

MINISTÈRE DE L'AGRICULTURE
ECOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

THÈSE

Pour obtenir le grade de

DOCTEUR

de

L' ECOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

Spécialité : Sciences agronomiques (Amélioration des Plantes)

Formation doctorale : Ressources phytogénétiques et interactions biologiques

Ecole doctorale : Biologie intégrative

Préparée à BIOTROP/CIRAD – UMR PIA (1096)

Présentée
par
Tatiana PUGH MORENO

**ETUDE DES DÉSÉQUILIBRES DE LIAISON DANS UNE COLLECTION DE
CACAOYERS (*THEOBROMA CACAO* L.) APPARTENANT AU GROUPE CRIOLLO/
TRINITARIO ET APPLICATION AU MARQUAGE GÉNÉTIQUE DES CARACTÈRES
D'INTÉRET**

DIRECTEUR DE THÈSE : C. LANAUD

Soutenue le 11 janvier 2005 devant le jury composé de :

A. CHARRIER

Professeur, AGRO-M, président du jury

D. PRAT

Professeur, Université Claude Bernard Lyon -1, Rapporteur

A. CHARCOSSET

Directeur de Recherche, INRA, Rapporteur

P. THIS

Chargé de recherches, INRA, Examinateur

C. LANAUD

Directeur de Recherche, CIRAD

MINISTÈRE DE L'AGRICULTURE
ECOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

THÈSE

Pour obtenir le grade de

DOCTEUR

de

L' ECOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

Spécialité : Sciences agronomiques (Amélioration des Plantes)

Formation doctorale : Ressources phytogénétiques et interactions biologiques

Ecole doctorale : Biologie intégrative

Préparée à BIOTROP/CIRAD – UMR PIA (1096)

Présentée

par

Tatiana PUGH MORENO

ETUDE DES DÉSÉQUILIBRES DE LIAISON DANS UNE COLLECTION DE
CACAOYERS (*THEOBROMA CACAO* L.) APPARTENANT AU GROUPE CRIOLLO/
TRINITARIO ET APPLICATION AU MARQUAGE GÉNÉTIQUE DES CARACTÈRES
D'INTÉRÊT

DIRECTEUR DE THÈSE : C. LANAUD

Soutenue le 11 janvier 2005 devant le jury composé de :

A. CHARRIER

Professeur, AGRO-M, président du jury

D. PRAT

Professeur, Université Claude Bernard Lyon -1, Rapporteur

A. CHARCOSSET

Directeur de Recherche, INRA, Rapporteur

P. THIS

Chargé de recherches, INRA, Examinateur

C. LANAUD

Directeur de Recherche, CIRAD

Remerciements

"Seigneur, je t'apporte un breuvage qui est bon et qui enivre celui qui le boit; il t'attendrira le coeur, te guérira et te fera connaître la route de ton prochain voyage au pays où tu retrouveras la jeunesse".

Quatre ans se sont écoulés et je mentirais si je disais que cela n'a rien été. Un long séjour sur ces terres si lointaines mérite de sincères remerciements à tous ceux qui m'ont apporté de l'aide, des connaissances, de la compréhension et en particulier de la tendresse. Conformément à la tradition, ma directrice de thèse et les membres du jury auront les premiers remerciements : Claire Lanaud, avec qui j'ai eu le plaisir de travailler et qui m'a apporté une petite partie de son expérience et de sa passion cacaotière. Sous sa direction, j'ai beaucoup travaillé avec grande autonomie : je l'en remercie profondément. Malgré son emploi du temps très chargé, elle s'est toujours montrée très disponible à mon égard. Qu'il s'agisse de mes rapporteurs, Daniel Prat et Alain Charcosset ou de mes examinateurs, André Charrier et Patrice This, je leur suis extrêmement reconnaissante.

J'exprime ma gratitude à Jean-Christophe Glaszmann, Directeur de l'UMR 1096 pour l'accueil de cette thèse au sein de son Unité de recherche. Je remercie aussi le Genopole pour la prise en charge du séquençage et le FIRC (Fond interprofessionnelle de la recherche sur le cacaoyer) pour leur contribution financière à ce travail.

Plusieurs personnes m'ont accompagné lors des réflexions. Je voudrais vivement remercier en particulier Brigitte Courtois pour son aide précieuse tout au long des analyses statistiques et de la phase de rédaction. Ce manuscrit n'aurait probablement jamais vu le jour sans son intervention. Je voudrais également remercier Ange-Marie Risterucci et Jean-Louis Noyer qui m'ont aussi accordé confiance et considération. Leurs conseils m'ont beaucoup apporté. Brigitte Gouesnard et Agnès Doliges pour leur aide précieuse. Mes remerciements s'adressent aussi à tous ceux qui ont contracté la « maladie » du déséquilibre de liaison et avec qui de longues discussions de couloirs, et de bureaux ont été plus qu'intéressantes : merci donc à Monique Deu, Claire Billot, Marc Seguin et Xavier Perrier.

Je voudrais également remercier Marguerite Rodier-Goud, Monique Costes, Corinne Poitout, Catherine Guillaume, Laurence Allemano, Zina Diaf, Marie Françoise qui m'ont beaucoup aidé avec leur amitié et leur sympathie.

Quelqu'un de très particulier, notre grand homme des caverne et mon ami, Olivier Fouet, avec qui j'ai partagé des réflexions, de bon moments et une très grande partie de ce parcours.

Je ne peux pas oublier la bande de copains : Alberto, José, Rogers, Billot, Loïc, Donaldo, Lucio, Daniel. J'ai eu le bonheur d'être entourée de gens exceptionnels. Merci à tous !

"Las cosas claras y el chocolate espeso "
(Ideas should be clear and chocolate thick.)

En esta lista que no es enumerativa, evidentemente que no puedo dejar de agradecer a la Universidad Central de Venezuela y a su Consejo de Desarrollo Científico y Humanístico. A la Facultad de Agronomía, mis colegas del Instituto de Genética, quienes han asumido las cargas de mi ausencia, en especial a Catalina Ramis; a mis amigos y compañeros.

En este camino que solemos recorrer en compañía, Gladys tendría motivos para estar contenta. A mi hermano David a quien le debo esta ausencia. Héctor, por las palabras que has dicho y las que faltaron por decir, por tu presencia y ayuda.

A Valentina, con quien he crecido y que me ha enseñado las cosas esenciales de la vida. Merci pour ta patience et pour ton amour. Ce doctorat est également l'aboutissement de tes efforts.

Sommaire

Avant-propos et remerciements

INTRODUCTION	1
--------------------	---

CHAPITRE I - Eléments de bibliographie : le cacaoyer, la cartographie génétique et le déséquilibre de liaison	4
--	---

1. Le cacaoyer	4
1.1 L'histoire du cacaoyer	4
1.2 L'économie du cacaoyer	5
1.3 Taxonomie et biologie du <i>Theobroma cacao</i> L.	7
1.4 Classification des variétés cultivées de cacaoyer	9
1.5 Diversité génétique des cacaoyers	11
1.6 Les collections des ressources génétiques	14
2. La cartographie génétique	15
2.1 Les marqueurs moléculaires	15
2.2 La cartographie génétique	17
2.2.1 Ségrégation des marqueurs	18
2.2.2 Estimation de la liaison entre marqueurs et ordonnancement des marqueurs	19
2.2.3 Estimation des distances entre les marqueurs	20
2.3 Cartographie génétique chez le cacaoyer	21
3. Déséquilibre de liaison et « association mapping »	22
3.1 Définition et mesures du DL	23
3.2 Mesure du déséquilibre de liaison à partir des données génotypiques	25
3.3 Evolution du Déséquilibre de liaison	28
3.4 Effet de la structure de la population sur l'estimation du DL	32
3.5 Le déséquilibre de liaison chez les plantes	33

CHAPITRE II - Cartographie génétique du cacaoyer : développement et cartographie de 201 nouveaux marqueurs microsatellite	37
--	----

A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers	38
Conclusions et Perspectives	49

CHAPITRE III - Diversité génétique des clones appartenant au groupe Criollo moderne/Trinitario de la collection du CATIE	50
---	----

Genetic diversity of Modern Criollo/Trinitario cacao clones (<i>Theobroma cacao</i> L.) from CATIE germplasm collection assessed by microsatellite markers	52
Conclusions et Perspectives	64

CHAPITRE IV - Études des associations entre marqueurs moléculaires et caractères morphologiques dans la collection de variétés de Criollo moderne/Trinitario du CATIE	66
Molecular marker-trait associations in a Criollo/Trinitario cacao (<i>Theobroma cacao</i> L.) germplasm collection	68
Conclusions et Perspectives	87
CHAPITRE V - Discussion générale et perspectives	88
Développement d'une carte génétique saturée en marqueurs SSR	89
Etude d'une collection de Criollo modernes/Trinitario	89
Etude des associations marqueurs/caractères	91
Références bibliographiques	94

CHAPITRE I : ETUDE BIBLIOGRAPHIQUE

CHAPITRE I

Eléments de bibliographie : le cacaoyer, la cartographie génétique et le déséquilibre de liaison

1. Le cacaoyer

1.1 L'histoire du cacaoyer

La culture du cacaoyer en Amérique centrale était déjà développée bien avant l'arrivée des espagnols (environ 400 av. J.-C.) (Paradis, 1979). Les peuples de Méso-Amérique, les Olmèques, les Mayas puis les Aztèques, furent les premiers à cultiver le cacaoyer. Ce sont les mots aztèques « cacahualte » et « xocoatl » qui ont donné naissance aux termes actuels de « cacao » et « chocolat ». En fait, les peuples préhispaniques attribuèrent une origine divine au cacaoyer. Ainsi, les Mayas et les Aztèques les offraient à leurs dieux, à leurs morts ou comme dot à leurs fiancés et s'en servaient comme médicament contre les morsures de serpents, pour soigner divers maux, pour la confection de certains cosmétiques et pour augmenter le tonus sexuel. Le cacao avait également une fonction importante d'échange, et ses fèves servaient de monnaie. Les unités de mesure se référaient à un nombre de fèves dont la main (5 fèves), le *zontle* (400 fèves), le *xiquipil* (8000 fèves) et la *corga* (24000 fèves) furent utilisés jusqu'au XVIII^e siècle (Torquemada, 1723). L'importance économique du cacao dans les sociétés précolombiennes est facilement comprit dans la phrase « tout pouvait s'acheter avec des fèves de cacao » (Oviedo, 1944).

Une boisson reconstituante était préparée à partir de ses fèves grillées et concassées sur une pierre incurvée (le « metate »). La pâte obtenue était délayée et mêlée au maïs, au piment et à d'autres épices comme la vanille et l'achiote (*Bixa orellana*). Cette boisson était considérée comme le « breuvage des Dieux ». De là viendrait le nom donné au genre *Theobroma* qui signifie en grec « nourriture des dieux ». Il est fortement probable que cette boisson, à haute valeur symbolique et religieuse, était réservée aux classes privilégiées. « Personne n'en buvait s'il n'était pas cacique, grand seigneur ou vaillant guerrier » (Garcia de Palacio, 1576) (Figures 1, 2 et 3).

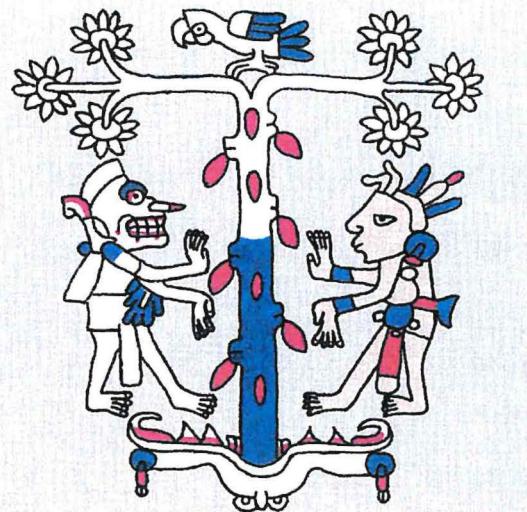


Figure 1. Détail d'un parchemin. Civilisation mixtèque. Codex Fejervary-Mayer. Free Public Museum, Liverpool



Figure 2. Glyphe maya signifiant « cacao »



Figure 3. Illustration tirée du Traitez nouveaux et curieux du café, thé et du chocolat (Phillippe Dufour, 1685)

Une des légendes qui font partie de la cosmogonie des peuples méso-américains, raconte que le dieu Quetzalcóatl, dont le nom signifie le Serpent à plumes s'enivra et perdit la tête après avoir bu un breuvage donné par le magicien Titlaucan. Il embarqua, paré de plumes et partit pour le pays promis, sur un radeau fait de serpents entrelacés, en promettant de revenir un jour pour rapporter à son peuple tous les trésors du paradis. Alors que Moctezuma, empereur des Aztèques attendait le retour de Quetzalcóatl, un homme et son armée argentée débarquèrent. La confusion était évidente. Celui qu'on croyait revenu portait un nom espagnol. Il s'appelait Hernán Cortez, brave et sanguinaire. Il n'était pas le Dieu attendu. On le couvrit d'or et de fèves du cacao et il couvrit la civilisation aztèque de sang.

Hernán Cortès comprit très vite la valeur économique du cacao et fit envoyer, une première cargaison de cacao à Charles Quint, souverain d'Espagne. Lors de ces premières exportations de fèves de cacao en Europe, le cacao fut associé à du sucre, de la vanille ou de la cannelle, ce qui donna naissance à une boisson très appréciée par l'aristocratie espagnole. Cette recette a été gardée en secret pendant une quarantaine d'années. La fermentation et la torréfaction des fèves furent les principales améliorations apportées ensuite aux techniques de préparation de produits alimentaires à base de cacao.

De par le succès du chocolat et sous la pression d'une demande croissante, les espagnols étendirent la culture du cacaoyer dans des vastes régions d'Amérique centrale et du sud, notamment au Venezuela et en Colombie et dans les îles des Caraïbes comme Trinidad et la République dominicaine. Vers 1750, les français firent la même chose en Martinique et à Haïti et les portugais, à Belém et à Bahia au Brésil. La culture du cacao fut également introduite en Asie, aux Philippines à partir de 1560 (Wood, 1991 ; Young, 1994) et, plus récemment, en Afrique où des cabosses de type Forastero Bas-Amazonien puis de type Trinitario furent introduites dans les îles de Sao Tomé et de Fernando Po en 1822 et 1855 (Burle, 1952).

1.2 L'économie du cacaoyer

Theobroma cacao L. est, économiquement, l'une des cultures pérennes les plus importantes dans les tropiques. La consommation de cacao a considérablement progressé à partir de 1850 avec l'invention du presse-beurre, qui a permis de diversifier les produits

Tableau 1. Evolution de la production des principaux pays producteurs.

Pays	1990		2002		Production cumulée 1998-2002	Production mondiale (%)	(%) Variation	
	(Tm)	(Tm/Ha)	(Tm)	(Tm/Ha)			Production	Rendement
1 Côte d'Ivoire	807.501	0,52	1.000.000	0,45	6.102.980	38,7%	4,4%	-0,6%
2 Ghana	293.355	0,42	380.000	0,30	2.033.660	12,9%	4,0%	-3,0%
3 Indonésie	142.347	0,90	348.000	0,97	2.027.200	12,8%	8,2%	2,0%
4 Nigeria	244.000	0,34	338.000	0,35	1.609.000	10,2%	1,9%	-1,0%
5 Brésil	256.246	0,39	172.743	0,30	1.035.530	6,6%	-5,0%	-5,4%
6 Cameroun	115.000	0,32	115.000	0,31	593.600	3,8%	1,1%	0,7%
7 Equateur	96.722	0,29	106.714	0,35	442.996	2,8%	-0,1%	1,6%
8 Malaisie	247.000	0,83	50.000	0,69	379.783	2,4%	-13,1%	-0,1%
9 Colombie	56.153	0,47	47.095	0,48	237.503	1,5%	-2,1%	0,6%
10 Rep. Dominicaine	43.157	0,36	44.906	0,36	220.455	1,4%	-2,0%	-3,5%
11 Mexique	44.045	0,59	56.233	0,67	216.040	1,4%	-0,5%	-1,7%
12 Papouasie N. Guinée	38.343	0,43	42.500	0,43	196.600	1,2%	1,7%	1,0%
13 Pérou	14.796	0,48	17.837	0,44	109.689	0,7%	3,9%	-0,5%
14 Venezuela	15.527	0,25	18.000	0,28	88.204	0,6%	0,9%	1,2%
15 Sierra Leone	24.000	0,42	10.920	0,36	56.680	0,4%	-1,8%	1,5%
Autres (44 pays)	93.845		82.776		435.817	2,8%		
Monde	2.532.037	0,44	2.830.724	0,41	15.785.737	100%	2,0%	-0,8%

Source FAO

Tableau 2. Evolution de l'importation des principaux pays broyeurs et transformateur dans le monde

Pays	1990	2001	Importation cumulée 1998-2002	Importation cumulée 1998-2002 (%)		Variation (%)
				Importation cumulée 1998-2002 (%)	Variation (%)	
1 Pays Bas	313.895	567.998	2.224.561	19,6%	4,6%	
2 Etat unis	337.195	434.105	2.138.564	18,8%	2,4%	
3 Allemagne	297.154	212.137	1.285.613	11,3%	-3,1%	
4 Angleterre	150.445	148.279	882.841	7,8%	0,2%	
5 France	67.475	162.295	666.728	5,9%	7,6%	
Belgique et Luxembourg	57.295	95.523	407.278	3,6%	3,1%	
7 Malaisie	101	139.440	390.395	3,4%	56,6%	
8 Italie	54.044	73.898	364.745	3,2%	3,0%	
9 Estonie		60.794	306.236	2,7%	61,6%	
10 Russie		63.279	294.624	2,6%	4,1%	
11 Espagne	42.818	48.059	265.571	2,3%	1,8%	
12 Canada	23.374	59.912	244.444	2,2%	5,8%	
13 Japon	47.599	49.065	238.028	2,1%	0,7%	
14 Brésil	0	33.931	206.735	1,8%	88,6%	
15 Pologne	10.975	29.199	163.063	1,4%	7,3%	
16 Singapour	92.444	29.222	145.335	1,3%	-12,6%	
17 Chine	10.074	21.134	132.093	1,2%	1,5%	
18 Turquie	4.651	37.470	130.541	1,1%	18,7%	
19 Suisse	21.366	24.109	111.861	1,0%	0,9%	
Autress (102 pays)	235.356	171.458	764.365	6,7%	-0,1%	
Monde	1.766.261	2.461.973	11.368.404	100,0%	3,0%	

Source FAO

cacaotés (boissons, tablettes, confiserie, poudre, pâte à tartiner). La production mondiale est passée, en effet, de 18 000 tonnes en 1850 à 155 000 tonnes vers 1900, et atteignit 672 000 tonnes en 1940 (Wood et Lass, 1985). Au début du XX^e siècle, près de 80% de la production mondiale de cacao était réalisée en Amérique centrale et en Amérique du Sud (Bradeau, 1969). Aujourd’hui la production est très concentrée; 60% de la récolte mondiale provient d’Afrique, principalement de Côte d’Ivoire, 40% d’Amérique Latine, et 20% d’Asie. Sur un marché évalué à 3 milliards de dollars, les sept plus grands pays producteurs (Tableau 1) cultivent 85% de la production mondiale de cacao. Bien que les structures de production soient différentes suivant les continents, 90% de la production de cacao provient de petites exploitations de moins de 5 hectares. Ainsi environ 20 millions de personnes dans le monde dépendent directement de la culture du cacao pour assurer leur subsistance.

Les importations mondiales de cacao s'élevaient à 2.4 millions de tonnes en 2001/2002. Près de 80 % de la production est exportée vers les Etats-Unis et l'Europe (spécialement au Pays-Bas et en Allemagne) (Tableau 2), où l'industrie de broyage s'est fortement développée. Les fèves de cacao, fermentées et séchées, constituent la matière première de cette industrie. Après torréfaction et élimination de la coque, elles sont utilisées pour fabriquer des produits semi-finis, comme la pâte et le beurre de cacao, ou des produits finis destinés directement à la consommation, comme la poudre, les tablettes ou les confiseries de chocolat. Le marché est, en effet, dominé à 80% par les multinationales américaines et européennes. Le prix international du cacao est fixé par les cotations faites à Londres (London Cocoa Terminal Market) et à New York (New York Coffee, Sugar and Cocoa Exchange). Le marché du cacao est bien connu pour son instabilité. Il est caractérisé par des périodes d'augmentation de la production appelées « boom du cacao » suivies de périodes de crise (Ruf, 1995). Les prix sont passés d'environ 4 000 dollars US par tonne en 1979 – leur niveau maximum – à environ 880 dollars US par tonne en octobre 2000. Cette situation, associée aux coûts élevés de la plantation a eu des conséquences terribles pour l'économie des principaux pays producteurs et particulièrement, sur les revenus des petits producteurs. Beaucoup d'entre eux ont quasiment abandonné leurs cacaoyers, en y consacrant le moins de temps possible et en n'investissant que le strict nécessaire dans leur entretien.

Le marché du cacao reconnaît le cacao « ordinaire », et le cacao « fin » ou « aromatique ». Les cacaos « fins » en particulier de type Criollo, diffèrent des cacaos courants par des spécificités aromatiques très prononcées. Son marché est réduit et représente environ 5 % du marché mondial (Petithuguenin et Daviron, 1995). Le cacao fin est produit par des variétés de Criollo modernes/Trinitario et de « Nacional ». Les principaux pays producteurs du cacao « fin » sont donc le Vénézuéla et l'Equateur. La production de cacao fin est considérée par ces pays comme un enjeu stratégique pour leur développement économique.

1.3 Taxonomie et biologie du *Theobroma cacao* L.

Le cacaoyer (*Theobroma cacao* L.) appartenait précédemment à la famille des Sterculiacées et a été reclassé récemment dans la famille des Malvacées (ordre des Malvales) (Whitlock *et al.*, 2001). C'est une espèce diploïde avec $2n = 2x = 20$ chromosomes (Carletto, 1946 ; Glicenstein et Fritz, 1989) et la taille du génome haploïde est environ de 0.4 pg/1C, soit $3.9 \cdot 10^8$ pb (Lanaud *et al.*, 1992, Figueira *et al.*, 1992), valeur qui correspond à environ deux fois et demi la taille du génome d'*Arabidopsis thaliana*. Le genre *Theobroma* comprend 6 sections et 22 espèces (Cuatrecasas, 1964). Parmi elles, la seule espèce cultivée est *T. cacao* bien que l'espèce *T. grandiflora* soit utilisé au Brésil pour préparer, à partir de la pulpe, une boisson appelée « cupuaçu ».

Theobroma est un genre distribué dans les forêts humides tropicales situées entre les latitudes 18° nord et 15° sud et avec de températures qui varient entre 16° et 40°C (Cuatrecasas, 1964).

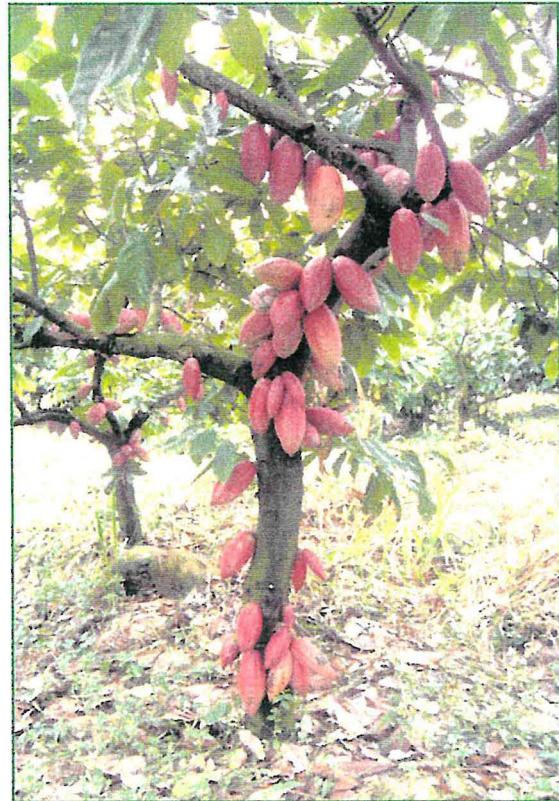
Le cacaoyer est un arbre qui, s'il est issu de semis, développe un axe orthotrope. Puis le bourgeon terminal stoppe sa croissance et cinq branches plagiotropes se développent en couronne. Des couronnes successives se forment à partir de nouveaux rejets orthotropes. Le cacaoyer peut atteindre 5 à 7 mètres de haut en plantation et jusqu'à 12 à 15 mètres à l'état sauvage. Son système racinaire présente un développement orthotrope avec un pivot de un à deux mètres et des racines latérales horizontales très ramifiées.

Le cacaoyer se reproduit à l'état naturel par voie sexuée bien que la multiplication végétative puisse intervenir par la formation de rejets orthotropes sur des troncs ou des branches tombés à terre. La multiplication végétative horticole est réalisée par bouturage



C. Lanaud

Photo 1 . Fleur de cacaoyer



C. Lanaud

Photo 2 . Arbre du cacaoyer



C. Lanaud

Photo 3 . Diversité morphologique trouvée
entre les cabosses des individus Criollo

ou greffage d'axes orthotropes et plagiotropes. Les arbres issus d'axes plagiotropes sont buissonnantes et nécessitent une taille de formation. Les boutures plagiotropes peuvent former, après quelques années, des rejets orthotropes. Les feuilles, longuement pétiolées, apparaissent suivant une phyllotaxie 3/8.

Les fleurs du cacaoyer sont groupées en cymes bipares aux ramifications très courtes (1 à 2 mm) qui se forment directement sur le tronc et les branches. Le renouvellement fréquent des inflorescences en des mêmes points de l'écorce forme de petits massifs renflés appelés « coussinets floraux ». La fleur, hermaphrodite et de petite taille, est constituée de cinq sépales, soudés à leur base, de couleur blanche ou rosée (Photo 1). Elle est bordée intérieurement de deux nervures violettes qui enveloppent les anthères. L'ovaire, supère, comprend cinq loges qui contiennent, chacune six à dix ovules de type anatrophe. Les cinq styles soudés sur la plus grande partie, mais séparés à leur extrémité, forment cinq stigmates tubulaires. La floraison du cacaoyer est très abondante à partir de 18 mois après le semis de la graine pour les variétés les plus précoces. L'intensité de la floraison est influencée par l'éclairement, par les régimes thermiques et hydriques (Alvim, 1965 ; Boyer, 1970) et par l'origine génétique des variétés (Paulin *et al.*, 1983). La pollinisation est entomophile. Elle est faite en particulier par des insectes Diptères du genre *Forcipomyia* et par des fourmis qui pollinisent en moyenne 5 à 10 % de toutes les fleurs produites (Soria 1970 ; Young 1982, 1983, 1984, 1985 ; Paulin *et al.* 1983 ; Lachenaud, 1991).

Le cacaoyer est une plante principalement allogame. Un système complexe gaméto-sporophytique d'autoincompatibilité qui se manifeste avant et après la formation du zygote a été décrit par Knight et Rogers (1955) et par Cope (1962). Des génotypes autocompatibles se rencontrent chez les Forastero bas-amazoniens, les Criollo et les Trinitario ; les Forastero haut-amazoniens sont, en général, auto-incompatibles. Pour certains génotypes, l'auto-incompatibilité n'est pas complète et, dans la plupart des cas, elle peut être levée par l'apport simultané de pollen compatible et incompatible (Lanaud, 1987). Cette situation, favorisée dans la nature par la pollinisation entomophile, peut conduire à un certain taux d'autogamie dans les populations naturelles auto-incompatibles.

Les fruits, appelés « cherelles » au jeune âge et « cabosses » à l'âge adulte, sont portés par le tronc et les plus grosses branches (Photo 2). Le fruit est indéhiscent et ressemble à une baie. Le péricarpe de la cabosse ou cortex est composé de trois couches : l'épicarpe,

charnu, épais et assez dur, le mésocarpe, mince et dur, plus ou moins lignifié, et l'endocarpe, mince, dur et relativement épais. L'extérieur du fruit montre cinq sillons, plus ou moins marqués selon les génotypes. La morphologie des fruits est très variable suivant l'origine des populations mais aussi au sein d'une même population. Ces différences de formes ont souvent été utilisées pour reconnaître les différents cultivars de cacao (Photo 3). Le nombre de fruits matures produits par le cacaoyer est très faible comparé au nombre de fleurs (0.7-4%) (Lachenaud, 1991). Les fruits comportent de 20 à 40 fèves, qui sont disposées en cinq rangées et portées par un rachis central. Ces fèves sont entourées de mucilage blanc, aqueux et sucré. Dès la maturité du fruit les fèves sont prêtes à germer. Après fermentation, séchage et torréfaction, elles sont utilisées pour la fabrication du chocolat.

Le cacaoyer est touché par un grand nombre de maladies et de ravageurs qui constituent souvent des facteurs limitant sa production. Les principales maladies sont la moliniose (*Moniliophthora roreri*), la pourriture brune des cabosses due à *Phytophthora spp.*, le VSD (« Vascular Streak Dieback ») due à *Oncobasidium theobromae* et le balai de sorcière (*Crinipellis perniciosa*). La pourriture brune des cabosses, due à *Phytophthora spp.*, est répandue dans le monde entier mais provoque des dégâts surtout en Afrique et en Papouasie-Nouvelle-Guinée. Le balai de sorcière et la moniliose ont dévasté la production de cacao en Amérique latine. La moliniose représente à l'heure actuelle environ 5% des pertes mondiales totales et est en passe de devenir un problème de plus en plus sérieux en Équateur, en Colombie et en Amérique centrale. Dans certaines régions du Pérou, des pertes de 100% ont été signalées (Evans *et al.*, 1998). Le balai de la sorcière représente environ 21% des pertes mondiales, en particulière en Brésil.

Les principaux insectes ravageurs du cacaoyer sont les mirides (*Sahlbergella sp.*, *Distantiella sp.* et *Helopeltis sp.*), qui attaquent les jeunes pousses et les cabosses, et les thrips (*Selanothrips rubrocinctus*). En Afrique, les mirides peuvent entraîner la dégénérescence totale du feuillage en quelques années. En Asie du Sud-Est, un insecte foreur des cabosses, le « cocoa pod borer » (*Conopomorpha cramerella*), provoque des pertes de fruits très importantes.

1.4 Classification des variétés cultivées de cacaoyer

Les variétés cultivées de *Theobroma cacao* ont été classiquement divisées en deux grands groupes morpho-géographiques : les Forastero et les Criollo. Les premiers taxonomistes considéraient ces groupes comme des espèces différentes (Morris, 1882, Pittier, 1930) alors que Cuatrecasass (1964) distingua ces deux groupes morpho-géographiques comme des sous-espèces différentes : *Theobroma cacao* subsp. *cacao* et *T. cacao* subsp. *sphaerocarpum* qui correspondaient respectivement aux groupes Criollo et Forastero. Un troisième groupe, Trinitario, résulte de l'hybridation entre les deux premiers.

Cette classification correspond à une classification pratique des producteurs de cacao de la fin du XIX^e siècle. Les termes Criollo et Forastero proviennent du vocabulaire vénézuélien qui qualifie de Forastero (étranger) toutes les variétés différentes des Criollo qui étaient les cacaoyers locaux traditionnellement cultivés.

Le groupe Forastero, originaire d'Amérique du Sud, regroupe la majorité des cacaoyers cultivés dans le monde. Ce groupe, très variable, rassemble un grand nombre de populations sauvages et de variétés cultivées. Les Forastero sont, généralement, des arbres vigoureux, précoces, de bon rendement, aux fèves aplatis de couleur violette et de taille variable (Photo 4). Ils sont aussi reconnus comme des sources de résistance aux principales maladies du cacaoyer. Au sein des Forastero, plusieurs groupes ont été différenciés selon leur origine géographique : les Forastero originaires de la Haute Amazonie (Pérou, Bolivie, Equateur et Colombie) eux mêmes très variables, les Forastero bas amazoniens dont la variété la plus typique, l'Amelonado est cultivée dans le bassin amazonien (Brésil), les Forastero du Vénézuéla qui se trouvent le long du fleuve Orénoque (Vénézuéla) et les Forastero des Guyanes.

Les Criollo ont été les premiers cacaoyers domestiqués par les civilisations préhispaniques. De façon générale, les Criollo sont des arbres moins vigoureux et plus sensibles aux maladies, mais qui donnent un chocolat fin de haute qualité. Les Criollo produisent des fruits de morphologie très variable. Certaines variétés ont de grands fruits allongés et pointus dont la surface rugueuse est marquée par des sillons profonds. D'autres variétés telles que les Porcelana produisent de petits fruits, de surface lisse et de forme pointue. Les fèves sont grandes, épaisses et claires, de couleur blanche ou rosée à maturité

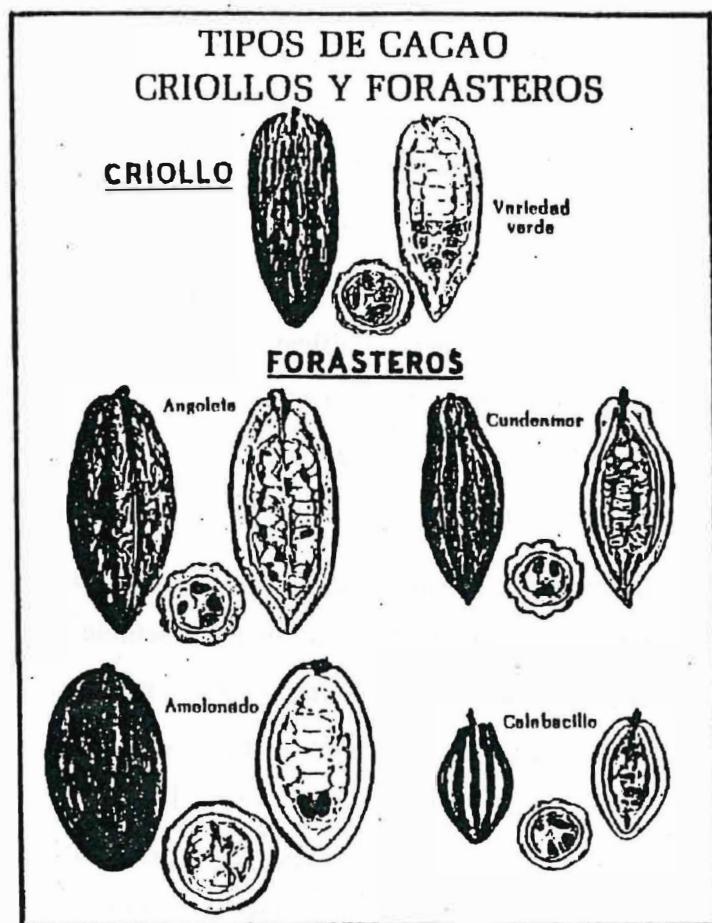


Figure 4. Principales formes de cabosses
(D'après Braudeau, 1969)

(Photo 5). La paroi des cabosses est peu lignifiée. Les Criollo sont réputés pour donner un cacao de bonne qualité caractérisé par des arômes de caramel et de noisette. Les Criollo se rencontrent du Mexique jusqu'en Colombie et au Vénézuéla (Reyes, 1992). Ils sont également cultivés à Madagascar et en Asie du Sud-Est.

Les Trinitario sont issus d'un croisement entre Criollo et Forastero qui a eu lieu sur l'île de Trinidad au XVIII^e siècle, alors que les plantations de Criollo avaient quasiment disparu de l'île suite à une catastrophe écologique (Cheesman, 1944). Ces hybrides, vigoureux, ont été diffusés dans de nombreux pays comme le Vénézuela et en Amérique centrale, se mélangeant progressivement aux populations de Criollo. Ils ont été ensuite introduits vers 1900 en Afrique de l'Ouest, où ils se sont croisés avec les Amelonado introduits précédemment.

En Equateur, les plantations anciennes étaient constituées de la variété « Nacional ». Cette variété constitue une population particulière (Pound, 1945 ; Lecerteau *et al.*, 1997). Elle est caractérisée par le goût « Arriba ».

1.5 Diversité génétique des cacaoyers

Les premières descriptions des cultivars se fondaient sur la forme de leurs cabosses : Amelonado (forme régulière ovale comme celle d'un petit melon, surface lisse, sillons peu marqués), Angoletta (cabosse pointue, base large, sillons profonds), Calabacillo (cabosse petite et presque sphérique), Cundeamor (cabosse allongée, pointue, base rétrécie en goulot de bouteille, surface verrueuse) (Figure 4). Ces appellations ne sont aujourd'hui utilisées que comme termes de référence pour décrire la forme des cabosses de certains cultivars par rapport à quelques formes caractéristiques, à l'exception du terme Amelonado, qui caractérise certains cultivars du bassin amazonien.

D'autres auteurs ont étudié la diversité morphologique à partir de caractères comme la couleur et la taille des fèves. Quelques études basées sur des caractères morphologiques ont montré une forte différentiation entre les Forastero d'une part et les Criollo et Trinitario d'autre part (Engels, 1986 ; N'Goran, 1994 ; Bekele et Bekele, 1996). Ces résultats ont été confirmés par une étude récente des accessions qui se trouvent dans la collection de ressources génétiques du CRU à Trinidad (Iwaro *et al.*, 2003). L'International Plant Genetic Resources Institute (IPGRI) a publié une liste qui comporte

65 descripteurs morphologiques pour caractériser les ressources génétiques du cacaoyer (IBGRI, 1981).

Au cours des dernières années, plusieurs auteurs ont entrepris des études dans le but de caractériser la diversité génétique de populations de cacaoyer par l'utilisation de marqueurs isoenzymatiques et de marqueurs moléculaires (RAPD, RFLP et microsatellites).

Malgré le faible nombre de locus révélés, Ronning et Schnell (1994) ont montré par une analyse de la diversité enzymatique de clones ou de populations de cacaoyers l'existence d'une grande différentiation génétique entre génotypes Criollo et Forastero. D'autres études ont établi aussi une forte diversité de clones provenant de la haute Amazonie (Warren, 1994), une variabilité réduite chez les populations du Venezuela et des Guyanes étudiées (Lanaud, 1987) ainsi qu'une forte hétérozygotie chez les Trinitario (Lanaud, 1987 ; Sounigo, 1996).

Des études de diversité ont été réalisés à l'aide des marqueurs moléculaires de type RAPD (Wilde *et al.*, 1992 ; Figueira *et al.*, 1994 ; Russel *et al.*, 1993 ; N'Goran *et al.*, 1994 ; Ronning *et al.*, 1995 ; De la Cruz *et al.*, 1995 ; Sounigo *et al.*, 1996 ; Lercetau *et al.*, 1997 ; Whitkus *et al.*, 1998), de type RFLP (Laurent *et al.*, 1994 ; Figueira *et al.*, 1994 ; Lercetau *et al.*, 1997 ; Motamayor *et al.*, 1997) et de marqueurs microsatellites (Lanaud *et al.*, 1999).

Ces analyses ont mis en évidence 1) une structuration claire entre Forastero et Criollo actuels ; 2) une grande diversité au sein des Forastero ; 2) l'originalité de la variété « Nacional » ; 3) Des niveaux élevés d'hétérozygotie chez les Trinitario et les Criollo actuellement cultivés et 4) Une grande proximité génétique entre les variétés actuelles de Criollo et les Trinitario cultivés en Amérique et en Afrique. Il faut rappeler que les individus appelés Criollo dans ces travaux correspondent en fait aux variétés de Criollo cultivés actuellement (Criollo modernes).

Les analyses les plus récentes se sont surtout focalisées sur l'étude des Criollo et de leur domestication (Motamayor *et al.*, 1997 ; Motamayor *et al.*, 2002, 2003). Des prospections réalisées dans cinq pays d'Amérique Latine ont permis d'échantillonner des Criollo dans des sites redevenus sauvages (forêt Lacandona, Mexique), et dans les plus vieilles plantations, là où pas ou peu d'introductions de Forastero avaient été faites.

L'étude de diversité a mis en évidence une très faible diversité génétique au sein des Criollo considérés comme « Criollo anciens » et qui ne sont pas introgressés par des gènes de Forastero. Malgré une diversité morphologique importante des formes de cabosse, ces Criollo anciens sont tous très homozygotes et identiques du Mexique jusqu'au Vénézuela.

Motamayor *et al.* (2002, 2003) ont montré que les Criollo actuellement cultivés (modernes) correspondent à des Criollo « anciens » introgressées par des Forastero bas-amazoniens. Par des analyses de paternité, ces auteurs ont pu identifier les parents à l'origine des Trinitario et de ces Criollo modernes. La diversité allélique de ce groupe peut être presque totalement expliquée par deux individus Criollo anciens et trois individus Forastero de Basse Amazonie. Ces analyses ont aussi montré que les Criollo présents dans les collections nationales et repérés dans les plantations locales pour leur bon comportement en champ, correspondent en fait à des hybrides Trinitario. Donc, la quasi totalité des Criollo modernes et des Trinitario, très utilisés dans les programmes de sélection et représentés en très grand nombre dans les collections, correspond à une base génétique extrêmement réduite. Ces résultats ont des conséquences importantes sur la gestion des ressources génétiques et sur les stratégies de sélection étant donné qu'une faible part de la diversité génétique a été jusqu'à présente exploitée en sélection.

Un autre résultat important de ces études est celui qui concerne l'origine des Criollo. Deux hypothèses ont été faites pour expliquer cette origine. Cuatrecasas (1964) suggère que les deux grands groupes « génétiques » de cacaoyer (Forastero et Criollo) auraient évolué indépendamment au cours de la différenciation de l'espèce : l'un, le groupe des Criollo qui est à l'origine des premiers cacaoyers cultivés en Amérique Centrale et l'autre, le groupe Forastero, en Amérique du Sud. Au contraire, Cheesman (1944) propose que les cacaoyers Criollo d'Amérique Centrale sont des cacaoyers domestiqués à partir d'une population d'Amérique du Sud et que les deux types de cacaoyers cultivés se seraient différenciés en Amérique du Sud. Quelques spécimens de cacao Criollo auront été transportés ensuite par l'homme en Amérique Centrale. Les résultats obtenus par Motamayor (2002) privilégient une origine sud-américaine des Criollo. En effet, les Criollo anciens sont plus proches des Forastero colombiens que ceux-ci ne le sont d'autres populations de Forastero d'Amérique du Sud. Ces résultats moléculaires suggèrent que la variété traditionnelle Criollo s'intègre dans la diversité des populations d'Amérique du sud bien qu'elle en soit un peu différenciée.

L'ensemble de ces résultats permettent de proposer une structure différente de la classification classique en trois groupes morphogéographiques (Lanaud *et al.*, 2003). Cette structure comprend :

- Les Criollo anciens, avec une base génétique très étroite et qui se rencontrent aujourd'hui en petites populations redevenues sauvages ou cultivées dans de petites plantations.
- Les variétés anciennes de cacaoyer « Nacional » d'Equateur à l'origine du goût Arriba.
- Les Forastero sauvages de Guyane.
- Les Forastero du bassin amazonien (de haute et basse Amazonie) et les cacaoyers situés dans la vallée de l'Orénoque qui forment un groupe dont la diversité et beaucoup plus vaste et continue.

1.6 Les collections des ressources génétiques

Dû à l'importance de la conservation des ressources génétiques de *T. cacao*, plusieurs collections nationales et internationales ont été créées. Une quarantaine de prospections ont été réalisées depuis 1930. Elles ont permis du collecter des cacaoyers cultivées et sauvages de type Forastero du Pérou, du Brésil, d'Equateur, de Colombie, du Venezuela et de Guyane et, plus récemment, des Criollo d'Amérique centrale, du Vénézuela et de Colombie. Chacune de ces collections a été récemment répertoriée par Motilal et Buttler (2003). Ces collections sont maintenues sous forme d'arbres rassemblés au champ.

La FAO au travers de l'Institut international des ressources phytogénétiques (IPGRI) a appuyé la mise en place des collections désignées comme des « Universal Collection Depositories » (IBGRI, 1981) : l'International Cocoa Genebank, Trinidad (ICG,T) avec environ 2500 génotypes, riche en Forastero de haute Amazonie ; et la collection du Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica (CATIE) (environ 800 génotypes) avec une représentation majoritaire des génotypes d'origine Criollo moderne/Trinitario.

La diversité génétique conservée dans ces collections a été évaluée en utilisant des caractères agronomiques et morphologiques et ces données ont été compilées dans

l'« International Cocoa Germplasm Database » (ICGD) (Wadsworth et Harwood, 2000) disponible sous forme de CD (ICGD 2000 v4.1 CD-ROM) et sur le site Internet : <http://www.icgd.reading.ac.uk/>. Cette base de données rassemble des informations sur plus de 14 000 génotypes conservés dans 43 collections. Une autre base de données, TropgeneDB existe au Centre de coopération international en recherche agronomique pour le développement (CIRAD) (<http://tropgenedb.cirad.fr>) ; elle regroupe les données et moléculaires d'environ 400 génotypes. Une nouvelle base de données, Cocoa GEN-DB (<http://cocoagendb.cirad.fr>) vient d'être mise en place et regroupe les données morphologiques et moléculaires de ces deux bases et permet des requêtes complexes combinant ces 2 types de caractères.

Malgré le grand nombre de génotypes de type Trinitario conservés dans ces collections, les différences morphologiques importantes observées ne semblent pas être corrélées à une forte diversité génétique. Elles sont, probablement, le résultat de mutations ponctuelles qui ont été sélectionnées par l'homme. La collection du CATIE contient un très grand nombre d'individus Criollo modernes et Trinitario ce qui pourra donner lieu à des redondances dues à la faible diversité existante parmi eux.

En effet, Motilal et Butler (2003), à partir des données rassemblées dans la base de données ICGD, ont reporté la présence de possibles identifications erronées et de duplications des accessions dans la majorité de ces collections.

2. La cartographie génétique

2.1 Les marqueurs moléculaires

Les premiers marqueurs génétiques étaient des gènes pour lesquels une variation allélique avait été mise en évidence par l'identification de mutants morphologiques ou physiologiques. Donc, dans sa définition primaire, un marqueur génétique est un caractère mesurable à hérédité mendélienne, dont l'observation permet de déterminer le génotype d'un individu à un locus. On oppose traditionnellement les marqueurs morphologiques (couleur, forme, etc.) aux marqueurs biochimiques (protéines, isozymes) et aux marqueurs moléculaires (au niveau de l'ADN).

De multiples propriétés définissent un bon marqueur génétique: il doit être polymorphe et discriminant, multiallélique, codominant , non épistatique, neutre, affranchi du milieu, facilement manipulable et de coût peu élevé.

Un marqueur moléculaire est une séquence d'ADN spécifiquement repérable. De façon générale, les marqueurs moléculaires sont classés en 1) marqueurs qui révèlent en masse plusieurs locus [marqueurs généralement dominants comme les RAPD (Random Amplified Polymorphism DNA) et les AFLP (Amplified Fragment Length Polymorphism)] et 2) marqueurs spécifiques de locus [marqueurs codominants comme les RFLP (Restriction Fragment Length Polymorphism) et les microsatellites].

Les microsatellites ou SSR (Simple Sequence Repeats) sont constitués de répétitions en tandem de motifs mono-, di-, tri- ou tétra-nucléotidiques (Tautz et Rentz, 1984). Leur polymorphisme repose sur la variation du nombre d'unités de répétition constituant le microsatellite. Les régions flanquant ces éléments répétés permettent de définir les amores utilisées pour l'amplification PCR (Polymerase Chain Reaction) du fragment contenant le SSR. Les fragments amplifiés sont ensuite révélés par électrophorèse. Les microsatellites regroupent l'ensemble des qualités d'un bon marqueur génétique: codominance, haut niveau de polymorphisme, multiallélisme et spécificité de locus. La possibilité du multiplexage au niveau de la PCR et le développement considérable de l'automatisation de l'électrophorèse donnent des avantages à ces types de marqueurs.

Par contre, l'inconvénient majeur de cette technique repose sur la nécessité d'un développement préalable et coûteux d'amores microsatellites pour chaque espèce à analyser. Ce développement nécessite de disposer de séquences microsatellites à partir desquelles seront définies les amores locus-spécifiques, soit en criblant une banque de clones d'ADN génomique (par hybridation à l'aide de sondes microsatellites synthétiques) soit en produisant directement une banque enrichie en séquences microsatellites. Cependant, Billote *et al.* (1999) ont développé une technique simple pour la création de banques enrichies en séquence microsatellites. Cette technique a été utilisé avec succès pour identifier des microsatellites chez différentes plantes tropicales (Billote *et al.* 2001; Aranzana *et al.* 2002; Dirlewanger *et al.* 2002 ; Viruel et Hormaza, 2004 ; Nguyen *et al.*, 2004). Chez le cacaoyer les premiers microsatellites ont été isolés par Lanaud *et al.* (1999). De façon générale, les motifs répétés plus abondants sont les motifs $(GA)_n$ et $(GT)_n$ chez les arbres tropicaux (Condit and Hubbel, 1991), les motifs $(CCG)_n$ chez les

monocotylédones et les motifs tri- ou tétra-nucléotidiques chez les dicotylédones (Zhao et Kochert, 1993).

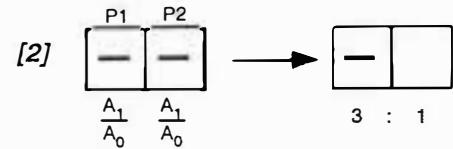
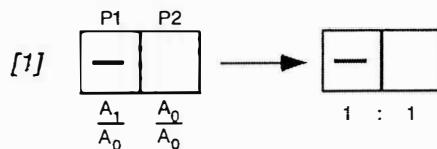
L'usage de ce type de marqueurs s'est répandu depuis les années 1990 grâce à leurs avantages sur d'autres techniques d'analyse. De nombreuses cartes génétiques basées sur des locus microsatellites ont été construites pour des espèces très différentes comme le sorgho (Bhatramakki *et al.*, 2000; Haussmann *et al.*, 2002), le blé (Gupta *et al.*, 2002), le riz (Temnykh *et al.*, 2000; McCouch *et al.*, 2002), l'amandier (Joobeur *et al.*, 2000), le pommier (Liebhard *et al.*, 2002), les pruniers (Aranzana *et al.*, 2003) ou la vigne (Adam-Blondon *et al.*, 2004).

2.2 La cartographie génétique

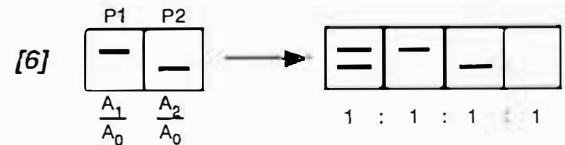
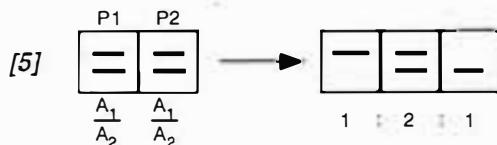
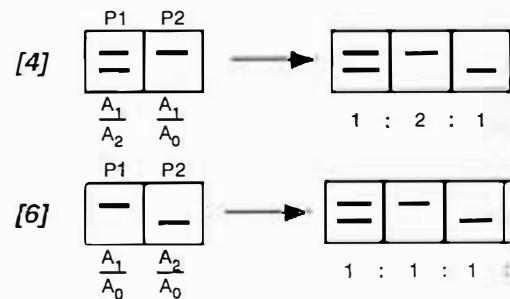
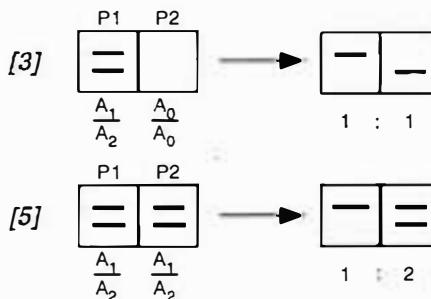
La cartographie génétique consiste à établir des groupes de liaison au sein desquels les marqueurs moléculaires sont ordonnés et les distances entre marqueurs calculées (Ritter *et al.*, 1990). Les cartes génétiques constituent donc, une représentation du génome d'une espèce, elles indiquent les positions relatives des marqueurs les uns par rapport aux autres. Les cartes génétiques sont construites selon un principe simple: la distance entre 2 locus est fonction du pourcentage de recombinaisons entre ces 2 locus lors de la méiose. Ce pourcentage est déduit de la proportion des différentes classes génotypiques dans la descendance. Dans le cas d'une plante diploïde, lors de la méiose, les gamètes obtenus peuvent être de type parental ou recombiné ; ces derniers sont le produit d'un événement de recombinaison entre deux chromosomes homologues grâce à un crossing-over. Ce phénomène a d'autant moins de chance de se produire entre deux locus que ceux-ci sont physiquement proches. Les distances génétiques entre locus sont basés sur les taux de recombinaison entre deux locus, de valeur comprise entre 0 (il n'y aura aucun gamète de type recombiné et les locus sont totalement liés) et 0.5 (les locus ne sont pas liés et ségrégent de façon indépendante). La fréquence de recombinaison est traduite en distance génétique dont l'unité est le centimorgan (cM).

L'établissement d'une carte génétique nécessite donc, tout d'abord, la création d'une descendance en ségrégation issue d'un croisement contrôlé de deux parents différents et la détermination des génotypes de chaque individu de la descendance pour un ensemble de locus marqueurs distribués le long du génome. L'utilisation d'outils d'analyse statistique et

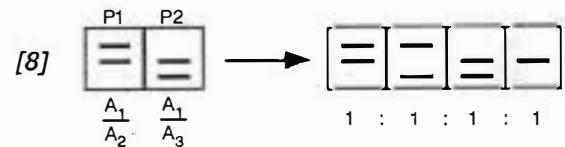
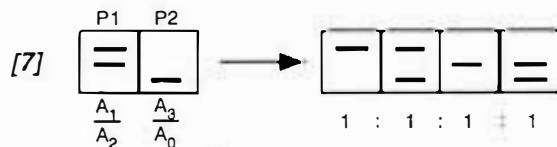
- Configurations à 1 allèle en ségrégation :



- Configurations à 2 allèles en ségrégation :



- Configurations à 3 allèles en ségrégation :



- Configurations à 4 allèles en ségrégation :

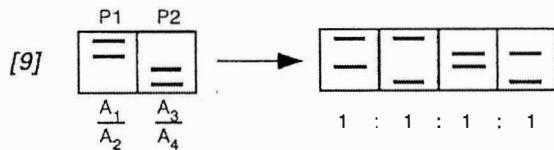


Figure 5. Configurations génotypiques observées lors d'un croisement entre deux plantes hétérozygotes [Lespinasse (1999) d'après Ritter *et al.* (1990)].

de méthodes de calculs mathématiques permet alors 1) de réunir les marqueurs qui sont sur le même groupe de liaison ou chromosome ; 2) d'ordonner les marqueurs au sein des groupes et ; 3) d'estimer les distances entre marqueurs.

Les populations de cartographie le plus couramment utilisées sont : les descendances F2, les populations issues de rétro-croisements ou backcross, les haploïdes doublés et les populations de lignées recombinantes. Le choix des parents est important. Il y aura, en effet, plus de chance d'obtenir du polymorphisme dans une descendance issue du croisement entre deux parents très différenciés.

Une stratégie d'établissement de la carte génétique chez des arbres pérennes hétérozygotes a été proposée par Gratacchia et Sederoff (1994). Cette stratégie est appelée Pseudo-test-cross et profite du niveau élevé d'hétérozygosité de certaines espèces. Comme les parents sont très hétérozygotes il y a deux possibilités pour la ségrégation de chaque locus : ceux qui ségrègent chez les deux parents (locus hétérozygotes chez les deux parents) et ceux qui ségrègent chez un seul des parents (locus hétérozygote pour un des parents et homozygote pour l'autre). La ségrégation de la descendance permet d'obtenir de façon simultanée les cartes génétiques de deux parents. On peut ensuite établir une carte consensus à l'aide de marqueurs communs et hétérozygotes chez les deux parents, et qui servent de marqueurs « ponts » entre les cartes génétiques des deux parents. Ritter *et al.* (1990) et Ritter et Salamini (1996) ont décrit les configurations génotypiques observées lors d'un croisement entre deux plantes hétérozygotes (Figure 5) (Lespinasse, 1999). Le logiciel JOINMAP (Stam, 1995) est souvent utilisé pour l'analyse de ce type de population. De nombreuses cartes de plantes tropicales et d'espèces forestières ont été construites avec cette approche (Gratacchia et Sederoff, 1994 ; Lespinasse *et al.*, 2000 ; Joobeur *et al.*, 2000 ; Chagné *et al.*, 2002 ; Wu *et al.*, 2004).

2.2.1 Ségrégation des marqueurs

La première étape, avant d'estimer la liaison entre marqueurs, est de tester l'hypothèse nulle d'une ségrégation mendélienne pour chaque marqueur à partir d'un test du χ^2 . Ce test permet de comparer les effectifs génotypiques observés aux effectifs théoriques, calculés sous l'hypothèse d'une ségrégation mendélienne. Cette hypothèse dépendra du type de population et de marqueurs utilisés. Par exemple, on testera la ségrégation 1:1 pour les

marqueurs codominants ou dominants dans des familles issues de « backcross », d'haploïdes doublés et la ségrégation 1:2:1 dans des lignées recombinantes, pour des marqueurs codominants, 3:1 pour des marqueurs dominants qui ségrègent dans des familles F₂. Dans le cas d'un pseudo-test-cross et pour chaque marqueur, des ségrégations de type backcross ou de type F₂ sont possibles.

2.2.2 Estimation de la liaison entre marqueurs et ordonnancement des marqueurs

Une fois testée la ségrégation mendéienne pour chaque marqueur, on peut détecter l'existence de liaison entre marqueurs. Si deux marqueurs sont sur deux chromosomes différents, ces marqueurs auront une ségrégation indépendante et il y aura autant de gamètes parentaux que de gamètes recombinés. S'ils sont sur le même chromosome, la proportion de gamètes recombinantes sera fonction de la distance entre ces deux marqueurs.

On peut tester la liaison entre locus marqueurs pris deux à deux par un test du χ^2 , sous l'hypothèse nulle d'une absence de liaison (taux de recombinaison = 0.5). Si les locus analysés sont suffisamment proches l'un de l'autre sur le même groupe de liaison, la valeur du χ^2 sera significative puisque les effectifs des individus recombinants seront plus faibles que les effectifs des types parentaux. Plus couramment, on teste cette liaison par la méthode du LOD score (Log of the odds ratio en anglais ou logarithme du rapport des vraisemblances) décrit par Morton (1955). L'estimation du taux de recombinaison est fondée sur la valeur du taux de recombinaison r qui maximise la probabilité de distribution observée des effectifs des différentes classes génotypiques. La résolution de l'équation de vraisemblance donne un estimateur de r (Fisher, 1921).

$$L(r) = \frac{n!}{\prod_j [n_j!]^{f_j}} \prod_j [f_j]^{n_j}$$

où n est l'effectif total de la descendance, n_j l'effectif observé de chaque classe de génotype j et f_j, la fréquence attendue de chaque classe de génotype j. On estime la valeur de r qui maximise la vraisemblance des observations.

Le test statistique du LOD score est égal au logarithme décimal du rapport des vraisemblances des deux hypothèses :

$$LOD = \log_{10} \frac{V(r)}{V(r_0)}$$

où $V(r)$ est la vraisemblance calculée avec la valeur la plus probable de r (estimateur r) à partir de la résolution de l'équation de vraisemblance et $V(r_0)$ la vraisemblance calculée pour $r = 0.5$, c'est-à-dire en absence de liaison.

Cet indice, exprimé en logarithme décimal exprime la probabilité de l'existence d'une liaison par rapport à celle de la non liaison. Un LOD égal à 3 indique qu'une liaison est 1.000 fois plus probable que la non-liaison. Généralement, on utilise des valeurs de 3.0 à 5.0 pour le seuil du test du LOD score et de 0.3 à 0.49 pour le seuil du taux de recombinaison r .

Une fois testés tous les marqueurs pris deux à deux, les marqueurs liés sont regroupés puis ordonnés au sein de groupes de liaisons. Les ordres que l'on attribue aux locus par l'analyse de la ségrégation de deux locus sur les groupes de liaison et les estimations des fréquences de recombinaison ne sont cependant pas corrects. Par contre, l'analyse multipoint, qui met en jeu tous les locus d'un group de liaison, permet d'établir l'ordre le plus probable car les recombinaisons multiples sont prises en compte. Si il y a suffisamment de marqueurs, le nombre de groupes de liaison sera égal au nombre haploïde de chromosomes. La méthode utilisée est toujours celle du maximum de vraisemblance donc la carte résultante est en fait, la solution de vraisemblance maximale parmi tous les ordres des marqueurs possibles.

2.2.3 Estimation des distances entre les marqueurs

Comme on l'a déjà décrit, la distance génétique entre deux marqueurs est estimée à partir de l'analyse de la ségrégation dans la descendance [fréquences de recombinaison (r) entre les différents locus]. Ces fréquences de recombinaison exprimées en pourcentage (%) sont ensuite transformées en distances additives exprimées en centimorgans (cM). Cette transformation est réalisée à l'aide de deux fonctions principales de cartographie : Haldane (1919) et Kosambi (1944). La fonction de Kosambi prend en compte l'effet d'un

« crossing over » dans un segment quelconque sur les recombinaisons possibles dans un intervalle adjacent (interférence), tandis que la fonction d'Haldane suppose qu'il n'y a pas d'interférence. La fréquence de recombinaison estimée entre deux locus ne prend en compte que les recombinaisons « impaires », car les effets des recombinaisons « paires » s'annulent (Lorieux, 1993). Les logiciels informatiques les plus couramment utilisés pour établir une carte génétique sont MAPMAKER (Lander *et al.*, 1987) et JOINMAP (Stam, 1995).

2.3. Cartographie génétique chez le cacaoyer

La première carte génétique du cacaoyer a été construite par Lanaud *et al.* (1995) à partir d'une descendance hybride de Côte d'Ivoire, issue d'un croisement entre deux parents hétérozygotes : un Forastero haut amazonien (UPA402) et un Trinitario (UF676). Cette première carte comportait 202 marqueurs, essentiellement RFLP, et a été établie à partir de 100 individus. Risterucci *et al.* (2000) ont augmenté la densité de marqueurs sur cette carte en y ajoutant des marqueurs microsatellites et AFLP. Cette carte de référence comporte 473 marqueurs, principalement RFLP, AFLP et microsatellites, répartis tout au long des 10 chromosomes. Ils couvrent une distance de 885 cM, soit en moyenne un marqueur tous les 2.1 cM

Les marqueurs moléculaires développés et cartographiés sur la carte de référence ont servi de base pour la recherche de QTL (Quantitative Traits Locus). Les premiers QTL du rendement chez le cacaoyer ont été détectés à partir du même croisement de la carte génétique de référence (UPA402xUF676) (Lanaud *et al.*, 1999). Des QTL du rendement ont été également cartographiés à partir d'un croisement Catongo x Pound12 (Crouzillat *et al.*, 1996, 2000a, 2000b). D'autre part, trois descendances issues de croisements entre trois parents femelles hétérozygotes (deux Trinitario et un Forastero haut amazonien) et un parent mâle homozygote bas amazonien ont également permis de détecter des QTLs pour de caractères liés au rendement (Clément *et al.*, 2003a, 2003b).

Des QTLs de facteurs impliqués dans la résistance à la pourriture brune des cabosses due aux *Phytophthora* ont été aussi identifiés (Flament, 1998 ; Flament *et al.*, 2001 ; Lanaud *et al.*, 2000, Paulin *et al.*, 2000 ; Risterucci *et al.*, 2003) ainsi que des QTL liés aux caractères de qualité (Lanaud *et al.*, 2003).

3. Déséquilibre de liaison et « association mapping »

Le développement des marqueurs moléculaires a permis de localiser sur une carte génétique des locus impliqués dans la variation des caractères d'intérêt (QTL pour Quantitative Trait Locus en anglais). Le principe est de corrélérer la variation phénotypique d'un caractère à la variation des allèles parentaux dans une population en ségrégation, c'est-à-dire une descendance contrôlée. Une nouvelle approche consiste à exploiter le déséquilibre de Liaison existant chez des individus issus de populations naturelles ou en sélection, afin d'optimiser la détection de QTLs sans faire appel à des structures familiales ou à des descendances en ségrégation (« association mapping »).

Le déséquilibre de liaison (DL) ou déséquilibre gamétique, mesure l'association non aléatoire d'allèles pris à des locus différents (Lewontin and Kojima, 1960). Le DL permet de quantifier la non indépendance des allèles, c'est à dire le fait que les allèles ne sont pas associés au hasard dans la population. Le terme « déséquilibre de liaison » est souvent associé par erreur à la liaison génétique ce qui laisse croire qu'il ne concerne que des locus physiquement liés.

Comme le DL décrit une association statistique entre deux locus, le DL observé regroupe les effets de liaison physique vraie sur le génome et les effets d'association apparente de locus indépendants dus aux forces évolutives comment les mutations, la dérive génétique, et la sélection ou dus à la structure de la population.

Si le DL est du aux liaisons physiques entre locus proches, ceci signifie la présence d'une combinaison d'allèles de multiple locus sur un chromosome (haplotypes) qui ne sont pas associés au hasard et qui tendent à être transmis ensemble car ces locus ont une généalogie similaire et provient d'un ancêtre commun (Mohlke *et al.*, 2001). Dans ces cas, on parle de **déséquilibre local** et le DL peut être utilisé pour localiser des gènes d'intérêt dans le génome. Par contre, deux locus situés sur des chromosomes différents peuvent aussi se trouver en déséquilibre, si cela est le cas on parle alors de **déséquilibre global**. Différents facteurs liés à l'histoire de la domestication et à la constitution du génome de l'espèce créent ou diminuent ces types de DL.

Le concept du DL est relativement ancien (Jenning, 1917) et très connu par les généticiens des populations. Récemment, le DL a été redécouvert et a reçu une attention considérable en particulier pour la cartographie des maladies complexes chez les humains. Quelques articles exclusivement consacrés au DL chez les plantes sont aussi récemment apparus (Flint-Garcia *et al.*, 2004 ; Rafalski et Morgante, 2004).

3.1 Définition et mesures du DL

Il existe de nombreux indices pour mesurer le déséquilibre de liaison entre paires de locus. L'expression du DL est formalisée par D, une mesure proposée par Lewontin (1964). D représente la différence entre la fréquence observée d'un haplotype donné à deux locus et la fréquence attendue de cet haplotype si les allèles s'associaient de façon aléatoire tel que cela est décrit ci-dessous :

Soient deux locus polymorphes :

le locus A avec les allèles A et a en fréquence p_A et p_a , respectivement

le locus B avec les allèles B et b en fréquence p_B et p_b ,

Il y a quatre combinaisons gamétiques possibles : AB, Ab, aB et ab. Si les associations alléliques se font au hasard dans la population gamétique, la fréquence du gamète AB est :

$$p_{AB} = p_A * p_B$$

C'est-à-dire qu'il y a indépendance statistique. On donne souvent le nom d'haplotype à ces combinaisons gamétiques, par exemple l'haplotype AB. La fréquence relative de l'haplotype sera alors, simplement le produit des fréquences relatives des allèles respectifs qui le constituent. Dans ce cas, la population est à l'équilibre d'association gamétique ou en équilibre de liaison. Par contre, si l'association des allèles n'est pas aléatoire lors de la formation des gamètes, les fréquences de certains types de gamètes seront supérieures ou inférieures à celles attendues à l'équilibre.

La valeur D représente le déséquilibre d'association gamétique ou déséquilibre de liaison comme l'écart à l'association aléatoire :

$$D = p_{AB} - p_A p_B$$

Tableau 3. Fréquences des gamètes en déséquilibre de liaison

Locus 1 Locus 2	A	a	Somme
B	$pApB + D$	$qapB - D$	pB
b	$pAqb - D$	$qaqb + D$	qb
Somme	pA	qa	

Les fréquences de quatre types de gamètes en déséquilibre dans la population se présentent dans le tableau 3. Si la population est en équilibre de Hardy-Weinberg, les fréquences haplotypiques peuvent changer lorsque les fréquences alléliques se maintiennent. La valeur de D est comprise entre -0.25 et 0.25. Le déséquilibre est maximal lorsque la connaissance de l'allèle présent à un locus permet de prédire l'allèle de l'autre locus, donc lorsqu'il n'y a que 2 types de gamètes sur les 4 possibles. Hill et Weir (1994) proposent d'utiliser le carré de D pour éliminer le signe arbitraire qu'on a introduit pour exprimer la valeur de D.

Les valeurs de D ainsi définies sont très fortement dépendantes des fréquences alléliques. Pour essayer d'obtenir des mesures indépendantes de ces fréquences, Lewontin (1964) a proposé une mesure alternative : le déséquilibre standard ou normalisé D', exprimant D par rapport à la valeur maximale (Dmax) que peut atteindre D pour un lot de fréquences alléliques données.

$$D'_{A,B} = \frac{D_{A,B}}{D \max_{A,B}} \quad \text{avec} \quad \begin{cases} D \max_{A,B} = \min\{p_A \cdot p_B, p_a \cdot p_b\} & D < 0 \\ D \max_{A,B} = \min\{p_A \cdot p_b, p_a \cdot p_B\} & \text{si } D > 0 \end{cases}$$

Le coefficient D' varie de -1 à +1 quelles que soient les valeurs de fréquences alléliques et il est égal à zéro si les locus sont en équilibre de liaison. La valeur de D' est très affecté par une échantillon de petite taille et par l'absence d'une ou de plusieurs combinaisons alléliques.

Etant donné que le DL correspond à une association non aléatoire entre allèles, on utilise aussi parfois une mesure de **coefficient de corrélation (r ou r2)** entre allèles, définie par :

$$r = \frac{D}{\sqrt{p_A \cdot p_a \cdot p_B \cdot p_b}} \quad (\text{Hill et Robertson, 1968})$$

Ce coefficient de corrélation est lui aussi très sensible aux fréquences alléliques.

Dans le cas où on considère plusieurs marqueurs multialléliques, il est possible de généraliser l'expression du D' comme :

$$D' = \sum_{i=1}^k \sum_{j=1}^l p_i q_j |D'_{ij}| \quad (\text{Hedrick, 1987})$$

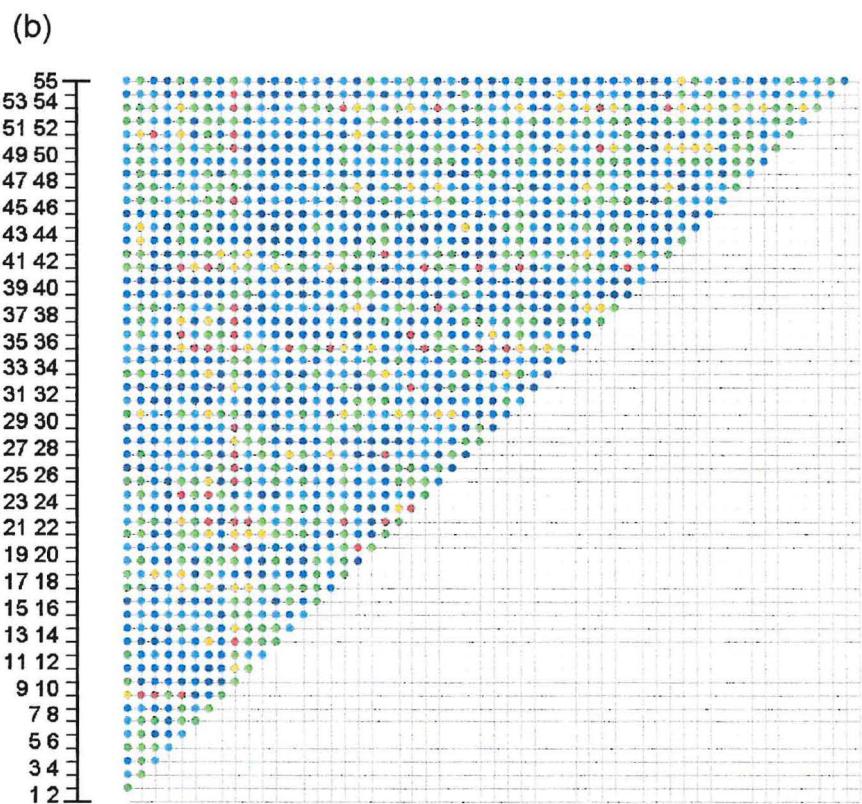
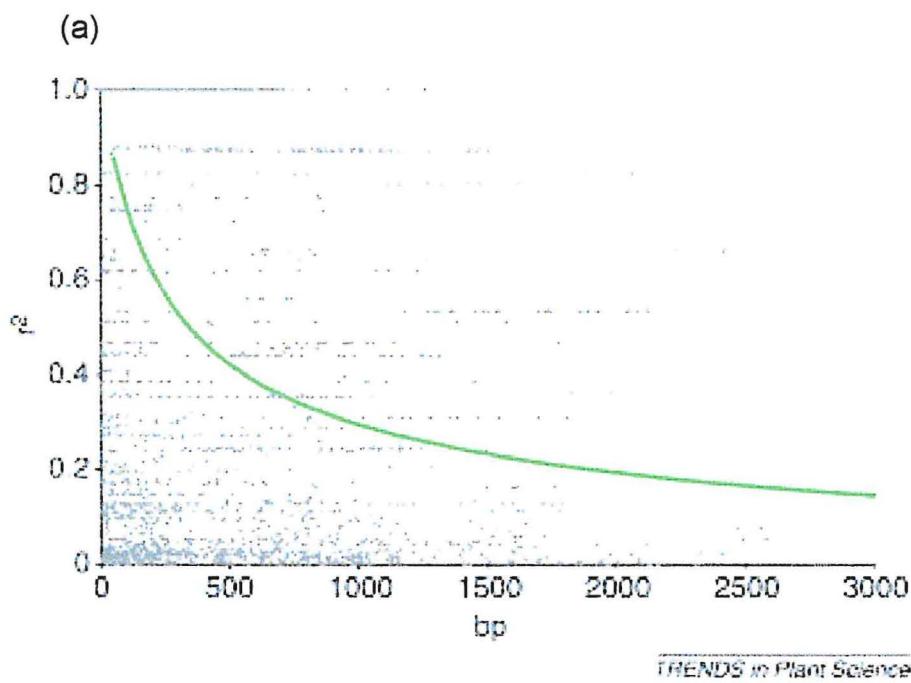


Figure 6. Représentation graphique de l’extension du DL a) Représentation du DL en fonction de la distance (physique ou génétique) (Neale et Savolainen, 2004); b) Matrice du DL au long d’un chromosome (Dechesne, 2002)

où $|D'_{ij}|$ est la valeur absolue du coefficient normalisé du DL de Lewontin (1964), k et l correspondent au nombre d'allèles à chaque locus, p_i est la fréquence de l'allèle i au premier locus et q_j la fréquence allélique de l'allèle j au deuxième locus.

Toutes ces mesures sont symétriques car elles ne permettent pas de connaître quel allèle d'un locus est associé de façon préférentielle à un allèle d'un autre locus.

L'hypothèse nulle d'indépendance entre les allèles aux paires de locus considérés peut être testée par un test du χ^2 sur des tables de contingence 2 x 2 dans le cas où les chromosomes simples ont été échantillonnés, ou lorsque les données familiales permettent d'inférer les haplotypes à partir des génotypes.

Sous l'hypothèse nulle que les locus sont en équilibre de liaison et que D vaut 0, on utilise la statistique :

$$\chi^2 = \frac{2N}{p_A \cdot p_a \cdot p_B \cdot p_b} \quad \text{avec } 2N \text{ nombre de gamètes}$$

qui soit une loi du χ^2 à 1 ddl (Hill, 1974).

Une probabilité inférieure à 5% indique qu'il y a une non indépendance entre les allèles des 2 locus, c'est-à-dire une association préférentielle entre certaines combinaisons alléliques. Cette hypothèse d'indépendance peut être aussi testée par un test exact de Fisher sur ces tables de contingence (Fisher, 1935).

Deux méthodes ont été utilisées pour visualiser l'extension du DL entre un grand nombre de locus: 1) Un graphique représentant le DL en fonction de la distance (physique ou génétique) qui permet d'évaluer la taille du fragment en DL et 2) Une matrice du DL qui est efficace pour visualiser le DL entre locus tout au long d'un chromosome (Figure 6).

3.2. Mesure du déséquilibre de liaison à partir des données génotypiques.

L'estimation la plus précise du DL nécessite la connaissance directe des fréquences des quatre haplotypes possibles pour deux locus bialléliques. Chez un organisme diploïde, sauf s'il est homozygote, il n'est pas facile d'inférer les haplotypes à partir de données génotypiques. Par exemple, pour un génotype AABB, l'haplotype est AB. Pour un

génotype AaBB, les haplotypes sont AB et aB avec des fréquences $\frac{1}{2}$ et $\frac{1}{2}$ pour chacun. Mais pour un génotype AaBb, les haplotypes peuvent être AB/ab ou Ab/aB. L'analyse des pedigrees des individus (Eaves *et al.*, 2001) par exemple, peut aider à résoudre ce problème. Cependant, ceci n'est pas toujours possible dans le cas des populations très hétérozygotes. On recourt alors à des méthodes statistiques d'estimation des haplotypes.

Une des méthodes les plus courantes utilisées est la méthode de maximum de vraisemblance mis en œuvre via l'algorithme EM (« Expectation-Maximizing ») (Hill, 1974 ; Excoffier et Slatkin, 1995). EM est un algorithme d'estimation de vraisemblance par itération successive de deux étapes : E (calcul d'Esperance) et M (Maximisation). Cet algorithme permet donc, de trouver les fréquences haplotypiques qui pourraient les mieux expliquer les données génotypiques. À partir d'une estimation initiale des fréquences gamétiques (p_{ij}), l'algorithme estime les fréquences gamétiques les plus vraisemblables aux données observées à une valeur L conformément à la supposition de la panmixie.

$$L \propto \prod (c_{hijk} * p_{hijk})^{n_{hijk}}$$

où p_{hijk} représente la probabilité que deux **gamètes** sélectionnés au hasard produisent le génotype $A_hA_iB_jB_k$; $c=1$ si $h=i$ et $j=k$ (pour les double homozygotes), 4 si $h \neq i$ et $j \neq k$ (double hétérozygotes) et 2 pour les combinaisons homozygotes-heterozygotes ; et n_{hijk} le nombre d'individus observés. Des itérations successives permettent d'arriver à la valeur maximale de L. Un inconvénient de cette méthode est que différentes fréquences alléliques initiales amènent à estimer différentes fréquences haplotypiques (Excoffier et Slatkin, 1995).

Une autre approche a été développée par Clark (1990). Cette méthode essaie d'établir le nombre minimum d'haplotypes possibles qui expliquent les génotypes observés dans la population. Des haplotypes facilement repérés sont déterminés à partir des individus homozygotes. Des échantillons plus complexes sont graduellement comparés avec les haplotypes initiaux. Les différences trouvées permettent alors d'établir à chaque fois les haplotypes possibles. Cette méthode est simple et efficace. Cependant, il est nécessaire d'avoir un nombre suffisant d'individus homozygotes et les fréquences des haplotypes estimés ne sont pas toujours exactes. En particulier dans le cas où un haplotype estimé est le résultat de la recombinaison entre deux vrais haplotypes. Cette approche est utilisée par le logiciel HAPINFERX

Tableau 4. Les probabilités génotypiques, dans la notation de Weir et Cockerham (1989)

	BB	Bb	bb	Total
AA	P^{AB}_{AB}	$2P^{AB}_{Ab}$	P^{Ab}_{Ab}	P^A_A
Aa	$2P^{AB}_{aB}$	$2P^{AB}_{ab} + 2P^{Ab}_{aB}$	$2P^{Ab}_{ab}$	$2P^A_a$
Aa	P^{aB}_{aB}	$2P^{aB}_{ab}$	P^{ab}_{ab}	P^a_a
Total	P^B_B	$2P^B_b$	P^b_b	1

Stephens *et al.* (2001) ont proposé une approche bayésienne d'estimation des haplotypes qui a été exploitée dans le logiciel PHASE disponible sur Internet (<http://www.stat.washington.edu/stephens/software.html>). Cette méthode permet d'obtenir à partir d'une simulation de Monte Carlo utilisant des chaînes de Markov (MCMC), un échantillon approximatif de la distribution postérieure des haplotypes (H) sachant les génotypes (G), c'est-à-dire une probabilité $\text{Pr}(H|G)$. L'algorithme commence avec une valeur initiale choisie au hasard pour H ($H^{(0)}$). Il choisit alors un individu au hasard et estime les haplotypes de chaque individu conformément à la supposition que tous les autres haplotypes ont été correctement reconstruits. Ces simulations sont répétées plusieurs fois jusqu'à que la convergence des résultats soit atteinte.

Toutes ces méthodes supposent que la population est en équilibre de Hardy-Weinberg, hypothèse qui devient leur inconvénient principal.

Pour des situations où on ne peut pas estimer les fréquences alléliques, Weir (1979) et Weir et Cockerham (1989) ont défini d'autres mesures, beaucoup plus complexes du déséquilibre de liaison à deux locus, appelées **déséquilibres composites (génotypiques)**, incluant des déséquilibres di-, tri-, quadrigéniques tenant compte du déséquilibre de liaison d'un allèle à un locus avec les différents allèles d'un autre locus.

Le déséquilibre de liaison est mesuré à partir d'un coefficient connu comme le coefficient de déséquilibre non gamétique digénique.

$$D_{A/B} = p_{A/B} - p_A p_B,$$

où le tirai ndique que deux allèles se trouvent sur des chromosomes différents.

Les probabilités génotypiques, dans la notation de Weir et Cockerham (1989) sont rapportées dans le tableau 4. Conformément aux probabilités pour chaque locus on peut dire que les probabilités digéniques sont :

$$p_{AB} = P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}$$

et

$$p_{A/B} = P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}$$

Toutes les deux dépendent de la probabilité des doubles hétérozygotes. Cette probabilité ne peut pas se déterminer d'une manière séparée si on ne connaît pas la phase des doubles hétérozygotes.

Pour des cas où la phase ne peut pas être déterminée et où les individus ne se croisent pas au hasard, Weir (1979) et Weir et Cockerham (1979) introduisent le concept du coefficient composé de déséquilibre de liaison

$$\Delta_{AB} = D_{AB} + D_{A/B} = p_{AB} + p_{A/B} - 2p_A p_B.$$

qui représente

$$\begin{aligned}\Delta_{AB} = & 2P_{AB}^{AB} + 2P_{Ab}^{AB} + 2P_{aB}^{AB} \\ & + (P_{aB}^{Ab} + P_{ab}^{AB}) - 2p_A p_B.\end{aligned}$$

la valeur de Δ_{AB} est comprise entre -0.5 et 0.5. Δ_{AB} permet d'estimer le DL à partir des données génotypiques.

3.3 Evolution du Déséquilibre de liaison

De nombreux facteurs peuvent créer du DL entre allèles de locus non liés. Les plus importants sont :

- **La dérive génétique**

L'une des forces principales engendrant du déséquilibre de liaison est la dérive génétique, ce qui affecte principalement les populations de taille restreinte (Hill et Robertson, 1968). Alors que dans des populations de taille « infinie » les fréquences alléliques sont stables au cours des générations en l'absence de sélection et de mutation, les fréquences alléliques varient aléatoirement dans des populations de taille finie. Cela est du à la variabilité du tirage aléatoire des allèles d'une génération à l'autre. En effet, le tirage aléatoire des gamètes qui vont former une nouvelle génération peut contenir un excès d'une combinaison particulière, créant un déséquilibre de liaison.

On s'attend théoriquement à

$$E(r^2) = \frac{1}{(1 + 4.N_e.c)} \quad (\text{Hill et Robertson, 1968})$$

où r est le coefficient de corrélation entre allèles. Si la taille efficace N_e et le taux de recombinaison c sont petits, r^2 tend vers 1, soit vers la valeur maximale du DL.

Tableau 5. Effet de la mutation sur le déséquilibre de liaison

	Fréquence de A_1B_1	Fréquence de A_1B_2	Fréquence de A_2B_1	Fréquence de A_2B_2	D
Avant mutation	q_1	q_2	--	--	0
Mutation sur B_1	q_1-p_2	q_2	p_2	--	$-q_2/2N$
Mutation sur B_2	q_1	q_2-p_2	--	p_2	$q_1/2N$

- **La migration et le mélange récent de populations**

Des flux de gènes entre populations de composition génétique différente, c'est-à-dire, comprenant des fréquences alléliques différentes, généreront du DL. Nei et Li (1973) montrent que si deux populations isolées commencent à échanger des gènes, le déséquilibre gamétique peut augmenter temporairement dans chaque population, même pendant beaucoup de générations. Ils montrent aussi que l'approche à l'équilibre (soit à $D=0$) est retardée si le taux de migration est faible ;

Ce facteur de décroissance est égal à :

$$[1 - \frac{2(m_1 + m_2)}{(m_1 + m_2)^2}],$$

où $m_1 = M / N_1$ et $m_2 = M / N_2$ pour deux populations, M étant égal au nombre de migrants échangés, et N_1 et N_2 aux effectifs des populations.

Par contre, si le taux de migration est élevé par rapport au taux de recombinaison, les changements de DL seront identiques à ceux d'une population panmictique.

Le cas extrême se présente si la population est constituée de deux populations ayant des fréquences gamétiques différentes à plusieurs locus. Il aura alors un fort déséquilibre de liaison entre ces locus.

- **La mutation**

Les mutations sont la base du polymorphisme. Lorsqu'un nouvel allèle apparaît par mutation à un locus monomorphe donné, une minorité d'individus portera cette mutation dans la population. Ceci va entraîner un déséquilibre de liaison avec les locus polymorphes voisins. Si, par exemple un allèle A1 mute en A2, la fréquence des gamètes A_2B_1 ou A_2B_2 est $p_2 = 1/2N$ où N est l'effectif de la population et $D = -q_2/2N$ (Tableau 5). La taille de la population est importante. Si la population est grande, le déséquilibre engendré par la mutation sera faible car une minorité d'individus seront porteurs de la nouvelle combinaison génique.

Si cette mutation entraîne un avantage sélectif, les nouveaux types gamétiques pourront augmenter par sélection générant du DL.

- **La sélection.**

La sélection peut engendrer et maintenir en permanence du déséquilibre de liaison entre locus malgré l'équilibre des fréquences géniques (Lewontin, 1964). Elle peut affecter le DL grâce à trois processus :

- si une association présente chez un individu lui confère un fort avantage sélectif favorisant par exemple des haplotypes particuliers et créant un DL entre les locus porteurs de ces allèles (sélection épistatique). En cas de forte épistasie, un déséquilibre peut même être maintenu entre gènes non liés (indépendants).
- la fixation d'un allèle, il n'y a plus de polymorphisme à ce locus et le DL disparaît.
- « l'auto-stop » qui permet qu'un allèle neutre ou quasi neutre, fortement lié à un locus qui a un avantage sélectif, change de fréquence sous l'effet de la sélection, si le déséquilibre de liaison initial est important entre les deux locus (Kojima et Schaffer, 1967).

- **L'effet de fondation ou une petite taille de la population due aux goulets d'étranglement**

L'effet de fondation et l'existence d'un goulot d'étranglement entraînent l'apparition d'un DL. Ces deux processus supposent nécessairement une étape de faible effectif des populations (Reich *et al.*, 2001). Lors d'un goulot d'étranglement, seules certaines combinaisons sont maintenues dans la population résultante. Ces facteurs sont d'une grande importance chez certaines espèces végétales où la domestication a induit la constitution de populations avec un fort DL initial.

- **Le système de reproduction**

D'une manière générale, on s'attend à un déséquilibre de liaison plus fort chez les plantes autogames que chez les plantes allogames. La vitesse de disparition du DL diminue chez les plantes autogames car la recombinaison n'est pas efficace lorsqu'elle se fait chez des individus homozygotes. Ce phénomène est facilement visualisé si on regarde la relation entre le taux de recombinaison effective (c) et le facteur d'autofécondation (s) :

$$c = \frac{1 - s}{2 - s}$$

Tableau 6. Effet de la recombinaison sur le déséquilibre de liaison

Génotypes possibles des individus	Fréquences génotypiques (sous hyp. HW)	Fréquence du gamète AB			
AB/AB	p_{AB}^2	p_{AB}^2	--	--	--
AB/Ab	$2 \cdot p_{AB} \cdot p_{Ab}$	$p_{AB} \cdot p_{Ab}$	$p_{AB} \cdot p_{Ab}$	--	--
Ab/Ab	p_{Ab}^2	--	p_{Ab}^2	--	--
AB/aB	$2 \cdot p_{AB} \cdot p_{aB}$	$p_{AB} \cdot p_{aB}$	--	$p_{AB} \cdot p_{aB}$	--
AB/ab	$2 \cdot p_{AB} \cdot p_{ab}$	$(1-r) \cdot p_{AB} \cdot p_{ab}$	$r \cdot p_{AB} \cdot p_{ab}$	$r \cdot p_{AB} \cdot p_{ab}$	$(1-r) \cdot p_{AB} \cdot p_{ab}$
Ab/aB	$2 \cdot p_{Ab} \cdot p_{aB}$	$r \cdot p_{Ab} \cdot p_{aB}$	$(1-r) \cdot p_{Ab} \cdot p_{aB}$	$(1-r) \cdot p_{Ab} \cdot p_{aB}$	$r \cdot p_{Ab} \cdot p_{aB}$
Ab/ab	$2 \cdot p_{Ab} \cdot p_{ab}$	--	$p_{Ab} \cdot p_{ab}$	--	$p_{Ab} \cdot p_{ab}$
aB/aB	p_{aB}^2	--	--	p_{aB}	--
aB/ab	$2 \cdot p_{aB} \cdot p_{ab}$	--	--	$p_{aB} \cdot p_{ab}$	$p_{aB} \cdot p_{ab}$
ab/ab	p_{ab}^2	--	--	--	p_{ab}^2
		$p'_{AB} = p_{AB} - r \cdot D^{(0)}$	$p'_{Ab} = p_{Ab} + r \cdot D^{(0)}$	$p'_{aB} = p_{aB} + r \cdot D^{(0)}$	$p'_{ab} = p_{ab} - r \cdot D^{(0)}$

Autrement, la **recombinaison** est la force qui permet la diminution du DL. Si la probabilité de recombinaison est r , c'est-à-dire la probabilité qu'un haplotype soit de type recombinant, cette probabilité dépendra de la position relative de deux locus dans le génome. Si les deux locus sont sur des chromosomes différents, la probabilité de recombinaison est égale à 0.5. Les quatre types d'haplotypes possibles sont produits à fréquences égales et sont transmis en accord avec les lois de Mendel.

Si les deux locus sont situés sur le même chromosome, la probabilité de recombinaison dépend de la distance entre ces deux locus. Les haplotypes pour un individu de génotype AB/ab seront :

$$AB (1-r)/2$$

$$ab (1-r)/2$$

$$Ab r/2$$

$$aB r/2$$

Le taux du déséquilibre de liaison diminue de ce facteur r après chaque génération. La fréquence de l'haplotype AB après une génération est la somme des haplotypes AB résultants d'un méiose. Ces haplotypes peuvent être un gamète de type parental, c'est-à-dire produit sans recombinaison entre les deux locus ou une gamète produit par recombinaison (probabilité r) (Tableau 6) De cette façon, la fréquence de l'haplotype AB à la génération t est égale à :

$$D_t = D_0 (1 - r)^t$$

où D_0 est le DL initial dans la population. D tend alors vers 0 à une vitesse qui dépend de r . Plus les locus sont distants (r est élevé), plus on atteint 0 rapidement. Pour des locus indépendants ($r = 0.5$), le déséquilibre de liaison décroît de moitié à chaque génération et devient rapidement négligeable. Ceci bien entendu en l'absence de forces de sélection et dans une population idéale en équilibre de Hardy-Weinberg. La figure 7 présente quelques exemples de la diminution du DL pour différents taux de recombinaison. Cependant, le taux de recombinaison n'est pas constant tout au long du génome. Chez l'homme, des points « chauds » de recombinaison (aux taux de recombinaison élevés) ont été observés (Goldstein, 2001 ; Shifman *et al.*, 2003).

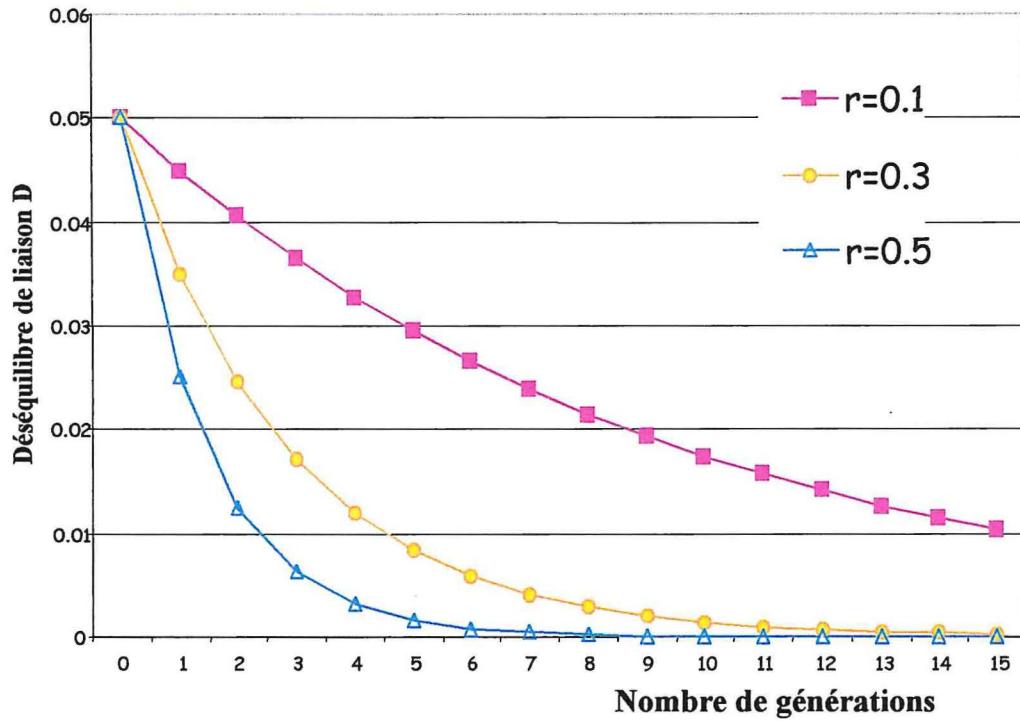


Figure 7. Décroissance du déséquilibre de liaison en fonction du temps

3.4. Effet de la structure de la population sur l'estimation du DL

Une population est structurée si les fréquences d'un caractère quelconque varient entre les sous-populations. Si la population est structurée, le DL tend à augmenter sur l'ensemble du génome, et le DL observé mélange un effet de liaison physique sur le génome et des effets d'association de locus indépendants. Pritchard et Prezworski (2001) ont développé une méthode qui prend en compte la structure de la population pour les études d'associations et qui est disponible sur Internet (<http://www.stats.ox.ac.uk/~pritch/home.html>). Une méthode de Monte-Carlo par chaîne de Markov (MCMC) permet de calculer des estimateurs empiriques à partir de réalisations aléatoires simulées du processus afin d'assigner les individus d'une population à K sous-populations. On suppose que chacune des populations est homogène et en équilibre de Hardy-Weinberg et qu'il y a un équilibre de liaison entre locus indépendants ; si des déséquilibres apparaissent, ils sont liés à des déséquilibres alléliques entre populations.

Si la structure de la population n'est pas contrôlée, il est très probable d'obtenir des faux positifs. Cet effet a été déjà observé chez les humains. Un exemple classique est le travail de Knowler *et al.* (1988) où ces auteurs ont trouvé de fausses associations entre un haplotype particulier du locus de l'immunoglobuline G et l'incidence de diabète chez certaines tribus américaines dues à la structure de la population. Chez les plantes, Thomsberry *et al.* (2001) ont trouvé une diminution des faux positifs en contrôlant la structure de la population.

Une autre méthode a été proposée par Deu et Glaszmann (2004). Cette méthode consiste à éliminer des individus trop éloignés qui contribuent le plus au DL global de manière à éliminer les traces de structure. L'arbre construit sur la base des coefficients de dissimilarités entre les individus étudiés prendra une forme étoilée.

Chez le riz, Garris *et al.* (2003), en étudiant la région autour du locus *Xa5* qui confère la résistance au *Xanthosoma oryzae* pv. *oryzae* ont trouvé des différences génétiques entre les groupes *indica* et *aus-boro*. Ces différences semblent avoir une influence importante sur les valeurs du DL estimées. Les auteurs mettent en avant la nécessité de prendre en compte la structure de la population au moment de la mise en place d'une étude de DL pour cette espèce.

3.5 Le déséquilibre de liaison chez les plantes

Des centaines d'études de QTL chez les plantes ont été publiés, toutes basées sur des descendances en ségrégation (Par exemple, des descendances F2, des populations issues de retrocroisements ou backcross, des haploïdes doubles ou des lignées recombinantes). En fait, une descendance en ségrégation est une population structurée en déséquilibre de liaison après un seul ou peu d'événements de recombinaison. On exploite alors, le DL local créé après le croisement entre caractère et marqueur dû au fait de leur proximité physique. Cependant, il n'est pas forcément facile de créer des descendances en ségrégation, en particulier chez certaines plantes arborées. Par ailleurs, la résolution de cette méthode « classique » est limitée par le faible nombre d'événements de recombinaison. En général, la détection des QTLs basée sur des populations en ségrégation sont détectés avec des intervalles de confiance de 10 à 30 cM sur le génome (Remington *et al.*, 2001). Un segment de 20 cM peut porter des centaines de gènes rendant l'identification du gène concerné difficile.

La cartographie basée sur le DL (« association mapping ») utilise des populations naturelles ou en sélection pour détecter les associations. En théorie, comme le nombre de recombinaisons accumulées dans l'histoire d'une population naturelle est plus élevé que celles observées dans une étude « classique » de détection de QTL basée sur des descendances contrôlées, la résolution obtenue est meilleure. De façon générale, la cartographie par DL se fonde sur l'hypothèse que les régions adjacentes à un gène d'intérêt sont transmises d'une génération à l'autre en même temps que ce gène. Ces régions peuvent être identifiées par les haplotypes qu'ils contiennent. Donc, le DL a une valeur prédictive car il permet de prévoir en observant le polymorphisme d'un marqueur moléculaire, celui d'un gène particulier que l'on sait proche de ce marqueur. La longueur de l'haplotype portant l'allèle associé à un caractère d'intérêt décroît à chaque génération. En effet, comme les individus échantillonnés ne sont pas nécessairement très proches et que la population a peut être, subi plusieurs générations de recombinaison, seuls les marqueurs ayant un fort DL s'associeront avec le caractère.

Le déséquilibre de liaison a déjà été largement utilisé chez les humains. Pritchard et Przeworski (2001) ont présenté une revue sur l'utilisation du DL. Cette approche est particulièrement utilisée pour la cartographie de maladies complexes selon l'hypothèse

que le DL entre locus est corrélé négativement à leur distance physique sur le chromosome (Kruglyak, 1999 ; Jorde, 2000 ; Horikawa *et al.*, 2000).

L'étude et l'exploitation du DL sont plus récentes chez les plantes. Malgré la relative nouveauté de ce sujet, les travaux scientifiques concernant ce thème se multiplient.

Un des aspects du DL le plus étudié est la détermination de la taille de fragment du génome qui est en déséquilibre de liaison. La cartographie basée sur le DL nécessite une carte avec une densité de marqueurs du même ordre ou inférieure à celle de l'étendue du DL dans la population étudiée. De façon générale, le déséquilibre de liaison, au niveau d'une population, s'étend sur de petites distances. Ceci est dû à l'érosion de l'association entre marqueurs résultant de la recombinaison sur des générations successives. Chez les humains, ces distances varient entre 3 cM (Gordon *et al.*, 2000) et quelques paires de bases sur certains régions chromosomiques (Taillon-Miller *et al.*, 2000). Ces différences dépendent des populations étudiées, du type de marqueur moléculaire et de la densité de marqueurs sur la carte génétique. Chez les bovins, le DL s'étend jusqu'au 10 cM (Tenesa *et al.*, 2003).

Les études réalisées chez les végétaux montrent que ces distances sont aussi très variables et dépendant de l'espèce et de la population choisie. D'une manière générale, on s'attend à un DL plus fort chez les plantes autogames que chez les plantes allogames. Ainsi, elle est d'environ 100 Kb chez le riz (Garris *et al.*, 2003), 500 kb chez le soja (Hyten *et al.*, 2004), entre 2 et 4 cM chez le sorgho (Deu et Glaszmann, 2004), 3 cM chez *Beta vulgaris* spp. maritima (Kraft *et al.*, 2000 ; Hansen *et al.*, 2001), entre 10 et 20 cM chez le orge (Stracke *et al.*, 2003) et le blé tétraploïde (Maccaferry *et al.*, 2004). Par contre, chez le maïs, une plante allogame qui a suivi une longue histoire de domestication, Remington *et al.* (2001) ont trouvé des associations entre marqueurs situés à 1500 pb en analysant un échantillon de lignées qui sont couramment utilisées par certains organismes publics dans leurs programmes d'amélioration. Dans le cas de groupes de maïs très variables et en étudiant 21 locus situés sur le chromosome 1, Tenaillon *et al.* (2001) estiment que cette distance ne s'étend que sur 200 pb. Ching *et al.* (2002) expliquent que l'existence du DL entre locus SNP séparés par moins de 500 pb est due aux goulots d'étranglement créés lors de la sélection des lignées élite étudiées. Un autre exemple est donné chez la canne à sucre, espèce polyploïde à propagation asexuée, pour laquelle le DL s'étend sur plusieurs dizaines de cM (Jannoo *et al.*, 1999). Le goulot d'étranglement à

l'origine de la majorité des variétés actuelles de canne à sucre et son taux de polyploidie peuvent être des explications satisfaisantes pour cette grande distance en DL.

Etablir cette distance permet de choisir le nombre et la densité de marqueurs nécessaires pour chercher des associations entre marqueurs et caractères d'intérêt grâce aux études d'associations. Chez l'homme, ce maillage du génome est extrêmement fin et il est d'environ un marqueur tous les 3 kb (Kruglyak, 1999). Chez les végétaux, en raison des goulots d'étranglement, des effets de fondation, et de la sélection par l'homme exercés lors de la domestication, le DL s'étend sur des régions beaucoup plus étendues et est alors détectable avec la densité de marqueurs qui existe déjà sur certaines cartes disponibles. Tenaillon *et al.* (2001) ont estimé que un SNP chaque 100 – 200 pb est nécessaire pour avoir des chances de mettre en évidence ces associations chez le maïs ; par contre chez *Arabidopsis* cette densité diminue à un marqueur chaque 50 kb (Nordborg *et al.*, 2002).

Le DL au sein de populations naturelles ou sélectionnées (par exemple, des collections de ressources génétiques) a été utilisé pour la mise en évidence d'associations entre les allèles des marqueurs et ceux de gènes d'intérêt agronomique. Jusqu'aujourd'hui, quelques travaux ont utilisé cette approche. La première étude d'association a été menée par Thornsberry *et al.* (2001). Ils ont trouvé que le gène de nanisme *Dwarf8* (d8) était associé au temps de floraison et à la hauteur des plantes chez 92 lignées de maïs. Chez *Arabidopsis thaliana*, une étude sur la variation des haplotypes dans la région chromosomique qui correspond au récepteur du photopériodisme CRYPTOCHROME2 et sur la variation des temps de floraison montre que ces « haplogroups » se limitent à une région de 65 Kb comportant 6 gènes autour du gène CRY2. Une simple substitution d'une sérine au CRY2 semble être responsable de la variation des temps de floraison (Olsen *et al.*, 2004). D'autre part, Simko *et al.* (2004), ont repéré un marqueur microsatellite très lié au gène StVel qui montre une association significative avec un QTL impliqué dans la résistance à *Verticillium dahliae* chez 137 cultivars de pomme de terre tétraploïde.

Le rendement et sa stabilité ont été associés aux marqueurs AFLP chez l'orge (Kraakman *et al.*, 2004). Cette étude a été réalisée sur une collection de ressources génétiques qui comporte 146 cultivars modernes couramment utilisés dans les programmes d'amélioration européenne. Comme le montre ce dernier travail, la cartographie DL envisage la possibilité d'exploiter les données agro morphologiques des banques de ressources génétiques qui ont été collectées dans certains cas, depuis des

CHAPITRE I. Eléments de bibliographie

dizaines d'années. L'« International Cocoa Germplasm Database » (ICGD) (Wadsworth et Harwood, 2000) contient les données de plus de 14000 clones du cacaoyer. Des études d'associations devraient permettre de valoriser ces données et de révéler l'existence des régions chromosomiques impliquées dans la variation des caractères d'intérêt. En particulier ceux des variétés de type Criollo Moderne/Trinitario. Un des objectifs de cette thèse est donc de montrer la faisabilité de ce type d'étude chez le cacaoyer.

CHAPITRE II : CARTOGRAPHIE GÉNÉTIQUE DU CACAOYER

CHAPITRE II

Cartographie génétique du cacaoyer : développement et cartographie de 201 nouveaux marqueurs microsatellite

Malgré le grand nombre de marqueurs déjà cartographiés, une grande partie d'entre eux sont des marqueurs de type dominants. Une carte saturée réalisée à l'aide uniquement de marqueurs codominants et polymorphes constituera un outil important pour mieux comprendre la structure de la diversité génétique. L'intérêt des marqueurs SSR est leur niveau de polymorphisme élevé et la possibilité qu'ils offrent de trouver des allèles spécifiques de différents groupes ou origines génétiques. Pour nos études d'association il était donc essentiel de produire et de cartographier un grand nombre de SSR bien repartis sur tout le génome et qui nous permettent d'identifier les allèles spécifiques des Criollo, groupe génétique particulièrement intéressant pour les caractères de qualité du chocolat.

Dans cet objectif, nous avons cartographié un set de 201 marqueurs microsatellites de cacao. Cette étude a abouti à l'obtention d'une nouvelle carte génétique de référence. L'article résultant qui a été publié dans *Theoretical and Applied Genetic* est présenté ci-après.

T. Pugh · O. Fouet · A. M. Risterucci · P. Brottier ·
 M. Abouladze · C. Deletrez · B. Courtois · D. Clement ·
 P. Larmande · J. A. K. N'Goran · C. Lanaud

A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers

Received: 24 June 2003 / Accepted: 10 November 2003 / Published online: 4 February 2004
 © Springer-Verlag 2004

Abstract A linkage map of cacao based on codominant markers has been constructed by integrating 201 new simple sequence repeats (SSR) developed in this study with a number of isoenzymes, restriction fragment length polymorphisms (RFLP), microsatellite markers and resistance and defence gene analogs (Rgenes-RFLP) previously mapped in cacao. A genomic library enriched for (GA)_n and (CA)_n was constructed, and 201 new microsatellite loci were mapped on 135 individuals from the same mapping population used to establish the first reference maps. This progeny resulted from a cross between two heterozygous cacao clones: an Upper-Amazon Forastero (UPA 402) and a Trinitario (UF 676). The new map contains 465 markers (268 SSRs, 176 RFLPs, five isoenzymes and 16 Rgenes-RFLP) arranged in ten linkage groups corresponding to the haploid chromosome number of cacao. Its length is 782.8 cM, with an average interval distance between markers of 1.7 cM. The new microsatellite markers were distributed throughout all linkage groups of the map, but their distribution was not random. The length of the map established with only SSRs was 769.6 cM, representing 94.8% of the total map. The current level of genome coverage is approximately one microsatellite every 3 cM. This new reference map

Communicated by H. Nybom

T. Pugh (✉) · O. Fouet · A. M. Risterucci · M. Abouladze · C. Deletrez · B. Courtois · D. Clement · P. Larmande · C. Lanaud
 UMR 1096, CIRAD-BIOTROP,
 TA 40/03, 34398 Montpellier Cedex 5, France
 e-mail: pugh@cirad.fr
 Tel.: +33-4-67655831
 Fax: +33-4-67615605

T. Pugh
 UCV-FAGRO, Av El Limón, 2101 Maracay, Venezuela

P. Brottier
 GENOSCOPE, Centre National de Séquençage,
 2 rue Gaston Crémieux, CP 5706, 91057 Evry, France

J. A. K. N'Goran
 CNRA, Centre National de Recherches Agronomiques,
 BP 582 Abidjan, Ivory Coast

provides a set of useful markers that is transferable across different mapping populations and will allow the identification and comparison of the most important regions involved in the variation of the traits of interest and the development of marker-assisted selection strategies.

Introduction

Genetic mapping is a basic tool of genomic research. Molecular linkage maps provide information about the organization of the genome and may be used for genetic studies and breeding applications. Several linkage maps have already been published for *Theobroma cacao* L. (Lanaud et al. 1995; Crouzillat et al. 1996; Risterucci et al. 2000). These maps were based on codominant markers such as restriction fragment length polymorphisms (RFLPs), isoenzymes and a small number of simple sequence repeats (SSRs) and, in some of the maps, associated with dominant amplified fragment length polymorphisms (AFLP) and random amplified polymorphic DNA (RAPD) markers. They have been used to locate quantitative trait loci (QTLs) affecting traits of interest, such as disease resistance and yield factors (Crouzillat et al. 2000; Flament et al. 2001; Clement et al. 2003; Queiroz et al. 2003; Risterucci et al. 2003).

Dominant markers require a relatively small investment with respect to time and cost and are commonly used to increase the density of linkage maps due to their multilocus properties. Nevertheless, the fact that they are dominant induces a loss of information that can be detrimental for a heterozygous crop. A linkage map for cacao based on codominant markers could provide more information. PCR-based codominant markers like microsatellite markers can easily be transferred to other populations and shared between research laboratories.

Microsatellites, or SSRs, are present in the majority of eukaryotic genomes and consist of simple, short tandemly repeated di- to penta-nucleotide sequence motifs (Beckman and Soller 1990). The allelic variation in microsatellite loci can easily be detected by PCR using specific

flanking primers. Polymorphism based on variation in the number of repeated motifs is probably due to slippage during DNA replication or unequal crossing-over (Levinson and Gutman 1987). Microsatellites have been widely used in many crop species due to their abundance, high degree of polymorphism, locus specificity, reproducibility, low amount of DNA required, suitability for multiplexing on automated systems and, above all, their codominant mode of inheritance. These characteristics make SSRs an attractive option for increasing the density of the cacao linkage map. Nevertheless, the development of microsatellite markers is generally time-consuming and costly. Several procedures have been developed to produce SSR-enriched gDNA libraries. Billote et al. (1999) described an easy method for developing microsatellite markers in tropical crops that has subsequently been used with success (Billote et al. 2001; Aranzana et al. 2002; Dirlewanger et al. 2002).

In the past few years, a new generation of linkage maps based on SSR markers has been constructed for several species—sorghum (Bhatramakki et al. 2000; Haussmann et al. 2002), wheat (Gupta et al. 2002), rice (Temnykh et al. 2000; McCouch et al. 2002), peach (Dettori et al. 2001), almond (Joobeur et al. 2000), apple (Liebhard et al. 2002) and *Prunus species* (Aranzana et al. 2003). In addition, microsatellite markers have been developed and mapped in cacao (Lanaud et al. 1999; Risterucci et al. 2000; Lanaud et al. 2004).

In the investigation reported here, we developed and mapped a new set of 201 cacao microsatellite markers onto the previous cacao linkage map.

Materials and methods

Production of microsatellite markers and primer design

A genomic library enriched for $(GA)_n$ and $(CA)_n$ was constructed from the cacao (*Theobroma cacao* L.) clone Catongo based on the procedure described by Billote et al. (1999). The enrichment step was pursued as described in Kijas et al. (1994) with some modifications. Genomic DNA was digested with *Rsa*I (Invitrogen, Carlsbad, Calif.), the fragments were ligated to 5'-end phosphorylated adaptors constituted with the following two following primers: *Rsa*21: CTCTTGCTTACCGCGTGGACTA and *Rsa*25: p-TAGTCCACCGCGTAAGCAAGAGCACA (Edwards et al. 1996). The ligated fragments were amplified with *Rsa*21 primer. Following hybridization with a 5'-biotinylated microsatellite oligoprobe $I_5(CT)_8$ and $I_5(GT)_8$, they were selected for the presence of microsatellites with Streptavidin MagneSphere Paramagnetic Particles (PMPs) (Promega, Madison, Wis.). The selected fragments were amplified with *Rsa*21, and the partial genomic library was then constructed by ligating the PCR products into a pGEM-T plasmid (Promega). Epicurian-coli XL1-Blue supercompetent cells (Stratagene, La Jolla, Calif.) were used for the transformation of the cloned DNA fragments. Subsequently, white transformant clones were transferred onto Hybond-N+ nylon membranes (Amersham, UK) and simultaneously hybridized using [^{32}P]-labelled microsatellite oligoprobes $(GA)_{15}$ and $(GT)_{15}$. Selected microsatellite-containing clones were sequenced by Genoscope (France) using its current sequencing protocol (Artiguenave et al. 2000).

Duplicates of sequenced SSRs were removed using SEQUENCHER 4.0 (Gene codes, Ann Arbor, Mich.). The primers flanking the microsatellite repeat sequences were designed using

the OLIGO 4.0 software (National Biosciences, USA). The main criteria for primer design were to produce well-matched primers that were 16–24 nucleotides long, had an average GC content ranging between 40% and 50% and an annealing temperature between 45°C and 55°C and were preferably G- or C-rich at the 3' end.

PCR amplification protocols and detection of polymorphism

PCR reactions were performed on a MJ Research PTC Thermal cycler (MJ Research, Waltham, Mass.) in 20- μ l volumes containing 10 ng of cacao DNA, 0.2 μ M of a 5'-endlabeled γ -[^{33}P]ATP forward primer, 0.2 μ M of reverse primer, 2 mM MgCl₂, 0.2 mM dNTP mix, 50 mM KCl, 10 mM Tris-HCl (pH 8.3) and 1 U *Taq* polymerase (Eurobio, France). The PCR profile was: an initial denaturation step at 94°C for 4 min, followed by 35 cycles of 30 s at 94°C, primer annealing at a specific annealing temperature (46°C or 51°C) for 1 min and 1 min at 72°C; this was completed with a final extension at 72°C for 5 min.

Twenty microliters of loading buffer (98% deionized formamide, 10 mM EDTA, bromophenol blue and xylene cyanol) was added to individual reactions. Samples were denatured at 94°C, and 5 μ l of each sample was subjected to electrophoresis at 55 W on 5% denaturing polyacrylamide gels containing 75 M urea in 0.5x TBE buffer (pH 8.0). The gels were dried and exposed for 48–72 h to X-ray film (Eastman Kodak). Polymorphism revealed by the newly synthesized primer pairs was tested using DNA from the parents of the mapping population.

Mapping population

The newly developed microsatellites were mapped on 135 individuals from the same mapping population as that used to establish the map described in Risterucci et al. (2000). This progeny resulted from a cross between two heterozygous cacao clones: an Upper-Amazon Forastero (UPA 402) and a Trinitario (UF 676). Because the parents were heterozygous, there were two possibilities for single-locus segregation—those which segregated in only one of the parents (1:1) (heterozygous for one parent and homozygous for the other parent), and those which segregated in both parents (1:2:1 or 1:1:1:1) (heterozygous in both parents). Codominant markers segregating according to the latter possibility can be used as genetic bridges for aligning the linkage information from each parental dataset to produce a consensus linkage map (Grattapaglia and Sederoff 1994).

Linkage map construction

The segregation of each microsatellite marker was tested with a chi-square test for goodness-of-fit to the expected Mendelian segregation ratio function of the parental configuration. The map was produced using JOINMAP version 3.0 (Van Ooijen and Voorrips 2001) by integrating the SSR loci reported in this paper, the RFLPs, SSRs, isoenzymes and resistance and defence gene analogues previously mapped in cacao by Risterucci et al. (2000) and Lanaud et al. (2004). JOINMAP software is able to combine data of several segregation types to construct an integrated map. When the parental phase is unknown, this software chooses the best option of association between coupling/repulsion phases of a set of markers based on the recombination frequencies observed within a linkage group. A LOD score of 5.0 was used to identify linkage groups. The Kosambi mapping function was used to convert recombination frequencies into map distances (Kosambi 1944).

Marker nomenclature

All loci are designated according to the nomenclature guidelines presented by Risterucci et al. (2000) and Lanaud et al. (2004): SSRs

were denoted as mTcCIRX where m corresponds to microsatellite, Tc to *Theobroma cacao*, CIR to CIRAD and X to the microsatellite number. RFLP probes were named cTcCIR, gTcCIR and rTcCIR corresponding to cDNA, genomic and isolated RAPD genomic fragments, respectively. TELX corresponds to telomeric markers; NX/X, PTX/X and PRx/X correspond to probes homologous to the N gene (Whitham et al. 1994), Pto gene (Martin et al 1994) and PR proteins (Kauffman et al. 1987), respectively. NX corresponds to genomic probes provided by Nestlé, and CX or CaX to microsatellite markers also provided by Nestlé.

Results

Microsatellite development

Of the 705 positive clones sequenced, 154 (22%) were redundant; 463 unique sequences (66%) were found to contain a microsatellite while 88 did not contain any microsatellite. We subsequently designed 387 primer pairs (63%) that flanked the microsatellite motifs; the remaining clones were not exploited because the SSRs were too short or too close to one sequence end.

Microsatellite amplification and polymorphism

Sequenced primer pairs flanking the microsatellite motifs were tested for amplification within the parents of the mapping population. Among the 387 primer pairs tested, 227 (59%) detected polymorphism between the two parents of the mapping population, with 203 (52%) providing a clear single-locus amplification and unambiguous profile. These latter primer pairs were used for the construction of the linkage map. All three types of repeats, as defined by Weber (1990), were found among the 203 SSRs. One hundred and fifty SSRs (74%) were derived from dinucleotide motifs with a perfect microsatellite stretch. The most frequent motifs were (CT)_n or (GA)_n (77%), as would be expected for a library based on selection for (GA)_n repeats, 21 (10%) were interrupted and 32 (16%) were compounds. Table 1 provides detailed information on the mapped markers developed in this study. The size of the amplified fragments was in agreement with that expected from sequenced data.

Linkage analysis and map construction

Ninety-one percent of all markers fitted the Mendelian ratio expected from the genotype of the parents. Only 42 of the markers (18 RFLPs, two isoenzymes, one Rgene and 21 SSRs) deviated significantly ($P < 0.05$) from the expected ratio; these included 20 newly developed microsatellite markers.

If all of the markers are taken into consideration, bridge loci (heterozygous in both parents) represented 26% of the scored patterns, 66% segregated for UF 676 only, and 8% segregated for UPA 402 only. When we constructed the two parent maps separately (data not

shown), good co-linearity between the two maps was observed, with only small modifications in the relative positions of closely linked markers, especially inversions—for example, mTcCIR18 and mTcCIR199 on linkage group (LG) 4. Genetic distances estimated between loci in the UF 676 or UPA 402 parent maps were generally of the same magnitude (e.g. 1.5 cM vs. 2.1 cM between mTcCIR188 and mTcCIR12, respectively or 4.1 cM vs. 5.4 cM between C104 and mTcCIR115 on LG4), even though genetic distances estimated in the UPA 402 map were slightly larger than those in the UF 676 map. This evidence allows us to use bridge loci as anchor markers to construct an integrated map.

The resulting order of the loci and map distances between markers (in centiMorgans) are shown graphically in the integrated linkage map reported in Fig. 1. The complete map contains 465 codominant markers (67 previously mapped SSRs, 201 new SSRs, 176 RFLPs, five isoenzyme loci and 16 Rgenes-RFLPs) arranged in ten linkage groups corresponding to the haploid chromosome number of cacao. Only one new SSR locus (mTcCIR146) remained unlinked at LOD 5.0. Another microsatellite marker, mtcCIR132, which produced an important rearrangement in the order of LG 2, was removed from the map.

The complete map is 782.8 cM long, with an average distance of 1.7 cM between markers. The number of mapped loci for each linkage group varies from 29 on LG8 to 61 on LG5. The most saturated linkage group in the map contains 61 markers and covers 73.2 cM with 1.2 cM between loci (LG5). Some distorted microsatellite markers appear to be clustered on LG3. Examination of the direction of segregation distortion showed that they all segregate for the male parent UF 676 only. Skewed segregations concerning both parents were also observed in loci of LGs 5 and 6.

Among the new SSRs, 116 (58%) were heterozygous for UF 676 only, 13 (6%) were heterozygous for UPA 402 only, and 72 (36%) were heterozygous in both parents (bridge markers) with two, three or four alleles (Table 2). Microsatellite markers were distributed throughout all linkage groups of the map, but their distribution was not random. The number of SSRs per group ranged from 11 in LG 10 to 36 in LG3. The map length with only SSRs was 769 cM, representing 94.8% of the total length of the map. The current level of genome microsatellite coverage is approximately one microsatellite every 3 cM. Nevertheless, some SSR markers were clustered and formed dense linkage blocks in LGs 1, 3, 5 and 9. No gap longer than 10 cM was found when all of the codominant markers were considered. With respect to microsatellite markers alone, linkage groups were well covered, but there were small gaps (10–20 cM) between microsatellites on LGs 2, 6, 7 and 10. The number of loci that were heterozygous in both parents varied between linkage groups from 23 on LG9 to zero on LG8. The relative order of the markers determined on previous maps (Lanaud et al. 1995; Risterucci et al. 2000) was preserved, and only a few inversions were observed (e.g. cTcCIR49

CHAPITRE II. Cartographie génétique du cacaoyer

1154

Table 1 Characteristics of SSR loci isolated from a cocoa genomic library enriched for $(GA)_n$ and $(CT)_n$ and integrated in the codominant marker-based map: locus designation, EMBL accession number, primer sequences, linkage group (LG) location,

expected PCR product size in the reference variety (Catongo), description of the repeat motif and calculated primer annealing temperature (T_m) using OLIGO 4.0 software

SSR name	EMBL accession number	5'-3' Forward primer	5'-3' Reverse primer	LG location	Expected size (bp)	Repeats	T_m (°C)
mTcCIR64	AJ566412	GAGAAAGTAAAAGGAGAGAG	TGTTAGAGAAATGAGAAATG	9	167	(GA) ₁₁	46.5
mTcCIR65	AJ566413	CATGAAAGCTAACGTGCCT	AAAAATGCGTTACAAGTGTG	3	242	(GA) ₁₀	50.7
mTcCIR66	AJ566414	TATCCGCCAGAAACAGA	CCAACAGTAGAGTCAGAGT	1	304	(AG) ₂₀	50.1
mTcCIR67	AJ566415	ATAGCTCTTTGACACGA	TTCTCTTTCACCTCTT	4	109	(GA) ₁₀	47.9
mTcCIR68	AJ566416	ATTAGCTGTAGCCGT	CCAGTTGATCTGCTTAATG	2	166	(GA) ₁₇	49.6
mTcCIR69	AJ566417	TCGGTTCCATCAGTA	CATGCTATGAGATTGAAAG	5	203	(CT) ₂₀	46.8
mTcCIR70	AJ566418	GGTATGAAGGATTGAGAG	TTCCTATTGATTTATGGG	8	107	(GA) ₁₁ AA (GA) ₄	44.4
mTcCIR71	AJ566419	CGACTAACCGCAGAAC	CTCCTCCTCCAT	6	170	(GA) ₁₀	47.5
mTcCIR73	AJ566420	CCAGTCAAGGAAGTATCT	AATGTCATGAACTGTTAGC	2	112	(CT) ₄ TT (CT) ₂ G (TC) ₈	45.8
mTcCIR75	AJ566421	CATTCATCTTTCTTCTCTC	TCCTCTCCAACCGA	8	121	(AG) ₁₀	48.5
mTcCIR76	AJ566422	AGCCAAAGAAAGGAT	TGAATCCGAGACAAAG	4	139	(TC) ₂₁ TTT (TC) ₈	46.2
mTcCIR77	AJ566423	GTTCTCCCCACTCTCT	AATAAATAAAACAAATACG	10	287	(TC) ₉	47.8
mTcCIR78	AJ566424	TGAAAATACGTCTGCTGA	CAAAAAGTTTCTGAAAGTC	3	159	(TC) ₂ T (TC) ₉	47.7
mTcCIR79	AJ566425	ATTTTCTTAGCGCACT	TAACTACCTCCACCTC	9	108	(TC) ₈	48.9
mTcCIR80	AJ566426	GCTGGGTTCTTGTATT	TTCTCATTTCTTATTGGTT	5	105	(CT) ₁₀ CC	46.8
mTcCIR81	AJ566427	TGAAAACCTCCATACTACTGA	ACAATCTGTCCATTATTCTG	3	216	(CT) ₁₅	47.9
mTcCIR82	AJ566428	ATCATGTGCCCTTCTAA	GGCAGCTAAGTGTTCATTC	3	174	(AG) ₆ AA	47.9
mTcCIR84	AJ566429	CATGGGACGCTGCCT	CTCTTATTAAATTGAATTCTCT	1	136	(GA) ₁₁	47.1
mTcCIR85	AJ566430	TTGAAGTAGAGAGTTGAAGAA	TTATGGTGTGGTGTGAT	1	211	(AG) ₁₆	46.7
mTcCIR86	AJ566431	TAACAAGGAAATGCTCTC	GTGAAACCGAAGGAAAG	5	370	(CT) ₈ TT	48.9
mTcCIR87	AJ566432	TAAGGGGCAACATAAAT	CAAATAGCGCAGAGACAAT	5	145	(AG) ₂₁	48.0
mTcCIR88	AJ566433	CTAGGATCCATAGAAGTAA	TTGGACCTCAATTATATGT	1	187	(CT) ₁₀	47.0
mTcCIR90	AJ566434	CCACTTCAAAACCATTCTA	GCAACTGTCAACCATTATCTA	9	291	(CT) ₁₀	47.6
mTcCIR91	AJ566435	TTTGCTGAGTGTGCTGT	ATCCGAGAAATAGAATAGGTTA	10	186	(CT) ₁₀	47.5
mTcCIR92	AJ566436	GTTCCAAATCATCTCACT	TCGCTATTCTCTTCACTCT	10	283	(AG) ₉	46.4
mTcCIR93	AJ566437	GTTGCCACTGCTCGCT	CCCTTTATGTTCCCAATTA	7	100	(CT) ₈	48.0
mTcCIR94	AJ566438	AATTGTAGGGTATTGAAGAG	CCATGCCAGTGAGTAG	1	196	(AG) ₁₆	51.0
mTcCIR95	AJ566439	CTCCTCCCTTCTCTC	CATCGTCTCTCTCATC	4	221	(TC) ₄ CC	47.9
mTcCIR96	AJ566440	ATAGAGAAAAGACCCAAATC	AGACAACAAATTATGTAATG	9	136	(TC) ₈	47.3
mTcCIR97	AJ566441	CTTCTTCTGGCAATCTTCT	GCTGCATCCATCATCC	1	122	(TC) ₃ C	46.9
mTcCIR98	AJ566442	CCAGTTGCTAACTTCTCTC	GCACATAGTITGGCAAT	9	140	(TC) ₈	46.2
mTcCIR99	AJ566443	CGGAAAATGAAACAGAC	AATAAAAGAAAAGAACATAC	8	249	(GA) ₉	49.0
mTcCIR100	AJ566444	TGATGGAATAAACTAAGAAC	TAAGAACGCCAGGTCAAG	2	244	(AG) ₆ C	48.2
mTcCIR101	AJ566445	GAACTCACCGAAGAGAAC	GAGCCGTACCAATGC	5	109	(TC) ₁₈	50.6
mTcCIR102	AJ566446	TTGTGAAAAGATTGCGA	TTGCTTGTATTGCTACTAT	1	124	(GA) ₉	46.0
mTcCIR103	AJ566447	GAGAGATGGCTTAAGGAT	ACCATACTATTGAAACATTG	8	112	(GA) ₁₀	46.5
mTcCIR104	AJ566448	AATAGGAAGGGTAAGTGAAT	CAAGCATATAAGCCAACA	10	167	(GA) ₁₀	47.8
mTcCIR105	AJ566449	GTTTACAATTATCGCTG	AATTGTATCCCTTATTATTA	3	201	(CT) ₈	46.1
mTcCIR106	AJ566450	ACGAAAATACCCAAAAA	TGCTGTGTGTCTTGT	1	143	(GA) ₉	45.8
mTcCIR107	AJ566451	TTGCCTGAAAGAGAGA	GATGGAAAGAGAAATAATAGT	4	120	(AG) ₁₁	46.8
mTcCIR109	AJ566452	GGAAAAGTGTAGGAAAGTAGAC	GGACAAAAAGAGCATA	5	162	(CT) ₁₂	46.4
mTcCIR110	AJ566453	GTGAAAAGTGGGATTG	TAAAGTAAGAGTGGTGTGATGGT	7	139	(AG) ₈	48.2
mTcCIR112	AJ566454	TTACTTGTAGGCTGCTG	CATTCCACTCATTTGCT	10	95	(TC) ₈	46.0
mTcCIR113	AJ566455	GGAAAGTTACAGCAAGAGAGA	ACAAGCCCGGTGAAGG	7	142	(AG) ₉	50.7
mTcCIR114	AJ566456	CAGATGAATGAAATACTT	GCATGAACACAAACACAC	9	207	(TC) ₉ (TG) ₅	48.7
mTcCIR115	AJ566457	GTGATTCAAATTCAAATG	AATAGCAAGAGAGTGTAG	4	191	(TC) ₁₁	46.3
mTcCIR117	AJ566458	TGTGGAATAAAAGAGCAAT	CACTGGTGTAGCAATGATA	4	168	(TC) ₁₀	47.2
mTcCIR118	AJ566459	TCTGCCGAAATGCTTC	TGGGGCACTAACCTTITG	1	165	(GA) ₁₀	47.4
mTcCIR119	AJ566460	TGGACTTGTGCTGGAAC	GCAAGAAATAAAATAGGAAC	5	123	(AG) ₁₂	47.8
mTcCIR120	AJ566461	TGGAAAGTGTACTCTTATG	TCTAGTTICAGGGGCTCT	3	95	(AG) ₁₃	46.2
mTcCIR121	AJ566462	CATGTGCAATTAGGTGTC	TCTGGCCTTCTAGTGTAC	1	138	(TG) ₁₂	46.8
mTcCIR123	AJ5664611	ATTCCTTAGCTTATGTTATG	CTCGGCCCTTCTCT	5	172	(CA) ₄ (TG) ₆	51.3
mTcCIR124	AJ566463	CAGCGTCTTGGAAATAAC	ACCCACACACAAAGACAC	9	131	(CT) ₁₂	46.2
mTcCIR125	AJ566464	CATGCAAATGCTTAGG	TGAACCAACAGCTGACAC	9	98	(TG) ₁₁	45.2
mTcCIR126	AJ566465	AACTCTCACTATCATCCAC	AACAAATCATCAACACCTT	9	212	(GA) ₁₁	46.3
mTcCIR127	AJ566466	CGTTTGTCTTGCCTTC	ATGTGTITTCGCCTTAC	5	130	(TC) ₈	48.6
mTcCIR129	AJ566467	CAGTGAGGATGAGGTTC	CGACATACCAAGTTACATAA	2	129	(TC) ₁₆	46.8
mTcCIR130	AJ566468	ACCGGCGGCTGATCTAC	CGCGCCAACCAATAAAG	1	133	(AG) ₁₇	51.0
mTcCIR131	AJ566469	TGAGTAAGAAAAAGTAGAAAA	GATCATCGTAAAGTAAAAT	3	204	(GA) ₉ C	46.9
						(GA) ₄	

Table 1 (continued)

SSR name	EMBL accession number	5'-3' Forward primer	5'-3' Reverse primer	LG location	Expected size (bp)	Repeats	T _m (°C)
mTcCIR133	AJ566470	GGATCACATCCGTTAGA	AATTTCAGCCCTCAA	3	155	(AG) ₁₁	49.0
mTcCIR134	AJ566471	CGTCCAAATCAACAC	ATAGTCTGCCCTCAA	8	176	(GT) ₁₅	47.1
mTcCIR135	AJ566472	ATTAGAGAGGGGTAGATGA	CTAGTGGGGTTGACATTG	3	246	(AG) ₂₀	50.0
mTcCIR136	AJ566473	GAGGAGGTGAGGCCA	GGTTTGATTTTGATTGAG	6	232	(GA) ₇ GC (GA) ₇	49.3
mTcCIR137	AJ566474	CAGCTGTACGGAAAC	GCCTCTTACCCCTATT	1	114	(TC) ₁₀	46.5
mTcCIR138	AJ566475	CTGCCAACGTAAGATTTC	CTGGGTATCAATCAATCTAAT	1	128	(CA) ₁₁	46.8
mTcCIR140	AJ566476	GATTCATAGTGGAAACACAGT	GGAAAACAGAGAGGAAGAGT	3	104	(CA) ₇	48.1
mTcCIR141	AJ566477	TGTTGCATAAAACACGAGTT	CCTAAAATCCTCTAACAGC	7	217	(TC) ₁₄	51.9
mTcCIR142	AJ566478	CCATITACAACCTCCATTACATA	AACTATCCATCCACCCCTACCTC	9	130	(GA) ₈	48.4
mTcCIR144	AJ566479	CCACTGACACGCCATGAA	CTAGGACTTAGGAAAGTGTGTTG	3	254	(TC) ₉	49.6
mTcCIR145	AJ566480	CAGACTTCAACTCAAACACT	TGAGAATAGATGGACCGAT	9	117	(CT) ₁₇	49.7
mTcCIR147	AJ566481	TGAAGCAATITGAAATCTGT	AACCACATCATAATGATITAAG	7	303	(TC) ₁₆	47.8
mTcCIR148	AJ566482	CGTCCTATCACTTCTCTTTC	TTGCCTTACAGCCATT	5	235	(TC) ₆ CC (TC) ₆	50.7
mTcCIR151	AJ566483	CAGGGGCTCTGTGTTT	ACCAAGAACGGGGAGA	2	139	(GA) ₁₂	50.5
mTcCIR153	AJ566484	GCCTCTCACACCATTATCTG	TACATTCAATTCTACTGCTG	3	217	(TC) ₉	50.1
mTcCIR154	AJ566485	CCTGTAAAGTGTGCGAAT	TGGAACAAAGAGGGTGTCA	9	167	(GA) ₁₄	49.0
mTcCIR155	AJ566486	CTTGGACTATTGGAAAAC	AAGGATACAATAAGGAAATAC	10	274	(TC) ₁₂	46.5
mTcCIR156	AJ566487	GGCAGGACCAAATGAT	AAAACCGGAAACACCG	5	216	(CT) ₁₁	51.3
mTcCIR157	AJ566488	ACTAATGCTGTTGGCTTC	TCACTCGACTGACTGTC	9	151	(AG) ₉	49.6
mTcCIR158	AJ566489	TGTAGGTTATGCACCGTGTTC	GATGGGGTGTAGCTGTTG	4	213	(CT) ₈	50.2
mTcCIR160	AJ566490	GATTGTTGTTGGTATGC	GTGAAGGTGAAGGTGTG	9	288	(GA) ₈	48.4
mTcCIR162	AJ566491	AAGATTGAGGTCACTCAGG	TAAGTTTGTCTTACTCTTC	2	162	(GA) ₁₉	48.0
mTcCIR163	AJ566492	CATAACCGAGACCAAGTGT	TTTGATCATGGCTTG	8	194	(AG) ₉	48.3
mTcCIR164	AJ566493	AGAACCGGTTAGGACAATC	AGGACAATGATGAAGAAATAAG	3	117	(CT) ₈	49.2
mTcCIR165	AJ566494	TTCACITCCCCCTCCCCAC	CTGGGTTGGAGTAGCTG	2	139	(CT) ₁₁	51.1
mTcCIR166	AJ566495	ATGAAACCACTATGTAAGACC	ATTCCAAAGGATTAGCAG	9	215	(CT) ₉ (CA) ₈	48.2
mTcCIR167	AJ566496	GTAGAACATAAAACACATT	ACAATCATTAAAAATACGAG	3	254	(GA) ₁₆	46.1
mTcCIR168	AJ566497	GGTACTATGAGGTGCGTAT	GTGAATGAATGGATGTGAAA	4	175	(TC) ₉	48.5
mTcCIR169	AJ566498	CTTTGGCTGTATGTTCG	CTGCCCTCTCTTCTCAC	5	178	(GA) ₉ AA (GT) ₇	51.0
mTcCIR170	AJ566499	CTCTTGACGGCACAGGA	TTGGCCCAACCCATACG	5	131	(TG) ₇	50.0
mTcCIR172	AJ566500	CGTTCCAGTGTGGGTGA	TGTTTCGCTCTACTGCTTC	9	127	(TG) ₉ (AG) ₄	51.1
mTcCIR173	AJ566501	TCCGGATGGCAATATGT	CTACCCCATGATTGTAACCT	3	148	(CA) ₇	49.9
mTcCIR174	AJ566502	TGGCAGCAATACTTCAAA	TCCCGATGTTCCACTC	1	167	(TG) ₇	49.5
mTcCIR175	AJ566503	TTACAATCAACAGAACCTC	TATATTGATGCGAAAGTC	3	248	(CA) ₇	47.2
mTcCIR176	AJ566504	TCACCAATTCTCTGTC	AATGAAATTACCTCCTAC	2	106	(TG) ₁₆	46.2
mTcCIR177	AJ566505	GATCCTTGAACACACACA	TAATTCTCTTACACATTCTC	7	128	(CA) ₇	47.4
mTcCIR178	AJ566506	CATCTTTGCACATATTG	GCTTGGCCCTTAACAC	9	138	(TG) ₈	48.5
mTcCIR179	AJ566507	TTTCCATTCTCATTCTCAAG	ATGTTITCATTTGCTATCCAA	7	288	(GT) ₁₆	50.7
mTcCIR180	AJ566508	ATGGTTTCTGATTGTCGT	CAAATCTAACGTCATAAAAC	3	186	(GT) ₉	46.4
mTcCIR181	AJ566509	CTTATGCTGCTCTGTA	CCAAGAATGTTTGTACTG	7	197	(CT) ₁₂ (CA) ₉	47.5
mTcCIR182	AJ566510	CTAATTGTTCAAGGAGGT	AACTGTTTGTGGCACTATC	6	148	(TG) ₉	46.3
mTcCIR183	AJ566511	GTTATCTTAGTTCTAGCCAC	GTAGTCTTACACCTGATTG	4	353	(AC) ₉	45.9
mTcCIR184	AJ566512	GGTTTCTAGCTCTCC	AGGAAAGAATGACTCATACTA	1	139	(CA) ₈ (CT) ₁₃	48.2
mTcCIR185	AJ566513	ATCCCCCTGCCAACAGAG	CCTGAATGAAGTAAAGCCAAAT	6	142	(CA) ₁₈	50.0
mTcCIR186	AJ566514	AAGGCTAAAGAACAAATG	CGTAGACGTACACAAATA	7	147	(TG) ₈	46.5
mTcCIR187	AJ566515	TTCACCTAGTGAATGGTCT	GCAGGCTTCAATTAGAG	9	262	(TG) ₈	49.4
mTcCIR188	AJ566516	GTCCTATTCCGGTAACTAC	TTCTGTTCTCTTGTGTC	4	118	(TC) ₉ (AC) ₈	48.0
mTcCIR189	AJ566517	GAATAGAAATTATGTCAC	TCAAAACACATAGTCAC	8	150	(GT) ₁₂	45.9
mTcCIR190	AJ566518	AAGAAACTGAAGCACAAAT	CACAAAGAGCATAAACTG	7	166	(TG) ₁₂	46.7
mTcCIR192	AJ566519	TCACCTACAATAATTCAAG	AAATGAAATTCCAGTGTAG	5	98	(TG) ₁₆	46.0
mTcCIR193	AJ566520	AACATGTTGATGGACCG	AAATGGTGAATTAGGCTC	6	134	(TG) ₉	48.9
mTcCIR194	AJ566521	ACACACGCTAAACAGAAA	GGGATGTGACGGATATTAC	1	192	(TG) ₁₄	47.7
mTcCIR195	AJ566522	CAAGTTGAATAAAAGCCTAAG	AAAATAAAGAAAATGAAGTAA	2	350	(CA) ₁₀	46.6
mTcCIR197	AJ566523	GGATTATTTATTGTAACACTCC	AATGATTCTACATTGTCACCA	5	162	(AT) ₉ (GT) ₁₈	46.2
mTcCIR198	AJ566524	TGGGACCATAGGAAATC	CCCAGGTGAAGTAAGACA	3	186	(CA) ₃ TA (CA) ₆	46.3
mTcCIR199	AJ566525	GATTCTTATTGATTTCCTTA	GCACGGTTACATTATTACA	4	211	(TG) ₁₄ TC	47.4
mTcCIR200	AJ566526	GCCAATTCTGACCCA	CTTAAAATAAGCCCAAATAC	8	238	(TG) ₈	47.3
mTcCIR202	AJ566527	TCTCTCATAGCTAACAGCA	CCTGAGTCAAAGTGTCT	3	172	(TG) ₇ (GA) ₉	48.3
mTcCIR203	AJ566528	GTGGATTITGGGTGGGAT	ATTGTGTTTGGCTATGTT	1	217	(AC) ₈	50.7
mTcCIR204	AJ566529	ATTACCTGCCGATGAAG	TGGGTITGGAATGATGT	3	131	(CT) ₁₀	47.3
mTcCIR205	AJ566610	GGGGTTTGTATTATGAT	TGTGGGATCTGCTTCT	9	198	(TC) ₈	47.5
mTcCIR207	AJ566530	TGGTTGACAAGGTAAAA	TGGATGTGCAAGTAAGT	4	174	(TC) ₉	45.4
mTcCIR208	AJ566531	GCAAGCCCCTAAACT	AAAAGCAAAAGAAGAAGA	6	209	(CT) _{10-25pb-} (AC) ₁₁	48.3
mTcCIR209	AJ566532	TACGGGCTAATGGTGA	AGGTATGCTGTATTGTTG	6	259	(TG) ₆ TAT (GA) ₉	47.8
mTcCIR210	AJ566533	CAAACCCCAAACCTCAA	CAGITATGGAAATTATGCTCTA	1	146	(AG) _{11-7pb-} (AAG) ₄	49.5
mTcCIR211	AJ566534	TGGTGTAACTCAAATC	CAAACAAGAAGGCTAAA	8	182	(TC) ₉	46.5
mTcCIR212	AJ566535	GAGAACACTCAGGATAC	GTCATITGGCAGATTAA	9	186	(TC) ₈	46.6

CHAPITRE II. Cartographie génétique du cacaoyer

1156

Table 1 (continued)

SSR name	EMBL accession number	5'-3' Forward primer	5'-3' Reverse primer	LG location	Expected size (bp)	Repeats	T _m (°C)
mTcCIR213	AJ566536	GATCTCGAAACTAACAA	TAAGTAAAATGAAGGTGTGA	4	261	(CT) ₂₆	47.9
mTcCIR215	-	GCTTCAACTCAAATCAC	TAGCATCCCGTATTGTG	9	197	(AG) ₁₃	49.1
mTcCIR216	AJ566537	ACTGCCAGGAATCA	TCITTGTTCTGCCITAT	5	158	(GA) ₁₂	47.4
mTcCIR217	AJ566538	AGTTTCCATCTATATTGTTA	TATTGTCTACGGTCTCT	4	132	(CT) ₁₁	43.4
mTcCIR218	AJ566539	TGACCAAGGAAGCTC	GGTGGGAAAGGTGGTA	8	187	(CT) ₁₁	48.9
mTcCIR219	AJ566540	GCGAACCAAGAACAAATAC	ATGGGTGCAATTCTCT	3	187	(AG) ₈	49.6
mTcCIR220	AJ566541	TGAAGTGTGTTGTGTA	CCAATAGAGGGATGTAATA	10	201	(TC) ₁₀	47.8
mTcCIR221	AJ566542	ATGTAGTTGGCTGTGA	TGTTAAGAGGGAAATGAA	4	273	(TC) ₉	48.6
mTcCIR222	AJ566543	CTACAGAAAATAGGCAATA	TCATTGTATTACAGGTAGA	4	220	(GA) ₉	45.2
mTcCIR223	AJ566544	GGTCCACACTCAACACT	TTATTCCATTTCATTTACT	10	202	(TC) ₄ GC (TC) ₂ GC (TC) ₁₅	45.0
mTcCIR224	AJ566545	TCAGAAAGCAATGTGGTA	AAGCAATATCAAGTGTAAAG	2	223	(TC) ₁₃ (AC) ₈	48.0
mTcCIR225	AJ566546	AAGACAAAGGGAAAGAAGA	AGGGAAAGAGCAAATC	8	302	(TC) ₁₀	49.8
mTcCIR226	AJ566547	TAACCCAAATTCAAAGTC	TITCAACAGCCTCATCT	3	246	(TC) ₁₁	47.3
mTcCIR227	AJ566548	ACATCATTAAAGGAGAAACA	CAAATCACCTCAAATAATC	2	142	(CT) ₈	46.4
mTcCIR228	AJ566549	CCCCCTGATACTGTGTG	GAAACCTAATCTGTAAATATGT	2	110	(GA) ₈	48.7
mTcCIR229	AJ566550	ATCTCGGTAAATGACATAA	CGCAATCTTACAACACA	10	307	(TC) ₈	47.9
mTcCIR230	AJ566551	GTGGAAGCCTTATGATTATGT	ATTATGCCATGCAGAC	2	231	(CT) ₈	49.5
mTcCIR231	AJ566552	AGGAGGAGTGTGAA	CAGGTTCCAAATTGTAT	4	226	(AG) ₈	47.3
mTcCIR232	AJ566553	GCTGTTGCTACTTTGAAAT	CACCCCTTGAATCAGTCTA	5	205	(TC) ₁₈	49.8
mTcCIR233	AJ566554	CCAGAACCAAAGAGA	ATGGATTAAAGAAGGAGGAA	4	211	(GA) ₁₅	48.9
mTcCIR234	AJ566555	TTGTGTCGGTTGATTTC	GAAAGAGAGGGAAAGTGA	4	123	(TC) ₉	48.1
mTcCIR235	AJ566556	TTCGGATGGCAACTAATC	AAAACAGCGGAACAGGTA	6	292	(AG) ₈	49.1
mTcCIR236	AJ566557	GAAGTCAAAGGAAAGTCAA	TCAGAAAACGCAAATAAA	8	193	(CT) ₅ TT (CT) ₁₄ T	50.1
mTcCIR237	AJ566558	GAAGACAAGGATGGAGACT	GCAAAGAGAGCAGGAGA	4	103	(TC) ₁₀	50.1
mTcCIR238	AJ566559	TTGGCTTCTTTAGTTA	AAATATAATCATTACTTCCTA	6	126	(AG) ₉	44.1
mTcCIR239	AJ566560	CTTCCACAGTCAAATAACAA	TIAAATCCCGAAAGT	5	203	(CT) ₉	47.4
mTcCIR240	AJ566561	CATACTACTACTGCTCTCT	AGTGATITATGGGACTTT	2	158	(CT) ₂₂	46.5
mTcCIR241	AJ566562	CAGTTGGAGGGCATT	ACGAGTGAAGAGAGTGAAGTT	4	146	(CT) ₂₃	50.2
mTcCIR242	AJ566563	TTTCGGCATTCCACTA	GTAAAAACAAATACTTCAACTA	4	287	(CT) ₉ (CA) ₉	48.2
mTcCIR243	AJ566564	ACAGCACTAGACGCATTC	AAAAGGCTTGGCACAG	4	141	(TC) ₉ -20pb- (CA) ₁₁	50.6
mTcCIR244	AJ566565	TGGCAATAACAAATGAACA	ATTTTGATGATTGATGAAGA	1	264	(TA) ₄ CATA (CA) ₁₇ (TA) ₄	47.2
mTcCIR245	AJ566566	GCAAATAGACAGCAAAT	TTCAAAGGAGTATAGGTAA	5	198	(GT) ₈	45.8
mTcCIR246	AJ566567	TATCCTCTCTCTGTGTATC	GCAGCACTAACCAACTA	1	169	(TG) ₈	47.6
mTcCIR247	AJ566568	CATTTTATAATTCTCTTCT	ACATTCTTATTTCACACT	3	111	(AC) ₉ ATAC (AT) ₃	39.9
mTcCIR248	AJ566569	TGATAGATTGCGTTACA	CCCAGAAAAGAAGAAGAT	5	190	(TG) ₈	48.0
mTcCIR249	AJ566570	TCTCAAGTTCAAGGTCT	GACACAAATGCCATTAT	1	246	(CT) ₄ TT (CT) ₂₈ (AC) ₁₆	47.9
mTcCIR250	AJ566571	CCCAGAGGACCACATCAC	ACTGCTCTCTCTACTCATC	9	237	(AC) ₂₂	49.6
mTcCIR251	AJ566572	TCTATGGATTTGATGAG	AGATACAGCAGGAACACA	9	188	(CT) ₇ (CA) ₁₂	46.8
mTcCIR252	AJ566573	AATGTGTGCTTTGTTCTA	TTCAAGGGCGTAAC	2	155	(AC) ₁₀	45.8
mTcCIR253	AJ566574	TGGCCTACTAACACCTACTA	GGGAGGGGAGTAGTT	2	155	(AC) ₇	45.4
mTcCIR254	AJ566575	ACAACTCCAAAGAACAAAG	GGTAAACCTCGTCATAAT	3	198	(AC) ₂₁ (AT) ₉	45.3
mTcCIR255	AJ566576	TTTACCTCCACCATCTT	TGGCACTTATCTTACTGT	6	203	(AC) ₁₁	47.5
mTcCIR256	AJ566577	AGAAGGCTGTCAACATTA	GAACAGTCAAACATAAGAGTA	5	185	(AC) ₁₃	46.1
mTcCIR257	AJ566578	CATACAGAAACCAGAAAAT	TATAGGGTAAAGCGAAAT	5	167	(CT) ₇ (CA) ₁₂	48.0
mTcCIR258	AJ566579	TAACTCACAAATCCATCAT	ATGGTCATTATCAAATC	8	116	(TC) ₇ (AC) ₁₀	44.5
mTcCIR259	AJ566580	TTTCCTGATTCACATTA	AGAGGTTCCAAATACAT	5	157	(CA) ₈	45.2
mTcCIR260	AJ566581	TGGCAACACATACATTA	GTGTATGCCAGATGAGA	2	112	(AC) ₁₇	44.6
mTcCIR262	AJ566582	GTTTCTTGTCCGTATCT	TTTGCCAAACCTGTGT	1	165	(TC) ₁₄	47.6
mTcCIR263	AJ566583	ACCAGGAGTTTCTTGT	ATTTAGTCAGCTCATCATTAT	3	244	(TC) ₉	47.3
mTcCIR264	AJ566584	TGCTATCCACAAACAGT	TAACTCACCTTGCCACTA	1	192	(CT) ₈	47.0
mTcCIR265	AJ566585	TGAATGCTGGAAAAATGT	GTGTCCTGTTGGTTGT	5	246	(AG) ₁₈	49.2
mTcCIR266	AJ566586	TCGTCGCCATCATAGA	GTCGTTATTCCGGAGTTC	9	192	(CT) ₁₅	50.7
mTcCIR267	AJ566587	CACTACCCCTTTCTT	TTCATGGCTTCTTCTAT	5	199	(GA) ₉	49.4
mTcCIR268	AJ566588	TGTAATCCAATAAAAGCAT	CAGTGAAGAGGCAAGAGA	2	316	(GA) ₁₇ GG	49.5
mTcCIR270	AJ566589	TTAGTGAAGATGGTGAACAT	AATCAAGGAAAAGTTATCA	1	224	(CT) ₁₀ (CA) ₉	46.1
mTcCIR271	AJ566590	GCACCTTTGTTATCTTG	GGAACCCACTGAAACT	5	173	(CT) ₁₂	45.6
mTcCIR272	AJ566591	TTTGCCTTCCTTCTT	TTTGTCAATTGGATAGTG	1	258	(CT) ₅ (TC) ₄ - 151 pb-(AT) ₄ (CA) ₇	48.5
mTcCIR273	AJ566592	AGAATGATCGCAGAGAG	ACGGCATTAGAGAGAGA	1	167	(CT) ₄ AC (CT) ₁₃ TT (CT) ₄	47.3
mTcCIR274	AJ566593	GAAAGGTTAAATGGCTGAA	CGATCATCACGACTGCT	5	184	(CT) ₆ CACG (CA) ₆ (CT) ₂	49.7
mTcCIR275	AJ566594	GGTTTGGTTGGTAAGAC	TAAGAGAGAGTGTGCTGACA	1	146	(CT) ₁₁	53.0

Table 1 (continued)

SSR name	EMBL accession number	5'-3' Forward primer	5'-3' Reverse primer	LG location	Expected size (bp)	Repeats	T _m (°C)
mTcCIR276	AJ566595	TCCTGCTTTAATACAT	GTCCTATCTGCCTCACT	6	124	(GA) ₁₄	46.3
mTcCIR277	AJ566596	ACCAAGATCAAAGTCAAGAA	GATAAGAACCAAGTGAAGAGA	7	304	(AG) ₁₆ AA	50.4
mTcCIR278	AJ566597	TGGCATCTGTCTGTC	GTATATGACCGTTGTAG	3	100	(AG) ₁₁ (TCTG) ₃ (TC) ₁₀ (TA) ₈	41.9
mTcCIR279	AJ566598	GTCCCATCTACATCATCAAGC	CAGCAACAGCATCACT	5	155	(CT) ₉	44.1
mTcCIR280	AJ566599	ATTGTCTATTGTTGTTGT	GCCTTGGTATTGACTGT	3	89	(CT) ₈ (CA) ₉	44.9
mTcCIR281	AJ566600	CCGCTGTTGGTATTTC	GGATGAGGGTGGTTG	2	194	(TC) ₁₂ (CA) ₁₄	51.4
mTcCIR282	AJ566601	TGGTGAGGGGAGAGAA	AGCAAAGCAATAATAATG	8	172	(GA) ₂ GG (GA) ₆	49.4
mTcCIR283	AJ566602	ATCAATACCCACCAACACA	CCCTTTCCCTTTTCT	1	239	(TC) ₁₁	49.3
mTcCIR285	AJ566603	TACTACCTCTACCCCTCTGT	ATAAAATTCCTCTCCCTCT	9	216	(AG) ₁₆	46.7
mTcCIR286	AJ566604	GTTCCTGCTTCATCTGTTTA	TTCAACCCACAAACCAT	1	119	(CT) ₁₈	46.2
mTcCIR287	AJ566605	TCCTTTCTGTTGTTCT	TTATCCGTGTCCTCTCT	9	301	(TC) ₉	48.2
mTcCIR288	AJ566606	ACAACACAAGGCAAAGA	CCCATTTAGCACCAAC	5	184	(GA) ₁₀ AA (GA) ₃ -36pb- (GA) ₁₁	48.8
mTcCIR289	AJ566607	CTTCCGCCACTAATAAA	CTATACATAACAGCAGCCA	3	123	(CT) ₁₀	46.8
mTcCIR290	AJ566608	AGCGAGAGACAAAGATAAT	GACTGAAATGGTGGTAAAG	6	175	(CT) ₁₉ CACC (CA) ₁₈	49.4
mTcCIR291	AJ566609	AGTCCCATAGGTTCCAAT	CGAGGTTATCCCCAAA	6	218	(CT) ₁₂	50.1

Table 2 Number of polymorphic SSRs obtained from the enriched genomic library and type of segregations observed in the mapping population

	Total	UPA 402 clone	UF 676 clone
Number of primer pairs screened for polymorphism	387		
Number of polymorphic loci	223		
Number of primer pairs analysed in the progeny	201		
Segregating 1:1	129	13	116
Segregating 1:1:1:1	72		
Two alleles	8		
Three alleles	34		
Four alleles	30		

and gTcCIR122 on LG3 where two new microsatellites segregating in both parents were mapped between them).

Discussion

We describe here the development and mapping of a new set of 201 microsatellite markers which were integrated with a set of previously mapped codominant markers (SSRs, RFLPs, isoenzymes and Rgenes-RFLPs). This enabled us to obtain a new codominant marker-based cacao linkage map with 465 codominant markers arranged in ten linkage groups.

A small proportion of the new loci showed segregation distortion when compared to the expected Mendelian ratio. The reasons behind this distortion remain unclear, but if we consider the fact that RFLP markers showing distorted segregation have previously been mapped in these regions (Risterucci et al. 2000), biological mechanisms such as selection against closely linked lethal or sublethal genes, linkage with genes that are subject to direct selection or the presence of incompatibility alleles, rather than chance or error, could be hypothesized. Segregation distortions have already been observed in tree species (Viruel et al. 1995; Kijas et al. 1997;

Barreneche et al. 1998; Dettori et al. 2001). Although the inclusion of distorted loci into the map increases the chance of type I errors of false linkage, these loci can be useful in increasing our knowledge of specific regions and in the mapping of QTLs.

Our map is shorter than previously published ones that include AFLP markers. It is known that AFLPs derived from certain combinations, such as *EcoRI/MseI* selective primers, are often clustered in some specific regions of the genome, especially AT-rich heterochromatic regions around centromeres and at chromosome ends (Boivin et al. 1999). In the previous maps, several AFLP markers were located around telomeric regions and their removal decreased map length.

The distribution of microsatellite markers within the linkage groups was not random. The tendency of SSR loci to cluster has been reported in several species, including barley (Ramsay et al. 2000), sorghum (Bhatramakki et al. 2000), rye grass (Jones et al. 2002) and rice (Mc Couch et al. 2002). It has been suggested that this is influenced by the non-uniform distribution of recombination events in the mapping population (reduced recombination frequency) (Castiglioni et al. 1999).

The addition of new SSR markers allowed us to fill some gaps present in the previous map (especially on

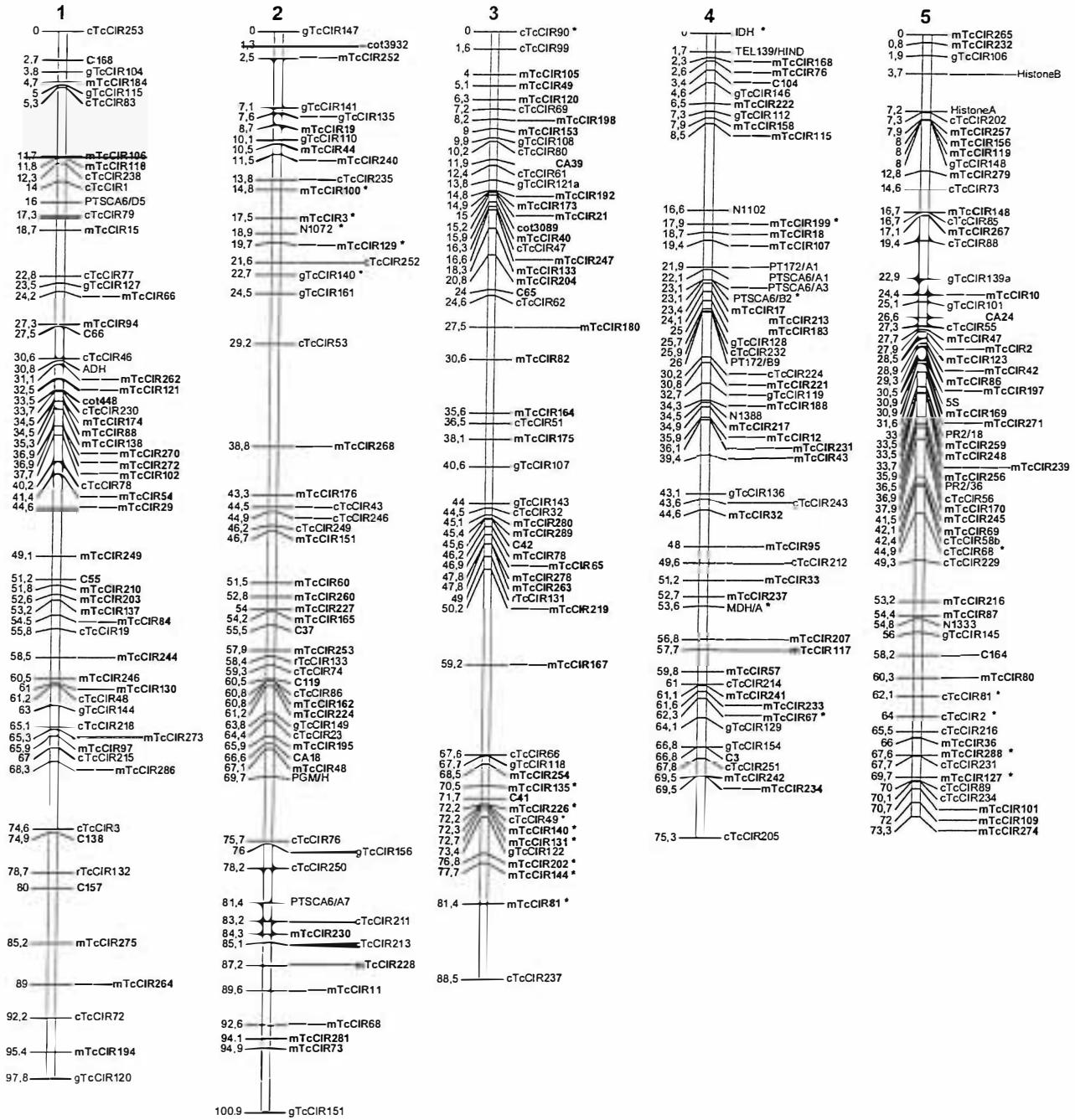


Fig. 1 Linkage map of cocoa based on the cross UPA 402 × UFA 676. The map consists of 465 codominant markers (269 SSR loci, 178 RFLPs, five isoenzymes and 16 Rgenes). Symbols for SSR loci are in bold type. Polymorphic markers in the gametes of the

UPA 402, in the gametes of UFA 676 and in the gametes of the both parents are designated on the right, left and middle respectively. Markers showing distorted segregation ratios are denoted with an asterisk

regions of LGs 3, 6 and 8) and to saturate some regions (LGs 4 and 5), but it did not enable the saturation of the distal region of LG 10, where only one new SSR was mapped in 15 cM and the average distance between markers is 3.5 cM longer than the average interval between markers (1.7 cM) in the whole map. This

suggests that low polymorphism exists between the two parents of the mapping population in this region. The non-random distribution of polymorphism along chromosomes may reflect some structural or functional properties of the DNA in these parts of the genome or the dynamics of domestication and breeding process accompanied by

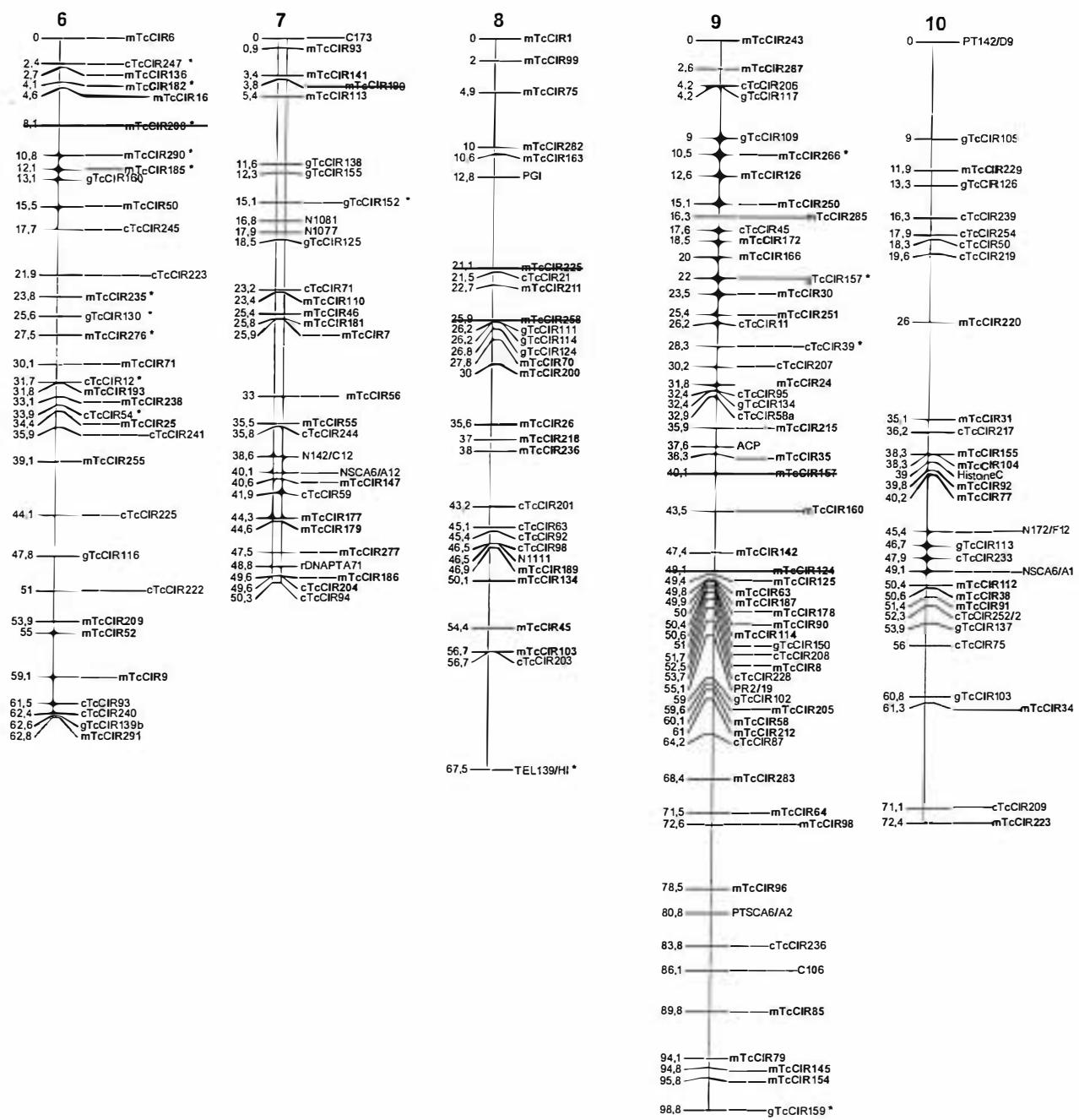


Fig. 1 (continued)

fixation of some segments of chromosome (Temnykh et al. 2000). Cregan et al. (1999) have proposed that targeted isolation of SSR loci using BAC (bacterial artificial chromosome) clones could populate these regions with a few SSRs. To this end, Bhatramakki et al. (2000) showed that BAC libraries were either equal or superior to enriched gDNA libraries as sources of microsatellites for sorghum.

Despite the high number of heterozygous markers segregating in both parents, the order of markers and their distances along LG10 of the UPA 402 map relative to those of the UF 676 map could be ambiguous. Indeed, only two closely linked markers segregated in both parents. LG8 was an exception, since all of the markers mapped there originated from UF 676 alone. We have no evidence that this is related to genes with specific functions located in this linkage group. The explanation

may lie in the pedigree of UPA 402. Indeed, UPA 402 is an Upper-Amazon Forastero clone obtained from a sib-mating involving two Forastero genotypes collected in Peru and may be highly homozygous for this linkage group. A similar situation, where a small number of markers were also assigned to LG8, was observed in other maps established by Clement et al. (2003) and Flament et al. (2001).

Nevertheless, we have produced a saturated microsatellite marker map, and the markers appear to be distributed throughout the genome. If we divide this map into 78 intervals of 10 cM each, there is at least one SSR in each interval, except for four intervals (5%) located on LGs 2, 6, 7 and 10. Macaulay et al. (2001) in barley and Aranzana et al. (2003) in peach have proposed the use of a set of single-locus, codominant and highly polymorphic markers as a framework set or 'genotyping set' that could be easily shared by different research groups. We consider that the marker density and distribution in our map could open the way for a similar approach in cacao.

The development of a saturated linkage map with the markers spaced at small intervals throughout the genome could help improve our knowledge of genome structure. The availability of 268 cacao microsatellites spread throughout the genome and with a coverage of approximately one microsatellite every 3 cM is an attempt in this direction. Studies with a similar approach have shown that the integration of SSRs into the previous linkage maps based mainly on RFLPs was also useful in extending the map length, filling gaps and improving genome coverage (Bhattamakki et al. 2000; Joobeur et al. 2000; Aranzana et al. 2003).

Due to their codominant nature, single-locus behaviour and high polymorphism, microsatellites constitute a set of useful markers that are transferable across different mapping populations, thereby allowing QTL position comparison, and easily transferable to laboratories in tropical regions. This linkage map based on codominant markers, especially SSRs associated with RFLP markers, will be used to localize the most important regions involved in the variation of the traits of interest, such as quality or disease resistance, and to develop marker-assisted selection strategies for this important crop in developing countries. Moreover, their intrinsic properties make microsatellites suitable for the analysis of linkage disequilibrium (LD). Most of the modern Criollo/Trinitario cacao varieties correspond to hybrids between a very small number of parents: a completely homozygous "ancient" Criollo and a Lower-Amazon Forastero individual. Only a few generations separate the first hybridizations from the present Criollo/Trinitario varieties, and allelic associations must have been maintained LD. Microsatellite markers could be used to evaluate the importance of this LD at the genome level in the Criollo/Trinitario varieties and to highlight the associations maintained between molecular markers and useful genes.

Acknowledgements We thank the CNS (Centre National de Séquençage) for sequencing the cacao DNA fragments needed to

develop microsatellites. We thank also USDA for their financial participation in these studies and D. Crouzillat (Nestlé) for providing us with 19 microsatellites and seven genomic probes.

References

- Aranzana M, Garcia-Mas J, Carbó J, Arús P (2002) Development and variability of microsatellite markers in peach. *Plant Breed* 121:87–92
- Aranzana M, Pineda A, Cosson P, Dirlewanger E, Ascasibar J, Cipriani G, Ryder C, Testolin R, Abbott A, King G, Iezzoni A, Arús P (2003) A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. *Theor Appl Genet* 106:819–825
- Artiguenave F, Wincker P, Brottier P, Duprat S, Jovelin F, Scarpelli C, Verdier J, Vico V, Weissenbach J, Saurin W (2000) Genomic exploration of the hemiascomycetous yeast: 2. Data generation and processing. *FEBS Lett* 487:6–13
- Barreneche T, Bodenes C, Lexer C, Trontin J-F, Fluch S (1998) A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5S rDNA markers. *Theor Appl Genet* 97:1090–1103
- Beckman JS, Soller M (1990) Toward a unified approach to the genetic mapping of eukaryotes based on sequence-tagged microsatellite sites. *Biotechnology* 8:930–932
- Bhattamakki D, Dong J, Chabra AK, Hart GE (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. *Genome* 43:988–1002
- Billote N, Lagoda PJL, Risterucci AM, Baurens FC (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits* 54:277–288
- Billote N, Risterucci AM, Barcelos E, Noyer JL, Amblard P, Baurens FC (2001) Development, characterisation and cross-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. *Genome* 44:413–425
- Boivin K, Deu M, Rami JF, Trouche G, Hamon P (1999) Towards a saturated sorghum map using RFLP and AFLP markers. *Theor Appl Genet* 98:320–328
- Castiglioni P, Ajmone-Marsan P, van Wijk R, Motto M (1999) AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. *Theor Appl Genet* 99:425–431
- Clement D, Risterucci AM, Motamayor JC, N'Goran JAK, Lanaud C (2003) Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46:103–111
- Cregan PB, Mudge J, Fickus EW, Marek LF, Danesh D, Denny R, Mathews BF, Jarvik T, Young ND (1999) Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. *Theor Appl Genet* 98:919–928
- Crouzillat D, Lerceteau E, Petiard V, Morera J, Rodríguez H, Walker D, Phillips W, Ronning C, Schnell R, Osei J, Fritz P (1996) *Theobroma cacao* L.: a genetic linkage map and quantitative trait loci analysis. *Theor Appl Genet* 93:205–214
- Crouzillat D, Phillips W, Fritz J, Petiard V (2000) Quantitative trait analysis in *Theobroma cacao* L. Using molecular markers: inheritance of polygenic resistance to *Phytophthora palmivora* in two related populations. *Euphytica* 114:25–36
- Dettori MT, Quarta R, Verde I (2001) A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome* 44:783–790
- Dirlewanger E, Cosson P, Tavaud M, Aranzana MJ, Poizat C, Zanetto A, Arús P, Laigret F (2002) Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). *Theor Appl Genet* 105:127–138
- Edwards KJ, Barker JHA, Daly A, Jones C, Karp A (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques* 20:758–760
- Flament MH, Kebe I, Clement D, Pieretti I, Risterucci AM, N'Goran JAK, Cilas C, Despréaux D, Lanaud C (2001) Genetic

- mapping of resistance factors to *Phytophthora palmivora* in cocoa. *Genome* 44:79–85
- Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-test-cross: mapping strategy and RAPD markers. *Genetics* 137:1121–1137
- Gupta PK, Balyan HS, Edwards KJ, Isaac P, Korzun V, Röder M, Gautier M-F, Joudrier P, Schlatter AR, Dubcovski J, de la Pena RC, Khairallah M, Penner G, Hayden MJ, Sharp P, Keller B, Wang RCC, Hardouin JP, Jack P, Leroy P (2002) Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *Theor Appl Genet* 105:413–422
- Haussmann BIG, Hess DE, Seetharama N, Welz HG, Geiger HH (2002) Construction of a combined sorghum linkage map from two recombinant inbred populations using AFLP, SSR, RFLP, and RAPD markers, and comparison with other sorghum maps. *Theor Appl Genet* 105:629–637
- Jones ES, Dupal MP, Dumsday JL, Hughes LJ, Förster JW (2002) An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). *Theor Appl Genet* 105:577–584
- Joobeur T, Periam N, de Vicente MC, King GJ, Arús P (2000) Development of a second-generation linkage map for almond using RAPD and SSR markers. *Genome* 43:649–655
- Kauffmann S, Legrand M, Geoffroy P, Fritig B (1987) Biological function of “pathogenesis-related” proteins: four PR proteins of tobacco have 1,3-glucanase activity. *EMBO J* 6:3209–3212
- Kijas JMH, Fowler JCS, Garbett CA (1994) Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques* 16:656–662
- Kijas JMH, Thomas MR, Fowler JCS, Roose ML (1997) Integrating of trinucleotides microsatellites into a linkage map of *Citrus*. *Theor Appl Genet* 94:701–706
- Kosambi DD (1944) The estimation of map distance from recombination values. *Ann Eugen* 12:172–175
- Lanaud C, Risterucci AM, N'Goran JAK, Clement D, Flament MH, Laurent V, Falque M (1995) A genetic linkage map of *Theobroma cacao* L. *Theor Appl Genet* 9:987–993
- Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Mol Ecol* 8:2142–2152
- Lanaud C, Risterucci AM, Pieretti I, N'Goran JAK, Fargeas D (2004) Characterisation and genetic mapping of resistance and defence gene analogs in cocoa (*Theobroma cacao* L.). *Mol Breed* (in press)
- Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221
- Liebhard R, Gianfranceschi L, Koller B, Ryder CD, Tarchini R, Van De Weg E, Gessler C (2002) Development and characterisation of 140 new microsatellites in apple (*Malus × domestica* Borkh.). *Mol Breed* 10:217–241
- Macaulay M, Ramsay L, Powell W, Waugh R (2001) A representative, highly informative ‘genotyping set’ of barley SSRs. *Theor Appl Genet* 102:801–809
- Martin GB, Frary A, Wu R, Brommonschenkel SH, Chunwongse J, Earle ED, Tanksley SD (1994) A member of the tomato *Pto* gene family confers sensitivity to fenthion in rapid cell death. *Plant Cell* 6:1543–1552
- McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, De Clerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res* 9:199–207
- Queiroz VT, Guimarães CT, Anhert D, Schuster I, Daher RT, Pereira MG, Miranda VRM, Loguerio LL, Barros EG, Moreira MA (2003) Identification of a major QTL in cocoa (*Theobroma cacao* L.) associated with resistance to witches’ broom disease. *Plant Breed* 122:268–272
- Ramsay L, Macaulay M, degli Ivanissevich S, MacLean K, Cardle L, Fuller J, Edwards KJ, Tuveson S, Morgante M, Massari A, Maestri E, Martinoli N, Sjakste T, Ganai M, Powell W, Waugh R (2000) A simple sequence repeat-based linkage map of barley. *Genetics* 156:1997–2005
- Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high density linkage map of *Theobroma cacao* L. *Theor Appl Genet* 101:948–955
- Risterucci AM, Paulin D, N'Goran JAK, Lanaud C (2003) Identification of QTL related to cocoa resistances to three species of *Phytophthora*. *Theor Appl Genet* 108:168–174
- Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organisation of microsatellites sequences in rice (*Oryza sativa* L.). *Theor Appl Genet* 100:697–712
- Van Ooijen JW, Voorrips R (2001) JOINMAP version 3.0, software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands
- Viruel MA, Messeguer R, De Vicente MC, Garcia-Mas J, Puigdomènech P (1995) A linkage map with RFLP and isozyme markers for almond. *Theor Appl Genet* 91:964–971
- Weber JL (1990) Informativeness of human (dC-dA)n (dG-dT)n polymorphism. *Genomics* 7:524–530
- Whitham S, Dinesh-Kumar SP, Choll D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic virus resistance gene N: Similarity to Toll and the interleukin-1 receptor. *Cell* 78:1101–1115

Conclusions et Perspectives

Une nouvelle carte génétique du cacaoyer a été construite en intégrant 201 nouveaux marqueurs microsatellites développés dans cette étude avec un certain nombre d'isoenzymes, de RFLP, de marqueurs microsatellites et de gènes analogues à des gènes de résistance et de défense déjà cartographiés. Cette carte a un avantage important par rapport aux cartes déjà existantes, car elle ne comporte que des marqueurs codominants.

La nouvelle carte contient 465 marqueurs (268 SSR, 176 RFLP, 5 isoenzymes et 16 Rgenes-RFLP) repartis dans les dix groupes de liaison correspondant au nombre de chromosomes du cacaoyer. Sa taille est de 782.8 cm, avec une distance moyenne entre les marqueurs de 1.7 cm.

La présence de marqueurs codominants bien repartis tout au long du génome et en particulier les marqueurs microsatellites, permettra l'obtention d'informations plus précises pour la détection du polymorphisme de l'espèce. Cette nouvelle carte de référence fournit un grand nombre des marqueurs utiles pour l'identification et la comparaison des régions les plus importantes impliquées dans la variation des caractères d'intérêt (QTL). Elle permettra en particulier de comparer de façon précise la localisation des QTL entre différentes cartes génétiques établies sur des populations différentes. Ces marqueurs sont maintenant largement distribués et adoptés au sein de la communauté internationale, ce qui permettra désormais, un réel échange entre les différentes équipes. Elle permettra aussi une analyse fine, tout au long du génome, de la diversité des populations de *Theobroma cacao* et en particulier des Criollo moderne/Trinitario.

En raison du polymorphisme important révélé par les microsatellites et d'allèles spécifiques qui peuvent être mis en évidence pour certains groupes comme les Criollo et Forastero bas amazonien, cette carte sera particulièrement importante dans le cadre de notre étude d'associations effectuée sur une population de Criollo moderne/Trinitario.

CHAPITRE III :
DIVERSITÉ GÉNÉTIQUE D'UNE COLLECTION DE
VARIÉTÉS CRIOLLO MODERNES/TRINITARIO

CHAPITRE III

Diversité génétique des clones appartenant au groupe Criollo moderne/Trinitario de la collection du CATIE

Ainsi que nous l'avons décrit dans la révision bibliographique (Chapitre1), une longue histoire lie l'homme au cacaoyer. Pour les peuples de Meso Amérique, premiers à le cultiver, le cacao faisait partie de leur cosmogonie, de leur économie, de leurs rites religieux et de tous les aspects de leur vie quotidienne. Une fois « découvert » par les européens et transformé en chocolat tel qu'on le consomme aujourd'hui, sa culture s'est répandue dans de vastes régions d'Amérique centrale et du sud, dans les îles des Caraïbes, en Asie et particulièrement en Afrique d'où provient 60% de la récolte mondiale.

Malgré cette histoire commune, l'intérêt pour la conservation des ressources génétiques ne commence qu'en 1930. De nos jours, une quarantaine de prospections ont été réalisées et les cacaoyers cultivés et sauvages collectés dans le bassin amazonien et en Amérique centrale ont été rassemblées dans plusieurs collections nationales et internationales.

Des caractères morphologiques, des marqueurs isoenzymatiques et des marqueurs moléculaires (RAPD, RFLP et microsatellites) ont été utilisés pour caractériser la diversité génétique des populations des cacaoyers. L'ensemble de ces résultats a mis en évidence que la classification des variétés cultivées, basée sur deux groupes morpho-géographiques, et qui a longtemps été utilisée pour l'ensemble de l'espèce n'était pas adaptée pour décrire et classifier la diversité totale de l'espèce.

Des études récentes sur la domestication et l'origine des Criollo actuellement cultivés (modernes) et des Trinitario, ont permis d'identifier les principaux parents à l'origine de la majeure partie de ce groupe. Motamayor *et al.* (2003) ont montré que la plus grande partie de ces clones correspondent en fait à des Criollo « anciens » introgressés en majorité par quelques génotypes de Forastero bas-amazoniens. La diversité allélique de ce groupe peut être en grande partie expliquée par uniquement deux individus Criollo anciens et trois individus Forastero de Basse Amazonie (Motamayor *et al.*, 2002, 2003). A la lumière de ces résultats, il est très probable que le quasi totalité des Criollo modernes et des Trinitario, représentés en très grand nombre dans les collections correspond à une base génétique extrêmement réduite. Cette situation est favorable pour développer des études

CHAPITRE III. Diversité génétique des clones de la collection du CATIE

d'associations au sein de ce group génétique où environ 6 à 7 générations séparent les variétés actuelles des premières hybridations entre Criollo et Forastero. Pour ces raisons, nous avons choisi d'étudier une population de Criollo moderne/Trinitario.

La collection du Centro Agronómico Tropical de Investigación y Enseñanza (CATIE) au Costa Rica maintien environ 800 génotypes avec une représentation majoritaire de génotypes d'origine Criollo modernes/Trinitario. Ces accessions ont été classifiées comme des Trinitario ou comme des Criollo modernes à l'aide de caractères morphologiques. Parmi ces accessions un certain nombre ont été également évalués pour des caractères morphologiques (Engels, 1981). L'étude de la diversité génétique a donc été faite sur une population de Criollo modernes/Trinitario de cette collection CATIE à l'aide des marqueurs moléculaires de type microsatellites. Ces analyses nous permettront aussi d'évaluer la possibilité d'utiliser cette population pour les tests d'association avec une approche de déséquilibre de liaison. Les résultats de cette étude de diversité sont présentés dans l'article ci-après.

Genetic diversity of Modern Criollo/Trinitario cacao clones (*Theobroma cacao* L.) from CATIE germplasm collection assessed by microsatellite markers

Pugh T^{1,2}, Phillips W.³, Astorga C³, Courtois B¹, Noyer JL¹, Risterucci AM¹, Fouet O¹, Lanaud C¹

¹ CIRAD BIOTROP TA 40/03 34398 Montpellier Cedex 5, France

² UCV-FAGRO , Av El Limón, 2101 Maracay, Venezuela

³CATIE, Turrialba, Costa Rica

Abstract

A cacao germplasm collection of about 800 accessions of different origin is conserved in the CATIE (Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica). A significant proportion of these accessions correspond to modern Criollo/Trinitario. The genetic diversity of 247 modern Criollo/Trinitario varieties from this germplasm collection was evaluated using 34 microsatellites markers widespread in the cacao genome. A total of 175 alleles were detected with an average of 5.2 allele number per locus. Two alleles more frequent than the others were always found at each locus. These alleles were identified as alleles originated from Ancient Criollo and Lower Amazon Forastero, individuals recognized as the founder genotypes of the Modern Criollo/Trinitario varieties. Some rare alleles were also identified. The factorial analysis based on dissimilarities distances grouped accessions in four subgroups according to alleles' origin: the first and second subgroup grouped accessions wearing mainly alleles from ancient Criollo or from Lower Amazon Forastero ancestors, respectively. The third subgroup contained accessions intermediate between clusters I and II and corresponding to hybrids forms between the 2 main ancestors. The fourth subgroup contained accessions out grouped from other clusters and contained rare alleles. Among intermediate accessions, a large number of redundant accessions was identified, especially accessions belonging to the RIM and UF groups. A subset of 37 accessions of the whole collection including 99.5% of the total allelic diversity was established. The narrow genetic base of the modern Criollo/Trinitario group and the small number of generations happened from the formation of the first hybrids cross is a

favourable situation to develop association mapping studies in this genetic group.

Introduction

Theobroma cacao L. ($2n=20$) is a perennial tree belonging to the Malvaceae family (Whitlock *et al.*, 2001). This species comprises a large number of highly morphologically variable populations from different origins. Archaeological records have revealed evidence of cacao plantations in Central America dated to 500 years B.C. (Bergman, 1969; Paradis, 1979). Religious, medicinal and economic uses of cacao beans by the Mayan and Azteque peoples are well-documented (Dillinger *et al.*, 2000; Motamayor *et al.*, 2002). Once introduced to the Spanish royal court in middle 16th century, the chocolate drink spread throughout Europe. The growing demand allowed extending cacao plantations to a large number of tropical regions. Currently cacao is cultivated on over five million hectares of humid tropical lowlands worldwide. It is mainly produced by small farmers in traditional farming systems. Only 30% of this surface is planted with selected varieties (Eskes, 2001).

Traditional cacao cultivars have been usually classified in three morphogeographic groups, Forastero and Criollo distinguished basically by fruit, seed and quality characteristics (Cheesman, 1944; Cuatrecasas, 1964) and their hybrid form: Trinitario. Early studies have revealed that these cultivars represent a small part of the *T. cacao* genetic diversity and that this classification is not adapted to describe the total diversity and the relationships between populations of the whole species. Indeed, Forastero grouped a large number of wild and cultivated populations from South America that can be found from Guyana, lower Amazonia (Brazil), and the Orinoco Valley (Venezuela) to upper Amazonia (Brazil, Bolivia, Peru, Ecuador, and Colombia). Several authors suggested that *T. cacao* originated in the Upper Amazon region (Cheesman, 1944; Purseglove, 1968). Motamayor *et al.*, (2003) focused on Criollo domestication and could identify with molecular tools the first Criollo clones, called “ancient Criollo”, domesticated by meso-american populations. According to these authors, these Ancient Criollo clones were probably originated from a few individuals transported by humans from South America and spread throughout Central America where they were domesticated by the Maya and Azteque civilisations. Molecular markers revealed that Ancient Criollo were highly

homozygous and poorly polymorphic contrary to the modern varieties of Criollo included in the germplasm collections (Motamayor *et al.*, 2002, 2003). In fact, in the 18th century, hybrids between Ancient Criollo and Forastero from the lower Amazon, called Trinitario, were produced in the Caribbean region. Due to their weakness and less productivity, ancient Criollo types were progressively replaced in the plantations by Trinitario more vigorous and productive leading to a large genetic mixing between cultivars (Pittier, 1935; Cheesman, 1944). Modern Criollo varieties represent in fact hybrid forms having conserved most of the quality traits of the ancient Criollo. The diversity presently observed in modern Criollo/Trinitario cultivars is the result of different levels of recombination between a limited number of parental genomes: mainly an ancient Criollo and a few Forastero clones (Motamayor *et al.*, 2003).

T. cacao germplasm has been extensively collected and assembled since 1930. Collections of cocoa trees in germplasm centres are maintained in several countries (Motilal and Buttler, 2003). The most important collections are The International Cocoa Genebank, Trinidad (ICG,T) (rich in Forastero clones) and the Collection of the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE) in Costa Rica which have been designated as “Universal Collection Depositories” (IBGRI, 1981). The germplasm collection of CATIE was initiated in 1942 with the introduction of the UF clones from the United Fruit Co. The current collection comprises almost 800 accessions of different origin and among them a significant proportion of Modern Criollo/Trinitario.

Germplasm collections were originally set up to preserve the genetic diversity of cultivated species for the benefit of plant breeders and researchers. Often because of the large number of accessions, problems are encountered in documentation, conservation, multiplication and evaluation. Otherwise, the size of many large germplasm collections may be an obstacle for full exploitation, evaluation and utilization (Holden, 1984). Numerous efforts have been realized to characterize the genetic diversity conserved in cacao germplasm collections. These data have been compiled in the International Cocoa Germplasm Database ICGD 2000 v. 4.1 CD-ROM. This database contains the published records of almost 14.000 separated cacao accessions in global holding (Wadsworth and Harwood, 2000). The data provide information about each clone (names, origin, synonyms or homonyms, presence in germplasm collections, etc) and mainly agronomic or morphological traits.

Table 1. Group name and origin of the 247 cacao accessions from the CATIE germplasm collection used in this study. ICGD 2000 database (Wardsworth and Harwood, 2000)

Group name	Origin	Accessions representing each group
CC	(Cacao Center) Selections from various populations.	10, 35, 37-41, 43-49, 54, 67, 71, 74, 79, 83, 99, 100, 106, 107, 120, 121, 124, 132, 137-139, 143, 144, 152, 169, 173, 223-226, 228, 231, 232, 234-236, 244, 251, 254, 255, 265, 266
CHUAO	(CHUAO) Froma Valle de Chuao, Venezuela.	120
CNS	(Caño Negro Selection) Froma Caño Negro, Barinas, Venezuela.	22
CRIOLLO	(CRIOLLO) also known as CRIOLLO LOLITA, from Costa Rica	1, 3, 5, 7, 8, 10-15, 17-19, 21-23, 27, 28, 33-37, 41, 43, 48, 50, 52, 54-56, 60, 62-66, 215, 216
CU	(CUyamel) from Cuyamel, Honduras.	1
DR	(Djati Roennggo) Selections from a cross: Cundeamor type (Venezuela) x Criollo (Venezuela) made by van Hall, 1913	1, 38
EET	(Estación Experimental Tropical) Pichiluengue, Ecuador	75
ESMIDA	(ESMIDA) Tabasco, Mexico	ROJO, AMARILLO
G	(Getas state) From Indonesia, Java	8.23
GA	(Grande Anse Bay), Haiti	11
GC	Originally reported as a Jamaican accession, but there are not GC identifiers recorded in the Jamaican accession list.	7
GS	(Grenade Selection) Selection made in late '40s in Grenade.	7, 17, 36, 50, 78
ICS	(Imperial College Selections) Selections from farm in Trinidad (1-100) and from Criollo progenies (101-107).	1, 6, 8, 16, 29, 39, 40, 43, 44, 47, 53, 60, 61, 84, 89, 91, 95, 100, 117, 135, 137, 138
IQ	(IQuiri region) Brazil.	1
LF	(La Fortuna) Recorded as Costa Rica criollo selections in Engels, 1981 but as Samoan Forastero in Morera <i>et al.</i> , 1991	3
MT	(MonTes) Selections from Honduras, Cuyamel	1
OC	(OCumare de la Costa) Selections from cacao plantations, Ocumare de la Costa, Aragua State.	61, 77
P	(Particular) 1940-45 selections from Chiapas, Mexico.	8, 10, 23, 43
PENTAGONA	(PENTAGONA) The name refers to the five-sided fruit character which results from the fusing of ridge pairs to give appearance of five single ridges. Collection of pentagona type.	2, 8, 17
PMCT	(Programa para el Mejoramiento de Cultivos Tropicales) Selections from Costa Rica, Belize, Honduras and Nicaragua.	5-18, 20-22, 25-29, 31-34, 42, 44-54
PV	(Porcelana Venezuela) From Maracaibo region, Venezuela.	5
RIM	(Rosario Izapa Mexico) Selections from Hacienda La Rioja, Tuxclachico, Chiapas, Mexico.	2, 6, 8-10, 13, 15, 19, 21, 23, 24, 30, 34, 39, 41, 43, 44, 52, 56, 68, 71, 75, 76, 78, 100, 101, 105, 106, 113, 189, 418
SC	(Selección Colombiana) Valle Palmira, Colombia.	5, 6, 13
SGU	(Selecciones GUatimaltecas) Selected from farms in the State of Suchitepéquez, Guatemala.	89
SNK	(Selection N° Koenvone)	12
SPA	(Selección PALmira) Collections made in somewhere in the Amazon Valley, Colombia by Pound.	5, 7, 9, 11, 12, 17
STICA	(Servicio Técnico Internacional de Cooperación Agrícola) By Minister of Agriculture, Costa Rica	100
TJ	(TauJica) Taujica, Cuyamel, Honduras.	1
TSH	(Trinitario Selected Hybrid) Selections from mixed populations of Amazon and Trinitario types.	644
UF	(United Fruit selections) Selections from Atlantic coast made in 1930's. UF company from Costa Rica.	10-12, 168, 221, 296, 601, 613, 650, 654, 666-668, 672, 677, 704, 707-710

Nevertheless, little is known about the genetic variability of these germplasm collections at the DNA level. PCR-based markers as microsatellites or single sequence repeats (SSR) provide a useful tool for estimating the extent of genetic diversity among and within populations and species. SSR have been used to study the genetic diversity of many species. SSR consist of short tandemly repeated di, tri or tetra motifs distributed over the eukaryotes genomes. The polymorphism between genotypes is due to the variation in the number of repeat units and probably due to slippage during DNA replication or unequal crossing over (Levinson and Gutman, 1987). Fragment containing microsatellites can be amplified by PCR using a pair of primers flanking the repeat sequence. In cacao, microsatellites were first isolated by Lanaud *et al.* (1999); recently a genetic map saturated with 268 SSR was produced (Pugh *et al.*, 2004).

In this study, we provide an analysis of genetic diversity for 247 clones of modern Criollo/Trinitario, belonging to the CATIE germplasm collection, using 34 microsatellite markers distributed throughout the cacao genome. Characterization of this germplasm could be useful for proper utilisation and conservation, to rationalise the collection and to identify duplications.

Materials and methods

Plant material, DNA extraction and microsatellite markers analysis

The plant material was collected from the germplasm collection maintained by the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE) in Turrialba, Costa Rica. Two hundred and forty-seven accessions representing most of the available cultivars from the modern Criollo/Trinitario cultivar group from this germplasm collection were studied. A comprehensive list of accessions used along with group names and origin from each of them is given in Table 1. Information about groups' origin was obtained from the ICGD 2000 database (Wadsworth and Harwood, 2000). In addition to these accessions, 10 cultivars from different origin were included in the analysis (MAT 1-6, SIAL 70, IFC1, Catongo, STA1, Atelier, LAN-17, Zea 4, Scavina 6, GU349).

Thirty-four microsatellite markers were chosen widespread in the cacao genome at about 15-30 cM intervals based on the last published cacao linkage map (Pugh *et al.*, 2004). This map contains 465 markers (268 SSR, 176 RFLP, 5 isoenzymes and 16

Table 2. List of the 34 simple sequence repeats (SSR) used in this study

SSR name	EMBL accession number	LG location	Expected size (bp)	Motifs	Nº alleles (rare alleles)	He	Ho	PIC
mTcCIR7	Y16981	7	160	(GA) ₁₁	6(3)	0.584	0.631	0.62
mTcCIR26	Y16998	8	298	(TC) ₉ C(CT) ₄ TT(CT) ₁₁	7(4)	0.569	0.582	0.55
mTcCIR32	AJ27185	4	198	(CA) ₁₀	3(1)	0.532	0.586	0.43
mTcCIR63	-	9	161	(CGT) ₇	3(0)	0.542	0.571	0.58
mTcCIR73	AJ566420	2	112	(CT) ₄ TT (CT) ₂ G (TC) ₈	4(2)	0.529	0.577	0.54
mTcCIR79	AJ566425	9	108	(TC) ₈	5(2)	0.600	0.656	0.71
mTcCIR82	AJ566428	3	174	(AG) ₆ AA (AG) ₇	4(2)	0.546	0.594	0.46
mTcCIR84	AJ566429	1	136	(GA) ₁₁	5(3)	0.434	0.462	0.53
mTcCIR91	AJ566435	10	186	(CT) ₁₀	3(1)	0.518	0.569	0.56
mTcCIR100	AJ566444	2	244	(AG) ₆ C (AG) ₄	6(3)	0.544	0.631	0.55
mTcCIR109	AJ566452	5	162	(CT) ₁₂	4(1)	0.559	0.625	0.60
mTcCIR120	AJ566461	3	95	(AG) ₁₃	5(2)	0.561	0.600	0.59
mTcCIR121	AJ566462	1	138	(TG) ₁₂	6(3)	0.547	0.631	0.56
mTcCIR140	AJ566476	3	104	(CA) ₇	5(2)	0.589	0.664	0.49
mTcCIR163	AJ566492	8	194	(AG) ₉	2(0)	0.484	0.587	0.53
mTcCIR167	AJ566496	3	254	(GA) ₁₆	8(5)	0.599	0.570	0.59
mTcCIR182	AJ566510	6	148	(TG) ₉	3(0)	0.598	0.617	0.58
mTcCIR184	AJ566512	1	139	(CA) ₈ (CT) ₁₃	5(1)	0.616	0.654	0.45
mTcCIR186	AJ566514	7	147	(TG) ₈	4(2)	0.477	0.559	0.56
mTcCIR189	AJ566517	8	150	(GT) ₁₂	7(5)	0.574	0.634	0.60
mTcCIR190	AJ566518	7	166	(TG) ₁₂	7(5)	0.560	0.607	0.53
mTcCIR195	AJ566522	2	350	(CA) ₁₀	3(1)	0.460	0.500	0.55
mTcCIR215	-	9	197	(AG) ₁₃	5(2)	0.555	0.645	0.48
mTcCIR220	AJ566541	10	201	(TC) ₁₀	6(3)	0.561	0.562	0.48
mTcCIR221	AJ566542	4	273	(TC) ₉	4(2)	0.487	0.506	0.57
mTcCIR222	AJ566543	4	220	(GA) ₉	6(3)	0.564	0.531	0.58
mTcCIR229	AJ566550	10	307	(TC) ₈	6(4)	0.530	0.569	0.56
mTcCIR238	AJ566559	6	126	(AG) ₉	8(6)	0.531	0.525	0.54
mTcCIR242	AJ566563	4	287	(CT) ₉ (CA) ₉	4(0)	0.587	0.662	0.30
mTcCIR265	AJ566585	5	246	(AG) ₁₈	8(5)	0.583	0.650	0.60
mTcCIR266	AJ566586	9	192	(CT) ₁₅	7(4)	0.608	0.653	0.53
mTcCIR268	AJ566588	2	316	(GA) ₁₇ GG (GA) ₉	6(2)	0.707	0.640	0.56
mTcCIR275	AJ566594	1	146	(CT) ₁₁	3(0)	0.581	0.611	0.52
mTcCIR291	AJ566609	6	218	(CT) ₁₂	7(4)	0.546	0.585	0.31

Linkage group (LG), expected heterozygosity (He), observed heterozygosity (Ho), Polymorphism Information Content (PIC)

Rgenes-RFLP) arranged in 10 linkage groups corresponding to the haploid chromosome number of cacao and covers about 782.8 cM. The length of the map established with only SSRs represent 94.8% of the total map with approximately 1 microsatellite every 3 cM. A list of these markers with their chromosomal location, expected size of the amplification product in clone Catongo, motif and number of repeats are provided in Table 2.

The accessions were genotyped at CIRAD (Montpellier, France). DNA extraction was performed from fresh leaves according to Risterucci *et al.* (2000). The details of genotyping method have been described in Pugh *et al.*, (2004).

Data analysis

Microsatellite allele positions were scored. The number of alleles per locus, the allele frequency as well as the expected and observed heterozygosity (H_e and H_o , respectively) was calculated using the GENETIX software (Belkhir *et al.*, 2002). Based on the frequency of microsatellite alleles (p_i), the polymorphic information content (PIC) was calculated for each locus using the following formula: $PIC = 1 - \sum p_i^2$.

Genetic relationships between cacao genotypes were studied on the basis of SSR data. A matrix of dissimilarity based on simple matching coefficients (Sneath and Sokal, 1973) was calculated, and a NJTree was constructed with the unweighted NeighborJoining method on the basis of the matrix of pairwise d_{ij} values between individuals (Saitou and Nei, 1987). These analyses were performed using the Darwin 4.2 software (Perrier *et al.*, 2003).

The genetic redundancy of the collection and a subset of genotypes displaying a large part of the allelic richness present in the collection were defined using MSTRAT program (Gouesnard *et al.*, 2001). This software is able to build a germplasm core collection by maximizing allelic or phenotypic richness following the method proposed by Schoen and Brown (1993).

Results

Microsatellite markers polymorphism

Thirty-four microsatellite markers previously mapped were used to characterize and evaluate the genetic diversity of two hundred and forty-seven accessions from CATIE's

Table 3. Total number of alleles (Tna), average number of alleles per locus (Ana), total number of rare alleles (Tnra), average number of alleles per accession (Anra), Group specific alleles (Gsa), Expected heterozygosity (He) and Observed heterozygosity (Ho) in each origin group.

Group name	Nº accessions	Tna	Ana	Tnra	Anra	Gsa	He	Ho
CC	52	142	4.18	50	0.96	2	0.534	0.569
CHUAO	1	54	1.59	1	1	0	0.294	0.588
CNS	1	62	1.94	0	0	0	0.469	0.938
CRIOLLO	40	110	3.24	24	0.6	2	0.461	0.366
CU	1	57	1.68	2	2	0	0.338	0.677
DR	2	80	2.35	8	4	0	0.493	0.691
EET	1	39	1.15	1	1	0	0.074	0.147
ESMIDA	2	64	1.88	0	0	0	0.375	0.485
G	2	71	2.09	4	2	0	0.404	0.397
GA	1	46	1.35	1	1	0	0.177	0.353
GC	1	57	1.78	5	5	0	0.391	0.781
GS	5	86	2.53	5	1	0	0.380	0.488
ICS	22	110	3.24	21	0.95	1	0.544	0.735
IQ	1	50	1.52	20	20	5	0.258	0.515
LF	1	50	1.52	8	8	0	0.258	0.515
MT	1	54	1.64	9	9	0	0.318	0.636
OC	2	73	2.15	1	0.5	0	0.478	0.838
P	4	85	2.50	11	2.75	1	0.521	0.738
PENTAGONA	3	69	2.03	1	0.33	0	0.407	0.260
PMCT	38	126	3.71	35	0.92	4	0.522	0.431
PV	1	66	2.00	0	0	0	0.500	1.000
RIM	31	83	2.44	6	0.19	1	0.506	0.917
SC	3	72	2.12	5	1.67	0	0.487	0.843
SGU	1	67	1.97	0	0	0	0.485	0.971
SNK	1	48	1.41	2	2	0	0.210	0.412
SPA	6	84	2.47	17	2.83	1	0.420	0.563
STICA	1	56	1.65	0	0	0	0.324	0.647
TJ	1	50	1.47	2	2	0	0.235	0.471
TSH	1	47	1.47	10	10	1	0.234	0.469
UF	20	95	2.79	9	0.45	0	0.515	0.797
All clones	247	175	5.15	83	0.33	18	0.555	0.556

genebank. Values for the number of alleles/locus, PIC and the observed and expected heterozygosities are provided in Table 2.

Altogether 175 alleles were identified. The number of alleles detected per locus was not equivalent among loci. It varied from 2 to 8 with an average number of 5.2 per locus. Nevertheless, for each microsatellite locus two alleles were always present with highest frequency compared to the others. Eighty-three alleles (47.4 % of the total) were classified as rare alleles, due to their low frequency across accessions (< 5%). A large number of unique alleles were found in only 17 of the 247 accessions. 20 rare alleles were found in IQ1 and 13 in SPA9 and SPA11.

The expected heterozygosity (H_e) varied from 0.460 to 0.616 per locus with an average of 0.555 whereas the observed heterozygosity (H_o) ranged from 0.462 to 0.664. The average observed heterozygosity and the expected heterozygosity were very similar (0.555 and 0.556 respectively). Considering all accessions, PIC values ranged from 0.32 for mTcCIR242 to 0.71 for mTcCIR79.

The clones were classified according to their group name. Hereby, 30 group names were identified. The largest group was 'CC' with 52 accessions whereas only one accession was from the following group names: 'CHUAO', 'CU', 'CNS', 'EET', 'GA', 'GC', 'IQ', 'LF', 'MT', 'PV', 'SGU', 'SNK', 'STICA', 'TJ' and 'TSH'. In spite of the large differences in the number of accessions in each group, we have calculated the total and average number of alleles per locus, the total number of rare alleles, the average number of alleles per accession, the group specific alleles and the expected and observed heterozygosity for each group name (Table 3). The highest average number of alleles per locus was detected in CC group (52 individuals) with 4.18, followed by PMCT group (38 individuals) with 3.71. Eighteen alleles classified as rare alleles were genotype or group-specific. IQ1 with 5 and PMCT group with 4 were the genotype and group with the highest number of unique alleles.

The expected heterozygosity per group or genotype varied between 0.074 in EET to 0.544 in ICS with an average of 0.555. The observed heterozygosity had a wider range of variation: between 0.147 (EET) to 1.000 (PV). For each group or genotype, the average H_o is higher than H_e , except for CRIOLLO, PENTAGONA and PMCT groups.

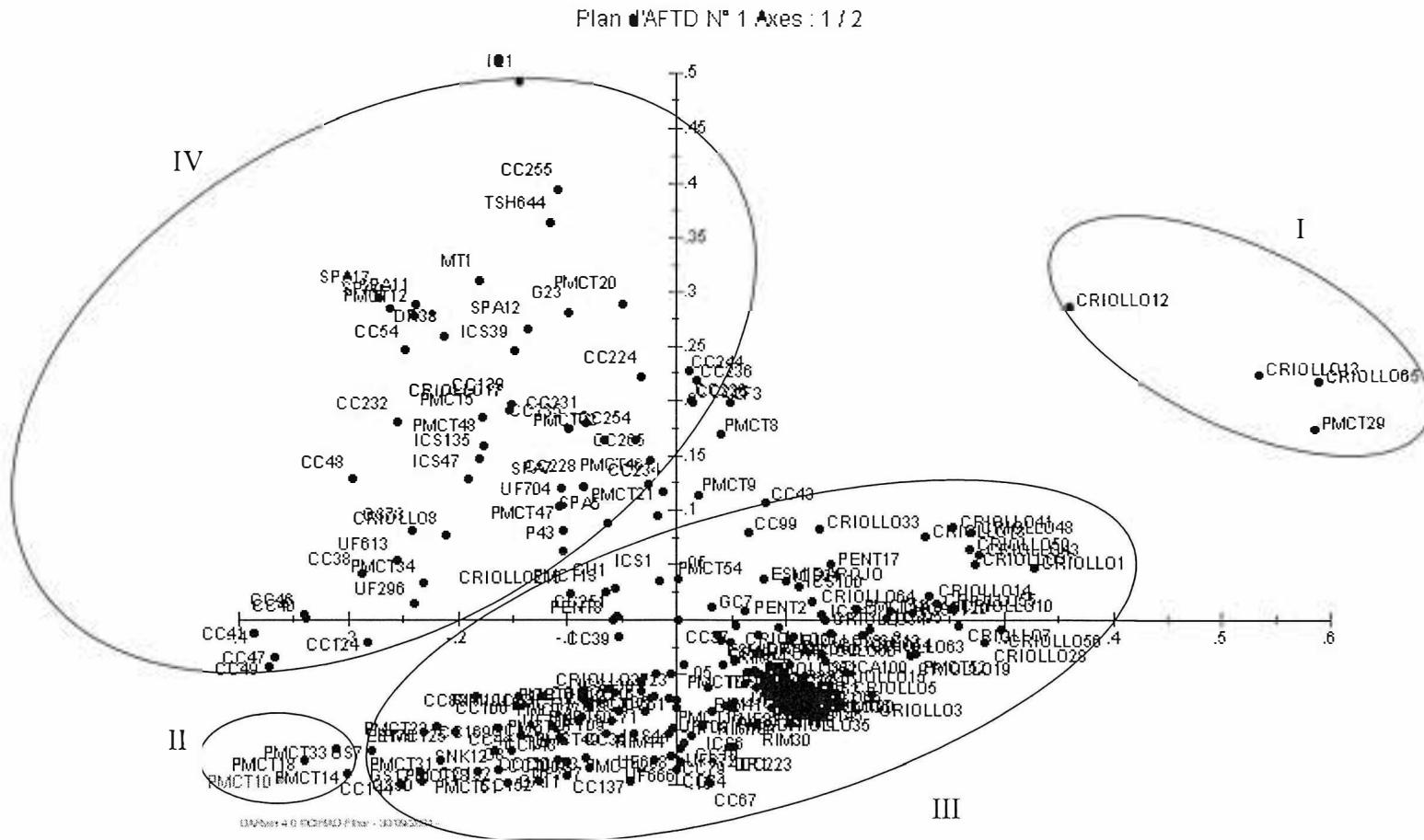


Figure 1. Factorial analyses based on dissimilarities distances evaluated using 34 SSR markers. The ellipses indicate I) accessions with mainly Ancient Criollo alleles; II) accessions with mainly Lower Amazon Forastero alleles; III) accessions intermediate (and representing hybrid forms) between I and II, and IV) accessions with other alleles than Lower Amazon Forastero allele.

Genetic diversity within accessions

To visualize the genetic relatedness among the cacao accessions, a factorial analysis based on a distance table was performed using the SSR data of 247 accessions (Figure 1). Four subgroups of accessions could be identified based on allele origin (labelled I, II, III and IV). Allele origin was established from reference genotypes carrying specific alleles such as LAN17 and Zea4 already identified as Criollo and MAT 1-6, SIAL 70, IFC1, Catongo identified as Lower Forastero ancestor (Motamayor *et al.* 2002, 2003). The first subgroup grouped accessions with mainly alleles from ancient Criollo and the second one grouped accessions with mainly alleles from Lower Amazon Forastero while the third subgroup contained accessions intermediate between subgroups I and II. The fourth subgroup contained accessions out grouped from the three other subgroups and that carried rare alleles. It showed that there was no correlation between the marker-based grouping and group name.

Redundancy and core set of accessions

We constructed a subset of accessions which will give a good representation of the allelic diversity present in the whole collection. This subset collection is represented by 37 clones (15 % of total) (Table 4). These accessions possessed alleles present in the 247 accessions tested, i.e. 99.5 % of the allelic diversity of the total allelic diversity and included all of the 83 rare alleles.

Discussion

Currently, microsatellite markers are commonly used to assess the genetic diversity. Microsatellites have several advantages over other types of markers, notably their ability to detect multiple alleles. In the present study, the number of alleles revealed for each locus ranged from 2 to 8 microsatellites with an average number of 5.2 per locus. A higher average number of alleles (18.1) was found when the whole CATIE germplasm collection was evaluated using 15 microsatellites (D. Zhang, personal communications). The whole collection comprises an important proportion of Forastero clones from the Upper and Lower Amazon, French Guyana and Ecuadorian Nacional varieties not included in our sample. Comparing with the whole collection, the accessions belonging to Modern

Table 4. Accessions included in the “core collection” defined in this study as containing 99.5% of the total allelic diversity of the Criollo moderne/Trinitario genetic group.

CC	40, 41, 48, 99, 107, 124, 139, 169, 226, 251, 255, 265, 266
CRIOLLO	19, 34, 66
DR	38
ESMIDA	AMARILLO
GS	36
ICS	6, 61, 95
IQ	1
P	43
PMCT	5, 47, 48
PV	5
RIM	24, 41, 76
SPA	5, 9, 11
TSH	644
UF	168, 296, 601, 613

Criollo/Trinitario seem to have a limited genetic diversity and correspond to a small proportion of the genetic diversity included in CATIE's germplasm collection

Allelic variation may be correlated with the number of repeats within a particular microsatellite locus. Microsatellite markers are often presumed to follow a stepwise mutation process in which alleles change in size by only one repeat unit (Shriver *et al.*, 1993). According to this mutation model, it seems that microsatellites with dimeric units should be more polymorphic than composed microsatellites. In our study, there was no correspondence between the number of alleles displayed and the type of microsatellites used. A composed microsatellite such as mTcCIR26 revealed the same number of alleles than a dimeric microsatellite such as mTcCIR 189, mTcCIR190 or mTcCIR 266. It suggests that the number of alleles detected in cacao could be more a function of the sample used than of the type of SSR motif. A positive correlation between type of motif and the number of alleles displayed have been found for tomato (He *et al.* 2003) and ryegrass (Jones *et al.*, 2002) but not in other species such Cucumis (Danin-Poleg, 2000).

Two alleles more frequent than the others were always found at each locus (Table 3). Comparing with some reference genotypes, we were able to identify these most-frequent alleles as alleles originating from Ancient Criollo (LAN-17 or Zea4) and Lower Amazon Forastero parental individuals (MAT 1-6, SIAL 70, IFC1 or Catongo). These individuals were recognized as the founder genotypes of the modern Criollo/Trinitario cultivars (Motamayor *et al.*, 2003). The global heterozygosity levels as well as the individual heterozygosity values also showed that almost all of Modern Criollo and Trinitario varieties presents in this collection are hybrid types resulting from introgression of Lower Amazon Forastero genotypes into ancestral Criollo (data not shown). In fact, almost 65% of all accessions correspond to hybrids between these two founder genotypes.

Globally, the alleles specific to only one group or one accession were rare except for some individuals as IQ1, SPA9 and SPA11 cumulating rare alleles. Unique alleles could also be an indication of the mutation at SSR loci or the inclusion of exotic germplasm in our sample. IQ1 comes from Iquira region, Brazil and the SPA group was collected by Pound in the Amazon Valley, Colombia. Putative Forastero parental genomes different from low Amazon Forastero should be involved. Thirteen percent of rare alleles were identified as Scavina 6 (Sca6) alleles and eight percent as GU349 alleles. Sca6 is an upper

Amazon Forastero representing a well known source of resistance to *Crinipellis perniciosa*. GU349 corresponds to a French Guiana clone.

Cacao germplasm has often been collected based on morphological characters as fruits, beans form and colour but also on agronomic characteristics of production or vigour. In the CATIE germplasm collection Modern Criollo/Trinitario clones were classified mainly based on these morphological observations. The first studies related to cacao genetic diversity considered these Modern Criollo/Trinitario clones as representative of the Criollo group. The modern Criollo varieties correspond in fact, to an ancestral Criollo type more or less introgressed by Forastero genes. Differences between clones could be the result of different levels of recombinations between the Criollo and Lower Amazon Forastero ancestors genome. The small number of alleles found in most of this groups' accessions confirms that Trinitario genetic basis is very narrow. These analyses agree with Motamayor *et al.*'s studies (2002, 2003) concluding that a small number of parents were involved in the genetic basis of modern Criollo cacao.

In the Factorial analysis, some of the genotypes were found overlapping each other depicting redundancy, especially in subgroup III. Many of these accessions belong to the UF and RIM groups. We constructed a neighbour-joining tree based on dissimilarities values for these overlapping accessions including all accessions from UF and RIM group (Figure 2). These groups of accessions include genotypes from different geographic origins as Mexico, Costa Rica Guatemala, Grenade, Trinidad or Venezuela. It reveals a high number of very close or similar varieties. Crouzillat *et al.* (2000) have found similar results with some of these clones. The high level of heterozygosity of same of these accessions indicates that they could be the results of the first generations of recombinations between an Ancient Criollo and a Lower Amazon clone.

The molecular characterisation of this germplasm collection would allow conserving an appropriate number of clones representing the genetic diversity of Trinitario cacao without duplicated or closely related accessions. It is impractical to characterize all collected accessions in detail. A sub set of the whole collection (core collection) including a maximum of the genetic variation contained in the whole collection with minimum of repetitiveness has been proposed (Frankel and Brown, 1984). Schoen and Brown (1993) have proposed a strategy, called M or maximization which examines all possible core collections and singles out those that maximize the number of observed alleles at the

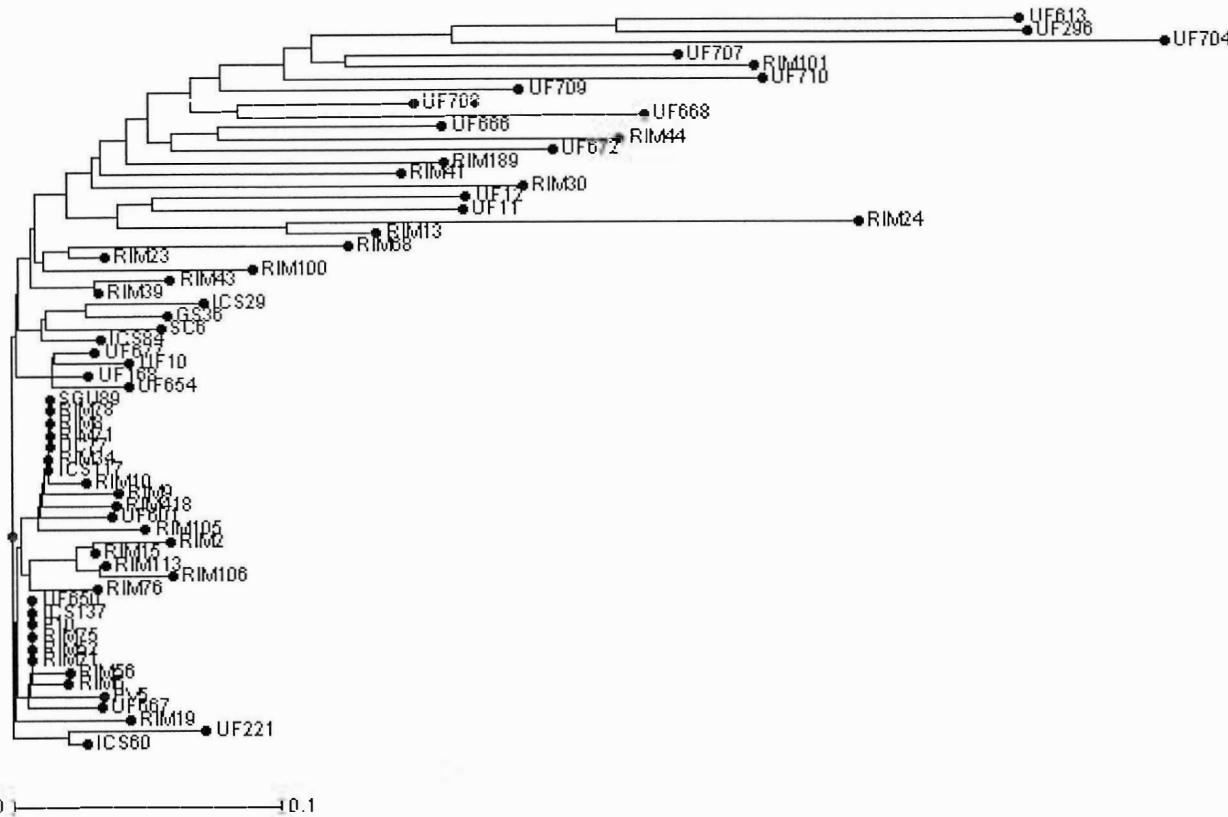


Figure 2. Neighbour-joining tree based on dissimilarity values of redundant accessions and including all accessions from UF and RIM group

marker loci. This is the basis of MSTRAT software (Gouesnard *et al.*, 2001). Only thirty seven accessions (15%) possessed 99.5 % of the total allelic diversity and included all of the 83 rare alleles (Table 5). Nevertheless, it is well-understood that these accessions represent the allelic diversity and not the genotypic diversity present in this collection. The core collection is not intended to replace the whole collection; it allows thinking about the right size of the collection and the number of existing duplicated accessions.

Genetic diversity is the basis of genetic improvement. During the last century, the main strategy used to improve yield and resistance to main diseases and pest in cacao was to cross a number of available Upper Amazon Forastero types collected by Pound with cultivated varieties, but this number was limited. It means that only a small part of the genetic diversity has been exploited in selection (Bartley 1979; Lockwood and End, 1992; Lockwood, 2003). In the case of Modern Criollo/Trinitario, a smaller number of Forastero genotype is at the origin of these hybrid varieties, and consequent genetic gain would be obtained exploiting the large diversity of Forastero group. Nevertheless, the structure of this particular genetic group resulting from a limited number of generation from the first hybrids, offers appropriate conditions to develop association mapping studies. Indeed, large cultivated areas or collections of Modern Criollo/Trinitario already exist and have been well characterised in the case of collections. The availability of both large adult trees population and data could accelerate and facilitate the identification of genetic bases of traits of interest for more effective crops-improvement programs based on marker assisted selection.

Acknowledgments

We thank the CATIE for providing us with the plant material studied.

References

- Bartley B.G. 1979. Global concepts for genetic resources and breeding in cacao. In: Seventh International Cocoa Research Conference, London, Cocoa Producers' Alliance, Douala, Cameroon, pp. 519-525.
- Belkhir K., Borsa P., Chikhi L., Raufaste N., Bonhomme F. 1996-2002 GENETIX 4b.04, logiciel sous windows TM pour la génétique des populations, pp. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier.
- Bergman, J.F. 1969. The distribution of cacao cultivation in pre-Columbian America. Annals of the Association of American Geographers, 59:85-96.
- Cheesman, E.E. 1944. Notes on the nomenclature, classification and possible relationships of cocoa population. Tropical Agriculture 21:144-159.

CHAPITRE III. Diversité génétique des clones de la collection du CATIE

- Crouzillat D., Bellanger L., Rigoreau M., Bucheli P., Pétiard V. 2000. Genetic structure, characterisation and selection of Nacional cocoa compared to other genetic groups. Proceedings of the 3rd International Group for Genetic Improvement of Cocoa (Ingenic) International Workshop on the New Technologies and Cocoa Breeding, 16th-17th October 2000, Kota Kinabalu, Malaysia, pp. 47-64.
- Cuatrecasas J. 1964. Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contributions from the United States Herbarium 35: 379-614.
- Danin-Poleg Y., Reis N., Baudracco-Arnas S., Pitrat M., Staub J.E., Oliver M., Arus P., deVicente C.M., Katir N. 2000. Simple sequence repeats in Cucumis mapping and map merging. Genome 43:963-74.
- Dillinger, T. L., Barriga P., Escarcega S., Jimenez M., Salazar Lowe D., Grivetti, L. E. 2000. Food of the gods: cure for humanity? A cultural history of the medicinal and ritual use of chocolate. J. Nutr. 130:2057S-2072S.
- Eskes, B. 2001. Introductory notes. Proceedings of the International Workshop on New Technologies for Cocoa breeding, Kota Kinabalu, Malaysia, INGENIC, London, UK, pp 8-11.
- Frankel O.H. and Brown A.H.D. 1984. Plant genetic resources today: a critical appraisal. In: Crop genetic resources: conservation & evaluation (Holden JHW and Williams JT, eds). London: George Allen & Unwin; 249–257.
- Gouesnard, B., Bataillon T.M., Decoux G., Rozale C., Schoen D.J., David J.L. 2001. MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. 92:93-94.
- He C., Poysa V., Yu K. 2003. Development and characterization of simple sequence repeat (SSR) markers and their use in determining relationships among *Lycopersicon esculentum* cultivars. Theor. Appl. Genet. 106:363-73.
- Holden J.H.W. 1984. The second ten years. In: Crop genetic resources: conservation and evaluation, (eds J.H.W Holden and J. Williams), George Allen and Unwin, Winchester, Massachusetts, pp. 277–285.
- IBGRI. 1981. Report of the IBPGR working group on genetic resources of cocoa. Rome, Italy, IBGRI, ACP: IBPGR /80/56, 28p.
- Jones S., Dupal P., Dumsday L., Hughes J., Forster W. 2002. An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). Theor Appl Genet. 105:577-584.
- Lanaud C., Risterucci A.M., Pieretti I., Falque M., Bouet A., Lagoda P.J.L. 1999. Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8:2142-2152
- Levinson G. and Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203-221
- Lockwood G. and End M. 1992. History, technique and future needs for Cocoa collection. In: International Workshop on the conservation, characterization and utilization of cocoa genetic resources in the 21st Century, Port of Spain, Trinidad and Tobago, pp 1-14.
- Lockwood, R. 2003. Who needs clothing? INGENIC News! 8:2-5.
- Motamayor, J.C., Risterucci A.M., Lopez P.A., Ortiz C.F., Moreno A., Lanaud C. 2002. Cacao domestication I. The origin of the cacao cultivated by the Mayas. Heredity 89:380-386.
- Motamayor, J.C., Risterucci A.M., Heath M., Lanaud C. 2003. Cacao domestication II. Progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322-330.
- Motilal L. and Butler D. 2003. Verification of identities in global cacao germplasm collections. Genetic resources and Crop Evolution. 50:799-807.

CHAPITRE III. Diversité génétique des clones de la collection du CATIE

- Paradis, L. 1979. Le cacao précolombien: monnaie d'échange et breuvage des dieux. *Journal d'agriculture traditionnelle et de botanique appliquée*, 26:3-4.
- Perrier, X., Flori A., Bonnot F. 2003 Methods of data analysis. In: Hamon P., Seguin M., Perrier X., Glaszmann J.C. (eds.) *Genetic Diversity of cultivated tropical plants*, pp 43-76, CIRAD, Montpellier, France.
- Pittier H. 1935. Degeneration of cacao through natural hybridization. *The journal of heredity* 36:385-390.
- Purseglove, J.W. 1968. *Theobroma* L. In: Purseglove, J.W. (ed) *Tropical Crops. Dicotyledons 2*. John Wiley & Sons, New York, pp 291-295.
- Pugh T., Fouet O., Risterucci A.M. , Brottier P., Abouladze M., Deletrez C., Courtois B., Clement D., Larmande P., N'Goran J.A.K., Lanaud C. 2004. A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* 108:1151-1161.
- Risterucci A.M., Grivet L., N'Goran J.A.K., Pieretti I., Flament M.H., Lanaud C. 2000. A high density linkage map of *Theobroma cacao* L. *Theor Appl Genet* 101:948-955
- Saitou N., and M. Nei. 1987. The neighbour joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and Evolution* 4(4), 406-425.
- Schoen D. J. and Brown A.H.D. 1993. Maximizing allelic diversity in core collections of wild crop relatives: the role of genetic markers. *Proc. Natl. Acad. Sci. USA* 90:1063-1067.
- Shriver, M.D., Jin, L., Chakraborty, R. and Boerwinkle, E. 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, 134, 983-993
- Wadsworth, R.M. and Hardwood, T. 2000. International Cocoa Germplasm Database ICDG version 4.1. London International Financial Futures and Options Exchange and the University of Reading, UK.
- Whitlock, B., Bayer C., Baum D. 2001. Phylogenetic relationships and floral evolution of the Byttherioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene ndhF. *Sys. Botany* 26:420-437.

Conclusions et Perspectives

La diversité génétique de 247 accessions de la collection du CATIE appartenant au groupe Criollo Modernes/Trinitario a été étudiée à l'aide de 34 marqueurs microsatellites repartis tout au long du génome du cacaoyer. Ces analyses ont montré une faible structuration de la diversité génétique au sein de ce groupe. En fait, malgré un nombre variable d'allèles détectés par locus (de 2 à 8), deux allèles étaient toujours plus fréquemment représentés que les autres. Ces allèles correspondent aux 2 ancêtres parentaux principalement à l'origine de ce groupe : un individu Criollo ancien et un individu Forastero bas amazonien (Motamayor *et al.*, 2003). Ces résultats confirment l'idée que les différences morphologiques observées entre variétés refléteraient principalement, la diversité de parties du génome Forastero introgressés dans les Criollo Anciens. D'autres allèles considérés comme allèles rares de par leurs basses fréquences, ont été aussi identifiés. Une analyse factorielle basée sur le tableau des distances génétiques a permis d'identifier quatre sous groupes sur la base de l'origine des allèles : deux groupes très proches des deux principaux ancêtres parentaux : Criollo ancien et Forastero bas Amazonien, un groupe intermédiaire hybride entre ces deux ancêtres et un quatrième group correspondant à des hybrides entre Criollo ancien et autres type de Forastero ayant apporté des allèles rares.

Ces résultats apportent des informations utiles pour la gestion de cette collection. Etant donné le grande nombre d'individus hybrides et la faible variabilité qui existe parmi eux, on a pu identifier des groupes d'individus très redondants, ainsi qu'un échantillon réduit d'accessions qui regroupe 99.5% de la diversité allélique. Il sera également important de prendre en compte des mutations ponctuelles portant sur des caractères d'intérêt agronomique pour aider à définir une « core collection ».

Par ailleurs, ces résultats permettent d'envisager une étude des déséquilibres de liaison entre marqueurs moléculaires et gènes d'intérêt dans cette population. En effet, ils confirment qu'un nombre restreint des génotypes parentaux sont à l'origine des ces hybrides Criollo/Trinitario après un petit nombre de générations de recroisement entre ces ancêtres parentaux. Ces hybrides ont été largement utilisés dans les programmes de sélection et conservés dans de nombreuses collections où ils ont été caractérisés pour un

CHAPITRE III. Diversité génétique des clones de la collection du CATIE

grand nombre de caractères morphologiques. Cette situation nous permet d'envisager de faire des études d'association en exploitant les données accumulées pour caractériser ces collections.

CHAPITRE IV :
ETUDES D'ASSOCIATIONS ENTRE MARQUEURS
MOLÉCULAIRES ET CARACTÈRES
MORPHOLOGIQUES DANS UNE COLLECTION DE
CRIOLLO/TRINITARIO

CHAPITRE IV

Études des associations entre marqueurs moléculaires et caractères morphologiques dans la collection de variétés de Criollo moderne/Trinitario du CATIE

Le développement des marqueurs moléculaires a permis la construction de cartes génétiques saturées qui constituent un outil important pour localiser les régions chromosomiques impliquées dans la variation des caractères d'intérêt (QTL). Cette approche (Cartographie classique des QTL) exploite le déséquilibre de liaison (DL) qui existe dans une population en ségrégation après une ou un faible nombre de générations de recombinaisons. Au cours de ces dernières années, la cartographie classique a été utilisée pour localiser des QTL chez de très nombreuses espèces.

De nouvelles approches (études d'association ou « association mapping ») exploitent le DL qui existe au sein des populations naturelles ou issues de sélection pour détecter des associations. Le déséquilibre de liaison ou déséquilibre gamétique, mesure l'association non aléatoire d'allèles pris à des locus différents. Le principe est d'exploiter le DL existant dans ces où des fragments de taille variable du génome ont pu être conservés au cours de l'évolution ou de la domestication de l'espèce. Chez des espèces pérennes, comme le cacaoyer, l'obtention de grandes descendances issues de croisements contrôlés n'est pas toujours évidente. Par contre, on peut disposer de populations cultivées ainsi que de grandes collections de ressources génétiques qui ont été très souvent évaluées pour un grand nombre de caractères morpho-agronomiques. C'est le cas de la collection internationale du CATIE (Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica) où un grand nombre d'individus de type Criollo moderne/Trinitario ont été évalués il y a 25 ans par Engels *et al.* (1981). Ces données sont toujours disponibles et stockés dans l'« International Cocoa Germplasm Database » (ICGD) (Wadsworth et Harwood, 2000). Les variétés actuelles de Criollo moderne/Trinitario sont le résultat de croisements entre un Criollo ancien et un nombre réduit d'individus Forastero. Le génome de ces variétés correspond en fait, à une mosaïque de régions chromosomiques correspondant principalement à ces 2 types de cacaoyer. Le nombre relativement faible de méioses qui séparent les types hybrides actuels des premières hybridations devrait avoir

maintenu un DL entre marqueurs moléculaires et caractères d'intérêt variables entre Criollo et Forastero. En se basent sur cette hypothèse, nous avons donc réalisé, sur la collection de Criollo modernes/Trinitario du CATIE, une étude d'associations entre marqueurs microsatellites repartis tout au long du génome et caractères morpho-agronomiques. Dans une première étape, nous avons étudié l'étendue du DL entre marqueurs microsatellites dans des régions ciblées du génome où des QTL ont été déjà identifiés, afin d'établir la finesse de maillage moléculaire du génome nécessaire aux études d'association. La seconde étape a consisté à étudier les associations entre marqueurs microsatellites et caractères d'intérêt. Les résultats de cette étude sont présentés dans l'article suivant.

Molecular marker-trait associations in a Criollo/Trinitario cacao (*Theobroma cacao* L.) germplasm collection

Pugh T^{1,2}, Courtois B¹, Engels, J.M.M³, Phillips W³, Astorga C³, Risterucci AM¹, Fouet O¹, Lanaud C¹

¹ CIRAD BIOTROP TA 40/03 34398 Montpellier Cedex 5, France

² UCV-FAGRO , Av El Limón, 2101 Maracay, Venezuela

³CATIE, Turrialba, Costa Rica

Abstract

Cocoa domestication has shaped the genome structure of modern Criollo/Trinitario varieties, a high quality and aromatic chocolate source. These varieties have originated from a reduced number of Criollo and Forastero ancestors first crossed 250 years ago. About 6 or 7 generations of recombinations separate the modern varieties from the first hybrids. One hundred and fifty modern Criollo/Trinitario varieties present in the CATIE germplasm collection (Costa Rica) were used to study the extend of linkage disequilibrium (LD) conserved along the genome. A first set of 24 SSR were used to verify the absence of population structure in this collection and the genome-wide LD. One hundred and ten SSR, with alleles specific to Criollo ancestors were used to evaluate the extend of LD. LD values decreased with an increasing genetic distance between loci, the genetic distance with LD being variable and up to 30 cM. This situation was appropriate to conduct molecular markers/traits association studies. The CATIE collection was characterised 25 years ago and data stored in the International Cocoa Gene Bank. The data, related to fruit, seed and flower traits were used for association studies. A total of 13 genomic regions were identified as involved in the variation of these traits. Among them 8 corresponded to genomic regions where QTL (quantitative trait loci) were already identified by classical QTL mapping studies. These results demonstrate that association studies approaches represent a valuable tool to help to identify the genetic or molecular bases of traits of interest and to valorise hundred of morphological and agronomic data accumulated to characterise large germplasm collections.

Introduction

Theobroma cacao L. ($2n=20$), a perennial tree belonging to the Malvaceae family (Whitlock *et al.*, 2001) is native to the American tropics. Cacaos seeds, fermented and dried, are a major cash crop for a large number of developing countries. Cheesman (1944) distinguished two morphogeographic groups, Criollo and Forastero. Pure Criollo varieties, now recognized as providing a high quality aromatic chocolate were the first domesticated by Mesoamerican populations more than 2000 years ago. Trinitario, a hybrid group between them was produced in the 18th in the Caribbean region. Recent studies on Trinitario diversity have shown that the Trinitario group resulted mainly from hybridisations between two almost completely homozygous Criollo and Forastero individuals (Motamayor *et al.*, 2002, 2003). Trinitario clones gradually spread into pure Criollo plantations, leading to further recombinations between Criollo and Trinitario individuals. It resulted that the modern Criollo varieties, selected by their quality traits, and Trinitario, both hybrid forms between the same ancestors, have a similar genetic structure.

The development of molecular markers and linkage maps has provided a powerful tool to map quantitative trait loci (QTL) controlling main agronomic traits using mapping populations (e.g. F2 or backcross). During the last decade, hundreds of these marker-trait association studies using mapping populations have been reported in plants. In cacao, several linkage maps have been published (Lanaud *et al.*, 1995; Crouzillat *et al.*, 1996; Risterucci *et al.*, 2000; Pugh *et al.*, 2004). These maps have been successfully used to locate QTL affecting traits of interest such as disease resistance and yield factors (Lanaud *et al.*, 1999 ; Crouzillat *et al.*, 1996, 2000a, 2000b; Clement *et al.*, 2003a, 2003b; Flament *et al.*, 2001 ; Lanaud *et al.*, 2003; Risterucci *et al.*, 2003). However, in cacao the long generation time needed to obtain large segregation populations makes this approach difficult to be applied. New approaches, exploiting linkage disequilibrium (LD) existing in natural or cultivated populations, and recently developed in other species could be applied.

Linkage disequilibrium is defined as the non-random association of population alleles at two or more loci (Lewontin and Kojima, 1960). A marker and a trait locus in LD will show an association between the marker locus and the phenotype controlled by the trait locus. Of particular interest are associations between loci tightly linked on the same chromosome. In the ideal case of an infinite random mating population, in the absence of

mutation, migration or selection, the LD value is only function of the recombination probability between loci. In natural populations, LD reflects the size of chromosome segments remaining intact in the population and that have been inherited together since the population was founded or subjected to a bottleneck.

The fundamental difference between a traditional QTL mapping and an association mapping studies using LD is the strength of linkage disequilibrium in the population. In an association mapping study, individuals belong to a population where recombinations over many generations diluted LD and has only maintained significant associations between markers and trait loci. In theory, LD mapping is expected to be more informative and with higher resolution than a traditional QTL mapping approach. LD is also a good indicator of recent mutations, genetic drift, bottlenecks, stratification or admixture, and of the demographic history of population (Hill and Robertson, 1968; Nei and Li, 1973; Kruglyak, 1999, Flint-Garcia *et al.*, 2003).

While LD has been extensively used for mapping diseases in humans, its use in plants has just begun as related in some reviews (Flint-Garcia *et al.*, 2003; Rafalski and Morgante, 2004). Conclusions indicate that this approach could improve molecular marker-trait associations studies. Association mapping studies can notably exploit pre-existing data accumulated in germplasm collection that have been extensively characterized.

The few recombination generations that have occurred since the formation of the modern Criollo/Trinitario varieties may have maintained associations between molecular markers and specific Lower Amazon Forastero and Criollo traits. A LD approach could be used to detect these associations. The aims of this study were 1) To evaluate the extent of LD genome-wide using unlinked SSR; 2) To measure the rates at which LD decays on different chromosomal regions using a high density SSR marker map 3) to test associations between these markers and trait of interest in a Modern Criollo/Trinitario population.

Table 1. List of accessions used in this study.

Group name	Accessions representing each group
CC	10, 35, 37-41, 43-45, 46, 47-49, 54, 67, 71, 74, 79, 83, 99, 100, 106, 107, 120, 121, 124, 132, 137, 138, 139, 143, 144, 152, 169, 173, 223-226, 228, 231, 232, 234-236, 244, 251, 254, 255
CHUAO	120
CNS	22
CRIOULLO	1, 3, 5, 7, 8, 10-15, 17-19, 21-23, 27, 28, 33-37, 41, 43, 48, 50, 52, 54-56, 60, 62-66, 215, 216
CU	1
DR	1, 38
EET	75
ESMIDA	AMARILLO
G	8, 23
GA	11
GS	36, 78
ICS	1, 6, 8, 16, 39, 40, 43, 44, 47, 53, 84, 89, 91, 95, 100, 117 , 135, 137 , 138
IQ	1
LF	3
MT	1
OC	77
P	8, 10, 23, 43
RIM	2, 6, 8, 9, 10, 13, 15, 19, 21, 23, 24, 30, 39, 41, 43, 44, 52, 56, 68, 71, 75, 76, 78, 100, 101, 105, 106, 113
SC	5, 6, 13
SGU	89
SNK	12
SPA	5, 7, 9, 11, 12, 17
STICA	100
TJ	1
TSH	644
UF	10, 11, 12, 168, 221, 296, 601, 613, 650, 654, 666-668, 672, 677, 704, 707-710

Individuals in bold were eliminated for the marker-trait association studies because of their redundancy.

Materials and methods

Plant material

One hundred and fifty cacao clones belonging to the Criollo/Trinitario group were obtained from the germplasm collection maintained by the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Turrialba, Costa Rica (Table 1). The 150 accessions are part of a larger collection of almost 800 accessions. These accessions were selected because of their characterisation already made for several traits by Engels (1981). The genetic diversity of these clones had also been previously analysed using SSR (Pugh *et al.*, in preparation).

Markers and genotyping

Two data sets were used a) Data set 1: Twenty-four unlinked simple sequence repeats (SSR) distributed on the 10 chromosomes of cacao. This data set was employed to analyse population structure and genome-wide LD; b) Data set 2: One hundred and twenty four SSR were used to genotype all individuals. The 124 markers had an average spacing of 14.3 cM across the cacao genome. Four linkage groups (LG) were explored with a higher density of markers: LG1 (20), LG3 (18), LG4 (24) and LG9 (21) spanning 80.5, 68.3, 63 and 91.5 cM with an average marker spacing of 2.5, 3.8, 2.5 and 4.4 cM respectively. QTLs for main agronomic traits have been still mapped in these LG (Lanaud *et al.*, 1999; Clement *et al.* 2003a, 2003b). SSR were used to evaluate the extent of LD across these regions. The same 124 markers were used for marker-trait association studies. Map positions for all SSR were based on the cacao linkage map built with codominant markers recently published (Pugh *et al.*, 2004). DNA extraction was performed from fresh leaves according to Risterucci *et al.* (2000). Genotyping was carried out as described by Pugh *et al.* (2004).

Phenotype trait evaluation

Phenotype information on the selected clones was obtained from the International Cocoa Germplasm Database (ICGD 2000 v. 4.1 CD-ROM) which contains the published records of almost 14.000 different cacao accessions in global holding (Wadsworth and Harwood,

2000). The data used in this study were provided and described by Engels *et al.* (1981). The following traits were considered: 1) For ripe fruit: length (FRL), width (FRWI), weight (FRWE), wall thickness at second ridge (FRWT) and ridge pair separation (FRRS); 2) For beans: Length (BL), width (BW) and dry weight (BDW) and 3) For flower: sepal length (FSL), sepal width (FLSW), ligule length (FLL) and width (FLW), style length (FSL) and ovary length (FOL) and ovary width (FOW).

Data analysis

For each SSR, we were able to identify specific Criollo alleles by comparison with Criollo founder genotype (LAN17 or Zea 4). The Modern Criollo/Trinitario population is derived from crosses between a Pure Criollo genotype and different parental Forastero genotypes, specially with Lower Forastero genotypes (Motamayor *et al.*, 2003; Pugh *et al.*, in preparation). In agreement to this, we grouped alleles in two classes for each SSR: Criollo or Non Criollo

Population structure analysis: An obstacle to successful association studies in plants is the nature of population structure. The presence of subgroups with different allelic frequencies can result in spurious associations (Pritchard *et al.*, 2000). The patterns of population structure were investigated using Data set 1 in two steps: a Factorial analysis of correspondences was performed using GENETIX software (Belkhir *et al.*, 2001). Redundant individuals were identified and eliminated for next analysis; Then, we used a model-based clustering method implemented in STRUCTURE program (Pritchard *et al.*, 2001) available at <http://pritch.bsd.uchicago.edu/>. This program assigns individual genotypes to a user-defined number of clusters (K), achieving linkage equilibrium within cluster. The most likely value of K is assessed by comparing the likelihood of the data into which the sample data (X) were fitted with posterior probability $\text{Pr}(X | K)$. The model assumes Hardy-Weinberg equilibrium (HWE) within each subpopulation. We used a model with admixture and uncorrelated allele frequencies. In an admixture model, for each individual, STRUCTURE estimated the proportion of ancestry for each of the K cluster. For each run, a burning period of 250 000 iterations and obtained probability estimates using 500 000 Markov Chain Monte Carlo repetitions were employed.

Linkage disequilibrium analysis: LD is often quantified using statistics of associations between alleles of pairs of loci or measures that reflect departures of two-locus haplotypes frequencies from those expected if the loci were in linkage equilibrium. A variety of measures have been discussed (Hedrick, 1987; Lewontin, 1988; Devlin and Risch, 1995; Jorde, 2000; Weir, 1996; Hamilton and Cole, 2004; Gorelick and Laubichler, 2004). From our genotypic data, it was not possible to distinguish the coupling from repulsion double heterozygotes and, therefore, to determine haplotypic frequencies. Weir (1979) and Weir and Cockerham (1989) defined a composite measure of linkage disequilibrium (also known as $\hat{\chi}_{ij}$) for the phase-unknown situations when random mating cannot be assumed. This composite estimator measures the association of alleles from different loci on the same haplotype (intrahaplotypic LD) as well as on different haplotypes (interhaplotypic LD). The composite $\hat{\chi}_{ij}$ was estimated using the Linkdos program (Garnier-Gere and Dillmann, 1992) adapted from Black and Krafsur's (1985) program used by GENETIX software (Belkhir *et al.*, 2001). To illustrate the extent of remaining LD, we have plotted LD values against genetic distance between SSR per chromosomal region.

Hardy-Weinberg equilibrium and haplotypes estimation: At every marker locus, genotype proportion for Hardy-Weinberg equilibrium were tested using a permutation version of the exact test given by Guo and Thompson (1992) implemented in Powermarker program (<http://statgen.ncsu.edu/powermarker>). Bootstrapping was performed (100 bootstrap). Sequential Bonferroni adjustments (Rice, 1989) were used to determine statistical significance. We used also this program to calculate the haplotypes frequencies using the expectation-maximization (EM) algorithm (Excoffier and Slatkin, 1995), an iterative method to reconstruct haplotypes and find frequencies to maximize the likelihood of the genotype data. We estimated the haplotypes in those regions where closely linked loci showing significant association were in Hardy-Weinberg equilibrium. The EM algorithm is based on the assumption that genotype frequencies at each locus are in HWE.

Molecular marker-trait associations: Normality of traits was tested with the Shapiro-Wilk test and variance homogeneity with the Levene test by the UNIVARIATE and GLM Procedure of SAS respectively (SAS Institute Inc., 1998). Marker data were compared with trait score by three methods: a one-way analysis of variance (ANOVA) performed with SAS, a non-parametric Kruskal-Wallis test and a likelihood of the odds (LOD) score

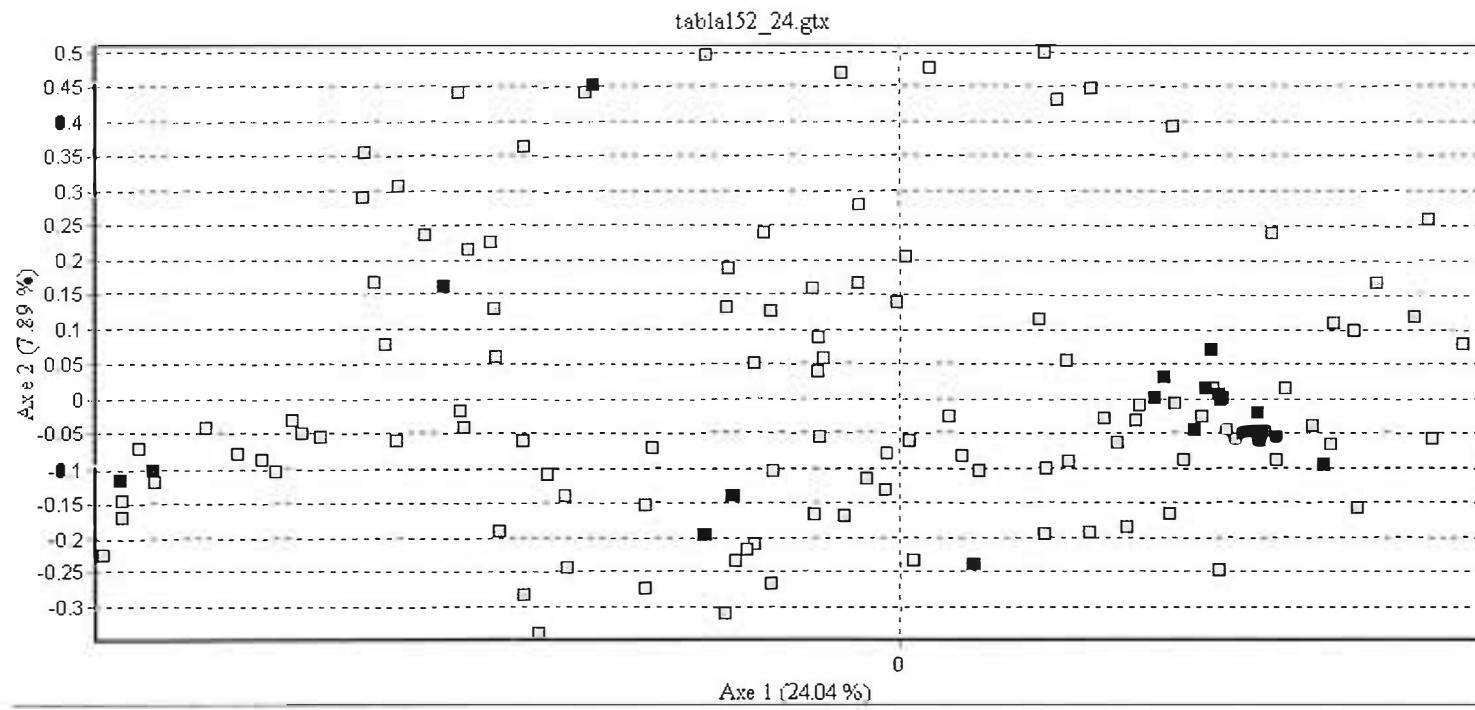


Figure 1. Factorial analysis of correspondence plot for 152 cacao accessions based on 24 unlinked SSR markers.
Accessions in bold were eliminated for association mapping studies.

both performed by MapQTL 4 (Van Ooijen and Maliepaard, 1996). For the ANOVA test, the marker-trait association was declared significant association for $p \leq 0.005$. To establish the LOD critical values, a permutation test with 1000 replications was conducted using MapQTL 4. The LOD threshold value at 5% significance at the genome level was fixed at 2.8. Associations over this level were considered to be significant. To evaluate the “confidence interval” of the marker trait associations identified, we choose a value of $\text{LOD}_{\max} - 1$ to delimitate this interval as generally used in classical QTL mapping studies. When adjacent markers to LOD_{\max} max position were lower than $\text{LOD}_{\max} - 1$, we consider their intervalle as a “maximum confidence interval” for the considered trait. Indeed, simple interval mapping did not allow to locate precisely the confidence interval delimited by $\text{LOD}_{\max} - 1$ at each side of the LOD_{\max} position.

Results

Structure of population studied

To investigate population structure we used two types of analysis. First, we performed a Factorial analysis of correspondence. According to this analysis, we eliminated 32 redundant individuals (Figure 1). This approach allowed us to select one hundred and eighteen non redundant individuals. Secondly, patterns of population structure in remaining sample were investigated using a model-based clustering method implemented in the Structure program. Analyses of subdivision using Data set 1 did not find a significant tendency towards substructuring in the sample. According to the clustering analysis, the most probable number of populations for interpreting the observed genotypes was $k=1$. Specifying K values of 2 or more led to minimal variation in likelihood values. Five runs of the program produced nearly identical results.

Polymorphisms and Hardy-Weinberg proportion

A total of 118 cacao clones were scored for their genotypes at 124 SSR. Among all SSR, only 110 enabled to identify alleles specific to Criollo and Non Criollo origin. Genetic position of the 110 discriminative markers, observed heterozygosities (H_o), expected heterozygosities under HWE for the observed population, allele frequencies and significant level of the test for departures from HWE proportions for each SSR are shown

Table 2. Description of the 110 SSR used to genotype the CATIE Criollo/Trinitario collection.

	LG	cM	f(a)	f(b)	H(o)	H(e)	HWE			f(a)	f(b)	H(o)	H(e)	HWE
mTcCIR 184	1	4.7	0.39	0.61	0.48	0.64		mTcCIR 231	4	36.1	0.08	0.92	0.15	0.08
mTcCIR 15	1	18.7	0.38	0.62	0.47	0.59	**	mTcCIR 32	4	44.6	0.38	0.62	0.47	0.57
mTcCIR 94	1	27.3	0.45	0.55	0.50	0.68	**	mTcCIR 33	4	51.2	0.33	0.67	0.44	0.41
mTcCIR 121	1	32.5	0.32	0.68	0.44	0.60	**	mTcCIR 57	4	59.8	0.33	0.67	0.44	0.51
mTcCIR 174	1	34.5	0.50	0.50	0.50	0.70	**	mTcCIR 233	4	61.6	0.31	0.69	0.43	0.44
mTcCIR 138	1	35.3	0.40	0.60	0.48	0.70	**	mTcCIR 67	4	62.3	0.51	0.49	0.50	0.50
mTcCIR 270	1	36.9	0.29	0.71	0.41	0.52	**	mTcCIR 242	4	69.5	0.31	0.69	0.43	0.47
mTcCIR 54	1	41.4	0.30	0.70	0.41	0.57	**	mTcCIR 265	5	0	0.38	0.62	0.47	0.60
mTcCIR 29	1	44.6	0.27	0.73	0.39	0.52	**	mTcCIR 267	5	17.1	0.32	0.69	0.43	0.45
mTcCIR 210	1	51.8	0.30	0.70	0.42	0.57	**	mTcCIR 109	5	72	0.34	0.66	0.45	0.59
mTcCIR 203	1	52.6	0.33	0.67	0.44	0.57	**	mTcCIR 6	6	0	0.29	0.71	0.41	0.45
mTcCIR 137	1	53.2	0.28	0.72	0.41	0.55	**	mTcCIR 136	6	2.7	0.27	0.73	0.39	0.44
mTcCIR 84	1	54.5	0.26	0.74	0.38	0.50		mTcCIR 182	6	4.1	0.44	0.56	0.49	0.62
mTcCIR 244	1	58.5	0.27	0.73	0.40	0.50		mTcCIR 290	6	10.8	0.27	0.73	0.40	0.44
mTcCIR 246	1	60.5	0.25	0.75	0.38	0.49	**	mTcCIR 71	6	30.1	0.27	0.73	0.40	0.45
mTcCIR 130	1	61	0.30	0.70	0.42	0.56	**	mTcCIR 238	6	33.1	0.26	0.74	0.39	0.40
mTcCIR 97	1	65.9	0.28	0.73	0.40	0.53	**	mTcCIR 209	6	53.9	0.26	0.74	0.39	0.41
mTcCIR 286	1	68.3	0.34	0.66	0.45	0.57	**	mTcCIR 291	6	62.8	0.27	0.73	0.39	0.49
mTcCIR 275	1	85.2	0.29	0.71	0.50	0.57	**	mTcCIR 93	7	0.9	0.35	0.65	0.45	0.55
mTcCIR 100	2	14.8	0.34	0.66	0.45	0.60		mTcCIR 141	7	3.4	0.35	0.65	0.46	0.51
mTcCIR 268	2	38.8	0.41	0.59	0.49	0.58		mTcCIR 190	7	3.8	0.34	0.67	0.45	0.52
mTcCIR 176	2	43.3	0.28	0.72	0.40	0.47	**	mTcCIR 110	7	23.4	0.33	0.67	0.44	0.52
mTcCIR 260	2	52.8	0.43	0.57	0.49	0.66	**	mTcCIR 7	7	25.9	0.26	0.74	0.39	0.49
mTcCIR 195	2	65.9	0.28	0.72	0.40	0.54	**	mTcCIR 55	7	35.5	0.40	0.61	0.48	0.67
mTcCIR 230	2	84.3	0.37	0.63	0.47	0.62	**	mTcCIR 179	7	44.6	0.04	0.96	0.07	0.04
mTcCIR 73	2	94.9	0.38	0.62	0.47	0.58		mTcCIR 186	7	49.6	0.39	0.62	0.47	0.59
mTcCIR 105	2	4	0.63	0.37	0.47	0.53		mTcCIR 163	8	10.6	0.53	0.47	0.50	0.66
mTcCIR 49	2	5.1	0.32	0.68	0.44	0.57		mTcCIR 225	8	21.1	0.33	0.67	0.44	0.56
mTcCIR 120	2	6.3	0.34	0.66	0.45	0.57		mTcCIR 211	8	22.7	0.35	0.65	0.46	0.58
mTcCIR 198	2	8.2	0.45	0.55	0.50	0.59		mTcCIR 258	8	25.9	0.38	0.62	0.47	0.67
mTcCIR 153	2	9	0.31	0.69	0.43	0.53		mTcCIR 26	8	35.6	0.29	0.71	0.41	0.51
mTcCIR 192	2	14.8	0.40	0.60	0.48	0.58		mTcCIR 218	8	37	0.20	0.80	0.32	0.35
mTcCIR 173	2	14.9	0.43	0.57	0.49	0.53		mTcCIR 189	8	46.9	0.33	0.67	0.44	0.54
mTcCIR 40	2	15.9	0.35	0.65	0.45	0.58		mTcCIR 287	9	2.6	0.37	0.63	0.47	0.51
mTcCIR 247	2	16.6	0.45	0.55	0.50	0.60		mTcCIR 266	9	10.5	0.33	0.67	0.44	0.54
mTcCIR 133	2	18.3	0.39	0.61	0.48	0.59		mTcCIR 251	9	25.4	0.26	0.74	0.39	0.46
mTcCIR 204	2	20.8	0.29	0.71	0.41	0.46		mTcCIR 30	9	23.5	0.32	0.68	0.43	0.57
mTcCIR 82	2	30.6	0.32	0.68	0.44	0.61	**	mTcCIR 24	9	31.8	0.32	0.68	0.43	0.45
mTcCIR 175	2	38.1	0.17	0.83	0.28	0.24		mTcCIR 215	9	35.9	0.29	0.71	0.41	0.50
mTcCIR 65	2	46.9	0.32	0.68	0.43	0.56	**	mTcCIR 35	9	38.3	0.25	0.75	0.38	0.49
mTcCIR 167	2	59.2	0.39	0.61	0.48	0.55		mTcCIR 157	9	40.1	0.28	0.72	0.41	0.43
mTcCIR 140	2	72.3	0.34	0.66	0.45	0.58	**	mTcCIR 142	9	47.4	0.39	0.61	0.47	0.66
mTcCIR 222	2	6.5	0.23	0.77	0.35	0.42		mTcCIR 124	9	49.1	0.30	0.70	0.42	0.52
mTcCIR 158	2	7.9	0.33	0.67	0.44	0.46		mTcCIR 63	9	49.8	0.41	0.59	0.48	0.56
mTcCIR 199	2	17.9	0.22	0.78	0.35	0.37		mTcCIR 187	9	49.9	0.36	0.64	0.46	0.57
mTcCIR 18	2	18.7	0.26	0.74	0.39	0.46		mTcCIR 178	9	50	0.34	0.66	0.45	0.56
mTcCIR 107	2	19.4	0.36	0.64	0.46	0.61	**	mTcCIR 90	9	50.4	0.31	0.69	0.43	0.55
mTcCIR 17	2	23.4	0.27	0.73	0.40	0.47		mTcCIR 114	9	50.6	0.32	0.68	0.44	0.57
mTcCIR 213	2	24.1	0.33	0.67	0.44	0.51		mTcCIR 8	9	52.5	0.36	0.64	0.46	0.55
mTcCIR 183	2	25	0.35	0.65	0.46	0.63		mTcCIR 205	9	59.6	0.23	0.77	0.35	0.37
mTcCIR 221	2	30.8	0.28	0.72	0.41	0.45		mTcCIR 58	9	60.1	0.06	0.94	0.11	0.09
mTcCIR 188	2	34.3	0.21	0.79	0.34	0.41		mTcCIR 79	9	94.1	0.32	0.68	0.44	0.54
mTcCIR 217	2	34.9	0.33	0.67	0.44	0.51		mTcCIR 229	10	11.9	0.34	0.66	0.45	0.59
mTcCIR 12	2	35.9	0.32	0.69	0.43	0.48		mTcCIR 220	10	26	0.26	0.74	0.38	0.43
								mTcCIR 91	10	51.4	0.40	0.60	0.48	0.53

For this table: Linkage group (LG); SSR position (Pugh *et al.*, 2004) (cM); Criollo and Non Criollo alleles frequency [f(a) and f(b)]; observed and expected heterozygosities [H(o) and H(e)] and ** HWE are significant departures to Hardy-Weinberg equilibrium. SSR in bold correspond to data set 1.

in Table 2. The distribution of allele frequencies varied considerably from locus to locus. Criollo allele frequencies ranged from 0.04 at mTcCIR179 on LG7 to 0.63 at mTcCIR105 on LG3. Of de 110 locus, 4 (3.6%) had a Criollo allele frequency below 20%, whereas 91 (82.7%) has a Criollo allele frequency between 20 to 40%, 14 (12.7%) between 40 to 60% and only one (0.9%) above 60%. The average observed heterozygosity was 0.492 and the expected heterozygosity under HWE was 0.410. Indeed, for most SSR loci, observed heterozygosities (H_o) were apparently higher than expected hereozygosities (H_e) under HWE.

Test for HWE indicated significant deviation for 40 of 110 loci (36.4%) after accounting for multiple tests using a Bonferroni correction ($p \leq 0.005$). SSR with departures from HWE were distributed throughout all linkage groups but their distribution was not random. The percentage of SSR with departures from HWE per group ranged from 84.2% on LG1 to 0% on LG6. LG3, LG4 and LG9 had almost 73% of SSR in HWE. Departures from HWE were examined as a condition to estimate haplotypes using the EM procedure.

Linkage disequilibrium

The extent of LD in the Modern Criollo/Trinitario population will determine the feasibility of LD mapping methods in this population and the marker density required for LD mapping to be effective. To visualize the extent of LD, we plotted LD values (Δ) against genetic distance between markers (cM) for each chromosome considered: LG1, LG3, LG4, LG6 LG9 (Figure 2). Along each of the chromosomes, LD values decreased as the distance between loci increased. The decay of LD did not seem to be variable among chromosomal regions. The maximal distance with LD seems to be about 30 cM. We used the LD values calculated between the 24 unlinked markers from Data set 1 to determine a threshold of significance. The LD values between 24 unlinked SSR were always below 0.08. Taking this value as a threshold, we consider that the LD maximum distance spans were about 30 cM.

Molecular marker-trait associations

Taking into account only the associations revealed as significant by the three analysis methods (ANOVA, K-W and LOD score), we detected a total of 69 single marker-trait

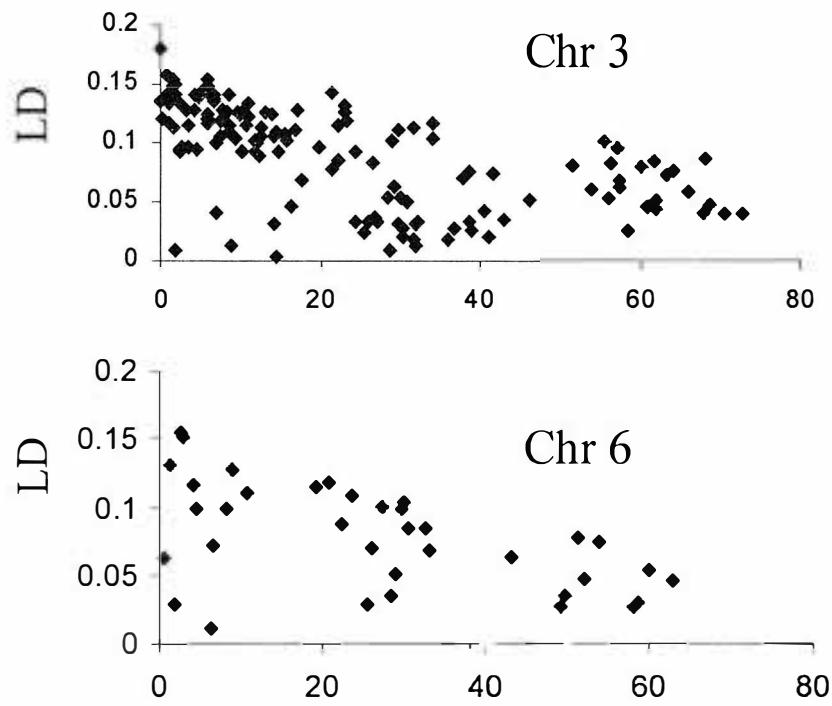
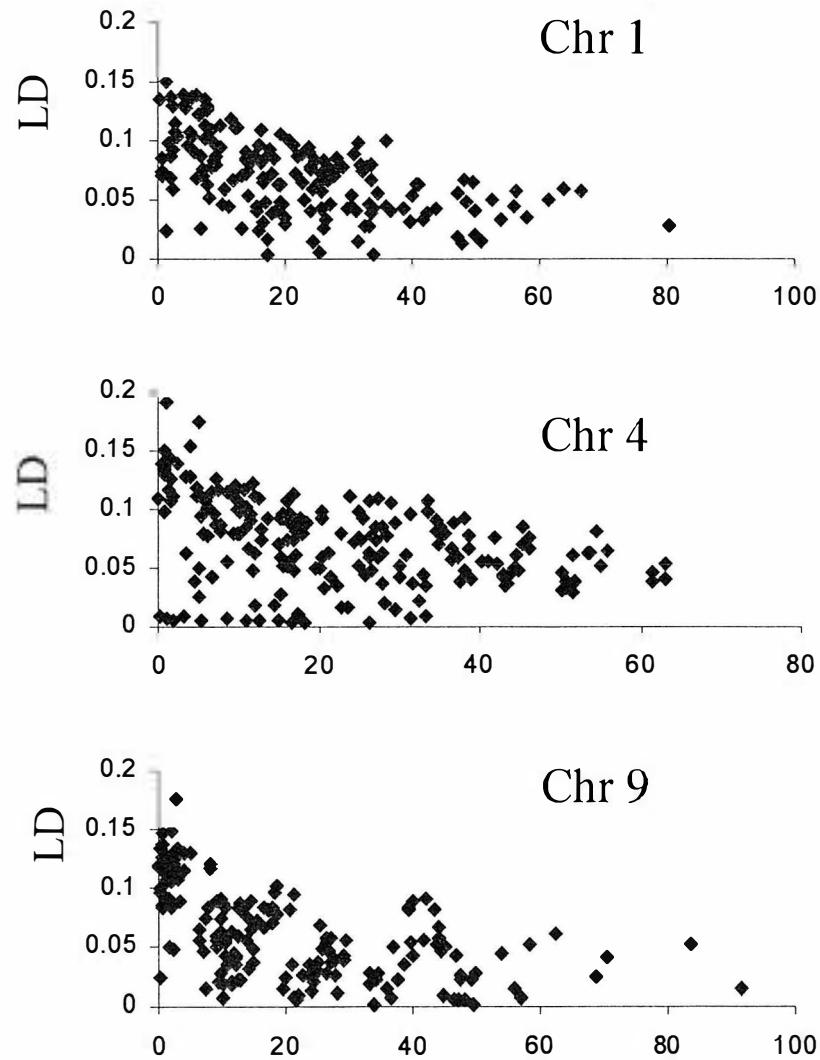


Figure 2. Linkage disequilibrium (Δ) as a function of genetic distance on 5 linkage groups.

associations corresponding to 13 genomic regions (GR) localised on 8 linkage groups. These associations mostly involved fruit (4 GR) and bean (9 GR) traits. Table 3 lists the SSR with their genome positions and traits associated. For fruit traits, these significant associations were detected on LG 3 and 4. GR related to several fruit traits were identified at (or near) mTcCIR222 on LG4. For beans traits, these significant associations were identified on LG1, LG2, LG5, LG6, LG8 and LG9. Several GR related to correlated beans traits (length, weight and width) were also identified. No molecular marker was associated with flower traits except for width ligule flower at mTcCIR182 on LG6.

Eight of the GR showing marker-trait associations involved only one SSR; 6 GR involved two closely linked SSR; 1 GR involved three closely linked SSR and 1 GR associated seven closely linked SSR. The maximal genetic distance showing molecular markers associated with a trait was 20.8 cM on LG6. On LG9, a GR contained 4 SSR and spanned 8.3 cM while an included SSR (mTcCIR35) did not show significant association with traits considered.

LOD scores for these significant associations ranged from 2.81 to 6.00. The associations with significant LOD score value were placed on a previous linkage map and are showed in Figure 3. LG4, LG6, LG8 and LG9 had more than a GR with significant SSR-trait associations. The contributions to phenotypic variations of these associations ranged from 12.7% to 31.4%. The major associations were found at or near mTcCIR174 on LG1 and mTcCIR176 on LG2 for width seeds and at or near mTcCIR198 on LG3 for fruit length respectively explaining 34.5, 33 and 30.6% of the observed variation. In spite of differences between populations, number of common markers and approach used, 8 significant associations among the 13 identified corresponded to the position of QTL already reported for the same traits in previous studies (Clement *et al.*, 2003a, 2003b; Lanaud *et al.*, 1999).

Haplotypes estimation

We used the EM algorithm to estimate the haplotypes in regions where closely linked loci showing significant association were in Hardy-Weinberg equilibrium. The HWE is an assumption of the EM algorithm and departures from HWE might lead to biased estimates of haplotypes frequencies (Excoffier and Slatkin, 1995). The haplotypes estimated and

Table 3. Associations between microsatellite markers and fruit traits indicating linkage group (LG), LG positions (cM), p value from ANOVA test (** p ≤ 0.005; *** p ≤ 0.001), significant Kruskal and Wallis test (**** p ≤ 0.005; ***** p ≤ 0.001; ***** p ≤ 0.0005), LOD score and QTL previously reported in the same region.

μsat	LG	cM	FRL				FRWE				FRWI				FRWT				QTL reported in the same region		
			p	KW	LOD	%	p	KW	LOD	%	p	KW	LOD	%	p	KW	LOD	%	Genotype	Traits	Author
mTcCIR 120	3	6.3	***	*****	4.34	20.7															
mTcCIR 198	3	8.2	***	*****	5.23	26.8															
mTcCIR 153	3	9.0	**	****	2.79	12.7															
mTcCIR 192	3	14.8	***	****	4.60	23.0															
mTcCIR 173	3	14.9	***	****	3.60	19.5															
mTcCIR 40	3	15.9	***	*****	5.35	25.9															
mTcCIR 247	3	16.6	***	*****	6.00	31.4															
mTcCIR 222	4	6.5	***	****	4.2	20.4	***	*****	5.47	25.4	***	****	3.87	18.7	***	*****	3.88	19.6	DRI	FRWE	(1)
mTcCIR 158	4	7.9	***	*****	4.77	22.8															
mTcCIR 17	4	23.4	**	****	2.94	15.9	**	*****	2.84	15.3											
mTcCIR 67	4	62.3	***	*****	4.84	26.7															

For ripe fruit: length (FRL), width (FRWI), weight (FRWE) and wall thickness at second ridge (FRWT)

(1) Clement *et al.* (2003a); (2) Clement *et al.* (2003b); (3) Lanaud *et al.* (2003)

Table 3 (cont.). Associations between microsatellite markers and beans and flower traits indicating linkage group (LG), LG positions (cM), p value from ANOVA test (** p ≤ 0.005; *** p ≤ 0.001), significant Kruskal and Wallis test (**** p ≤ 0.005; ***** p ≤ 0.001; ***** p ≤ 0.0005), LOD score, variation explained (%) and QTL previously reported in the same region.

	LG	cM	BDW				BW				BL				FLGW				QTL reported in the same region		
			p	KW	LOD	%	p	KW	LOD	%	p	KW	LOD	%	p	KW	LOD	%	Genotype	Traits	Author
mTcCIR 94	1	27.3	***	*****	3.65	17.4	**	****	3.39	16.3	***	*****	3.62	17.3					DRI, S52	BL, BW	(2)
mTcCIR 176	2	43.3	***	*****	4.86	25.0	***	*****	6.12	31.4	***	*****	3.94	21.1					DR1	BL, BW	(2)
mTcCIR 265	5	0	***	*****	3.41	17.8													IMC 78	BL	(2)
mTcCIR 267	5	17.1	***	***	3.31	21.7															
mTcCIR 182	6	4.1					***	*****	2.82	20.0					**	***	3.68	17.8	UF 676	BWE	(3)
mTcCIR 238	6	33.1	***	*****	3.61	17.2	***	*****	5.33	24.3	***	*****	4.67	21.6					S52 ; IMC 78	BL	(2)
mTcCIR 209	6	53.9	***	*****	3.17	16.1	***	*****	3.88	19.4	***	*****	4.17	20.5							
mTcCIR 225	8	21.1					**	*****	3.04	16.0											
mTcCIR 211	8	22.7					***	*****	4.67	23.9											
mTcCIR 258	8	25.9					***	*****	3.61	18.7									DRI	BW	(2)
mTcCIR 287	9	2.6					***	*****	4.65	24.6											
mTcCIR 24	9	31.8	***	*****	4.15	19.8	***	*****	4.55	21.7									UF 676	BWE	(3)
mTcCIR 215	9	35.9	***	*****	4.73	21.9	***	*****	3.71	17.7	***	*****	3.50	16.7							
mTcCIR 35	9	38.3																			
mTcCIR 157	9	40.1	***	*****	5.51	27.9	***	*****	5.02	26.9	***	*****	4.47	23.4							
mTcCIR 124	9	49.1	***	*****	4.27	22.1	***	*****	4.45	22.9	**	*****	3.00	16.2							
mTcCIR 63	9	49.8	***	*****	5.71	25.9	***	*****	4.85	22.6	***	*****	3.97	18.8							
mTcCIR 187	9	49.9	**	*****	2.74	14.7															
mTcCIR 178	9	50.0	**	*****	3.00	15.7															
mTcCIR 90	9	50.4	***	*****	4.13	20.7	***	*****	3.52	12.8											
mTcCIR 14	9	50.6	**	*****	2.81	13.1															
mTcCIR 8	9	52.2	**	*****	3.08	15.9															

For beans: dry weight (BDW), fresh weight (BFW), width (BW) and length (BL)

(1) Clement *et al.* (2003a); (2) Clement *et al.* (2003b) ; (3) Lanaud *et al.* (2003)

Table 4. Estimated Haplotypes from genomic regions where significant associations were detected and where SSR markers were in Hardy-Weinberg equilibrium.

Marker order (10.3 cM)		
Estimated Haplotypes		120-198-153-192-173-40-247
#	Haplotype	Frequency
1	b-b-b-b-b-b-b	39.08
2	a-a-a-a-a-a-a	20.70
3	b-a-b-b-b-b-b	4.70
4	b-b-b-a-a-a-a	4.26
5	b-a-b-b-b-b-a	2.50
6	a-a-a-b-a-b-a	2.16
7	b-a-b-a-b-a-b	1.90
8	b-b-b-b-a-b-b	1.80
9	a-a-b-b-a-b-a	1.60
10	a-a-b-a-a-a-a	1.47
11	a-b-b-b-b-b-b	1.45
12	b-a-b-a-b-b-b	1.41
13	b-b-a-b-b-b-b	1.34
14	a-b-a-a-a-a-a	1.31
15	b-a-b-b-a-b-a	1.25
16	b-b-b-a-a-b-a	1.23
17	a-a-a-a-a-b-a	1.04
18	b-b-b-b-b-b-a	0.98
19	b-a-a-a-a-a-a	0.97
20	a-a-a-b-a-a-a	0.94
21	a-b-b-b-b-b-a	0.88
22	b-a-a-a-a-b-a	0.84
23	b-a-a-b-b-b-b	0.70
24	b-b-b-b-a-a-a	0.61
25	b-a-b-a-a-b-b	0.59
26	b-a-b-b-b-a-a	0.55
27	a-a-a-a-a-a-b	0.54
28	b-b-b-b-b-a-b	0.54
29	a-a-b-a-a-b-a	0.53
30	a-a-b-a-b-b-b	0.52
31	b-a-b-b-b-a-b	0.52
32	b-b-b-b-a-a-b	0.46
33	a-a-b-b-b-b-a	0.31
34	a-b-a-b-a-a-a	0.27
35	a-b-a-b-a-b-a	0.05
36	b-b-b-b-b-a-a	1.22E-05
81	b-a-a-b-b-b-a	1.02E-44

Marker order (1.4 cM)		
Estimated Haplotypes		222-158
#	Haplotype	Frequency
1	b-b	63.07
2	a-a	18.47
3	a-b	4.43
4	b-a	4.10

Marker order (17.1 cM)		
Estimated Haplotypes		265-267
#	Haplotype	Frequency
1	b-b	58.50
2	a-a	27.37
3	a-b	10.20
4	b-a	4.10

Marker order (17.1 cM)		
Estimated Haplotypes		238-209
#	Haplotype	Frequency
1	b-b	69.96
2	a-a	22.10
3	a-b	4.15
4	b-a	3.79

their frequencies for 4 GR (on LG3, LG4, LG5 and LG6) are shown in Table 4. Analysis of haplotype diversity showed that in all cases and independently from marker number, the most frequent haplotypes were constituted by “Non Criollo alleles only” followed by “Criollo alleles only”.

Discussion

To our knowledge, this study is the first report studying the extent of LD on cacao genome and identifying molecular marker-trait associations in a cacao germplasm collection using an association mapping approach.

In this study we first examined the presence of population structure. Indeed, among others factors, the extent of LD is strongly dependent on population stratification. The presence of subgroups within population with an unequal distribution of alleles make more difficult to locate genes responsible for phenotypic variation because it can result in spurious associations between markers and traits. False positive associations resulting of population structure, have been found in an admixed human population (Knowler *et al.*, 1988). In plants LD studies, the importance of population structure in rice was demonstrated by an association identified between the bacterial blight resistance allele (*Xa5*) and the *aus-boro* subpopulation (Garris *et al.*, 2003). High levels of LD were also observed in *Oryza glaberrima* but primarily caused by population structure (Semon *et al.*, 2004). On the other hand, Thornsberry *et al.* (2001) have also demonstrated that the number of false positive decreased when population structure was corrected in maize. No population structure could be detected in our sample after eliminating redundant individuals. This result seems to agree with earlier studies that showed that most of Modern Criollo/Trinitario varieties have a unique Criollo ancestor and a limited number of Forastero parents (Motamayor *et al.*, 2002, 2003, Pugh *et al.*, in preparation). Differences between individuals from our sample are mainly due to the variable proportion of alleles shared with the Criollo and Forastero original parents. Genome-wide LD results using unlinked markers also showed low LD values, reflecting no structure in the population.

The feasibility of association mapping studies depends strongly on the level of LD. In a general way, LD is expected to extend over small distances. In a large, randomly mated population, in the absence of selection, mutation or migration, the LD value is only

function of the recombination probability between loci. The extent of LD determines the number and diversity of markers required for association studies. In a population with extensive LD, few markers could be necessary to detect associations; meanwhile if LD extends to small distances, a higher marker density is required. In this study, we have showed that LD spanned long genetic distances across chromosome regions considered (until 30 cM), as expected in a population that has undergone recent admixture (< 6 or 7 generations). With random mating, LD is expected to decline very rapidly at a rate of $(1 - r)^n$ over n generations, where r is the recombinant fraction between markers. Despite few generations since first cross between Trinitario parents, we observed a decline of LD with genetic markers distances up to 30 cM. The large extent of LD limits the number of markers needed for LD mapping studies. Our results suggest that LD approach could be successfully feasible in this population using the recently published cacao linkage map based on codominant markers which have 1 SSR every 3 cM. (Pugh *et al.*, 2004).

In humans, the strength of LD and the distance over which it extends have been already studied. LD is highly variable across human genome and between populations studied (Abecasis *et al.*, 2001; Stephens *et al.*, 2001; Reich *et al.*, 2001; Ardlie *et al.*, 2002; Wall and Pritchard, 2003) but might extend over 50kb. Our current knowledge of the extent of DL along chromosomal regions in plants is still limited. Some studies have shown that these distances are also very variable and dependent of the species, chosen population and the area of the genome studied. In selfing species, LD declines over a distance of 100 Kb in rice (Garris *et al.*, 2003), 500 kb in soybean (Hyten *et al.*, 2004) or between 2 and 4 cM in sorghum (Deu and Glaszmann, 2004). In outcrossers plants, LD seems to decay more rapidly than in selfers plants, as expected. In maize, LD can decay quite rapidly within a few hundred bases in landraces, but contradictory results show that it can also span several hundreds kb in other genomic regions and populations (Remington *et al.*, 2001; Tenaillon *et al.* 2001). Sugarcane exhibits LD covering 10 cM. The bottleneck at the origin of the majority of the modern sugarcane and the few generations of recombination could explain this large distance (Jannoo *et al.*, 1999). Tenaillon *et al.* (2001) considered that molecular marker such as Single Nucleotide Polymorphism (SNP) every 100-200pb are necessary to have chances to find associations in maize while for Arabidopsis this density decreases to a marker every 50 kb (Nordborg *et al.*, 2002).

Significant marker-trait associations could be identified in this study. From our data, in order to avoid type I errors, we have used stringent significance threshold for each test performed indicating the robustness of associations found. An other evidence of this robustness is that among the 13 genomic regions associated, 8 correspond to QTL already identified by Clement *et al.* (2003a, 2003b) and Lanaud *et al.* (2003). In spite of large extent of LD across genome regions considered, we have found a higher number of genomic regions with only a single marker-trait association. Associations for closely linked marker loci extended for a maximum of 20 cM. A particular situation was found in a genomic region of LG9 which contained four closely linked SSR spanning 8.3 cM. In this genomic region, a SSR (mTcCIR35) did not show significant association with the three beans traits considered (BDW, BW, BL). To explain this result, different reasons could be hypothesized such as problems in the relative order of the markers in the linkage map used, recent mutations or higher mutation rate in this area. The linkage map used was constructed using a mapping population derived from the cross between two heterozygous parents (Pugh *et al.*, 2004). Despite the high number of heterozygous markers segregating in both parents, the order of markers and their distances on the consensus map could be ambiguous for some markers. We have examined in each parent map analysed separately the relative markers' order in this region including the four SSR showing significant association, without finding differences in marker order (data not shown). Similarly, another study with barley cultivars showed that the association profiles, made by plotting p-values of correlation between markers and trait against the position of the marker, jumped up and down in a 46-48 cM area of chromosome 6 (Kraakman *et al.*, 2004). In this same region of LG9 associations involved in the same beans traits are identified in two very closed regions separated by 9 cM with only one marker (mTcCIR 142) with a lower significative association in this interval (LOD=2.17). We can wonder if these results correspond really to two different associations regions or to only one association region still to be defined with a larger or another population.

One of the interests of the association mapping approaches is the mapping resolution. If linkage disequilibrium is too large, resolution may be low, but genome scans are viable. In theory, the precision with which a QTL can be localized is directly proportional to the number of meioses sampled. In mapping populations, a few number of meioses have generally happened whereas LD mapping in populations with a higher number of meioses

since admixture may permit more precise molecular marker traits associations localisations. According to our results, in spite of long LD distances founded, the confidence interval for association mapped ranged from 0.7 cM to 24 cM. Comparing only with the common associations localised by others authors, the confidence interval ranged from 3.1 to 23.8 cM in our results whereas the confidence interval were always upper than 9 cM from QTL mapping. It ranged from 10 to 23 cM in QTL localised by Clement *et al.* (2003a, 2003b) and from 9 to 30 cM in QTL localised by Lanaud *et al.* (2003).

Some studies have also found associations between traits and markers across germplasm collections. Thornsberry *et al.* (2001) found associations between markers and flowering times in 92 maize inbred lines. In barley, Ivandic *et al.*, (2003) identified associations between markers and water-stress tolerance and powdery mildew resistance in 52 wild barley lines; Kraakman *et al.* (2004) have localised 18-20 AFLP markers that accounted for 40-58% of the yield and yield stability in 146 modern two-row spring barley cultivars representing the current commercial germplasm in Europe. Simko *et al.*, (2004a) have found a significant association between a candidate gene marker (StVe1) and resistance to *Verticillium dahliae* in 137 tetraploid potato genotypes. The usefulness of haplotypes at StVe1 locus for marker-assisted selection was then investigated. Simko *et al.*, (2004b) concluded that tagging of additional genes for resistance to *Verticillium* with molecular markers will be required for efficient marker-assisted selection.

Another aspect investigated was haplotype diversity in some genomic regions. Samples of gametes or haplotypes are necessary for accurate estimation of linkage disequilibrium. In selfing species, where the frequency of double heterozygotes with unknown phase is low, estimate of haplotype frequency is relatively easy. In outcrossing species, sampling gametes from natural populations is a difficult task. Statistical haplotypes estimations and evaluation of linkage disequilibrium may often be biased due to departures from Hardy-Weinberg expectations when these estimations are obtained from genotypic data (Weir, 1996). From our data, haplotype frequencies can not be directly obtained. However, we estimated haplotypes from regions where associations were found and where most of markers were in HWE. In all cases, the most frequent haplotypes were constituted by Non Criollo alleles followed by Criollo allele haplotypes. It represents the evidence of linkage disequilibrium. A haplotype represents a chromosomal region from an ancient founder

that was left intact through several generations. However, since we have grouped alleles different from Criollo alleles as “Non Criollo” alleles, a potential source of bias could be expected. It is known that SSR are very polymorphic markers for which each allele summarizes an evolutionary history (Kruglyak, 1999). This polymorphism based on variation in the number of repeated motifs is probably due to slippage during DNA replication or unequal crossing-over (Levinson and Gutman 1987). Multiallelic SSR could give more information about specific haplotypes associated to traits. However, linkage disequilibrium measures between multiallelic markers are difficult to interpret. To facilitate the studies that imply multiallelic markers, different strategies have been used. For example, to generate di-allelic data, Semon *et al.* (2004) identified the most frequent alleles and combined all the remaining alleles into a second “allele” class. Tenesa *et al.* (2003) have proposed grouping SSR alleles for biological more than for statistical reasons. In our case, we were interested by associations with Criollo traits. Variations in fruits and bean traits as well as chocolate quality products are related to the genetic origin of genotypes. The Criollo group generally has large, round and white or pink beans and also produces a high quality chocolate. For these reasons, to identify genomic regions associated to phenotypic variations corresponding to Criollo group we choose to classify alleles on two classes: Criollo/Non Criollo alleles.

With one-way analysis, associations found were based on phenotypic means and variances within each of the genotypic classes at the marker locus. One of challenge in association mapping is evaluating phenotypes. The phenotype data used in this study was obtained from the ICGD data where are conserved data obtained by Engels *et al.* (1981) 25 years ago. In spite of the lack of replicated genetic test to minimize environmental variation, this data represent a valuable source of data with average values estimated from several evaluations.

We have shown that it is possible to identify molecular marker-trait associations directly on already available Modern Criollo/Trinitario cacao varieties, without developing a new mapping population. Exploiting LD remains a promising strategy to exploit available data from germplasm collections evaluations to identify loci associated to phenotypic variation. These regions could be examined in further studies and be used for marker-assisted selection. This work is a first step toward that goal.

Acknowledgments

We thank the CATIE for providing us with the plant material studied.

References

- Abecasis G.R., Noguchi E., Heinzmann A., Traherne J.A., Bhattacharyya S., Leaves N.I., Anderson G.G., Zhang Y., Lench N.J., Carey A., Cardon L.R., Moffatt M.F., Cookson W.O. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68:191-197.
- Ardlie K.G., Kruglyak L., Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 2002, 3:299-309.
- Belkhir K., Borsig P., Chikhi L., Raufaste N., Bonhomme F. 2001 GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions: CNRS UMR. 5000, Université de Montpellier II, Montpellier, France.
- Black W. C. and Krafur E. S. 1985. A Fortran program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theoretical and Applied Genetics* 70: 491-491
- Cheesman E.E. 1944. Notes on the nomenclature, classification possible and relationships of cocoa populations. *Tropical Agriculture* 21:144-159.
- Clement D., Risterucci A.M., Motamayor J.C., N'Goran J.A.K., Lanaud C. 2003a. Mapping QTL for yield components, vigor, and resistance to Phytophthora palmivora in *Theobroma cacao* L. *Genome* 46:204-12.
- Clement D., Risterucci A.M., Motamayor J.C., N'Goran J.A.K., Lanaud C. 2003b. Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46:103-111.
- Crouzillat D., Lerceteau E., Petiard V., Morera J., Rodríguez H., Walker D., Phillips W., Ronning C., Schnell R., Osei J., Fritz P. 1996. *Theobroma cacao* L.: a genetic linkage map and quantitative trait loci analysis. *Theor Appl Genet* 93:205-214.
- Crouzillat D., Menard B., Mora A., Phillips W., Petiard V. 2000a. Quantitative trait analysis in *Theobroma cacao* L. using molecular markers. *Euphytica* 114:13-23.
- Crouzillat D., Phillips W., Fritz P.J., Petiard V. 2000b. Quantitative trait loci analysis in *Theobroma cacao* L. using molecular markers. Inheritance of polygenic resistance of Phytophthora palmivora in two related cacao populations. *Euphytica* 114:25-36.
- Deu M. and Glaszmann J.C. 2004. Linkage disequilibrium in sorghum. In: Plant & Animal Genomes. XII Conference, 10-14 January, Town & Country Convention Center, San Diego, W10.
- Devlin B. and Risch A. 1995. Comparison of Linkage Disequilibrium Measures for Fine-scale Mapping. *Genomics* 29 : 311-322.
- Engels J.M.M. 1981. Genetic resources of cacao. A catalogue of the CATIE collection. Technical Bulletin N° 7. Tropical Agriculture and Research Training Center, Turrialba, Costa Rica. pp 191.
- Excoffier L. and Slatkin M. 1995. Maximum- likelihood estimation of molecular haplotypes frequencies in a diploid population. *Mol. Biol. Evol.* 19:921-927.

CHAPITRE IV. Études des associations

- Flament M.H., Kebe I., Clement D., Pieretti I., Risterucci A.M., N'Goran J.A.K., Cilas C., Despréaux D., Lanauud C. 2001. Genetic mapping of resistance factors to Phytophtora palmivora in cocoa. *Genome* 44:79-85
- Flint-Garcia, S., Thornsberry J.M., Buckelr E.S. 2003. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54:354-374.
- Garnier-Gere, P., and Dillmann C. 1992. LinkDos. *J. Heredity* 56:409-415
- Garris A.J., Mc Couch S.R., Kresovich S. 2003. Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the Xa5 locus of rice *Oryza sativa* L. *Genetics* 165:759-769.
- Gorelick R., and Laubichler M.D. 2004. Decomposing multilocus linkage disequilibrium. *Genetics* 166:1581-3.
- Guo, S.W. and Thompson, E.A. 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics* 48: 361-372.
- Hamilton, D.C. and Cole, D.E.C. 2004. Standardizing a Composite Measure of Linkage Disequilibrium. *Annals of Human Genetics*, 68:234-239.
- Hedrick, P.W. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117:331-341.
- Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 226-231.
- Hyten D.L., Song Q., Cregan P.B. 2004. Linkage disequilibrium in four soybean populations. In: *Plant & Animal Genomes. XII Conference*, 10-14 January, Town & Country Convention Center, San Diego, CA p534.
- IBGRI. 1981. Report of the IBPGR working group on genetic resources of cocoa. Rome, Italie, IBPGRI, ACP: IBPGRI /80/56, 28p.
- Ivandic V., Thomas W.T.B., Nevo E., Zhang Z., Forster B.P. 2003. Associations of simple sequence repeats with quantitative trait variation including biotic and abiotic stress tolerance in *Hordeum spontaneum*. *Plant Breeding* 122: 300-3004.
- Jannoo N., Grivet L., Dookun A., D'Hont A., Glaszmann J.C., 1999. Linkage disequilibrium among modern sugarcane cultivars. *Theoretical and Applied Genetic* 99 : 1053-1060.
- Jorde L.B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10 : 1435-1444. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22: 139-144.
- Knowler W.C., Williams R.C., Pettitt D.J., Steinberg A.G. 1988. Gm^{3-5,13,14} and type 2 diabetes mellitus: an association with genetic admixture. *Am. J. Hum. Gen.* 43:520-526.
- Kraakman, A.T.W., Rients E. N., van den Berg, P.M.M.M., Stam P., Van Eeuwijk F.A. 2004. Linkage disequilibrium mapping of yield and yield stability in Modern Spring Barley Cultivars. *Genetics* 168:436-444.
- Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22: 139-144.
- Lanauud C., Risterucci A.M., N'Goran J.A.K., Clement D., Flament M.H., Laurent V., Falque M. 1995. A genetic linkage map of *Theobroma cacao* L. *Theor Appl Genet* 9:987-993.
- Lanauud C., Kébé I., Risterucci A.M., Clement D., N'Goran J.A.K., Grivet L., Tahí M., Cilas C., Pieretti I., Eskes A.B., Despréaux D. 1999. Mapping quantitative loci (QTL) for resistance to Phytophtora palmivora in *T. cacao* L. 12th International Cocoa Research Conference, November 17-23, Salvador, Bahia, Brazil pp. 99-105.

CHAPITRE IV. Études des associations

- Lanaud C., Boult E., Clapperton J., Cros E., Chapelin M., Risterucci A.M., Allaway D., Gilmour M., Cattaruzza A., Fouet O. N'Goran J.A.K., Petithuguenin P. 2003. Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma Cacao* L. In: 14ème Conférence internationale sur la recherche cacaoyère. Accra, Ghana, 13-17 octobre 2003.
- Levinson G., Gutman G.A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203-221
- Lewontin R.C. and Kojima K. 1960. The evolutionary dynamics of complex polymorphism. *Evolution* 14:458-472.
- Lewontin R.C. 1988. On measures of gametic disequilibrium. *Genetics* 120:849-852.
- Motamayor, J.C., Risterucci A.M., Lopez P.A., Ortiz C.F., Moreno A., Lanaud C. 2002. Cacao domestication I. the origin of the cacao cultivated by the Mayas. *Heredity* 89:380-386.
- Motamayor, J.C., Risterucci A.M., Heath M., Lanaud C. 2003. Cacao domestication II. Progenitor germplasm of the Trinitario cacao cultivar. *Heredity* 91:322-330.
- Nei, M. and Li, W.H. 1973. Linkage disequilibrium in subdivided populations. *Genetics* 75: 213-219.
- Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 30, 190–93.
- Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
- Pritchard J. K., and Przeworski, M. 2001. Linkage disequilibrium in Humans: models and data. *American Journal Human Genetics*, 69: 1-14.
- Pugh T., Fouet O., Risterucci A.M. , Brottier P., Abouladze M., Deletrez C., Courtois B., Clement D., Larmande P., N'Goran J.A.K., Lanaud C. 2004. A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* 108:1151-1161.
- Pugh T., Phillips W., Astorga C., Courtois B., Noyer J.L., Risterucci A.M., Fouet O., Lanaud C. Genetic diversity of Modern Criollo/Trinitario cacao clones (*Theobroma cacao* L.) from CATIE germplasm collection assessed by microsatellite markers (in preparation)
- Rafalski, A. and Morgante, M. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* 20:103-11.
- Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R., Lander E.S. 2001. Linkage disequilibrium in the human genome. *Nature*, 411:199-204.
- Remington D.L., Thornsberry J.M., Matsuoka Y., Wilson L.M., Whitt S.R., Doebley J., Kresovich S., Goodman M.M., Buckler E.S. 4th. 2001. Structure of linkage disequilibrium and phenotypic associations in maize genome. *Proceedings of the National Academy of Science* 98: 11479-11484.
- Rice W.R. 1989. Analysing tables of statistical tests. *Evolution*, 43, 223-225.
- Risterucci A.M., Grivet L., N'Goran J.A.K., Pieretti I., Flament M.H., Lanaud C. 2000. A high density linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.* 101:948-955.
- Risterucci A.M., Paulin D., N'Goran J.A.K., Lanaud C. 2003. Identification of QTL related to cocoa resistances to three species of *Phytophtora* *Theor. Appl. Genet.* 108:168-74.
- SAS. 1998. SAS User's Guide (Version 7 ed). SAS Inst Inc, Cary NC.

- Simon M., Nielsen R., Jones M., McCouch S. 2004. The population structure of African cultivated rice (*Oryza Glaberrima* (Steud.): evidence for elevated levels of LD caused by admixture with *O. sativa* and ecological adaptation. Published Articles Ahead of Print, published on November 15, 2004 as 10.1534/genetics.104.033175
- Simko I., Costanzo S., Haynes K.G., Christ B.J., Jones R.W. 2004a. Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. *Theor. Appl. Genet.* 108:217-24.
- Simko I., Haynes KG, Ewing EE, Costanzo S, Christ BJ, Jones RW. 2004b. Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Mol. Genet. Genomics.* 271:522-31.
- Stephens J.C., Schneider J.A., Tanguay D.A., Choi J., Acharya T., Stanley S.E., Jiang R., Messer C.J., Chew A., Han J.H., Duan J., Carr J.L., Lee M.S., Koshy B., Kumar A.M., Zhang G., Newell W.R., Windemuth A., Xu C., Kalbfleisch T.S., Shaner S.L., Arnold K., Schulz V., Drysdale C.M., Nandabalan K., Judson R.S., Ruano G., Vovis G.F. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489-493.
- Tenaillon M.I., M. C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doeley et al., 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize. *Proceedings of the national academy of science*.98: 9161-9166.
- Tenesa A., Knott S.A., Carothers A.D., Visscher P.M. 2003. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. *Ann Hum Genet.* 67:557-66.
- Thornberry J.M., Goodmen M.M., Doebley J., Kresowich S., Nielsen D., Buckler E.S. IV. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genet.* 28:286-289.
- Van Ooijen, J.W., and Maliepaard C. 1996. MapQTL (tm) Version 3.0: Software for the Calculation of QTL Positions on Genetic Maps. DLO-Centre for Plant Breeding and Reproduction Research, Wageningen, The Netherlands.
- Wadsworth, R.M. and Hardwood, T. 2000. International Cocoa Germplasm Database ICDG version 4.1. London International Financial Futures and Options Exchange and the University of Reading, UK.
- Wall J.D. and Pritchard J.K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet*, 4:587-597.
- Weir, B.S. and Cockerham C.C. 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 42:105-111
- Weir, B.S. and Cockerham C. C. 1989. Complete characterization of disequilibrium at two loci. In: *Mathematical Evolutionary Theory*, (Ed.) M.W. Feldman, Princeton Univ. Press. pp 86-110.
- Weir B.S. 1996. *Genetic data analysis II*. Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts, 445 pp.
- Whitlock, B., Bayer C., Baum D. 2001. Phylogenetic relationships and floral evolution of the Byttherioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene *ndhF*. *Sys. Botany* 26:420-437.

Conclusions et Perspectives

Une étude d'associations a été réalisée à partir d'une collection de variétés de cacaoyer de type Criollo/Trinitario présente au CATIE (Costa Rica). Cette collection a été caractérisée il y a 25 ans par Engels *et al.* (1981) et les données de caractérisation sont maintenant accessibles via l'ICGD (International Cocoa Gene Bank). Après élimination de 32 individus redondants, l'absence de structuration de cette collection, préalable indispensable à une étude d'association, a tout d'abord été vérifiée à l'aide de 24 SSR répartis dans tout le génome. Un déséquilibre de liaison global faible a confirmé la non structuration de cette collection.

Les 118 individus restants ont été génotypés à l'aide de 124 SSR répartis dans tout le génome mais présents avec une densité plus élevée dans 5 régions du génome où des QTL portant sur les caractères étudiés ont déjà été identifiés par l'étude de descendances contrôlées. Grâce au polymorphisme élevé des marqueurs SSR, des allèles spécifiques des Criollo ont pu être identifiés pour 110 loci. L'étendue du DL évaluée par ces marqueurs est élevé et peut atteindre 30 cM. La densité de marqueurs utilisée (en moyenne un SSR tous les 14 cM) semble donc adaptée pour les études d'associations faites dans cette collection.

Des associations hautement significatives entre marqueurs et caractères ont été identifiées dans 13 régions chromosomiques pour des caractères de fruit, de fève et de fleur. Huit de ces régions correspondent à des QTL déjà identifiés par Clément *et al.* (2003) et Lanaud *et al.* (2003) dans des descendances contrôlées.

Ces résultats nous démontrent la faisabilité et l'intérêt de cette approche pour aider à connaître les bases génétiques et moléculaires des caractères d'intérêt chez le cacaoyer. Une sélection des données morphoagronomiques à utiliser dans ces études devra toutes fois être faite afin de garantir la robustesse des mesures de caractérisation et donc celle des associations identifiées.

Le peu de descendances contrôlées à fort effectif est souvent un frein à l'étude de QTL chez le cacaoyer. Ces approches ouvrent de nouvelles perspectives pour l'étude du génome du cacaoyer et permettront de mieux valoriser les données de caractérisation de collection accumulées depuis un grand nombre d'années.

CHAPITRE V : DISCUSSION GÉNÉRALE ET PERSPECTIVES

CHAPITRE V

Discussion générale et perspectives

Produire de nouvelles variétés productives, résistantes aux maladies et ayant conservé les caractères de qualité des variétés telles que le Criollo et le Nacional, est devenu un enjeu stratégique pour certains pays comme le Venezuela ou l'Equateur.

L'amélioration de ces variétés passe par une meilleure connaissance des bases génétiques des caractères à améliorer. Le développement des marqueurs moléculaires a ouvert de nouvelles perspectives pour localiser les régions du génome contrôlant les caractères morpho-agronomiques ou les caractères de qualité des variétés et en accélérer la sélection. La plupart des caractères d'intérêt agronomique sont sous le contrôle génétique de plusieurs locus ou QTL (Quantitative Trait Locus). La localisation des QTL a été traditionnellement faite à l'aide de populations issues de croisements contrôlées. Cependant, chez le cacaoyer, la disponibilité de grandes populations en production, permettant d'observer avec précision tous les caractères souhaités, est réduite et longue à mettre en place.

L'objectif principal de ce travail a été de tester de nouvelles approches permettant d'étudier les associations entre marqueurs moléculaires et caractères d'intérêt en exploitant des populations cultivées ou issues de plusieurs générations de sélection. Ces approches se basent sur la mise en évidence du déséquilibre de liaison (DL) qui a pu être maintenu au cours de la domestication ou de la sélection, et qui reflètent les fragments chromosomiques « ancestraux » encore conservés.

Notre travail a donc consisté tout d'abord à produire des outils efficaces pour évaluer ce DL, les SSR (single sequence repeat), puis à étudier la structure génétique et les associations marqueurs/caractères dans une collection de Criollo modernes/Trinitario établie au CATIE (Costa Rica).

Développement d'une carte génétique saturée en marqueurs SSR

Une nouvelle carte génétique a été établie en intégrant 201 nouveaux marqueurs SSR à la carte de référence préexistante. La nouvelle carte dont la taille est de 783 cM contient 465 marqueurs codominants (268 SSR, 176 RFLP, 5 isoenzymes et 16 Rgenes-RFLP) repartis dans les dix groupes de liaison et avec une distance moyenne entre les marqueurs de 1.7 cM. et de 3cM. entre les microsatellites. Les 268 SSR sont bien répartis sur l'ensemble du génome et couvrent en effet à eux seuls 94,8% de la carte totale.

L'intérêt des marqueurs SSR est leur degré de polymorphisme élevé. Grâce à cela, on a pu combler ou saturer certaines régions du génome peu marquées au préalable. Ces marqueurs ont été diffusés à l'ensemble de la communauté internationale qui les utilisent maintenant largement. On dispose ainsi désormais d'un outil efficace pour identifier et comparer la localisation de QTLs d'une population à l'autre. Les SSR sont des marqueurs co-dominants, aisément transférables dans de petits laboratoires situés en régions tropicales, et constituent de ce fait un bon outil pour faire de la sélection assistée par marqueurs.

De par leur grand nombre d'allèles possibles (de 12 à 25 par locus selon D. Zhang, pers. Com.), et avec certains d'entre eux spécifiques de populations particulières, ces SSR permettront aussi une analyse fine de la diversité tout au long du génome et une identification plus aisée des génotypes fondateurs à l'origine de certaines populations.

Dans notre étude, 110 SSR ont apporté des allèles spécifiques des Criollo qui permettent ainsi de bien connaître la structure génétique des variétés et l'origine des allèles impliqués dans les associations mises en évidence.

Etude d'une collection de Criollo modernes/Trinitario

Du fait de son histoire, les Criollo/Trinitario constituent un bon modèle d'étude pour les tests d'associations. En effet, un petit nombre de générations de recombinaison sépare ce groupe hybride des premières hybridations entre ancêtres parentaux Criollo et Forastero qui ont eu lieu il y a 250 ans. Notre choix s'est donc porté sur l'étude d'une collection présente au CATIE (Costa Rica) et qui rassemble des variétés de Criollo/Trinitario sélectionnées dans les différents pays d'Amérique latine. La structure de la diversité

génétique de 247 de ces variétés a tout d'abord été étudiée à l'aide de 34 SSR répartis sur tous les chromosomes du cacaoyer. Ces analyses ont montré une faible structuration de la diversité génétique au sein de ce groupe. 175 allèles ont pu être observés dans cette collection. Pour chaque locus, deux allèles étaient toujours plus fréquemment représentés que les autres. Ces allèles correspondent aux allèles des 2 ancêtres parentaux principalement à l'origine de ce groupe : un individu Criollo ancien et un individu Forastero bas amazonien (Motamayor et al., 2003). Des allèles rares, de fréquence inférieure à 5% ont également été observés et correspondent à des hybridations entre Criollo et Forastero d'origines différentes.

Par une analyse globale de la diversité faite à l'aide d'une analyse factorielle faite sur tableau de distances, nous avons observé l'existence de quatre sous-groupes qui présentent toutefois une variation continue entre eux : Deux groupes d'accessions comprenant principalement des allèles Criollo ou des allèles Forastero bas amazonien, tous deux avec une faible diversité génétique et qui correspondent aux ancêtres parentaux; un grand groupe d'accessions hybrides, intermédiaires entre les deux premiers, et un quatrième groupe d'accessions avec des allèles Forastero différents du type bas amazonien, mais qui ont été classifiés comme des Trinitario par certains de leurs caractères morphologiques typiques des Criollo.

Des informations utiles pour la gestion de cette collection ont donc pu être apportées au cours de ce travail. En effet, un grand nombre de clones redondants ont pu être identifiés dans cette collection au cours de nos analyses et on a pu proposer un sous-ensemble de 37 accessions qui représente toute la diversité allélique existante dans les variétés étudiées.

Ces analyses mettent en avant la base génétique très étroite de ce groupe génétique qui a été largement utilisé par tous les programmes d'amélioration génétique du cacaoyer. Une moyenne de 5,2 allèles par locus a en effet été observée dans ce groupe par rapport à 18,5 alleles par locus observé en moyenne sur l'ensemble de la collection du CATIE (D. Zhang, pers. Com.). De très nombreuses autres combinaisons hybrides associant Criollo et autres types de Forastero seraient ainsi possibles.

La base génétique très étroite de ce groupe hybride crée toutefois une situation favorable pour des études d'associations.

Etude des associations marqueurs/caractères

La collection du CATIE a été évaluée pour un grand nombre de caractères morphoagronomiques (Engels, 1981) et les données sont disponibles par le biais de la base de données internationale ICGD. Les données de 150 accessions de type Criollo/Trinitario ont pu être récupérées sur cette base de données.

Chez le cacaoyer, des dispositifs permettant un contrôle précis des effets environnementaux sont difficiles et longs à mettre en place. Engels (1980) a proposé une méthode et une liste de descripteurs qui sont toujours utilisées afin de palier au mieux à ce manque de dispositifs. Des valeurs moyennes obtenues après un grand nombre d'observations défini en fonction de la variance intra échantillon, ont été calculées pour chaque caractère. Ce sont ces valeurs qui ont été utilisées dans notre étude.

On sait que la structure de la population a une forte influence sur le DL. Le DL observé dans une population structurée est le reflet du déséquilibre global (sur l'ensemble du génome) et du déséquilibre local (effet d'un liaison physique). Si la structure de la population n'est pas contrôlée, l'obtention de faux positifs est très probable. Cet effet a été déjà observé chez les humains (Knowler *et al.*, 1988). Chez les plantes, Thornskey *et al.* (2001), et Garris *et al.* (2004) ont alerté sur l'effet de la structure de la population dans les études d'association. L'étude de la diversité nous avait déjà indiqué le manque de structure claire de la population Criollo Moderne/Trinitario de la collection CATIE. Par contre, un grand nombre d'individus redondants avait été détecté. Une analyse factorielle de correspondance a permis de « visualiser » les individus redondants et de les éliminer de nos analyses ultérieures. La non structuration de l'échantillon obtenu, composé de 118 individus, a ensuite été testé avec l'approche bayésienne développé par Pritchard et Prezworski (2001) dans le logiciel STRUCTURE.

Un autre aspect intéressant était le choix de la mesure du déséquilibre de liaison. Il existe de nombreux indices pour mesurer le déséquilibre de liaison entre paires de locus. La plus grande partie de ces indices requièrent la connaissance des haplotypes à deux locus. Des méthodes statistiques d'estimation des haplotypes ont été proposées (Hill, 1974 ; Excoffier et Slatkin, 1995 ; Clark, 1990 ; Gibbs, 2001 ; Stephens *et al.*, 2001). Toutes ces méthodes supposent que la population est en équilibre de Hardy-Weinberg, ce qui n'est pas toujours le cas. Le taux élevé d'hétérozygotie des clones de cacaoyer utilisés ne permettait pas de connaître la phase des doubles hétérozygotes (« coupling ou

repulsion ») et donc, d'estimer de façon appropriée les haplotypes. Une mesure du DL appelée déséquilibre composite, proposée par Weir (1979) et Weir et Cockerham (1989), a été utilisée dans notre étude. Cette mesure, qui est la somme des déséquilibres intra- et inter gamétiques offre une alternative pour mesurer le DL chez les espèces qui, comme le cacaoyer, sont pérennes, hétérozygotes et où il est difficile d'estimer les haplotypes. Des logiciels comme GDA (Lewis et Zaikin, 2001) ou Genetix (Belkhir et al., 1996) permettent d'estimer cette mesure.

Les analyses de DL entre marqueurs sur l'ensemble du génome ont été faites en deux temps :

- Un premier lot de 24 marqueurs non liés et répartis sur tous les chromosomes de cacaoyer ont permis de mesurer un déséquilibre global faible sur l'ensemble du génome, déséquilibre qui a confirmé la non structuration de cette population et permis de fixer un seuil permettant d'identifier un déséquilibre local.
- Un deuxième lot de 110 marqueurs (avec en moyenne 1 SSR tous les 14 cM) permettant d'identifier spécifiquement les allèles d'origine Criollo a été utilisé pour mesurer le DL sur l'ensemble du génome, et en particulier sur des groupes de liaison où des QTL avaient déjà été identifiés dans des études précédentes faites sur des descendances contrôlées. La structure particulière de cette population, où un génotype de Criollo ancien a été hybridé avec différents parents Forastero, parmi lesquels un génotype majoritaire bas amazonien, nous a amené à ne considérer que 2 classes d'allèles : Criollo ou non Criollo. De ce fait, les marqueurs SSR multialléliques ont été codés comme des marqueurs bialléliques et ont facilité les estimations du DL en le ciblant préférentiellement sur la conservation des fragments de génome de type Criollo.

Les résultats ont fait apparaître un DL élevé s'étendant sur des distances pouvant atteindre parfois 30 cM. Ces résultats nous indiquent que la densité de marqueurs utilisée pour nos études d'association (1SSR/14 cM) est suffisante.

Treize régions chromosomiques portant des associations marqueurs/caractères très significatives ont été identifiées pour des caractères de fruit, de fève et de fleur. Parmi ces régions, neuf correspondent à des QTL déjà identifiés par Clément et al. (2003) et Lanaud et al. (2003). Ces associations ont été détectées avec des intervalles de confiance variables

selon les régions du génome et les caractères, et compris entre 0,7 cM et 24 cM. Si l'on compare nos résultats avec les 8 QTL communs identifiés par Clément et al. (2003) et Lanaud et al. (2003) par l'étude de descendances contrôlées d'effectif semblable, les intervalles de confiance apparaissent semblables avec toutefois 2 cas pour lesquels cet intervalle est compris entre 3 et 5 cM dans nos études, alors qu'il est toujours supérieur à 9 cM dans les analyses de Clement et al. (2003a) ou Lanaud et al. (2003).

Jusqu'aux années 50, la culture du cacaoyer s'est développée essentiellement à partir de variétés traditionnelles. Les premières étapes dans l'amélioration génétique du cacaoyer furent la sélection de variétés locales principalement résistantes aux principales maladies. Cette sélection clonale s'est faite à partir des Forastero bas amazoniens et des Trinitario en Afrique et des Criollo/Trinitario en Amérique. Les programmes de sélection d'hybrides débutèrent dans les principaux pays producteurs dans les années 50 et 60. Ces programmes utilisèrent principalement des cacaoyers originaires de Haute Amazonie (Pérou) et collectés par Pound (1938, 1943). Ces arbres, plus vigoureux, précoce et résistants aux maladies furent croisés avec les clones sélectionnés localement de type Forastero bas amazonien ou Trinitario. La plus grande partie des variétés sélectionnées et utilisées de nos jours proviennent de ces programmes d'hybridation. Des caractères associés à la morphologie des cabosses et des fèves, au rendement et à la résistance aux principales maladies ont été sélectionnés lors de ces programmes. Afin de rechercher des marqueurs de sélection précoce et de pouvoir faire de la sélection assistée par marqueurs, le déterminisme génétique de ces caractères a été parfois exploré et certains QTL ont été localisés en utilisant des marqueurs moléculaires et des descendances en ségrégation. Cependant, le peu de descendances contrôlées à effectifs importants est souvent un frein à l'étude de QTL chez le cacaoyer. Les résultats que nous avons obtenus par les études d'associations faites sur cette collection nous montrent la faisabilité et l'intérêt de cette approche pour aider à connaître les bases génétiques des caractères d'intérêt chez le cacaoyer. Ces résultats, ainsi que les outils (SSR) développés au cours de cette thèse nous ouvrent de nouvelles perspectives pour l'analyse du génome du cacaoyer et son application pour développer des programmes de sélection assistée par marqueurs.

De nombreuses collections de ressources génétiques ont été caractérisées et les données assez peu exploitées. L'approche développée dans notre étude sera une valorisation supplémentaire apportée à ces collections et à leur caractérisation.

RÉFÉRENCES BIBLIOGRAPHIQUES

Références bibliographiques

- Adam-Blondon A.F., Roux C., Claux D., Butterlin G., Merdinoglu D., This P. 2004. Mapping 245 SSR markers on the *Vitis vinifera* genome: a tool for grape genetics. *Theor. Appl. Genet.* 109(5):1017-27.
- Abecasis G.R., Noguchi E., Heinzmann A., Traherne J.A., Bhattacharyya S., Leaves N.I., Anderson G.G., Zhang Y., Lench N.J., Carey A., Cardon L.R., Moffatt M.F., Cookson W.O. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68:191-197.
- Alvim P.T. 1965. Ecophysiology of cacao tree. International Cocoa Research Conference, Abidjan, Côte d'Ivoire, Novembre 15-20. pp. 23-35.
- Aranzana M., Garcia-Mas J., Carbó J., Arús P. 2002. Development and variability of microsatellite markers in peach. *Plant Breed.* 121:87-92.
- Aranzana M., Pineda A., Cosson P., Dirlewanger E., Ascasibar J., Cipriani G., Ryder C., Testolin R., Abbott A., King G., Iezzoni A., Arús P. 2003. A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. *Theor. Appl. Genet.* 106:819-825.
- Ardlie K.G., Kruglyak L., Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 2002, 3:299-309.
- Artiguenave F., Wincker P., Brottier P., Duprat S., Jovelin F., Scarpelli C., Verdier J., Vico V., Weissenbach J., Saurin W. 2000. Genomic exploration of the hemiascomycetous yeast: 2. Data generation and processing. *FEBS Lett.* 487:6-13.
- Bartley B.G. 1979. Global concepts for genetic resources and breeding in cacao. In: Seventh International Cocoa Research Conference, London, Cocoa Producers' Alliance, Douala, Cameroon, pp. 519-525.
- Barreneche T., Bodenes C., Lexer C., Trontin J-F., Fluch S. 1998. A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5S rDNA markers. *Theor. Appl. Genet.* 97:1090-1103.
- Beckman JS, Soller M. 1990. Toward a unified approach to the genetic mapping of eukaryotes based on sequence-tagged microsatellite sites. *Biotechnology* 8:930-932.
- Bekele F. and Bekele I. 1996. A sampling of the phenetic diversity of cacao in the international cocoa gene bank of Trinidad. *Crop Science* 36:57-64.
- Belkhir K., Borsa P., Chikhi L., Raufaste N., Bonhomme F. 2001 GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions: CNRS UMR. 5000, Université de Montpellier II, Montpellier, France.
- Bergman, J.F. 1969. The distribution of cacao cultivation in pre-Columbian America. *Annals of the Association of American Geographers*, 59:85-96.
- Bhattaramakki D, Dong J, Chhabra AK, Hart GE. 2000. An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. *Genome* 43:988-1002.

- Billote N., Lagoda P.J.L., Risterucci A.M., Baurens F.C. 1999 Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits* 54:277-288.
- Billote N., Risterucci A.M., Barcelos E., Noyer J.L., Amblard P., Baurens F.C. 2001. Development, characterisation and across-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. *Genome* 44:413-425.
- Black W. C. and Krafur E. S. 1985. A Fortran program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor. Appl. Genet.* 70: 491-491.
- Boivin K., Deu M., Rami J.F., Trouche G., Hamon P. 1999. Towards a saturated sorghum map using RFLP and AFLP markers. *Theor. Appl. Genet.* 98:320-328
- Boyer, J. 1970. Influence des régimes hydrique, radiatif et thermique du climat sur l'activité végétative et la floraison de cacaoyers cultivés au Cameroun. *Café, cacao, Thé* 14 :189-200.
- Braudeau J. 1969. Le cacaoyer. G.P. Maisonneuve et Larose, Paris, France, 304 p.
- Burle, L. 1952. La production du cacao en Afrique occidentale française. Centre de recherches agronomiques de Bingerville. Bulletin n° 5, pp 3-21.
- Carletto, A. 1946. O numero de cromosômios em cacaueiros. *Boletin tecnico, Instituto de cacau de Bahia* 6 :33-43.
- Castiglioni P., Ajmone-Marsan P., van Wijk R., Motto M. 1999. AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. *Theor. Appl. Genet.* 99:425-431.
- Cheesman E.E. 1944. Notes on the nomenclature, classification possible and relationships of cocoa populations. *Tropical Agriculture* 21:144-159.
- Chagné D., Lalanne C., Madur D., Kumar S., Frigerio J.M., Krier C., Decroocq S., Savouré A., Bou-Dagher-Kharrat M., Bertocchi E., Brach J., Plomion C. 2002. A high density genetic map of Maritime pine based on AFLPs. *Ann. For. Sci.* 59: 627-636.
- Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S. Tingey, S., Morgante M. Rafalsky, A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics* 3:19.
- Clark, A.G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 7:111-122.
- Clement D., Risterucci A.M., Motamayor J.C., N'Goran J., Lanaud C. 2003a. Mapping QTL for yield components, vigor, and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. *Genome* 46:204-12.
- Clement D., Risterucci A.M., Motamayor J.C., N'Goran J.A.K., Lanaud C. 2003b. Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46:103-111.
- Condit, R., and Hubbell, S.P. 1991. Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. *Genome* 34: 66-71.
- Cope F.W. 1962. The mechanism of pollen incompatibility in *Theobroma cacao*. *Heredity* 17:157-182.

- Cregan P.B., Mudge J., Fickus E.W., Marek L.F., Danesh D., Denny R., Mathews B.F., Jarvik T., Young N.D. 1999. Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. *Theor. Appl. Genet.* 98:919-928.
- Crouzillat D., Lerceteau E., Petiard V., Morera J., Rodríguez H., Walker D., Phillips W., Ronning C., Schnell R., Osei J., Fritz P. 1996. *Theobroma cacao* L.: a genetic linkage map and quantitative trait loci analysis. *Theor. Appl. Genet.* 93:205-214.
- Crouzillat D., Menard B., Mora A., Phillips W., Petiard V. 2000a. Quantitative trait analysis in *Theobroma cacao* L. using molecular markers. *Euphytica* 114:13-23.
- Crouzillat D., Phillips W., Fritz P.J., Petiard V. 2000b. Quantitative trait loci analysis in *Theobroma cacao* L. using molecular markers. Inheritance of polygenic resistance of *Phytophtora palmivora* in two related cacao populations. *Euphytica* 114:25-36.
- Crouzillat D., Bellanger L., Rigoreau M., Bucheli P. Pétard V. 2000c. Genetic structure, characterisation and selection of Nacional cocoa compared to other genetic groups. Proceedings of the 3rd International Group for Genetic Improvement of Cocoa (Ingenic) International Workshop on the New Technologies and Cocoa Breeding, 16th-17th October 2000, Kota Kinabalu, Malaysia, pp. 47-64.
- Cuatrecasas J. 1964. Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contributions from the United States Herbarium 35: 379-614.
- Danin-Poleg Y., Reis N., Baudracco-Arnas S., Pitrat M., Staub J.E., Oliver M., Arus P., deVicente C.M., Katzir N. 2000. Simple sequence repeats in *Cucumis* mapping and map merging. *Genome* 43:963-74.
- De la Cruz, M. Whitkus R., Gómez-Pompa A., Mota-Bravo L. 1995. Origins of cacao cultivation. *Nature* 375:542-543.
- Dechesne, F. 2002. Étude du déséquilibre de liaison local chez le sorgho entre marqueurs très liés: microsatellites et RFLP. Diplôme d'Études Supérieures spécialisées. Gestion de la biodiversité : Méthodologies d'Étude et de Valorisation des ressources génétiques. Université Pierre & Marie Curie. 39 p.
- Deu M. and Glaszmann J.C. 2004. Linkage disequilibrium in sorghum. In: Plant & Animal Genomes. XII Conference, 10-14 January, Town & Country Convention Center, San Diego, W10.
- Dettori M.T., Quarta R., Verde I. 2001. A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome* 44:783-790
- Devlin B. and Risch A. 1995. Comparison of Linkage Disequilibrium Measures for Fine-scale Mapping. *Genomics* 29 : 311-322.
- Dillinger, T. L., Barriga P., Escarcega S., Jimenez M., Salazar Lowe D., Grivetti, L. E. 2000. Food of the gods: cure for humanity? A cultural history of the medicinal and ritual use of chocolate. *J. Nutr.* 130:2057S-2072S.
- Dirlewanger E., Cosson P., Tavaud M., Aranzana M.J., Poizat C., Zanetto A., Arús P., Laigret F. 2002. Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.) *Theor. Appl. Genet.* 105:127-138.
- Eaves I.A., Merriman T.R., Barber R.A., Nutland S., Tuomilehto-Wolf E., Tuomilehto J., Cucca F., Todd J.A. 2000. The genetically isolated populations of Finland and

- Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat Genet* 25:320–323
- Edwards K.J., Barker J.H.A., Daly A., Jones C., Karp A. 1996. Microsatellite libraries enriched for several microsatellite sequences in plants. *BioTechniques* 20:758-760.
- Engels, J.M.M, Bartley B., Enríquez G. 1979. Descriptores de cacao, sus clases y modus operandi. CATIE, Costa Rica.
- Engels J.M.M. 1981. Genetic ressources of cacao. A catalogue of the CATIE collection. Technical Bulletin N° 7. Tropical Agriculture and Research Training Center, Turrialba, Costa Rica. pp 191.
- Engels J.M.M. 1986. The systematic description of cacao clones and its significance for taxonomy and plant breeding. PhD thesis, Agricultural University, Wageningen, Netherlands.
- Eskes B. 2001. Introductory notes. Proceedings of the International Workshop on New Technologies for Cocoa breeding, Kota Kinabalu, Malaysia, INGENIC, London, UK, pp 8-11.
- Evans, H.C., Krauss, K., Rios, R.R., Zecevich, T.A. & Arevalo-Gardini, E. 1998. Cocoa in Peru. *Cocoa Growers' Bulletin* 51, 7-22.
- Excoffier L. and Slatkin, M. 1995. Maximum-likelihood estimation of molecular haplotypes frequencies in a diploid population. *Mol. Biol. Evol.* 1921-927.
- Figueira A., Janick P., Goldsbrough P. 1992. Genome size and DNA polymorphism in *Theobroma cacao*. *J. Am. Soc. Hort. Sci.* 117:673-677.
- Figueira A., Janick J., Morris L., Goldsbrough P. 1994. Re-examining the classification of *Theobroma cacao* L. using molecular markers. *Journal of the American Society for Horticultural Science* 119:1073-1082.
- Fisher, R.A. 1921. On the ‘probable error’ of a coefficient of correlation deduced from a small sample. *Metron* 1 pt 4, 3-32.
- Fisher, R. A. 1935. The logic of inductive inference. *Journal of the Royal Statistical Society Series A* 98: 39–54.
- Flament M.H. 1998. Cartographie génétique de facteurs impliqués dans la résistance du cacaoyer (*Theobroma cacao* L.) à *Phytophtora megakaria* et à *Phytophtora palmivora*. These de Doctorat, Ecole Nationale Supérieure agronomique de Montpellier, Ministère de l’Agriculture, Montpellier, France. 113p.
- Flament M.H., Kebe I., Clement D., Pieretti I., Risterucci A.M., N’Goran J.A.K., Cilas C., Despréaux D., Lanaud C. 2001. Genetic mapping of resistance factors to *Phytophtora palmivora* in cocoa. *Genome* 44:79-85.
- Flint-Garcia, S.A., Thornsberry, J.M., Buckler, E.S. 2003. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant. Biol.* 54, 357-74.
- Frankel O.H. and Brown A.H.D. 1984. Plant genetic resources today: a critical appraisal. In: *Crop genetic resources: conservation & evaluation* (Holden JHW and Williams JT, eds). London: George Allen & Unwin; 249–257.
- García de Palacio, D. 1576. Relación de Diego García de Palacio. Universidad Nacional autónoma de México (UNAM), Centro de Estudio Maya, ed. 1983.

- Garris A.J., Mc Couch S.R., Kresovich S. 2003. Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the *Xa5* locus of rice *Oryza sativa* L. *Genetics* 165:759-769.
- Garnier-Gere, P., and C. Dillmann. 1992. LinkDos. *J. Heredity* 83:409-415
- Glicenstein L.J. and Fritz P.J.. 1989. Meiosis in *Theobroma cacao*. *Turrialba* 39:497-500.
- Goldstein D.B. 2001. Islands of linkage disequilibrium. *Nature Genetics* 29:109-211.
- Gordon D., Simonic I., Ott J. 2000. Significant evidence for linkage disequilibrium over a 5-cM region among Afrikaners. *Genomics* 66:87-92.
- Gorelick R., and M.D. Laubichler. 2004. Decomposing multilocus linkage disequilibrium. *Genetics* 166:1581-3.
- Gouesnard, B., Bataillon T.M., Decoux G., Rozale C., Schoen D.J., David J.L. 2001. MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. *J. Hered.* 92(1):93-94.
- Grattapaglia D. and Sederoff, R. 1994. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-test-cross: mapping strategy and RAPD markers. *Genetics* 137:1121-1137.
- Guo, S.W. and Thompson, E.A. 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics* 48: 361-372.
- Gupta P.K., Balyan H.S., Edwards K.J., Isaac P., Korzun V., Röder M., Gautier M-F., Joudrier P., Schlatter A.R., Dubcovski J., De la Pena R.C., Khairallah M., Penner G., Hayden M.J., Sharp P., Keller B., Wang R.C.C., Hardouin J.P., Jack P., Leroy P. 2002. Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *Theor. Appl. Genet.* 105:413-422
- Hagenblad J. and Nordborg, M. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. *Genetics*, 161, 289–98.
- Haldane, J.B.S. 1919. The mapping function. *J. Genet.* 8: 299-309.
- Hamilton D.C. and Cole D.E.C. 2004. Standardizing a composite measure of linkage disequilibrium. *Annals of human genetics* 68:234-239.
- Hansen M., Kraft T., Ganestam S., Sall T., Nilsson N-O. 2001. Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. *Genet. Res.* 77:61-66.
- Haussmann B.I.G., Hess D.E., Seetharama N., Welz H.G., Geiger H.H. 2002. Construction of a combined sorghum linkage map from two recombinant inbred populations using AFLP, SSR, RFLP, and RAPD markers, and comparison with other sorghum maps. *Theor. Appl. Genet.* 105:629-637.
- Hedrick, P.W. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117:331-341.
- Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 226-231.
- Hill, W.G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33:229-238.

- Hill W.G. and Weir, B.S. 1994. Maximum likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. 54:705-714.
- Holden J.H.W. 1984. The second ten years. In: Crop genetic resources: conservation and evaluation, (eds J.H.W Holden and J. Williams), George Allen and Unwin, Winchester, Massachusetts, pp. 277285.
- Horikawa Y., Oda N., Cox N., Li X., Orho-Melander M., Hara M., Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI. 2000. Genetic variation in the gene encoding calpain-10 is associated with type diabetes mellitus. Nat. Genet. 26:163-175.
- Hyten D.L., Song Q., Cregan P.B. 2004. Linkage disequilibrium in four soybean populations. In: Plant & Animal Genomes. XII Conference, 10-14 January, Town & Country Convention Center, San Diego, CA p534.
- IBGRI. 1981. Report of the IBPGR working group on genetic resources of cocoa. Rome, Italie, IBPGR, ACP: IBPGR /80/56, 28p.
- Ivandic V., Thomas W.T.B., Nevo E., Zhang Z., Forster B.P. 2003. Associations of simple sequence repeats with quantitative trait variation including biotic and abiotic stress tolerance in *Hordeum spontaneum*. Plant breeding 122:300-304.
- Iwaro A.D., Bekele F.L., Butler D.R. 2003. Evaluation and utilisation of cacao (*Theobroma cacao* L.) germplasm at the International Cocoa Genebank, Trinidad. Euphytica 130:207-221.
- Jannoo N., Grivet L., Dookun A., D'Hont A., Glaszmann J.C., 1999. Linkage disequilibrium among modern sugarcane cultivars. Theor. Appl. Genet. 99: 1053-1060.
- Jennings, H.S. 1917. The numerical results of diverse systems of breeding with respect to two pairs of characters, linked or independent with special relation to the effects of linkage. Genetics 2:97-154.
- Jones E. S., Dupal M.P., Dumsday J.L., Hughes L.J., Forster J.W. 2002. An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). Theor. Appl. Genet.. 105:577-584.
- Joobeur T., Periam N., de Vicente M.C., King G.J., Arús P. 2000. Development of a second generation linkage map for almond using RAPD and SSR markers. Genome 43:649-655
- Jorde L.B. 2000. Linkage disequilibrium and the search for complex disease genes. Genome Research, 10 : 1435-1444.
- Kauffmann S., Legrand M., Geoffroy P., Fritig B. 1987. Biological function of "pathogenesis-related" proteins: four PR proteins of tobacco have 1,3 – glucanase activity. EMBO J. 6:3209-3212
- Kijas J.M.H., Fowler J.C.S., Garbett C.A.. 1994. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. BioTechniques 16:656-662.

Références bibliographiques

- Kijas J.M.H., Thomas M.R., Fowler J.C.S., Roose M.L. 1997. Integrating of trinucleotides microsatellites into a linkage map of *Citrus*. *Theor. Appl. Genet.* 94:701-706.
- Knight R. and Rogers, H.H. 1955. Incompatibility in *Theobroma cacao*. *Heredity* 9:69-77.
- Knowler W.C., Williams R.C., Pettitt D.J., Steinberg A.G. 1988. Gm^{3-5,13,14} and type 2 diabetes mellitus: an association with genetic admixture. *Am. J. Hum. Gen.* 43:520-526.
- Kosambi, D.D. 1944. The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172-175.
- Kraakman, A.T.W., Rients E. N., van den Berg, P.M.M.M., Stam P., Van Eeuwijk F.A. 2004. Linkage disequilibrium mapping of yield and yield stability in Modern Spring Barley Cultivars. *Genetics* 168:436-444.
- Kraft T., Hansen M., Nilsson N.O. 2000. Linkage disequilibrium and fingerprinting in sugar beet. *Theor. Appl. Genet.*, 101: 323-326.
- Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22: 139-144.
- Lachenaud P. 1991. Facteurs de la fructification chez le cacaoyer *Theobroma cacao* L. Thèse de Doctorat, Ecole Nationale Supérieure Agronomique Paris-Grignon, France, 188 p.
- Lanaud C. 1987. Nouvelles données sur la biologie du cacaoyer (*Theobroma cacao* L.) : diversités de populations, systèmes d'incompatibilité, haploïdes spontanés ; leurs conséquences pour l'amélioration génétique de cette espèce. Thèse de doctorat, Université Paris XI, Orsay, France, 106 p.
- Lanaud C., Boult E., Clapperton J., Cros E., Chapelin M., Risterucci A.M., Allaway D., Gilmour M., Cattaruzza A., Fouet O. N'goran J.A.K., Petithuguenin P. 2003. Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma Cacao* L. In: 14ème Conférence internationale sur la recherche cacaoyère. Accra, Ghana, 13-17 octobre 2003.
- Lanaud C., Ham P., Duperray C. 1992. Estimation of nuclear DNA content of *Theobroma cacao* L. by flow cytometry. *Café, Cacao, Thé* 36:3-8.
- Lanaud C., Kébé I., Risterucci A.M., Clement D., N'Goran J.A.K., Grivet L., Tahí M., Cilas C., Pieretti I., Eskes A.B., Despréaux D. 1999. Mapping quantitative loci (QTL) for resistance to Phytophtora palmivora in *T. cacao* L. 12th International Cocoa Research Conference, November 17-23, Salvador, Bahia, Brazil pp. 99-105.
- Lanaud C., Motamayor J.C., Sounigo O. 2003. Cacao. In: Hamon P., Seguin M., Perrier X., Glaszmann J.C. (eds.) Genetic Diversity of cultivated tropical plants, pp 125-156, CIRAD, Montpellier, France.
- Lanaud C., Risterucci A.M., N'Goran J.A.K., Clement D., Flament M.H., Laurent V., Falque M. 1995. A genetic linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.* 9:987-993
- Lanaud C., Risterucci A.M., Pieretti I., Falque M., Bouet A., Lagoda P.J.L. 1999. Isolations and characterization of microsatellites in *Theobroma cacao* L. *Mol. Ecol.* 8:2141-2152.

- Lanaud C., Risterucci A.M., Pieretti I., N'Goran J.A.K., Fargeas D. 2003. Characterisation and genetic mapping of resistance and defence gene analogs in cocoa (*Theobroma cacao* L.) Molecular Breeding 13:211-227.
- Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln and L. Newburg, 1987. MAPMAKER: an interactive computer package of constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.
- Laurent V., Risterucci A.M., N'Goran A.K.J., Clement D., Flament V., Falque M. 1994. Genetic diversity in cocoa revealed by cDNA probes. Theor. Appl. Genet. 91:987-993.
- Lespinasse D. 1999. Cartographie génétique de l'hévéa (*Hevea* spp) et déterminisme de la résistance au champignon pathogène *Microcyclus ulei*. Thèse. Université Montpellier II. 96 p.
- Lespinasse D., Rodier-Goud M., Grivet L., Leconte A., Legnate H., Seguin M. 2000. A saturated genetic linkage map of rubber tree (*Hevea* spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. Theor. Appl. Genet. 100: 127-138
- Lecertau E., Robert T., Pétiard V., Crouzillat D. 1997. Evaluation of the extent of genetic variability among *Theobroma cacao* accessions using RAPD et RFLP markers. Theor. Appl. Gen. 95:10-19.
- Levinson G. and Gutman G.A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203-221.
- Lewis P.O. and Zaykin D. 2001. Genetic data analysis: computer program for the analysis of allelic data, version 1.0 (d16c). Disponible on <http://lewis.eeb.uconn.edu/lewishome/software.html>.
- Lewontin R.C. and Kojima K. 1960. The evolutionary dynamics of complex polymorphism. Evolution 14:458-472.
- Lewontin R. 1964. The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49:49-67.
- Lewontin R.C. 1988. On measures of gametic disequilibrium. Genetics 120:849-852.
- Liebhard R., Gianfranceschi L., Koller B., Ryder C.D., Tarchini R., Van De Weg E., Gessler C. 2002. Development and characterisation of 140 new microsatellites in apple (*Malus x domestica* Borkh.) Molecular Breeding 10:217-241.
- Lockwood G. and End M. 1992. History, technique and future needs for Cocoa collection. In: International Workshop one the conservation, characetrization and utilization of cocoa genetic resources in the 21st Century, Port of Spain, Trinidad and Tobago, pp 1-14.
- Lockwood, R. 2003. Who needs clothing? INGENIC Newslet. 8:2-5.
- Lorieux M. 1993. Cartographie des marqueurs moléculaires et distorsions de ségrégation : modèles mathématiques. Thèse Université Montpellier II. 133 p.
- Macaulay M., Ramsay L., Powell W., Waugh R. 2001. A representative, highly informative 'genotyping set' of barley SSRs. Theor. Appl. Genet. 102:801-809.

- Maccaferry M., Sanguinetti M.C., Noli E., Tuberosa R. 2004. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. In: Plant & Animal Genomes. XII Conference, 10-14 January, Town & Country Convention Center, San Diego, p 416.
- Martin G.B., Frary A., Wu R., Brommonschenkel S.H., Chunwongse J., Earle E.D., Tanskley S.D. 1994. A member of the tomato *Pto* gene family confers sensitivity to fenthion in rapid cell death. *Plant Cell* 6:1543-1552
- McCouch S.R., Teytelman L., Xu Y., Lobos K.B., Clare K., Walton M., Fu B., Maghirang R., Li Z., Xing Y., Zhang Q., Kono I., Yano M., Fjellstrom R., De Clerck G., Schneider D., Cartinhour S., Ware D., Stein L. 2002. Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Research* 9:199-207.
- Mohlke K.L., Lange E.M., Valle T.T., Ghosh S., Magnusson V.L., Silander K., Watanabe R.M., Chines P.S., Bergman R.N., Tuomilehto J., Collins F.S., Boehnke M. 2001. Linkage disequilibrium between microsatellite markers extends beyond 1cM on chromosome 20 in finns. *Genome research*, 11: 1221-1226.
- Morris D. 1882. Cocoa: how to grow and how to cure it. Jamaica, pp. 1-45.
- Morton, N E. 1955. Sequential tests for the detection of linkage Amer. J. Hum. Gen 7: 277-318.
- Motamayor J.C., Risterucci A.M., Laurent V., Moreno A., Lanaud C. 1997. The genetic diversity of Criollo cacao (*Theobroma cacao* L.) and its consequence in quality breeding. Conference in the 1st Venezuelan Cocoa Congress. Maracay, November 17-21.
- Motamayor, J.C., Risterucci A.M., Lopez P.A., Ortiz C.F., Moreno A., Lanaud C. 2002. Cacao domestication I. The origin of the cacao cultivated by the Mayas. *Heredity* 89:380-386.
- Motamayor, J.C., Risterucci A.M., Heath M., Lanaud C. 2003. Cacao domestication II. Progenitor germplasm of the Trinitario cacao cultivar. *Heredity* 91:322-330.
- Motilal L. and Butler D. 2003. Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution* 50: 799-807.
- Neale D.B. and Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* 9:325-329.
- Nei M. and. Li, W.H. 1973. Linkage disequilibrium in subdivided populations. *Genetics* 75:213-219.
- N'Goran, J.A.K., Larent V., Risterucci A.M., Lanaud C. 1994. Comparative genetic diversity of *Theobroma cacao* L. Using RFLP and RAPD markers. *Heredity* 73:589-597.
- Nguyen T.B., Gibaud M., Brottier P., Risterucci A.M., Lacape J.M. 2004. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *Theor. Appl. Genet.* 109:167-75.
- Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial selffertilization. *Genetics* 154:923-29.

- Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 30, 190–93.
- Olsen K.M., Halldorsdottir S.S., Stinchcombe J.R., Weinig C., Schmitt J., Purugganan M.D. 2004. Linkage disequilibrium mapping of *Arabidopsis* CRY2 flowering time alleles. *Genetics* 167:1361-1369.
- Oviedo G.F. 1944. Historia general y natural de las Indias, Madrid. P. 246.
- Paradis, L. 1979. Le cacao précolombien: monnaie d'échange et breuvage des dieux. *Journal d'agriculture traditionnelle et de botanique appliquée*, 26:3-4.
- Paulin D., Decazy B., Coulibaly N. 1983. Etude des variations saisonnières des conditions de pollinisations et de fructification dans un cacaoyer. *Café, Cacao Thé* 27 :165-175.
- Perrier, X., Flori A., Bonnot F. 2003 Methods of data analysis. In: Hamon P., Seguin M., Perrier X., Glaszmann J.C. (eds.) *Genetic Diversity of cultivated tropical plants*, pp 43-76, CIRAD, Montpellier, France.
- Petithuguenin P. et Daviron B. 1995. Les tendances du marché mondial du cacao, quelques points de repères sur les volumes, les qualités et les prix. *Journée MITECH*. CIRAD CP.
- Pittier H. 1935. Degeneration of cacao through natural hybridization. *The journal of heredity* 36:385-390.
- Pittier H. 1930. A propos des cacaoyer spontanés. *Revue de botanique appliquée* 10:777.
- Pound F. J. 1945. A note of the cacao population of South America. *Rep. and Proc. Cocoa Res. Conf.*, London 1945. Colonial, 192:95-7. Reprinted 1982 in *Arch. Cocoa Res.*, 1:93-97.
- Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
- Pritchard J. K. and Przeworski, M. 2001. Linkage disequilibrium in Humans: models and data. *American Journal Human Genetics*, 69: 1-14.
- Purseglove, J.W. 1968. *Theobroma* L. In: Purseglove, J.W. (ed) *Tropical Crops. Dicotyledons* 2. John Wiley & Sons, New York, pp 291-295.
- Pugh T., Fouet O., Risterucci A.M. , Brottier P., Abouladze M., Deletrez C., Courtois B., Clement D., Larmande P., N'Goran J.A.K., Lanaud C. 2004. A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* 108:1151-1161.
- Queiroz V.T., Guimarães C.T., Anhert D., Schuster I., Daher R.T., Pereira M.G., Miranda V.R.M., Loguerio L.L., Barros E.G., Moreira M.A. 2003. Identification of a major QTL in cocoa (*Theobroma cacao* L.) associated with resistance to witches' broom disease. *Plant Breeding* 122:268-272
- Rafalski, A. and Morgante, M. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* 20:103-11.
- Ramsay L., Macaulay M., degli Ivanissevich S., MacLean K., Cardle L., Fuller J., Edwards K.J., Tuveson S., Morgante M., Massari A., Maestri E., Marmiroli N., Sjakste T., Ganal M., Powell W., Waugh R. 2000. A simple sequence repeat-based linkage map of barley. *Genetic* 156:1997-2005.

- Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R., Lander E.S. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199-204.
- Remington D.L., Thornsberry J.M., Matsuoka Y., Wilson L.M., Whitt S.R., Doebley J., Kresovich S., Goodman M.M., Buckler E.S. 4th. 2001. Structure of linkage disequilibrium and phenotypic associations in maize genome. *Proceedings of the National Academy of Science* 98: 11479-11484.
- Reyes, H.E. 1992. Criollo cacao germplasm in Venezuela. In: International workshop on conservation, characterisation and utilisation of cocoa genetic in the 21st century, Trinidad and Tobago. pp 244-252.
- Rice W.R. 1989. Analysing tables of statistical tests. *Evolution*, 43, 223-225.
- Risterucci A.M., Grivet L., N'Goran J.A.K., Pieretti I., Flament M.H., Lanaud C. 2000. A high density linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.* 101:948-955.
- Risterucci A.M., Paulin D., N'Goran J.A.K., Lanaud C. 2003. Identification of QTL related to cocoa resistances to three species of *Phytophtora*. *Theor. Appl. Genet.* 108:168-174.
- Ritter, E., Gebhardt C., Salamini, F. 1990. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* 125, 645-654.
- Ritter E. and Salamini F. 1996. The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. *Genet. Res.* 67: 55-65.
- Ronning C.M., et Schnell R.J. 1994. Allozyme diversity in a germplasm collection of *Theobroma cacao* L. *J. Hered.* 85:291-295.
- Ronning, C.M., Schnell R.J., Kuhn D.N. 1995. Inheritance of random amplified polymorphic DNA (RAPD) markers in *Theobroma cacao* L. *J. Am. Soc. Hort. Sci.* 120:681-686.
- Ruf, F. 1995. Booms et crises du cacao. Les vertiges de l'or brun. In : Ministère de la Coopération, CIRAD-Sar et KARTHALA, Montpellier-Paris, France 35 p.
- Russel J.R., Hosein F., Johson E., Waugh R., Powell W. 1993. Genetic differentiation of cocoa (*Theobroma cacao* L.) populations revealed by RAPD analysis. *Mol. Ecol.* 2:89-97.
- Saitou N., and M. Nei. 1987. The neighbour joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and Evolution* 4: 406-425.
- Schoen D. J. and Brown A.H.D. 1993. Maximizing allelic diversity in core collections of wild crop relatives: the role of genetic markers. *Proc. Natl. Acad. Sci. USA* 90:1063-1067.
- Simons M., Nielsen R., Jones M., McCouch S. 2004. The population structure of African cultivated rice (*Oryza Glaberrima* (Steud.)): evidence for elevated levels of LD caused by admixture with *O. sativa* and ecological adaptation. *Genetics (on line)*

- Shifman S., Kuypers J., Kokoris M., Yakir B., Darvasi A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics* 12:771-776.
- Shriver, M.D., Jin, L., Chakraborty, R. and Boerwinkle, E. 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, 134, 983–993.
- Simko I., Costanzo S., Haynes K.G., Christ B.J., Jones R.W. 2004a. Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. *Theor. Appl. Genet.* 108:217-24.
- Simko I., Haynes K.G., Ewing E.E., Costanzo S., Christ B.J., Jones R.W. 2004b. Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Mol. Genet. Genomics.* 271:522-31.
- Soria J.V. 1970. Principal varieties of cocoa cultivated in tropical america. *Cocoa Growers' Bull.* 19:12-21.
- Sounigo O., Christopher Y., Umaharan R. 1996 Genetic diversity assessment of *Theobroma cacao* L. using isoenzyme and RAPD analyses. In: Annual Report 1996, CRU.
- Stam P. 1995. Construction of integrated genetic linkage maps by means by a new computer package: JoinMap. *The Plant Journal* 3:739-744.
- Stephens J.C., Schneider J.A., Tanguay D.A., Choi J., Acharya T., Stanley S.E., Jiang R., Messer C.J., Chew A., Han J.H., Duan J., Carr J.L., Lee M.S., Koshy B., Kumar A.M., Zhang G., Newell W.R., Windemuth A., Xu C., Kalbfleisch T.S., Shaner S.L., Arnold K., Schulz V., Drysdale C.M., Nandabalan K., Judson R.S., Ruano G., Vovis G.F. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489-493.
- Stephens M., Smith N., Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978-989.
- Taillon-Miller P., Bauer-Sardina I., Saccone N., Putzel J., Laitinen T., Cao A., Kere J., Pilia G., Rice J.P., Kwok P.Y. 2000. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* 25:324-328.
- Tautz D. and Rentz.M. 1984. Simple sequences are ubiquitous repetitive components of eucariotic genome. *Nucleic Acids Research*, 12: 4127-4138.
- Temnykh S., Park W.D., Ayres N., Cartinhour S., Hauck N., Lipovich L., Cho Y.G., Ishii T., McCouch S.R. 2000. Mapping and genome organisation of microsatellites sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100:697-712
- Tenaillon M.I., M. C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doeley et al., 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize. *Proceedings of the national academy of science.*98: 9161-9166.
- Tenesa A., Knott S.A., Carothers A.D., Visscher P.M. 2003. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. *Ann Hum Genet.* 67:557-66.

- Tenesa A., Knott S.A., Ward D., Smith D., Williams J.L. Visscher P.M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *J. Anim. Sci.* 81:617-623
- Thornberry J.M., Goodmen M.M., Doebley J., Kresowich S., Nielsen D., Buckler E.S. IV. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genet.* 28:286-289.
- Torquemada J. de. 1723. Monarquía Indiana. Madrid, éd. 1989.
- Van Ooijen, J.W., and C. Maliepaard. 1996. MapQTL (tm) Version 3.0: Software for the Calculation of QTL Positions on Genetic Maps. DLO-Centre for Plant Breeding and Reproduction Research, Wageningen, the Netherlands.
- Van Ooijen JW, Voorrips E. 2001. Joinmap ® Version 3.0, Software for the calculation of genetic linkage maps Plant Research International, Wageningen, the Netherlands.
- Viruel M.A., Messeguer R., De Vicente M.C., Garcia-Mas J., Puigdomènech P. 1995. A linkage map with RFLP and isozyme markers for almond. *Theor. Appl. Genet.* 91:964-971.
- Viruel M.A. and Hormaza J.I. 2004. Development, characterization and variability analysis of microsatellites in lychee (*Litchi chinensis* Sonn., Sapindaceae). *Theor. Appl. Genet.* 108:896-902.
- Wadsworth, R.M. and Hardwood, T. 2000. International Cocoa Germplasm Database ICDG version 4.1. London International Financial Futures and Options Exchange and the University of Reading, UK.
- Wall J.D., Pritchard J.K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet.* 4:587-597.
- Warren, J.M. 1994. Isozyme variation in a number of populations of *Theobroma cacao* L. obtained through various sampling regimes. *Euphytica* 72:121-126.
- Weber JL. 1990. Informativeness of human (dC-dA)n (dG-dT)n polymorphism. *Genomics* 7:524-530
- Weir B.S. 1979. Inferences about linkage disequilibrium. *Biometrics* 35:235-254.
- Weir B.S. 1996. Genetic data analysis II. Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts, 445 pp.
- Weir, B.S. and Cockerham C.C. 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 42:105-111
- Weir B.S. and Cockerham C.C. 1989. Complete characterisation of disequilibrium at two loci. In: Mathematical Evolutionary theory, edited by M.W. Feldman. Princeton University Press. Princeton, N.Y., pp 86-110.
- Whitlock, B., Bayer C., Baum D. 2001. Phylogenetic relationships and floral evolution of the Byttherioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene ndhF. *Sys. Botany* 26:420-437.
- Wilde J., Waugh R., Powell W. 1992. Genetic fingerprinting of *Theobroma* clones using randomly amplified polymorphic DNA markers. *Theor. Appl. Genet.* 83:871-877.
- Whitkus R. de la Cruz M., Mota-Bravo L. 1998. Genetic diversity and relationships of cocoa (*Theobroma cacao* L.) in southern Mexico. *Theor. Appl. Genet.*, 96: 621-627.

Références bibliographiques

- Witham S, Dinesh-Kumar SP, Choll D, Hehl R, Corr C, Baker B. 1994. The product of the tobacco mosaic virus resistance gene N: Similarity to Toll and the interleukin-1 receptor. *Cell* 78:1101-1115.
- Wood G.A.R and Lass R.A. 1985. *Cocoa*. Longman. London
- Wood G.A.R. 1991. A history of early cocoa introductions. *Cocoa Growers' Bulletin* 44:7-12.
- Wu S.B., Collins G., Sedgley, M. 2004. A molecular linkage map of olive (*Olea europaea* L) based on RAPD, microsatellite, and SCAR markers. *Genome* 47:26-35.
- Young A.M. 1982. Population biology of tropical insects. Plenum press.
- Young A.M. 1983. Seasonal difference in abundance and distribution of cocoa-pollinating midges in relation to flowering and fruit-set between shaded and sunny habitats of "La Lola". *Cocoa Farm in Costa Rica. J. Appl. Ecol.* 20:801-831.
- Young A.M. 1984. Pollen-collecting by singles bees on cocoa flowers. *Experientia*, 41:760-762.
- Young A.M. 1985. Research on the natural pollination of cocoa in central America: overview of current directions. In: Proc. 9 Int. Cocoa Res. Conf. Lome, Togo. Stephen Austin 1 Sons Ltd.
- Young A.M. 1994. *The chocolate tree: a natural history of cacao*. Washington, Etats-Unis, Smithsonian Institution Press, 200p.
- Zhao, X. P. and Kochert, G. 1993. Phylogenetic distribution and genetic mapping of a (GGC)(n)microsatellite from rice (*Oryza sativa* L.). *Plant Mol. Biol.* 21: 607-614.

Abstract

Cocoa domestication has shaped the genome structure of modern Criollo/Trinitario varieties, a high quality and aromatic chocolate source. These varieties have originated from a reduced number of Criollo and Forastero ancestors first crossed 250 years ago. A cacao germplasm collection of about 800 accessions of different origin is conserved in the CATIE (Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica). A significant proportion of these accessions correspond to modern Criollo/Trinitario. The narrow genetic base of the modern Criollo/Trinitario group and the small number of generations happened from the formation of the first hybrids cross is a favourable situation to develop association mapping studies in this genetic group. In order to have a set of useful markers to evaluate genetic diversity and to develop an association mapping study, a linkage map of cacao based on codominant markers has been constructed by integrating two hundred and one new simple sequence repeats (SSR) with a number of codominant markers previously mapped. The new map contains 465 markers (268 SSR, 176 RFLP, 5 isoenzymes and 16 Rgenes-RFLP). Its length is 782.8 cM, with an average interval distance between markers of 1.7 cM. The current level of genome coverage is approximately 1 microsatellite every 3 cM. The genetic diversity of 247 modern Criollo/Trinitario varieties from this germplasm collection was evaluated using 34 microsatellites markers widespread in the cacao genome. An average of 5.2 allele number per locus total was detected. Two alleles more frequent than the others were always found at each locus. These alleles were identified as alleles originated from Ancient Criollo and Lower Amazon Forastero, individuals recognized as the founder genotypes of the Modern Criollo/Trinitario varieties. Among the 247 accessions, 150 were characterised to fruit, seed and flower traits 25 years ago and data stored in the International Cocoa Germplasm Database. These varieties were used to study the extend of linkage disequilibrium (LD) conserved along the genome. A first set of 24 SSR were used to verify the absence of population structure in this collection and the genome-wide LD. One hundred and ten SSR, with alleles specific to Criollo ancestors were used to evaluate the extend of LD. LD values decreased with an increasing genetic distance between loci, the genetic distance with LD being variable and up to 30 cM. A total of 13 genomic regions were identified as involved in the variation of these traits. Among them 8 corresponded to genomic regions where QTL (quantitative trait loci) were already identified by classical QTL mapping studies. These results demonstrate that association studies approaches represent a valuable tool to help to identify the genetic or molecular bases of traits of interest and to valorise hundred of morphological and agronomic data accumulated to characterise large germplasm collections.

Tatiana PUGH MORENO (2005) - Ecole Nationale Supérieure Agronomique de Montpellier

Titre de la thèse : Etude des déséquilibres de liaison dans une collection de cacaoyers (*Theobroma cacao L.*) appartenant au groupe Criollo/ Trinitario et application au marquage génétique des caractères d'intérêt

Résumé :

La structure du génome des variétés de Criollo modernes et Trinitario de cacaoyer résulte d'une domestication récente datant de 250 ans et qui a mis en jeu un nombre réduit d'ancêtres parentaux : Criollo ancien et Forastero bas amazoniens. Une collection de 247 accessions de Criollo modernes et Trinitario conservées au CATIE (Centro Agronómico Tropical de Investigación y Enseñanza, Costa Rica) a été étudiée. La base génétique très étroite de ce groupe génétique et le petit nombre de générations de recombinaisons qui se sont produites depuis les premières hybridations entre ancêtres parentaux en font un bon modèle pour des études d'associations.

Afin de disposer d'un outil d'analyse performant pour ces études d'association, une carte génétique saturée en marqueurs SSR (single sequence repeat) a été produite. Cette carte contient 465 marqueurs (268 SSR, 176 RFLP, 5 isoenzymes et 16 Rgenes). Sa longueur est de 783 cM, avec une distance moyenne entre marqueurs de 1,7 cM et une distance moyenne entre marqueurs SSR de 3 cM.

La diversité génétique des 247 accessions de Criollo modernes /Trinitario a été évaluée à l'aide de 34 SSR répartis sur tout le génome du cacaoyer. Une moyenne de 5,2 allèles par locus a été observée. Deux allèles, toujours plus fréquents que les autres, correspondent à ceux des principaux ancêtres fondateurs: Criollo ancien et Forastero bas-amazoniens. Un sous ensemble de 37 accessions représentant 99,5% de la variabilité allélique totale de cette collection de Criollo/Trinitario a été identifiée.

Une caractérisation morphologique de 150 de ces accessions a été faite il y a 25 ans par Engels et al. (1981), et ces données sont accessibles par la base de données ICGD (International Cocoa Germplasm Database). Ces 150 accessions ont été utilisées pour nos études de déséquilibre de liaison. Un premier lot de 24 SSR a servi à vérifier l'absence de structure de cette collection ainsi que le déséquilibre de liaison (DL) global qui est apparu faible. Puis 110 SSR, pour lesquels des allèles spécifiques des Criollo ont été identifiés, ont été utilisés pour évaluer l'étendue du DL. Celui est variable mais s'étendre jusqu'à 30 cM. Des associations entre caractères de fruit, de fève et de fleur ont été identifiées sur 13 régions chromosomiques. Parmi elles 8 correspondent à des régions où des QTL portant sur les mêmes caractères ont déjà été identifiés par des approches classiques. Ces résultats nous démontrent l'intérêt et la faisabilité de ces études d'association pour identifier les bases génétiques des caractères d'intérêt chez le cacaoyer et représentent une valorisation supplémentaire des données de caractérisation de collection qui existent.

Mot-clefs : *Theobroma cacao L.*, carte génétique, diversité génétique, déséquilibre de liaison, étude d'associations.