Les Actes du BRG, 6 (2006) 57-74


**Article original**

# Analyse et prédiction des patrons de déséquilibre de liaison dans les collections de ressources génétiques de plantes pérennes ou annuelles, autogames ou allogames

Marc SEGUIN[(1)*], Agnès ATTARD[(1)], Thomas BATAILLON[(2)], Claire BILLOT[(1)], Alberto CENCI[(2)], Nathalie CHANTRET[(2)], Brigitte COURTOIS[(1)], Jacques DAVID[(2)], Monique DEU[(1)], Najoi EL AZHARI[(1)], Jean-Christophe GLASZMANN[(1)], Sylvain GLEMIN[(3)], Annabelle HAUDRY[(2)], Vincent LE GUEN[(1)], Marie MAYNADIER[(2)], Virginie POMIÈS[(1)], Joëlle RONFORT[(2)], Anne TSITRONNE[(2)], Christelle WEBER[(1)]

[(1)] Cirad, UMR 1096, Polymorphismes d'Intérêt Agronomique, Avenue Agropolis, 34398 Montpellier Cedex 5, France
[(2)] INRA, UMR 1097, Diversité et Génome des Plantes Cultivées, Domaine de Melgueil, 34130 Mauguio, France
[(3)] Université Montpellier 2, UMR Génome populations Interactions Adaptation, CC 63 Bât. 24, place Eugène Bataillon, 34095 Montpellier, Cedex 5, France

**Abstract: Analysis of linkage disequilibrium patterns in perennial or annual, autogamous or allogamous plant species.** Statistical dependence between alleles at a pair of loci, or linkage disequilibrium (LD) was studied in rice, sorghum, wheat, rubber tree and *Medicago truncatula* using samples drawn from germplasm collections. LD was studied in restricted genomic areas which required to develop closely linked markers. Approaches using BAC resulted in the development of microsatellite markers in localized genomic area. In rubber tree, BACs were localized using mapped markers. The BAC end sequences were used to detect microsatellites motifs. On wheat and *Medicago*, primers were designed to allow direct sequencing of gene fragments. The high number of alleles for the microsatellite markers necessitated to group alleles with close sizes into the same allelic class in order to estimate LD. The genetic structure observed on the cultivated crops yielded a strong noise on LD, while LD was low in the samples drawn from collections of wild selfing species. Among the varietal groups defined for the cultivated crops, LD was lower than in the global sample and decayed rather rapidly, sometimes differently between groups. The decay is around 100 kb in rice and 500 kb in sorghum. Therefore, a good knowledge of the species diversity appears as a necessary pre-requisite. Those conditions fulfilled, association genetics studies should be successful in the different species studied here.

* Correspondance et tirés à part : marc.seguin@cirad.fr

**germplasm/ linkage disequilibrium/ microsatellite/ SNP/ genetic structure**

**Résumé** : L'analyse de la liaison statistique entre allèles à 2 locus, ou déséquilibre de liaison (DL), a été abordée simultanément chez le riz, le sorgho, le blé, l'hévéa et *Medicago truncatula* en utilisant des échantillons tirés de collections de ressources génétiques. Des stratégies utilisant des BACs ont permis de développer des marqueurs dans des zones localisées. L'utilisation des séquences terminales des BACs ciblés a été efficace pour définir des marqueurs microsatellites. Par ailleurs, des amorces ont été développées pour séquencer directement des fragments géniques. Le haut niveau de polymorphisme des marqueurs microsatellites a pu nécessiter des adaptations pour le calcul du DL. La structure génétique observée sur les plantes cultivées génère un fort bruit de fond de DL, tandis que dans les collections d'espèces sauvages autogames, le DL reste faible même à courte distance, décroissant rapidement au delà de 20 kb. A l'intérieur des groupes définis pour les espèces cultivées, le DL est plus faible et décroît assez rapidement, parfois de manière différente selon les groupes, de 100 kb environ pour le riz à 500 kb pour le sorgho. Une bonne connaissance de la diversité de l'espèce apparaît donc nécessaire avant de réaliser des études d'associations qui paraissent dès lors réalisables en utilisant les accessions contenues dans les collections de ressources génétiques.

**ressources génétiques/ déséquilibre de liaison/ microsatellite/ SNP/ structure génétique**

## 1. INTRODUCTION

La détection de QTL (Quantitative Trait Locus) a permis de localiser des gènes contrôlant des caractères agronomiques importants sur la plupart des génomes végétaux. Cette approche a deux défauts : elle oblige à développer une population de cartographie et, à moins de disposer d'effectifs de plusieurs milliers de descendants, la position du QTL est imprécise et il n'est généralement pas possible de déterminer le gène qui, parmi ceux co-ségrégeant avec le QTL, est impliqué dans la variation du caractère. Des méthodes récentes, inspirées de la génétique humaine, proposent d'utiliser directement une collection de génotypes faiblement apparentés pour détecter des QTLs. Ces méthodes, dites « d'association » (association studies or LD mapping, [1]), reposent sur le déséquilibre de liaison (DL), qui mesure l'association statistique entre allèles observés à deux sites polymorphes du génome. Cette association dépend de la distance physique qui sépare ces deux sites, mais aussi des propriétés génomiques de la région considérée (taux de recombinaison local, taux de mutation) et de l'histoire évolutive de l'espèce ou de la population étudiée (goulots d'étranglement, expansion démographique, mode de reproduction). Jusqu'à récemment le DL ne pouvait que rarement être étudié entre des marqueurs physiquement liés. L'avancée des connaissances sur le génome et l'accès au polymorphisme de

séquences permet aujourd'hui d'aborder l'étude du DL le long de segments chromosomiques de taille physique connue. Pour les études d'association, un DL limité à des sites physiquement proches (quelques kb) peut permettre une cartographie fine de gènes d'intérêt. À l'inverse, si le DL se maintient sur des dizaines de kilobases ou plus, la résolution de l'analyse sera réduite mais un maillage relativement lâche du génome pourra permettre de repérer les régions génomiques qui participent à la variation phénotypique d'un caractère donné (approche de type « genome scan », [2]).

Pour poursuivre l'analyse et l'exploitation des ressources génétiques aujourd'hui présentes au sein des collections de ressources génétiques, des développements méthodologiques de génétique des populations fondés sur le DL et adaptés aux caractéristiques particulières des plantes cultivées (autogamie *vs* allogamie, échantillons tirés de collections) sont nécessaires. Des données empiriques sur les patrons de DL effectivement observables au sein des collections sont, par ailleurs, indispensables pour mieux comprendre comment le DL est organisé au sein des populations et des collections. Les travaux présentés ici concernent des situations assez différentes, tant sur le plan des espèces comparées que sur les outils utilisés. Pour l'ensemble de ces espèces, il s'agissait i) de déterminer l'effet de la structure génétique des échantillons sur l'étendue du DL et ii) d'évaluer la faisabilité d'études d'association au sein des collections de ressources génétiques établies pour ces espèces. Au-delà de ces questions générales, les différents cas étudiés fournissent un ensemble d'exemples méthodologiques pour le développement de marqueurs génétiques pour l'étude du déséquilibre de liaison, et adaptés à des situations variées (information plus ou moins précises sur le génome, régions génomiques simples- versus multi- copies,…).

## 2. MATÉRIEL ET MÉTHODES

### 2.1. Un panel de modèles biologiques

Le riz (*Oryza sativa*) est une plante autogame. Il est établi que deux formes domestiquées majeures (*indica* et *japonica*) structurent fortement la diversité génétique de cette espèce. L'échantillon utilisé ici est une core collection de 217 accessions, analysée avec 15 marqueurs microsatellites répartis sur 12 chromosomes. Elle est représentative de la diversité de l'espèce *O. sativa* et de ses 7 groupes enzymatiques (0 à 6) [3]. Cette collection comprend notamment 83 variétés de type *indica* (groupe 1) et 67 variétés de type *japonica* (groupe 6).

Les sorghos cultivés (*Sorghum bicolor* ssp. *bicolor*) ont une large répartition en Afrique et Asie et présentent un large panel de variation phénotypique.

Une core collection de 205 accessions représentatives de la diversité de l'espèce cultivée, a été établie sur la base de critères raciaux, géographiques et moléculaires. Elle a été analysée au moyen de 74 sondes RFLP réparties sur le génome. La différenciation est marquée entre les sorghos africains cultivés au nord et au sud de l'équateur [4].

L'effet de la domestication d'une plante autogame est abordé chez le blé en étudiant le passage de la forme sauvage (*Triticum dicoccoides)* à la forme cultivée de blé dur. La diversité ayant été fortement réduite par un effet d'échantillonnage [5], ce type d'étude transversale permet de documenter l'impact d'un goulot d'étranglement sur le déséquilibre de liaison à très faible distance. L'échantillonnage est tiré d'une core collection représentant l'espèce dans son ensemble (48 accessions dont 28 appartenant à la forme sauvage).

L'hévéa (*Hevea brasiliensis*) est une espèce pérenne allogame, originaire d'Amazonie mais domestiquée très récemment, à la fin du XIX$^{e}$ siècle, en Asie. Pour cette espèce, l'objectif était de développer des marqueurs fortement liés pour l'analyse du DL dans des zones cibles du génome. Par rapport aux autres espèces, les ressources génomiques sont peu développées et une stratégie de développement de marqueurs est proposée.

*Medicago truncatula* est une espèce diploïde et autogame, aujourd'hui considérée comme l'espèce modèle principale pour la génétique et la génomique des Légumineuses. Dans ce contexte, le séquençage systématique des régions riches en gènes de son génome a été engagé. L'analyse de la diversité naturelle au sein de cette espèce a permis de repérer des groupes génétiques, mais l'histoire évolutive de l'espèce reste difficile à cerner (Ronfort *et al.* comm. pers.). L'analyse du déséquilibre de liaison au sein de cette espèce devrait donc nous permettre de préciser l'histoire évolutive de *M. truncatula* [6] - par exemple, en révélant des situations de mélanges de populations (« admixture ») - et de mieux comprendre l'impact de la structure génétique d'un échantillon sur l'étendue du DL.

Pour appréhender l'effet de la structure génétique, les échantillons ont été basés sur des connaissances établies sur la différenciation génétique des espèces concernées. Ces structurations avaient été établies par des études préalables basées sur des études de polymorphisme de marqueurs moléculaires. Pour le riz, cette structuration portait sur les formes *indica* et *japonica*, pour le sorgho sur les formes Nord et Sud et sur la séparation en races. Sur *M. truncatula*, deux échantillons de 30 individus ont été définis afin de représenter deux niveaux de structure : un échantillon représentant l'ensemble de l'espèce et donc supposé structuré et un échantillon représentant un des groupes génétiques détectés à l'aide de marqueurs neutres et supposé plus homogène.

### 2.2. Différentes échelles d'analyse

Le déséquilibre de liaison a été calculé sur des distances assez larges, de plusieurs centimorgans (cM) ou bien dans des régions physiques très courtes de l'ordre d'une centaine de kb. Pour le riz, 3 segments de 600 kb à 1 100 kb sur le bras court du chromosome 6 ont été choisis autour de trois gènes d'intérêt (*waxy* contrôlant la teneur en amylose du grain; *hd1* et *hd3a*, deux QTL récemment clonés jouant un rôle dans la durée de cycle).

Pour le sorgho, deux régions ont été étudiées : la première de 5 cM, située à l'extrémité du bras court du chromosome 4, orthologue de la région portant un gène majeur de résistance à la rouille chez la canne à sucre [7]. Dans cette région, deux contigs de BACs distants d'environ 200 kb sont disponibles : le premier couvre une région de 750 kb et le second 250 kb. Douze marqueurs RFLP et 5 marqueurs microsatellites y sont déjà placés et leur position physique déterminée. La seconde zone correspond à la région contenant le gène *waxy*, elle est couverte par un BAC de 130 kb.

Sur le blé et sur *M. truncatula*, l'objectif était de suivre la décroissance du DL dans une zone physique très réduite. Sur le blé, la région génomique choisie est située en position distale du bras court du chromosome 5 du génome A. Il s'agit d'une zone impliquée dans la dureté du grain et sur laquelle se focalisent de nombreux efforts de génomique comparative [8]. Pour *M. truncatula*, la zone étudiée est un contig de BACs de 120 kb, portant le gène *NorK* (Nodulation Receptor Kinase, [9]) impliqué dans la mise en place de la symbiose plante/*Rhizobium*. Quatre fragments génomiques (dont le gène *NorK*) de taille variable (entre 600 et 1 500 pb) et balisant ce contig de BACs ont été définis et séquencés sur un ensemble de 60 individus.

Pour l'hévéa, 2 zones du génome ont été ciblées prioritairement, l'une localisée autour d'un gène majeur de résistance à une maladie foliaire [10], l'autre correspondant à un cluster d'homologues de gènes de résistance (Rga, [11]). La cartographie du génome de l'hévéa est en cours, au laboratoire, à l'aide de microsatellites issus de banques enrichies et le développement ciblé de microsatellites, devrait fournir des paires de marqueurs liés à des distances allant de quelques dizaines de kb à plusieurs dizaines de cM, pour l'analyse du DL.

### 2.3. Développement de marqueurs

La densification des régions étudiées en marqueurs génétiques a été faite selon différentes stratégies et selon les informations génomiques disponibles. Pour certaines études, les marqueurs microsatellites ont été retenus pour leur polymorphisme et leur faible coût de développement. Sur le riz, la

stratégie a consisté à utiliser la séquence du chromosome 6 disponible pour définir *in silico* des marqueurs microsatellites autour des 3 gènes cibles.

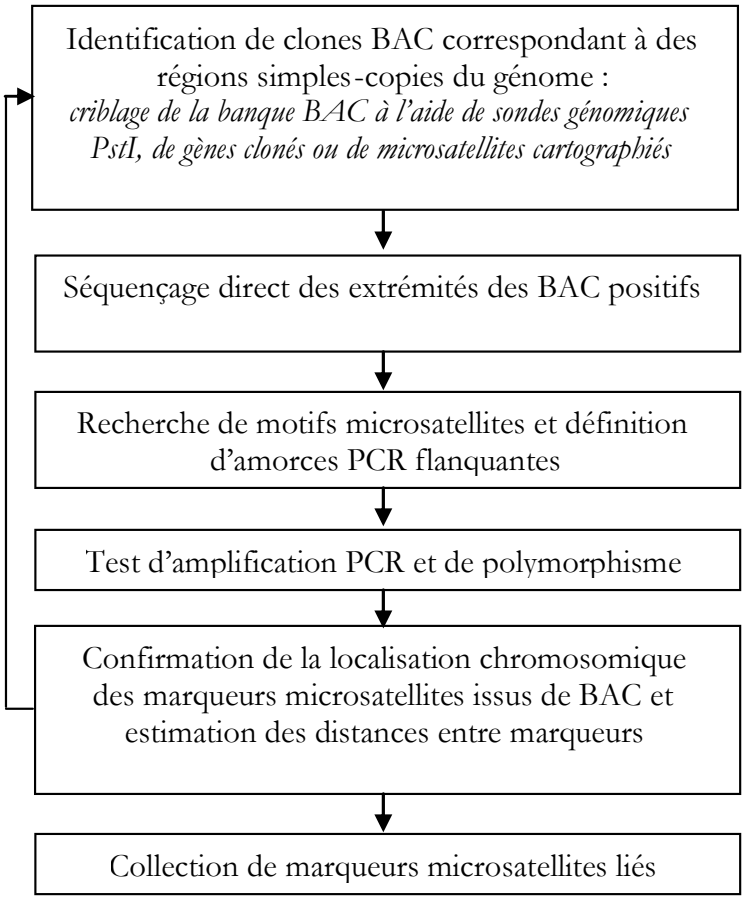



**Figure 1 :** Stratégie d'identification ciblée de marqueurs microsatellites (SSR) à partir de ressources BAC.

Sur le sorgho, l'approche a été de développer des banques enrichies en motifs microsatellites à partir de BACs situés dans les contigs d'intérêts ou d'utiliser des séquences d'extrémités de BAC. Sur l'hévea, une méthode d'identification ciblée de marqueurs microsatellites à partir de ressources BAC a été développée. La stratégie proposée, illustrée figure 1, repose sur les densités en microsatellites observées chez les plantes dont le génome a été séquencé. D'une part, cette densité apparaît plus élevée sur l'ensemble du génome que ce qui avait été rapporté auparavant et, d'autre part, les microsatellites apparaissent préférentiellement localisés dans la fraction simple copie du génome, *i.e.* la fraction contenant l'ADN non répété [12], [13].

Pour tester cette stratégie sur l'hévéa, nous avons sélectionné d'une part des clones BAC par hybridation sur la banque BAC hévéa, disponible au laboratoire, avec 22 sondes cartographiées correspondant à des séquences non répétées [11] : 16 génomiques *Pst*I, 1 ADNc et 5 homologues de gènes de résistance (Rga) et, d'autre part un échantillon de 37 clones BAC pris au hasard. Les inserts des clones BAC sélectionnés ont été extraits suivant un protocole standard et séquencés à partir des deux extrémités (sous-traitance Genome Express, France). Les motifs microsatellites retenus sont des répétitions de 1 à 6 bases, parfaites ou imparfaites, d'au moins 12 bases de longueur [13]. Les amorces PCR encadrant les microsatellites ont été testées pour leur aptitude à révéler du polymorphisme sur 4 accessions d'hévéa, utilisées comme géniteurs des 3 descendances et la localisation chromosomique des microsatellites polymorphes a été contrôlée (vrais positifs) par cartographie sur ces descendances.

Pour le blé et *M. truncatula*, des amorces ont été développées sur des séquences génomiques pour permettre un séquençage direct des produits d'amplification et obtenir le polymorphisme de séquence sur des fragments de 600 à 2 000 pb. En raison de la polyploïdie du blé, des amorces spécifiques du génome À ont du être définies en alignant des séquences d'EST (Expressed Sequence Tag) (http://www.ncbi.nlm.nih.gov/), et des séquences de BACs homéologues de différents génomes (Chantret comm. pers.). Pour *M. truncatula*, les fragments à séquencer ont été choisis dans des gènes et de façon à amplifier principalement des introns. Pour ces deux espèces, les conditions de PCR ont été optimisées pour chaque paire d'amorces afin d'obtenir les amplifications spécifiques recherchées. Les séquences obtenues ont été alignées avec le logiciel Staden Package 2003.0.1 (ftp://ftp.mrc-lmb.cam.ac.uk/pub/staden). Les indices de polymorphismes nucléotidiques et les D de Tajima ont été calculés avec le logiciel DnaSP v4.0 (http://www.ub.es/dnasp).

### 2.4. La mesure du déséquilibre de liaison

Deux mesures de DL multiallélique, $r^2$ et $D'$, ont été calculées avec le logiciel Tassel v1.9.0 (http://www.maizegenetics.net/index.php?page=bioinformatics/tassel/index.html) pour chaque paire de locus. Les représentations du $r^2$ et $D'$ sont données en fonction des distances physiques entre paires de locus. La signification des mesures de DL a été évaluée en utilisant le test de permutations (n = 1 000). Les statistiques ont été adaptées en fonction des marqueurs utilisés. Notamment chez le riz, un regroupement d'allèles a du être effectué pour les microsatellites pour réduire le nombre d'allèles rares. Sous hypothèse d'un modèle mutationnel pas à pas (Stepwise Mutation Model), des allèles de taille voisine peuvent résulter de mutations récentes à partir du même allèle ancestral et ont été regroupés lorsque les

distributions de taille d'allèles montraient des discontinuités nettes. Les allèles ayant une fréquence inférieure à 5 % ont été remplacés par des données manquantes. Pour l'analyse du polymorphisme de séquence, les singletons ont été éliminés.

## 3. RÉSULTATS

### 3.1. Développement de marqueurs

Sur le riz, 68 marqueurs microsatellites ont été définis et ont conduit à identifier 47 microsatellites polymorphes qui ont ensuite été génotypés sur la core collection.

La stratégie menée chez le sorgho a donné des résultats contrastés. Dans la première zone, l'hybridation d'oligonucléotides (CT)15 et (GT)15 sur les extrémités de 30 BACs choisis dans le grand contig de 750 kb et de 14 BACs dans le contig de 250 kb n'a pas été couronnée de succès, de même que le séquençage des extrémités. Au final aucun microsatellite nouveau n'a pu être obtenu dans cette zone alors que la stratégie de construction de banques enrichies spécifiques y avait permis le développement de 5 marqueurs microsatellites. Pour la seconde région, sept couples d'amorces permettant l'amplification de marqueurs microsatellites ont été définis d'après la séquence du BAC disponible et les données de génotypage ont été acquises pour 5 d'entre eux.

**Tableau I :** Identification de microsatellites chez l'hévéa : nombre de séquences analysées et densité en microsatellites pour des BAC sélectionnés au hasard et des BAC sélectionnés par hybridation avec 22 séquences non répétées.

| | séquençage | | | microsatellites identifiés* | | |
|---|---|---|---|---|---|---|
| sélection des BAC | Clones BAC | extrémités de BAC | longueur séquencée (kb) | extrémités de BAC avec ≥ 1 microsat. | séquences microsat. | distance moyenne (kb) |
| hasard | 37 | 55 | 30,5 | 3 | 5 | 6,1 |
| hybridation | 87 | 141 | 91,6 | 38 | 56 | 1,6 |

* tous motifs mono-, di-, tri-, tetra- ou penta-nucléotidiques, parfait ou imparfaits, d'au moins 12 pb de longueur.

Sur l'hévéa, les résultats d'identification de motifs microsatellites, dans les extrémités de 37 BAC sélectionnés au hasard et de 87 BAC localisés dans la zone d'intérêt par hybridation avec les 22 sondes, sont donnés dans le tableau I. La densité pour les BAC sélectionnés au hasard (distance moyenne de 6,1 kb entre 2 microsatellites) est comparable à la densité moyenne sur l'ensemble du génome observée chez différentes espèces de plantes (6-7 kb, [12]). Conformément à notre hypothèse, la densité apparaît 4 fois plus éle-

vée dans les BAC sélectionnés par hybridation et correspondant à la fraction non répétée du génome de l'hévéa (1,6 kb). Ceci a permis d'identifier des séquences microsatellites pour 21 des 22 sondes testées, malgré le petit nombre de clones BAC positifs par sonde et la faible longueur des séquences analysées (650 bases en moyenne).

La fréquence relative des différents motifs microsatellites – mono- à penta-nucléotidiques - identifiés hévéa (données non montrées) est comparable à ce qui a été rapporté pour les dicotylédones dont le génome a été largement séquencé [12], [13]. À partir des 56 séquences microsatellites identifiées, 25 ont donné des marqueurs polymorphes. Pour 11 de ces marqueurs, la localisation chromosomique a été confirmée par cartographie génétique, la cartographie des 14 autres est en cours. Au final, des microsatellites polymorphes ont été obtenus pour 18 des 22 locus ciblés, confirmant la faisabilité de la stratégie proposée.

Pour le blé, quatre couples d'amorces gène- et génome-spécifiques ont été retenus pour le séquençage des fragments amplifiées sur l'ensemble de l'échantillon. Pour *M. truncatula*, quatre couples d'amorces gène-spécifiques ont été retenus et utilisés sur les deux échantillons.

### 3.2. Influence de la nature des marqueurs sur le niveau de DL

Chez le sorgho, le DL trouvé entre marqueurs microsatellites, assez polymorphes (de 3 à 17 allèles par locus) s'est révélé très faible et non significatif, que ce soit au sein du BAC « *waxy* » ($r^2$ maximal = 0,017, d = 46 kb) ou sur les 4 cM portés par les 750 kb situés sur le chromosome 4 ($r^2$ maximal = 0,16, d = 200 kb). Sur cette zone, le maximum observé est un $r^2$ de 0,16 pour 2 microsatellites distants d'environ 200 kb. À contrario, le DL entre marqueurs RFLP est marqué, le $r^2$ atteint un maximum de 0,67 pour des locus distants de 17 kb et 12 couples de RFLP montrent un DL $\geq$ 0,3 (fig. 2). De même, certains marqueurs microsatellites montrent un DL fort avec des marqueurs RFLP sur des distances variant de 75 à 567 kb.

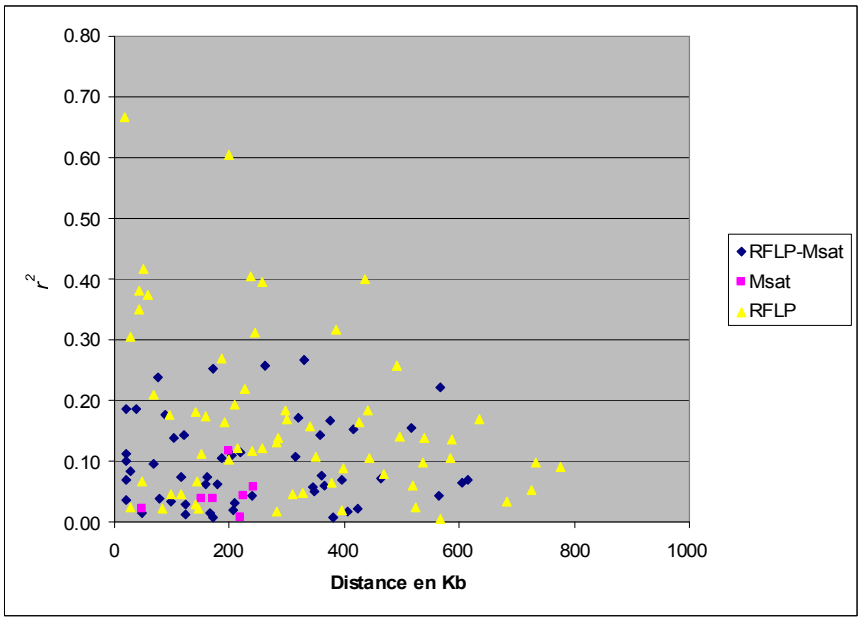
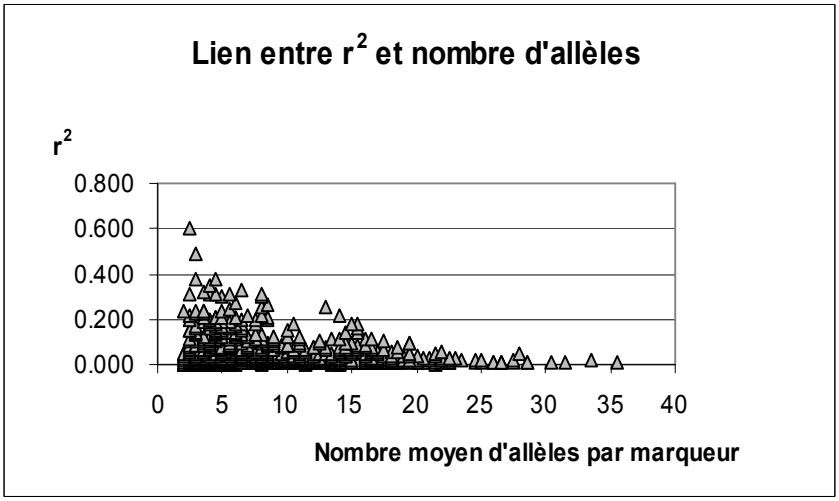



**Figure 2 :** Évolution de $r^2$, dans la zone du chromosome 4, en fonction de la distance physique et du type de marqueurs, au sein de la core collection de sorgho. L'ensemble des valeurs significatives ($p < 0,05$) est représenté.

La question du traitement des locus comptant de nombreux allèles s'est posée de manière encore plus prégnante chez le riz. La diversité allélique des locus microsatellites y est considérable (10,3 allèles en moyenne). Sur les 444 allèles observés, 61 % ont une fréquence allélique inférieure ou égale à 5 %. Le nombre d'allèles est négativement corrélé au nombre de nucléotides du motif ($r = -0,38^*$) mais est surtout très fortement corrélé à son nombre de répétitions ($r = 0,72^{***}$), ce qui suggère des taux de mutation très différents entre locus. Ces nombreux allèles rares sont difficiles à gérer dans le calcul de certains indices de DL [2]. Ceci transparaît dans la forme des graphes représentant les indices de DL en fonction du nombre d'allèles, la valeur de D' ne diminuant plus au delà d'un certain nombre d'allèles contrairement à celle de $r^2$ (fig. 3). Les allèles ayant une fréquence inférieure à 5 % ont été remplacés par des données manquantes ce qui a conduit à éliminer les microsatellites présentant plus de 12 allèles. Au final, 30 locus ont été analysés pour les riz *indica* et 26 pour les riz *japonica*.



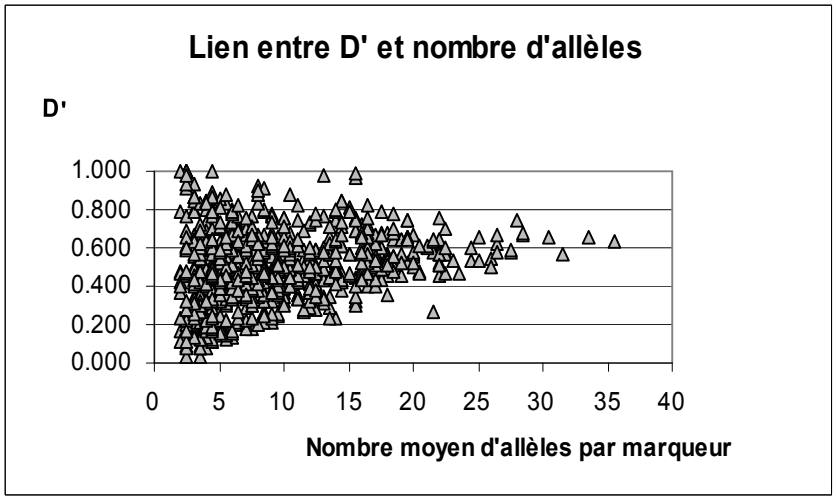




**Figure 3 :** Lien entre indice de DL et nombre moyen d'allèles par marqueur chez le riz.

### 3.3. Influence de la structure sur le DL

Chez le riz, les groupes *indica* et *japonica* apparaissent fortement différenciés pour les marqueurs microsatellites (Fst *indica/japonica* = 0,54, compris entre 0,20 et 0,84) et le DL décroît lentement avec la distance quand toutes les accessions sont prises en compte. Pour tester l'effet de la structure *indica/japonica* sur le DL, des analyses séparées portant sur le critère de $r^2$ ont été

conduites pour chacun des 3 segments (voir fig. 4 pour l'exemple de *hd3a)*. Sur le segment portant *hd3a*, $r^2$ est faible, même pour des sites proches et décroît plus lentement chez *indica* que chez *japonica*. Il atteint la valeur de 0,1 à 300 kb chez les indica pour seulement 150 kb chez les *japonica*. Autour de *waxy*, $r^2$ part de valeurs élevées mais décroît rapidement ($r^2$ = 0,1 pour d = 100 kb dans les deux groupes). Autour de *hd1*, le DL s'étend plus loin avec $r^2$ réduit à 0,1 après 350 kb dans les deux groupes.

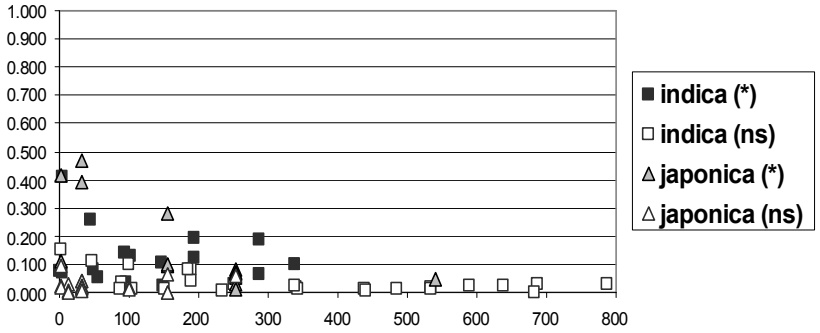



**Figure 4 :** Évolution de $r^2$ autour de *hd3a* en fonction de la distance physique entre locus chez le riz ; * : valeurs de $r^2$ significatives au seuil de 5 % ; ns : valeurs de $r^2$ non significatives.

Chez le sorgho, l'analyse Bayesienne conduite avec le logiciel Structure v2.1 [14] révèle une forte différenciation en 2 pôles géographiques (Fst = 0,29), correspondant aux variétés africaines originaires du nord et du sud de l'équateur, les variétés asiatiques se répartissant dans les 2 grands pôles (fig. 5). Le groupe nord, constitué de 129 variétés, présente une plus grande diversité génétique (H = 0,395 ; richesse allélique : 2,68), que le groupe sud incluant 51 variétés (H = 0,237 ; richesse allélique : 1,91). Pour étudier l'incidence de cette structuration sur le DL chez le sorgho, nous avons calculé le $r^2$ entre ces 60 marqueurs RFLP non liés. Le $r^2$ varie fortement entre 0 à 0,51 avec 95 % des valeurs inférieures à 0,18. Cette valeur a été prise comme seuil dans l'analyse des marqueurs liés. La projection des haplotypes obtenus avec les locus liés (12 RFLP) montre que, souvent, un haplotype est caractéristique d'un groupe génétique, défini par classification sur des marqueurs indépendants. Cela montre la très forte incidence de la structure sur le DL à courte distance.

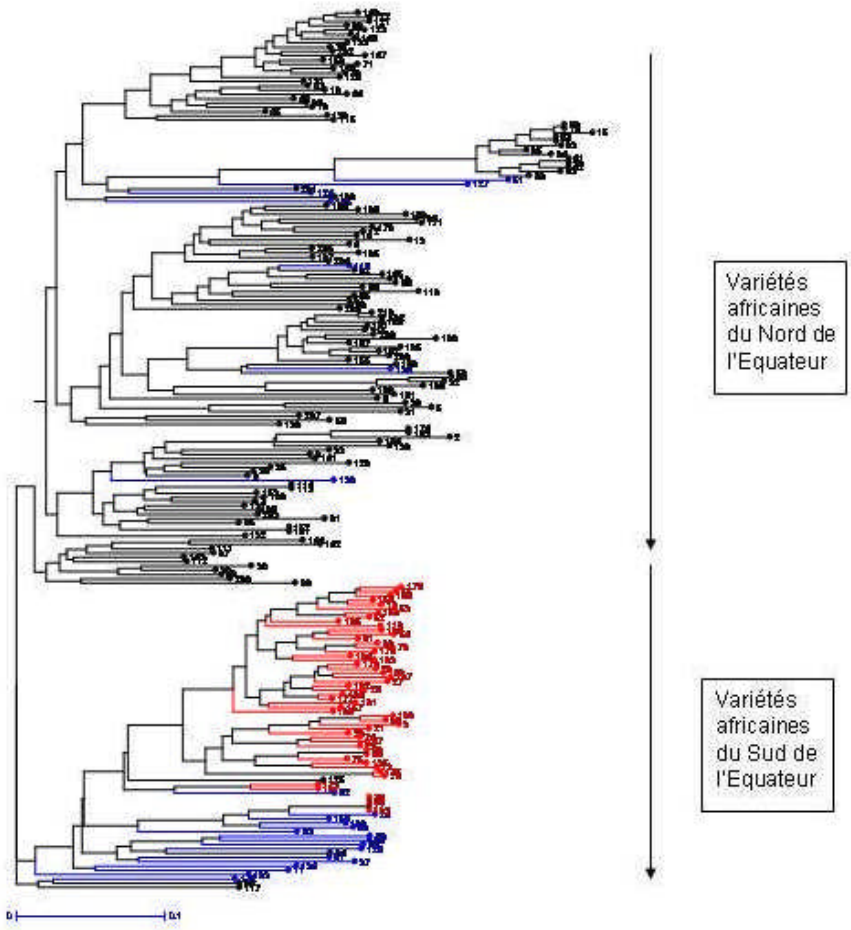



**Figure 5 :** Structuration de la core collection obtenue avec les 74 marqueurs RFLP répartis sur le génome (NJ tree, indice de dissimilarité de Dice). En noir, sont représentées les accessions attribuées au groupe 1, en rouge les accessions attribuées au groupe 2, en bleu les accessions « hybrides » entre les 2 groupes, l'attribution à un groupe résultant de l'analyse conduite avec Structure [14] sur 60 locus non liés (distance minimale de 10 cM).

Lorsqu'on restreint l'analyse au niveau intra-groupe, la situation change. Au sein du groupe Nord, un nombre plus réduit de marqueurs apparaît en DL, par comparaison avec l'analyse effectuée sur la core collection globale. En particulier, des DL existant entre marqueurs distants d'environ 50 et 200 kb ne sont pas conservés au sein de ce groupe. Mais, un DL élevé entre 2 marqueurs distants d'environ 400 kb a été observé. Le maintien du DL au sein de ce groupe sur d'assez longues distances peut encore être lié à une structuration au sein de ce groupe dans lequel 5 ou 6 sous-groupes ont été

identifiés lorsque l'analyse Bayesienne est conduite au sein du groupe Nord. Au sein du groupe Sud, seules 9 paires de locus sont significativement liées au seuil de 5 % et un fort DL est révélé ($r^2$ de 0,48) entre 2 locus RFLP distants d'environ 500 kb. Des valeurs de $r^2$ de 1 ont été observées entre plusieurs locus distants de 100 kb maximum, indiquant une histoire similaire de mutation (apparition de 2 mutations liées sur le même haplotype) sans recombinaison entre ces locus à l'intérieur de ce groupe et/ou un fort effet de goulot d'étranglement lors de la domestication [2]. Cette dernière hypothèse semble en accord avec la domestication supposée plus récente de ces sorghos africains du sud de l'Équateur à partir d'un pool génétique réduit [15], confirmé par une forte réduction de la richesse allélique comparée à la core collection.

Chez *M. truncatula* et *T. dicoccoides*, le polymorphisme de séquence observé à l'échelle de l'espèce est relativement élevé dans les segments étudiés ($\theta_S$~3,10$^{-3}$ pour *M. truncatula* et $\theta_S$~6,10$^{-3}$ pour *T. dicoccoides)*. Pour les deux espèces, les patrons de diversité observés pour les différents fragments étudiés ne montrent pas d'écarts significatifs au modèle neutre (équilibre mutation-dérive, [16]). Pour la plupart des fragments génomiques étudiés, les valeurs du *D* de Tajima apparaissent cependant négatives. Ce biais général vers les valeurs négatives, même s'il n'est pas significatif, atteste d'un (faible) excès d'allèles rares à l'échelle de l'espèce, ce qui peut être interprété comme la signature d'une expansion démographique, de pressions de sélection positive ou d'un fonctionnement global de type métapopulation, associé à de forts taux de migration entre dèmes [17].

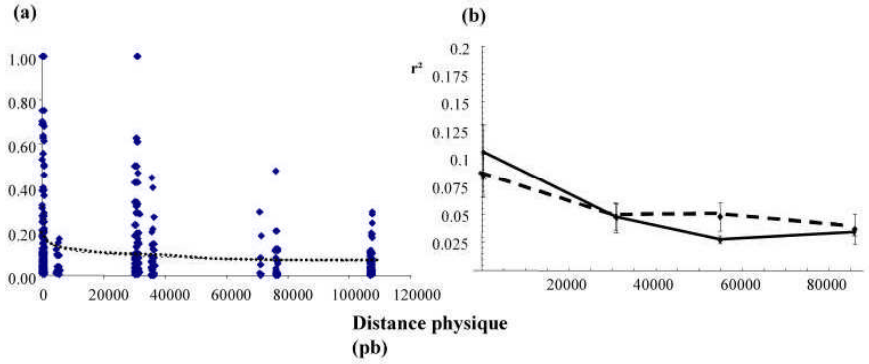



**Figure 6 :** Décroissance du déséquilibre de liaison en fonction de la distance physique dans deux régions génomiques chez le blé sauvage (a) et *M. truncatula* (b). En (a), chaque point représente la corrélation de fréquence ($r^2$) de SNP (Single Nucléotide Polymorphism) détectés dans un échantillon de 28 accessions séquencées pour quatre fragments de gènes situés dans le locus Ha. La ligne en pointillés représente la régression logarithmique entre $r^2$ et la distance physique. En (b), les courbes représentent les moyennes du $r^2$ estimées entre SNP détectés dans deux

échantillons de *M. truncatula* séquencés pour 4 fragments génomiques situés sur le BAC contenant le gène *NorK*. La courbe en plein représente les données obtenues sur l'échantillon global ($n$ = 30) potentiellement structuré ; la courbe en pointillé, les données obtenues sur un échantillon ($n$ = 30) plus homogène. Les barres d'erreur donnent l'écart type (divisé par 10).

Malgré le fort taux d'autogamie de ces deux espèces, le DL observé est faible ($r^2$<0,2), décroît rapidement sur 10 à 20 kb. Chez *M. truncatula,* il est peu dépendant du type d'échantillon considéré (fig. 6). Chez le blé, nous n'avons pu étudier l'effet de la domestication sur le déséquilibre de liaison car dans le segment considéré du chromosome 5A, le polymorphisme restant au sein de l'espèce cultivée s'est révélé, et de loin, nettement insuffisant : de 96 SNPs au départ chez *T. dicoccoides*, il ne reste que 9 SNP dont 6 singletons au sein de la forme cultivée (blé dur).

## 4. DISCUSSION

L'étude du DL dans une zone donnée nécessite d'y développer des marqueurs. Nous avons ici utilisé les possibilités offertes par les clones BAC, dont la séquence pouvait être complètement ou partiellement connue. Des microsatellites ont été développés sur l'hévéa à moindre coût et de manière ciblée, ce qui démontre la faisabilité de la démarche proposée. À contrario, l'utilisation de banques enrichies à partir de mélanges de BAC ciblés n'a pas donné de très bons résultats sur le sorgho. Pour le riz, la connaissance de la séquence génomique complète de la région était un atout évident dans la mise en place de nouveaux marqueurs. L'étude a notamment montré qu'utiliser la fraction simple copie – *i.e.* riche en gènes – du génome donne accès à une forte densité de microsatellites. La possibilité de cibler cette fraction non hautement répétée suggère que cette méthode devrait avoir la même efficacité pour toute espèce, quelle que soit la taille de son génome haploïde. L'application à l'hévéa, espèce au génome (2,1 pg/1C) environ 13 fois plus grand que celui d'*Arabidopsis* conforte cette hypothèse. La méthode proposée ici permet, de plus, l'identification de tous les motifs microsatellites présents dans le génome, en fonction de leur fréquence relative, et l'on pourra ainsi comparer le taux d'allélisme en fonction de la longueur du motif répété et l'incidence sur la mesure du DL. Pour l'hévéa, les marqueurs microsatellites développés dans cette étude, associés à ceux déjà disponibles permettent désormais de couvrir localement des distances allant de quelques bases à plusieurs dizaines de centi-Morgans. Le génotypage est en cours sur une collection de 450 accessions et permettra l'analyse du DL chez cette espèce.

Pour autant, les marqueurs microsatellites ne sont pas toujours simples à utiliser dans les études de DL. Leur taux de mutation extrêmement variable

et le grand nombre d'allèles rares généralement observé conduisent à éliminer les plus polymorphes (cas du riz dans notre étude) et à regrouper des allèles par taille. D'autres méthodes statistiques de regroupement d'allèles sont en cours d'exploration (C. Billot comm. pers.) et seront évaluées pour leur incidence sur la mesure du DL.

Dans nos études, le DL est clairement apparu dépendant de la structure génétique de l'échantillon. Cette dernière génère des associations statistiques indépendamment de la liaison physique qu'il faut donc pouvoir contrôler en génétique d'association. Le problème est de connaître l'influence de la structure non repérée. Par exemple, la tendance du DL à être plus élevé chez le groupe des riz *japonica* que chez les *indica* pourrait être due à une structure résiduelle dans le groupe *japonica* qui est constitué d'un mélange d'accessions tempérées et tropicales.

Par contre, l'autogamie ne semble pas pouvoir à elle seule expliquer le maintien d'un DL sur une longue distance physique lorsque l'état d'équilibre est proche. Chez les deux espèces représentatives de cette situation, *M. truncatula* et *T. dicoccoides*, le DL entre SNP décroît très rapidement et devient très faible au delà d'une vingtaine de kilobases. Des résultats similaires ont été obtenus chez *Arabidopsis thaliana*. Pour cette dernière, il a par ailleurs été démontré que l'autogamie est une acquisition récente de l'espèce [18]. Pour le blé et *M. truncatula*, le faible DL observé pourrait également signifier que la taille efficace de l'espèce est élevée. Concernant la faisabilité des études d'association dans ces espèces, nos résultats suggèrent qu'un maillage de l'ordre de 1 marqueur tous les 10 ou 20 kb pourrait permettre de réaliser des études d'association à une très forte résolution quasiment quel que soit le type d'échantillon. Nos données sont cependant limitées à une très faible portion du génome et ne peuvent pas, pour l'instant, être généralisées. Malgré tout, dans l'échantillon plus local de *M. truncatula*, moins susceptible d'être structuré, le DL entre sites proches est moins fort que dans l'échantillon global et sa décroissance moins rapide. Par ailleurs, dans le cas du blé, les études d'association seraient peut-être plus efficace chez l'ancêtre sauvage (polymorphisme et faible DL) que dans la forme cultivée.

L'observation d'un DL à longue distance chez les plantes domestiquées autogames, comme ici sur le riz, doit donc s'interpréter sur la base d'évènements démographiques plus ou moins récentes et de structurations marquées à l'échelle géographique. D'autres études conduites sur le riz ont trouvé un DL disparaissant après 100 kb dans une population [19], alors que le DL ne décroît pas avec la distance dans une population d' *Oryza glaberrima* trouvée très structurée [20]. La situation autour de nos trois gènes est proche de celle observée autour de *xa5*. Du fait de leur importance agronomique, *waxy*, *hd1*, et, dans une moindre mesure, *hd3,* représentent des gènes qui ont pu subir l'effet de la sélection et au voisinage desquels la sélection pour-

rait avoir créé du DL. La distance à laquelle le DL peut être considéré comme relativement faible correspond à 1 à 2 cM chez le riz. Cela permet une meilleure résolution que la plupart des analyses de QTL et permet d'envisager l'utilisation d'études d'association pour réduire le nombre de gènes candidats sous-jacents à un QTL. Chez le sorgho, le DL s'étend sur une zone assez longue d'environ 400-500 Kb mais il est possible que cette situation reflète une structuration des formes cultivées, non prise en compte dans les groupes géographiques Nord/Sud. La race est par exemple un facteur important dans la structuration de la diversité génétique des sorghos cultivés et elle n'a pas encore été prise en compte en particulier au sein du groupe 1, comportant des groupes raciaux relativement différenciés.

En conclusion, les études d'association paraissent tout à fait réalisables dans la plupart des collections/échantillons étudiés ici ; ceci à condition de se donner les moyens de décrire la structure génétique sous-jacente et de pouvoir développer un jeu de marqueurs adaptés. De nouveaux types de marqueurs SNP haut débit ou les DArT [21] pourraient permettre de lever les problèmes liés à l'utilisation des marqueurs microsatellites.

## RÉFÉRENCES


[1] Remington D.L., Thornsberry J.M., Matsuoka Y., Wilson L.M., Whitt S.R., Doebley J., Kresovich S., Goodman M.M., Buckler E.S., Structure of linkage disequilibrium and phenotypic associations in the maize genome, Proc. Natl. Acad. Sci. USA 98 (2001) 11479-11484.

[2] Flint-Garcia S.A., Thornsberry J.M., Buckler E., Structure of linkage disequilibrium in plants, Annu. Rev. Plant Biol. 54 (2003) 357-374.

[3] Glaszmann J.C., Isozymes and classification of Asian rice varieties, Theor. Appl. Genet. 74 (1987) 21-30.

[4] Deu M., Rattunde F., Chantereau J., A global view of genetic diversity in cultivated sorghums using a core collection, Genome 49 (2006) 168-180.

[5] Thuillet A.C., Bataillon T., Poirier S., Santoni S., David J.L., Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data, Genetics 169 (2005) 1589-1599.

[6] Nordborg M., Tavare S., Linkage disequilibrium: what history has to tell us, Trends Genet. 18 (2002) 83-90.

[7] Asnaghi C., Paulet F., Kaye C., Grivet L., Deu M., Glaszmann J.C., Hont A.D., Application of synteny across Poaceae to determine the map location of a sugarcane rust resistance gene, Theor. Appl. Genet. 101 (2000) 962-969.

[8] Chantret N., Salse J., Sabot F., Rahman S., Bellec A., Laubin B., Dubois I., Dossat C., Sourdille P., Joudrier P., Gautier M.-F., Cattolico L., Beckert M., Aubourg S., Weissenbach J., Caboche M., Bernard M., Leroy P., Chalhoub B., Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*), Plant Cell 17 (2005) 1033-1045.

[9] de Mita S., Santoni S., Hochu I., Ronfort J., Bataillon T., Molecular evolution and positive selection of the symbiotic gene *NORK* in *Medicago truncatula*, J. Mol. Biol. 62 (2006) 234-244.

[10] Le Guen V., Lespinasse D., Oliver G., Rodier-Goud M., Pinard F., Seguin M., Molecular mapping of genes conferring field resistance to South American Leaf Blight (*Microcyclus ulei*) in rubber tree., Theor. Appl. Genet. 108 (2003) 160-167.

[11] Lespinasse D., Rodier-Goud M., Grivet L., Leconte A., Legnate H., Seguin M., A saturated genetic linkage map of rubber tree (*Hevea* spp.) based on RFLP, AFLP, microsatellite, and isozyme markers, Theor. Appl. Genet. 100 (2000) 127-138.

[12] Cardle L., Ramsay L., Milbourne D., Macaulay M., Marshall D., Waugh R., Computational and experimental characterization of physically clustered simple sequence repeats in plants, Genetics 156 (2000) 847-854.

[13] Morgante M., Hanafey M., Powell W., Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes, Nature Genet. 30 (2002) 194-200.

[14] Pritchard J.K., Stephens M., Donnelly P., Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.

[15] Doggett H., Sorghum, 2nd edition, Longman Scientific and Technical, New York, 1988. p.

[16] Tajima F., Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics 123 (1989) 585-595.

[17] Wakeley J., Aliacar N., Gene genealogies in a metapopulation, Genetics 159 (2001) 893-905.

[18] Charlesworth D., Vekemans X., How and when did *Arabidopsis thaliana* become highly self-fertilising, BioEssays 27 (2005) 472-476.

[19] Garris A.J., McCouch S.R., Kresovich S., Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.), Genetics 165 (2003) 759-769.

[20] Semon M., Nielsen R., Jones M.P., McCouch S.R., The population structure of african cultivated rice *Oryza glaberrima* (Steud.): Evidence for elevated levels of linkage disequilibrium caused by admixture with *O. sativa* and ecological adaptation, Genetics 169 (2005) 1639-1647.

[21] Jaccoud D., Peng K., Feinstein D., Kilian A., Diversity Arrays: a solid state technology for sequence information independent genotyping, Nucl. Acids Res. 29 (2001) e25-.