Mémoire de fin de cycle

en vue de l'obtention du titre de

Master Professionnel, *Spécialité* **Sciences et Productions Végétales,**

Option **Amélioration des Plantes et Semences**,

Sous-option Amélioration des Plantes

Développement de méthodes de génétique d'association et application à l'analyse de la qualité et de la floraison chez le mil

Présenté par **SAIDOU Abdoul-Aziz**

Soutenu devant la commission d'examen le 20 septembre 2007 :

Membres du jury

Dominique BARLOY, Enseignante-chercheur à Agrocampus-Rennes
Maria MANZANARES-DAULEUX, Enseignante-chercheur à Agrocampus-Rennes
Yves VIGOUROUX, Chercheur à l'IRD, Maître de stage

Remerciements

Je remercie très sincèrement Dr Yves VIGOUROUX qui m'a encadré et suivi le long de ce stage. Ses conseils, ses critiques, ses corrections et tous les échanges réguliers avec lui au cours de ce stage ont certainement été la lumière qui a guidé et éclairé mes pas.

J'exprime aussi ma reconnaissance pour les membres de l'équipe d'accueil, qui m'ont offert la collaboration et l'ouverture, en particulier Cédric MARIAC, Djibo MOUSSA, Moussa TIDJANI, Pierre SIRE.

Mes remerciements vont aussi au représentant de l'IRD au Niger, M. Gilles Bezançon, et à tout le personnel à Niamey, qui m'ont offert un cadre agréable et une ambiance quotidienne favorable à un meilleur travail.

Merci également à Ibrahim AMOUKOU, Jean-Louis PHAM et Anne-Céline THUILLET pour leurs critiques constructives lors de la finalisation de ce travail.

Merci de tout coeur à mes amis, qui m'ont accueilli et accompagné lors de tout mon séjour à Niamey: Abass A.I., Zoulkarnaini A. W., Habibou D., Lawali D. , Mahamadou S., et tous les autres.

Merci aux amis de Rennes, pour l'excellente compagnie.

Très particulièrement, j'exprime ma profonde reconnaissance pour le corps professoral et le personnel de mon école à Rennes, qui ont su donner une formation de qualité. Très particulièrement, mes remerciements et ma profonde gratitude vont à l'endroit de Mme Dominique BARLOY et Mme Maria MANZANARES-DAULEUX.

Dédicaces

A ma mère Fatouma Sadou, l'être le plus cher, qui dans la vie m'a tout donné,

A la mémoire de mon père, Saïdou Danjima, dont le souvenir éclaire mon coeur,

A ma très chère tante Hadjia Zeinabou,

A mes aimables aînés: Abdou et sa femme Ousseina, Adama et son époux Dr ILLO A., qui ont toujours su m'apporter attention, conseil et soutien,

A tous mes frères et soeurs, l'energie et la joie de ma vie.

SOMMAIRE

PRESENTATION DE L'INSTITUT D'ACCUEIL

Créé en 1944, l'Institut de recherche pour le développement (IRD, ex ORSTOM) est un établissement public français à caractère scientifique et technologique (EPST), placé sous la double tutelle des ministères chargés de la Recherche et de la Coopération. L'IRD conduit des programmes scientifiques centrés sur les relations entre l'homme et son environnement dans les pays du Sud, dans l'objectif de contribuer à leur développement. Il remplit les missions fondamentales de : recherche, expertise et valorisation, soutien et formation, information scientifique. L'IRD mène des recherches en partenariat avec les acteurs scientifiques, sociaux et politiques des pays du Sud, d'où l'importance d'une représentation physique à l'étranger. Il dispose de 35 implantations dans le monde, dont 5 en France métropolitaine, 5 dans les ROM-COM, 25 représentations dans des pays étrangers. Les chercheurs de l'IRD interviennent dans une cinquantaine de pays. Les travaux effectués par les chercheurs de l'IRD sont coordonnés par trois départements scientifiques :

Milieux et Environnement (DME) : Les recherches visent à comprendre certains phénomènes comme la variabilité climatique, l'interaction entre océan et atmosphère… Une meilleure perception du climat permet d'évaluer ses effets sur les ressources en eau et végétales ainsi que sur les risques naturels de l'environnement tels que les séismes ou les volcans.

Ressources Vivantes (DRV) : Les travaux portent sur les ressources et écosystèmes des milieux naturels terrestres et des milieux aquatiques, continentaux et marins, dans une optique de développement et de gestion durables. Certaines unités de recherche se consacrent à l'amélioration des productions végétales et tropicales, d'autres à la défense des cultures contres les parasites et les prédateurs. Enfin, de nombreuses recherches portent sur l'écologie aquatique et les sciences de la pêche.

Sociétés et Santé (DSS) : Les études menées couvrent deux domaines, les sciences sociales et la santé ainsi que leur interface dans un large spectre de disciplines. Elles concernent les grandes endémies (dengue, paludisme, sida…), le développement urbain, la pauvreté et ses déterminants et plus récemment les questions relatives aux interactions société/risques environnementaux tels que la migration ou les conflits.

En 2006, l'institut a mobilisé un budget de 200 M€. Il totalisait plus de 2200 agents, dont 830 chercheurs, 1000 ingénieurs et techniciens et 400 personnels sur contrats locaux. La vocation de cet institut pour les partenariats avec les équipes du Sud se traduit notamment par l'affectation de 43 % d'es gents hors métropole (2006). En matière d'enseignement supérieur, 140 thèses étaient encadrées en 2006, et 6000 heures d'enseignement dispensées par des chercheurs et ingénieurs de l'IRD.

L'unité d'accueil pour ce stage est l'UMR Diversité et Adaptation des Plantes Cultivées (DIA-PC), rattachée au département des Ressources Vivantes (DRV). Cette UMR fédère des recherches conduites par l'IRD, l'INRA et l'université de Montpellier 2. L'implantation à Niamey (Niger) travaille depuis une longue période sur les espèces sahéliennes, notamment le mil et le sorgho, qui sont les deux principales cultures au Niger. Elle étudie la diversité de ces espèces, la gestion des ressources génétiques et la dynamique de cette diversité dans le temps et dans l'espace. Ces études se rapportent notamment à l'interaction avec les facteurs anthropiques, mais également avec les variations climatiques. Des études sont aussi menées pour comprendre les phénomènes évolutifs liés à la domestication et à la sélection, grâce aux empreintes moléculaires présentes dans les populations sauvages et cultivées. Des études sur les gènes d'intérêt sélectionnés au cours du temps sont également développées.

INTRODUCTION

Le mil (*Pennisetum glaucum* L.) fait partie des principales céréales de la zone sahélienne. Au Niger, le mil couvre plus de 60% de la surface cultivée et représente environ 73% de la production céréalière totale (IRD, 2004); il constitue l'aliment de base de la majorité de la population. Les superficies cultivées en mil couvrent à travers le monde 33,9 millions ha environ et s'étendent plus de 70 pays. Elles se repartissent principalement sur les zones arides et semi-arides d'Afrique et d'Inde (FAO, 2006b).

Le mil est une espèce diploïde (x = 7, 2n = 2x = 14). Son système de reproduction est marqué par une allogamie préférentielle, avec une protogynie fortement marquée (Bezançon et *al.,* 1994). Les cultivars adoptés sont en général des variétés populations locales (*landraces*), maintenues et sélectionnées par les agriculteurs. Le développement et la diffusion de cultivars modernes améliorés de mil restent encore limités au Sahel. Cette culture traditionnelle du mil fait face à de nombreux défis dont l'un est de parvenir à nourrir une population en forte croissance. Au Niger par exemple, la population a doublé au cours des 25 dernières années. Durant la même période la surface cultivée a elle aussi doublé alors que le rendement stagnait voire régressait. La possibilité d'extension de la surface cultivée étant de plus en plus limitée, la sécurité alimentaire du Niger se jouera par une diversification de l'agriculture et l'augmentation des rendements. L'agronomie a une large place à jouer dans ce progrès agricole, mais l'approche sera d'autant plus efficace si des variétés améliorées sont disponibles. Ce défi s'inscrit aussi dans un contexte climatique délicat. A l'échelle de la planète, une hausse moyenne de température d'environ 0.2°C par décade a été enregistrée ces trente dernières années (Hansen et *al.,* 2006). A ces changements globaux s'ajoutent des variations climatiques au niveau régional. Ces 30 dernières années ont été particulièrement sèches sur toute la zone sahélienne. On observe ainsi un déplacement des isohyètes d'un degré de latitude environ vers le Sud (~100 km par endroit) sur cette période comparée aux 30 années précédentes. A cette baisse de pluviométrie s'ajoute une variabilité de plus en plus forte dans la période d'installation de la saison pluvieuse, le raccourcissement de la durée agricole de la saison de pluies, et une sécheresse assez prononcée touchant surtout les mois de cœur de saison (Agrhymet et CIRAD, 2005). Des interrogations demeurent encore, si ce changement dans la zone sahélienne est juste une composante de variations cycliques régionales (et donc temporaire) ou s'il traduit les conséquences de changements globaux. Dans tous les cas, l'évolution adaptative des variétés en adéquation avec l'évolution de l'environnement est une condition *sine qua non* pour maintenir voir améliorer la production.

L'amélioration génétique sera facilitée chez le mil si des connaissances sont acquises sur le déterminisme de caractères d'intérêt agronomique. Parmi les nouvelles méthodologies permettant d'étudier de tel traits, l'utilisation de méthodes d'association phénotype/génotype au niveau populationnel (Thornsberry et *al.,* 2001) parait une voie intéressante. Ces méthodes peuvent permettre d'utiliser les connaissances déjà acquises sur d'autres céréales beaucoup plus étudiées comme le maïs et le riz.

Le présent travail a pour but le développement chez le mil de méthodes de génétique d'association, l'estimation du pouvoir de ces méthodes et leur application à des jeux de données disponibles. Une discussion autour des méthodes de génétique d'association auxquelles nous nous intéressons est développée dans le chapitre *Synthèse bibliographique*. Ce paragraphe ferra également le point des connaissances acquises sur le plan génétique concernant des voies de recherche développées par le laboratoire d'accueil. Les données génétiques disponibles concernent des gènes de la voie de floraison, un gène impliqué dans la voie de synthèse des protéines (Opaque2) et un gène impliqué dans la qualité de l'amidon (Waxy). Waxy et Opaque2 ont déjà révélé un signal de sélection lors de la domestication du mil, ce qui laissait présager un rôle dans les caractères sélectionnés (Lauret, 2006).

Notre démarche méthodologique se décompose en quatre étapes principales. La première comporte une analyse statistique de la variabilité phénotypique du mil pour les caractères étudiés (principalement floraison, taux de protéines, taux d'amylose), sur la base d'un dispositif d'essai comportant 3 réplications.

La deuxième étape est une analyse de la structure génétique (inconnue *à priori*) au sein de notre panel des lignées. L'étude est conduite par une approche comparée, qui met en œuvre 2 méthodes différentes d'analyses de structure : la méthode STRUCTURE largement utilisée actuellement (Pritchard et *al.,* 2000a; Falush et *al.,* 2003) et une méthode nouvelle (INSTRUCT) développée très récemment (Gao et *al.,* 2007). L'avantage est aussi de mettre en interface ces deux méthodes et d'étudier la corrélation entre leurs résultats respectifs.

Notre troisième étape s'intéresse à la mise en oeuvre d'une méthode d'association phénotype/génotype à partir du logiciel TASSEL. Ces méthodes sont relativement récentes et peu de recherches ont été menées pour connaître leur pouvoir. Nous avons donc mené une étude par simulation visant à évaluer de la capacité de la méthode à détecter une association existante (estimation de l'efficacité du test**)**. Enfin, nous mettons en oeuvre dans la dernière étape les tests d'association pour la recherche d'une corrélation entre les données génotypiques disponibles et les phénotypes d'intérêt.

1. Effet de la structure génétique et association

L'association entre variation génétique et variation phénotypique a été abordée chez les plantes en premier lieu par les analyses QTL (Doebley, 1991). Ces approches permettent un contrôle assez fin de l'histoire des croisements et plus généralement de l'histoire *évolutive* des ségrégations. Au niveau populationnel, les associations génotype/phénotype font face à un défi : la connaissance très imparfaite de l'histoire évolutive. Dans le cas de ces méthodes, l'histoire évolutive peut être limitée dans un premier temps à la connaissance de la structure génétique des populations. L'effet de la structure génétique des populations a constitué un problème sérieux à l'utilisation des approches d'association au niveau population chez les plantes cultivées (Pritchard et *al.,* 2001). Des populations soumises à un processus de migration, d'échange de gènes et aussi d'adaptation et de sélection se différencient et se structurent génétiquement. Cette structuration se caractérise par des différences en terme de fréquences d'allèles entre populations. Ces différences sont le jeu de forces neutres (migration, dérive, mutation) mais peuvent aussi pour certains loci être liés à une adaptation (sélection). Lorsque l'on analyse l'association entre une variabilité phénotypique et une variabilité génétique au sein d'un panel qui regroupe des individus appartenant à différents groupes génétiquement structurés, la structure génétique agit comme un effet confondant. Un allèle dont la distribution est liée à la structure génétique peut alors montrer une corrélation significative avec un caractère différenciant les populations sans que ne soit forcément vérifié la réalité biologique d'un lien fonctionnel entre le locus et le caractère. L'association statistique ne recoupe pas dans ce cas une réalité génétique.

Si les approches classiques (de type QTL) permettent un contrôle strict de l'histoire évolutive et évite ce problème de la structure génétique, les méthodes d'association au niveau population doivent y faire face. Très récemment, un développement novateur a été d'appliquer aux plantes des méthodes d'association au niveau population sans connaissance (*à priori*) de la relation entre plantes, de leur histoire évolutive (Thornsberry et *al.,* 2001). Un progrès important a été fait pour ces études; une méthodologie statistique a été développée pour la prise en compte de la structure des populations (Thornsberry et *al.,* 2001; Yu et *al.,* 2006).

Le nombre de populations structurées dans un échantillon n'est pas souvent connu *à priori*. La définition des populations sur des bases linguistiques, morphologiques ou géographiques ne

correspond pas toujours, non plus, à une structuration génétique. Deux groupes géographiquement proches peuvent être génétiquement distants, et inversement. Il est aussi difficile de détecter une structure en utilisant des caractères phénotypiques, surtout lorsqu'il s'agit de structure cryptique. Heureusement, des progrès méthodologiques récents permettent de définir la structure des populations à partir de données génotypiques multilocus (Pritchard et *al.,* 2000a ; Gao et *al.,* 2007). Aujourd'hui une des analyses les plus fréquentes se résume à définir le nombre de groupes d'individus différents (le nombre K de clusters) et l'association probabiliste des individus aux K groupes. Sachant que des migrations et des mélanges sont possibles dans l'histoire évolutive des populations, un individu peut ainsi posséder des fractions q de son génome provenant de ces K différentes populations (*ancestry*).

Il existe 2 groupes principaux de méthodes utilisées pour la détection de la structure génétique (revue par Pritchard et *al.,* 2000a). Les méthodes basées sur la distance (*distance-based methods*) partent du calcul d'une matrice de distance entre les individus pris 2 à 2 ; une représentation graphique (par exemple type *arbre*) peut par la suite permettre la détermination visuelle des clusters. Le second type d'approche est basé sur l'hypothèse selon laquelle les observations provenant de chaque cluster sont réparties suivant un modèle paramétrique (*model-based methods*). Dans ces approches, on infère la valeur des paramètres du modèle conjointement avec l'appartenance des individus aux clusters sachant leur génotype, en utilisant des méthodes statistiques et des algorithmes adaptés (*Markov Chain Monte Carlo*).

Nous avons opté dans le cadre de ce travail pour des méthodes de détection de structure basées sur des modèles. Dans ce type d'approche, les hypothèses qui sous-tendent le modèle sont clairement formulées. La validité des hypothèses peut être discutée, *à priori* et *à posteriori*, sur le plan biologique. Ceci selon les connaissances relatives sur les particularités du matériel biologique étudié (mode de reproduction, caractéristiques génétiques…) et sur la complexité du schéma écologique dans lequel aurait évolué ce matériel (relations familiales, flux de gènes entre populations…). Les modèles reposent sur les méthodes de génétique des populations, ce qui rend même les résultats plus interprétables biologiquement. Ces méthodes de détection de la structure des populations sont basées sur une approche bayésienne (Pritchard et *al.* 2000a, Falush et *al.* 2003 ; Gao et *al.* 2007). Elles permettent d'estimer l'assignation des individus aux populations et les fréquences alléliques dans chaque population à partir du génotype multilocus des individus. Cette approche a été implémentée sous le logiciel STRUCTURE (Pritchard et *al.,* 2000a). Tous les modèles implémentés sous le logiciel STRUCTURE (Pritchard et *al.* 2000a, Falush et *al.,* 2003) supposent l'hypothèse de l'équilibre de Hardy-Weinberg au sein des populations, c'est-à-dire une reproduction au

hasard au sein de chaque groupe. Une option possible à prendre en compte est de considérer (ou non) les individus comme des individus potentiellement hybrides entre les K populations (*admixture model*). Enfin une dernière option est de considérer qu'il existe une corrélation entre les fréquences alléliques des différents loci. Cette corrélation modifie les fréquences alléliques par rapport à ce qui est attendu sous l'hypothèse d'indépendance des loci. La corrélation entre allèles de deux loci est dénommée déséquilibre de liaison (LD). Cette corrélation entre loci peut être physique (loci proches sur un même chromosome) ou simplement statistique (des loci sur différents chromosomes peuvent aussi être en LD). Trois sources de LD peuvent être distinguées : *mixture LD, admixture LD* et *back-ground LD* (Falush et *al.,* 2003). Un individu provenant du mélange (croisement) d'individus de 2 ou plusieurs populations présentera un excès d'association d'allèles communs dans ces populations. Ce mélange crée une corrélation des fréquences alléliques, même lorsque les loci concernées ne sont pas physiquement liés (*mixture LD*). Le LD présent au sein même des populations (*back-ground LD*) est lié simplement au jeu de la recombinaison, de la mutation et de la dérive en population à l'équilibre. Il diminue en général à une échelle plus courte, quelques centaines de paires de bases par exemple pour le maïs (Tenaillon et *al.,* 2001). Enfin, certains fragments chromosomiques comportant des loci liés introgressées au sein d'une population, sont transmis en bloc (*chunks*) à la descendance. Ceci cause un excès de DL par rapport au *back-ground LD*, ce déséquilibre additionnel est le DL d'introgression (*admixture LD*). Les méthodes d'association consistent à prendre en compte (contrôler) le *mixture LD* et *admixture LD* pour pouvoir associer la variation chromosomique locale (*background LD*) avec un phénotype.

Cependant, la reproduction par autofécondation n'est pas prise en compte par la méthode du logiciel STRUCTURE. L'autofécondation conduit à une modification des fréquences génotypiques au sein des populations. La présence d'un effet d'autofécondation peut conduire, avec la méthode STRUCTURE, à de faux signaux de structure génétique et de faux signaux de mélange entre populations (Gao et *al.,* 2007). Récemment une nouvelle méthode a été élaborée, qui permet une analyse en prenant en compte l'autofécondation (Gao et *al.,* 2007). A la différence de STRUCTURE, INSTRUCT infère le taux d'autofécondation et définit les fréquences alléliques sur la base de la consanguinité plutôt que sur la base de l'équilibre de Hardy-Weinberg.

2. Les voies métaboliques étudiées

La floraison. Une des plantes modèles des études de la floraison reste aujourd'hui la dicotylédone *Arabidopsis thaliana*. Des études ont montré une forte conservation de la fonction générale des gènes entre monocotylédones et dicotylédones dans la voie de la floraison (Hayama et *al.,* 2003) suggérant que la voie de la floraison est relativement conservée au sein des espèces. Des gènes identifiés chez Arabidopsis sont ainsi impliqués dans la variation du temps de floraison chez d'autres espèces (PhyB chez le sorgho, Childs et *al.,* 1997 ; Hd1 chez le riz, Yano et *al.,* 2000). La multitude des gènes connus chez Arabidopsis et autres espèces (riz, maïs) peut servir à identifier des gènes orthologues potentiels contrôlant les phénotypes liés à la floraison. L'exploitation de la synthénie pourrait ultérieurement renforcer cette approche. Des relations synthéniques entre les régions chromosomiques du mil et celles du riz ont été établies à partir d'une cartographie basée sur RFLPs (Gale et *al.,* 2006). Pour le maïs, il a été montré une colocalisation entre certains gènes de la floraison chez Arabidopsis et des QTLs liés à la floraison (Chardon et *al.,* 2004), ce qui suggère que ces gènes seraient de bons candidats pour ce caractère. Certains QTLs liés à la floraison ont déjà été détectés chez le mil dans les groupes de liaison 2, 4, 5, 6 (2 QTLs) et 7 (Poncet 1998; Yadav et *al.,* 2002, 2003). Cependant le nombre d'études de type QTL reste encore faible chez le mil et la localisation des régions impliquées peu précise.

L'amidon et les protéines. Les gènes de la voie de synthèse de l'amidon sont relativement bien connu au moins chez le mais et le riz (James et *al.,* 2003). Le laboratoire a en effet préalablement réalisé une étude pilote sur le gène Waxy (Lauret, 2006). Les résultats de cette étude sur le gène Waxy (voie de synthèse de l'amylose) et le gène Opaque2 (protéines) ont permis de détecter un signal de sélection à ces gènes (Lauret, 2006). Dans les deux cas, la diversité observée semble valider une sélection stabilisante pour des formes différentes du gène. Des travaux récents (Jideani, 2005) ont montré l'existence d'un phénotype Waxy chez le mil (variation de la teneur en amylose). Jusqu'ici, il n'a pas été montré d'association entre le gène et le phénotype Waxy chez le mil, contrairement au riz chez lequel il a été montré une relation entre le polymorphisme au gène Waxy et la variation du rapport amylose/amylopectine.

Tableau 1 : Variables phénotypiques notées sur les lignées de mil étudiées au cours des essais 2005, 2006a et 2006b.

	Variable	Code	Type	Répétitions
1	Diamètre du rachis	DRA	Continue	2005, 2006a, 2006 b
2	Diamètre de la tige principale à la récolte	DTP	Continue	2005, 2006a, 2006 b
3	Début d'épiaison	EPI	Continue	2005, 2006a, 2006 b
4	Début floraison femelle	FLO	Continue	2005, 2006a, 2006 b
5	Hauteur totale de la plante à la récolte	HPR	Continue	2005, 2006a, 2006 b
6	Largeur de la chandelle	LAC	Continue	2005, 2006a, 2006 b
7	Longueur de la chandelle	LOC	Continue	2005, 2006a, 2006 b
8	Nombre de talles aériennes à maturité	TAA	Continue	2005, 2006a, 2006 b
9	Nombre de talles au début de l'épiaison	TAE	Continue	2005, 2006a, 2006 b
10	Nombre de talles productives à maturité	TAP	Continue	2005, 2006a, 2006 b
11	Nombre d'inter-nœud	NIN	Continue	2006a, 2006b

MATERIEL ET METHODES

1. Matériel végétal

Le matériel végétal initial provient des collections de mil (*Pennisetum glaucum* L.) qui ont été fourni par J. Chantereau (CIRAD, Montpellier), T. Hash (ICRISAT à Hyderabad), A. Sarr et T. Robert (Université de Paris IV). Ces accessions sont originaires de différentes zones d'Afrique et d'Inde. Ce matériel a été reproduit à Niamey par autofécondation en 2004, 2005 et 2006. L'année 2004 a permis de multiplier les lignées et de réaliser une première étude génétique. L'essai expérimental comprenait une seule date de semis en 2005 et 2 dates semis en 2006.

2. Caractérisation phénotypique

Une caractérisation phénotypique a été effectuée au cours des 3 essais en champ : l'essai de 2005, puis 2 essais à des dates de semis différentes en 2006 (2006a, 2006b). Au total, 10 variables sont concernées en 2005, et 11 variables pour les 2 essais 2006. La liste des caractères notés est donnée *tableau 1*. Les données pluriannuelles correspondant à ces notations étaient disponibles en début de stage. Nous avons effectué au cours du stage une analyse statistique de la distribution de ces variables. Les corrélations entre variables sont étudiées par ACP (SPAD, v6). Nous avons également complété ces données par des analyses de qualité. La teneur en protéines dans les graines a été analysée pour les 2 essais de 2006, par colorimétrie à l'auto-analyseur (laboratoire d'analyses, ICRISAT Niamey). Le taux d'amylose a été dosé par méthode spectrométrique (laboratoire CRSBAN, Université de Ouagadougou).

3. Analyse génétique

Chaque individu a été génotypé à l'aide de marqueurs microsatellites (Mariac et *al.* 2006).Vingt-sept loci SSRs préalablement définis (Mariac et *al.,* 2006) ont ainsi été caractérisés sur le génome pour chacune des accessions. Ces données étaient disponibles lors de mon arrivée en stage. Le génotypage microsatellite a été répliqué lors des essais 2004, 2005 et lors des 2 essais 2006. Les résultats ont été confrontés entre eux afin de définir pour chaque lignée un *génotype consensus*. La démarche permet de corriger des erreurs éventuelles survenues au cours des manipulations (PCR, gel, lecture). Pour une lignée donnée, si un des

génotypes est différent des 3 autres, il est exclu car attribuable à une erreur de saisie, d'étiquetage, ou autre erreur technique éventuelle. Si pour une lignée, le génotype multilocus est globalement identique sauf variation à un seul loci, le génotype le plus fréquent à ce locus est conservé. La variation est attribuée à de la diversité résiduelle ségrégant dans la lignée. Sur l'ensemble des accessions, on ne retient pour la constitution de notre jeu de données que des lignées fixées à 82 % au minimum (soit, sur les 27 loci SSRs génotypés, un nombre de loci homozygotes supérieur à 22).

4. Inférence de la structure génétique des populations par une approche bayésienne

L'analyse de la structure des individus consiste dans le présent cas à déterminer les groupes génétiques différents (populations) observés dans notre échantillon de lignées. Nous avons procédé à l'analyse par la méthode STRUCTURE en choisissant les modèles de mélange entre populations (*admixture*) et la corrélation entre fréquences alléliques (Falush et *al.,* 2003). Les runs de STRUCTURE sont réalisés avec 10^6 itérations (runs) précédés d'une période de burning de 30.000 itérations (Pritchard et *al.,* 2000a). Dix simulations indépendantes sont effectuées pour chaque valeur de K. L'intervalle des valeurs testées de K varie de 1 à 10.

INSTRUCT a été paramétré sur la méthode inférant à la fois le mélange de populations et le taux individuel d'autofécondation. Un million (10^6) d'itérations sont effectuées, dont 500.000 burns. Deux chaînes MCMC (*Markov Chain Monte Carlo*) sont réalisées par simulation. On effectue 10 simulations pour chaque valeur de K. Une valeur de K variant de 1 à 8 est testée.

5. Choix de K et comparaison des méthodes

Le but de ces méthodes reste la détermination du nombre de populations présentes dans l'échantillon (la valeur de K) et l'assignation des individus à ces populations. Le problème de l'inférence de K semble encore une étape quelque peu délicate. La méthode initiale de STRUCTURE proposait un critère *ad hoc* basé sur la distribution du log likelihood L(K). Le log likelihood le plus élevé (ou celui à partir duquel s'observe un plateau) était supposé correspondre au K optimal (Pritchard et *al.,* 2000a). Cette méthode a montré des limites et nous appliquerons un nouveau critère de choix *ad hoc* qui semble plus efficace. Celui-ci repose sur la variation de second ordre du log likelihood (Evanno et *al.,* 2005). La méthode INSTRUCT a elle introduit un nouveau critère de choix du K optimal, le DIC (*Deviance*

Information Criterion). Ce critère permet de choisir le K associé au modèle le plus concordant aux données (Gao et *al.* 2007).

L'assignation des individus aux différents groupes par STRUCTURE et par INSTRUCT respectivement a été comparée. Le résultat d'une simulation est la détermination pour chaque individu des pourcentages q de génome provenant d'un ou plusieurs des K groupes. Un pourcentage d'attribution à un groupe donné pour un individu est nommé l'*ancestry* et noté q. L'ensemble de ces valeurs définit une matrice Q. Pour comparer les résultats de STRUCTURE et INSTRUCT nous allons comparer ces matrices.

Comparaison graphique. Pour chaque valeur de K on retient pour STRUCTURE la matrice Q correspondant à la simulation affichant le plus grand log likelihood parmi les 10 simulations. Pour INSTRUCT, la matrice retenue pour chaque valeur de K est celle ayant le plus faible déviance (DIC) sur les 10 répétitions. Les matrices Q issues des 2 méthodes sont analysées et comparées par une méthode exploratoire graphique.

Chacune des matrices est triée sous Excel, de façon à classer dans un même cluster tous les individus ayant la même population d'origine principale. On considère comme population principale pour chaque individu celle pour laquelle q est maximal. Puis on agence dans chaque cluster les individus en fonction de q décroissant. On représente graphiquement la répartition des individus dans les clusters, en plaçant tous les individus dans un histogramme où les ordonnées représentant les valeurs de q. Les proportions q liées à chaque population d'origine sont marquées d'une couleur spécifique. On visualise graphiquement l'organisation des différents clusters inférés pour chaque K donné. Une comparaison visuelle est ensuite faite entre le résultat des deux méthodes. Nous avons aussi effectué de façon supplémentaire une analyse en composantes principales (ACP) sur les matrices d'*ancestry* (Q) sorties de STRUCTURE et INSTRUCT à K = 7. Cette ACP (SPAD, v6) étudie la corrélation deux à deux des matrices STRUCTURE/INSTRUCT relatives à chacun des 7 clusters.

Comparaison statistique. Nous avons calculé deux mesures de similarité. Pour la première, chaque individu est attribué à une population principale. Celle-ci a été définie comme celle dont est issue la plus grande proportion du génome de cet individu (la valeur de q maximum pour l'individu). Les populations principales obtenues avec INSTRUCT et STRUCTURE sont ensuite couplées deux à deux. Ces couples de populations présentent généralement les mêmes individus (les mêmes lignées). Pour chaque couple de populations principales, on détermine le nombre Nc d'individus communs entre INSTRUCT et STRUCTURE. Le

nombre d'individus communs pour l'ensemble des couples i de populations principales couplées est :

$$\sum_{i=1}^{K} Nc_i \ .$$

La similarité SN est simplement calculée en divisant cette somme par le nombre total d'individu N. Ce coefficient de similarité est ainsi borné entre 0 et 1. Il est donné par :

$$SN = \frac{\sum_{i=1}^{K} Nc_i}{N} .$$

La deuxième mesure de similarité repose sur la distance euclidienne entre les valeurs de q affectées à chaque individu par les 2 méthodes. On note q_{ik} la proportion de génome de l'individu i dans la population k, inférée sous STRUCTURE et q'_{ik} la valeur correspondante inférée sous INSTRUCT. La distance euclidienne entre les quantités q affectées à l'individu i dans chaque population par les 2 méthodes respectives est:

$$di = \sqrt{\sum_{k=1}^{K} (q_{ik} - q'_{ik}2)^2}$$

La similarité à l'échelle individuelle est ensuite calculée selon la formule :

$$Sdi = 1 - \frac{di}{\sqrt{K}} .$$

En effet, le nombre total de populations étant K, cette valeur di doit être divisée par la racine carrée de K pour être normalisée entre 0 et 1. La similarité globale SD est ensuite calculée comme la moyenne des valeurs individuelles Sdi.

6. Analyse du pouvoir de la méthode par simulation

Nous avons voulu évaluer la capacité de la méthode à détecter un effet sur une donnée phénotypique en faisant varier deux paramètres : la fréquence de l'allèle et l'importance de l'effet de cet allèle sur le phénotype. Cette simulation repose sur les données de base de structuration et phénotype propres à notre échantillon.

6. 1. Simulation de phénotypes et de génotypes candidats

Nous simulons le polymorphisme à un locus en affectant un génotype selon la présence ou l'absence au sein de chaque lignée d'un allèle A. Les fréquences simulées de l'allèle A dans l'échantillon sont : 3,12%, 6.25%, 12,5%, 25% et 50%. Les lignées portant l'allèle sont choisies au hasard. Cent jeux de données aléatoires différents sont créés pour chaque valeur

Tableau 2 : Niveaux d'effets affectés aux polymorphismes simulés

A. Floraison

% moy.	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	15%	20%	25%	30%	35%	40%	45%	50%
FLO	0.55	1.10	1.66	2.21	2.76	3.31	3.86	4.41	4.97	5.52	8.28	11.04	13.80	16.56	19.32	22.07	24.83	27.59

B. Taux de protéines

% moy	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	20%	25%	30%	35%	40%	45%	50%
PRO	0.13	0.25	0.38	0.50	0.63	0.75	0.88	1.00	1.13	1.25	1.38	1.50	1.63	1.75	1.88	2.50	3.13	3.75	4.38	5.00	5.63	6.25

Les premières lignes expriment l'effet additif en pourcentage de la moyenne de base des caractères. La distribution de base utilisée pour simuler l'effet sur la floraison **(A)** est celle de l'essai 2005. Pour le taux de protéines **(B)**, on a utilisé comme base la distribution de l'essai 2006a. Le même effet est traduit (deuxièmes lignes des tableaux) en date de floraison (jours) et en taux de protéines.

de fréquence allélique. Chaque lignée portant l'allèle voit son phénotype augmenter d'une valeur exprimée en pourcentage de la moyenne du caractère. On suppose que les polymorphismes simulés ont un effet additif.

Le phénotype *date de floraison* (FLO) est simulé en rajoutant pour chaque lignée présentant l'allèle candidat un nombre de jours correspondant à un pourcentage donné de la date moyenne de floraison de l'échantillon. On réalise différentes distributions en faisant varier ce pourcentage dans l'intervalle allant de 0% (FLO de base de la lignée) à 50% (FLO de base de la lignée + 28 jours). La variation ajoutée (*Tableau* 2) évolue d'une distribution à la suivante au pas régulier de 5%, excepté pour certaines parties de l'intervalle que l'on affine en évoluant au pas de 1%.

La distribution du taux de protéine (PRO) est simulée de façon similaire au caractère date de floraison, en rajoutant en présence de l'allèle candidat une variation correspondant à pourcentage du taux moyen de protéines de l'échantillon. L'effet ajouté (*Tableau 2*) variera ainsi, selon les distributions, entre 0% du taux de protéine moyen (PRO de base de la lignée) et 50% de ce taux moyen (PRO de base de la lignée + 6,25).

6. 2. Test du pouvoir de détection

Les données génotypiques et phénotypiques simulées sont soumises à un test d'association par régression logistique sous TASSEL en prenant en compte la matrice de structure génétique optimale. Le modèle de régression logistique (Thornsberry et *al.,* 2001) teste deux hypothèses. Dans l'hypothèse H_0, les polymorphismes candidats sont indépendants des phénotypes. Dans l'hypothèse alternative H_1, les polymorphismes sont associés aux phénotypes. La probabilité des hypothèses est ensuite comparée sous la forme :

$$\Lambda = \frac{\Pr_1(C;T;\hat{Q})}{\Pr_0(C;\hat{Q})}$$

C est le génotype du polymorphisme candidat pour toutes les lignées, T est la valeur du phénotype, $\hat{Q}$ est l'estimation bayésienne de la structure génétique. Les deux probabilités sont estimées en utilisant le régression logistique du modèle SAS, dans lequel la variable réponse est la présence ou l'absence du polymorphisme candidat, T et Q étant considérés comme des variables indépendantes (Thornsberry et *al.,* 2001).

Le pouvoir de détection (estimation de l'efficacité du test) est évalué par le pourcentage de tests indiquant une valeur de probabilité critique inférieure à $\alpha = 0.05$.

7. Analyse de la distribution de la probabilité critique avec un jeu de marqueurs neutres

La prise en compte de la structuration est faite par l'utilisation des données de génotypage microsatellites. Ces marqueurs sont supposés non liés aux caractères d'intérêt. La prise en compte de la structure par la méthodologie précédente peut ne pas être totalement parfaite. Dans le cas d'une prise en compte parfaite de la structure et avec des allèles microsatellites non liés aux caractères étudiés, nous attendons environ 5% des allèles présentant une significativité d'association au seuil 5%. Cependant si la prise en compte de la structure est imparfaite ce seuil peut être plus élevé. Il est intéressant d'évaluer le seuil expérimental approprié qui délimite 5% des observations dans la distribution des *p-values* issues d'un jeu de marqueurs neutres. Cette *p-value* ainsi obtenue (que nous nommerons seuil expérimental) peut être alors interprétée comme un seuil de significativité corrigé.

Pour cela, un jeu de données génotypique est réalisé en partant du génotypage microsatellite effectué sur les 90 lignées. Pour cela, on *haploïdise* sous TASSEL le génotype diploïde aux 27 loci microsatellites. Une des deux séries de données résultant de cette étape étant choisie, on en extrait sous TASSEL les allèles ayant une fréquence minimale de 2,5 % dans l'échantillon. Procédant allèle par allèle, on recode le génotype des lignées selon la présence ou l'absence de l'allèle. Ensuite nous avons considéré les matrices de structures (K=1 à K=7) inférées respectivement sous STRUCTURE et INSTRUCT. Cette analyse permet de comprendre comment la prise en compte de la structure fait évoluer le seuil.

L'association entre chacun des caractères et toute la série des allèles microsatellites est analysée sous TASSEL (régression logistique). Les tests d'association sont effectués d'abord en l'absence d'une matrice de structure génétique, puis en prenant en compte respectivement les différentes matrices de structure (Q) pour chaque valeur de K. Nous avons considéré pour cette analyse les caractères : date de floraison (FLO), nombre d'inter-nœuds (NIN), hauteur de la plante à la récolte (HPR), taux de protéines (PRO), et deux variables composites provenant de l'ACP (CP1 et CP2). Ces associations sont effectuées et analysées essai par essai (jusqu'à 3 essais selon la disponibilité des données). La distribution expérimentale pour chaque caractère des *p-value*s est établie en classant par ordre croissant les valeurs de *p* obtenus pour les 119 allèles microsatellites. On détermine le seuil expérimental, que l'on définit comme la valeur de *p* en dessous de laquelle se placent 5% des *p-value*s de la distribution.

8. Application du test d'association aux données expérimentales

Nous avons soumis différentes séquences de polymorphismes candidats à la régression logistique sous TASSEL pour la recherche d'association avec les caractères d'intérêt. Plusieurs gènes candidats ont été identifiés et séquencés par le laboratoire d'accueil avant mon arrivée en stage. Des marqueurs SNPs présents sur le fragments 1 (27 sites) du gène Opaque2 et le fragment 2 (22 sites) du même gène ont été testés vis-à-vis de la teneur en protéines (Lauret, 2006). Enfin, une deuxième série concerne des candidats à la date de floraison, à savoir des marqueurs SNPs appartenant aux séquences des gènes candidats HD3a, PhyA, PhyB, PhyC, Gigantea, Floricaula (C. Mariac and Y. Vigouroux, données non publiées)**.** Pour tous les tests d'association, un contrôle de l'effet de la structure a été associé en intégrant au jeu de données la matrice de structure correspondant au K optimal détecté.

La probabilité critique *(p)* associé aux couples site polymorphe/caractère est déterminée par le logiciel TASSEL sur la base de 100 permutations. Le seuil de décision permettant de valider la significativité selon le caractère et l'essai de chaque probabilité critique est fixé sur la base de notre correction empirique (cf. *Matériel et Méthodes*, paragraphe *7*).

Tableau 3 : Description sommaire de variables morphologiques et phénologiques chez le mil (P. *glaucum*)

Caractère	Nombre d'obs.	Unité	Moyenne	Ecart-type	Médiane	Minimum	Maximum
Essai 2005							
DRA	444.00	mm	0.44	0.15	0.40	0.10	1.00
DTP	450.00	cm	1.07	0.29	1.10	0.40	2.00
EPI	446.00	jours	51.53	9.28	50.00	32.00	94.00
FLO	453.00	jours	**54.79**	**9.10**	**53.00**	**35.00**	**87.00**
HPR	450.00	cm	83.00	32.62	76.75	26.00	197.00
LAC	450.00	cm	2.19	0.61	2.20	0.40	4.20
LOC	450.00	cm	26.32	14.22	21.50	6.50	93.00
TAA	449.00	-	1.95	3.26	0.00	0.00	21.00
TAE	456.00	-	8.75	4.06	8.00	2.00	28.00
TAP	449.00	-	3.47	2.71	3.00	0.00	21.00
Essai 2006a							
DRA	534.00	mm	0.46	0.15	0.40	0.20	1.00
DTP	534.00	cm	1.08	0.28	1.00	0.50	2.50
EPI	540.00	jours	58.41	8.15	57.00	46.00	90.00
FLO	549.00	jours	**61.18**	**8.28**	**60.00**	**46.00**	**92.00**
HPR	534.00	cm	87.30	33.08	81.75	27.40	200.50
LAC	534.00	cm	2.18	0.40	2.10	1.10	3.80
LOC	534.00	cm	24.72	11.12	21.50	6.00	87.50
TAA	534.00	-	1.88	5.15	0.00	0.00	73.00
TAE	550.00	-	8.09	3.89	7.00	1.00	26.00
TAP	534.00	-	5.06	5.83	4.00	1.00	82.00
NIN	533.00	-	7.66	1.56	8.00	4.00	13.00
Essai 2006b							
DRA	541.00	mm	0.47	0.14	0.40	0.20	0.90
DTP	548.00	cm	0.98	0.28	0.90	0.30	2.00
EPI	550.00	jours	56.75	8.83	56.00	40.00	96.00
FLO	560.00	jours	**59.28**	**9.00**	**58.00**	**42.00**	**98.00**
HPR	548.00	cm	81.23	26.52	76.40	32.50	166.00
LAC	547.00	cm	2.20	0.95	2.10	0.70	19.00
LOC	548.00	cm	25.12	11.81	21.40	2.00	90.20
TAA	548.00	-	0.30	1.05	0.00	0.00	9.00
TAE	560.00	-	8.89	4.42	8.00	1.00	28.00
TAP	548.00	-	4.27	2.86	4.00	1.00	20.00
NIN	547.00	-	6.23	1.43	6.00	3.00	12.00

La caractérisation phénotypique a été effectuée sur la base d'un échantillon de 90 lignées de mil, notées en plein champ à Sadoré (40 km de Niamey). Les statistiques sont présentées séparément pour les 3 répétitions : 2005 (85 lignées), 2006a (84 lignées) et 2006b (81 lignées). En 2005, la valeur des variables pour chaque lignée correspond à la moyenne pour 6 individus de la lignée, tandis que 7 individus par lignée ont été notés en 2006.

Variables	Valeur Test	Probabilité Critique (Pc)
NIN	6.75	$1.5\ 10^{-11}$*
EPI	5.71	$5.6\ 10^{-09}$*
FLO	**5.32**	**$5.3\ 10^{-08}$***
TAA	3.85	$5.9\ 10^{-05}$*
TAP	2.80	$2.6\ 10^{-03}$*
DTP	1.94	0.026*
TAE	0.56	0.29
DRA	0.48	0.32
HPR	0.29	0.39
LAC	-0.63	0.74
LOC	-1.81	0.96

Tableau 4 : Effet environnemental sur les variables phénotypiques du mil (P. *glaucum*).

* Test significatif au seuil $\alpha = 5\%$. Les valeurs tests calculées par SPAD (v6) correspondent à un nombre d'écart-types de la distribution normale. On considère comme significatif un test lorsque la valeur test qui lui est lié est supérieure en valeur absolue à ~1,96. La décision sur la base des valeurs-tests confirme ici le modèle de décision classique ($Pc < \alpha$). L'effet environnement (essai) a été analysé sur les données concernant un sous-échantillon de 78 lignées communes aux essais 2005, 2006a et 2006b (ANOVA). La moitié des caractères étudiés ne montrent une sensibilité significative à l'effet de l'environnement, contrairement à la floraison par exemple, qui varie significativement d'un essai à l'autre.

RESULTATS

1. Variabilité phénotypique de la date de floraison, du taux de protéines et de quelques caractères morphologiques au sein du panel de 90 lignées de mil

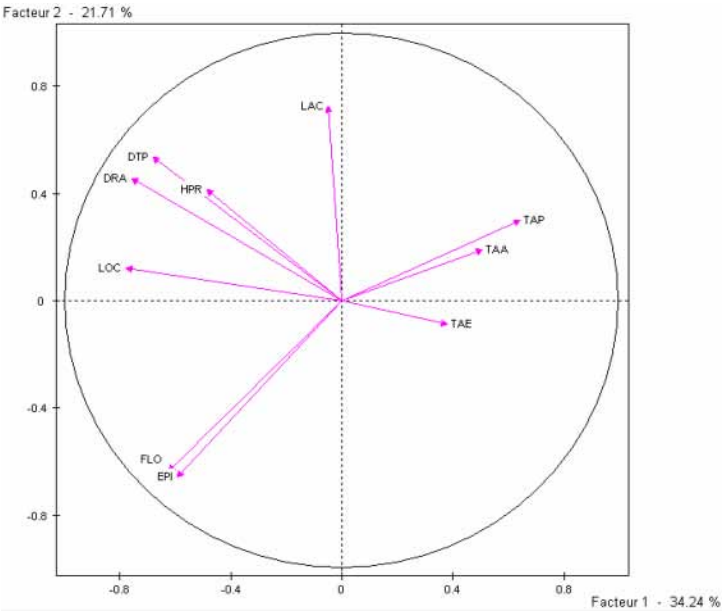
Les variables phénotypiques ont été notées sur un dispositif pluriannuel. . Nous avons analysé statistiquement la distribution de ces variables pour chacun des essais indépendamment. Le *tableau 3* présente les statistiques sommaires correspondantes. Nous allons nous concentrer dans ce rapport sur la variable date de floraison. Ce caractère varie suivant les lignées et les essais entre 35 jours et 98 jours. La médiane est de l'ordre de 58 jours.

Sur la base des 3 essais nous avons voulu évaluer l'effet environnement. Cet effet environnement est significatif pour 6 parmi les 11 variables morphologiques notées, notamment sur la date de floraison (*Tableau 4*). La distribution des dates de floraison varie ainsi suivant l'essai (*Figure 2*). Toutefois, la corrélation entre les dates de floraison des lignées entre essais demeure forte et très significative pour les 3 essais (*Tableau 5*). Cette corrélation traduit partiellement l'héritabilité de ce caractère.

Les différentes variables étudiées ne sont pas indépendantes. Il est donc intéressant de créer des variables composites qui résument les morphologies observées. Pour cela, nous avons réalisé une analyse en composante principale (ACP). Le premier axe de l'ACP se révèle particulièrement structurant (inertie : 34,2%). Plusieurs variables lui sont significativement corrélées (DRA, DTP, EPI, LOC, FLO, TAA, TAP). La composante principale (CP1) liée à cet axe sera donc une variable composite résumant l'essentiel de la variation de ces caractères. La deuxième composante explique 21.7% de la variation. A eux deux ces deux axes résument 55.9% de la variation. En outre, la corrélation établie par ACP entre les variables reste du même ordre quelque soit l'année.

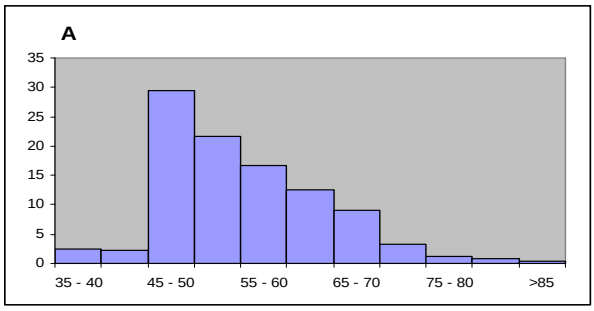
Les analyses du taux d'amylose, lancées au cours du stage, ne sont pas encore à point. Notre analyse sur la qualité s'est donc finalement limitée à la voie des protéines. Les analyses de quantité de protéines ont été réalisées sur les deux essais 2006a et 2006b. La valeur moyenne des lignées est de 12,46% pour le premier essai (2006a) et 13,77 pour le second (2006b). Les teneurs en protéines observées varient entre 8,75% et 17,78% de. La corrélation entre les deux essais (R^2=0,4) est significative.

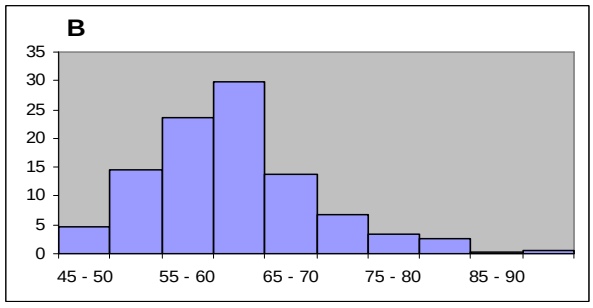
Figure 1 : Corrélation entre variables phénotypiques chez le mil (*P. glaucum*).

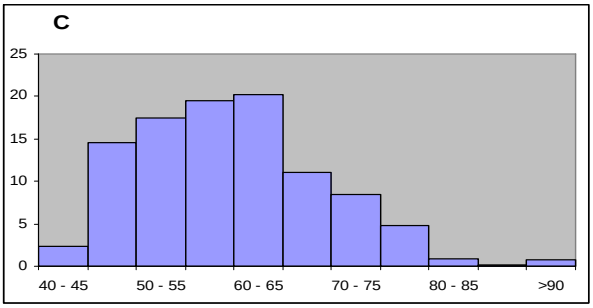


La variable FLO est positivement corrélée à EPI, LOC notamment. La plupart des variables sont bien liée à l'axe 1, qui affiche une inertie assez élevée de 34,24%. Le plan factoriel-ci est celui de l'essai 2005. Les mêmes tendances sont observées en 2006 (essais 2006a et 2006b).

Figure 2 : Distribution de la date de floraison de lignées de mil (P. *glaucum*) dans un dispositif d'essai pluriannuel.







La date de floraison, calculée en nombre de jours à partir de la date de semis, a été caractérisée sur un panel de plus de 80 lignées évaluées au champ à Sadoré (Niger) au cours d'un essai réalisé en 2005 **(A)**, puis au cours de 2 essais réalisés en 2006 **(B & C)**. L'ordonnée représente la fréquence (en %) des classes correspondantes.

Tableau 5 : Corrélation entre les dates de floraison des lignées.

Essais comparés	R	R^2	Probabilité critique (Pc)
2005-2006a	0.717	0.51	$7.4\ 10^{-14}$
2005-2006b	0.758	0.57	$4.1\ 10^{-16}$
2006a-2006b	0.711	0.51	$3.1\ 10^{-13}$

La liaison entre les dates de floraison observées d'un essai à l'autre a été analysée sur la base d'un sous échantillon de 78 lignées communes aux 3 essais. Les corrélations observées sont toutes significatives ($Pc < 0.001$) et traduisent une corrélation forte ($R > 0.7$).

2. Analyse comparée de la structure génétique

2.1. Détection du K optimal

A partir des simulations effectuées sous STRUCTURE, nous avons représenté (*Figure 3)* le log likelihood L(K), la probabilité postérieure des données (Pritchard et *al.,* 2000a). On n'observe pas de valeur modale, ni de plateau sur la courbe de L(K). Tout de même, on mentionne une augmentation de la variance de L(K) entre les simulations après la valeur K = 2 (*Figure 3*). Ces éléments ne sont pas suffisamment concluants quant à la détection du K optimal. Nous avons aussi déterminé la variation de second ordre de ce log likelihood (*Figure 3*) qui est la fonction ΔK (Evanno et *al.,* 2005). Cette fonction de variation de second ordre du log likelihood (ΔK) montre une valeur modale claire, à K = 6. Le nombre de clusters optimal détecté par STRUCTURE dans notre panel serait donc de 6, conformément à Evanno et *al.* (2005).

Sous INSTRUCT, le choix de K a été établi sur la base du critère de déviance DIC (*Deviance Information Criterion*) implémenté dans le logiciel. Les 10 répétitions prévues pour l'intervalle de K allant de 2 à 8 sous INSTRUCT n'ont pas pus être complétées[1]. Ainsi, nous ne rapportons ici que les résultats d'une simulation accomplie avec succès. Ce résultat révèle un nombre optimal de 7 clusters, le DIC le plus faible étant associé à cette valeur de K. Des runs supplémentaires ont été effectués pour un intervalle de K ciblé et plus restreint (K = 6 à 7). Le K optimal, pour ces simulations, apparaissait tantôt à K = 6, tantôt à K = 7.

2.2. Evolution de la structure génétique selon K et similarité des résultats STRUCTURE et INSTRUCT

Nous avons représenté graphiquement l'*ancestry* q (fraction de génome issue de chaque population d'origine) en fonction des lignées, à partir des matrices Q de STRUCTURE et INSTRUCT respectivement, pour K variant dans un intervalle de 1 à 7 (*Figure 4*). Chaque individu est représenté par une barre verticale et les différentes valeurs d'*ancestry* dans les K

[1] La plateforme en ligne utilisée pour effectuer ces runs interrompait les simulations en chemin, sans renvoyer explicitement la cause de l'erreur. Plusieurs tentatives ont été vouées à l'échec, alors que la durée nécessaire pour un seul run était de plusieurs jours, avec une limitation du nombre de tâches par utilisateur. Par ailleurs, la version téléchargeable du logiciel que nous avons installé sous LINUX sollicite de la mémoire importante (2.0242E+09 bytes) et plus de 2 semaines par run sur le serveur à notre possession (PC ordinaire). Nous avons transmis les fichiers correspondants aux erreurs au développeur du logiciel, afin d'en vérifier la cause. A terme, cette information permettrait certainement de compléter au besoin les répétitions pour confirmer (ou non) la stabilité de la détection.

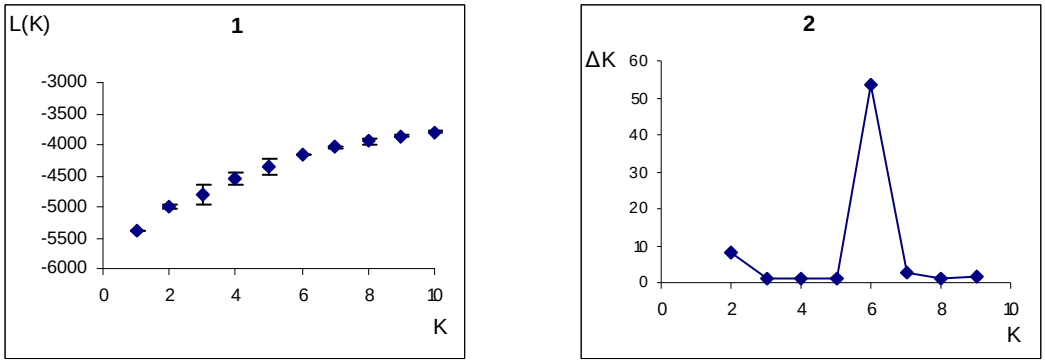


Figure 3 : Détection du nombre de clusters dans un panel de lignées analysées par STRUCTURE. L'échantillon comporte 90 lignées de mil (*P. glaucum* L.) génotypées à 27 loci microsatellites et analysées sous STRUCTURE (Falush et *al.,* 2003). (1) Les points représentent la moyenne L(K) du log likelihood sur 10 simulations et la barre d'erreur correspond à l'écart type entre les simulations (2) Variation de second ordre du log likelihood ΔK calculé selon la formule d'Evanno et *al.* (2005). Une augmentation de la variance de L(K) est observée à partir de K=2, mais cet élément n'est pas suffisamment concluant. Par contre, ΔK permet de détecter sans ambiguïté une structure optimale à K = 6, correspondant à la valeur modale de la fonction.

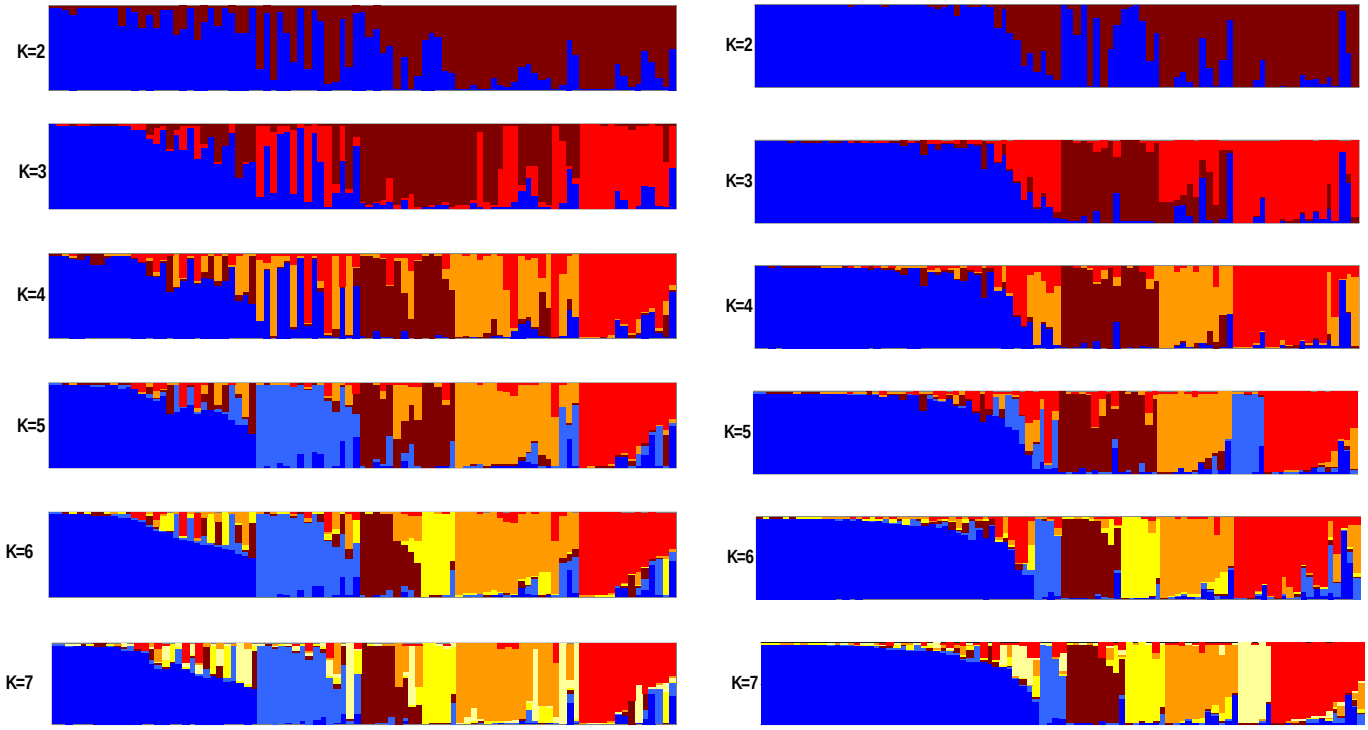


Figure 4 : Inférence de la structure des populations par une approche bayésienne. Pour chaque nombre de cluster donné (K), la structure présenté a été inférée sous STRUCTURE (Pritchard et *al.*, 2000a; Falush et *al.*, 2003), puis sous INSTRUCT (Gao et *al.*, 2007) indépendamment. En plus du flux de gènes ancestraux pris en compte par STRUCTURE (*admixture model*), INSTRUCT intègre dans son modèle le taux d'autofécondation (*selfing rate for individuals*). Les individus en abscisse représentent 91 lignées de mil génotypées à 27 loci microsatellites; l'axe de ordonnées représente la proportion q du génome de chaque individu provenant d'une population d'origine donnée. Chaque population ou sous-population est caractérisée par une couleur distincte. En partant de K=2 (a) à K=7, on observe la subdivision progressive des groupes en sous-populations structurées. On compare ici la structure globale des clusters, l'agencement des individus en ordonnée étant légèrement différent pour les 2 représentations.

groupes sont représentées par différentes couleurs. Les groupes similaires entre INSTRUCT et STRUCTURE sont représentés avec la même couleur. Le passage d'une structure comportant K populations à une structure comprenant K+1 populations révèle à chaque fois la scission de l'une des K populations en deux sous-populations génétiquement différenciées.
Nous avons ensuite observé le positionnement individuel des lignées dans les clusters. En permutant les agencements horizontaux au sein des clusters et en gardant les valeurs de q pour chaque individu, les représentations obtenues ne reflètent plus les limites entre toutes les populations (*Figure 6, zone fléchées*). Ce réarrangement observé dans quelques zones stipule que certains individus sont assignés à des populations différentes par les deux méthodes. Toutefois, ce flottement entre les populations selon la méthode d'inférence est limité, car le coefficient de similarité basée sur la population principale des lignées (SN) est égal à 0,83 en moyenne (*Tableau 6*). Le coefficient de similarité basée sur la distance euclidienne (SD) montre une moyenne de 0,74 (*Tableau 6*). Ce coefficient étant également borné entre 0 et 1, la présente valeur reflète un niveau de similarité assez élevé entre les matrices Q d'INSTRUCT et de STRUCTURE. Il existe certes de valeur de K pour lesquelles la similarité entre les variables Q d'INSTRUCT et STRUCTURE est faible (à K = 5, SD = 0,4), ce qui a contribué à hausser l'écart-type du coefficient de similarité. Mais dans l'ensemble les résultats des deux méthodes sont assez similaires.

3. Association phénotype/marqueurs neutres dans des populations structurées

L'analyse d'association doit se prémunir contre les corrélations erronées. La prise en compte de la structure permet-elle de faire baisser la fréquence de faux positifs?

3. 1. Effet de la structure génétique sur la significativité des tests d'association

On a évalué dans notre dispositif le pourcentage de *p-values* significatives (PPS) obtenu par hasard, en testant la significativité de la régression logistique entre les traits et des allèles microsatellites neutres (*Figure 7*). La quasi-totalité des histogrammes de résultats (21 sur 22) montrent un PPS nettement plus élevé pour la modalité K = 1 qui ne prend pas en compte la structure. Cette observation graphique est confirmée par les tests statistiques formels. Nous montrons que la structure a un effet très significatif ($p < 0.001$) sur le taux de tests significatifs (PPS) qui dépend aussi, certes, du caractère étudié et de l'effet environnemental (*Tableau 7*). L'intervalle de confiance de cette moyenne (MINITAB, *v14*) ne se recoupe pas avec celui des moyennes de PPS prenant en compte la structure (*Figure 8*).

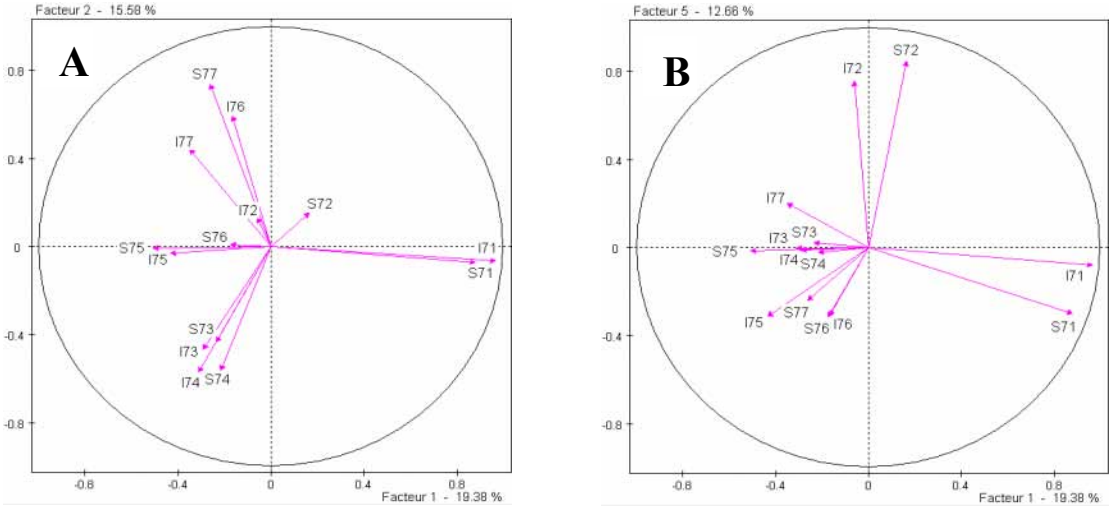
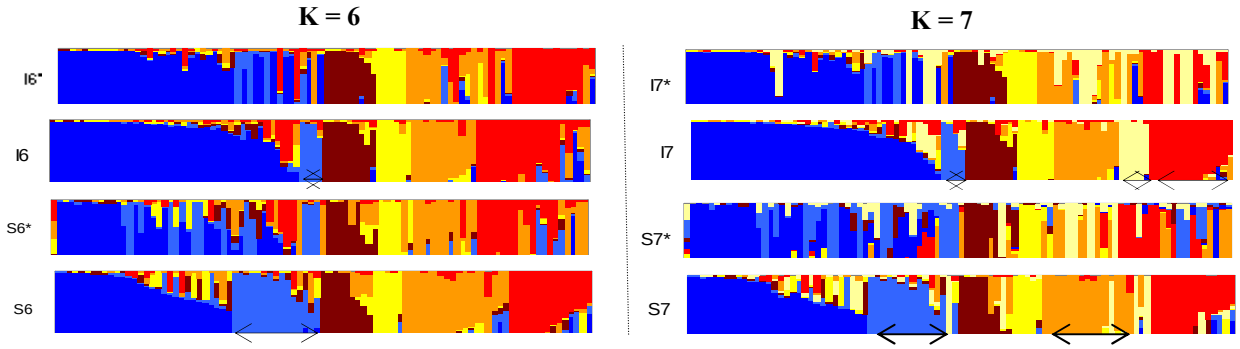


Figure 5 : Corrélation (ACP) entre les matrices de structure inférées par STRUCTURE et INSTRUCT. La structures génétique pour le panel de 90 lignées de mil (*P. glaucum* L.) génotypées à 27 loci SSRs a été inférée avec STRUCTURE (Pritchard et al., 2000a ; Falush et *al.*, 2003) et INSTRUCT (Gao et *al.*, 2007). Le nombre de populations a été fixé à 7. Les vecteurs représentent la distribution de la variable Q_k, qui définit la fraction de génome de chaque individu dans la kième population inférée par STRUCTURE **(S7k)** ou par INSTRUCT **(I7K)**. Les 2 premiers axes factoriels **(A)** de l'analyse en composantes principales (ACP) expliquent un niveau élevé de la variabilité (pourcentage d'inertie totale : 34,96%). Ils montrent donc une bonne corrélation à la plupart des vecteurs. Notre deuxième plan factoriel **(B)** se prête mieux à la représentation de certaines variables qui étaient mal représentées sur le premier plan (exemple : S72, I72). La projection des variables Q sur ces deux plans respectifs révèle, pour chaque valeur de K, une forte corrélation entre les couples STRUCTURE /INSTRUCT de vecteurs associés. Les deux méthodes conduisent donc à des résultats (Q_k) très corrélés quelque soit par ailleurs le niveau de structure (K) considéré.

Figure 6 : **Inférence des clusters et assignation des individus par les méthodes STRUCTURE et INSTRUCT.** La



fraction de génome dans chaque population d'origine (Q) est représentée pour un panel de 91 lignées de mil génotypées à 27 loci microsatellites. STRUCTURE (Pritchard et *al.*, 2000a ; Falush et *al.*, 2003) a été paramétrés avec les options *admixture model* & *correlated allel frequences*. INSTRUCT (Gao et *al.*, 2007) a été paramétré sur modèle a*dmixture and selfing rate for individuals*. L'agencement optimal des individus pour chacune des 2 méthodes permet de visualiser de façon limpide les clusters inférés par chacune de celles-ci **(S6, I6, S7, I7).** Lorsque l'on représente les valeurs de Q inférées par INSTRUCT en imposant l'agencement horizontal obtenu à partir des résultats de la méthode STRUCTURE **(I6*, I7*),** les groupes obtenus ne sont pas très homogènes. Certains groupes **(zones fléchées)** s'éclatent, et les individus correspondants se retrouvent dans les populations voisines **(I6*, I7*)**. La même chose est observée lorsque la distribution de Q inférée par STRUCTURE est représentée en réarrangeant les individus conformément à l'agencement optimal obtenu pour INSTRUCT (**S6*, S7***). Cette permutation réciproque des individus permet de voir si les mêmes individus sont systématiquement associés aux mêmes groupes par les 2 méthodes. Il vient que les populations formées apparaissent globalement similaires; toutefois, certains individus flottent entre des populations voisines, selon qu'ils sont assignés en partant de la matrice Q d'INSTRUCT ou de celle de STRUCTURE.

Tableau 6 : Coefficient de similarité entre les résultas obtenus par différentes méthodes d'estimation de la structure des populations (INSTRUCT et STRUCTURE)

K	2	3	4	5	6	7	Moyenne	Ecart-type
SD	0.97	0.98	0.98	0.38	0.81	0.84	**0.83**	**0.21**
SN	0.91	0.85	0.93	0.40	0.78	0.58	**0.74**	**0.19**

La similarité entre deux méthodes d'estimation de la structure des populations est présentée ici. Deux coefficients sont utilisés pour comparer les deux méthodes (SN et SD, voir texte pour détails). La similarité de résultat est élevée entre les 2 méthodes.

Le contrôle de la structure génétique dans notre dispositif limite la moyenne des tests significatifs (PPS) à 8,15 % (± 0,85), alors que cette moyenne atteint 17,18 % lorsque l'on néglige l'effet structure. Par ailleurs, les analyses faites en utilisant les résultats de STRUCTURE ou INSTRUCT ne différent guère.

3. 2. Amélioration du modèle de décision du test d'association par l'introduction d'une correction empirique du seuil de significativité α

L'analyse précédente permet de mettre en évidence un pourcentage d'association supérieur au 5% classique. Une seconde source d'erreur de type I dans la méthode d'association est donc l'application d'un seuil α non approprié. Pour déterminer un seuil approprié ; nous avons établi la distribution nulle empirique des probabilités critiques en utilisant le jeu de marqueurs microsatellites. Nous avons vu précédemment qu'en moyenne 8.15% des allèles microsatellites présentent une association significative au seuil 5%. Nous aimerions ici obtenir la valeur de seuil *p* corrigée pour laquelle 5% seulement des microsatellites sont significatifs. Pour cela on construit la distribution des valeurs de *p* pour les allèles microsatellites par ordre croissant (*Figure 9*). Les allèles qui ont un rang de 5% permettent de définir la valeur de *p* corrigée. A partir de chaque distribution de ce type, on a déterminé le seuil empirique, qui est la valeur de probabilité critique *p* en dessous de laquelle se situent 5% des valeurs de *p* observées (*Tableau 8*). Ce seuil est spécifique des combinaisons caractère/essai. En résumé, cette valeur correspond sur la distribution à la valeur de *p* dont la fréquence cumulée (ordonnée) est égale à 5%. Tous les seuils empiriques qui sortent de cette analyse (sauf un seul) se situent en deçà du seuil standard.
Par ailleurs, la méthode par laquelle la structure génétique a été inférée n'influe pas sur le pourcentage de tests significatifs (PPS) avec le jeu de marqueurs microsatellites (ANOVA, $p >> 0.05$). Les matrices Q d'INSTRUCT ou STRUCTURE ont donc permis le contrôle de la structure génétique avec une efficacité équivalente.

4. Etude par simulation du pouvoir de la méthode TASSEL

Les simulations effectuées montrent que le pouvoir de la méthode TASSEL (régression logistique) est dépendant de la fréquence de l'allèle candidat dans la population et de l'effet de cet allèle (*Figure 10*). Pour les deux caractères (taux de protéines, date de floraison) sur lesquels nous avons modélisé un effet additif, la méthode assure une détection certaine des

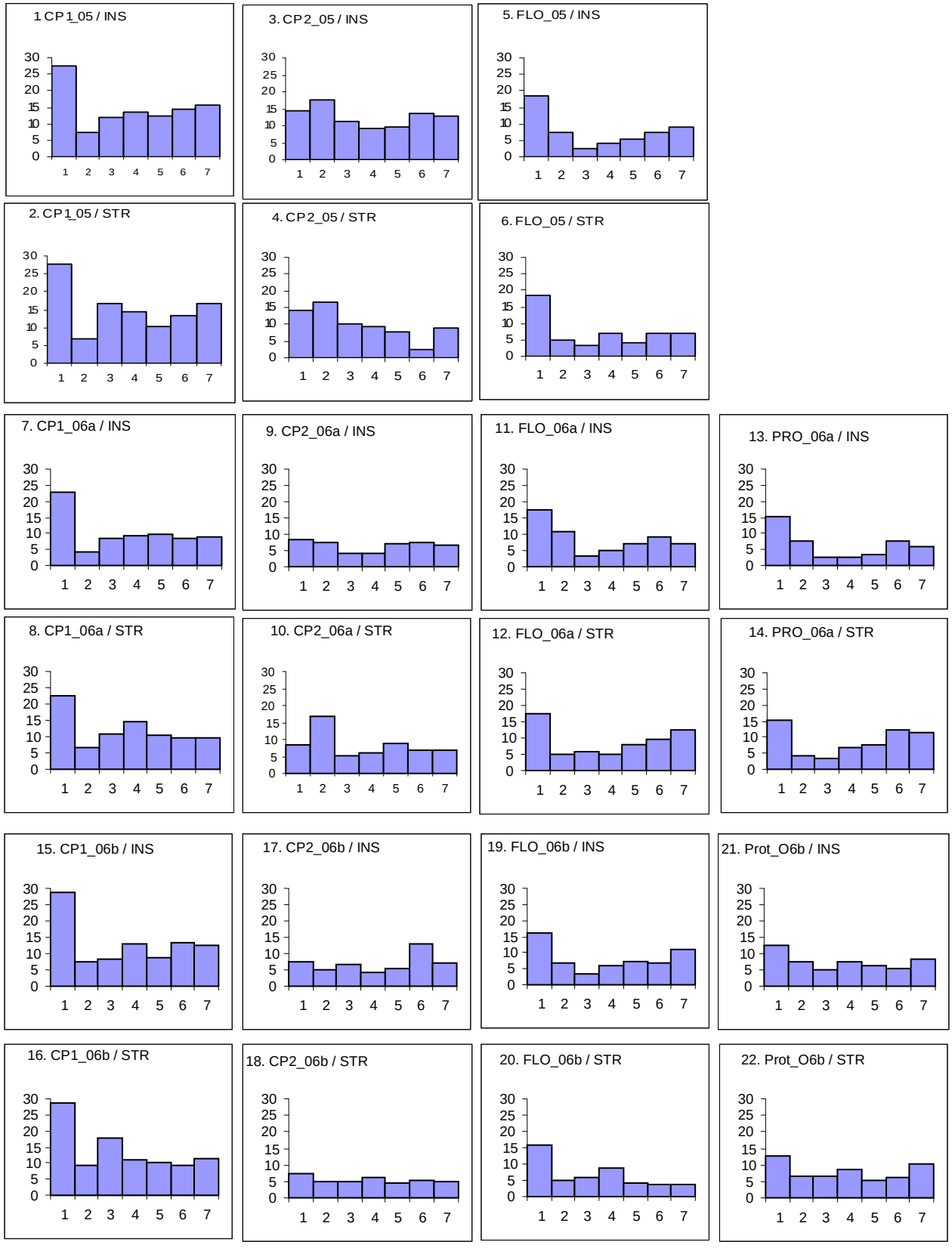


Figure 7 : Fluctuation du pourcentage de tests d'association significatifs sous l'effet de la structure génétique inférée par deux méthodes bayésiennes. L'analyse d'association (TASSEL) teste la corrélation entre des variables phénotypiques et 119 allèles correspondant à 27 loci microsatellites par hypothèse non liés aux caractères. Les traits analysés sont la date de floraison (FLO), le taux de protéines (Prot) et 2 composantes principales (CP1 & CP2). Les tests d'association sont réalisées essai par essai sur un panel de 90 lignées de mil (*P. glaucum* L.) notées au champ lors de l'essai 2005 (graphes 1 à 6), 2006a (graphes 7 à 14) et 2006b (graphes 15 à 22). Les graphes représentent le pourcentage de tests significatifs ($pc < 0.05$) lorsque l'on ignore la structure génétique (abscisse K = 1) ou lorsque l'on prend en compte la structure génétique inférée par STRUCTURE (STR) ou INSTRUCT (INS) pour des valeurs de K variant de 2 à 7. Ce taux de tests significatifs peut être interprété comme un indicateur du taux de faux positifs. Il est très élevé (jusqu'à 28%) lorsque l'on ignore l'effet structure (K=1), mais peut être considérablement limité selon la matrice de structure considérée.

associations significatives créées entre les polymorphismes et les phénotypes dès lors que la fréquence de l'allèle est assez élevée. Cela est vrai même pour des effets faibles à très faibles. Dans le modèle, un effet de 6 jours (ou plus) sur la date de floraison est détectée de façon fiable (niveau de confiance>95%) dès lors que l'allèle est présent dans l'échantillon à une fréquence allélique supérieure à 25 %. Ce niveau de fréquence allélique garantit aussi la détection certaine de tout effet sur la variation du taux de protéine, même lorsque cet effet est inférieur à 2% de taux de protéines (*Figure 10*). Les allèles rares à effet trop faible sont difficilement détectables. Des effets faibles systématiquement détectés par la méthode car étant portés par des allèles suffisamment fréquents ne sont détectés que dans très peu de cas quand ils sont affectés à des allèles rares. Aux fréquences alléliques très faibles, la perte de pouvoir de la méthode est très marquée. Toutefois, à partir d'un certain seuil d'effet sur le phénotype, la méthode est capable de mettre en évidence des associations même avec des allèles rares. Par exemple, quasiment toutes les répétitions indépendantes simulant une augmentation de 6% sur le taux de protéines conduisent à un test significatif sous TASSEL même pour des allèles candidats présents sur moins de 3,5% des individus (3 lignées sur 90).

5. Association entre polymorphismes candidats et caractères d'intérêt chez le mil

La régression logistique sous TASSEL nous a permis de tester l'association de plusieurs gènes candidats avec les caractères en contrôlant la structure génétique (Thornsberry et *al.*, 2001). Le gène Opaque2 (fragments 1 et 2) n'a pas révélé de corrélation significative avec la variation du taux de protéines dans les lignées, aux seuils empiriques définis *tableau 8*.

Les SNPs séquencés sur PhyA, HD3a, Floricaula et Gigantea affichent des valeurs de *p* très élevées (>0.05, voir annexe) ; la méthode ne détecte donc aucune corrélation entre ces sites et la variation de la date de la floraison.

Les valeurs de *p* nettement significatives sont obtenues avec le gène PhyC. Plusieurs SNPs de ce gène montrent une corrélation significative avec la date de floraison et d'autres caractères liés à la date de floraison. La première composante principale (CP1) liées aux variables morphologiques est corrélée à la quasi-totalité des sites SNPs séquencés, quelque soit l'essai considéré (2005, 2006a et 2006b). En 2005, la part de variabilité expliquée par les SNPs pour CP1 est de 12,61% en moyenne pour chaque marqueur SNP, et peut atteindre 14,34% (*Tableau 9*). Cette corrélation avec CP1 reste significative pour toutes les 3 réplications (essais 2005, 2006a et 2006b).

Pour la date de floraison, la corrélation la plus significative avec les polymorphismes de PhyC a été observée sur l'essai 2006b (*Tableau 9*). La part de variabilité expliquée varie selon les SNPs de 6,01% à 10,68%.

En dehors de ces corrélations que nous avons validées pour PhyC en tenant compte de la correction empirique du seuil, nous avons des tests pour lesquels les *p-value*s se situent entre les valeurs des seuils empiriques (inférieur à 5%) et le seuil standard (5%). C'est le cas

Tableau 7 : Effet de différents facteurs sur le taux de valeurs de probabilités critiques inférieures au seuil standard dans un test d'association avec des marqueurs choisis aléatoirement

Facteur	ddl	Sommes des carrés des écarts (type III)	Carrés moyens	F	Probabilité critique (Pc)
Caractère	3	763.1	254.4	22.7	$4.4\ 10^{-12}$
Essai	2	133.4	66.7	6.0	$3.2\ 10^{-03}$
Structure (K)	6	1619.7	269.9	24.1	$1.6\ 10^{-19}$
Méthode d'inférence de structure	1	1.7890	1.79	0.16	0.69

SPAD, v6

Des allèles microsatellites (119) supposés sans effet sur les caractères (4) ont été testés par régression logistique sous TASSEL avec des données phénotypiques de 3 essais. On a fait varier la covariable structure de la valeur K = 1 (pas de structure) à K = 7. La structure génétique a un effet significatif sur le pourcentage de probabilités critiques (Pc) significatives au seuil de 5%. Ce pourcentage dépend aussi du caractère analysé et de l'effet environnemental (essai). Par contre, les 2 méthodes utilisées pour l'inférence des matrices de structures (INSTRUCT ou STRUCTURE) conduisent à des résultats statistiquement équivalents.

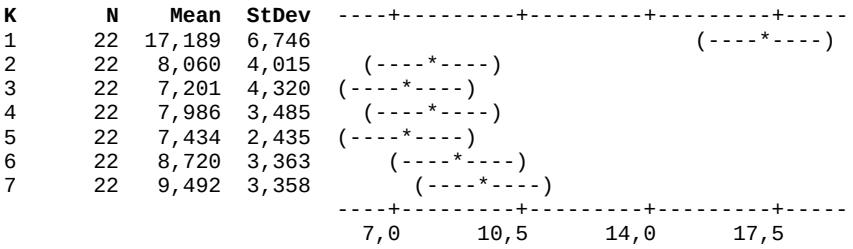


Figure 8 : Intervalle de confiance du taux de tests d'association significatifs dans un jeu de données comportant des allèles microsatellites. Les allèles microsatellites (119) ont été corrélés à 4 caractères différents dont la notation a été répliquée 3 fois indépendamment, l'hypothèse à priori étant l'absence de lien avec les allèles. On a déterminé, selon chaque niveau de la covariable structure (K = 1 à 7), la moyenne du pourcentage de tests significatifs (PPS). L'intervalle de confiance de cette moyenne a été calculé pour toutes les modalités de K, sur la base de l'écart-type général (Pooled StDev = 4,157). Tous les intervalles de confiances se recoupent, exception faite pour la modalité K = 1 où PPS est exagérément grand. La non prise en compte de la structure conduit donc à un taux potentiellement élevé de fausses associations, dont PPS est ici un indicateur.

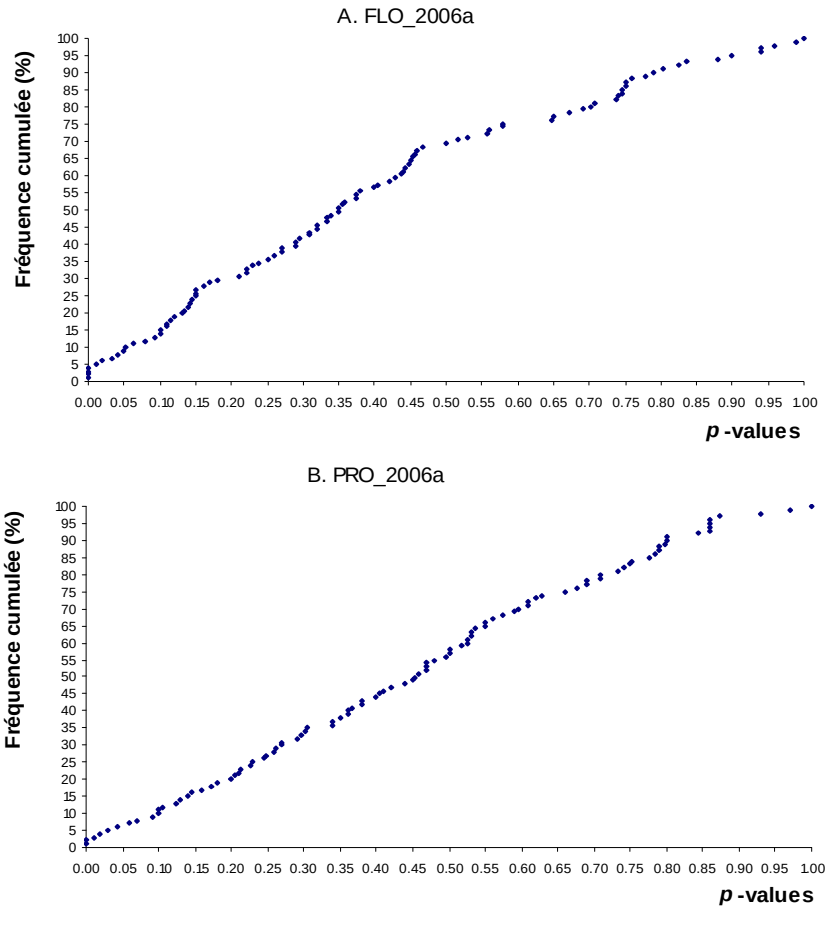


Figure 9 : Exemples de distributions nulles des probabilités critiques pour le test d'association par régression logistique de TASSEL. La distribution permet de déterminer la valeur de probabilité critique (*p-value*) en dessous de laquelle se situent 5% des valeurs de p. Les associations ont été effectuées avec un jeu de données de 119 allèles SSRs**. (A)** Distribution empirique de *p* pour l'analyse de la date de floraison dans l'essai 2006a. 5% des valeurs de *p* observées sont inférieures à 0.0101. Cette valeur constitue le seuil empirique α pour ce jeu de donnée particulier. **(B)** Distribution de *p* pour l'analyse du taux de protéines dans l'essai 2006a. Le seuil empirique correspond ici à la valeur 0.0309 au niveau de laquelle la fréquence cumulée est de 5%.

notamment du gène PhyB. Notre modèle de décision ne permet pas de conclure avec certitude dans cette zone, au sein de laquelle les tests pourraient, mais pas forcément, induire de fausses associations.

Tableau 8 : Seuils empiriques pour l'analyse d'association phénotype/génotype sur divers caractères chez le mil

Essai	CP1	CP2	FLO	HPR	NIN	PRO
2005	0.0100	0.0200	0.0200	0.0515	-	-
2006a	-	0.0241	0.0101	0.0114	0.0200	0.0309
2006b	-	0.0213	0.0100	0.0200	0.0300	0.0300

Les seuils ont été déterminés à l'aide de la distribution des probabilités critiques de tests de régression logistique (TASSEL) sur un jeu de données comportant des marqueurs microsatellites, des variables phénotypiques et une matrice de structure (INSTRUCT). Les variables phénotypiques, mesurés sur un panel de 90 lignées, sont la date de floraison (FLO), la hauteur de la plante à la récolte (HPR) ; le nombre d'inter-nœuds (NIN), le taux de protéines (PRO) et deux composantes principales (CP1 et CP2) basées sur la combinaison de plusieurs variables morphologiques.

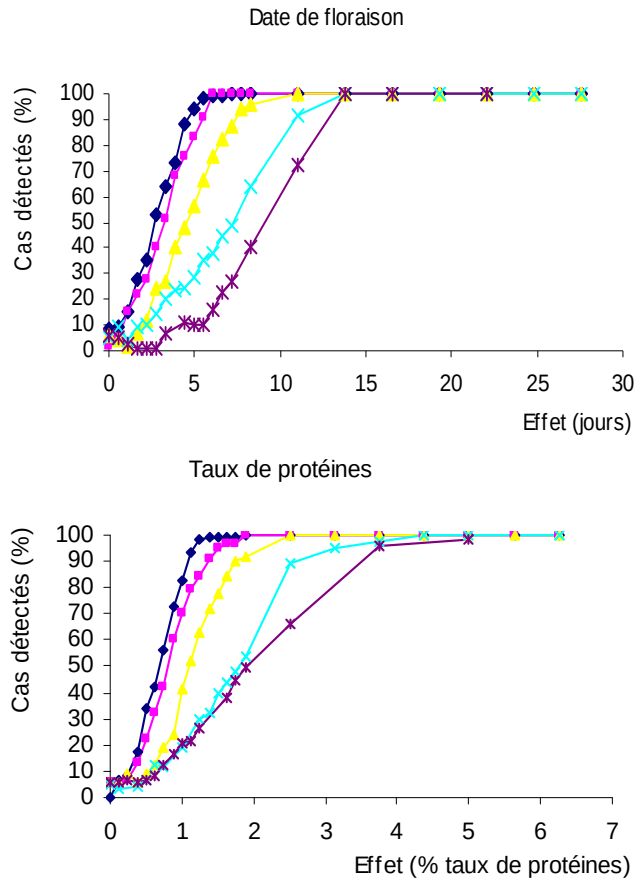


Figure 10: Pouvoir de la méthode d'association par régression logistique sous TASSEL. Les données phénotypiques (date de floraison et taux de protéines) ont été simulées de sorte que les polymorphismes candidats expliquent un effet additif sur le caractère. Cent (100) répétitions indépendantes ont été effectuées. L'axe des abscisses représente la variation expliquée par l'allèle candidat. Pour la floraison, l'effet varie de 0 à 28 jours (28 jours correspondent à 50% de la date de floraison moyenne lorsque l'effet de l'allèle est totalement exclu). L'axe des ordonnées représente le pourcentage de cas où l'effet simulé a été dûment détecté par la méthode ($p < 0.05$). Les courbes correspondent, de gauche à droite respectivement, à un modèle où l'allèle candidat est fréquent à 50%, 25% ; 12.5%, 6.25% et 3.12%. Le taux de détection donne une estimation de la puissance du test, qui apparaît très sensible à la fréquence allélique et à la force de l'effet. Les allèles rares seront quand bien même détectés à coup sûr si leur effet atteint un seuil de variation très élevé. Un effet agronomiquement exploitable créant un écart à la date moyenne de floraison de 7-8 jours, par exemple, est facilement détecté par la méthode (puissance > 95%), tant que l'allèle est présent chez seulement 1/8 des individus ou plus.

DISCUSSION

1. La détection du nombre K de clusters : une étape toujours délicate ?

Les algorithmes bayésiens, notamment STRUCTURE, arrivent à assigner les individus avec succès à leurs populations d'origine si le nombre K de ces populations est connu *à priori* et fixé pour le logiciel (Evanno et *al.,* 2005). Mais ce nombre K est souvent inconnu, et il est nécessaire de l'inférer au même titre que l'assignation des individus aux K différents groupes.
Dans notre cas, la distribution du log likelihood issu de STRUCTURE ne répond pas aux critères permettant une détection du K optimal, à savoir l'apparition d'une valeur modale ou d'une phase de plateau (Pritchard et *al.,* 2000). Ce même cas de difficulté a été mentionné dans des travaux ultérieurs (Camus-Kulandaivelu, 2007; Evanno et *al.,* 2005). La fonction L(K) présente souvent une distribution multimodale ou, à l'image de notre cas (*Figure 3*), une croissance quasiment constante sur tout l'intervalle (Camus-Kulandaivelu, 2007). L'augmentation prompte de la variance de L(K), à l'image de celle que nous avons observée à K = 2, a constitué cependant un élément de description parmi d'autres du comportement de la fonction L(K) au voisinage du K optimal (Evanno et *al.,* 2005). L'utilisation du critère L(K) sans le coupler à la fonction ΔK aurait peut être conduit à considérer cet élément comme un signal potentiel de structure optimale, même si ce signal ne serait pas suffisamment concluant. Ce faible signal de L(K) converge par ailleurs avec l'observation d'un pic sur ΔK, à la même abscisse K = 2 (*Figure 3*), sauf que ce pic n'est qu'un optimum local, moins élevé que celui observé à K = 6. Son observation ne remet donc pas en cause la considération de K = 6 comme optimum sur tout l'intervalle.
Notre analyse graphique a révélée que la septième population détectée par INSTRUCT est d'un effectif très faible et qu'elle est se serait différenciée à partir de la sixième population (*Figure 4*). La fluctuation du résultat des simulations entre 6 et 7 clusters optimaux pourrait donc relever du fait que ces 2 sous-populations, tantôt distinguées lors de l'inférence (K= 7) et tantôt confondues (K = 6) auraient des fréquences alléliques particulièrement similaires. Cette tendance chez la méthode INSTRUCT à mélanger deux sous-populations lorsqu'elles sont très proches a été mentionnée par Gao et al. (2007). La même tendance s'observe chez d'autres méthodes bayésiennes d'analyse de structure (Corander et *al.,* 2003 ; François et *al.,* 2006). En général, cela arrive lorsque le nombre de populations supposé est supérieur au nombre réel, ou lorsque le nombre d'individus pour certaines sous-populations réelles est très faible dans l'échantillon (Gao et *al.,* 2007)**.** L'augmentation du nombre d'individus et des

Tableau 9 : Association entre le gène PhyC et la variation des caractères

A.

Gène	SNP	Caractère	DiffLL	*p-value*	R^2_Marker
PhyC	101	CP1	24.2	<0.001	14.34%
PhyC	128	CP1	23.1	<0.001	14.06%
PhyC	155	CP1	23.1	<0.001	14.06%
PhyC	456	CP1	23.1	<0.001	14.06%
PhyC	615	CP1	20.5	<0.001	9.85%
PhyC	645	CP1	19.6	<0.001	9.70%
PhyC	697	CP1	18.5	<0.001	12.23%

B.

Gène	SNP	Caractère	DiffLL	*p-value*	R^2_Marker
PhyC	101	CP1	9.8	0.004	7.85%
PhyC	128	CP1	10.2	0.006	8.52%
PhyC	155	CP1	10.2	0.006	8.52%
PhyC	456	CP1	10.2	0.006	8.52%
PhyC	615	CP1	8.0	0.015	6.03%
PhyC	645	CP1	7.6	0.016	6.08%
PhyC	697	CP1	7.4	0.016	6.10%

C.

Gène	Site	Caractère	DiffLL	*p-value*	R^2_Marker
PhyC	101	CP1	25.2	<0.001	13.10%
PhyC	128	CP1	20.5	<0.001	11.59%
PhyC	155	CP1	20.5	<0.001	11.59%
PhyC	456	CP1	20.5	<0.001	11.59%
PhyC	615	CP1	18.2	0.001	8.76%
PhyC	645	CP1	17.6	0.001	8.98%
PhyC	697	CP1	16.7	0.001	8.87%

D.

Gène	Site	Caractère	DiffLL	*p-value*	R^2_Marker
PhyC	101	FLO	13.1	<0.001	10.68%
PhyC	128	FLO	8.2	0.005	6.99%
PhyC	155	FLO	8.2	0.005	6.99%
PhyC	456	FLO	8.2	0.005	6.99%
PhyC	615	FLO	7.0	0.009	6.01%
PhyC	645	FLO	7.9	0.007	7.06%
PhyC	697	FLO	7.5	0.012 (NS)	7,03%

Les résultats de l'analyse logistique sont présentés pour le gène PhyC testé vis-à-vis des traits: composante principale 1 de l'ACP (CP1) pour l'essai 2005 (**A**), 2006a (**B**), 2006 (**C**) et date de floraison (FLO) pour 2006b (**D**). Sept des huit SNPs que nous avons séquencés sont corrélés significativement (seuil de 1%) à la variable composite CP1 ($\alpha = 1\%$). Les réplications (essais 2005, 2006a et 2006b) confirment ce résultat (**A, B, C**). La part de variabilité R^2 expliqué par le modèle (Somme des Carrés des Ecarts de type III) est en moyenne de 10,64% pour l'essai 2006a, et 7,37% pour l'essai 2006b, et 12,61% pour l'essai 2005. La date de floraison (essai 2006b) est corrélée significativement à 6 marqueurs SNPs sur les 8 testés, (seuil 1%) avec une part de variabilité expliquée variant de 6,01% à 10,68% selon le site **(D)**.

groupes aux effectifs similaires peut permettre d'améliorer la précision et de la stabilité de la détection de K. Bozdogan (1993) préconise de fixer la limite supérieure des valeurs de K testées à $1 + [n^{0.3}]$, pour un échantillon de taille n. Cela dit, une structure atteignant le niveau que nous avons détecté lors de cette étude (K =7) aurait été détectée dans des conditions idéales, théoriquement, si l'échantillon comportait environ 392 individus (soit ~4 fois la taille de notre panel actuel).

2. Comparaison des 2 méthodes d'analyse de structure

Les matrices de structures inférées par les deux méthodes (INSTRUCT et STRUCTURE) sont fortement corrélées. Les résultats sont similaires en terme de groupes génétiques détectés et de nombre d'individus dans ces groupes. Le léger flottement observé sur une minorité d'individus entre différentes populations procède d'un écart résiduel. Cet écart pourrait être imputable en partie à l'absence de barrière étanche entre les populations, du fait des introgressions réciproques. Cela est vrai surtout pour des individus placés à la limite des groupes, c'est-à-dire ceux présentant des valeurs d'*ancestry* (Q) assez proches pour 2 ou 3 différentes populations. En plus, STRUCTURE amplifie dans le cas d'individus consanguins les signaux d'admixture (Gao et *al.,* 2007). Une variation même moindre sur les valeurs limites de Q ferraient basculer alors l'individu d'une population à l'autre.
D'autre part, nous avons montré que sur notre jeu de données les matrices d'INSTRUCT et celles de STRUCTURE conduisent à peu près à la même correction de l'effet de la structure dans les tests d'association (*Tableau 7*). Ce résultat est en phase avec les observations sur la forte similarité entre ces deux méthodes d'inférence de la structure. Même si INSTRUCT possède le meilleur modèle sous-jacent pour des lignées, les résultats d'association obtenus avec STRUCTURE restent assez proches. Ceci conforte la fiabilité de STRUCTURE, qui a été pendant longtemps le seul logiciel disponible et utilisé (Thorsnberry et *al.,* 2001).

3. Reconstitution de l'histoire évolutive

Notre analyse graphique (*Figure 4*) rend compte de la façon dont se subdivisent les clusters pour passer de K populations à K + 1 populations. Les modèles de STRUCTURE et INSTRUCT, regardés sous cet angle, pourraient s'interpréter comme une reconstitution éloquente de l'histoire évolutive des populations. Chaque augmentation du nombre de cluster est assimilable à la différenciation, sur une échelle évolutive, des fréquences alléliques chez

un certain nombre d'individus, qui redéfinissent un niveau hiérarchisé de structure. Des introgressions à partir de différentes populations d'origines sont détectées au sein des lignées. L'évolution de ces lignées se serait donc faite d'un brassage plus ou moins important entre les populations d'origines. Ces mélanges, quantitativement évalués au sein des matrices de structure, reflètent des flux de gènes qui se seraient produits par le passé entre les différentes populations. Pour une espèce naturellement allogame comme le mil, ces évènements biologiques sont des plus probables, les échanges entre populations étant favorisés par la pollinisation croisée.

4. Le contrôle de l'effet de la structure permet-il de limiter efficacement le taux de faux positifs en génétique d'association ?

Les tests d'association conduits avec des allèles neutres (microsatellites) ont révélé un taux de tests significatifs variable selon que l'on prenne ou non en compte la structure génétique des lignées. Cet estimateur confirme *à posteriori* le postulat selon lequel la structure génétique conduit inévitablement à de fausses associations (Pritchard et *al.,* 2000b ; Gao et *al.,* 2007;Yu et *al.,* 2005). Le contrôle de l'effet structure limite à moins de 9% en moyenne les associations détectées avec le jeu de marqueurs neutres, contre environ 17,19% lorsque l'on néglige la structure. Il est cependant intéressant de noter que la prise en compte simplement de K=2, une structure globale, est similaire dans notre cas voire meilleure que la structure fine K=7. L'évolution du taux de faux positif est variable suivant le phénotype considéré. Ceci est facilement compréhensible : un phénotype covariant avec la structure des populations doit montrer la plus forte correction d'association si la structure est prise en compte. A l'inverse un phénotype indépendant de la structure doit montrer une évolution indépendante de K.

Notre hypothèse est que les allèles microsatellites pris au hasard ne sont pas liés *à priori* à la variation des caractères étudiés. Sous cette hypothèse, on devrait s'attendre à ce que les valeurs de probabilité critique des tests d'association soient pour leurs quasi-totalités non significatives. Cela dit, on a considéré le taux de tests significatifs (PPS) dans l'analyse d'association avec ces allèles neutres comme un estimateur du taux de faux positifs. Ce taux semblerait d'ailleurs surestimé car certains des allèles microsatellites peuvent être réellement associés à des traits.

Le choix du seuil demeure un problème crucial dans les analyses d'association phénotype/génotype, y compris en cartographie QTL où par ailleurs l'effet de la structure est maîtrisé. La difficulté pour déterminer un seuil approprié est accentuée notamment par des

facteurs pouvant différer d'une expérience à l'autre. Parmi ces facteurs, on note la taille de l'échantillon, la taille du génome de l'organisme étudié, les données manquantes... (Revue par Churchill et Doerge, 1994). L'intérêt d'un seuil empirique, c'est son adaptation spécifique au jeu de données sur lequel se fait la détection de l'association (Churchill et Doerge, 1994). Pour chaque combinaison de jeu de données (caractère/essai), nous estimons alors que l'utilisation des seuils empiriques qui ont été définis lors de ces travaux serait bien appropriée. Ces seuils empiriques sont aussi une alternative à des corrections trop conservatives qui, tout en limitant les faux positifs, augmentent les faux négatifs.

Le pouvoir de la méthode TASSEL

Nous avons montré que dans le modèle considéré (effet additif), la capacité de détection de la régression logistique implémentée sous TASSEL est dépendante de la fréquence de l'allèle candidat et aussi de l'effet qu'il produit. Certaines configurations fréquence allélique/effet notamment se sont montrées très favorables à une détection sans faille des associations. Notre jeu de données a été simulé en partant de distributions de base provenant d'un jeu de données réel (date de floraison, taux de protéines), afin de maintenir sciemment la structure génétique. Le contrôle de cette structure est un élément qui pourrait potentiellement baisser la puissance du test, par une induction de faux négatifs. Les polymorphismes participant à la structure ne sont souvent pas détectés par de telles méthodes car confondues avec un effet de la structure. Les conditions du test ne sont donc pas des plus favorables à la détection. La balance en jeu est donc celle de l'antinomie entre un souci d'évitement des fausses associations (dont l'extrême serait une démarche trop conservative) et une nécessité de pouvoir pour détecter mêmes les associations limites tendant vers les zones de prudence.

Validation de gènes candidats par génétique d'association

Nous avons développé au cours de ce stage des approches permettant le contrôle de l'erreur de type I dans les analyses d'association. La démarche, sans être trop conservative, a montré sur un jeu de marqueurs neutres assez large une capacité à limiter les fausses associations. Ainsi, nous avions pu détecter des associations significatives en corrigeant l'effet de la structure et le seuil de détection. La répétabilité de cette corrélation est assez bonne pour notre variable composite (CP1). Pour la floraison, la significativité du test demeure tranchée pour l'essai 2006b. Pour les essais 2005 et 2006a, des probabilités critiques assez basses (<0.05)

sont observées. Si l'on applique strictement les seuils empiriques correspondants ces tests ne sont pas significatifs. Cependant, comme nous l'avons dit ces seuils peuvent être trop conservatifs si certains allèles microsatellites sont associés aux caractères. Les vrais seuils se trouvent donc entre ces deux extrêmes 5% et le seuil empirique corrigé. Nos obervations traduisent sans doute un effet détecté sur les deux essais 2005 et 2006a mais avec moins de force. Il est à noter que l'essai 2006b correspond à une date de semis tardive (mi juillet) alors que 2006a est une date assez précoce (mi juin). Si le caractère de floraison est influencé par la photopériode ceci peut aussi expliquer la différence entre essais. La date de semis pour 2005 est elle intermédiaire (début juillet). Cette différence peut donc s'expliquer par l'effet environnemental. Enfin, le jeu des 90 lignées (au niveau phénotype et génotype) n'est pas toujours complet, un certain nombre de données sont manquantes. Or la méthode, comme nous l'avons montré, est sensible à la fréquence des allèles, qui peut être modifiée d'un essai à l'autre à cause des manquants.

La répétabilité des associations est indiscutable avec PhyC pour CP1 sur tous les essais, et aussi sur au moins un essai pour la date de floraison. Les niveaux de variabilité expliquée par les polymorphismes de ce gène sur le phénotype sont comparables à ceux de certains polymorphismes validés chez Dwarf8 (Thornsberry et *al*., 2001). Nous avons aussi observé des corrélations entre PhyC et d'autres variables morphologiques, notamment le diamètre du rachis, la longueur de la chandelle, le nombre de talles productifs à maturité, le nombre de talles aériens à maturité, le diamètre de la tige principale, la date d'épiaison. Toutes ces variables sont assez corrélées à CP1 et à la floraison lors des 3 essais effectués. Le fait que PhyC soit un gène de floraison connu chez d'autres espèces laisse penser que l'association avec ces caractères passent par la floraison ; en tant normal, une floraison précoce est associé, par exemple, à un faible nombre de talles.

Les phytochromes participent à plusieurs voies du métabolisme végétal, notamment à la photomorphogénèse. Le gène PhyC a été identifié chez *Arabidopsis thaliana*. Le phytochrome C associé à ce locus est impliqué chez Arabidopsis dans la variation de la date de la floraison (Balasubramanian et *al.,* 2006). Il joue un rôle notamment dans la réception du stimulus lumineux et la régulation de l'horloge circadienne chez la plante. Chez le riz, les phytochromes permettent le contrôle photopériodique de la floraison. Ils interviennent plus ou moins en amont dans la voie de la floraison, pour laquelle on dénombre plus d'une soixantaine de gènes chez Arabidopsis.

CONCLUSION ET PERSPECTIVES

Les deux méthodes bayésiennes INSTRUCT et STRUCTURE ont permis d'inférer avec succès le nombre et la structure génétique des populations dans le panel de 90 lignées de mil. Nous avons ensuite développé une démarche pour le contrôle de l'erreur de type I dans la méthode d'association, qui inclut le contrôle de l'effet de la structure et la mise au point d'un modèle de décision basé sur la détermination empirique des seuils empiriques appropriés à chaque jeu de données spécifique.

D'autre part, la simulation effectuée montre que la méthode TASSEL dispose d'une potentialité de puissance appréciable pour mettre en évidence des associations entre polymorphismes et caractères d'intérêt. Cette étude simulée renseigne aussi sur les configurations de jeu de données qui seraient optimales pour une détection effective.

L'application de la méthode d'association TASSEL sur nos données expérimentales s'est révèle fructueuse. Nous avons mis en évidence des associations significatives entre les polymorphismes (SNPs) du gène candidat PhyC et la variation notamment de la date de floraison. Ces associations semblent pertinentes, et confirment l'intérêt de l'approche gène candidat développée par le laboratoire.

Aussi, des perspectives intéressantes existent-elles pour les approches développées. Nous avons montré que la structure inférée sur ce panel par les méthodes INSTRUCT et STRUCTURE est assez similaire. Toutes ces deux méthodes sont bayésiennes. L'utilisation pour l'inférence de la structure d'une approche non bayésienne, à l'exemple de la méthode *SMART-PCA* basée sur l'ACP (Patterson et *al.,* 2006) serait une voie intéressante à explorer, dans l'espoir d'un éventuel perfectionnement. De même, l'extension du contrôle de la structure à l'effet de l'apparentement entre les lignées (Yu et *al.,* 2006) pourrait-il rendre plus pointue l'analyse d'association, par un meilleur contrôle des liens entre individus. La réestimation du pouvoir de la méthode en intégrant des effets comme l'épistasie entre les locus rendrait notre modèle de simulation plus proche de la complexité réelle de certaines voies génétiques. Enfin, l'identification en masse de gènes candidats permettra de tester un plus grand nombre de marqueurs. Le laboratoire travaille déjà sur les stratégies de *genome scan*. L'exploitation de la synthénie pourrait également être probante, vu des colocalisations déjà mises en évidence entre le riz et le mil, par exemple (Gale et *al.*, 2006). A terme, les résultats de ces projets seront utiles pour la gestion des ressources génétiques et l'innovation variétale.

BIBLIOGRAPHIE

- Agrhymet, Comité permanent Inter-états de Lutte contre la Sécheresse dans le Sahel – Centre Régional Agrhymet (CILSS – CRA) ; CIRAD, Centre de coopération Internationale en Recherche Agronomique pour le Développement ; 2005. Après la famine au Niger…Quelles actions de lutte et de recherche contre l'insécurité alimentaire au sahel ? Dossier de presse, décembre 2005, 41 p.
- Balasubramanian et *al.,* 2006. The phytochrome C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. Nature Genetics 38: 711-715.
- Bezançon Gilles, Renno Jean-François, Kumar K. Anand, 1994. *Le mil*. *In :* L'*Amélioration des Plantes Tropicales*. Edition CIRAD & OSRTOM. *La Librairie du CIRAD*, Montpellier France, 1994; 457-481. ISSN 1251-7224.
- Bozdogan, H., 1993. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix, Information and Classification, O. Opitz, B. Lausen, and R. Klar (eds.). Heidelberg: Springer-Verlag.
- Camus-Kulandaivelu, 2007. Thèse de doctorat en génétique végétale, UMR 8120 (Gif-sur-Yvette, France). Titre: " Evolution génomique du maïs durant son adaptation aux conditions européennes". Encadrement : Dr. Charcosset, Dr. Manicacci (UMR de Génétique Végétale du Moulon, Gif-sur-Yvette, France) and Dr. Gouesnard (UMR GGPC, Montpellier, France).
- Chardon F., Virlon B., Moreau L., Falque M., Joets J., Decousset L., Murigneux A., Charcosset A., 2004. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics 168(4): 2169-2185.
- Childs K. L., Miller F. R., Cordonnier-Pratt M. M., Pratt L. H., Morgan P. W., Mullet J. E., 1997. The Sorghum Photoperiod Sensitivity Gene Ma3 Encodes a Phytochrome B. Plant Physiol. 113: 611-619.
- Churchill G.A., Doerge R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 138: 963-971.
- Corander J., Waldamann P., Sillanpaa M., 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163: 367–74.
- Evanno G., Regnaut S., Goudet J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611-2620.

- Falush D., Stephens M., Pritchard J.K., 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics 164: 1567–1587.
- FAO, 2006. FAOSTAT, [En ligne] http://faostat.fao.org/site/336/DesktopDefault.aspx?PageID=336 (Déc. 2006)
- François O., Ancelet S., Guillot G., 2006. Bayesian clustering using hidden markov random fields in spatial population genetics. Genetics 174: 805–16.
- Gale M.D., Devos K.M., Zhu J.H., Allouis S., Couchman M.S., Liu H., Pittaway T.S., Qi X.Q., Kolesnikova-Allen M., Hash C.T., 2006. *New Molecular Marker technologies for Pearl millet Improvement.* Research Paper; John Innes Centre, Norwich, UK & ICRISAT, Patancheru, India; 7 p.
- Gao H., Williamson S., Bustamante C.D., 2007. An MCMC Approach for Joint Inference of Population Structure and Inbreeding Rates from Multi-Locus Genotype Data. Genetics: 10.1534/genetics.107.072371.
- Hansen James, Makiko Sato Makiko, Ruedy Reto, Lo Ken, W. Lea David, Medina-Elizade Martin, 2006. Global temperature change. PNAS 103: 14288-14293.
- Hayama Ryosuke, Yokoi Shuji, Tamaki Shojiro, Yano Masahiro, Shimamoto Ko, 2003. Adaptation of photoperiodic control pathways produces short-day flowering in rice. Nature 422 : 719 – 722.
- IRD, 2004. *Sciences du Sud*- Le journal de l'IRD- N° 26 sept/oct 2004, p. 8.
- James et *al.*, 2003. Starch synthesis in the cereal endosperm. Current Opinion in *Plant Biology* 6:21-222.
- Jideani, 2005. Characteristics of local pearl millet (*Pennisetum glaucum*) grains. *Nigerian Food Journal* 23:193-204.
- LAURET Stéphane, 2006. Diversité moléculaire, domestication et sélection chez le mil. Application à des gènes impliqués dans la qualité de la graine. Master 2 Recherche Ressources Phytogénétiques et Interactions Biologiques. Université Montpellier 2.
- Mariac C., Luong V., Kapran I., Mamadou A., Sagnard F., Deu M., Chatereau J., Gerard B., Ndjeunga J., Bezançon G., Pham J.-L., Vigouroux Y., 2006. Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assed by microsatellite markers. Theor Appl Genet 114:49-58.
- Olsen K. J., Purugganan M. D., 2002. Molecular Evidence on the Origin and Evolution of Glutinous Rice. Genetics 162: 941–950.

- Patterson N., Price A.L., Reich D., 2006. Population Structure and Eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190
- Poncet V., 1998. Organisation génétique du syndrome de domestication du mil (*Pennisteum glaucum*, Poacea). Thèse, Université de Paris-Sud, Orsay. 116 p.
- Pritchard J.K., Stephens M., Donnelly P., 2000a. Inference of Population structure Using Multilocus Genotype Data. Genetics 155: 945-959.
- Pritchard J.K., Stephens M., Rosenberg N. A., Donnelly P., 2000b. Association Mapping in Structured Population. Am. J. Hum. Genet. 67:170–181.
- Pritchard J.K., Donnelly P., 2001. Case–Control Studies of Association in Structured or Admixed Populations. Theoretical Population Biology 60, 227–237.
- Tenaillon M., Sawkins M.C., Gaut R.L., Long A.D., Doebleyb J.F., Gaut B.S.,2001. Patterns of DNA sequence polymorphism along chromosom 1 of maize (Zea mays ssp. Mays). Proc. Natl. Acad. Sci. USA 98, 9161-9166.
- Thornsberry J.M., Goodman Major M., Doebley J., Kresovich S., Nielsen D., Buckler IV E. S., 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28: 286-289.
- Wilson L.M., Whitt S.R., Ibanez A. M., Rocheford T. R., Goodman M.M., Buckler E. S., 2004. Dissection of Maize Kernel Composition and Starch Production by Candidate Gene Association. The Plant Cell, 16: 2719–2733.
- Yadav R.S., Hash C.T., Bidinger F.R., Cavan G.P., Howarth C.J., 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 R.S.., Bidinger F.R., Hash C. T., Yadav Y. P., Yadav O. P., Bhatnagars S. K., Howarth J., 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.
- Yano, M., Y. Katayose, M. Ashikari, U. Yamanouchi, et L. Monna, (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell, 12: 2473-2483.
- Yu J., Pressoir G., Briggs W. H., Bi I. V., Yamasaki M., Doebley J.F., McMullen M. D., Gaut B. S., Holland J. B., Kresovish S., Buckler E., 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 233-208.

ANNEXE

Résultats de tests d'association phénotype/génotype sous TASSEL

On reporte ici les résultats de l'analyse l'association (régression linéaire) pour les polymorphismes (SNPs) séquencés sur 6 gènes candidats : **(A) :** Opaque2, candidat pour la variation du taux de protéines, et 5 gènes candidats pour la date de floraison **(B à F)** : Floricaula, Gigantea, HD3A, PhyA et PhyB. **(G) :** corrélations entre le gène PhyC et 7 variables morphologiques corrélées à la composante principale 1 (CP1). Pour la description de CP1 et des corrélations entre le gène PhyC et CP1, voir le texte principal.

A. Test d'association par régression logistique entre des SNPs sur les fragments 1 et 2 du gène Opaque-2 et le taux de protéines des lignées de mil

Essai 2006a

Locus	Site	DiffLL	*P-value*
opaque2_1	68	NaN	0.00
opaque2_1	118	NaN	0.00
opaque2_1	129	0.17	0.41
opaque2_1	143	0.00	1.00
opaque2_1	183	NaN	0.00
opaque2_1	184	NaN	0.00
opaque2_1	185	NaN	0.00
opaque2_1	221	4.19	0.03
opaque2_1	231	0.30	0.52
opaque2_1	240	0.42	0.59
opaque2_1	249	0.30	0.52
opaque2_1	284	0.00	1.00
opaque2_1	314	0.00	1.00
opaque2_1	317	0.00	1.00
opaque2_1	324	0.00	1.00
opaque2_1	334	NaN	0.00
opaque2_1	387	NaN	0.00
opaque2_1	397	0.32	0.56
opaque2_1	407	0.32	0.56
opaque2_1	420	0.04	0.90
opaque2_1	422	0.36	0.61
opaque2_1	436	0.00	1.00
opaque2_1	438	0.32	0.56
opaque2_1	457	NaN	0.00
opaque2_1	466	0.00	1.00
opaque2_1	476	0.00	1.00
opaque2_1	588	0.05	0.86
opaque2_2	70	0.05	0.84
opaque2_2	81	0.00	1.00
opaque2_2	128	NaN	0.00
opaque2_2	131	0.00	1.00
opaque2_2	157	0.15	0.72
opaque2_2	183	0.35	0.30
opaque2_2	201	0.38	0.45
opaque2_2	202	0.32	0.43
opaque2_2	245	NaN	0.00
opaque2_2	310	0.00	1.00
opaque2_2	319	0.00	1.00
opaque2_2	399	0.32	0.52
opaque2_2	402	0.16	0.66
opaque2_2	413	NaN	0.00
opaque2_2	513	NaN	0.00
opaque2_2	547	0.00	1.00
opaque2_2	594	0.16	0.66
opaque2_2	594	NaN	0.00
opaque2_2	596	NaN	0.00
opaque2_2	613	0.39	0.48
opaque2_2	643	0.00	1.00
opaque2_2	661	0.00	1.00

Essai 2006b

Locus	Site	DiffLL	*P-value*
opaque2_1	68	0.02	0.70
opaque2_1	118	0.02	0.70
opaque2_1	129	#NOM?	0.82
opaque2_1	143	0.00	1.00
opaque2_1	183	0.01	0.88
opaque2_1	184	0.01	0.89
opaque2_1	185	0.01	0.88
opaque2_1	221	4.99	0.06
opaque2_1	231	0.00	0.88
opaque2_1	240	0.00	0.99
opaque2_1	249	0.00	0.88
opaque2_1	284	0.00	1.00
opaque2_1	314	0.00	1.00
opaque2_1	317	0.00	1.00
opaque2_1	324	0.00	1.00
opaque2_1	334	0.08	0.40
opaque2_1	387	0.08	0.40
opaque2_1	397	0.11	0.72
opaque2_1	407	0.11	0.72
opaque2_1	420	0.14	0.73
opaque2_1	422	0.20	0.70
opaque2_1	436	0.00	1.00
opaque2_1	438	0.11	0.72
opaque2_1	457	NaN	0.00
opaque2_1	466	0.00	1.00
opaque2_1	476	0.00	1.00
opaque2_1	588	0.00	0.96
opaque2_2	70	0.08	0.79
opaque2_2	81	0.00	1.00
opaque2_2	128	0.26	0.48
opaque2_2	131	0.00	1.00
opaque2_2	157	0.22	0.63
opaque2_2	183	0.02	0.68
opaque2_2	201	0.21	0.60
opaque2_2	202	0.34	0.57
opaque2_2	245	2.88	0.10
opaque2_2	310	0.00	1.00
opaque2_2	319	0.00	1.00
opaque2_2	399	0.00	0.95
opaque2_2	402	0.07	0.81
opaque2_2	413	NaN	0.00
opaque2_2	513	0.00	1.00
opaque2_2	547	0.00	1.00
opaque2_2	594	0.07	0.81
opaque2_2	594	NaN	0.00
opaque2_2	596	NaN	0.00
opaque2_2	613	0.05	0.77
opaque2_2	643	0.00	1.00
opaque2_2	661	0.00	1.00

B. Test d'association par régression logistique entre des polymorphismes du gène Floricaula et la date de floraison chez le mil

Essai 2005

Site	DiffLL	*P-value*
104	1.70	0.30
232	0.00	1.00
456	0.00	1.00
457	0.00	1.00
495	0.00	1.00
499	#NOM?	0.79
504	0.00	1.00
571	0.00	1.00
575	1.15	0.31
686	0.00	1.00
728	0.00	1.00

Essai 2006a

Site	DiffLL	*P-value*
104	4.07	0.12
232	0.00	1.00
456	0.00	1.00
457	0.00	1.00
495	0.00	1.00
499	#NOM?	0.88
504	0.00	1.00
571	0.00	1.00
575	1.74	0.21
686	0.00	1.00
728	0.00	1.00

Essai 2006b

Site	DiffLL	*P-value*
104	2.01	0.18
232	0.00	1.00
456	0.00	1.00
457	0.00	1.00
495	0.00	1.00
499	#NOM?	0.95
504	0.00	1.00
571	0.00	1.00
575	1.73	0.24
686	0.00	1.00
728	0.00	1.00

C. Test d'association par régression logistique entre les polymorphismes du gène Gigantea et la date de floraison chez le mil

Essai 2005

Site	DiffLL	*P-value*
121	1.34	0.33
284	0.00	1.00
373	0.01	0.75
385	0.01	0.75
439	0.25	0.55
509	3.52	0.03
509	3.52	0.03
509	3.52	0.03
509	3.52	0.03
624	0.25	0.55
706	0.15	0.50
825	0.00	1.00
960	0.33	0.39
1086	0.00	1.00
1087	0.31	0.42
1123	0.00	1.00
1208	0.00	1.00
1215	0.75	0.36

Essai 2006a

Site	DiffLL	*P-value*
121	3.09	0.07
284	0.00	1.00
373	1.43	0.26
385	1.43	0.26
439	0.99	0.39
509	1.24	0.30
509	1.24	0.30
509	1.24	0.30
509	1.24	0.30
624	0.99	0.43
706	0.20	0.51
825	0.00	1.00
960	NaN	0.00
1086	0.00	1.00
1087	NaN	0.00
1123	0.00	1.00
1208	0.00	1.00
1215	1.17	0.33

Essai 2006b

Site	DiffLL	*P-value*
121	0.03	0.86
284	0.00	1.00
373	0.04	0.77
385	0.04	0.77
439	0.16	0.68
509	0.04	0.80
509	0.04	0.80
509	0.04	0.80
509	0.04	0.80
624	0.16	0.64
706	0.12	0.64
825	0.00	1.00
960	2.54	0.18
1086	0.00	1.00
1087	2.55	0.22
1123	0.00	1.00
1208	0.00	1.00
1215	0.05	0.58

D. Test d'association par régression logistique entre des polymorphismes du gène HD3A et la date de floraison chez le mil

Essai 2005

Site*	DiffLL	*P-value*
0	1.35	0.31
1	0.00	1.00
2	0.01	0.92
3	2.29	0.14
4	2.29	0.14
5	2.29	0.14
6	1.46	0.24

Essai 2006a

Site	DiffLL	*P-value*
0	3.57	0.06
1	0.00	1.00
2	0.51	0.52
3	1.74	0.21
4	1.74	0.21
5	1.74	0.21
6	NaN	0.00

Essai 2006b

Site	DiffLL	*P-value*
0	0.01	0.85
1	0.00	1.00
2	0.04	0.78
3	0.26	0.60
4	0.26	0.60
5	0.26	0.60
6	0.57	0.29

* Pour HD3A, ces noms de sites ne correspondent pas à la position sur le chromosome, contrairement à la notation pour les autres gènes ici présentés.

E. Test d'association par régression logistique entre des polymorphismes du gène PhyA et la date de floraison chez le mil

Essai 2005

Site	DiffLL	*P-value*
146	0.02	0.92
170	0.02	0.92
337	0.00	0.95

Essai 2006a

Site	DiffLL	*P-value*
146	3.10	0.07
170	1.89	0.15
337	1.48	0.21

Essai 2006b

Site	DiffLL	*P-value*
146	0.00	0.96
170	0.00	0.96
337	0.01	0.92

F. Test d'association par régression logistique entre des polymorphismes du gène PhyB et la date de floraison chez le mil

Essai 2005

Site	DiffLL	*P-value*
166	0.00	0.41
175	NaN	0.00
265	7.76	0.01
302	NaN	0.00
325	NaN	0.00
548	3.89	0.03
775	1.15	0.32
794	0.00	1.00
893	0.00	1.00

Essai 2006a

Site	DiffLL	*P-value*
166	0.00	1.00
175	NaN	0.00
265	6.23	0.02
302	NaN	0.00
325	NaN	0.00
548	#NOM?	0.60
775	0.04	0.76
794	0.00	1.00
893	0.00	1.00

Essai 2006b

Site	DiffLL	*P-value*
166	0.00	1.00
175	NaN	0.00
265	3.99	0.08
302	NaN	0.00
325	NaN	0.00
548	0.34	0.29
775	0.18	0.70
794	0.00	1.00
893	0.00	1.00

G. Test d'association par régression logistique entre des polymorphismes du gène PhyC et différentes variables morphologiques corrélées à la composante principale 1 :

Essai 2005

Trait	Site	DiffLL	p-value
DRA	101	15.30	0.00
DRA	122	#NOM?	0.82
DRA	128	18.79	0.00
DRA	155	18.79	0.00
DRA	456	18.79	0.00
DRA	615	15.92	0.00
DRA	645	14.95	0.00
DRA	697	14.40	0.00
DTP	101	8.86	0.00
DTP	122	Infinity	0.00
DTP	128	10.18	0.00
DTP	155	10.18	0.00
DTP	456	10.18	0.00
DTP	615	7.95	0.01
DTP	645	7.11	0.02
DTP	697	6.47	0.02
EPI	101	9.30	0.00
EPI	122	Infinity	0.00
EPI	128	6.52	0.03
EPI	155	6.52	0.03
EPI	456	6.52	0.03
EPI	615	4.67	0.06
EPI	645	5.29	0.05
EPI	697	5.21	0.04
FLO	101	7.36	0.01
FLO	122	Infinity	0.00
FLO	128	5.11	0.04
FLO	155	5.11	0.04
FLO	456	5.11	0.04
FLO	615	3.34	0.10
FLO	645	3.87	0.08
FLO	697	3.86	0.08
LOC	101	11.55	0.00
LOC	122	NaN	0.00
LOC	128	11.87	0.00
LOC	155	11.87	0.00
LOC	456	11.87	0.00
LOC	615	9.87	0.00
LOC	645	8.84	0.01
LOC	697	8.66	0.01
TAA	101	7.65	0.02
TAA	122	NaN	0.00
TAA	128	10.10	0.01
TAA	155	10.10	0.01
TAA	456	10.10	0.01
TAA	615	7.59	0.01
TAA	645	7.34	0.01
TAA	697	8.23	0.01
TAP	101	20.89	0.00
TAP	122	Infinity	0.00
TAP	128	19.35	0.00
TAP	155	19.35	0.00
TAP	456	19.35	0.00
TAP	615	16.84	0.00
TAP	645	16.94	0.00
TAP	697	15.89	0.00

Essai 2006a

Trait	Site	DiffLL	p-value
DRA	101	11.84	0.00
DRA	122	Infinity	0.00
DRA	128	14.23	0.00
DRA	155	14.23	0.00
DRA	456	14.23	0.00
DRA	615	11.81	0.00
DRA	645	11.32	0.00
DRA	697	11.47	0.00
DTP	101	7.77	0.01
DTP	122	NaN	0.00
DTP	128	8.44	0.01
DTP	155	8.44	0.01
DTP	456	8.44	0.01
DTP	615	6.72	0.03
DTP	645	6.08	0.03
DTP	697	5.71	0.04
EPI	101	5.43	0.05
EPI	122	Infinity	0.00
EPI	128	5.66	0.04
EPI	155	5.66	0.04
EPI	456	5.66	0.04
EPI	615	3.91	0.08
EPI	645	4.12	0.08
EPI	697	3.82	0.08
FLO	101	5.10	0.01
FLO	122	Infinity	0.00
FLO	128	5.33	0.02
FLO	155	5.33	0.02
FLO	456	5.33	0.02
FLO	615	3.49	0.06
FLO	645	3.75	0.05
FLO	697	3.47	0.08
LOC	101	10.36	0.00
LOC	122	Infinity	0.00
LOC	128	10.78	0.00
LOC	155	10.78	0.00
LOC	456	10.78	0.00
LOC	615	8.76	0.00
LOC	645	8.30	0.01
LOC	697	8.17	0.01
TAA	101	2.85	0.09
TAA	122	Infinity	0.00
TAA	128	2.37	0.18
TAA	155	2.37	0.18
TAA	456	2.37	0.18
TAA	615	0.96	0.32
TAA	645	1.06	0.31
TAA	697	0.77	0.41
TAP	101	7.98	0.00
TAP	122	NaN	0.00
TAP	128	7.88	0.01
TAP	155	7.88	0.01
TAP	456	7.88	0.01
TAP	615	6.15	0.01
TAP	645	6.12	0.01
TAP	697	5.57	0.01

Essai 2006b

Trait	Site	DiffLL	p-value
DRA	101	16.09	0.00
DRA	122	Infinity	0.00
DRA	128	19.73	0.00
DRA	155	19.73	0.00
DRA	456	19.73	0.00
DRA	615	17.45	0.00
DRA	645	16.67	0.00
DRA	697	15.87	0.00
DTP	101	10.92	0.00
DTP	122	Infinity	0.00
DTP	128	12.17	0.00
DTP	155	12.17	0.00
DTP	456	12.17	0.00
DTP	615	10.01	0.00
DTP	645	8.92	0.01
DTP	697	8.38	0.01
EPI	101	12.83	0.00
EPI	122	Infinity	0.00
EPI	128	7.86	0.01
EPI	155	7.86	0.01
EPI	456	7.86	0.01
EPI	615	6.77	0.02
EPI	645	7.47	0.02
EPI	697	7.27	0.02
FLO	101	13.12	0.00
FLO	122	Infinity	0.00
FLO	128	8.17	0.01
FLO	155	8.17	0.01
FLO	456	8.17	0.01
FLO	615	6.96	0.01
FLO	645	7.85	0.01
FLO	697	7.57	0.01
LOC	101	10.31	0.01
LOC	122	NaN	0.00
LOC	128	10.46	0.01
LOC	155	10.46	0.01
LOC	456	10.46	0.01
LOC	615	8.46	0.01
LOC	645	7.83	0.01
LOC	697	7.53	0.02
TAA	101	0.97	0.38
TAA	122	Infinity	0.00
TAA	128	0.66	0.45
TAA	155	0.66	0.45
TAA	456	0.66	0.45
TAA	615	0.16	0.71
TAA	645	0.08	0.78
TAA	697	0.23	0.66
TAP	101	10.66	0.00
TAP	122	Infinity	0.00
TAP	128	7.56	0.01
TAP	155	7.56	0.01
TAP	456	7.56	0.01
TAP	615	6.02	0.01
TAP	645	6.30	0.02
TAP	697	5.70	0.02

Développement de méthodes de génétique d'association et application à l'analyse de la qualité et de la floraison chez le mil. SAIDOU Abdoul-Aziz. Mémoire de Master 2. Equipe d'accueil :IRD, UMR DIA-PC, soutenu le 20/09/2007.

Nous nous sommes intéressés au développement chez le mil de méthodes permettant l'analyse des associations phénotype/génotype au niveau populationnel. L'objectif était de mettre en place un contrôle de l'erreur de type I (faux positifs) et aussi d'estimer le pouvoir de l'analyse d'association par régression logistique (TASSEL). L'écueil de la non connaissance *à priori* de la structure génétique au sein du panel, commun aux études de ce genre, est levé en inférant cette structure génétique par des algorithmes bayésiens ((INSTRUCT et STRUCTURE). L'analyse a été conduite de façon comparée, ce qui a aussi permis de faire l'interface entre les deux méthodes mises en oeuvre. Nous montrons ensuite que la structure des populations augmente significativement le risque de fausses associations, mais qu'un outil de contrôle permet de limiter efficacement ces faux positifs. Cet outil inclut le contrôle de la structure, et une amélioration du modèle de décision par la correction du seuil α. D'autre part, une simulation a été faite afin d'estimer la capacité de la méthode d'association TASSEL à détecter des associations significatives. Cela révèle que la méthode dispose d'une puissance potentielle globalement élevée. Enfin, l'application de cette méthode à notre jeu de données a permis de mettre en évidence une association significative entre des polymorphismes du gène candidat PhyC et la variation de la date de floraison chez le mil.

Mots clefs : association, structure génétique, contrôle des faux positifs, simulation, PhyC.

Development of association study methods and application to the study of grain quality and flowering time variation in pearl millet. SAIDOU Abdoul-Aziz. Master 2 report. IRD, UMR DIA-PC, defended on 20/09/2007.

This work deals with the use of association studies in pearl millet. Our goal was to develop an approach for controlling type I error (spurious associations) and to evaluate the power of the method. The genetic structure is assess using bayesian clustering algorithms. Two methods implemented in the software INSTRUCT and STRUCTURE were used and compared. We show that genetic structure leads to spurious association, but this may be effectively avoided using an adequate control. Our control consists of using structure matrix in association test and correcting threshold by assessing experiment-wise values. On the other hand, our results on simulated data suggest that the regression model of TASSEL used for these association analyses could have in general enough power to detect significant correlations. Finally, the application of this method to our data gives evidence of association between PhyC polymorphisms and flowering time variation in pearl millet.

Key words: association studies, population structure, false positif control, simulation, PhyC.