, AA Saïdou, AC Thuillet, M Couderc, S Sire, H Jugdé, H Adam, G Bezançon, JL Pham, and Y Vigouroux (2011). Genetic basis of pearl millet population adaptation along an environmental gradient investigated by a combination of genome scan and association mapping. Mol Ecol 20: 81-91.

Mariac C, T Robert, C Allinne, MS Remigereau, A Luxereau, *et al.* (2006b). Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. Theor Appl Genet 113: 1003-1014.

Mathews S, R C. Tsai, and EA Kellogg (2000). Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. Am. J. Bot. 87: 96-107.

McMullen M D, S Kresovich, HS Villeda, P Bradbury, Huihui Li, *et al.* (2009). Genetic properties of the maize nested association mapping population. Science 325: 737: 740.

McMullen M D, S Kresovich, H S Villeda, P Bradbury, Huihui Li, *et al.* (2009). Genetic properties of the maize nested association mapping population. Science 325: 737: 740.

Monte, E, JM Alonso, JR Ecker, Y Zhang, X LI, *et al.* (2003 Isolation and characterization of phyC mutants in Arabidopsis reveals complex crosstalk between phytochrome signaling pathways. Plant Cell 15: 1962-1680.

Myles S, J Peiffer, P J Brown, ES Ersoz, Z Zhang, DE Costich, and ES Buckler (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell 21: 2194-2202.

Nordborg M and D Weigel (2008). Next-generation genetics in plants. Nature 456: 720-723.

Olsen KM, AL Caicedo, N Polato, A McClung, S McCouch, and MD Purugganan (2006). Selection under domestication: evidence for a sweep in the rice Waxy genomic region. Genetics 173: 975-983.

Oumar I, C Mariac, JL Pham, and Y Vigouroux (2008). Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L] R. Br.) as revealed by microsatellite loci. Theor Appl Genet 117: 489-497.

Patterson N, A L Price and DReich (2006). Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi: 10.1371/journal.pgen.0020190.

Phillips PC (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9: 855-867.

Poncet V (1998). Organisation génétique du syndrome de domestication du mil (*Pennisteum glaucum*, Poacea). Thèse de l'université de Paris-Sud, Orsay. 116 p.

Posada D and T R Buckley (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53 (5): 793-808.

Price AL, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

Pritchard JK, M Stephens, and P Donnelly (2000b). Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Pritchard JK, M Stephens, NA Rosenberg, and P Donnelly (2000). Association mapping in structured populations. Am J Hum Genet 67: 170-181 .

Putterill J, R Laurie, and R Macknight (2004). It's time to flower: the genetic control of flowering time. BioEssays 26 (4): 363–373.

R Development Core Team (2008). R: A language and environment for statistical computing R. Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http: //www.R-project.org.

Rafalski JA (2010). Association genetics in crop improvement. Curr Opin Plant Biol 13: 1-7.

Remington DL, JM Thornsberry, Y Matsuoka, LM Wilson, SR Whitt, *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. PNAS 98: 11479-11484.

Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res 67: 175-185.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. Camb. 67: 175-185. .

Roche Applied Science. 454 Sequencing. http: //454com/.

Roux F, P Touzet, J Cuguen, and V Le Corre (2006). How to be early flowering: an evolutionary perspective. Trends Plant Sci. 11: 1360-1385.

Rozas J, JC Sanchez-Delbarrio, X Messeguer, and R Rozas (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

Saïdou AA, C Mariac, V Luong, JL Pham, G Bezancon, and Y Vigouroux (2009). Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. Genetics 182: 899-910.

Sambandan D, MA Carbone, RRH Anholt, and TFC Mackay (2008). Genetics 179(2): 1079-1088

Samis KE, KD Heath, and JR Stinchcombe (2008). Discordant longitudinal clines in flowering time and phytochrome C in *Arabidopsis thaliana*. Evolution 62: 2971-2983. .

Schabenberger O (2004). Mixed model influence diagnostics. SUGI 29 Proceedings: Paper 189-29.

Schwarz G (1978). Estimating the dimension of a model. The Annals of Statistics 6: 461-464.

Seguin B (2010). Le changement climatique: conséquences pour l'agriculture et la forêt. Rayonnement du CNRS 54: 36-47.

Serpentier G et P Milleville 1993. Les systèmes de culture paysans à base de mil (*Pennisetum glaucum*) et leur adaptation aux conditions sahéliennes. *In*: Le mil en Afrique. Diversité génétique et agro-physiologique: potentialités et contraintes pour l'amélioration génétique et l'agriculture. Editeur scientifique: S Hamon. Editions de l'ORSTOM, Paris 1993.

Shi P and C-L Tsai (2002). Regression model selection - A residual likelihood approach. J. R. Stat. Soc. B 64: 237-252.

Shin JH, S Blay, N Lewin-Koh, B McNeney and J Graham (2010). R package: Graphical display of pairwise linkage disequilibria between SNPs. http: //www.R-project.org.

Sivakumar M V K (1988). Predicting rainy season potential from the onset of rains in southern sahelian and sudanian climatic zones of West Africa. Agr and Forest meteorology 42 (4): 295-305.

Sokal RR, and FJ Rohlf (1991). Biometry. Third edition. Ed. WH Freeman and Co, New York.

Spiegelhalter DJ, NG Best, BP Carlin and A van der Linde (2002). Bayesian measures of model complexity and fit. J. R. Stat. Soc. B 64 (4): 583-639.

Stich B, J Mohring, HP Piepho, M Heckenberger, ES Buckler and AE Melchinger (2008). Comparison of mixed-model approaches for association mapping. Genetics 178: 1745-1754.

Stich B (2009). Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. Genetics 183: 1525–1534.

Stich B, BIG Haussmann, R Pasam, S Bhosale, CT Hash, *et al.* (2010). Patterns of molecular and phenotypic diversity in pearl millet [*Pennisetum glaucum* (L.) R. Br.] from West and Central Africa and their relation to geographical and environmental parameters. BMC Plant Biology 10: 216-225.

Stracke S, G Haseneyer, J-B Veyrieras, HH Geiger, Sauer S, Graner A, and H-P Piepho (2009). Association studies reveals gene action and interactions in the determination of flowering time in barley. Theor Appl Genet 118: 259-273.

Szyda J, E Grindflek, Z Liu, and S Lien (2003). Multivariate mixed inheritance models for QTL detection on porcine chromosome 6. Genet Res Camb 81: 65-73.

Tajima, F, 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Takano M, N Inagaki, X Xie, N Yuzurihara, F Hihara, *et al.* (2005). Distinct and cooperative functions of phytochromes A, B, and C in the control of deetiolation and flowering in rice. Plant cell 17: 3311-3325.

Thornsberry JM, MM Goodman, J Doebley, S Kresovich, D Nielsen and E S Buckler (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28: 286-289.

Tian F, P J Bradbury, PJ Brown, H Hung, Q Sun, *et al.* (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics. doi: 10.1038/ng.746.

Uwatoko N, A. Onishi, Y. Ikeda, M. Kontani, A. Sasaki, K. Matsubara, Y. Itoh, and Y. Sano (2008). Epistasis among the three major flowering time genes in rice: coordinate changes of photoperiod sensitivity, basic vegetative growth and optimum photoperiod. Euphytica 163: 167-175.

Verbeke G and G Molenberghs (2000). Linear mixed models for longitudinal data, New York: Springer-Verlag.

Vonesh EF, VM Chinchilli and K Pu (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics 52: 572-587.

Vos P, R Hogers, M Bleeker, M Reijans, T Van De Lee, *et al.* (1995). AFLP: a new technique for DNA fingerprinting. Nucl. Acids Res. 23: 4407-4414.

Walther GR, E Post, P Convey, A Menzel, C Parmesan, *et al.* (2010). Ecological responses to recent climate change. Nature 416: 389-395.

Wang, RL, A Stec, J Hey, L Lukens, and J Doebley (1999). The limits of selection during maize domestication. Nature 398: 236-239.

Wang J (2007). Selecting the best linear mixed model using predictive approaches. Master of Science. Brigham Young University. .

Weigel D, and M Nordborg (2005). Natural Variation in Arabidopsis. How do we find the causal genes? Plant Physiol 138: 567-568.

White GM, MT Hamblin, and S Kresovich (2004). Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution. Mol. Biol. Evol. 21: 716-723.

Yadav RS, CT Hash, FR Bidinger, GP Cavan, and CJ Howarth (2002). Quantitative trait loci associated with traits determining grain and stover yield pearl millet under terminal drought-stress conditions. Theor Appl Genet 104: 67-83.

Yadav RS, FR Bidinger, CT Hash, YP Yadav, OP Yadav, SK Bhatnagars, et J Howarth (2003). Mapping and characterisation of QTL x E interactions for traits determining grain and stover yield in pearl millet. Theor Appl Genet 106: 512-520.

Yu J, Z Zhang, C Zhu, DA Tabanao, G Pressoir, MR Tuinstra, S Kresovich, RJ Todhunter, and ES Buckler (2009). Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Gen. 2: 63-77.

Yu J and ES Buckler (2006). Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17: 155-160.

Yu J, G Pressoir, WH Briggs, IV BI, M Yamasaki, *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genetics 38: 233-208.

Zhang Z, E Ersoz, CQ Lai, RJ Todhunter, HK Tiwari, *et al.* (2010a). Mixed linear model approach adapted for genome-wide association studies. Nature Genetics, 42: 355-360.

Zhang N, A Gur, Y Gibon, R Sulpice, S Flint-Garcia, *et al.* (2010b). Genetic analysis of central carbon metabolism unveils an amino acid substitution that alters maize NAD-dependent isocitrate dehydrogenase activity. PLoS ONE 5(4): e9991. doi: 10.1371/journal.pone.0009991.

Zhao K, MJ Aranzana, S Kim, C Lister, C Shindo, *et al.* (2007). An Arabidopsis example of association mapping in structured samples. PLoS Genetics 3 (1): 0071-0082.

Zhu C, M Gore, ES Buckler, and J Yu (2008). Status and prospects of association mapping in plants. The Plant Genome 1: 5-20.

Zhu C and J Yu (2009). Nonmetric multidimensional scaling corrects for population structure in whole genome association studies. Genetics 182: 875-888.

ANNEXE 1.

Genetic basis of pearl millet population adaptation along an environmental gradient investigated by a combination of genome scan and association mapping.

MOLECULAR ECOLOGY

Molecular Ecology (2010) doi: 10.1111/j.1365-294X.2010.04893.x

Genetic basis of pearl millet adaptation along an environmental gradient investigated by a combination of genome scan and association mapping

CÉDRIC MARIAC,*† LÉA JEHIN,*† ABDOUL-AZIZ SAÏDOU,*†‡§ ANNE-CÉLINE THUILLET,* MARIE COUDERC,* PIERRE SIRE,† HÉLÈNE JUGDÉ,† HÉLÈNE ADAM,‡§ GILLES BEZANÇON,† JEAN-LOUIS PHAM* and YVES VIGOUROUX*

*Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/Sup-Agro, BP64501, 34394 Montpellier, Cedex 5, France, †Institut de Recherche pour le Développement, BP 11416, Niamey, Niger, ‡Institut de Recherche pour le Développement, UMR DIAPC IRD/INRA/Université de Montpellier II/Sup-Agro, BP11416, Niamey, Niger, §University Abdou Moumouni, BP 11040, Niamey, Niger

Abstract

Identifying the molecular bases of adaptation is a key issue in evolutionary biology. Genome scan is an efficient approach for identifying important molecular variation involved in adaptation. Association mapping also offers an opportunity to gain insight into genotype–phenotype relationships. Using these two approaches coupled with environmental data should help to come up with a refined picture of the evolutionary process underlying adaptation. In this study, we first conducted a selection scan analysis on a transcription factor gene family. We focused on the MADS-box gene family, a gene family which plays a crucial role in vegetative and flower development. Twenty-one pearl millet populations were sampled along an environmental gradient in West Africa. We identified one gene, i.e. *PgMADS11*, using Bayesian analysis to detect selection signatures. Polymorphism at this gene was also associated with flowering time variation in an association mapping framework. Finally, we found that *PgMADS11* allele frequencies were closely associated with annual rainfall. Overall, we determined an efficient way to detect functional polymorphisms associated with climate variation in non-model plants by combining genome scan and association mapping. These results should help monitor the impact of recent climatic changes on plant adaptation.

Keywords: adaptation, flowering time, genome scan, MADS-box genes, *Pennisetum glaucum*

Received 28 April 2010; revision received 17 August 2010; accepted 22 August 2010

Introduction

Genetic changes underlying adaptation are a central topic of evolutionary biology. Adaptation could occur when populations experience new selective pressures because of an environmental change or because they have moved into a new environment. In such an adaptation process, we should expect to find selection signatures at the molecular level, a phenotypic change associated with this evolution and a link between causative environmental changes, phenotype and genotype.

Selection signatures are common at the molecular level. The story of *Drosophila melanogaster* illustrates a case of worldwide colonization, with an estimated rate of adaptive substitutions of around 10^{-11} per nucleotide site and per generation (Stephan & Li 2007). Domestication events have also been extensively addressed in terms of the effects of selection on genetic diversity (Vigouroux *et al.* 2005; Wright *et al.* 2005). A notable example concerns maize, the most recent results prove that a large genomic region of maize is involved in the response to selective pressures under domestication (Tian *et al.* 2009). These selective signatures have been linked to phenotypic variation in a number of cases. For instance, in plants, recent studies identified the genetic

Correspondence: Yves Vigouroux, Fax: 33 (0) 467416222; E-mail: yves.vigouroux@ird.fr

bases of flowering time variation associated with the spread of barley in Europe (Jones *et al.* 2008) or maize in America (Camus-Kulandaivelu *et al.* 2006, 2008; Ducrocq *et al.* 2008). In other cases, these selective signatures were associated with environmental variation. For example, in *Arabidopsis* (Balasubramanian *et al.* 2006; Samis *et al.* 2008) or *Drosophila* (Umina *et al.* 2005), a genetic cline appears to be associated with an environmental gradient. However, few studies have been carried out to correlate molecular selection signatures, phenotypic variation and potential environmental causes (but see Linnen *et al.* 2009).

There is growing interest in using natural populations for the detection of molecular selection signatures and genotype–phenotype associations. The first approach focuses on genome-wide selection scan (Vigouroux *et al.* 2002; Wright *et al.* 2005). Genome scans of genes or markers linked to genes have been performed on plant and animal models for which substantial genetic resources are available, e.g. maize (Vigouroux *et al.* 2002; Wright *et al.* 2005), *Drosophila* (Harr *et al.* 2002; Li & Stephan 2006) and human (Sabeti *et al.* 2006). For species with less genomic information, studies generally focus on random anonymous markers like AFLP (Campbell & Bernatchez 2004; Murray & Hare 2006; Manel *et al.* 2009; Meyer *et al.* 2009; Poncet *et al.* 2010).

Selection induces a very particular signature in terms of genetic diversity (Nielsen 2005; Storz 2005). The principle of these studies is to analyse variation at a large number of loci to identify selection signatures (Storz 2005). Detection of selection in populations sampled along environmental gradients is a possible strategy for identifying genes linked to adaptation along environmental clines (Endler 1986). However, a caveat of markers or genes identified by selection scan is the often complete lack of knowledge about their functional effect (Sabeti *et al.* 2002; Vigouroux *et al.* 2002; Stephan & Li 2007).

The second approach using natural variation is association mapping in populations (Thornsberry *et al.* 2001; Yu *et al.* 2006; Saïdou *et al.* 2009). These studies allows to link molecular variation and phenotypic traits (Thornsberry *et al.* 2001; Yu *et al.* 2006; Saïdou *et al.* 2009). Combining selection scan and association studies thus seems to be a promising approach because it pools the advantages of both methods (Stinchcombe & Hoekstra 2008): selection scans can lead to the identification of functional variation at the genome level, and association mapping establishes links between allelic variation and phenotypic variation.

This study is about the adaptation of pearl millet, *Pennisetum glaucum* L. (Cyperales, Poaceae), to spatial climatic variation. Pearl millet is a crop species growing throughout the Sahel in West Africa. Pearl millet varieties are characterized by several distinctive traits: flowering time, spike length (SpL), spike diameter (SpD) and seed colour. In the Sahel, rainfalls occur during a brief time period from two to four consecutive months. So for an annual plant like pearl millet, this rainfall regime imposed a very strong constraint on the flowering time period. *Pennisetum glaucum* is distributed along a climatic gradient, notably characterized by decreasing rainfall along a south/north axis. We previously observed that flowering time is correlated with latitude which covaried with rainfall (Haussmann *et al.* 2006). Late flowering varieties are found in wetter and costal West African countries, and early flowering varieties are found in the driest area in the northern limit of rainfed agriculture around 200–250 mm annual rainfall. Other development traits like SpL are also associated with this environmental gradient (Haussmann *et al.* 2006).

The aim of this pioneer study was to identify gene associated with these adaptations. We overcame the lack of genomic data on this species by developing a method to perform genome-wide selection scanning on targeted gene families. We focused on the MADS-box gene family for which over a hundred genes have been identified in *Arabidopsis* (Nam *et al.* 2004). This transcription factor gene family plays a crucial role in vegetative and flower development (De Bodt *et al.* 2003). Several studies have identified different genes from this family associated with plant adaptation and evolution (Vigouroux *et al.* 2002; Wang *et al.* 2009). In this study, populations were sampled along an appropriate environmental gradient to identify functional variation associated with climatic variation. Thus, we first identified markers showing a selection signature (Foll & Gaggiotti 2008) that covaries with environmental climatic factors (Endler 1986;) and then correlated this variation with phenotypic variation using an independent association mapping panel (Saïdou *et al.* 2009).

Materials and methods

Study system and plant material

Pearl millet is an allogamous crop that is cultivated throughout West Africa, East Africa and India (Oumar *et al.* 2008). Pearl millet varieties are characterized by several distinctive traits: flowering time, SpL, SpD and seed colour. Twenty-one pearl millet varieties were sampled in 2003, with each variety originating from a different village (Bezançon *et al.* 2009). The 21 villages span the whole rainfed cropping area in Niger (Fig. 1). In 2004 and 2005, the varieties were characterized for several morphological and phenological traits in two

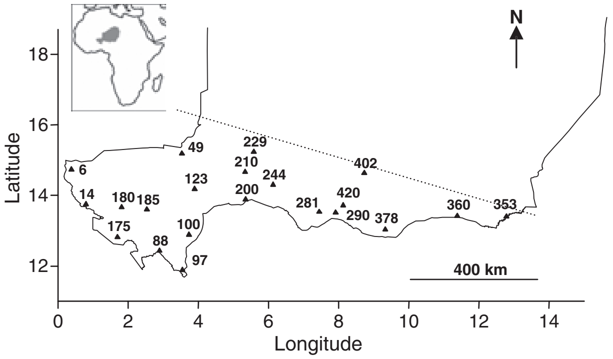


Fig. 1 Sampling location. Map of Niger with the 21 chosen villages sampled in 2003 spanning the whole cultivated area. One variety per village was genotyped. The dash line represents the limit of rainfed agriculture in Niger.

field trials (see Saïdou *et al.* 2009 for details). More than 90% of the genetic variation is found among plants within varieties (Allinne *et al.* 2008). So, for each variety, 21–31 different plants were used for DNA extraction, as previously described (Mariac *et al.* 2006; Oumar *et al.* 2008). A total of 530 different plants were used, with an average of 25.2 per variety (SE 2.5). The DNA samples were then randomly distributed in six 96-well plates with three negative controls (water) in each.

Targeting the MADS-box gene family across the genome

The MADS-box gene family is characterized by a conserved domain corresponding to the DNA-binding domain of the protein (N terminal). This domain (i.e. the MADS-box domain) is approximately of 60 amino acids. Based on the number of genes by chromosome in rice (Goff *et al.* 2002) and the total number of annotated MADS-box genes in rice, we could calculated an expected number of MADS-box genes per chromosome. This expected number is highly correlated to the observed number of MADS-box genes ($R = 0.87$, $P < 3 \times 10^{-4}$). In conclusion, the MADS-box genes are relatively well scattered across the rice genome (Arora *et al.* 2007). Our hypothesis is that this pattern might be conserved across cereals.

Our technique of genetic marker development is derived from the well-known AFLP method (Vos *et al.* 1995) adapted to pearl millet (Allinne *et al.* 2008). Primers were defined based on *Zea mays* sequences downloaded from NCBI to target MADS-box genes (http://www.ncbi.nlm.nih.gov/). A set of 43 sequences was aligned and analysed. Alignments were performed on protein sequences. Three highly conserved protein stretches in the MADS-box domain were used to design primers: RQVT, KKA and LCDA. Three primers specific to the RQVT domain were designed: RQVT1 (ACSARCMGSCARGTSACCT), RQVT2 (ARCCGSCAGGTSACVTTC) and RQVT3 (GSCAGGTSACSTTCT). Two primers were specific to the KKA domain: KKA1 (GGVMTVNTSAAGAAGGC) and KKA2 (CGCAAYRGVCTSCTCAAGAAGGC). Finally, two primers were specific to the LCDA domain: LCDA1 (AGMTCDSCRTSCTCTGCGA) and LCDA2 (AGCTSDCSSTGYTSTGCGAYGC).

DNA digestion by *Eco*RI restriction enzymes and adapter ligation were performed as described in a previous publication (Allinne *et al.* 2008). A first PCR, called preselective amplification, was performed with forward primers targeting the RQVT domain, coupled with a reverse primer named EPA (GTAGACTGCGTACCAATTC) targeting the *Eco*RI adapter. The reaction was performed in a 10-μL volume, including 3 μL digested and ligated DNA, 370 μM dNTP, 1 U Taq polymerase, 1× GoTaq buffer (Promega), 0.5 mM $MgCl_2$, 1 μM of the three bulked forward primers and 0.15 μM of the reverse primer. PCRs were performed on a Biometra thermocycler with the following programme: 3 min at 94 °C, 20 cycles of 30 s at 94 °C, 1 min at 50 °C, 1 min at 72 °C and a final 10 min extension step at 72 °C. PCR products were purified with magnetic beads according to the supplier's instructions (Ampure, Agencourt). Purified DNA was eluted in 40 μL water.

A second PCR, so-called selective amplification, was carried out on 3 μL of the purified first PCR product. Four independent PCRs were performed using KKA1, KKA2, LCDA1 and LCDA2 as forward primer and an IRD-700 labelled EPA primer as reverse primer. The PCR programme consisted of 3 min cycles at 94 °C, 35 cycles of 30 s at 94 °C, 30 s at 55 °C, 1 min at 72 °C and a final 10 min extension step at 72 °C.

PCR products were diluted (1:9) in a formamide solution (96% formamide, 4% EDTA 0.5 M, pH 8, bromophenol blue) and denatured by heating at 95 °C for 5 min. Samples were loaded on 5.5% acrylamide gels (41 cm × 0.2 mm) on an automated Li-Cor sequencer. Images were analysed using the AFLP-Quantar software package (Keygen) to identify and measure DNA fragments. For each individual, variation was coded as the absence (−) or presence (+) of each genetic marker.

A subset of amplified fragments was cloned and sequenced to check whether they belonged to the MADS-box gene family. This analysis was performed on two randomly selected primers KKA1 and LCDA2. PCR amplification was repeated as previously described except that the EPA primer of the second PCR was not labelled. Amplified fragments were cloned using the pGEM-T easy Vector system kit (Promega). Ligation and cell transformation were performed as recommended by the supplier on 3 μL of purified PCR products (Ampure). Positive clones were amplified with M13 forward and reverse primers. Amplicons of different sizes were sequenced after purification (Ampure). Sequencing was performed using the BigDye® Terminator v3.1 Cycle Sequencing kit (Applied Biosystem) according to the supplier's recommendations. CleanSeq kit (Agencourt) was used for the purification of the sequence reactions, which were then eluted in 40 μL of water. Migration was carried out on an ABI PRISM® 3100 Genetic Analyzer. Sequences were aligned using SeqMan II version 5.07 software and checked against the NCBI database through a tblastx algorithm. We considered that a sequence had some MADS-box family homology if the first matching sequence was a MADS-box gene (even for a high blast E value). The percentage of genes targeting the MADS-box family was assessed as the number of sequences having MADS homologies divided by the number of sequenced fragments.

Diversity analysis and selection detection

Analysis of diversity and structuration. The polymorphism information content (PIC) among varieties was calculated as $1-\left(\sum p_i^2\right)-\sum_i \sum_{j>1} 2p_i^2 p_j^2$, where p_i and p_j were the frequencies of the *i* and *j* alleles (Botstein *et al.* 1980). We used AFLP SURV 1.0 (Vekemans 2002) to calculate F_{ST} between populations. Three different methods are proposed by this software: (i) the square root method, (ii) a Bayesian approach assuming a uniform prior of the allele frequency (iii) and a Bayesian approach assuming a nonuniform prior of the allele frequency. For all methods, a fixed prior value of F_{IS} is needed. We investigated the impact of this parameter on F_{ST} estimations using $F_{\mathrm{IS}} = 0$ and $F_{\mathrm{IS}} = 0.13$. The value $F_{\mathrm{IS}} = 0.13$ was calculated based on a previous microsatellite data set (Mariac *et al.* 2006) with the same twenty-one pearl millet varieties used here.

Correlation analysis. Rainfall data were retrieved for the 21 sampling points by an estimation based on average rainfall from 1976 to 2003 (Ali *et al.* 2005). Isohyets and genotype isofrequencies were calculated using SURFER V7.02 software and the default linear variogram. Genotype frequency was correlated respectively with latitude and rainfall at the sampling site. The correlation significance was evaluated with the Student's test (Sokal & Rohlf 2000; Hancock *et al.* 2008).

Bayesian analysis. Foll & Gaggiotti (2008) proposed a Bayesian approach to separate neutral effects from adaptive effects in genomic surveys, which is also suitable for dominant markers like AFLP markers. Their method is based on a statistical framework previously described to separate inbreeding coefficients into a population-specific parameter β_j and a locus-specific parameter α_i, considering *j* populations and *i* loci (Beaumont & Balding 2004). Beaumont & Balding's (2004) initial Bayesian method proposed an informal criterion to assess the significance of locus showing an excess of differentiation. The improvement of the Bayesian approach is to derive a posterior probability of a locus being selected: $P(\alpha_i \neq 0)$ (Foll & Gaggiotti 2008) using a reversible-jump MCMC algorithm. So, this new Bayesian analysis allows a refined statistical assessment of significance. If the locus-specific parameter α_i is significantly different from zero, this indicates that its differentiation value differs more than one would expect based on the overall value from all markers. This higher or lower differentiation indicates the singularity of the diversity found at this locus. Selection is one of the mechanisms that could lead to such higher or lower differentiation values but other processes like a locus-specific mutation rate could also modify the expected F_{ST}. We are working with coding sequences, so F_{ST} could also be biased owing to possible evolutionary constrains (e.g. purifying selection.). Further studies thus have to be performed to confirm the selection explanation, e.g. by linking genetic variation and phenotypic variation. To identify loci showing an excess of differentiation (potential selected locus), we used the BAYESCAN v 1.0 software package (Foll & Gaggiotti 2008). A total of 100 000 iterations was performed after a burn-in period of 50 000 iterations. We chose a thinning interval of 20 iterations and thus 5000 results were recorded. For each marker, the analysis led to the estimation of α_i and the assessment of its significance using the log of the Bayes factor.

Associations between phenotype and genotype

Association mapping was performed on a panel of 90 inbred lines genotyped using 25 microsatellite loci and 306 classic AFLP markers (Saïdou *et al.* 2009). This panel of inbred lines is not related to the populations used for the selection scan and represents the diversity of a worldwide sample of pearl millet plants (Saïdou *et al.* 2009).

The analysis of population structure, kinship relationship between the 90 inbred lines as well as field trial data is lengthily reported elsewhere (Saïdou *et al.* 2009). To help readers in the understanding of current results, a brief summary of the previously performed analysis is given here. The population structure was assessed using microsatellite data set and the STRUCTURE procedure (Pritchard *et al.* 2000) with 30 000 burn-in iterations and 10^6 runs. We assumed that there were $K = 1$–10 populations and performed 10 simulations for a given number of assumed populations. The matrix of ancestry (Q) corresponding to the selected number of populations was used for further analysis of genotype/phenotype associations. A matrix of kinship coefficients (K) between each inbred line pair was calculated using SPAGeDI and AFLP data set (Hardy & Vekemans 2002). Morphological traits were evaluated on the basis of the findings of three different field trials: one trial in 2005 and two trials in 2006, named hereafter 2006a and 2006b (see Saïdou *et al.* 2009 for field trial details). We recorded flowering time from planting to female flowering stage (FT), the number of tillers at head emergence (NTHE), plant height (PH), stem diameter (SD), primary SpL and primary SpD from 6 to 10 plants per inbred line for each field trial. We used the average value for each trait and for each field trait, and a value calculated across experiments. To obtain a value across all three experiments, we calculated the best linear unbiased effect (BLUE) for each trait and inbred line. We considered y_{ijkl} as being the phenotype of plant l corresponding to the inbred line i, measure in plot k of field trial j. We fitted a mixed model: $y_{ijkl} = \mu + x_i + z_j + v_{jk} + \varepsilon_{ijkl}$, where μ is the grand mean and ε_{ijkl} is the residual error. We were interested in x_i the inbred effect (considered as a fixed effect), and field trial (z_j) and subplot (v_{jk}) were considered as random effects. The model was fitted using the lmer() function (R: http://cran.r-project.org/). The BLUE of each trait (FT, NTHE, PH, SD, SpL, SpD) was calculated for each inbred line as $\hat{\mu} + \hat{x}_i$(Saïdou *et al.* 2009).

The association analysis was run using a mixed model incorporating a SNP effect, the genetic structure of the line panels and their kinship. The mixed model was implanted using the kinship() function (Atkinson & Therneau 2008) in R using a maximum likelihood method. The complete model used was $y = X\beta + S\alpha + Qv + Zu + e$ (Yu *et al.* 2006; Saïdou *et al.* 2009), where y is the phenotype vector, β is a fixed effect other than SNP or population structure, α is the SNP effect, v the population structure effect, u the genetic background effect and e the residual error. The different parameters β, α, v, u and e are vectors related to the y vector of the phenotype by matrices: the ancestry matrix Q and the binary (0/1) matrices X, S, Z. The variance of the random effect u was expected to be $\mathrm{Var}(u) = K \times V$, where V is the variance and K the kinship matrix (Yu *et al.* 2006).

We assessed how taking the population structure and kinship into account leads to a better statistical model using likelihood ratio tests. We thus fit four different models: the complete model considering the kinship matrix (K) and ancestry matrix (Q), the model considering only Q, the model considering only K and a null model without K and Q. The likelihood of each model was recorded, and likelihood ratio tests were used to compare the different models. Likelihood ratio tests were appropriate as we considered nested model fit using a maximum likelihood approach. The complete model was compared to a model without the population structure (Q), without the kinship matrix (K) or a null model (without kinship and an ancestry matrix).

Moreover, to assess how taking population structure and kinship into account helps control type I error, we calculated the P-value of the association between phenotypes and AFLP markers. For each AFLP allele, their probability of being associated with each phenotype was calculated considering the complete model, the model without kinship (K), the model without an ancestry matrix (Q) and the null model without Q and K. For a given model, using the probability of association for all AFLP markers, we sorted them from their lower to higher probabilities. The rank of the markers divided by the number of markers was calculated. Then, this 'expected probability' was plotted with the observed probability of the considered markers. If the population structure and kinship were actually found to control the type I error, a direct relationship between the observed P-value and expected P-value ($x = y$) was expected. On the contrary, for the null model (without K and Q), a high number of false positives was expected. The P-values were calculated and plotted for the null model, the model considering K only, the model considering Q only and the model considering $Q + K$.

Although the completed model ($Q + K$) greatly reduced the false-positive rate, it could be higher than the commonly used 5% threshold (Yu *et al.* 2006; Saïdou *et al.* 2009). Thus, a conservative corrected threshold was estimated for each trait as follows: for each of the 306 AFLP markers, the P-value of its association with the phenotype was calculated. The AFLP markers were then ranked from the lowest to the highest P-value of the association. The corrected probability threshold was the P-value observed for the AFLP marker at the 5% ranks (Saïdou *et al.* 2009).

For the 21 populations used for the genome scan, we performed a simple correlation between the genotype frequency and morphological data. Morphological data (FT, SPL) were recorded in two field trials in 2004 and 2005 for each of the 21 populations (see Saïdou *et al.* 2009 for field trial details). To obtain a value across field trials for each population, a mixed model was fitted: $y_{ijk} = \mu + x_i + z_j + \varepsilon_{ijk}$, where y_{ijk} is the phenotype of plant k in variety i from field trial j, μ is the mean effect, x_i is the variety effect (fixed effect), z_j is the field trial effect (random effect) and ε_{ijk} the residual error. The BLUE for SpL and flowering time was estimated as $\hat{\mu} + \hat{x}_i$(Saïdou *et al.* 2009). The correlation between genotype frequency and the BLUE of SpL and flowering time was tested using a *t*-test with *n*−2 degrees of freedom (d.f.).

Results

Genetic diversity and differentiation

A total of 182 fragments were amplified by the four primer combinations among which 47 polymorphic markers were obtained. The average PIC value of the 47 polymorphic markers was 0.19. A total of 55 amplified fragments were cloned and sequenced (EMBL numbers: FN552468–FN552522). The percentage of enrichment was measured at 69%, as calculated by the number of markers blasting MADS-box genes divided by the total number of loci. The overall differentiation among the 21 populations was weak (0.0292) but significant, based on the Bayesian method with a uniform allele frequency prior and an F_{IS} value of 0.13 ($P < 0.001$, Table 1). The considered method and the initial F_{IS} value only had a minor effect on the differentiation and its significance (Table 1).

Table 1 Overall differentiation between the 21 populations

Method	F_{IS}	F_{ST}
Square method	0	0.0292**
	0.13	0.0358**
Bayesian uniform prior	0	0.0215**
	0.13	0.0236**
Bayesian nonuniform prior	0	0.0253**
	0.13	0.0292**

The overall differentiation (F_{ST}) between the 21 different populations was calculated using AFLP SURV 1.0 (Vekemans 2002). Three different methods were used: square method and two Bayesian approaches assuming or not a uniform prior distribution of allele frequencies and two F_{IS} values of 0 or 0.13 (see text for details). All differentiations calculated were significant (**$P < 0.01$).

Correlation analysis

For each marker, the correlation between the allele frequency in the 21 populations and their latitude was calculated. The average correlation over the 47 markers was $R = 0.273$, with a standard error of 0.183, and was not significant. M9LCDA2 and M13KKA1 showed a significant correlation (Bonferroni corrected *P*-value $P < 0.001$, $R = 0.72$ and $R = 0.73$ respectively, Fig. 2). The average correlation with rainfall was $R = 0.279$, with a standard error of 0.204, and was not significant. The same two markers also showed a significant correlation with rainfall (M9LCDA2 $R = -0.79$, $P < 0.001$ and M13KKA1 $R = -0.76$, $P < 0.001$), in addition to two others, i.e. M11KKA1 ($R = 0.67$, $P < 0.001$) and M13KKA2 ($R = -0.71$, $P < 0.001$). Rainfall covaried with latitude ($R = -0.84$, $P < 0.001$).

Bayesian analysis of differentiation

For each marker, the overall F_{ST} and the log of the Bayes factor (BF) were calculated (Fig. 3). The log of BF measures the posterior probability of selection. Threshold values for logBF were very high at 1.7 and showed decisive evidence at 2.0. A threshold of 1.7 corresponds to a posterior probability P ($\alpha_i \neq 0$) of 0.97, a threshold of 2 to a posterior probability of 0.99 (Foll & Gaggiotti 2008, BAYESCAN help file). Two markers showed a very high logBF, M13KKA1 (logBF = 1.81 alpha = 1.24) and decisive logBF, M9LCDA2 (logBF = 2.16 alpha = 1.27). Alpha was positive in both cases, characterizing a selection for adaptive divergence between populations.

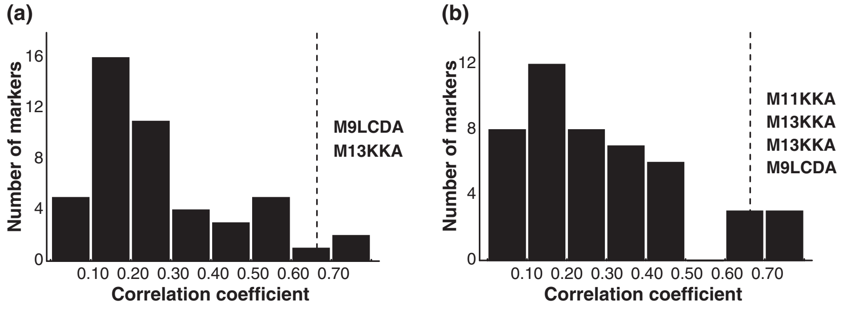


Fig. 2 Correlation between genotype frequency and latitude and average rainfall. Correlations of genotype frequency and latitude (a) or average rainfall (b) were calculated for each of the 47 markers. The correlation coefficient (*R*) distributions are given. For 47 markers, the Bonferroni corrected threshold corresponds to $P \leq 0.001$ or a correlation higher than $R = 0.665$ (dashed line). The name of the markers showing a significant correlation are reported on the graph for allele frequency and latitude (M9LCDA2: $R = 0.72$, $P < 0.001$; M13KKA1: $R = 0.73$, $P < 0.001$) and for allele frequency and rainfall (M9LCDA2: $R = -0.79$, $P < 0.001$; M13KKA1: $R = -0.76$, $P < 0.001$; M11KKA1: $R = 0.67$, $P < 0.001$; M13KKA2: $R = -0.71$, $P < 0.001$).

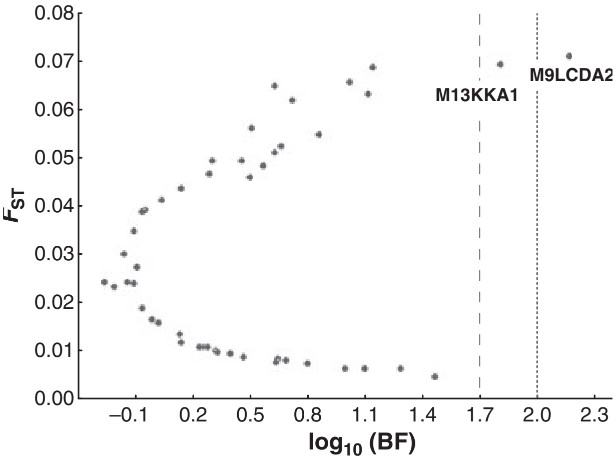


Fig. 3 Detection of selection using a Bayesian approach. Selected loci were detected using BAYESCAN V1.0 (Foll & Gaggiotti 2008). The figure represents the estimated F_{ST}, the log of the Bayes factor and the posterior selection probability. Threshold values for very strong and for decisive selection evidence were at 1.7 (dashed line) and 2.0 (dotted line), respectively. Two markers showed very strong (M13KKA1) and decisive (M9LCDA2) selection evidence.

Identification of MADS-box genes

The correlation of allele frequencies in the 21 populations was high between M11KKA1 and M13KKA1 ($R = -0.55$, $P < 0.001$) as well as between M8LCDA2 and M9LCDA2 ($R = -0.76$, $P < 0.01$). A 24-bp length differentiated the M11KKA1 and M13KKA1 alleles, and the same size difference was noted between M8LCDA2 and M9LCAD2. Finally, the frequencies of M13KKA1 and M9LCDA2 alleles were highly correlated ($R = 0.97$, $P < 10^{-12}$). Correlations between the frequencies of M11KKA1 and M8LCDA2 alleles were significant but weaker ($R = 0.57$, $P = 0.006$). Altogether, these correlations suggest that the two primers (KKA1 and LCDA2) might amplify the same two alleles. This was confirmed on the 90 inbred lines, where the four markers showed a clear pattern of being different alleles of the same gene (data not shown).

AFLP fragments amplified by KKA1/EPA primers were sequenced to identify one cloned fragment having the same size as the M13KKA1 marker. To check whether the sequenced fragment corresponded to the AFLP markers, primers were designed to amplify this sequence (MADS11F2 GGCTGAGGAGAAAACAGTGC, MADS11R2 CCAAAACCAAACCCTAGCAA). PCR using this new set of primers on preselective AFLP-PCR led to a single amplified fragment, and we sequenced eight individuals (EMBL numbers: FN551253–FN551260). The sequences were aligned and showed two alleles with an INDEL of 24 bp. A new set of primers (MADS11F3 CCAAAACCAAACCCTAGCAA, MADS11R3 GCAGTT CGCCAACTCCAG) designed to amplify a shorter fragment was used to genotype the INDEL on the 90 inbred lines. The two observed alleles showed a 97.8% match with the four markers (M11KKA1, M13KKA1, M8LCDA2, M9LCDA2). Polymorphisms outside the INDEL, or genotyping errors, could explain the slight difference with respect to the expected 100% match. The sequenced gene returned a hit on the *StMADS11* subgroup through a BLAST algorithm (Table S1, Supporting Information). Altogether, these results suggest that the four markers correspond to the same gene (named *PgMADS11* hereafter), which had two alleles that differed from each other by an INDEL polymorphism of 24 bp.

Association mapping

The population structure results were explained in length in a previous paper (Saïdou *et al.* 2009), and $K = 7$ was thus selected as the number of different subgroups of individuals. We then statistically compared the different mixed models: the complete model ($Q + K$), the kinship model (K), the population structure model (Q) and a null model. Although we considered type I errors, based on the relationship between observed P-value vs. expected P-value (Fig. S1, Supporting Information), we observed that the $Q + K$ model was generally better for all traits but often not very different from the K model. One noticeable exception was the SpD, where the K model led to a better type I error control than the $Q + K$ model. Likelihood ratio tests (Table S2, Supporting Information) showed that the mixed model that considered both the population structure (Q matrix) and the kinship matrix (K) was better than simple models for flowering time, SpD and basal spike diameter (BSpD). The most complete model was not significantly different from the K model for the number of tillers and plant height. Finally, the most complete model was not significantly different from the Q model for stem diameter and SpL. For further analysis, we thus used the most complete mixed model ($Q + K$) because it generally had the highest likelihood.

The analysis of association between the identified INDEL polymorphism and morphological traits was performed on the panel of 90 inbred lines (Table 2). A significant association was observed with flowering time ($t = 2.66$, $P \leq 0.009$) and SpL ($t = 3.03$, $P \leq 0.003$). The variance explained by the marker was $R^2 = 0.056$ and 0.067 for flowering time and SpL, respectively. This association analysis was performed using the BLUE value of each trait for each inbred line. Similar results were obtained for SpL when the analysis was

Table 2 Association between *PgMADS11* and phenotype

Field trials	Effect	*t*	*P*-value	*P*-value threshold
Traits				
Flowering time				
FT	4.21 ± 1.58	2.66	**0.009**	0.021
Plant morphology				
PH	0.40 ± 5.93	0.067	0.95	0.003
SD	0.11 ± 0.042	2.74	0.008	0.006
NTHE	0.77 ± 0.68	1.13	0.26	0.033
Spike morphology				
SpD	0.19 ± 0.09	2.10	0.039	0.027
BSpD	0.0093 ± 0.025	0.37	0.71	0.008
SpL	6.17 ± 2.03	3.03	**0.003**	0.015

The statistical results of the association between the best linear unbiased estimates (BLUE) of phenotype and the two *PgMADS11* alleles are presented. The effect of one allele over another (absolute value) for each trait and the standard error for flowering time (FT), plant height (PH), stem diameter (SD), number of tillers at head emergence (NTHE), spike diameter (SpD), spike length (SpL) and basal spike diameter (BSpL) are given. The associated t-test and its *P*-value for significance are also given. Finally, for each trait, a *P*-value empirical threshold is given. This corrected *P*-value threshold is a trait-specific threshold, taking into account an imperfect control of the false-positive rate at the 5% standard threshold (see text for details). Significant *P*-value, lower than the *P*-value corrected threshold, is highlight in bold. A significant association of the gene was observed with flowering time and SpL.

performed on the three field trials (Table S2, Supporting Information). For flowering time, only the 2006b field trial was highly significant ($t = 2.73$, $P \leq 0.008$), i.e. the two other field trials are lower than the 5% threshold but slightly higher than the empirical *P*-value threshold (Table S3, Supporting Information). For the 21 populations, a correlation was calculated between the M9LCDA2 allele frequency and the BLUE estimate of flowering time and found a significant correlation (Fig. 4, $R = -0.69$, $t = -4.16$, d.f. = 19, $P < 0.001$). A significant correlation was also obtained with SpL ($R = -0.49$, $t = -2.45$, d.f. = 19, $P < 0.03$). *PgMADS11* allelic variation was thus significantly associated with the flowering time variation.

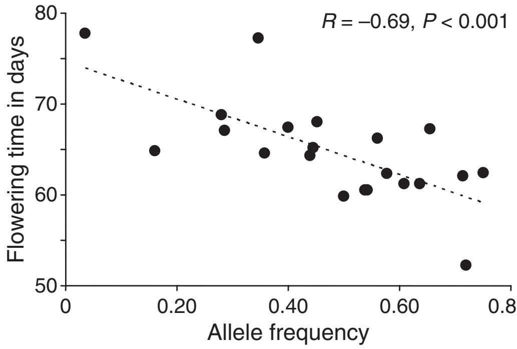


Fig. 4 Flowering time variation and *PgMADS11* allele frequency. The allele frequency of *PgMADS11* (M9LCDA2 marker) and flowering time were correlated for 21 pearl millet varieties. A highly significant correlation was obtained ($R = -0.69$, $n = 21$, $P < 0.001$).

Discussion

Genome scan using differentiation between populations

In this study, we used a differentiation-based approach to identify outlier loci. Early work proposed variance of F_{ST} as a metric to assess if some loci showed unusual diversity, suggesting selection (Lewontin & Krakauer 1973). This early test was criticized (Nei & Maruyama 1975; Robertson 1975) because correlations of allele frequency between populations generally inflate the expected variance. To avoid relying on a fixed formula, several authors later proposed simulation-based approaches (Bowcock *et al.* 1991; Beaumont & Nichols 1996) to build up the expected distribution of differentiation. Simulations are based on a simple island model (Beaumont & Nichols 1996) which provides a good approximation of the F_{ST} distribution obtained with some more complex population structure models (Beaumont & Nichols 1996). More recently, new statistical approaches were proposed which separated differentiation into gene-specific effects and population-specific effects (Beaumont & Balding 2004; Foll & Gaggiotti 2008). However, some specific population structures could lead to high levels of false-positive loci (Excoffier *et al.* 2009) like the hierarchical island model (Slatkin & Voelm 1991; Vigouroux & Couvet 2000). In our study, the hierarchical population structure was not expected. Sampling was performed at a relatively small scale, on a single region, where long-range migration was observed (Allinne *et al.* 2008), and differentiation between populations did not show specific sample clusters, i.e. a signature of different migration rates in different areas. So our sampling did not seem to fit the extreme population structure which might strongly bias the statistics we used.

The idea underlying all of these different methods is that the population structure and history shaped genome-wide diversity, while selection is locus specific. However, in some special cases of colonization of new habitats, mutations arising at the front of colonization could increase to a high frequency (Klopfstein *et al.* 2006). The frequency pattern observed for genes harbouring these mutations would mimic an increase in frequency similar to positive selection, even though the process is purely neutral. The historical spread of agriculture after domestication might exhibit some loci

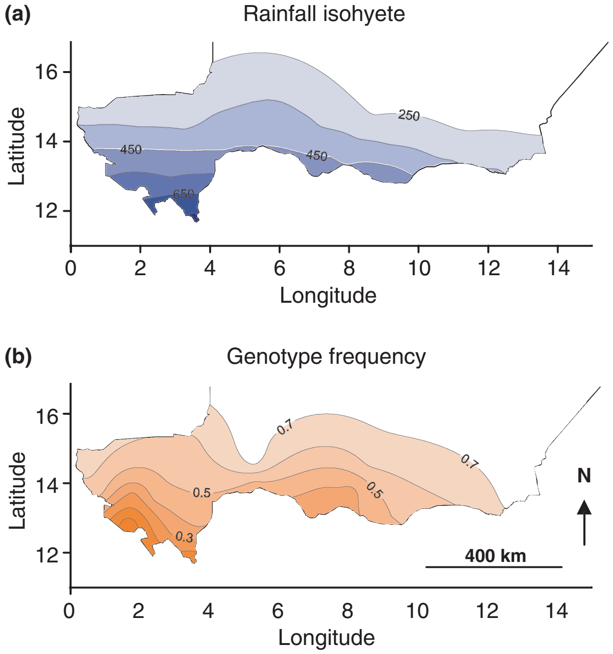


Fig. 5 Rainfall and *PgMADS11* allele frequency. Annual rainfall (in mm) was estimated based on average pluviometric data from 1976 to 2003. Genotype isofrequencies for the *PgMADS11* (M9LCDA2 marker) were calculated using SURFER V7.02. The genotype frequency in populations ranged from <4% to 75% and covaried with rainfall.

which show a 'surfing mutation' pattern. However, our analyses are not limited to the detection of positive signals using genome scans. We also used an association mapping strategy to link molecular polymorphisms to an associated phenotypic variation.

PgMADS11 *and flowering time phenotype*

We identified some *PgMADS11* polymorphism associated with phenotypic variation (flowering, SpL). However, it is still possible that the real functional polymorphism(s) is located in the neighbouring genomic region or in a genomic region that is in sufficient linkage disequilibrium with the revealed polymorphism. For instance, in maize, an early study found *Dwarf8* polymorphism associated with flowering time differences (Thornsberry *et al.* 2001), and later work suggests the causative polymorphism might be located 100 kb upstream of the *Dwarf8* gene (Camus-Kulandaivelu *et al.* 2008). In pearl millet, linkage disequilibrium seems to decrease rapidly (Saïdou *et al.* 2009). One can only conclude that the revealed polymorphism is linked with the studied phenotypic variation in flowering time and not that the amplified gene is directly responsible for the observed trait variation. Further analysis is thus required to obtain genomic data around the *PgMADS11* gene and further experimental evidence to identify the exact functional variation associated with the observed phenotype. Based on the actual sequence data, the INDEL polymorphism is located in *PgMADS11* intron and is not likely responsible for obvious functional variation (Fig. S2, Supporting Information). *PgMADS11* has a high sequence similarity with a small clade (*StMADS11* clade) of the MADS-box gene family (Becker & Theissen 2003; Kane *et al.* 2005; Zhao *et al.* 2006). In *Arabidopsis*, two genes are found in this clade, i.e. *AGL24* and *SVP* (Becker & Theissen 2003). It was demonstrated that a *SVP* loss of function leads to an early flowering phenotype (Becker & Theissen 2003) and that natural variation at this gene is also associated with flowering differences (Atwell *et al.* 2010). In monocotyledons, including species like maize, rice or wheat, only one to two genes per species are found in this particular clade (Kane *et al.* 2005). *TaVRT2* is one gene whose function is well documented in monocotyledons. This gene is known to repress the transition from the vegetative to the reproductive phase in wheat (Kane *et al.* 2005, 2007). In conclusion, several members of the *StMADS11* family in both monocotyledons and dicotyledons are associated with phase transition (Liu *et al.* 2009), and their mutants have a phenotypic effect on flowering time. *PgMADS11* is thus a very likely candidate for a direct role in the observed phenotypic variation.

Polymorphism at PgMADS11 *and climate variation*

We found that polymorphism revealed at the *PgMADS11* gene was closely associated with rainfall isohyets (Figs 3 and 5). Our hypothesis is that flowering time is directly or indirectly governed by environmental variables, including rainfall. We observed a close association between rainfall and flowering time (Haussmann *et al.* 2006). *PgMADS11* molecular polymorphism and flowering phenotype variation make this gene an ideal candidate for monitoring adaptive changes to climatic variation. We now have to assess whether the drastic reduction in rainfall over the last 40 years in the Sahel has had a selective effect on allele frequencies at this gene.

Transcription factor genome scan

We applied our method to focus on transcription factors that are part of an interesting gene family because several of these genes have been identified as playing a key role in the evolution of plant morphology (Wang *et al.* 1999, 2005, 2009; Vigouroux *et al.* 2002). Fifty transcription factor families (Davuluri *et al.* 2003) and 1738 DNA-binding domain genes have been identified to date in *Arabidopsis* (Wilson *et al*. 2008). This method can be applied to any plant, including species that have not been extensively studied, and it targets any gene family insofar as primers can be designed and validated. It allows whole genome scans for the identification of interesting genes. Moreover, genome scan to detect selection does not require any knowledge on gene functions. It is thus possible to find new genes for which the functions remain totally unknown in model plants or animals (Vigouroux *et al*. 2002). The genotypic profiles obtained from the MADS primers KKA1, KKA2, LCDA1 and LCDA2 had around 182 bands. With an enrichment rate evaluated at 69%, we amplified 124 markers matching to MADS genes. In *Arabidopsis*, which has been entirely sequenced, the number of putative MADS proteins was estimated to be around 100 (Nam *et al*. 2004) and around 80 in rice (Arora *et al*. 2007). This suggests either that pearl millet has 25% more MADS genes than *Arabidopsis* or, more likely, that there was redundancy among the amplified markers. A single locus having more than one allele could be revealed by several bands, with each allele being potentially amplified. However, on the contrary, an AFLP marker of a given size could also pool different genes of the exact same size. In the present case, it was hard to assess the total number of different genes that our method is able to screen. However, next-generation sequencing approaches could easily be used to complement the approach adopted in this study. Such studies would enable the identification of a larger number of polymorphisms like SNPs and not only the presence/absence of alleles. Combining this approach with next-generation sequencing methods could be particularly powerful and help in determining the number of different genes assessed by this approach.

Combining genome scan, association mapping and environmental data would be particularly useful to obtain a complete picture of the adaptation process *in situ*. Association studies are particularly powerful when the alleles are relatively common and have a strong phenotypic effect (Yu *et al*. 2006; Saïdou *et al*. 2009). A simulation study highlighted that selection scans can especially identify genes that have a major effect on the phenotype (Chevin & Hospital 2008). Consequently, a successful selection scan that identifies functional variation could be easily followed up by an association study. We performed a study that combined these different approaches and detected a very likely candidate gene that included a molecular selection signature, a link between molecular variation and a phenotype and a link between molecular variation and environmental data. Further studies are needed to validate the role of this particular gene, but this study paves the way to faster and more efficient identification of the molecular basis of adaptation in a population of non-model plants and animals.

Acknowledgements

We thank D. Moussa and M. Tidjani for help in the field studies and laboratory experiments. This project was funded by a core grant from the French Research Institute for the Development (IRD) and by a grant from the French National Research Agency (ANR) to Y.V. (ANR-07-JCJC-0116-01). Y.V., ACT and JLP were supported by the Agropolis Foundation through the Agropolis Resource Center for Crop Conservation, Adaptation and Diversity project (ARCAD). A.-A. S. was supported by a PhD fellowship from the IRD and the Agropolis Foundation.

References

Ali A, Lebel T, Amani A (2005) Estimation of rainfall in the Sahel. Part 1: error function. *Journal of Applied Meteorology*, **44**, 1691–1706.

Allinne C, Mariac C, Vigouroux Y *et al*. (2008) Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. *Genetica*, **133**, 167–178.

Arora R, Agarwal P, Ray S *et al*. (2007) MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics*, **8**, 242.

Atkinson B, Therneau T (2008) Kinship: mixed kinship:mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package, Versions 1.1.0–21. http://cran.r-project.org.

Atwell S, Huang YS, Vilhja'lmsson BJ *et al*. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Balasubramanian S, Sureshkumar S, Agrawal M *et al*. (2006) The phytochrome C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. *Nature Genetics*, **38**, 711–715.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society London B: Biological Sciences*, **263**, 1619–1626.

Becker A, Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution*, **29**, 464–489.

Bezançon G, Pham JL, Deu M *et al*. (2009) Changes in the diversity and geographic distribution of cultivated millet (*Pennisetum glaucum* [L.] R. Br.) and sorghum (Sorghum bicolor (L.) Moench) varieties in Niger between 1976 and 2003. *Plant Genetic Resources and Crop Evolution*, **56**, 223–236.

Bolstein D, White RL, Skolnick MH (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

Bowcock AM, Kidd JR, Mountain JL *et al.* (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 839–843.

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, **21**, 945–956.

Camus-Kulandaivelu L, Veyrieras JB, Madur D *et al.* (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. *Genetics*, **172**, 2459–2463.

Camus-Kulandaivelu L, Chevin LM, Tollon-Cordet C *et al.* (2008) Patterns of molecular evolution associated with two selective sweeps in the Tb1-Dwarf8 region in maize. *Genetics*, **180**, 1107–1121.

Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics*, **180**, 1645–1660.

Davuluri RV, Sun H, Palaniswamy SK *et al.* (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

De Bodt S, Raes J, Van de Peerl Y, Theißen G (2003) And then there were many: MADS goes genomic. *Trends in Plant Science*, **8**, 475–483.

Ducrocq S, Madur D, Veyrieras JB *et al.* (2008) Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. *Genetics*, **178**, 2433–2437.

Endler JA (1986) *Natural Selection in the Wild*. Princeton University Press, New-Jersey.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Goff SA, Ricke D, Lan T-H *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

Hancock AM, Witonsky DB, Gordon AS *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, **99**, 12949–12954.

Haussmann BIG, Boubacar A, Boureima SS, Vigouroux Y (2006) Multiplication and preliminary characterization of West and Central African pearl millet landraces. *International Sorghum and Millet Newsletter*, **47**, 110–112.

Jones H, Leigh FJ, Mackay I *et al.* (2008) Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. *Molecular Biology and Evolution*, **25**, 2211–2219.

Kane NA, Danyluk J, Tardif G *et al.* (2005) TaVRT-2, a member of the StMADS-11 clade of flowering repressors, is regulated by vernalization and photoperiod in wheat. *Plant Physiology*, **138**, 2354–2363.

Kane NA, Agharbaoui Z, Diallo AO *et al.* (2007) TaVRT2 represses transcription of the wheat vernalization gene TaVRN1. *Plant Journal*, **51**, 670–680.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genetics*, **2**, e166.

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science*, **325**, 1095–1098.

Liu C, Thong Z, Yu H (2009) Coming into bloom: the specification of floral meristems. *Development*, **136**, 3379–3391.

Manel S, Conord C, Després L (2009) Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology*, **9**, 288.

Mariac C, Luong V, Kapran I *et al.* (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**, 49–58.

Meyer CL, Vitalis R, Saumitou-Laprade P, Castric V (2009) Genomic pattern of adaptive divergence in *Arabidopsis halleri*, a model species for tolerance to heavy metal. *Molecular Ecology*, **18**, 2050–2062.

Murray MC, Hare MP (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. *Molecular Ecology*, **15**, 4229–4242.

Nam J, Kim J, Lee S, An G, Ma H, Nei M (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences*, **101**, 1910–1915.

Nei M, Maruyama T (1975) Letters to the editors: Lewontin–Krakauer test for neutral genes. *Genetics*, **80**, 395.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Oumar I, Mariac C, Pham JL, Vigouroux Y (2008) Phylogeny and origin of Pearl Millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theoretical and Applied Genetics*, **117**, 489–497.

Poncet BN, Herrmann D, Gugerli F *et al.* (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology*, **19**, 2896–2907.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Robertson A (1975) Letters to the editors: remarks on the Lewontin–Krakauer test. *Genetics*, **80**, 396.

Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.

Saïdou AA, Mariac C, Luong V *et al.* (2009) Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics*, **182**, 899–910.

Samis KE, Heath KD, Stinchcombe JR (2008) Discordant longitudinal clines in flowering time and phytochrome C in *Arabidopsis thaliana*. *Evolution*, **62**, 2971–2983.

Slatkin M, Voelm L (1991) F_{ST} in a hierarchical island model. *Genetics*, **127**, 627–629.

Sokal R, Rohlf FJ (2000) *Biometry*. W.H. Freeman and Co, New York.

Stephan W, Li H (2007) The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity*, **98**, 65–68.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Thornsberry JM, Goodman MM, Doebley J *et al.* (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, **28**, 286–289.

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proceedings of the National Academy of Sciences*, **106**, 9979–9986.

Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA (2005) A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science*, **308**, 691–693.

Vekemans X (2002) AFLP-SURV version 1.0. Distributed by the author. Laboratoire de Génétique et EcologieVégétale, Université Libre de Bruxelles, Belgium.

Vigouroux Y, Couvet D (2000) The hierarchical island model revisited. *Genetic Selection Evolution*, **32**, 395–402.

Vigouroux Y, McMullen M, Hittinger CT *et al.* (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences*, **99**, 9650–9655.

Vigouroux Y, Mitchell S, Matsuoka Y *et al.* (2005) An analysis of genetic diversity across the maize genome using microsatellites. *Genetics*, **169**, 1617–1630.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. *Nature*, **398**, 236–239.

Wang H, Nussbaum-Wagler T, Li B *et al.* (2005) The origin of the naked grains of maize. *Nature*, **436**, 714–719.

Wang R, Farrona S, Vincent C *et al.* (2009) PEP1 regulates perennial flowering in *Arabis alpina*. *Nature*, **459**, 423–427.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, **36**, D88–D92.

Wright SI, Bi IV, Schroeder SG *et al.* (2005) The effects of artificial selection on the maize genome. *Science*, **308**, 1310–1314.

Yu J, Pressoir G, Briggs WH *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Zhao T, Ni Z, Dai Y *et al.* (2006) Characterization and expression of 42 MADS-box genes in wheat (*Triticum aestivum* L.). *Molecular Genetics and Genomics*, **276**, 334–350.

C.M. and H.J. are research associates interested in plant genetics. L.J. is a master student interested in plant biodiversity. M.C. and P.S. are research assistants working on plant molecular biology. A.-C.T. and G.B. are researchers interested in crop diversity and adaptation. H.A. is a researcher interested in molecular evolution and development. J.-L.P. is a researcher interested in crop diversity and *in-situ* conservation of crop genetic resources. Y.V. is a researcher interested in plant evolutionary biology, domestication and plant adaptation.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Blast of FN551253 on different species

Table S2 Likelihood and model comparison for association studies

Table S3 *P*-value of associations between *PgMADS11* and phenotype

Fig. S1 Type I error comparison on different pearl millet traits.

Fig. S2 Sequence of two alleles of the *PgMADS11* gene.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

ANNEXE 2.

Selection for earlier flowering crop associated to climatic variations in the Sahel.

Title: Selection for earlier flowering crop associated to climatic variations in the Sahel

Short title: Selection for earliness in the Sahel

Authors

Yves Vigouroux[1,2,*], Cédric Mariac[1,2], Jean-Louis Pham[1], Bruno Gérard[3], Issoufou Kapran[4], Fabrice Sagnard[5], Monique Deu[5], Jacques Chantereau[5], Abdou Ali[6], Jupiter Ndjeunga[3], Viviane Luong[1], Anne-Céline Thuillet[1], Abdoul-Aziz Saïdou[1,2,7], Gilles Bezançon[2]

Affiliations

1 Institut de Recherche pour le Développement, IRD, UMR DIAPC, BP64501, 34394 Montpellier, France ; **2** Institut de Recherche pour le Développement, IRD, BP11416, Niamey, Niger ; **3** International Center of Research for the Semi-Arid Tropics, ICRISAT, Niamey, Niger; **4** Institut National de la Recherche Agronomique du Niger, INRAN, Niamey, Niger ; **5** Centre de coopération internationale en recherche agronomique pour le développement, CIRAD, Montpellier, France ; **6** Centre Régional AGHRYMET, Niamey, Niger ; and **7** Université Abdou Moumouni de Niamey, Niger

* Corresponding author:

Y. Vigouroux
Institut de Recherche pour le Développement,
UMR DIAPC
911, avenue Agropolis,
BP64501
34394 Montpellier, France.
Phone: 33 (0) 467416165
Fax: 33 (0) 467416222
Email: yves.vigouroux@ ird.fr

Abbreviations

PHYC, phytochrome C gene.

Abstract

Climatic changes will have an impact on food production and will necessitate important and costly adaptative procedures. Adaptation to a changing environment will be particularly acute in sub-Saharan Africa where climatic change is expected to have an important impact. However, one of the important yet often overlooked and poorly documented phenomena is the ability for agro-systems to adapt rapidly to environmental variations. Such adaptation could proceed by the adoption of new varieties or by the adaptation of varieties to a changing environment. In this study, we analyzed these processes in one of the driest agro-ecosystems of Africa, the Sahel. We performed a detailed study in Niger. In Niger, pearl millet is the major crop covering 65% of the cultivated area. To assess how the agro-system responds to recent recurrent drought, we analyzed samples of pearl millet landraces collected in the same villages in 1976 and 2003 throughout the entire cultivated area of Niger. We studied the phenological and morphological evolutions between 1976 and 2003 collections by comparing them over three crop seasons in a common garden experiment. We found no major changes in the main cultivated varieties or in term of genetic diversity. But, we observed a significant shift in adaptive traits. Compared to the 1976 samples, samples collected in 2003 displayed a shorter lifecycle, and a reduction in plant and spike size. We also found that an early flowering allele at the *PHYC* locus increase in frequency from 1976 to 2003. The increase exceeded drift and sampling effect, suggesting a direct effect of selection for earliness on this gene. We conclude that recurrent drought lead to selection for earlier flowering in a Sahelian major crop. These results astonishingly suggest that diffusion of crop varieties is not the main driver of short term adaptation to climatic variation.

Introduction

Feeding 9 billions people in the next few decades is a significant challenge (1). This is particularly acute in developing countries where population growth is the highest. In developing countries, farming communities are a large fraction of the population and ensure their food security through their own production (2). Agriculture in developing countries is strongly dependent on rain fall conditions, unfavorable changes in environmental conditions have then a direct impact on food security (3). As a consequence, the impact of climatic change on food security is expected to be the highest in developing country (3-5). Face to this challenge, several study highlight the necessity of adaptation of agriculture to future conditions (3). Sub-Saharan Africa is foreseen as one of the most susceptible area (4). Among Sub-Saharan regional climatic forecast, the Sahel is one where uncertainty is the highest (6), and models give drastic opposite outcome: important increase or decrease of annual rainfall (6-7). However a more recent study about the Sahel suggests a common phenomena observed across climatic models : a delay of the rainy season and a likely shortening of the rainy season (8). So, there still a strong uncertainty on the forecast of climatic variation in the Sahel. Better forecasts of the impact of climatic variation on agriculture production will need subtle regional climatic model, but also a better knowledge of how climate and agriculture production interact and also how agro-ecosystem respond to climatic variation (9-10).

Today, data on the response of agro-ecosystem and society to current changes are scarce. How for example spread or adaptation of varieties might already mitigate climatic variations? Understanding this process is of utmost importance as it could already suggest strategy to reinforce agro-ecosystem adaptation. In this study, we address this query in the Sahel. The Sahelian agro-ecosytem is dominated by two crops: sorghum and pearl millet. In Niger, pearl millet is the main crop covering more than 65% of the cultivated area and producing more than 80% of all caloric intakes in the country. To understand how traditional cropping systems respond to environmental variation we focused on the changes in the adaptation and spread of cultivated pearl millet varieties at the whole country scale. We analyzed the impact on the presence of cultivated pearl millet varieties and their adaptation during the drier period that has been under way in the Sahel since the early 1970s. Pearl millet cultivation is based on rain fed agriculture, and no irrigation system is generally used. Farmers sow their fields after the first significant rain which may occur between May and July in Niger (Fig. 1*A*). Several sowing dates are sometimes necessary as high plantlet mortality often occurs because of rain scarcity at the onset of the rainy season. Different traditional varieties are grown in Niger and they are variable with respect to spike shape, spike size, grain color and flowering time, among other traits (Fig. 1*B*). Plants are harvested after the end of the rainy season, which is generally in October in Niger (Fig. 1*C*). Spike bundles are then stored in granaries built in the field or close to the village during the dry season (Fig. 1*D*). Farmer selected spikes are then stored and used as seed for the next growing season (Fig. 1*E*). Sahelian farmers rely mainly on their own seeds, but when there are shortages they obtain seeds from family members, neighbors or sellers on local markets (11-12). Few modern pearl millet varieties are currently grown in Niger (11).

We used a sample of 192 accessions collected in 1976 throughout Niger (13). In 2003, a second sampling (14) was performed in the same set of 79 villages covering the entire rainfed cultivated area of this country (Fig. S1). These two samples were compared in a common garden experiment with respect to their phenological and morphological changes. We also compare the changes associated with these two samples in term of their genetic diversity at neutral locus and at a quantitative trait locus. And we observed an evolution toward earlier flowering varieties.

Results

Characterization of a recent climatic change in Niger.

Sahelian countries have experienced a significant climatic shift to drier climatic conditions over the last 50 years. On the country scale, we actually observe a major shift in average isohyetes for the 1977-2003 period compared to 1950-1976 (Fig. 2). The last 27-year period was drier than the previous period.

Evolution between 1976 and 2003

The two collections show no significant changes in term of the varieties cultivated (14). The 1976 sample presented a very good average germination rate (i.e. 87.2%). The 2003 sample show a slightly higher germination rate of 90.0% ($t=1.62$, $p\leq0.044$). This slight difference in germination rate (2.8%) for seeds preserved at 4°C for 27 years showed that the conservation conditions were very good. We first performed an analysis using these initial seeds.

We first compare the two collections globally. We found that plants collected in 2003 had a significantly earlier flowering date ($F_{1,1500}=11.4$, $P\leq0.001$), were shorter ($F_{1,1500}=40.8$, $P<0.001$), and have smaller spikes ($F_{1,1500}=10.5$, $P\leq0.001$) than the plants collected in 1976. No significant interactions between sample and field effects were detected but there was a strong and significant field effect for each morphological trait: flowering time ($F_{2,1500}=131.9$, $P<0.001$), plant height ($F_{2,1500}=174.3$, $P<0.001$) and spike length ($F_{2,1500}=5.2$, $P\leq0.006$).

We also perform a statistical analysis which considered the geographical origin of the varieties. To do so, we first calculated an average flowering time, spike length and plant height for each sampling point based on all the varieties collected on a given village. Then, we compared of the 1976 and 2003 samples by comparing the average values using a paired Wilcoxon test based on their geographical origin. To combine the three field trials into a single test, we used a Fischer combining probability test. Taking into account geographical origin, we observed a significant shift in flowering time (Fisher's combined probability, $\chi^2=42.0$, $n=6$, $P<0.001$), plant height (Fisher's combined probability, $\chi^2=58.9$, $n=6$, $P<0.001$), and spike length (Fisher's combined probability, $\chi^2=47.4$, $n=6$, $P<0.001$). The decrease in flowering time and plant height is important in the southern eastern and central parts of the country (Fig. 1 B).

These initial analyses of changes in phenological and morphological changes over time cannot separate two phenomena: expansion or contraction of the distribution area of some varieties and possible adaptation of varieties *per se*. The main reason is that a different number of varieties were collected in 1976 and 2003 in a given village. As a consequence, changes in the average value of some traits (e.g. flowering date) could be due to a change in the richness of certain types of varieties (e.g. early varieties), instead of reflecting adaptation of varieties *per se*. To assess whether adaptation does play a role in the observed changes, a further analysis was conducted considering, for each village, only the varieties that were sampled both in 2003 and 1976. Each variety sampled in 2003 was compared with the "same" variety sampled in 1976. We observed again a significant shift for all traits: the 2003 samples display a shorter flowering time by 1.2 days (Fisher's combined probability, χ^2=53.4, n=6, P<0.001), a shorter plant size by 6.6 cm (Fisher's combined probability, χ^2=61.8, n=6, P<0.001) and a shorter spike by 4.5 cm (Fisher's combined probability, χ^2=48.7, n=6, P<0.001). We finally assess if a possible seed conservation effect or an increase in earlier flowering weedy plants in the 2003 sample could be a factor explaining the morphological difference between samples.

Analysis of maternal, sampling year and conservation effects.

A new field trial analysis was performed using seeds produced in the same field in 2004 for the 2003 and 1976 samples. Germination rate of seeds produced in 2004 did not present any significant difference between the 1976 and 2003 samples (t=1.16, n=100, p=0.25). The average germination rate was 87.6% (SE ±0.021) and 84.4% (SE ±0.024) for the 2003 and 1976 samples, respectively. The average seed weight of the 2003 sample was 0.978 (SE ±0.006) and 0.971 (SE ±0.010) for the 1976 sample. The two samples did not show any difference in weight between the 1976 and 2003 (t=-0.60, n=607, p=0.55).

A new field experiment was performed in 2007 using these seeds to compare the 1976 and 2003 samples. The comparison of both 1976 and 2003 samples showed highly significant differences in flowering time ($F_{1,1806}$=19.6, P<0.001), plant height ($F_{1,1806}$ =23.3, P<0.001) and spike length ($F_{1,1806}$ =9.4, P≤0.002) between samples. A significant block effect ($F_{2,1806}$ =5.3, p≤0.005) was detected for flowering time only, but there was no interaction between blocks and samples ($F_{2,1806}$ =0.015, P=0.985). No other effect of blocks or block/sample interactions were detected for both other traits, plant height and spike length. We found an average difference between both samples with a shorter flowering time of 2.1 days, shorter plant size of 6.5 cm and shorter spike of 2.4 cm for plants generated from the 2003 sample versus those generated from the 1976 sample. We obtained similar results in the sampling site comparison, i.e. a significant decrease for average flowering time, (Fisher's combined probability, χ^2=37.0, n=6, P<0.001), plant height (Fisher's combined probability, χ^2=33.1, n=6, P<0.001), and spike length (Fisher's combined probability, χ^2=44.1, n=6, P<0.001). Finally, when we compared the same varieties sampled in 2003 and 1976 in the same village, we also observed a significant difference in flowering time (Fisher's combined probability, χ^2=35.4, n=6, P<0.001), plant height (Fisher's combined probability, χ^2=14.3, n=6, P<0.05) and spike length (Fisher's combined probability, χ^2=34.1, n=6, P<0.001). Using paired varieties, the estimated average changes in flowering time was 1.4 days, in plant height 4.7 cm and in spike length 3.34 cm. All these results suggest that conservation effect has no major impact on our initial results based on the seeds directly collected in the fields.

Analysis of the number of intra-varietal weedy plants.

A second confounding factor is the changes in frequency of weedy pearl millet plants. Weedy plants are commonly found in pearl millet seedlots in Niger (15). Weedy plants are characterized by shorter ears, thinner stems, branched stems, partial or total shattering phenotype and long bristles (Fig. S2) and earlier flowering. We wanted to assess if the number of weedy plants had increased or not between the 1976 and 2003 samplings. The number of weedy plants (Fig. S3) was higher in the 1976 sample than in the 2003 sample (Field trial 2004, G=17.3, p<0.009; Field trial 2005, G=19.7, p<0.004). This result suggests that there was a significant decrease in the number of weedy plants in 2003 compared to 1976.

Analysis of correlations between morphological traits and environmental data.

We also assessed for each sampling plot if there is a correlation between morphological traits and average rainfall (Fig. 4). We noted first that average rainfall values were highly correlated between the 1950-1976 and 1977-2003 periods (r=0.99, t=67.3, n=79, p<<0.001). We found a significant correlation between average rainfall and flowering time for the 2003 sample (r=0.72, t=9.1, n=79, P<0.001) and the 1976 sample (r=0.60, t=6.5, n=78, P<<0.001). The correlation between average rainfall and spike length was r=0.54 for the 2003 sample (t=5.6, n=79, p<0.001) and r=0.46 for the 1976 sample (t=4.6, n=78, p<0.001). Finally, plant height was also significantly correlated with the average rainfall for both the 2003 (r=0.62, t=7.0, n=79, P<0.001) and 1976 samples (r=0.49, t=4.6, n=78, P<0.001). No significant correlation differences between 1976 and 2003 were noted for flowering time (t=1.36, p=0.17), spike length (t=0.60, p=0.54) or plant height (t=1.17, p=0.24).

Genetic comparison of 1976 and 2003 samples. The genetic differentiation (F_{ST}) between the two samples (1976 vs 2003), by measuring the difference in allele frequencies at microsatellite loci, was very low (F_{ST}=0.0015) but statistically significant (P<0.001). A principal component analysis (PCA) showed slight genetic differences between the 1976 and 2003 samples. The difference between the two samples was not significant on the first PCA axis (Mann-Whitney test, χ^2=3.62, p=0.057) but was significant on the second axis (Mann-Whitney test, χ^2=5.26, p=0.022). The average allelic richness was 8.29 (SE 1.32) and 8.39 (SE 1.20) alleles per locus for the 1976 and 2003 samples, respectively. (Wilcoxon paired test, Z=-0.37, n=25, p=0.71). The average genetic diversity was 0.487 (SE 0.052) and 0.488 (SE 0.051) for 1976 and 2003, respectively. The number of alleles (Wilcoxon paired test, Z=-0.37, n=25, p=0.71) and genetic diversity (Wilcoxon paired test, Z=-0.21, n=25, p=0.83) did not differ between the two samples.

Finally, if selection plays a role in the present evolution, frequencies of alleles associated with quantitative variation in pearl millet are expected to change. We tested the change in allele frequency at the flowering locus *PHYC* (16) between the two samples. The earliness allele at the *PHYC* locus increased (G test, p<0.001) from 9.9% to 18.3% from 1976 to 2003 (Fig 5A).

To assess if the differentiation observed at the *PHYC* locus exceed sampling and drift between the two samples, we first constructed an empirical F_{ST} distribution. Finally, we analyzed population structure to also build a model based F_{ST} distribution. Using STRUCTURE (17-18), the likelihood is higher for K=1 for the 1976 sample and ancestry at K=2 do not show any clear population structure. The likelihood for the 2003 sample is higher for K=2, however no clear genetic group is observed in terms of ancestry (Fig. S4). Using STRUCTURAMA (19) with the 1976 sample, we found that the probability of having one population knowing the dataset (X) was Pr[K =1| X]=1. For the 2003 sample, Pr[K =1| X] was 0.9998. So for the two samples, K=1 is the most likely result. So in the subsequent analysis we considered the two samples unstructured. Considering the two populations unstructured, we estimated the population effective size at 12813 with a 95% lower bound confidence interval of 7248 and a 95% higher bound confidence interval of 34280 (Fig. S5). We used this effective size to build a model base F_{ST} distribution. We found that the allele frequency difference at the *PHYC* locus (Fig. 5B) exceeds the simple effect of drift and sampling based on empirical (P<0.02) or model-based distribution of differentiation (p<0.001). These results suggest a positive selection for the earliness allele.

Discussion

Climatic variation and sampling.

Western and Eastern Africa have experienced a significant decrease in rainfall in the last four decades. The origin of this drier period is assumed to be partially driven by abnormally warmer temperatures in tropical oceans (20-21), but also by vegetation dynamics and anthropogenic changes in land use (22-23). Sahel droughts began rather abruptly in the early seventies (22-23). The initial sample for this study was collected in 1976, a few years after the serious Sahelian droughts of the early 1970s. This initial sampling was meant to minimize the impact of these experienced droughts (13). The comparison of the 1976 variety distribution and composition was very similar to the 2003 distribution (14), and in agreement with the findings of the distribution study carried out in 1950 (24). Moreover, only slight overall genetic differences were noted between the 1976 and the 2003 samplings. We stress that sampling is a key component to allow a meaningful comparison of varieties over time. In 2003, we were careful to copy the initial sampling strategy in order to avoid or at least minimize the impact of differences in sampling processes on our results.

Crop adaptation to climatic variation

Globally the frequencies of the most important groups of varieties were not statistically different between 1976 and 2003 (14). However, there was declining trend in the relative occurrence of the later flowering variety group from 18.2% in 1976 to 10.2% in 2003 (14). Some earlier relatively minor varieties have also shown a geographical spread (14). The disappearance or spread of varieties certainly played a role in the actual changes that took place between 1976 and 2003, and is directly shaped by human choice in the varieties they cultivated. However, when we compared samples of the same varieties in the same village in 1976 and 2003, we found a significant change, which suggests that some changes in the adaptation of the varieties had actually occurred. Human and environmental constraints have certainly paving the way to this adaptation. We showed that this effect could not be explained by seed conservation effects, an increase in the occurrence of weedy pearl millet plants or sampling effects alone. We observed a correlation between pearl millet variety phenology and annual rainfall. We proposed that the observed change in annual rainfall is associated with the selection for earlier flowering varieties. The flowering time variation on a 27 period was an overall decrease of 1.44 days. Recent climatic middle of the road scenario suggests a change of the rainy season of 7 days for the end of the 21th century (8). So, by a simple extrapolation, a change of 1.9 days of the growing season in 27 years. So adaptation of local variety could be a significant contribution to adaptation of varieties to project climate changes.

However, adaptation to future climate will depend on the type of selection imposed by the environment (25). Decrease in rainfall might lead to selection favoring an escape strategy (i.e. rapid flowering) or water use efficiency (25). Recent models suggest that climatic change in the Sahel leads to shortening of the rainy season (8) while the total annual rainfall predictions are not very reliable (6). Rapid flowering or evolution of photosensibility might be an evolutionary road to adaptation to short rainy season. Finally, selection on a particular gene might also be hampered by pleiotropic effects (26). We actually observed that the earlier flowering *PHYC* allele was also associated with shorter spike length (Saïdou et al. 2009). This earlier flowering allele therefore certainly had an adverse fitness effect (and so a yield effect). However, this pleiotropic effect has not impaired adaptation from 1976 to 2003.

Some recent study investigated the role that varieties' substitution could play for a response to future climatic change. Surprisingly, adaptation without substitution in the present case seems a significant strategy. One of the possible reasons might be the out breeding reproduction system of pearl millet, allowing a strong intra-varietal variability to exist (12). This strong diversity inside a given varieties allows adaptation on standing variation. It remains to be assessing if the adaptation we observed in pearl millet is also observed for selfing Sahelian crop like sorghum. Whatsoever, this study illustrated how the diversity found in local landraces could be an asset to response to a variable environment.

Materials and Methods

Rainfall data.

The isohyet estimation was performed as previously described (27) for two periods: 1950-1976 and 1977-2003. Isohyets across Niger were built using datasets from the Centre Regional de Formation et d'Application en Agrométéorologie et Hydrologie Opérationnelle (AGRHYMET) international research center in Niamey, Niger. The study area was limited to Niger: longitudes 0° - 16°E and latitudes 12°N – 18°N and concerned 65 rainfall stations. Rainfall estimations for each sampling site were inferred according to their latitude and longitude coordinates, and the isohyets estimated for the two periods. We used the data obtained for each period 1950-1976 and 1977-2003, to estimate the average rainfall for each sampling site.

Seed collection.

Millet landrace accessions were sampled from 27 October to 26 December 1976 in Niger (13). These samples were conserved at 4°C at IRD, Bondy and then Montpellier, France, without multiplication. Out of the 184 villages investigated in 1976, we selected 79 villages. These villages were distributed evenly throughout the country (Fig. S1). A set of 192 different samples from 79 different villages was available from the 1976 initial sampling trip (Fig. S1).

In 2003, a new sampling survey (14) was carried out in the same 79 villages and 420 different accessions were collected (Fig. S1). The 2003 sampling scheme was modeled on the 1976 sampling scheme. Dr Borgel who performed the 1976 sampling helped in designing the 2003 sampling scheme. In 2003 and 1976, sampling was carried out in fields or village granaries after harvest. When possible, 30 different spikes were collected from a single farmer population for each variety name grown in the village. In 1976, villages having seeds of unknown origin due to complete crop failure in previous years were not sampled (13). The 2003 samples were conserved at 4°C before the July 2004, 2005 and 2007 field trials.

Out of the 192 varieties sampled in 1976, some varieties having the same name and thus sharing common characteristics (spike size, seed color, flowering behavior) were also found in 2003 in the same villages. We built a file associating a variety sample in the same village in 1976 and 2003 based on shared names, and for some troublesome accessions on morphology observed in the field. A total of 136 paired varieties were kept (Table S2).

We estimated the germination rate of a random sample of 50 accessions for the 2003 collections and 48 accessions for the 1976 collection. A hundred seeds per accession were placed in a petri dish and the number of germinated seeds after 3 days was recorded. The germination rate ranged from 0 to 1. To compare the germination rate between samples, we used an arcsin square root transformation of this variable. Differences in the transformed variable were then assessed using a t-test. The average germination rate is 87.2% for the 1976 sample and 90.0% for the 2003 sample. The difference in germination rate is significant ($t=1.62$, $p\leq0.044$), but is low (2.8%). This slight difference in germination rate for seeds preserved at 4°C for 27 years shows that the conservation conditions were very good. We also estimated the hundred seeds weight of a set of 129 of the 136 paired varieties (Table S2). The two samples were compared (1976 vs. 2003) using a paired Wilcoxon nonparametric test. Seeds from the 1976 sample are significantly heavier than seeds from the 2003 samples (Wilcoxon paired test, $n=129$, $Z=3.51$, $P<0.001$). The average 100-seeds weights are 1.037 g (SE ±0.015) for the 1976 sample and 0.980 g (SE ±0.013) for the 2003 sample. We first performed a field experiment using these initial seeds.

Field experiments.

Seeds were planted in the field station of the International Center of Research for the Semi-Arid Tropics (ICRISAT), Sadore, Niger. The planting dates were 7 July 2004, 7 July 2005 and 23 June 2007. Plants were sustained by natural rainfall and irrigated if needed at the beginning or end of the rainy season. For each seedlot accession, 25 plots were sown in 2004 and 2005, 15 plots in 2007. Around 30 seeds per plot were sown and hand-thinned to three plants 2 weeks after emergence. Four weeks after emergence, plants were hand-thinned to only one plant. Plants were spaced 0.7 m apart within rows and 0.7 m apart between rows. Field borders were planted with three to six rows of sorghum or pearl millet varieties to avoid side effects. Plants were hand-weeded twice at the beginning of the experiment. Treatments against mildew and insects were performed when needed. In 2004, the experiment was a pseudo-random experiment where villages were randomly picked and then samples of 1976 and 2003 for a given village were sown aside**.** The 2005 and 2007 experiments were complete random experiments where each sampling site and sample was randomly picked. Morphological data of five plants randomly chosen per accession were recorded individually in 2004 and 2005, and of 15 plants in 2007. We recorded the flowering time from planting date to the female flowering stage, plant height and spike length for the three field experiments. In 2004 and 2005, we also recorded the hundred seed weight. In 2004 and 2005, the number of weedy plants out of 25 plants per accession was assessed by two independent observers. The number of accessions with 0, 1, 2, 3, 4, 5 and 6 or more weedy plants was calculated for the 2003 and 1976 samples and for the 2004 and 2005 field trials. The distributions were then compared for significance using a G test (28).

In 2004, five plants out of a total of 607 accessions were selfed and individually harvested. These plants were used in 2007 to perform a new field experiment with three repetitions. The germination rate for these selfed seeds was assessed on 50 random seedlots derived from the 2003 sample and 50 random seedlots derived from the 1976 sample. Each individual’s selfed progeny (five per accession) was planted in Sadore ICRISAT field station on June 2007. Three repetitions of the experiment were performed and planted on 15 June 2007, 16 June 2007 and 19 June 2007. Days from planting to female flowering, plant height and spike length at maturity were recorded for each individual plant.

Statistical analysis.

To compare morphological differences between the 1976 and 2003 samples, we first calculated, for each field trial (2004, 2005, 2007), an average value for each accession.

We then compared the two samples (1976/2003) using an analysis of variance : $y_{ijk}=\mu+s_i+b_j+ s_i*b_j +\varepsilon_{ijk}$ where y_{ijk} is the phenotypic value, μ is the overall mean, s_i is the sample effect (1976/2003), b_j is the repetition effect (year or block), s_i*b_j is the sample and repetition interaction effect and ε_{ijk} is the residual error.

We also compared, for a given sampling site (village), the differences in phenological and morphological traits. To do so, for each sampling site we calculated an average value for the different morphological traits, i.e. the mean of all varieties sampled in each village. Comparisons between the two sets of sampling sites (1976 vs. 2003) were performed using a paired Wilcoxon nonparametric test for each field trial. Overall significance was obtained using Fisher's combined probability test (28).

A total of 136 varieties sampled in 1976 were found in the same village in 2003. We performed an analysis based on these paired varieties. Average values for each morphological trait (flowering time, plant height and spike length) were compared by a Wilcoxon paired test for each field trial. To obtain an overall test effect across repetitions (field trial for the original seeds or block for selfed seeds), a Fisher's combined probability test was performed (28).

Finally, we calculated a correlation between the annual rainfall previously calculated and the average morphological and phenological traits of each sampling site for the 1976 and 2003 samples, respectively. The correlation was tested using a t-test (28).

DNA extraction and Genetic analysis.

For each accession, DNA was extracted from one seed (29). DNA was dispatched on 96-well plates and on each plate two negative controls were included (no DNA). A set of 25 microsatellite loci were used to genotype the different accessions as previously described (29-30). Ten out of the 25 used microsatellite loci are mapped: PSMP2237, PSMP2201, PSMP2206 (linkage group 2), PSMP2216, PSMP2214 (linkage group 3), PSMP2005 (linkage group 4), PSMP2208, PSMP2202, PSMP2202 (linkage group 5) and PSMP2266 (linkage group 7). These microsatellite loci are spread throughout the pearl millet genome. We also genotyped a SNP at the *PHYC* gene (16). A fragment at the 5' of the gene was amplified and a restriction assay using PvuII was performed to recognize a C/G SNP (see 16 for primer sequences, PCR and digestion conditions). Genotypes were scored as C/C, G/G and C/G according to the digestion pattern. For microsatellite loci, each PCR product was migrated on an ABIPrism 3100 sequencer and the scoring was manually checked by two different persons. *PHYC* assay was scored on agarose gel (16).We used allelic richness to compare the number of alleles between the two samples which presented a different number of individuals. Allelic richness (R) was calculated using the formula:

$$R = \sum_{i=1}^{k} \left(1 - \frac{C_{2N-N_i}^{2n}}{C_{2N}^{2n}} \right)$$

where N_i was the number of i alleles in the population of the largest size N (2N chromosomes), and n was the number of individuals analyzed for the smallest population (2N chromosmes), and k was the total number of alleles for the locus studied (31). Allelic richness and gene diversity were calculated for the 1976 and the 2003 samples using FSTAT (32) and compared using a non-parametric Wilcoxon paired test. A principal component analysis was also performed on the allele frequencies of each individual. Differentiation (F_{ST}) between the two samples was calculated using Power marker (33). Population structure for the two samples was assessed using STRUCTURE V2.3.1 (17-18), based on the 25 microsatellite markers. Burnin period was set to 50000 and run length to 100000. The number of assumed populations (K) varied from 1 to 5. Five different runs were performed for each K value. The highest likelihood was used for further analysis and ancestry plotting. To assess the number K of populations, we also used STRUCTURAMA (19). The analysis is based on STRUCTURE (17) but the number K of populations could be sampled in a Dirichlet distribution. We used the software with the number K of populations sampled in a Dirichlet distribution, with a mean number of expected populations of 5. The output of the analysis is the probability to have i populations knowing the dataset Pr[K = i | X]. The analysis was run with 100000 MCMC steps and 5 chains.

Analysis of differentiation at the PHYC locus.

We first calculated the differentiation for each microsatellite allele between the two samples 1976 and 2003 using Powermarker (33). The distribution obtained is referred to as the empirical distribution. The differentiation obtained for the *PHYC* locus was ranked in this distribution. This rank gave an empirical p-value.

For the simulation, we used the two temporal samples (1976 and 2003) to calculate a population effective size N_e based on the 25 microsatellite loci. We used the approach of Wang (34-35) to obtain a likelihood point estimate of N_e the effective size and the distribution of the log-likelihood around this estimated effective size. We then simulated a F_{ST} distribution between two datasets sampled 27 generations apart considering only the drift effect associated with the whole population estimated effective size (N_e). For this simulation, we used a classical Wright-Fischer model assuming a binomial law to draw individuals from one generation to the next. We considered bi-allelic loci (SNP) with a uniform frequency distribution in the initial population and simulated the frequency of the allele in the last population drawing a binomial law to update the frequency from one generation to the next. We then simulated a dataset based on the initial population and the final population allele frequencies, and the number of individuals sampled in our real data set in these two populations. So this simulation took into account both drift and sampling effect. We then used these two samples to calculate the differentiation F_{ST} (36). A hundred thousand different simulations were performed, and a hundred thousand F_{ST} were calculated. We obtain a distribution of F_{ST} for bi-allelic markers. These simulations were performed by developing R code. The *PHYC* F_{ST} was then compared to this simulated distribution, and the probability to observe this particular value was estimated. We used the rank of the *PHYC* F_{ST} in the distribution and reported this rank as a p-value by dividing the number of simulations (100000).

Acknowledgments

We are grateful to the farmers of Niger for their willingness to participate in this experiment. We are grateful to IRD for logistical support in Niamey and Montpellier over the course of this project.

Author contributions

Conceived and design the experiment: YV, JLP, BG, IK, FS, MD, JC, JJ, GB. Perform experiment: YV, CM, VL, ACT, AAS, GB. Contribute reagents/materials/analysis tools: AA. Performed statistical analysis: YV, CM, VL. Wrote the paper: YV, ACT, JLP.

References

1. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, et al. (2010) Food security: the challenge of feeding 9 billion people. Science 327: 812-818.

2. Morton JF (2007) The impact of climate change on smallholder and subsistence agriculture. Proc Natl Acad Sci U S A 104:19680-19685.

3. Howden SM, Soussana JF, Tubiello FN, Chhetri N, Dunlop M et al. (2007) Adapting agriculture to climate change. Proc Natl Acad Sci U S A 104 : 19691-19696.

4. Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP et al. (2008) Prioritizing climate change adaptation needs for food security in 2030. Science 319: 607-610.

5. Brown ME, Funk CC (2008) Food security under climate change. Science 319:580-581.

6. Cook KH (2008) The mysteries of Sahel droughts. Nature Geoscience 1: 647-648.

7. Biasutti M, Held IM, Sobel AH, Giannini A (2008) SST forcings and sahel rainfall variability in simulations of the twentieth and twenty-first centuries. J Climate 21: 3471-3486.

8. Biasutti M, Sobel AH (2009) Delayed Sahel rainfall and global seasonal cycle in a warmer climate. Geophys Res Lett 36: L23707.

9. Challinor AJ, Ewert F, Arnold S, Simelton E, Fraser (2009) Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation. J Exp Bot 60: 2775-2789

10. Piao S, Ciais P, Huang Y, Shen Z, Peng S, et al. (2010) The impacts of climate change on water resources and agriculture in China. Nature 467: 43–51

11. Ndjuenga J (2002) Local village seed system and pearl millet seed quality in Niger, Exp Agr 38: 149-162.

12. Allinne C, Mariac C, Vigouroux Y, Bezançon G, Couturon E et al. (2008) Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger, Genetica 133: 167-178.

13. Borgel A, Sequier J (1977) Prospection of pearl millet and sorghun in West-Africa. Field trip of 1976 in Niger (french). ORSTOM ed., Paris.

14. Bezançon G, Pham JL, Deu M, Vigouroux Y, Sagnard F et al. (2009) Changes in the diversity and geographic distribution of cultivated millet (*Pennisetum glaucum* [L.] R. Br.) and sorghum (*Sorghum bicolor* (L.) Moench) varieties in Niger between 1976 and 2003, Genet Res Crop Evol 56: 223-236.

15. Mariac C, Robert T, Allinne C, Remigereau MS, Luxereau A et al (2006) Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. Theor Appl Genet. 113:1003-1014.

16. Saïdou AA, Mariac C, Luong V, Pham JL, Bezançon G et al (2009) Association studies identify natural variation at *PHYC* linked to flowering time and morphological variation in Pennisetum glaucum [(L.) R. Br.], Genetics 182: 899-910.

17. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

18. Falush D, Stephens M, Pritchard JK (2003) Inferences of population structure using multilocus genotype data : linked loci and correlated allele frequencies. Genetics 164:1567-1587.

19. Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. Genetics. 175:1787-1802.

20. Funk C et al. (2008) Warming of the Indian Ocean threatens eastern and southern African food security but could be mitigated by agricultural development. *Proc Natl Acad Sci USA* 105:11081-11086.

21. Giannini A, Saravanan R, Chang P (2003) Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. *Science* 3021027-1030.

22. Foley, J.A., M.T. Coe, M. Scheffer, and G. Wang (2003). Regime Shifts in the Sahara and Sahel interactions between ecological and climatic systems in Northern Africa. *Ecosystems* 6:524-539.

23. Narisma GT, Foley JA, Licker R, Ramankutty N (2007) Abrupt changes in rainfall during the twentieth century. *Geophys Res Lett* 34, L06710, doi:10.1029/2006GL028628.

24. Marshal A. (1950) Pearl millet varieties cultivated in Niger (french). *L'agronomie tropicale* 5 : 582-592.

25. Hoffmann AA, Willi Y. (2008) Detecting genetic responses to environmental change. Nat Rev Genet. 9:421-432.

26. Scarcelli N, Cheverud JM, Schaal BA, Kover PX (2007) Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci U S A 104:16986-16991.

27. Ali A, Lebel T, Amani A (2005) Estimation of rainfall in the sahel. Part 1 : error function. J Appl Meteorology 44: 1691–1706.

28. Sokal RR, Rohlf FJ (1995) Biometry. Ed. WH Freeman and Co., NY.

29. Mariac C., Luong V, Kapran I *et al.* (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**, 49-58.

30. Oumar I, Mariac C, Pham JL, Vigouroux Y (2008) Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. Theor Appl Genet. 117:489-497.

31. Petit RJ et al. 1998. Conserv Biol 12:844–855.

32. Goudet J (2002) FSTAT, a program to estimate and test gene diversity and fixation indices (version 2.3.9.2). Available from http://www.unil.ch./izea/softwares/fstat.html.

33. Liu K, Muse S (2005). PowerMarker : new genetic data analyis software. Version 3.25. free program distributed by author over the internet from http://www.powerMarker.net

34. Wang J (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet Res 78: 243-257.

35. Wang J (2005) Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society Biological Sciences 360: 1395-1409.

36. Weir BS (1996) Genetic data analysis II. Ed. Sinauer Associates Inc. MA.

Figure Legends

Figure 1. Pearl millet cultivation in Niger

Pearl millet is planted at the beginning of the rainy season after a significant rain (a). Different traditional varieties exist in Niger and have different flowering phenotypes (b) from early flowering (b, left), to late flowering (b, right). Pearl millet is generally harvested in September or October in Niger (c), bundles of pearl millet spikes are stored in granaries during the dry season (c, d). After harvest, farmers often select the best spike to use as seed for the next planting year (e). Photo by Y.Vigouroux and C. Mariac.

Figure 2. Pluviometry isohyetes for 1950-1976 and 1977-2003 periods.

The pluviometry isohyets are calculated and plotted for the 1950-1976 periods (blue) and the 1977-2003 periods (red). The isohyets shift from 100 to 150km south between the two periods illustrating an average rainfall decrease.

Figure 3. Morphological changes in pearl millet varieties between 1976 and 2003.

The average values of morphological and phenological traits are plotted for each of the 79 villages sampled throughout Niger. Significant variations were observed between villages, i.e. varieties collected in 2003 flowered earlier, were shorter and had shorter spikes.

Figure 4. Correlation between average phenological and morphological traits and average annual rainfall.

The average flowering time (in days), average spike length (in cm) and average plant height (in cm) for each sampling plot and each year 1976 (red) and 2003 (bleu) are plotted against the annual rainfall (in mm). Annual rainfall is calculated for 1976 on the period 1950-1976, for 2003 on the period 1977-2003. Significant correlation is observed for flowering time, spike length and plant height for the 1976 samples (r=0.60, p<0.001; r=0.46, p<0.001; r=0.49, p<0.001 respectively) and the 2003 samples (r=0.72, p<0.001; r=0.54, p<0.001; r=0.62, p<0.001 respectively).

Figure 5. Selection for earlier flowering PHYC allele.

A significant increase (A) of the early flowering allele at the PHYC gene (SNP G) was observed between 1976 and 2003. The differentiation (FST) between the two samples for PHYC alleles exceeded the effect of drift or sampling (B) expected based on empirical (dark) or model-based (grey) FST distributions. These results suggest that selection lead to the increase of the early flowering allele at the PHYC locus.

Figure 1.

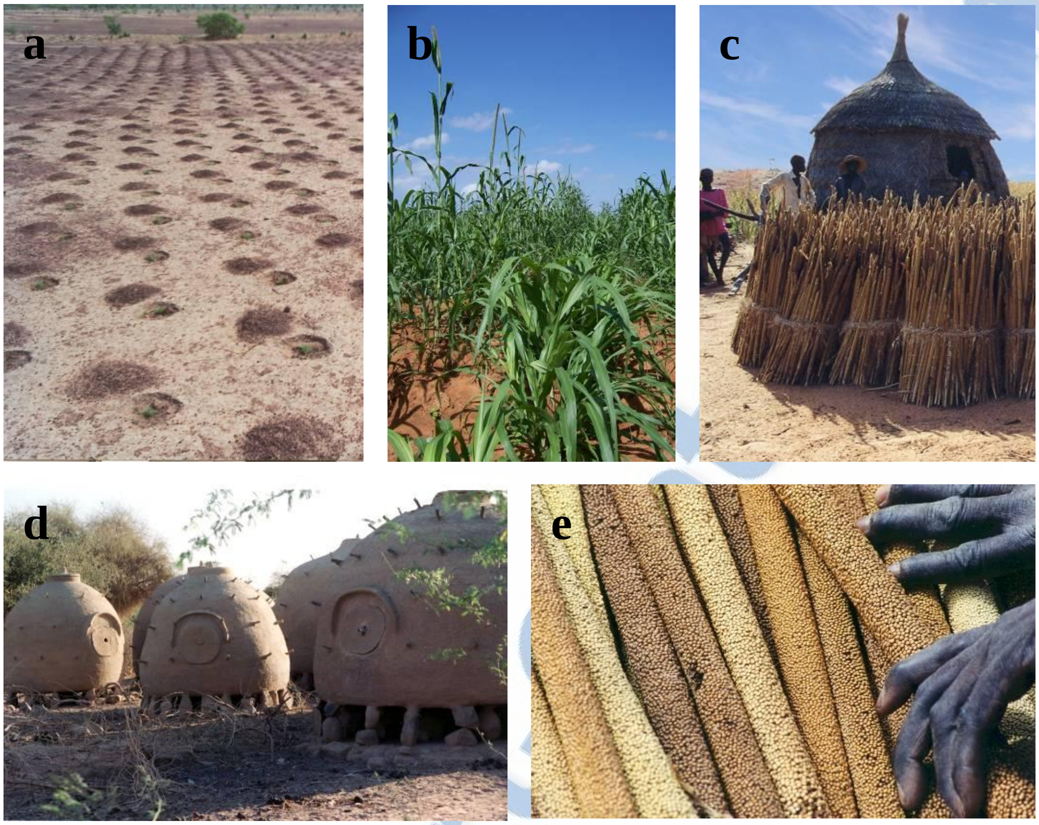


Figure 2.

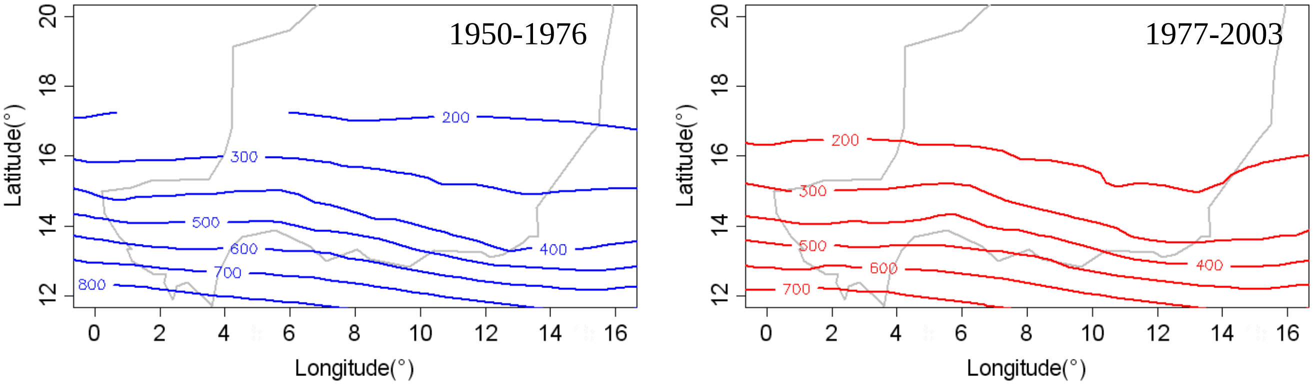

1950-1976
Latitude(°)
20
18
16
14
12
200
300
400
500
600
700
800
0
2
4
6
8
10
12
14
16
Longitude(°)
1977-2003
Latitude(°)
20
18
16
14
12
200
300
400
500
600
700
0
2
4
6
8
10
12
14
16
Longitude(°)

Figure 3.

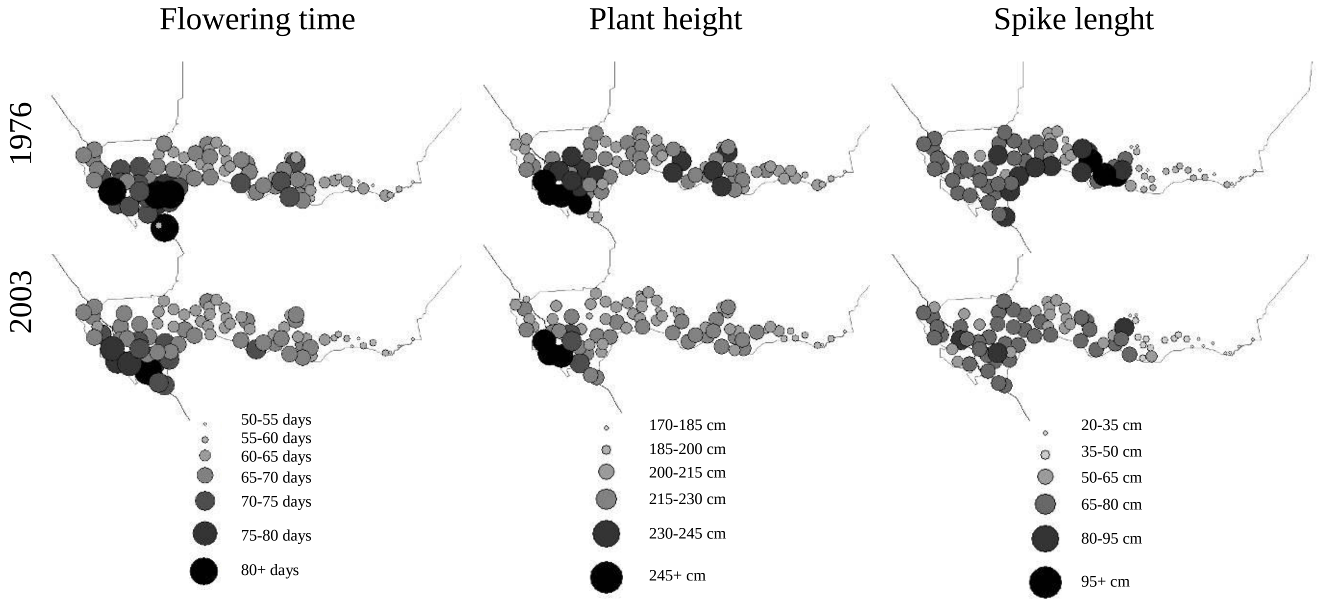

Flowering time
Plant height
Spike lenght
1976
2003
50-55 days
55-60 days
60-65 days
65-70 days
70-75 days
75-80 days
80+ days
170-185 cm
185-200 cm
200-215 cm
215-230 cm
230-245 cm
245+ cm
20-35 cm
35-50 cm
50-65 cm
65-80 cm
80-95 cm
95+ cm

Figure 4.

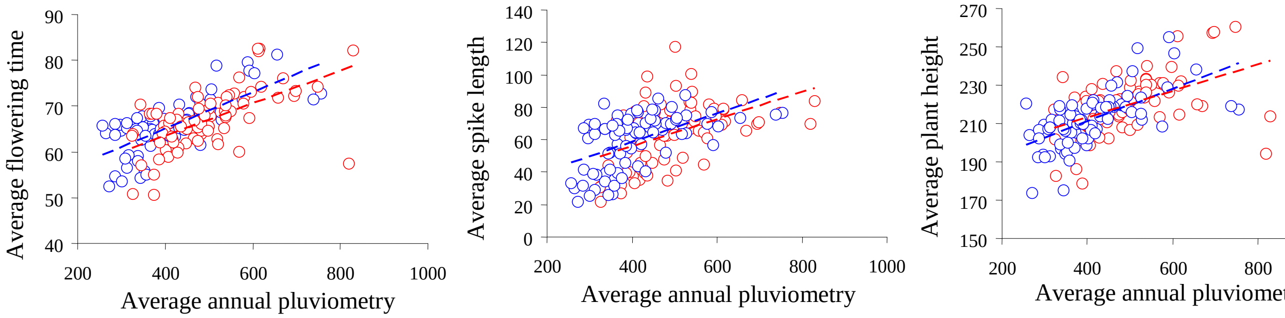
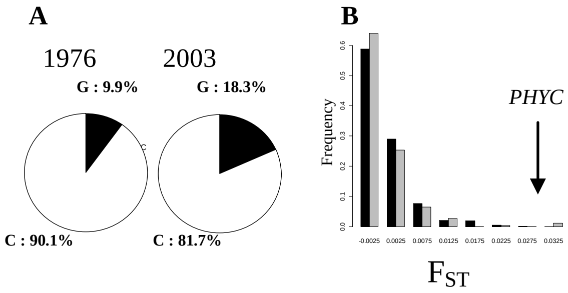

Average flowering time
90
80
70
60
50
40
200
400
600
800
1000
Average annual pluviometry
Average spike length
140
120
100
80
60
40
20
0
200
400
600
800
1000
Average annual pluviometry
Average plant height
270
250
230
210
190
170
150
200
400
600
800
1000
Average annual pluviometry

Figure 5.

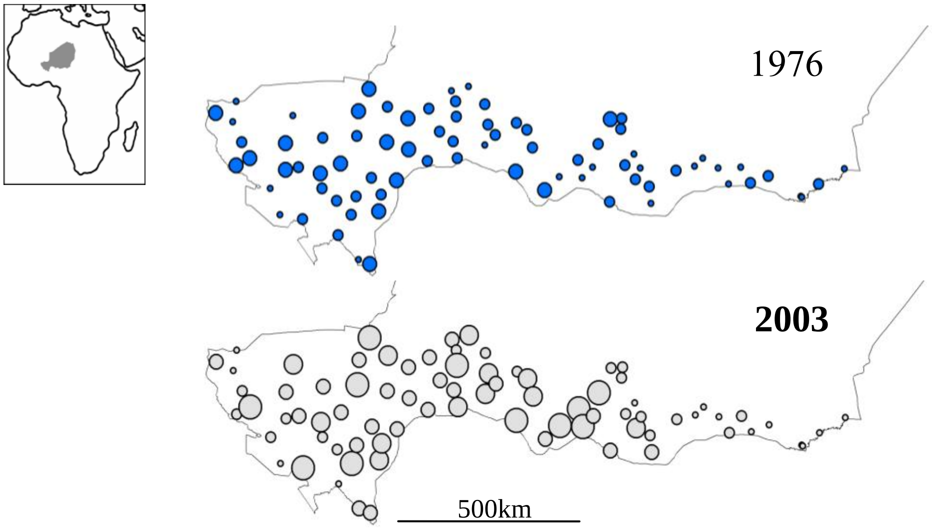

A
1976
2003
G : 9.9%
G : 18.3%
C : 90.1%
C : 81.7%
B
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
PHYC
-0.0025
0.0025
0.0075
0.0125
0.0175
0.0225
0.0275
0.0325
F_{ST}

Supporting information

Table S1 List of accessions from the 1976 and 2003 samples.
Joint file.

Table S2 List of varieties found in the same village in 1976 and 2003.
Joint file.

Fig. S1. Geographical sampling locations in Niger.



A total of 192 pearl millet accessions from 1976 and 420 accessions from 2003 were sampled in the same villages. The number of samples collected in a given village is represented by the size of the dot.

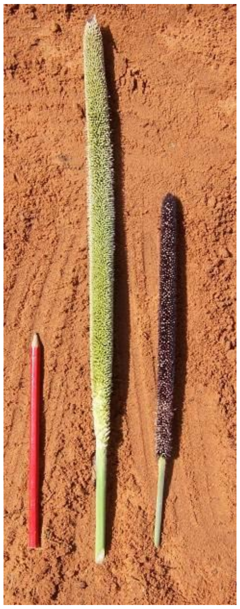
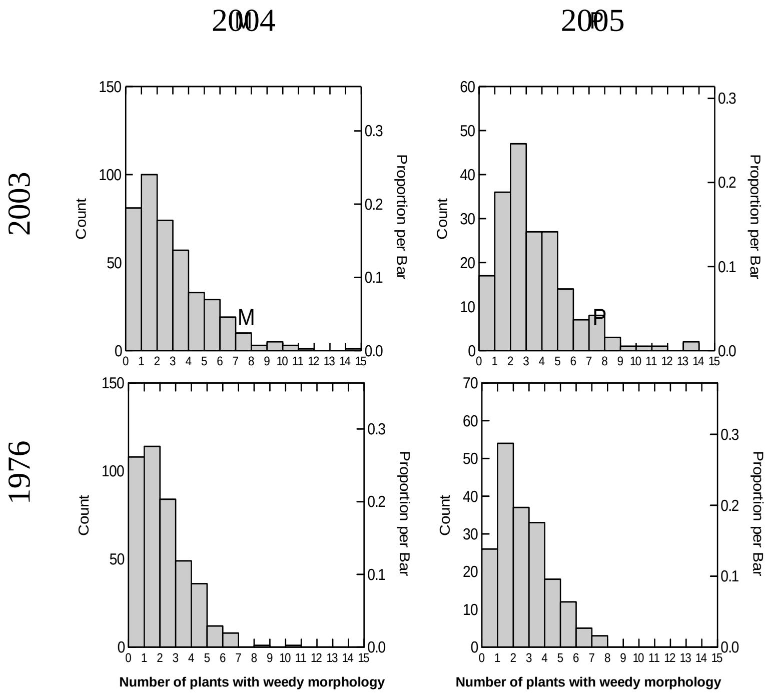


Fig. S2. Morphological differences observed between cultivated and weedy morphotypes.

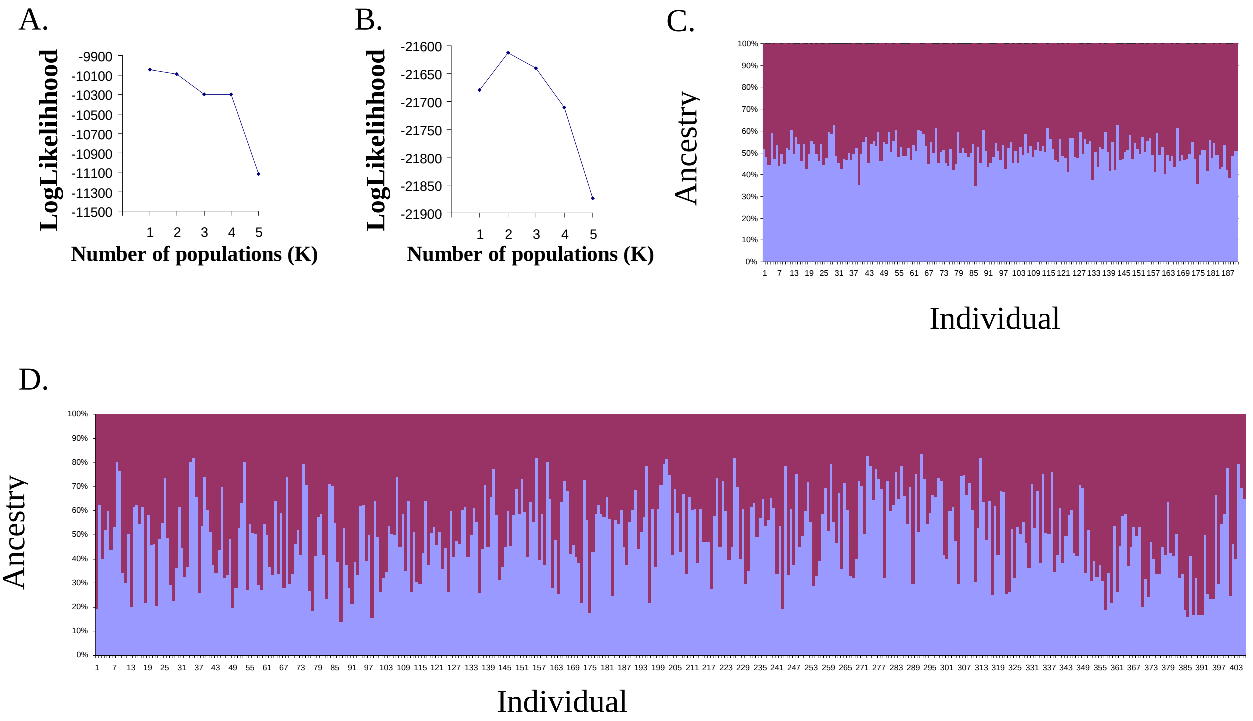
In this picture, two different individuals from the same accession seedlot (Pe02783) with a cultivated (left) and weedy spike morphology (right) are presented. Weedy plants are generally characterized by shorter ears, thinner stems, higher branching morphology, partial or total shattering and long bristles.

Fig. S3. Comparison of the number of weedy plants between the 2003 and the 1976 samples.



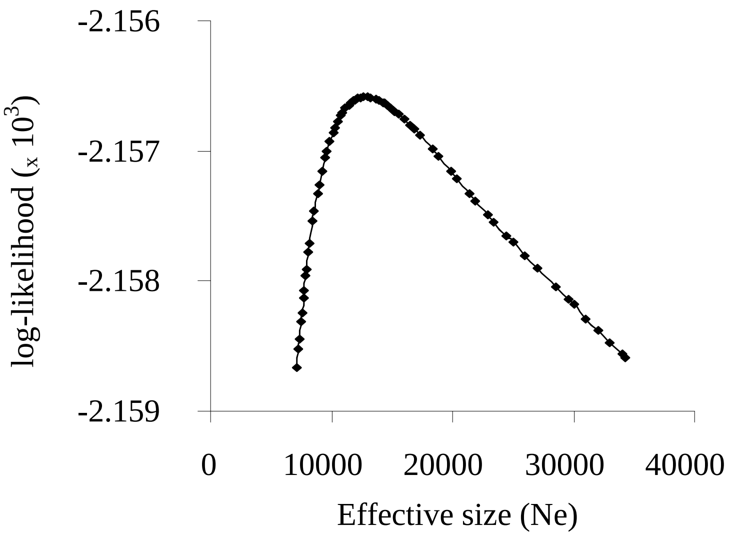
The number of weedy plants was assessed for each seedlot for the 2003 and 1976 samples. The figure represent the number of seedlot presenting zero, 1, 2 etc… weedy plants out of 25 individuals. Two field trials were performed one in 2004 and one in 2005.

Fig. S4. Population structure for the 1976 and 2003 samples.



The log-likelihood for different number of assumed populations is given for the 1976 sample (**A**) and the 2003 sample (**B**). The ancestry for two assumed populations is given for the 1976 (**C**) and 2003 (**D**) samples. Even if the likelihood is higher for K=2 for the 2003 samples, the ancestry did not show a clear two group ancestry structure. The structure signal detected for the 2003 sample is very weak so we considered the absence of population structure (K=1) as a likely hypothesis for the 2003 sample.

Fig. S5. Estimation of the effective size (N_e) based on the two temporal samples.



The analysis is based on allele frequency of the 25 different microsatellites for the two samples (1976 and 2003). The log-likelihood is reported for a different value of effective size. The highest log-likelihood (log-L=-2156.6) is observed for Ne=12813.

Table S1 List of accessions from the 1976 and 2003 samples.

Table S2 List of varieties found in the same village in 1976 and 2003.