Le génome du cacaoyer : du décodage de sa séquence jusqu'à l'étude des gènes impliqués dans des caractères agronomiques d'intérêt

Xavier Argout

THÈSE

Pour obtenir le grade de

Docteur

Délivré par **Montpellier SupAgro**

Préparée au sein de l'école doctorale **GAIA**
Et de l'unité de recherche **AGAP**

Spécialité : **Biologie, Interactions, Diversité Adaptative des Plantes**

Présentée par **Xavier ARGOUT**

Le génome du cacaoyer : du décodage de sa séquence jusqu'à l'étude de gènes impliqués dans des caractères agronomiques d'intérêt.

Soutenue le 08 mars 2017 devant le jury composé de

M. Patrick WINCKER, Chercheur-HDR, Genoscope/CEA	Rapporteur
M. Christophe PLOMION, Directeur de recherche, INRA	Rapporteur
M. Jacques DAVID, Professeur, Montpelier SupAgro	Examinateur
M. Francis QUETIER, Professeur émérite, Univ. Evry	Examinateur
Mme Claire LANAUD, Chercheur-HDR, CIRAD	Directeur de thèse

Remerciements

Tout d'abord je tiens à remercier les membres du jury, Patrick Wincker et Christophe Plomion, rapporteurs de ce travail de thèse et Jacques David et Francis Quétier, examinateurs. Merci d'avoir sacrifié une partie de votre précieux temps pour lire et juger ce travail.

J'adresse également mes remerciements à Philippe Lachenaud qui a présidé mon comité de thèse et qui a partagé avec moi ses formidables connaissances sur le cacaoyer, notamment durant nos missions de terrain.

Je tiens à exprimer toute ma gratitude à Brigitte Courtois, Olivier Fouet, Gaetan Droc, Guillaume Martin, Manuel Ruiz, Baptiste Guitton, Frédéric De Lamotte, Stéphanie Sidibe-Bocs, André Clément-Demange, Vincent le Guen et Emannuel Guiderdoni pour leur collaboration dans les travaux de thèse.

Merci également à Fabien De Bellis pour ses coups de main et son soutien dans la dernière ligne droite.

Merci à tous mes amis du bâtiment 3 qui m'ont aidé et soutenu, au travail et en dehors.

Mes remerciements également aux collègues de la filière Cacao pour leur disponibilité.

Enfin je tiens à remercier profondément Claire Lanaud qui m'a donné l'opportunité de rejoindre son équipe de recherche et qui a encadré cette thèse. Ta passion pour le cacaoyer est contagieuse! Un grand MERCI pour m'avoir guidé, fait confiance et soutenu depuis maintenant de nombreuses années.

TABLE DES MATIÈRES

Avant-propos

Theobroma cacao L. est un arbre fruitier tropical, diploïde (2n=2x=20), à petit génome (430Mb) et endémique des forêts tropicales d'Amérique du Sud. Ses fruits contiennent des fèves qui après fermentation, séchage et torréfaction servent à la fabrication du chocolat. La culture du cacaoyer (cacaoculture) permet à des millions de petits producteurs répartis dans les zones tropicales du monde entier de dégager des revenus substantiels. La résistance aux maladies et l'amélioration de la qualité des fèves sont deux défis extrêmement importants pour l'ensemble des acteurs impliqués dans la cacaoculture et la production de chocolat. Par exemple la perte de rendement liée aux différentes maladies affectant le cacaoyer est estimée à environ 30%. De plus l'amélioration des qualités nutritionnelles et organoleptiques est de plus en plus demandée par les consommateurs de chocolat. Si certains de ces caractères ont un déterminisme génétique simple, la plupart possèdent une hérédité polygénique et leur manipulation demande d'établir leurs bases génétiques par des approches de génétique quantitative et moléculaire.

Depuis plusieurs années, les programmes de recherche ont mis l'accent sur l'étude des bases génétiques de plusieurs caractères d'intérêt agronomique chez le cacaoyer. De nombreux outils moléculaires ont été élaborés pour ces études : cartes génétiques, marqueurs moléculaires, QTL, etc…

En 2008 les transcriptomes de plusieurs variétés de cacaoyers ont été séquencées et en 2010 le génome d'une variété Criollo. Cette avancée majeure pour la recherche cacaoyère offre de nouvelles perspectives pour identifier les régions géniques qui sous-tendent les caractères agronomiques d'intérêt.

Plus récemment, le pouvoir de résolution des méthodes de génotypage s'est fortement accru et de nouvelles méthodes de séquençage ont permis de progresser dans la qualité de la séquence du génome.

Dans le cadre de cette thèse, je présenterai dans les 2 premiers chapitres, la constitution des ressources moléculaires que nous avons effectuée et qui a abouti à la publication du transcriptome et de la séquence complète du génome de la variété Criollo.

Un troisième chapitre portera sur l'analyse d'une combinaison de nouvelles données de génotypage et de séquençage utilisée pour produire une version améliorée de la séquence de référence du génome du Criollo.

Dans un quatrième chapitre, je m'appuierai sur la version améliorée du génome pour rechercher les gènes candidats impliqués dans la variation de la couleur des fèves (par l'étude des voies de biosynthèse des anthocyanes). Ces études s'appuieront sur la cartographie fine de ces caractères à partir d'une grande descendance implantée en Guyane et conduira à la recherche de gènes candidats dans les régions génomiques concernées.

Enfin un cinquième chapitre présentera les conclusions et perspectives qui peuvent être faites à partir des résultats obtenus.

INTRODUCTION

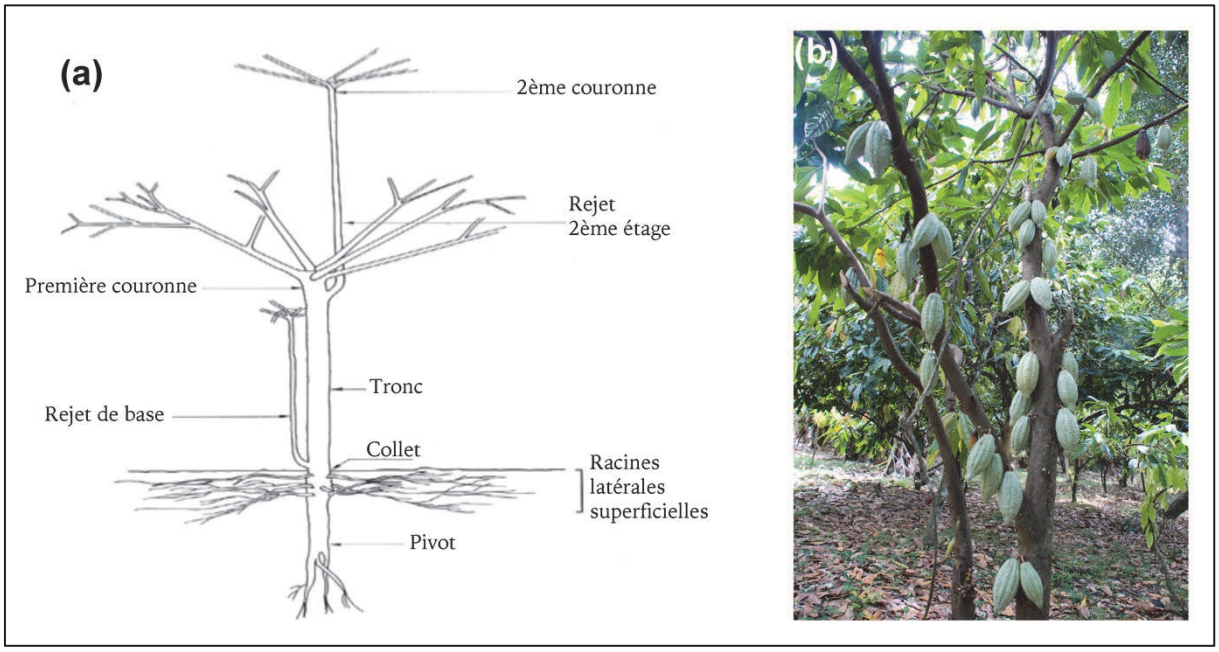


Figure 1 : le cacaoyer. (a) Représentation schématique de l'architecture du cacaoyer (d'après Mossu 1990. (b) Photographie d'un cacaoyer (X. Argout).

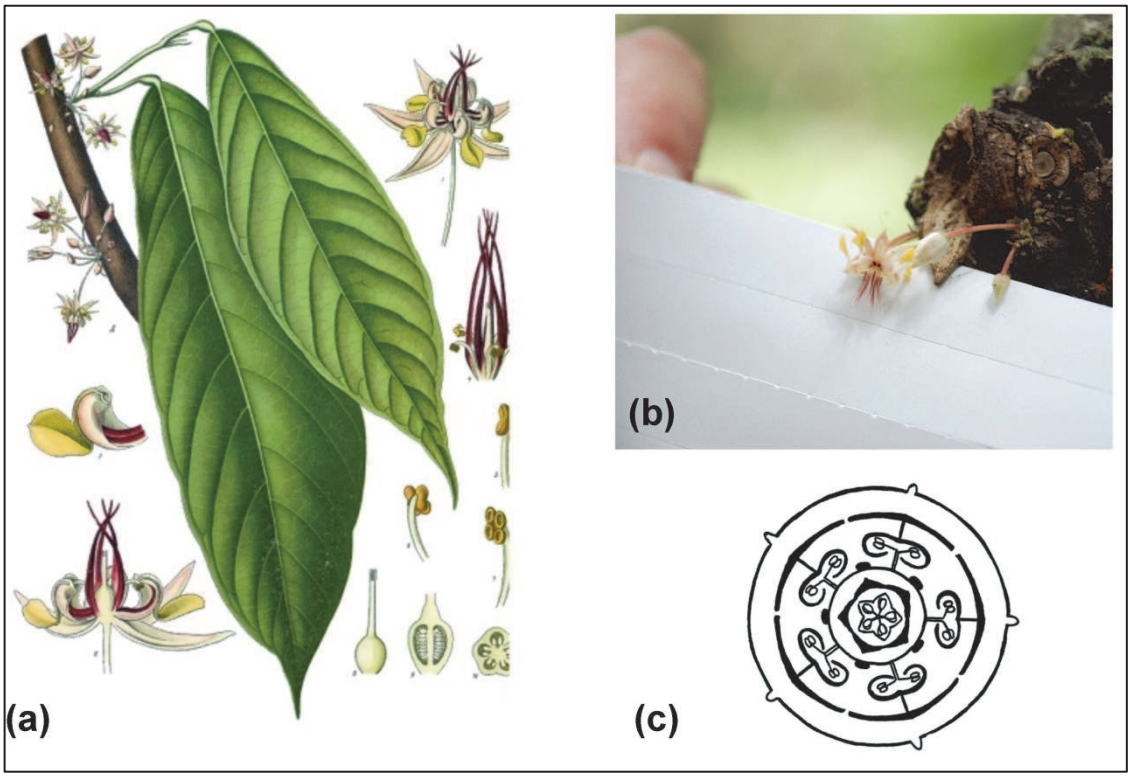


Figure 2 : la feuille et la fleur du cacaoyer. (a) Représentation schématique de la feuille et de la fleur du cacaoyer (d'après Köhler, 1897). (b) Photographie d'une fleur de cacaoyer (X. Argout). (c) Diagramme floral du cacaoyer (P. Lachenaud)

1. Biologie du cacaoyer

1.1 Description botanique

Le cacaoyer, *Theobroma cacao* L., est un arbre tropical appartenant à la famille des *Malvaceae* (Alverson et al., 1999). Le genre *Theobroma* comprend 22 espèces originaires d'Amérique du Sud et d'Amérique centrale. *T. cacao* est une espèce diploïde et a un nombre chromosomique de 2n=2x=20. A l'état naturel, l'arbre peut atteindre 12 à 15 mètres de haut et pousse en groupe le long des rivières des forêts tropicales amazoniennes.

Après germination, l'arbre développe un axe orthotrope dont le bourgeon terminal dégénère après environ 18 mois et 5 bourgeons axillaires donnent alors naissance à une couronne de 5 branches plagiotropes en verticille (Figure 1). Les jeunes feuilles sont très souvent pigmentées et leur couleur peut varier du vert pâle plus ou moins rosé au rouge foncé. Après maturation, elles prennent une couleur vert foncé, sont de forme oblongue et le pétiole est muni à ses extrémités de deux renflements caractéristiques (Figure 2).

Les fleurs sont groupées en inflorescence et sont situées sur un coussinet floral, directement sur le tronc ou les plus grosses branches. Elles sont hermaphrodites, de type 5, de petite taille (diamètre de 0,5 à 1cm) et très nombreuses (certains génotypes peuvent produire plus de 100 000 fleurs par an). L'ovaire est supère et comprend cinq carpelles contenant chacun 6 à 12 ovules. Etamines et staminodes sont soudés à leur base, formant une gaine tubulaire.

La floraison s'effectue par périodes successives et l'intensité de floraison est influencée par l'éclairement et par les régimes thermiques et hydriques (Alvim, 1965; Boyer, 1970). L'origine génétique des clones influence également l'intensité et l'étalement de la floraison (Mossu and Reffye, 1981; Paulin et al., 1983).

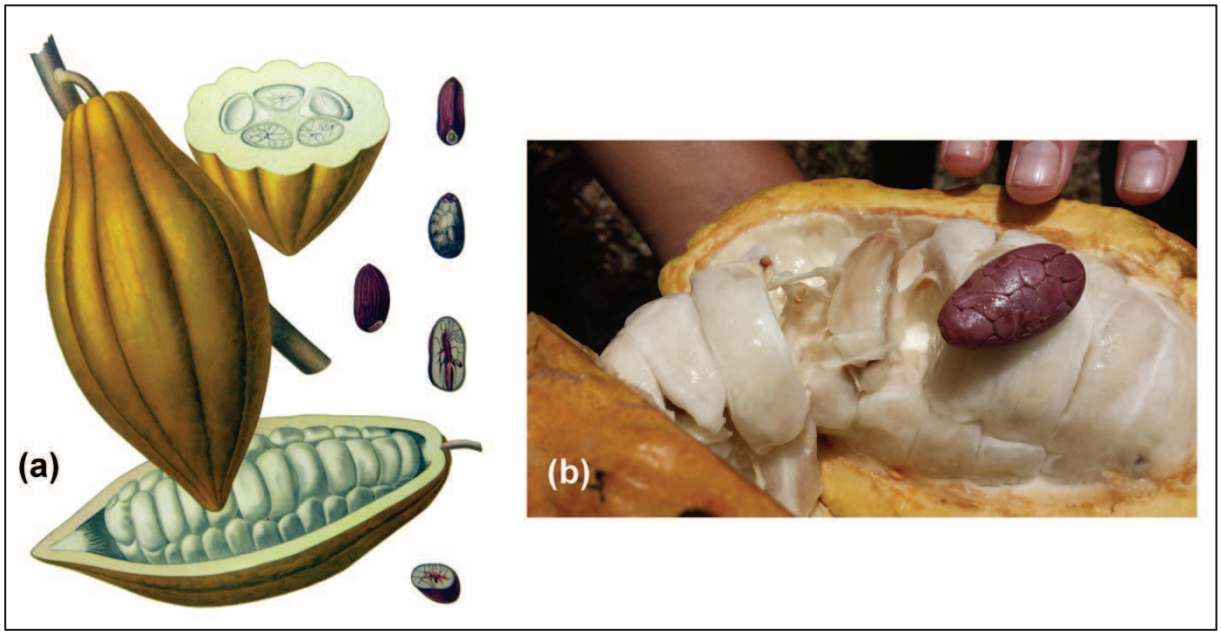


Figure 3 : le fruit du cacaoyer. (a) Représentation schématique d'une cabosse et fève de cacaoyer (d'après Köhler, 1897). (b) Photographie d'une cabosse de cacaoyer ouverte et fève "fraiche" de cacao (X. Argout).

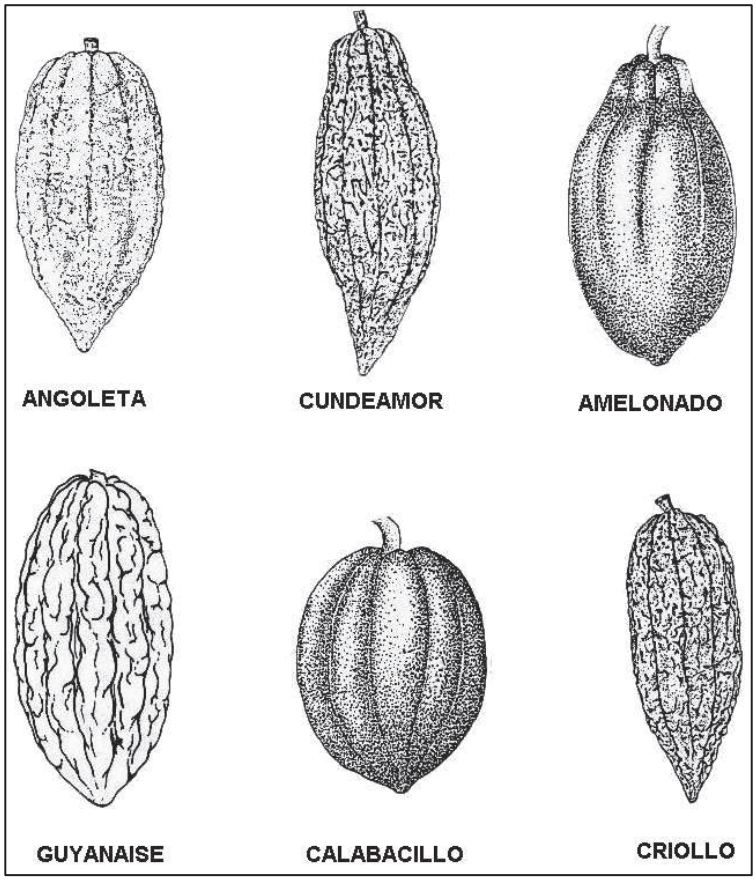


Figure 4 : diversité morphologique des cabosses du cacaoyer (d'après Cuatrecasas, 1964)

La pollinisation est essentiellement entomophile, réalisée majoritairement par des moucherons du genre *Forcypomyia* et des fourmis du genre *Crematogaster*. La nouaison et le développement des fèves dépendent de facteurs liés à la pollinisation, à la nutrition et à la fertilité ovulaire (Falque, 1994; Lachenaud, 1995).

Le cacaoyer est partiellement allogame, l'autogamie étant régie par un système d'auto-incompatibilité chez certains génotypes. Ce système d'auto-incompatibilité a été découvert par Harland en 1925 et confirmé par Pound en 1932. Ce phénomène d'incompatibilité intervient dans l'ovaire où les gamètes sont déversés dans le sac embryonnaire, sans pour autant que la fusion gamétique ne soit effective. Il est contrôlé génétiquement par un locus S (Knight and Rogers, 1955) avec une interaction possible avec 2 autres loci A et B (Cope, 1958) et avec plusieurs allèles en relation de dominance (S1>S2=S3>S4>S5). Récemment, une étude d'association pangénomique a permis de mettre en évidence des associations positives entre marqueurs et tenue du fruit sur le chromosome 4 (da Silva et al., 2016).

Le fruit du cacaoyer est appelé chérelle durant sa période de croissance, puis cabosse lorsqu'il a sa taille définitive (Figure 3). La maturité du fruit varie de 4 à 6 mois en fonction du génotype et des conditions environnementales. La cabosse est indéhiscente et ses caractéristiques morphologiques sont assez variables en fonction des origines génétiques (Figure 4). Elle contient de nombreuses graines ou fèves (de 20 à 60), qui, à maturité du fruit, sont entourées d'une pulpe appelée mucilage. Celui-ci dérive de la testa des graines, est blanc, sucré, acidulé et aqueux.

La graine de cacaoyer a la forme d'une amande plus ou moins dodue et ne possède pas d'albumen. Elle comprend, de l'extérieur vers l'intérieur, une coque mince, résistante et rosée provenant du développement de l'ovule, une fine membrane translucide, vestige de l'endosperme et l'embryon dont deux cotylédons occupent quasiment tout le volume de la graine. Les cotylédons sont fortement plissés et leur couleur varie du blanc au violet foncé en passant par toutes les teintes intermédiaires en fonction des génotypes. Ils sont reliés à leur base à une radicule et à une gemmule, insérées entre les deux cotylédons, formant le "germe" de la fève de cacao.

1.2 Domestication et diversité génétique

Avant l'avènement des marqueurs génétiques moléculaires, 2 groupes morphogéographiques résultants des différences observées entre les cacaoyers d'Amérique du Sud et les cacaoyers d'Amérique centrale ont été proposés par Cheesman (1944) : Criollo et Forastero. Ce même auteur a émis l'hypothèse d'un troisième groupe, dénommé Trinitario et provenant de l'hybridation naturelle des cacaoyers Criollo et Forastero réalisée à Trinidad et Tobago au XVIIIème siècle. Le centre d'origine géographique de l'espèce *Theobroma cacao* L. a quant à lui été pendant longtemps controversé.

D'un côté, certains auteurs, dont Cuatrecasas (1964), ont émis l'hypothèse d'une origine simultanée dans les 2 parties du continent américain, séparée par l'isthme de Panama. Les 2 populations auraient évolué indépendamment et correspondraient à 2 sous espèces, *T. cacao* ssp *cacao* et *T. cacao* ssp *sphaerocarpum*, correspondant aux groupes Criollo et Forastero.

D'un autre côté, d'autres auteurs, dont Van Hall (1914) et Cheesman (1944) ont suggéré que le cacaoyer avait été introduit en Amérique Centrale, son centre d'origine serait unique et situé dans le bassin du fleuve Orinoco et Amazone. Bien que l'archéologie et la linguistique aient identifié le premier centre de domestication et de culture du cacaoyer en Amérique centrale, notamment chez les populations Maya, aucune population de cacaoyers réellement sauvage n'était présente dans la région. Ainsi l'homme, par l'intermédiaire des populations indiennes, aurait diffusé le cacaoyer depuis l'Amérique du Sud vers l'Amérique centrale et le sud du Mexique.

Le débat entre ces 2 hypothèses est resté ouvert jusqu'à ce que Motamayor et al. (2002) identifient par des marqueurs génétiques des représentants des Criollo anciens et démontrent que les cacaoyers Criollo de génotype ancien d'Amérique centrale et d'Amérique du sud présentent une diversité génétique extrêmement faible et sont très proches des autres génotypes originaires d'Equateur et de Colombie. Ces résultats suggèrent donc que *T. cacao* est bien originaire d'Amérique du Sud et qu'un nombre limité de cacaoyers d'Amérique du Sud aurait été introduit par l'homme en Amérique Centrale.

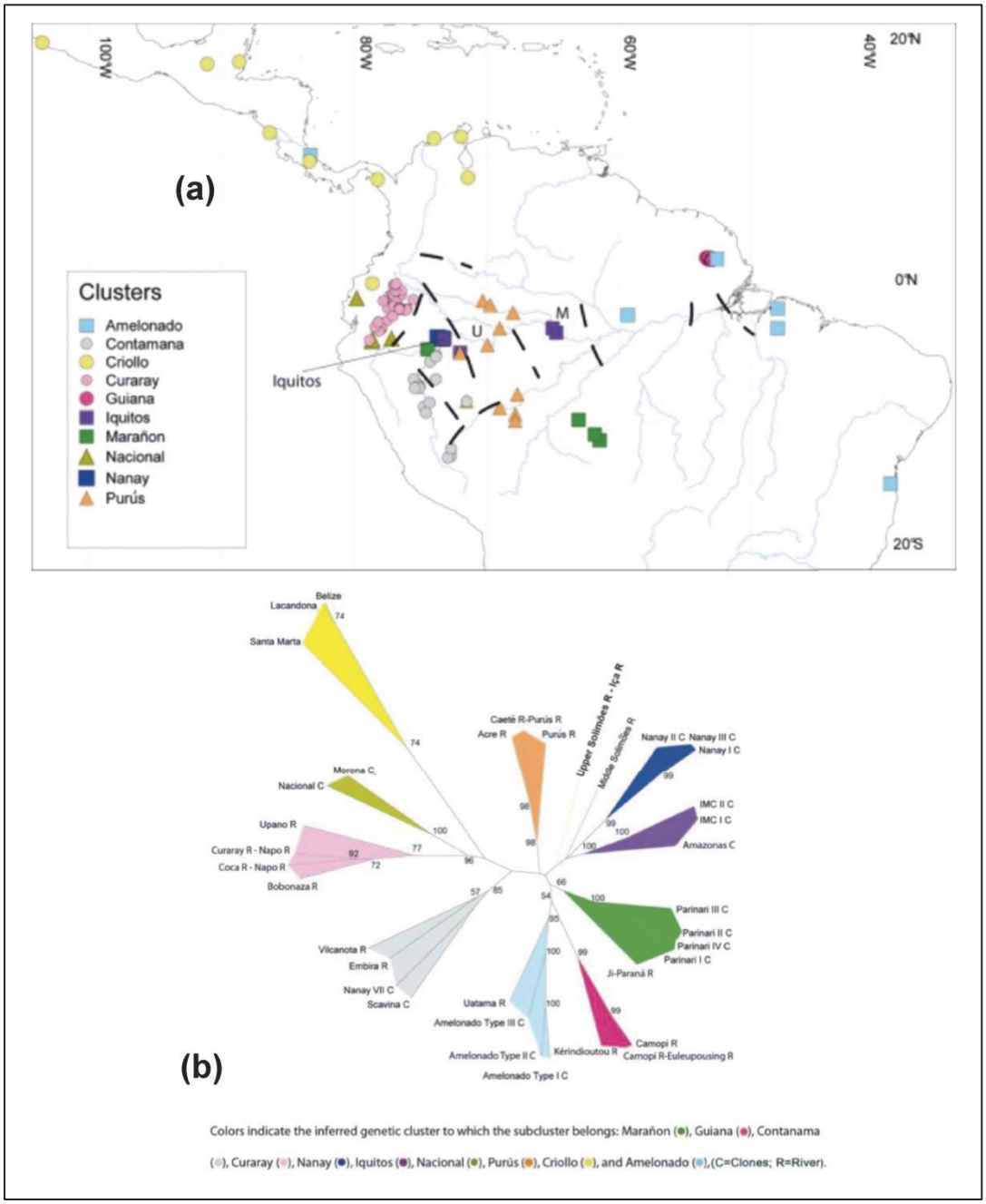


Figure 5 : diversité génétique du cacaoyer (Motamayor et al., 2008). (a) Localisation des individus analysés; la couleur indique le cluster génétique d'appartenance; les traits noirs représentent les paléo-barrières géographiques. (b) Arbre phylogénétique (Neighbor Joining) issu du logiciel Structure.

Par ailleurs le groupe génétique correspondant au Forastero ne semblait pas être homogène. En effet, les premières études de diversité réalisées à l'aide de marqueurs izo-enzymatiques et moléculaires ont mis en évidence une large diversité parmi les Forastero haut-amazoniens et Forastero bas-amazoniens (Lanaud, 1984; Laurent, 1993; N'Goran et al., 1994; Solorzano et al., 2012).

Motamayor et al. (2008) présenteront la classification de référence utilisée actuellement (Figure 5). Cette classification a été réalisée par génotypage de 1241 génotypes à l'aide de 106 marqueurs microsatellites couvrant une très large diversité géographique. Cette étude a conduit à la structuration de l'espèce en 10 groupes génétiques, supportés par des évidences de paléo-barrières géographiques (rivières, crêtes montagneuses). Ces 10 groupes sont : Marañon, Curaray, Criollo, Iquitos, Nanay, Contamana, Amelonado, Purús, Nacional et Guiana.

Figure 6 : Technologie du cacao. (a) Scène d'écabossage du cacao (source photographique, journal "la Tribune de Diego", Madagascar). (b) Bac de fermentation de type coffres en cascade (photographie X. Argout, plantation Millot, Madagascar). (c) Fèves de cacao en cours de fermentation (photographie X. Argout, plantation Millot, Madagascar). (d) Séchoir solaire de type "autobus" (photographie X. Argout, plantation Millot, Madagascar)

2. Technologie du cacao

Les fèves de cacao fermentées et séchées, ou cacao marchand, constituent la matière première de l'industrie du chocolat. Après la cueillette, les fèves fraiches sont extraites des cabosses après ouverture ou écabossage et subissent des opérations de fermentation et de séchage pour conduire au cacao marchand (Figure 6).

La fermentation est une étape primordiale, elle permet l'élimination de la pulpe mucilagineuse entourant les fèves, supprime le pouvoir germinatif des fèves et provoque les réactions chimiques nécessaires à l'apparition des précurseurs d'arômes. Le temps de fermentation dépend de l'origine génétique des fèves et peut varier de 3 à 8 jours. Durant cette étape se succèdent 2 types de fermentation, une première phase anaérobie de fermentation alcoolique sous l'action de levures qui transforment les sucres contenus dans la pulpe en alcool et une fermentation aérobie après brassage au cours de laquelle les bactéries transforment l'alcool en acide acétique. Les réactions de fermentation sont exothermiques et la chaleur générée (autour de 50°C) altère les tissus de la coque et laisse pénétrer dans la fève les produits de dégradation de la pulpe, conduisant aux réactions d'hydrolyse et à la formation des précurseurs aromatiques.

Une fois fermentées, les fèves sont séchées, le but étant de faire baisser la teneur en eau des fèves fermentées de 60% à 7%. Cela confère au cacao de bonnes conditions de conservation pour le stockage et le transport. Ce séchage peut être réalisé de manière naturelle (solaire) ou artificielle (four).

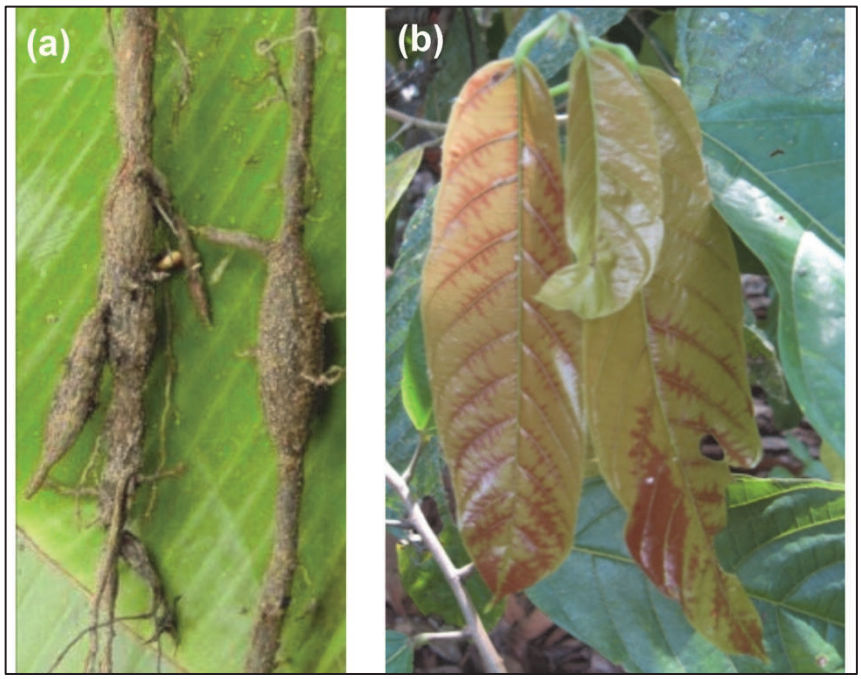


Figure 7 : Symptomatologie du Cacao Swollen-Shoot Virus (CSSV). (a) Gonflement des racines (photographie Koffié Kouakou, CNRA). (b) Nervures des feuilles de couleur rouge

3. Les maladies du cacaoyer et ravageurs

On estime que la perte due aux maladies et ravageurs représente entre 30 et 40% de la production mondiale (source International Cocoa Organization http://www.icco.org/about-cocoa/pest-a-diseases.html). Parmi les nombreuses maladies et ravageurs affectant la production de cacao marchand, voici un descriptif non exhaustif des principaux problèmes sanitaires des cacaoyers.

3.1 Maladie virale du Cacao Swollen-Shoot Virus

Une maladie impliquant un virus, Le Cacao Swollen-Shoot Virus (CSSV), touche particulièrement l'Afrique. Cette maladie provoque une perte de rendement les 2 premières années puis peut conduire à la mort de l'arbre. Le CSSV est un virus du genre Badnavirus et est responsable de la maladie virale la plus étendue en Afrique. Les symptômes de la maladie du CSSV sont très variables et dépendent de la souche virale et du stade d'infection. La plupart des symptômes caractéristiques sont l'apparition de liserés et de nervures de couleur rouge sur les feuilles et des gonflements des tiges et racines (Figure 7). On estime que le virus est responsable d'au moins 15% de perte de production de cacao au niveau mondial (Kokutse, 2008).

3.2 Maladies fongiques

Les maladies fongiques constituent la contrainte majeure pour la cacaoculture mondiale. Elles ont une distribution mondiale contrairement aux autres types de maladies qui touchent un continent particulier. Les pertes de rendement peuvent être très importantes (parfois plus de 80%) et les maladies peuvent conduire à la mort de l'arbre.

Figure 8 : Symptomatologie de la maladie du balai de sorcière. Balai vert terminal (photographie X. Argout, Equateur)

Figure 9 : Symptomatologie de la moniliose. Cabosse atteinte de Moniliose, "frosty pod" (photographie Philippe Lachenaud © Cirad)

Figure 10 : Symptomatologie des maladies provoquées par le genre *Phytophthora*. Pourriture brune sur cabosse, Phytophthora palmivoraCabosse (photographie Philippe Lachenaud © Cirad)

3.2.1 Maladies provoquées par les champignons du genre *Moniliophthora*

Les champignons Basidiomycètes du genre *Moniliophthora* sont responsables de maladies qui touchent les pays d'Amérique du Sud et d'Amérique Centrale.

3.2.1.1 La maladie du balai de sorcière

L'agent causal, *Moniliophthora perniciosa*, induit de nombreux symptômes sur les parties végétatives, les fleurs, les coussinets floraux et les cabosses. La manifestation la plus caractéristique de la maladie est l'hypertrophie des méristèmes végétatifs infectés qui sont alors appelés "balais" (Figure 8). La perte de production dans les régions d'infection est très importante (50-90%) et l'arrivée de la maladie dans la région de Bahia au Brésil en 1989 a fait passer la production du pays (alors 3ème producteur mondiale) de 347 000 tonnes à 141 000 tonnes en 2000 (Meinhardt et al., 2008).

3.2.1.2 La moniliose

Le champignon Basidiomycète responsable de la maladie est *Moniliophthora roreri*. En condition naturelle, la Moniliose affecte seulement les cabosses. La manifestation la plus caractéristique est la formation d'un stroma fongique blanc recouvrant le secteur infecté de la cabosse ("frosty pod"), avec une nécrose et une agglutination des fèves qu'elle contient (Figure 9). A un niveau mondial, les pertes de production engendrée par la maladie sont relativement faibles. En revanche, à un niveau local, les pertes peuvent être extrêmement importantes, allant jusqu'à 100% de la production. Dans la région de Santander en Colombie, on estime la perte annuelle de production due à la maladie à 40% (Phillips-Mora, 2003).

3.2.2 Maladies provoquées par les champignons du genre *Phytophthora*

Les champignons du genre *Phytophthora* appartiennent à la classe des Oomycètes. Plusieurs espèces de *Phytophthora* sont des agents causaux de maladies chez le cacaoyer. Les plus importantes sont *P. palmivora*, *P. megakarya*, *P. capsici* et *P. citrophthora*. Les *Phytophthora* spp. peuvent attaquer toutes les parties de l'arbre, cependant les symptômes principaux sont la pourriture des cabosses (Figure 10), le chancre de la tige et la rouille des plantules.

Figure 11 : Adulte et larves de Sahlbergella singularis, miride du cacaoyer. (photographie Régis Babin © Cirad)

Leur distribution géographique est mondiale et la majorité des pays producteurs est confronté à la maladie (End et al., 2014). La perte annuelle mondiale causée par *Phytophthora* spp est estimée à 44% (Appiah et al., 2003).

3.2.3 Vascular Streak Dieback (VSD)

La maladie est causée par le champignon de la classe des Basidiomycètes *Ceratobasidium theobroma*. Les syndromes les plus caractéristiques sont une chlorose des feuilles et une destruction du système vasculaire conduisant à une nécrose des branches et à la mort de l'arbre si le tronc est infecté. La maladie a été observée dans la plupart des zones de culture du cacaoyer d'Asie du Sud Est et de Mélanésie. Comme la Moniliose, le VSD a un impact modéré sur les pertes annuelles au niveau mondial (environ 30 000 tonnes) mais localement (Indonésie, Malaisie) cause des pertes extrêmement fortes, obligeant certains producteurs à délaisser le cacao pour le maïs, le clou de girofle ou le palmier à huile (Bailey and Meinhardt, 2016).

3.2.4 Flétrissement du à *Ceratocystis* ou Mal de machete

La maladie est causée par le champignon de la classe des ascomycètes *Ceratocystis cacaofunesta*. C'est un pathogène important du cacaoyer causant le flétrissement, la pourriture du système racinaire et la mort des arbres en s'attaquant aux cellules du xylème. La maladie a uniquement été décrite en Amérique du Sud et Amérique Centrale mais elle représente une menace très forte pour les pays producteurs de cacao (Bailey and Meinhardt, 2016).

3.3 Insectes ravageurs

3.3.1 Les Mirides

Ces insectes suceurs de sève de la famille des *Miridae* sont des ravageurs qui touchent toutes les zones géographiques de la cacaoculture. Ils percent la surface des tiges et des cabosses provoquant des lésions nécrotiques qui favorisent le développement des maladies fongiques (Figure 11). On estime que l'impact des Mirides sur la production de cacao peut réduire le rendement de 75% (source : ICCO).

3.3.2 Foreurs de cabosses ou maladie du "Cocoa Pod Borer"

L'agent causal de la maladie, *Conopomorpha cramerella* est un lépidoptère de la famille des Gracillaridae. Les cabosses immatures qui sont infestées présentent une agrégation des fèves en une masse solide impossible à extraire. Cette maladie est largement distribuée à travers l'Asie du Sud-Est. En 2000 en Indonésie, la maladie était étendue sur 60 000 ha, provoquant une perte estimée à 40 millions de dollars US (source : ICCO).

3.4 Nématodes parasites

Plus de 25 genres de nématodes ont été décrits comme affectant le cacaoyer (Campos and Villain, 2005). Cependant les nématodes du genre *Meloidogyne* sont les plus virulents, provoquant le dépérissement de la tige et le flétrissement des feuilles dus à la destruction du système racinaire. Les pertes de production estimées varient de 15 à 30 % mais peuvent atteindre 40 à 60 % (Fademi et al., 2006). L'infestation par les nématodes sur le cacaoyer est répertoriée dans la plupart des régions productrices de cacao du monde

3.5 Vertébrés

Les mammifères (*ie.* écureuils, rats, singes), les oiseaux (*ie*. pic vert) et certaines espèces de singes peuvent conduire à des pertes de rendement significatives (Ploetz, 2007). Cependant la problématique est généralement très localisée.

3.6 Moyen de lutte

Le maintien d'une plantation en bon état nécessite la plupart du temps, en dehors des soins culturaux habituels, des traitements phytosanitaires pour lutter contre les dégâts occasionnés par les parasites ou maladies. Outre les surcoûts engendrés par l'utilisation des fongicides et insecticides, leur impact sur la santé humaine et l'environnement est extrêmement préoccupant.

La nature et l'importance des traitements sont très variables selon les pays et les maladies. Au Ghana par exemple, où la culture du cacao représente plus de la moitié des revenus du secteur agricole, la fréquence de l'application des pesticides par les producteurs de cacao varie de 1 à 9 par saison (Denkyirah et al., 2016). Dans ce contexte, les compréhensions des processus biologiques et des caractères génétiques impliqués dans les mécanismes de résistance aux maladies sont des priorités majeures pour la recherche cacaoyère.

3.7 Facteurs génétiques

Alors qu'environ un tiers de la production mondiale de cacao est perdue chaque année à cause des maladies et des insectes, seuls 30% des cacaoyers cultivés actuellement dans le monde sont issus de variétés améliorées (Gutiérrez et al., 2016). Des sources de résistance naturelle ont pourtant été identifiées dans des fonds génétiques variés. Pour les maladies les plus graves, telles que la maladie du balai de sorcière, la moniliose ou la pourriture brune des cabosses, les pathogènes ont le plus souvent co-évolué avec *T. cacao* dans les aires d'origine du cacaoyer (Leppik, 1970). Par conséquent, les génotypes originaires de Haute-Amazonie présentent une fréquence plus importante des caractères de résistance associés à ces maladies que les génotypes Bas-amazonien et ceux cultivés en Amérique centrale (Pires et al., 2000).

Les programmes d'amélioration pour la résistance au balai de sorcière ont débuté dans les années 1950 à partir des cultivars Scavina (SCA) originaire du Pérou (principalement SCA6 et SCA12), collectés lors d'une expédition de Pound en 1938 et utilisés comme parents donneurs de gènes de résistance. Les parents Scavina sont présents dans le fond génétique de plusieurs lignées d'amélioration et d'hybrides qui ont été créés dans de nombreuses zones géographiques (Lopes et al., 2011). Cependant, la tolérance à la maladie dans les populations impliquant les parents Scavina commence à s'éroder dans le bassin amazonien au Brésil, en Equateur et au Pérou, probablement à cause d'une évolution du pathogène. Selon la base de données internationale ICGD (International Cocoa Germplasm Database, http://www.icgd.reading.ac.uk/), plus de 400 clones ont été identifiés comme présentant une résistance ou une tolérance à la maladie du balai de sorcière. Des QTLs d'association à la maladie ont été découvert, à partir d'une population F2 obtenue à partir d'un croisement issu de SCA6 (résistant) et ICS1 (sensible).

Un QTL majeur est situé sur le groupe de liaison LG9 et un QTL mineur sur le groupe de liaison LG1 (Brown et al., 2005; Faleiro et al., 2006; Feitosa Jucá Santos et al.; Queiroz et al., 2003). Cette résistance des clones Scavina semble être contrôlée par un nombre limité de gènes à effet fort (Brown et al., 2005).

Des fonds génétiques de résistance à la moniliose ont été identifiés chez des cacaoyers sauvages prospectés en Haute Amazonie (UF 273 er UF 712), dans des descendances hybrides (ex. ICS 95) et dans des collections nationales issues de prospections locales (ex. Colombie, FEC 2). Dans les génotypes qu'ils ont étudiés, Phillips-Mora et Castillo (1999) ont montré que la résistance à *M. roreri* apparaissait plutôt comme un caractère récessif. Cinq QTLs de résistance ont par la suite été identifiés sur les groupe de liaison LG2, LG7 et LG8 (Brown et al., 2007; Cervantes-Martinez et al., 2006), indiquant un déterminisme polygénique.

Les programmes d'amélioration pour la résistance à la pourriture brune des cabosses ont débuté il y a de très nombreuses années. L'héritabilité de la résistance a été estimée à 0.33 (sens stricte) et 0.51 (sens large) (Iwaro et al., 2005). De très nombreuses sources de résistance ont été trouvées dans des fonds génétiques variés. Thevenin et al. (2012) ont par exemple caractérisé la résistance au pathogène *Phythophthora palmivora* de 186 accessions du groupe génétique Guiana. Cinquante neuf clones de ce groupe génétique se sont révélés résistants et la résistance à l'espèce *P. megakarya* a également été établie dans du matériel végétal sauvage collecté en Guyane française (Paulin et al., 2008). Le cluster génétique Guiana semble donc être une bonne source de résistance à la pourriture brune. De nombreuses études ont également permis de localiser des QTLs de résistance à la pourriture brune sur plusieurs cartes génétiques. Ainsi, jusqu'à présent, 65 QTLs de résistance ont été identifiés sur les groupe de liaison LG1, LG2, LG4, LG5, LG8 et LG10 (Brown et al., 2007; Clement et al., 2003; Lanaud et al., 2009). Parmi ces QTLs, seuls 13 sont consensus et sont localisés sur les groupes de liaison LG1, LG2, LG4 et LG5.

Quatre clones ont été identifiés comme résistant à *Ceratocystis* par une évaluation réalisée par Soria et Salazar (1965), SPA 9, IMC 67, Pound 12, et PA 121. Gardella et al. (1982), ont montré que la résistance à *Ceratocystis* des clones SPA 9 et IMC 67 était contrôlée par un locus majeur dominant. Plus récemment, Santos et al. (2012) ont trouvé 2 QTLs de résistance sur les groupes de liaison LG3 et LG9 à partir d'une population F2 issue de l'autofécondation du clone TSH 516.

Par ailleurs, Marelli et al. (2014), ont trouvé une région de 1,2 Mb sur le chromosome 6 impliquée dans la résistance à *Ceratocystis*, en étudiant une descendance réalisée à partir du croisement entre TSH 1188 (résistant) et CCN 51 (sensible).

De nombreuses sources de résistance aux maladies du cacaoyer ont donc été identifiées dans des fonds génétiques variés. La recherche d'une résistance durable cumulant les différents gènes de résistance est l'un des défis majeurs des programmes d'amélioration génétique du cacaoyer.

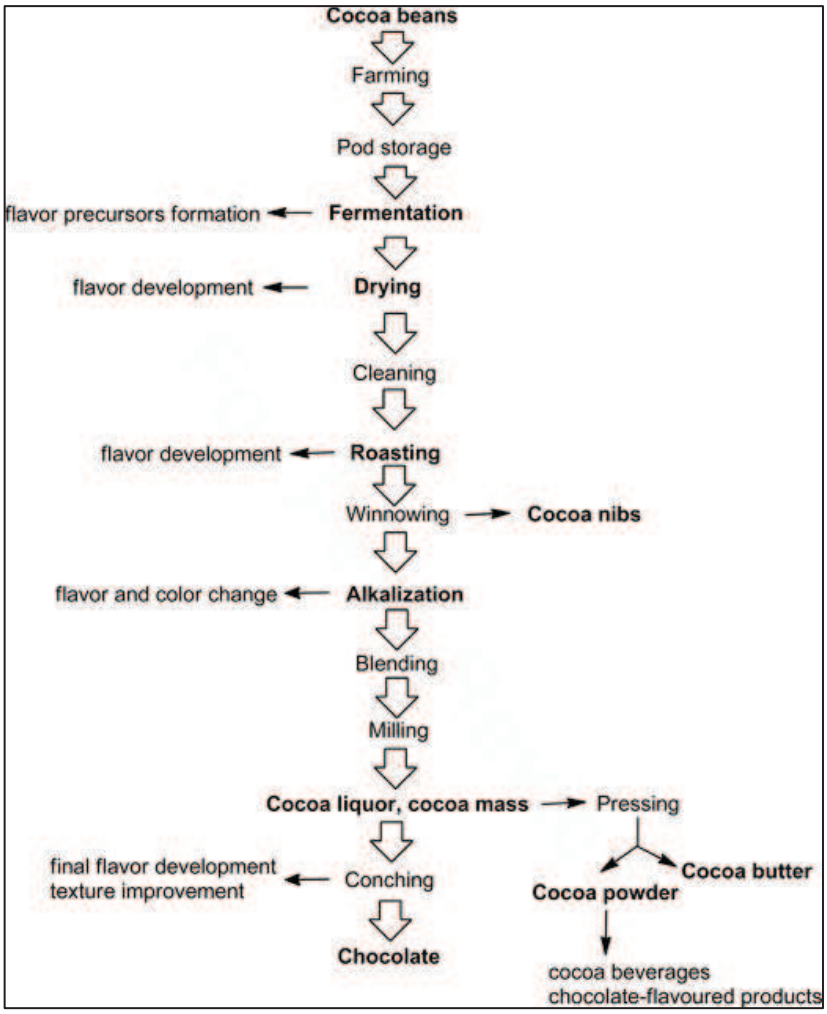


Figure 12 : Mise en place des composés aromatiques dans la chaine de transformation des fèves de cacao (d'après Aprotosoaie et al., 2015).

4. Cacao et qualité

Comme le décrit Michel Barel dans son livre relatif à la qualité du cacao (Barel, 2013), la notion de qualité du cacao n'est pas la même selon le point de vue des différents acteurs qui constituent la filière cacao; par exemple le producteur va plutôt s'intéresser à la productivité, l'exportateur à la présence de moisissure, le transformateur à la stabilité et à la qualité aromatique du produit et le consommateur à la qualité organoleptique. Dans ce chapitre, nous nous intéresserons à une partie des attributs de la qualité des fèves de cacao liés à leur origine génétique.

Depuis quelques années, les consommateurs et industriels du cacao s'intéressent de plus en plus aux qualités organoleptiques et nutritionnelles du cacao. Pour répondre à cette demande, la compréhension des bases génétiques des caractères de qualité est un défi important pour la recherche cacaoyère.

4.1 Composés aromatiques

Le cacao torréfié est l'un des produits alimentaires les plus aromatiques, avec près de 400 composés volatiles identifiés (Bonvehí, 2005). Il a été montré depuis de nombreuses années, que les traitements post-récolte jouaient un rôle primordial dans la formation des précurseurs d'arômes (Figure 12). Durant l'étape de fermentation, les fèves fraiches de cacao subissent de profondes transformations : 1) les sucres contenus dans la pulpe sont rapidement métabolisés en composés organiques acides volatiles et non volatiles; 2) les protéines sont dégradées en peptides et acides aminés libres; 3) les polyphénols sont oxydés en composés non solubles; 4) les glycosides, principalement ceux impliqués dans la voie de biosynthèse des anthocyanes sont hydrolysés (Clapperton, 1994; Kirchhoff et al., 1989). De plus, les levures et bactéries impliquées dans la fermentation contribuent également à la formation des arômes et précurseurs d'arômes (Schwan and Wheals, 2004). La torréfaction des fèves permet la libération des arômes définitifs du cacao. Durant cette étape, les réactions de Maillard produisent les arômes en combinant la réduction des sucres et les peptides et acides aminés libres formés durant l'étape de fermentation (Schnermann and Schieberle, 1997).

Terpenoids compounds	Odor quality	Sensory perception	Reference
Geraniol	Floral, rose, fruity	Floral, fruity	Bonvechi (2005)
Geranyl acetate	Rose, lavender	Floral	Bonvechi (2005)
α-Terpenyl formate	Herbaceous, citrus	Herbal, fruity	Bonvechi (2005)
Linalool (cis-pyranoid)	Floral, green	Floral, herbal	Bonvechi (2005)
Linalool (trans-pyranoid)	Floral	Floral	Bonvechi (2005)
Linalool oxide (cis-furanoid)	Nutty	Nutty	Bonvechi (2005)
Linalool oxide (trans-furanoid)	Floral, citrus	Fruity, floral	Bonvechi (2005)

Table 1. Principaux composés terpénoïdes identifiés dans les composés aromatiques du cacao (d'après Aprotosoaie et al., 2015).

Par ailleurs, le marché distingue deux types de cacao marchand : un cacao de type "standard" (environ 95% de la production mondiale) et un cacao "fin" ou "aromatique" (5% de la production), principalement issu des variétés Criollo et Nacional. Ces cacaos fins contiennent des arômes particuliers tels que des notes florales, fruitées, de noisette ou épicées. Clapperton (1994) a montré que l'origine génétique influence grandement la composition aromatique du cacao et ce indépendamment des traitements post-récolte.

Dans ces variétés aromatiques, la classe des terpènes ou terpenoïdes semble jouer un rôle clé (Table 1). Il a par exemple été trouvé une grande quantité de linalool, un monoterpène, dans des variétés Nacional équatoriennes et des clones Criollo vénézuéliens (Ziegleder, 1990). Combinés à d'autres composés volatiles, le linalool pourrait être responsable de l'arôme typique floral de ces cacaos.

Très peu de travaux sur le cacaoyer ont été réalisés pour étudier le déterminisme génétique des composés aromatiques. Lanaud et al. (2003) à partir de la descendance issue du croisement UPA402 (haut amazonien) x UF676 (Trinitario) ont identifié une région d'environ 20cM sur le groupe de liaison LG1 qui rassemble des QTLs impliqués dans l'arôme floral et les arômes fruités. Deux autres QTLs significatifs pour l'arôme floral ont également été identifiés sur les groupes de liaison LG4 et LG9 et un QTL pour l'arôme fruité sur le groupe de liaison LG7.

Le décryptage des voies de biosynthèse conduisant à ces composés aromatiques et l'étude de la régulation des gènes impliqués sont des objectifs clés pour la compréhension du déterminisme génétique des caractères de qualité.

	South America %	North and Central America %	Africa %	Asia %
POP	19.0	18.6	18.4	18.6
POSt	38.0	38.9	39.1	40.0
StOSt	26.0	26.9	28.2	30.8
AOSt	0.5	0.6	0.6	0.8
POO	3.4	2.7	2.2	1.2
StOO	5.7	5.3	4.7	2.9
PLiP	1.1	1.0	1.0	0.8
PLiSt	3.5	3.3	3.2	2.9
StLiSt	2.8	2.7	2.5	2.2

Table 2 : Composition des triglycérides contenus dans le beurre de cacao issu de différentes origines (d'après Chaiseri et Dimick, 1989. Abréviations : P=acide palmitique; O= acide oléique; St=acide stéarique; A=acide arachidique; Li= acide linoléique

4.2 Acides gras

La graine de cacao est très riche en matière grasse. La teneur en beurre de cacao des fèves séchées est généralement supérieure à 50%. Ce beurre de cacao est principalement composé de triglycérides (triacylglycérols) c'est à dire d'esters de glycérol et d'acides gras. En comparaison avec d'autres graisses végétales, sa composition en acide gras est relativement simple (Chaiseri and Dimick, 1989). En effet, les acides gras sont en majorité l'acide palmitique (P), l'acide stéarique (St), et l'acide oléique (O). L'acide linoléique vient compléter ce trio mais en plus faible quantité. De cette simple composition en acides gras découlent une composition simple en triglycérides, avec 3 triglycérides prédominants, POP, POSt et StOSt. Cette composition riche en acide stéarique confère au beurre de cacao un point de fusion plutôt élevé (34-38°C). Par ailleurs ce profil unique en acide gras donne au beurre de cacao une bonne capacité de libération des arômes dans le chocolat.

Chaiseri and Dimick (1989) ont montré une variation dans la composition des beurres de cacao en fonction de leur origine (Table 2) indiquant un certain déterminisme génétique dans la biosynthèse des triglycérides.

Lanaud et al. (2003) ont mis en évidence un QTL significatif lié à la quantité d'acide gras du beurre de cacao sur le groupe de liaison LG9. Araújo et al. (2009), à partir d'une descendance F2 dérivant du croisement ICS1 x SCA6 ont confirmé la présence de ce QTL lié à la quantité d'acide gras du beurre de cacao, il expliquerait 51% de la variation phénotypique. Ce même auteur a également mis en évidence 2 autres QTLs liés à la dureté du beurre sur les groupes de liaison LG7 et LG9, expliquant 28,8% de la variation.

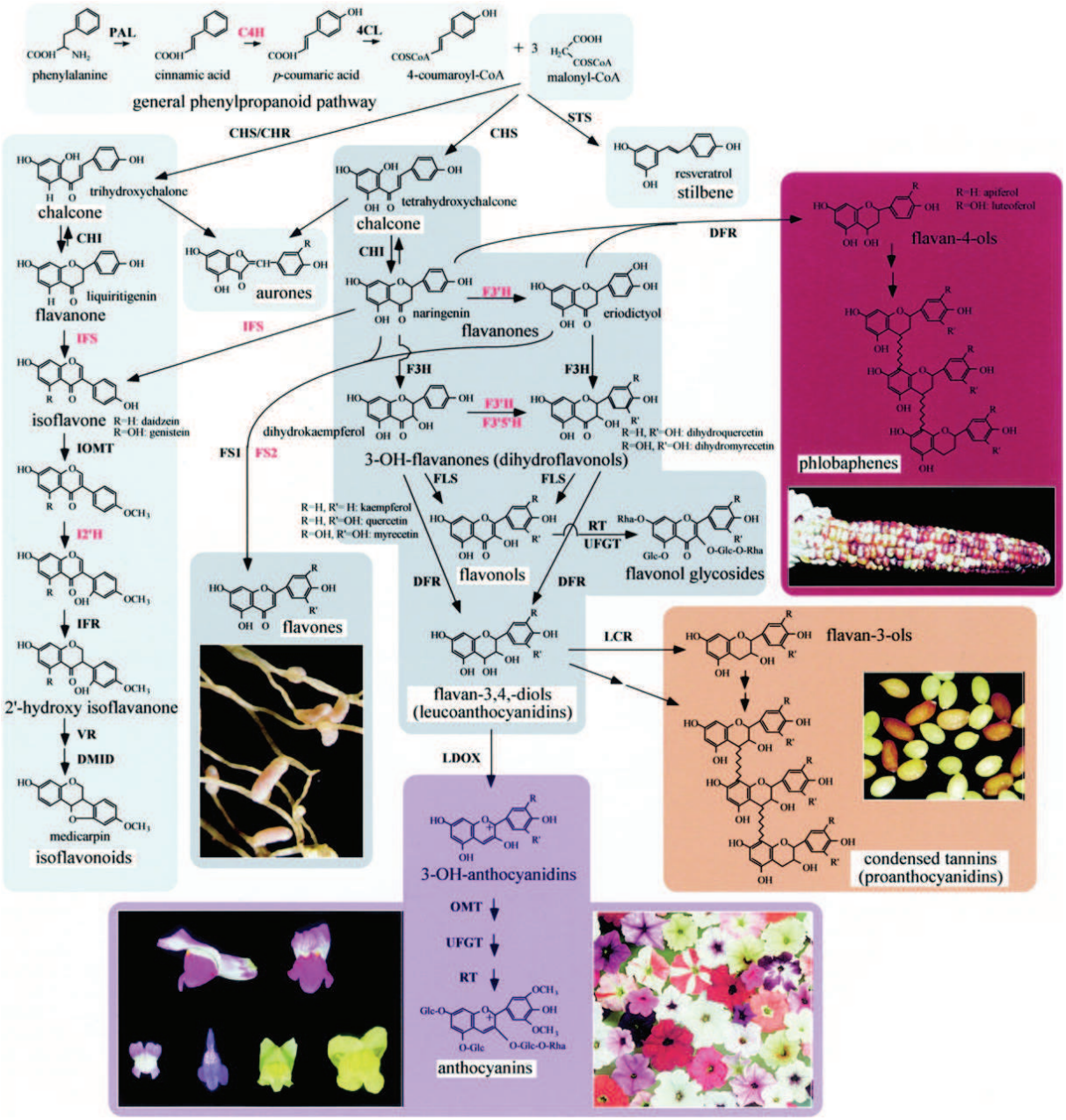


Figure 13 : Représentation schématique de la voie de biosynthèse des flavonoïdes (d'après Brenda Winkel-Shirley, 2001).

Les 9 sous-groupes majeurs des flavonoïdes sont indiqués par des rectangles : chalcones, aurones, isoflavonoides, flavones, flavonols, et flavandiols (rectangles gris), et anthocyanines, proanthocyanidines et pigments phlobaphene (rectangles colorés). Les hydrolases P450 sont indiquées en rouge sur la figure. Les photographies illustrent les trois classes majeures de pigments rencontrées chez le Muflier, *Arabidopsis thaliana*, le Maïs et le Pétunia. Est également illustré la nodulation du cortex racinaire par les bactéries du genre *Rhizobium* chez *Melilotus alba* et impliquant les flavones et isoflavones.

Abréviations : cinnamate-4-hydroxylase (C4H), chalcone isomerase (CHI), chalcone reductase (CHR), chalcone synthase (CHS), 4-coumaroyl:CoA-ligase (4CL), dihydroflavonol 4-reductase (DFR), 7,2'-dihydroxy, 4'-methoxyisoflavanol dehydratase (DMID), flavanone 3-hydroxylase (F3H), flavone synthase (FSI and FSII), flavonoid 3' hydroxylase (F3'H) or flavonoid 3'5' hydroxylase (F3'5'H), isoflavone*O*-methyltransferase (IOMT), isoflavone reductase (IFR), isoflavone 2'-hydroxylase (I2'H), isoflavone synthase (IFS), leucoanthocyanidin dioxygenase (LDOX), leucoanthocyanidin reductase (LCR), *O*-methyltransferase (OMT), Phe ammonia-lyase (PAL), rhamnosyl transferase (RT), stilbene synthase (STS), UDPG-flavonoid glucosyl transferase (UFGT), vestitone reductase (VR).

4.3 Polyphénols

Les polyphénols sont des composés issus du métabolisme secondaire de nombreuses plantes et jouent des rôles variés. Ils sont par exemple impliqués dans le développement cellulaire, les mécanismes de défense ou les voies de signalisation. Leur intérêt a été souligné ces dernières années en médecine préventive de par leurs propriétés antioxydantes (Rimbach et al., 2009; Spencer, 2009).

Les fèves de cacao sont exceptionnellement riches en polyphénols et leur composition détaillée a été révélée depuis de nombreuses années (Bastide, 1987; Duthie, 1938; Kim and Keeney, 1984; Zumbé, 1998). Les flavonoïdes sont une sous-classe de polyphénols particulièrement bien représentés dans les fèves de cacao. Ces composés phénoliques sont stockés dans les cellules pigmentaires des cotylédons. Ils interviennent de façon directe dans les processus d'hydrolyse, d'oxydation, de polymérisation et de tannage des protéines au cours de la fermentation. Ils influencent également les propriétés organoleptiques puisque certains dérivés flavoniques ont des propriétés d'amertume et d'astringence.

On distingue trois classes principales de flavonoïdes dans les fèves de cacao : les catéchines ou flavonols (37%), les anthocyanines (4%) et les proanthocyanidines (58%). La quantité totale de flavonoïdes dans la partie non grasse des fèves fraiches est de 15 à 20%. Cette quantité est variable en fonction de l'origine génétique du cacao. Par exemple, les fèves de Criollo, de couleur blanche, ne contiennent qu'environ 2/3 de la quantité de composés phénoliques observée chez les autres groupes génétiques à fèves violettes, et les anthocyanines sont absentes (Lange and Fincke, 1970).

Les voies de biosynthèse conduisant à la production des flavonoïdes ont été largement étudiées et décrites chez de nombreuses espèces de plantes (Heim et al., 2003; Lepiniec et al., 2006; Winkel-Shirley, 2001). Les composés flavoniques dérivent de la Phénylalanine et de la Malonyl-coenzyme A (COA) via la biosynthèse des acides gras (Figure 13). La première étape de la voie de biosynthèse des flavonoïdes est catalysée par la chalcone synthase (CHS), qui utilise le malonyl CoA et le 4-coumaroyl CoA comme substrats.

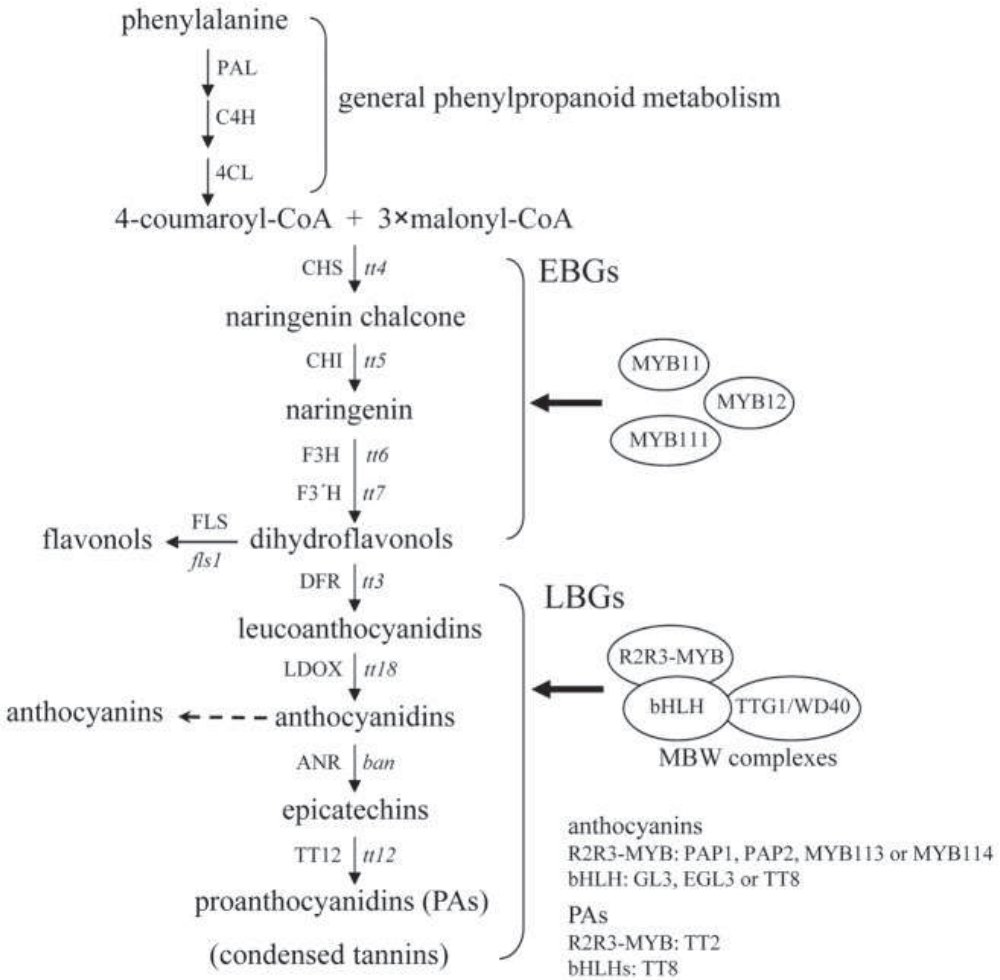


Figure 14 : régulation de la voie de biosynthèse des flavonoïdes chez *Arabidopsis thaliana* (d'après Li, 2013).

Abréviations : EBGs=Early Biosynthetic Genes; LBGs=Late Biosynthetic genes; PAL=phenylalanine ammonia-lyase; C4H=cinnamate 4-hydroxylase; 4CL=4-coumarate:CoA ligase; CHS=chalcone synthase; CHi=chalcone isomerase; F3H=flavanone 3-hydroxylase; F3'H=flavanone 3'-hydroxylase; DFR=dihydroflavonol 4-reductase; LDOX=leucoanthocyanidin dioxygenase; AnR=anthocyanidin reductase; *tt=transparent testa*; *ban=banyuls.*

La biosynthèse des flavanols implique principalement la flavone synthase (FLS) alors que la biosynthèse des proanthocyanines et des anthocyanines partage une étape commune conduisant aux flavan-3,4-diols (comme la leucoanthocyanidine) qui peuvent être convertis en catéchines (2,3-*trans*-flavan-3ol) par LCR ou en anthocyanidines par LDOX. Les anthocyanidines peuvent alors servir de substrat pour la synthèse des anthocyanines par glycosylation.

La régulation de ces voies de biosynthèse sont également bien connus (Jaakola, 2013; Li, 2014; Figure14). En résumé, trois complexes protéiques R2R3-MYB (MYB11, MYB12 et MYB111) contrôlent la biosynthèse des flavon-3-ols en activant les gènes de la voie de biosynthèse précoce (EBGs) alors que la production d'anthocyanines et de proanthocyanidines requiert le complexe protéique MYB-bHLH-WD40 (MBW) pour activer les gènes de la voie de biosynthèse tardive (LBGs). D'autres régulateurs de cette voie de biosynthèse ont été décrits plus récemment. Ils interagissent en favorisant ou en empêchant la formation des complexes protéiques R2R3-MYBs ou MBW.

Des études sur le cacaoyer ont permis de corréler la teneur en anthocyanine des fèves avec la teneur en polyphénols totaux (Cakirer et al., 2003, 2010). Les fèves claires contiennent moins de polyphénols que les fèves foncées. Chez le cacaoyer, les fèves blanches sont présentes dans plusieurs types de population :

- les populations qui découlent de la variété Criollo et donc dans certains hybrides Trinitario
- des individus prospectés dans certaines régions de haute Amazonie d'Equateur et du Pérou (Allen et Lass, 1983; Solorzano et al., 2012)
- des clones bas amazoniens "albinos" dont le clone Catongo, qui ont une absence complète d'anthocyanes dans les fèves, feuilles et fleurs.

Le déterminisme génétique de la couleur des fèves a été peu étudié chez le cacaoyer. Wellensiek (1931) suggère qu'un gène majeur dominant est responsable de la pigmentation des fèves de cacao dans les populations hybrides Trinitario. Crouzillat et al. (1996) ont identifié dans une population impliquant le parent "albinos" Catongo un locus majeur sur le groupe de liaison LG4 qui à l'état récessif confère la couleur blanche. Enfin, Marcano et al. (2009), ont identifié par une étude d'association réalisée dans une population d'hybrides Criollo, 3 régions d'association entre marqueurs et pigmentation des fèves sur les groupes de liaison LG1, LG4 et LG10.

CHAPITRE 1 - L'ANALYSE DU TRANSCRIPTOME DU CACAOYER

Une aide précieuse pour l'analyse des mécanismes moléculaires impliqués dans les caractères agronomiques d'intérêt

1. Limite des études de détection de locus affectant les caractères quantitatifs (QTLs)

Une des applications directes des études de détection de QTLs est la sélection assistée par marqueurs. La construction de génotype cumulant les allèles favorables des marqueurs qui composent les QTLs a été validée théoriquement et expérimentalement chez les plantes dès les années 1990 (Lande and Thompson, 1990; Stuber, 1989). Cependant, peu de travaux de ce type ont été publiés à ce jour chez le cacaoyer.

Les caractères agronomiques d'intérêt chez le cacaoyer, que ce soit pour la résistance aux maladies ou pour la qualité sont des caractères le plus souvent à déterminisme complexe et polygénique. La détection de QTLs de tels caractères souffre de plusieurs handicaps :

- Sa méthodologie est statistique et l'interprétation des résultats est parfois délicate.
- La distance physique à parcourir entre les marqueurs moléculaires et le QTL peut être très grande.
- Le nombre, la localisation précise, l'effet et le mode d'action des gènes qui sous-tendent ces QTLs ne sont pas connus.

Au début des années 2000, de nouvelles méthodes moléculaires basées sur le séquençage des ARN messagers ont ouvert de nouvelles perspectives concernant l'identification des gènes et l'analyse spatio-temporelle ou conditionnelle de leur expression.

2. L'approche ESTs

Une étiquette de gène exprimé ou EST (Expressed Sequence Tag) est une séquence nucléotidique plus ou moins partielle d'ADN complémentaire (ADNc) correspondant à une des extrémités d'un ARN messager. Elle donne, par conséquent, directement accès à une partie de la séquence codante du gène dont elle est issue. Par comparaison avec les séquences protéiques ou nucléotidiques contenues dans les bases de données internationales, il est alors possible de lui inférer une fonction potentielle. L'analyse de banques d'ESTs, réalisée à partir de populations cellulaires à étudier reflète l'état du transcriptome pour une situation expérimentale donnée et permet d'appréhender l'étude de l'expression des gènes.

Chez les plantes, les premières ESTs ont été publiées dans les bases de données internationales en 1992/1993 pour le riz (Uchimiya et al., 1992), le maïs (Keith et al., 1993) et *Arabidopis thaliana* (Höfte et al., 1993). Puis l'automatisation et l'augmentation des capacités de séquençage Sanger, favorisant la baisse des coûts ont permis à de nombreux programmes de séquençage d'être engagés. Parallèlement, de nombreux outils bioinformatiques ont vu le jour afin de faciliter les analyses et annotations. Les approches ESTs ont ainsi été développées chez de nombreuses espèces de plantes; par exemple une collection d'ESTs chez le *citrus* (Luro et al., 2008) a permis l'identification de plus de 15 000 transcrits et la découverte de nombreux gènes potentiellement impliqués dans la qualité des fruits, la production et la tolérance à la salinité. Un travail de plus grande ampleur a été mené chez le cotonnier (Udall et al., 2006) et a conduit à l'identification de plus de 50 000 transcrits à partir de 180 000 ESTs, conférant une boite à outil conséquente pour les futures recherches en génomique du cotonnier.

Chez le cacaoyer, seules quelques petites collections d'ESTs ont été produites au début des années 2000 et se sont limitées à l'étude des gènes impliqués dans les mécanismes de résistance et de défense, principalement concernant la maladie du balai de sorcière. Les premières banques d'ESTs ont été publiées en 2002 (Jones et al., 2002) et ont été réalisées à partir de 5 clones Amelonado sensibles à la maladie du balai de sorcière et 2 tissus, fève et feuille. Elles ont permis l'identification de 1380 transcrits qui ont été caractérisés par homologie de séquences. Plusieurs gènes impliqués dans les mécanismes de défense, de stockage d'énergie et de photosynthèse ont été annotés. Les profils d'expression de ces gènes ont été précisés par une analyse réalisée avec des puces ADN (microarray).

En 2004, Verica et al. ont publié une étude conduisant à l'identification de 1256 gènes à partir d'une banque d'ESTs réalisée à partir de feuilles de cacaoyer traitées par des inducteurs de réponse à des stress biotiques. L'expression de ces gènes a été étudiée par macroarray. Leal et al. ont identifié en 2007, 187 gènes dans des banques ESTs soustractives construites spécifiquement pour mettre en évidence les gènes exprimés entre du matériel sensible (ICS39) et résistant (CAB214) à la maladie du balai de sorcière. Enfin Gesteira et al. ont publié en 2007 une étude détaillant l'analyse de 6884 ESTs provenant de deux banques réalisées à partir de clones résistant (TSH1188) et sensible (Catongo) inoculés par l'agent pathogène de la maladie du balai de sorcière.

Pour complémenter l'étude du transcriptome sur d'autres caractères que la résistance à la maladie du balai de sorcière, un consortium international impliquant 15 instituts de recherche et coordonné par le CIRAD, a été créé en 2005. Le but de ce consortium était d'obtenir une collection exhaustive des gènes exprimés chez le cacaoyer, pour de nombreux organes, différents génotypes et de nombreuses conditions environnementales. Ce projet a abouti à la construction de 56 banques d'ADNc qui ont été séquencées par la méthode Sanger avec la collaboration du Genoscope. L'analyse de ce travail est détaillée sous la forme de l'article scientifique suivant, publié dans la revue BMC Genomics en 2008 et intitulé : **"Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions**".

BMC Genomics

BioMed Central

Research article

Open Access

Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions

Xavier Argout*[1], Olivier Fouet[1], Patrick Wincker[2], Karina Gramacho[3], Thierry Legavre[1], Xavier Sabau[1], Ange Marie Risterucci[1], Corinne Da Silva[2], Julio Cascardo[4], Mathilde Allegre[1], David Kuhn[5], Joseph Verica[6], Brigitte Courtois[1], Gaston Loor[7], Regis Babin[8,9], Olivier Sounigo[8,9], Michel Ducamp[10], Mark J Guiltinan[6], Manuel Ruiz[1], Laurence Alemanno[11], Regina Machado[12], Wilberth Phillips[13], Ray Schnell[5], Martin Gilmour[14], Eric Rosenquist[15], David Butler[16], Siela Maximova[6] and Claire Lanaud[1]

Address: [1]Biological Systems Department – UMR DAP TA 40/03, CIRAD, Montpellier, France, [2]GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France, [3]CEPLAC, Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00, Bahia, Brazil, [4]Laboratório de Genômica e Expressão GênicaRodovia Ilhéus-Itabuna, UESC, Km 16, Ilhéus, Brazil, [5]USDA-ARS, 13601 Old Cutler Rd. Miami, Florida, USA, [6]Department of Horticulture, The Pennsylvania State University, 422 Life Sciences Building, University Park, PA, 16802, USA, [7]EET-Pichilingue, INIAP, Código Postal 24 Km 5 vía Quevedo El Empalme, Ecuador, [8]IRAD, Nkolbisson, BP 2067, Yaoundé, Cameroon, [9]UPR 31 TA 80/02, CIRAD, Montpellier, France, [10]UMR BGPI TA41/K, CIRAD- 34398 Montpellier France, [11]UMR BEPC TA 80/03, CIRAD, Montpellier, France, [12]MASTERFOODS, Almirante, Brazil, [13]CATIE, P.O.Box 7170, Turrialba, Costa Rica, [14]Mars Inc., Dundee Road, Slough, SL1 4JX, UK, [15]National Program Staff, USDA-ARS, Beltsville, Maryland 20705, USA and [16]Cocoa Research Unit, The University of the West Indies, St. Augustine, Trinidad and Tobago

Email: Xavier Argout* - xavier.argout@cirad.fr; Olivier Fouet - olivier.fouet@cirad.fr; Patrick Wincker - pwincker@genoscope.cns.fr; Karina Gramacho - karina@cepec.gov.br; Thierry Legavre - thierry.legavre@cirad.fr; Xavier Sabau - xavier.sabau@cirad.fr; Ange Marie Risterucci - ange-marie.risterucci@cirad.fr; Corinne Da Silva - dasilva@genoscope.cns.fr; Julio Cascardo - cascardo@labbi.uesc.br; Mathilde Allegre - maallegre@yahoo.fr; David Kuhn - David.Kuhn@ARS.USDA.GOV; Joseph Verica - joeverica@yahoo.com; Brigitte Courtois - brigitte.courtois@cirad.fr; Gaston Loor - reyloor@yahoo.es; Regis Babin - regis.babin@cirad.fr; Olivier Sounigo - olivier.sounigo@cirad.fr; Michel Ducamp - michel.ducamp@cirad.fr; Mark J Guiltinan - mjg9@psu.edu; Manuel Ruiz - manuel.ruiz@cirad.fr; Laurence Alemanno - laurence.alemanno@orange.fr; Regina Machado - regina.machado@effem.com; Wilberth Phillips - wphillip@catie.ac.cr; Ray Schnell - rschnell@ars-grin.gov; Martin Gilmour - martin.gilmour@eu.effem.com; Eric Rosenquist - Eric.Rosenquist@ARS.USDA.GOV; David Butler - dbutler@intrepidequipment.com; Siela Maximova - snm104@psu.edu; Claire Lanaud - claire.lanaud@cirad.fr

* Corresponding author

Published: 30 October 2008

BMC Genomics 2008, **9**:512 doi:10.1186/1471-2164-9-512

Received: 4 June 2008
Accepted: 30 October 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/512

Abstract

Background: *Theobroma cacao* L., is a tree originated from the tropical rainforest of South America. It is one of the major cash crops for many tropical countries. *T. cacao* is mainly produced on smallholdings, providing resources for 14 million farmers. Disease resistance and *T. cacao* quality improvement are two important challenges for all actors of cocoa and chocolate production. *T. cacao* is seriously affected by pests and fungal diseases, responsible for more than 40% yield losses and quality improvement, nutritional and organoleptic, is also important for consumers. An international collaboration was formed to develop an EST genomic resource database for cacao.

Results: Fifty-six cDNA libraries were constructed from different organs, different genotypes and different environmental conditions. A total of 149,650 valid EST sequences were generated corresponding to 48,594 unigenes, 12,692 contigs and 35,902 singletons. A total of 29,849 unigenes shared significant homology with public sequences from other species.

Gene Ontology (GO) annotation was applied to distribute the ESTs among the main GO categories.

A specific information system (ESTtik) was constructed to process, store and manage this EST collection allowing the user to query a database.

To check the representativeness of our EST collection, we looked for the genes known to be involved in two different metabolic pathways extensively studied in other plant species and important for *T. cacao* qualities: the flavonoid and the terpene pathways. Most of the enzymes described in other crops for these two metabolic pathways were found in our EST collection.

A large collection of new genetic markers was provided by this ESTs collection.

Conclusion: This EST collection displays a good representation of the *T. cacao* transcriptome, suitable for analysis of biochemical pathways based on oligonucleotide microarrays derived from these ESTs. It will provide numerous genetic markers that will allow the construction of a high density gene map of *T. cacao*. This EST collection represents a unique and important molecular resource for *T. cacao* study and improvement, facilitating the discovery of candidate genes for important *T. cacao* trait variation.

Background

Theobroma cacao is a diploid species (2n = 2X = 20) with a small genome size of 380 Mbp [1,2]. It is a tree fruit originating from the tropical rainforest of South America. According to Cheesman (1944) [3], its center of origin is the lower eastern equatorial slopes of the Andes. *T. cacao* is now cultivated in all tropical lowlands of the world and its beans are used to produce chocolate and cocoa butter after a post harvest treatment including fermentation, drying and torrefaction steps. *T. cacao* is one of the major cash crops for several tropical countries. Its economic importance is high and presently cocoa is the third most important internationally traded raw material after sugar and coffee.

Cocoa is mainly produced on smallholdings. It is estimated that approximately 14 million people around the world rely on cacao plantations for income. *T. cacao* production is seriously affected by several fungal diseases and insect attacks. Oomycetes and especially *Phytophthora*, spp., (black pod) are responsible, worldwide, for 30% of losses. Several species are involved. *P. palmivora* is present in the entire cacao growing area, whereas *P. capsici* and *P. citrophthora* are prevalent in South America. *P. megakarya* is limited to some countries in West Africa, however it is by far the most aggressive species causing losses of production up to 50% Harvest losses due to *Phytophthora* species were estimated to be 450,000 tons [4].

Two basidiomycetes, *Moniliophthora roreri* (frosty pod) and *Moniliophthora perniciosa* (witches' broom) are also responsible for important harvest losses. In Brazil, *M. perniciosa* was responsible for a drastic yield loss with a fall in production from 405,000 tons in 1986 to less than 130,000 tons in 1998. *Moniliophthora roreri* causes a very destructive pod rot and has already had dramatic effects in some countries such as Ecuador [5] and Costa Rica [6]. *M. roreri* was confined to several countries of Central and northern South America, but is continuously spreading towards other Central American countries like Mexico or southward towards countries like Peru.

Several sources of disease resistance have been identified in different genetic backgrounds, and the search for a sustainable disease resistance, cumulating the different resistance genes is one of the major challenges of *T. cacao* genetic breeding programs [7].

Other traits of importance in *T. cacao* are quality traits. Food quality improvement, nutritional as well as organoleptic, is now a strong demand of consumers. Fundamental knowledge of the genetic basis of quality is an important challenge that can address this demand.

Flavor is among the main criteria of quality for chocolate manufacturers, but these characteristics are largely understudied by the cocoa research and breeding community due to their complexity and a dramatic lack of fundamental knowledge about these traits. Flavour components depend strongly on conditions of post-harvest processing [8]. After pod harvests, fresh seeds need to be fermented for 4 to 6 days, then dried and roasted to develope good

cocoa aromas. Raw seeds, embedded in a pulp rich in sugar, undergo biochemical changes under the effect of various microorganisms present in the environment. The initial anaerobic, low pH and high sugar conditions of the pulp favour yeast activity, converting sugars in the pulp to alcohol and carbon dioxide. Bacteria then start oxidising the alcohol into lactic acid and then into acetic acid as conditions become more aerobic. These biochemical changes are accompanied by changes of amount and composition of several compounds having a major effect on cocoa flavor such as peptide aroma precursor formation, procyanidines or terpenes content.

However, it is now well recognized that the genetic origin is also a strong determinant of flavor, independent of the conditions of post-harvest processing [9].

Although some aromas are prominently defined by a single molecule, most aromas are composed of a bulk of volatile compounds responsible for aroma perception, and belonging to different classes of organic compounds. Interestingly, despite the vast number of chemical structures involved, the large majority of scent compounds are biosynthesized by a surprisingly small number of metabolic pathways. Parts of these metabolic pathways are ubiquitous, and have been developed by small but important modifications of ancestral genes and pathways [10]. In *T. cacao* more than 500 volatile compounds have been detected. However, only a small number are thought to play a key role in natural aroma variations.

Cocoa is classified into two classes: the «standard quality cocoa» corresponding to 95% of the total market, and the «fine flavor cocoa» produced by *T. cacao* trees originated from two main varieties: Criollo and Nacional, which bring a higher price in the market.

An important class of volatile compounds, the terpenes, plays an important role in the aromatic flavor of these varieties.

For example, a high level of linalool, a monoterpene, has been observed in Nacional varieties [11] from Ecuador, characterised by a floral taste, and could be at the origin of this specific flavor which represents an important economic «niche» for the country. However, the modern and hybrid Nacional varieties present a wide range of flavor variations due to introgressions of foreign and more vigorous varieties, leading to a dilution of this specific floral flavor, and recently a part of Ecuador cocoa production was declassified from fine flavor to "bulk cocoa" with a lower price. An increased knowledge of the metabolic pathways and expression of genes involved in terpene synthesis could help to improve the aromatic flavor of new "Nacional" varieties.

Independent to volatile compounds, some other biochemical compounds are known to interact with *T. cacao* organoleptic traits. This is the case with polyphenols. Catechin, epicatechin and procyanidines are the main polyphenols present in *T. cacao*. They have well known antioxidant biological activities and beneficial effects on the cardiovascular system [12-14]. Contributing to bitterness and astringency, polyphenols influence *T. cacao* organoleptic quality [15,12]. They influence aromatic profiles of *T. cacao* in restricting Maillard's reactions, which generates a majority of the aromatic compounds of *T. cacao*.

Genomic research provides new tools to study the genetic and molecular bases of important trait variations: EST sequencing projects carried out on other plant models have allowed the characterization of the transcriptome and facilitated the gene discovery of important trait variations [16]. In tree crops, except for poplar whose genome has been recently sequenced [17], genomic resources are generally limited, and few large EST collections have been produced. Recently, a *citrus* EST collection comprising 15,664 putative transcription units [18] has been produced, allowing the identification of clusters associated with fruit quality, production and salinity tolerance. A cotton study identified 51,107 unigenes from a global assembly of 185,000 cotton ESTs, [19] providing a framework for future investigation of cotton genomics. The same approach was used to characterize the grape transcriptome during berry development by the analysis and annotation of 25,746 unigenes from 146,075 ESTs [20].

In *T. cacao*, only small collections of ESTs have been produced so far and used to study gene expression related to stress or disease resistance and defense [21-24]

The objective of this study was to produce a large *T. cacao* EST collection from a wide range of organs, providing a good representation of *T. cacao* genes expressed during *T. cacao* development and suitable for further analysis of all kind of traits in *T. cacao*. Moreover, we emphasized the production of tools to further study *T. cacao* diseases, a major constraint for cocoa production, and quality features. Therefore, we also produced cDNA libraries relevant to disease resistance and quality traits. ESTs were produced from *T. cacao* tissues interacting with various pest and fungal diseases, from seeds at different stages of development and during the fermentation steps. This large EST collection will provide valuable tools to carry out functional genomic studies and discover genes essential to important agronomic and quality trait variation in *T. cacao*, aiming to accelerate *T. cacao* improvement. A multidisciplinary approach combining functional genomic and quantitative genetic approaches could lead to a better understanding of gene function involved in disease resistance mechanisms or quality trait variations. *T. cacao*'s phy-

logenetic proximity to the model plant *Arabidopsis* will facilitate our understanding of most metabolic pathways. However, *T. cacao* is a tree, and expresses traits not found in *Arabidopsis*, thus we hypothesize that genes not found in *Arabidopsis* play important roles in cacao development.

Results and Discussion

Library construction

Fifty-six libraries were constructed from two main genotypes representing three contrasting genetic origins: ICS1, a hybrid between Criollo and Forastero from Lower Amazonia of Brazil, and Scavina 6, a Forastero from Upper Amazonia of Peru. A few other genotypes characterized by specific resistance or quality traits and belonging to various genetic origins were also used. The plant materials were provided from a various panel of different *T. cacao* L. organs (Table 1). Among them, 25 libraries corresponded to *T. cacao* tissues introduced to different biotic stresses: pods inoculated by *Phytophthora palmivora, Phytophthora megakarya, Moniliophthora perniciosa* and *Moniliophthora roreri*, leaves inoculated by *Phytophthora palmivora* and *Phytophthora megakarya*, stems inoculated by *Moniliophthora perniciosa* and *Ceratocystis fimbriata*, and stems attacked by *Sahlbergella singularis* (mirids). Among these libraries, 17 are suppressive subtractive hybridization (SSH) libraries. Finally, two libraries corresponded to *T. cacao* tissues introduced to drought stresses and 11 corresponded to seed development and fermentation stages.

EST sequencing and assembly

From the 56 libraries, 8565 clones were first sequenced on both strands using forward and reverse primers, to have an overview of the quality of the libraries, and then 163,868 clones were single-pass sequenced from 5' or from 3' end (Table 1). This represented a total number of 180,998 chromatograms that were used in this analysis. After low quality, vector and adapters trimming, 149,650 sequences longer than 100 bp remained as good quality sequences. The average sequence length was 472 bp and 62% were longer than 400 bp. These individual ESTs (available through EMBL-Bank [25]) were assembled using the TIGR Gene Indices clustering tools (TGICL) [26]. The assembly process produced 12,692 contigs and 35,902 singletons that represented a total of 25.6 Mb of transcripted sequences. The combined set of contigs and singletons resulted in 48,594 unigenes which might correspond to different putative transcripts or different parts of the same transcript found in the *Theobroma cacao* transcriptome. The average length of this *T. cacao* non redundant sequences dataset was 527 bp.

An assembly of ESTs has already been published for *Theobroma cacao* but has been limited to 1380 unigenes (4433 ESTs) from two leaf and bean cacao libraries [21], to the isolation of 1256 unigenes (2114 ESTs) from cacao leaves treated with inducers of defense response [23] and to 2926 non redundant sequences from libraries of cacao meristems inoculated by *Moniliophthora perniciosa* [24].

The results of this study are more comparable to a cotton EST project [19], involving 30 cDNA libraries. This analysis detected 51,107 unigenes in approximately 185,000 *Gossypium* ESTs.

Analysis of EST abundance in a contig can provide insights to gene expression levels, although this information must be taken with caution due to cloning and replication bias resulting form library construction and propagation steps. The number of ESTs in the *T. cacao* contigs ranged from 2 to 5102 (Figure 1) and 65.3% were composed of 4 or less ESTs. 98% of the contigs contained less than 50 ESTs.

We evaluated the redundancy of transcripts in each library and among all libraries by studying the distribution of ESTs in contigs across multiple libraries. 11,226 had members from more than one library (Figure 2) and 1466 contigs were specific from one library. No contigs had members from all 56 libraries. Two contigs were found in 52 libraries: the contig CL1Contig269 was similar to the mitochondrial large subunit ribosomal RNA gene and the contig CL1Contig513 to the 18S ribosomal RNA gene. The contig CL18Contig2, CL2Contig3 and CL15Contig2, similar to an ATP Synthase beta subunit, a metallothionein-like protein and a photosystem II D1 protein respectively, were found in 47 libraries.

Unigene set annotation

BLASTN against cacao ESTs

The unigene dataset was used to detect how many cacao sequences had not been already described in public databases. To answer this question, we collected all 2539 *T. cacao* unique sequences already published by the Dana Farber Cancer Institute (DFCI) gene index [27] and we did a BLASTN search against our unigenes. An e-value cutoff of $1e^{-50}$ was used to ensure that only highly similar sequences were detected. A total of 3901 unigenes produced a significant hit with 1788 unique sequences from the DFCI gene index, therefore these sequences may correspond to *T. cacao* sequences already published or may match different parts of the gene index sequences. They may be also produced by closely related genes (multigenic families). Finally, 44,693 unique sequences did not produce a significant hit, therefore these sequences may be new.

BLASTX and BLASTN annotation

The unigenes were first translated into amino-acid sequences and then searched for similar protein with the BLASTX program using an e-value cutoff of 1e-5 against

Table 1: Summary of *T. cacao* libraries

Genotype	Library	Library description	Good quality ESTs	Unigenes
Jaca	CERATOJ_KZ0ACI	stem tissues inoculated by Ceratocystis fimbriata	1729	1270
Scavina6	CHERELS_KZ0AAC	cherels from 1 week to 1 month stage of development	4252	2836
Scavina6	COPHAS_KZ0AAL	pod tissue inoculated by Phytophthora palmivora	4905	2621
Scavina6	CORTEXS_KZ0AAT	cortex tissue, external part	3817	2227
Scavina6	CORTINS_KZ0AAV	cortex tissue internal part with lignified chanels	5096	3331
ICS1	COSSHPPI_KZ0AA	SSH library from tissues inoculated/non inoculated by Phytophthora palmivora	1721	955
Scavina6	COSSHPPS_KZ0AA	SSH library from tissues inoculated/non inoculated by Phytophthora palmivora	1702	1129
ICS1	COTYLEI_KZ0ABB	cotyledons from germinated seeds (1 to 3 weeks)	5153	2961
B97 C-C-2	CUSHIONC_KZ0ACAC	young cushions	2849	2120
Scavina6	DROUGHTLS_KZ0ACAF	leaves submited to drought stresses	2766	1290
Scavina6	DROUGHTRS_KZ0ACAE	roots submited to drought stresses	2685	1563
ICS1AF	EMBR1WI_KZ0ABA	epicotyle and hypocotyle from 1 week germinated seeds	3246	2473
ICS1AF	EPIC23I_KZ0AAS	epicotyle from 2–3 week germinated seeds	3005	2459
Scavina6	FLOWERS_KZ0AAD	flowers at different stages of development	3511	2434
Scavina6	FLPOLSSH_KZ0ABL_M	SSH library from ovaries submitted to compatible/incompatible pollinations	2398	431
ICS1AF	HYPO23I_KZ0AAP	hypocotyle from 2–3 week germinated seeds	5111	2955
Scavina6	LEAVES_KZ0ABE	young and adult leaves at different stages of development	4698	3069
GU255V	LEAVPAGU_KZ0ACQ	leaves inoculated by Phytophthora palmivora	3030	2139
PNG seedlings	LEPAPNGR_KZ0ACP	leaves inoculated by Phytophthora palmivora	1021	862
PNG seedlings	LESSHMEPNGa_KZ0ACAP	SSH library from leaves inoculated by Phytophthora megakarya from susceptible-resistant PNG seedlings	356	169
PNG seedlings	LESSHMEPNGb_KZ0ACV	SSH library from leaves inoculated by Phytophthora megakarya from resistant – susceptible PNG seedlings	1244	749
PNG seedlings	LESSHPNGRSb_KZ0ABP	SSH library from leaves inoculated by Phytophthora palmivora from resistant – susceptible PNG seedlings	701	438
UF676	MIRIDUFS_KZ0ACAD	young shoot tissues attacked by Sahlbergella singularis (mirids)	3011	1908
P7	MONILIOP_KZ0AB	pod tissues inoculated by Moniliophthora roreri	3074	2217
UF273	MONILIOU_KZ0ABV	pod tissues inoculated by Monilia roreri	3159	1871
IMC47	OVULI_7M_KZ0ACAK	ovaries from 1 to 7 days after pollinations	1565	1218
ICS1	OVULEI_KZ0AAB	ovules collected 2 to 3 months after pollination	4942	3315
UPA134	PODMEUPA_KZ0ACAB	pod tissues inoculated by Phytophthora megakarya	3492	2093
Scavina6	PODSSHWB1Sb_KZ0ACD	SSH library from pod tissues inoculated-non inoculated by Moniliophthora perniciosa less than 60 days after inoculation	652	534
Scavina6	PODSSHWB2Sb_KZ0ACF	SSH library from pod tissues inoculated-non inoculated by Moniliophthora perniciosa between 60 to 120 days after inoculation	1399	912
Scavina6	PODWB1S_KZ0ACM	pod tissues inoculated by Moniliophthora perniciosa less than 60 days after inoculation	1704	1213
Scavina6	PODWB2S_KZ0ACN	pod tissues inoculated by Moniliophthora perniciosa between 60 to 120 days after inoculation	1718	1217
PNG seedlings	RESSHMEPNGb_KZ0AC	SSH library from leaves of resistant seedlings inoculated-non inoculated by Phytophthora megakarya	1287	931
Scavina6	ROOTS_KZ0ABF	roots	3567	2892
PNG seedlings	RPPSSHPNGa_KZ0ACAL	SSH library from leaves of resistant seedlings non inoculated-inoculated by Phytophthora palmivora	344	266
PNG seedlings	RPPSSHPNGb_KZ0ACR	SSH library from leaves of resistant seedlings inoculated-non inoculated by Phytophthora palmivora	1407	823
ICS1	SEED34I_KZ0AAH	seeds 3 to 3,5 months after pollinations	3942	2637
ICS1	SEED45I_KZ0AAE_F	seeds 4 to 5 months after pollinations	3296	1902
33–49	SEEDFERB_KZ0ACAG	Cotyledons from seeds fermented between 6 H and 4 days	1664	465
ICS1	SEEDMAI_KZ0AAG	seeds from mature pods 5,5 to 6 months after pollinations	3068	1844

Table 1: Summary of *T. cacao* libraries *(Continued)*

BE240	SEEDNAB_KZ0ABH	seeds 2 to 5 months after pollinations	4988	3101
ICS1	SEFERMI_A_KZ0AAR	fermented seeds during 6 to 26 H	1798	844
ICS1	SEFERMI_B_KZ0AAM	fermented seeds during 32 to 40 H	3931	2110
Jaca	SSHCERATOJb_KZ0ACS	SSH library from stems inoculated-non inoculated by Ceratocystis fimbriata	339	327
Jaca	SSHCERATOJa_KZ0ACAM	SSH library from stems non inoculated-inoculated by Ceratocystis fimbriata	1364	918
UF676	SSHMIRUFa_KZ0ACAN	SSH library from young shoots non attacked-attacked by Sahlbergella singularis	320	296
UF676	SSHMIRUFb_KZ0ACT	SSH library from young shoots attacked-non attacked by Sahlbergella singularis	1393	1051
Scavina6	STEMS_KZ0AAA	complete disc of stems 1 cm diameter	4938	2880
Scavina6	STSSHWB1S_KZ0ABI_K	SSH library from (and reverse sens) shoot tissues inoculated/non inoculated by Moniliophthora perniciosa less than 18 days after inoculation	1594	370
Scavina6	STSSHWB2Sb_KZ0ACB	SSH library from shoot tissues inoculated-non inoculated by Moniliophthora perniciosa between 18 to 120 days after inoculation	1408	1056
33–49	TEGFERB_KZ0ACAH	testa from seeds fermented between 6 H and 4 days	1649	808
ICS1	TEGPULI_KZ0AAI_K	testa with pulp from mature seeds	5017	3254
Scavina6	TISCIVS_KZ0AAQ	embryogenic and non embryogenic callus in vitro culture	3434	2389
ICS1	TPFERMI_A_KZ0AAN	fermented testa during 6 to 40 H	4005	2164
P7	WILTP_KZ0ACL	young wilted cherels 7 to 10 days after pollination	1706	1247
Scavina6	WOODS_KZ0ACAA	bark and cambium part of wood	3478	2234

the non-redundant protein sequence database (NR) with entries from GenPept, Swissprot, PIR, PDF, PDB and NCBI RefSeq. The 10 best hits were retained for the annotation, providing an annotation for 27,245 cacao sequences (56.1%). The 43.9% of the unigenes that did not have any match were searched for similar nucleotide sequences from the Genbank nucleotide collection NT with the BLASTN program. An e-value cutoff of 1e-5 was also used and the 10 best hits were used for the annotation. 2604 unigenes exhibited a significant similarity with nucleotide sequences providing a BLASTX or BLASTN annotation for 29,849 unigenes. The 10 BLASTX hits were used to classify the unigenes according to the species associated with the annotation (Figure 3A). A total of 140,270 hits (56%) involved proteins from *Vitis vinifera*, *Arabidopsis thaliana* or *Oryza sativa*, while 1955 hits involved proteins from *Gossypium hirsutum*, a closely related species from the Malvaceae family [28]. Although fewer protein sequences from *Vitis vinifera* than from *Arabidopsis thaliana* (54,395 and 58,061 respectively) were present in the non redundant database we used for BLASTX, and although the evolutionary distance between *Vitis vinifera* and *Theobroma cacao* is higher than the distance between *Arabidopsis thaliana* and *Theobroma cacao* [28], we found more similarities with *Vitis vinifera* (50,315 hits) than with *Arabidopsis thaliana* (41,766 hits).

To further investigate this unexpected result we compared with the BLASTX program the cacao unigenes dataset against the two proteomes of *Arabidopsis thaliana* and *Vitis vinifera* (Figures 3B, C). For each Blast result, we selected the species found in the first hit having an expected value lower than 1^{e-15} to detect similar sequence. A total of 25,049 *Theobroma cacao* sequences (56%) presented at least a significant hit with an *Arabidopsis thaliana* or *Vitis vinifera* protein. The results showed that 18,643 *Theobroma cacao* sequences presented a higher similarity to the *Vitis vinifera* proteome whereas only 6406 *Theobroma cacao* sequences presented a first Blast hit similar to the *Arabidopsis thaliana* proteome (Figure 3B). Moreover, it was determined that these first significant hits involved 9943 *Vitis vinifera* proteins (33% of the proteome) and 4246 *Arabidopsis thaliana* proteins (12% of the proteome) (Figure 3C).

These surprising results suggest that the genes expressed in *Theobroma cacao* are more similar to *Vitis vinifera* proteins than to those of *Arabidopsis thaliana*. These findings could be explained by the fact that *Theobroma cacao* and *Vitis vinifera* are both fruit trees. This idea could be supported by the large amount of Blast hits found with other tree crops such as *Populus trichocarpa* (8605 Blast hits), despite a small number of non redundant proteins in the databases for this species.

Gene Ontology annotation

We used BLAST2GO [29], a program that retrieves GO terms based on BLAST definition, to assign gene ontology

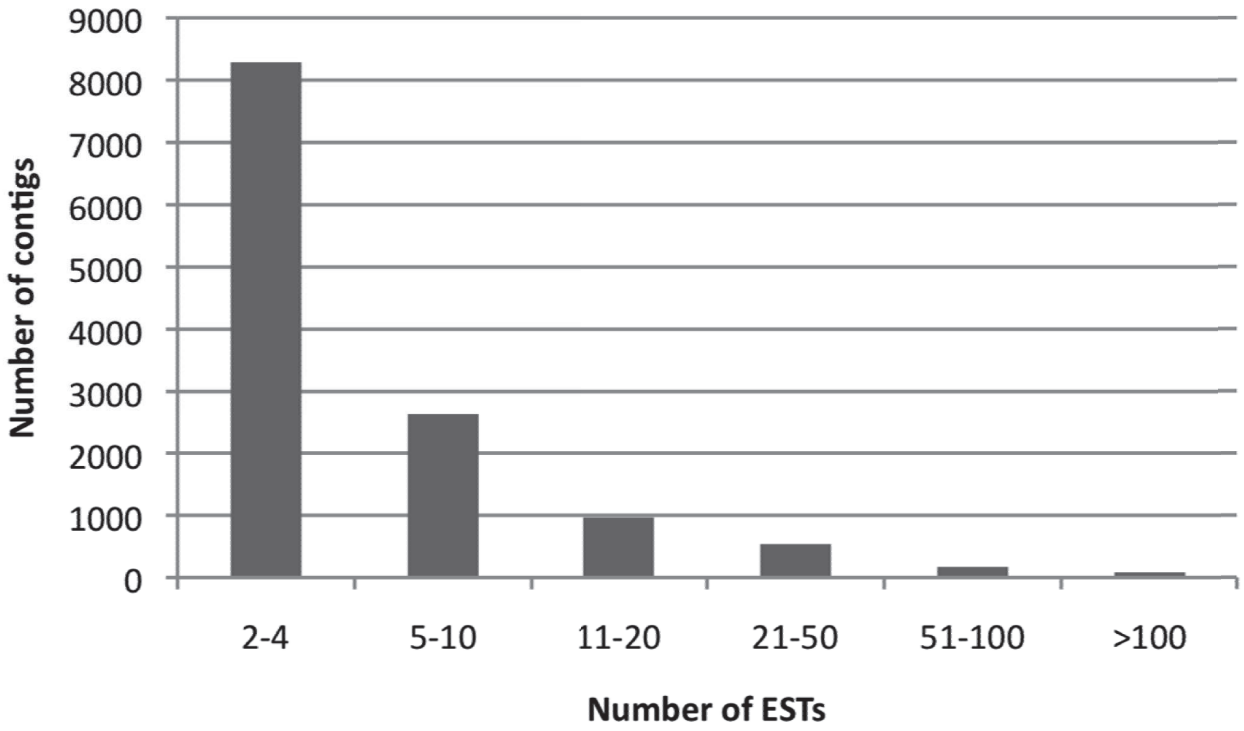


Figure 1
Distribution of *T. cacao* EST members in contigs after the assembly process.

(GO) annotation [30] to the unigene dataset. To best exploit GO results, we built a local AmiGO browser [31]. A total number of 49,364 annotations were found and 16,364 unigenes were characterized by at least one annotation. These annotations were distributed among the main GO categories into 16,448 Biological Process (P), 14,696 Cellular Component (C) and 18,219 Molecular Function (F) (Figure 4A). The most abundant high-level direct GO counts within these categories were C: mitochondrion (1924), C: membrane (1218), C: plastid (1173), F: ATP binding (1017) and C: chloroplast (1001) (Figure 4B).

Genes involved in defense and resistance mechanisms

Some of the libraries provide an important resource to study plant/pathogens interactions. Using the annotations provided by Blast and Gene Ontology, we specifically focussed on genes known to play a crucial role in plant pathogen resistance and defense mechanisms [32]. Using the AmiGO browser, we identified 1001 gene product associations to "response to stress" (GO:0006950). Both searches with Blast result and Gene Ontology annotation resulted in the identification of unigenes similar to known proteins involved in resistance or defense mechanisms such as LRR-NBS [33] (8 contigs and 32 singletons), chitinase [34] (19 contigs and 37 singletons), 1–3 beta glucanase [35] (5 contigs and 7 singletons) or pathogenesis-Related protein (24 contigs and 24 singletons).

Other genes related to resistance/defense mechanisms were also found more specifically in libraries produced from pathogen infected tissues, such as those involved in regulation of pathogen-induced genes like transcription factors (6 contigs and 7 singletons), in signal transduction (like MAPKinase with 5 contigs and 3 singletons) or in the cell death program.

The identification of a unigene set gathering sequences from all genes known to be involved in plant resistance and defense mechanisms, and the construction of a corresponding microarray could constitute a valuable tool to progress in the understanding of plant/pathogens interactions.

Genes involved in particular metabolic pathways or biological activities

To check the representativeness of our EST collection, we looked for ESTs encoding proteins known to be involved in the flavonoid and the terpene pathways, already studied in other plant species, and at the basis of important

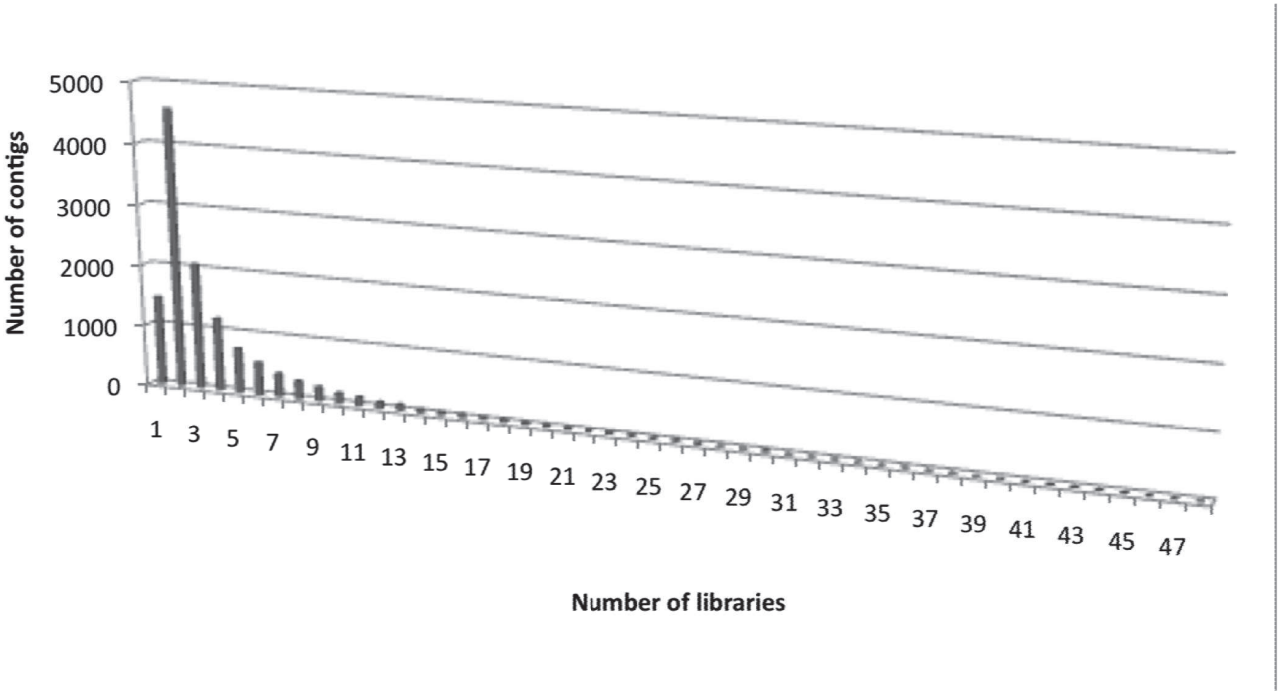


Figure 2
Number of contigs composed from sequence originated from one ore more libraries.

traits of interest in *T. cacao*. Generally, polyphenols play a major role in chocolate quality, acting as colour precursors or taste agents [36]. Moreover, they are strongly implicated in health benefits associated with chocolate consumption [37-40].

The flavonoid pathway

The flavonoid pathway has been already studied in several plants [41]. In *T. cacao*, this pathway is the source of numerous essential components for human health benefits of chocolate [37-40] and resistance against pathogens [42].

Gene Ontology analysis highlighted 99 EST sequences implicated in "phenylpropanoid biosynthetic process" (GO:0009699), most of them implicated in flavonoid biosynthesis. For example, the GO analysis, together with keyword ESTtik database searching (see material and methods) into Blast Results allowed us to find sequences encoding phenylalanine ammonia lyase (5 contigs and 12 singletons), cinnamate-4-hydroxylase (4 contigs and 11 singletons), the 4-coumarate-CoA ligase (14 contigs and 12 singletons), chalcone synthase (6 contigs and 25 singletons) and chalcone isomerase (8 contigs and 13 singletons), all major enzymes of the general flavonoid pathway (Figure 5). Most specific enzymes, implicated in anthocyanin biosynthesis (flavanone-3-hydroxylase, dihydroflavonol reductase, anthocyanidin synthase, flavonoid-3-glucosyltransferase) were also represented in this *T. cacao* EST resource.

The terpene pathway

Terpenoid compounds, synthesized in the isoprenoid pathway (Figure 6), are compounds of importance for specific scent and aromatic qualities of chocolates classified as "fine and flavor". For example, linalool, a monoterpenol, is found in high quantity in Arriba Nacional varieties from Ecuador and in some Criollo clones from Venezuela [11,43,44]. Linalool, together with other volatiles, could be responsible for the typical floral aroma [45] of these chocolates.

One of our goals was to identify enzymes involved in the terpenoid pathway that could be responsible for linalool content variations among Nacional clones. As a first step we identified sequences encoding isoprenoid pathway enzymes (42 contigs and 55 singletons). The final step enzyme for linalool synthesis, linalool synthase, was represented by 2 contigs and 4 singletons. Nearly all enzymes reported to be involved in this biochemical pathway were present in our ESTtik database, allowing the analysis of the *T. cacao* terpene pathway based on oligonucleotide microarrays derived from these ESTs.

The fact that nearly all of the genes involved in these two pathways as described in other plant species were identi-

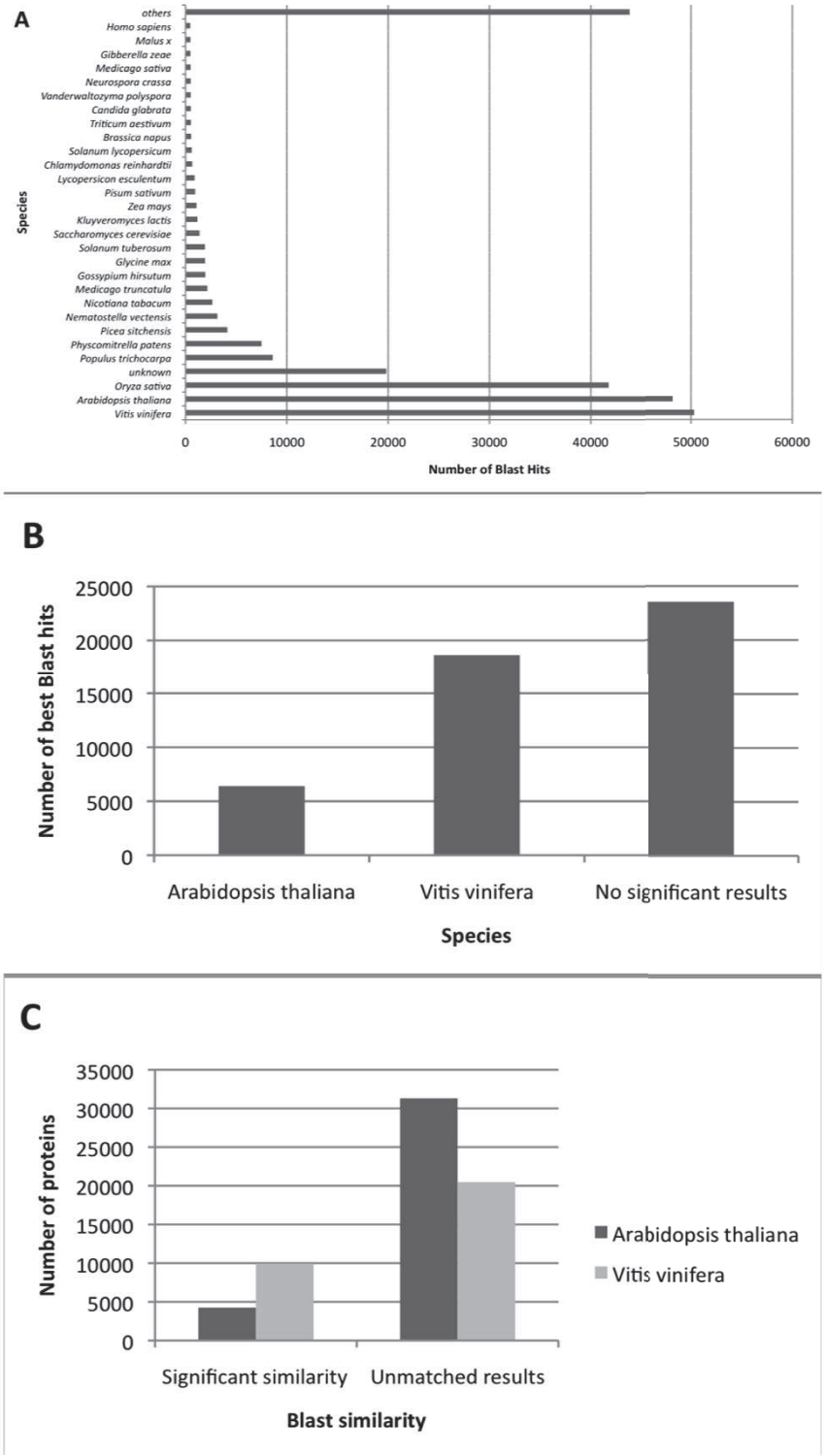


Figure 3
Species distribution among the Blast results of *T. cacao* unigenes. A – Distribution of species represented in the 10 first Blast hits against NCBI Non redundant protein database. B – Number of best Blast hits against *Arabidopsis thaliana* and *Vitis vinifera* proteomes. C – *Arabidopsis thaliana* (black columns) and *Vitis vinifera* (grey columns) proteome coverage.

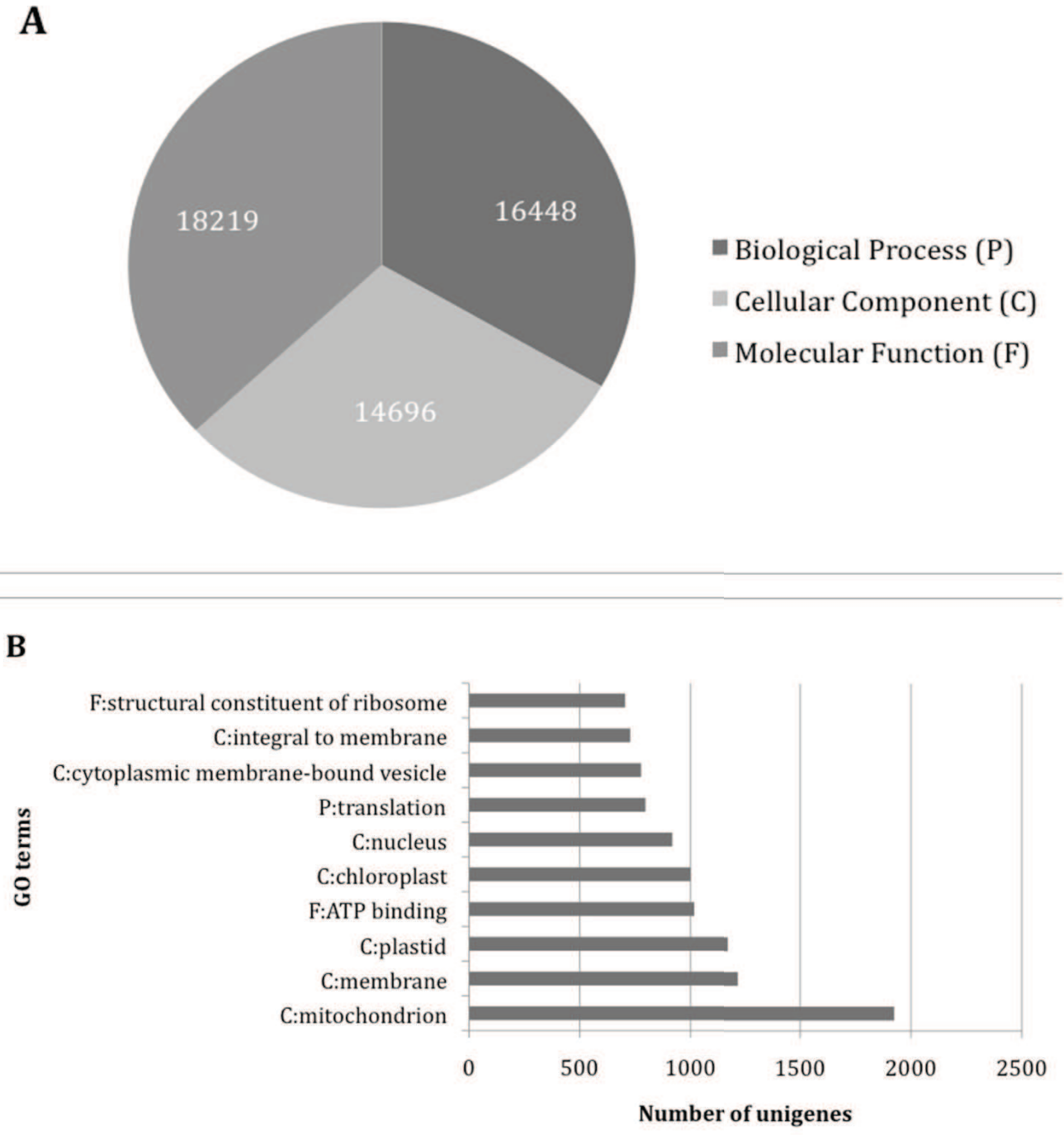


Figure 4
Gene Ontology annotation results. A – Distribution of the unigenes among the main Gene Ontology categories (Biological Process, Cellular Component and Molecular Function). B – Distribution of the unigenes among the 10 best Gene Ontology terms.

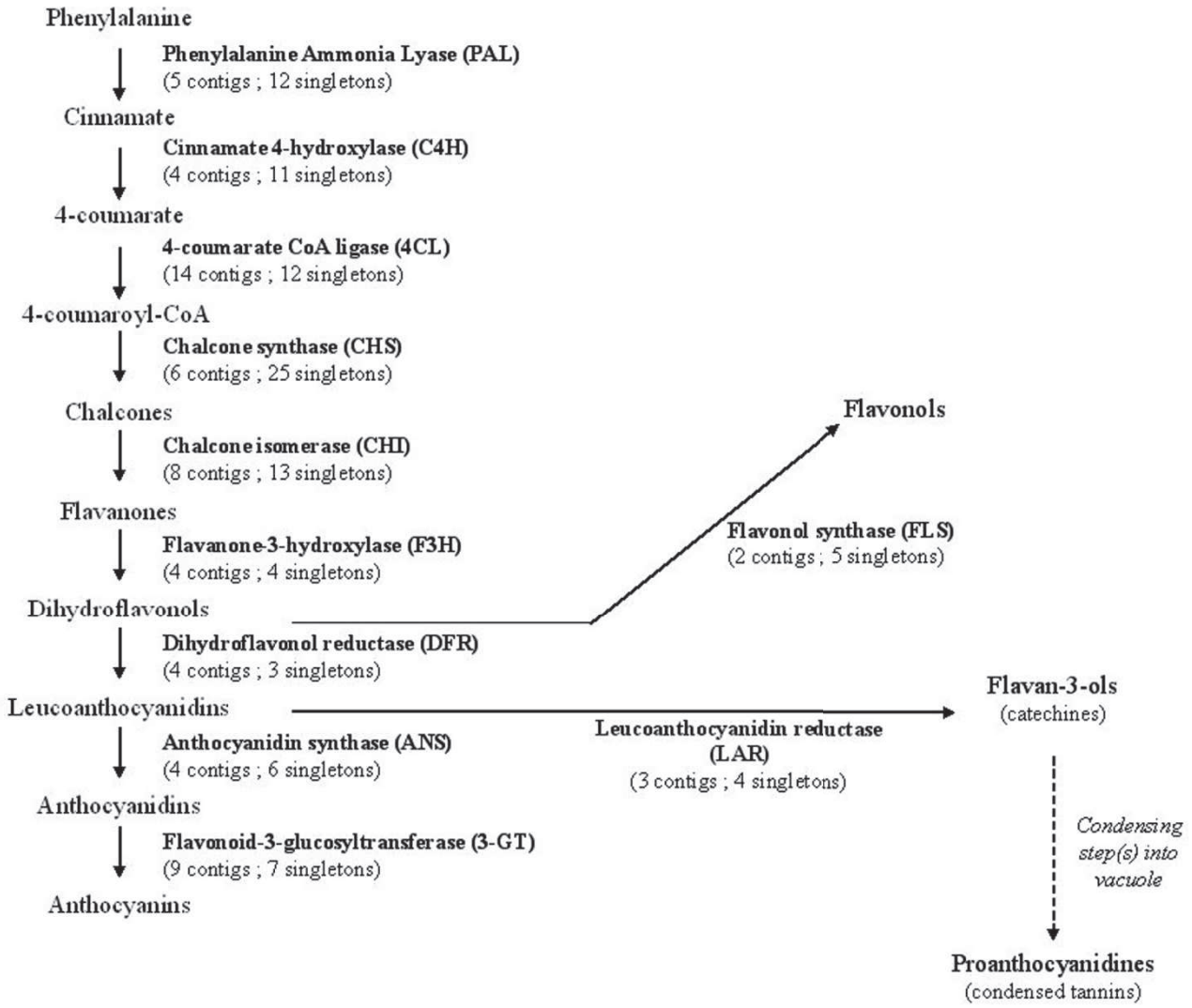


Figure 5
Schematic overview of the general flavonoid biosynthesis pathway (according to Schijlen et al., 2004; Marles et al., 2003). The number of contigs and singletons present in our EST dataset was added between brackets for each enzyme.

fied in ESTs from our collection demonstrates the high level of representation of this resource and suggests that the majority of cacao genes have been sampled. Thus, this EST collection offers a comprehensive resource to search for candidate genes involved in quality traits and other important agronomical traits variation.

Production of SSR and SNP markers

Molecular markers derived from ESTs are part of, or adjacent to genes, and therefore they provide an efficient means of gene mapping.

Simple Sequence Repeats (SSRs) were identified in the unigene dataset with the MISA pipeline [46]. In this study, SSRs were defined as dimers with at least 6 repetitions and trimers, tetramers, pentamers and hexamers with at least 5 repeats. Microsatellites were considered compound when two SSRs were not separated by more than 100 bp. A total of 2252 SSRs were identified as 2164 unigenes, and 204 unigenes had more than 1 SSR. Dimers and trimers were the most common types (Table 2) and represented 94.2% of SSRs found in unigenes. The distribution of all possible dimer and trimer motifs found in the unigenes is

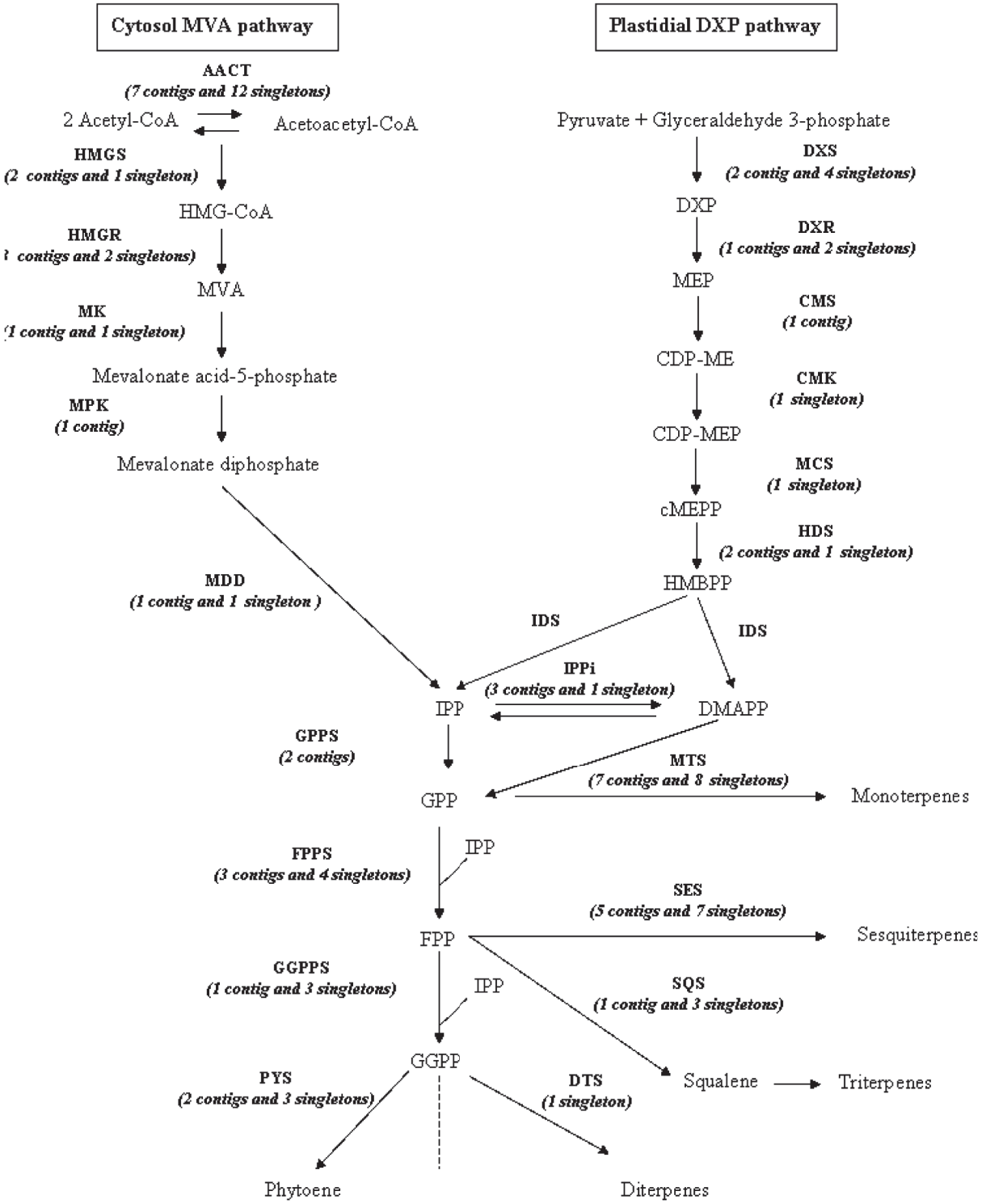


Figure 6 *(see legend on next page)*

Figure 6 *(see previous page)*
The biosynthesis pathway of isoprenoïdes. (according Liu *et al.*, 2005). Pathway Mevalonate (MVA) cytoplasmic in left and pathway 1-deoxyxylulose-5-phosphate (DXP) chloroplastic in right. **AACT**, acetoacetyl-coenzyme A (CoA) thiolase; **CMS**, 2-*C*-methyl-*D*-erythritol 4-phosphaate cytidyl transferase; **DTS**, diterpene synthase; **DXR**, 1- deoxy-*D*-xylulose 5-phosphate reductoisomerase; **DXS**, 1-deoxy-*D*-xylulose 5-phosphate synthase; **FPPS**, farnesyl diphosphate synthase; **GGPPS**, geranylgeranyl diphosphate synthase; **GPPS**, geranyl diphosphate synthase; **HMGR**, 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase; **IPPi**, isopentenyl diphosphate isomerase; **MTS**, monoterpene synthase; **SES**, sesquiterpene synthase; **SQS** squalene synthase; **MK**, mevalonate kinase; **MPK**, mevalonate-5-phosphate kinase; **CMK**, 4-(cytidine 5'-diphospho)-2-*C*-methyl-*D*-erythritol kinase; **MDD**, mevalonate diphosphate decarboxylase; **IDS**, isopentenyl diphosphate/dimethylallyl diphosphate synthase; **MCS**, 2-*C*-methyl-*D*-erythritol 2,4-cyclodiphosphate synthase; **HDS**, 1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate synthase; **PSY**, phytoene synthase; **HMGS**, HMG-CoA synthase; **HMG-CoA**, 3S-hydroxy-3-methylglutaryl coenzyme A; **DXP**, 1-deoxy-*D*-xylulose 5-phosphate; **MVA**, 3R-Mevalonic acid; **MEP**, 2-*C*-methyl-*D*-erythritol 4-phosphate; **CDP-ME**, 4-(cytidine 5'-diphospho)-2*C*-methyl-*D*-erythritol; **CDP-MEP**, 4-(cytidine 5'-diphospho)-2*C*-methyl-*D*-erythritol 2-phosphate; **cMEPP**, 2*C*-methyl-*D*-erythritol 2,4-cyclodiphosphate; **DMAPP**, Dimethylallyl diphosphate; **HMBPP**, 1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate; **IPP**, isopentenyl diphosphate; **GPP**, geranyl diphosphate; **FPP**, farnesyl diphosphate; **GGPPS**, geranylgeranyl diphosphate. The number of contigs and singletons present in our EST dataset was added between brackets for each enzyme.

listed in Table 3. The poly(AG)n and poly(AAG)n groups were the most abundant motifs in *T. cacao* unigenes.

For each SSR identified, if possible, 3 couples of primers were defined using Primer3 [47]. A total of 5265 flanking sequences were designed and it was possible to define at least one couple of primers for 1755 SSRs.

The exploration of redundant ESTs in contigs was shown to be a valuable resource of Single Nucleotide Polymorphisms (SNP) [48]. SNPs were detected using QualitySNP [49] pipeline from unigene contigs. We assumed that contigs with at least 100 members contained paralogous sequences [50,51] therefore we selected 4818 contigs that contained at least 4 sequences but no more than 100 sequences. A preliminary study assembled 5246 SNPs into 2012 contigs. Transitions (A/T-G/C) represented 54.2% of the SNPs found, transversions 32.1% and InDels 13.7%.

Conclusion

The present assembly of 149,650 *T. cacao* ESTs produced from 56 cDNA libraries constructed from different organs and environmental conditions is the largest transcriptome dataset produced so far for *T. cacao*, and among the largest ones generated for any tree fruit crop. It provides a major resource for *cacao* genetic and functional genomic analyses of important *T. cacao* traits, with the identification and annotation of 48,594 different putative transcripts.

Table 2: Distribution of motifs length in SSRs dataset

Motif Length	Number of SSRs	Frequency
2	1132	50.3
3	857	38.1
4	82	3.6
5	35	1.6
6	14	0.6
compound	132	5.9

The improved knowledge of the *T. cacao* transcriptome will enhance our understanding of main disease resistance mechanisms and will be useful to improve new varieties and establish a sustainable *T. cacao* resistance to pests and diseases. Towards this goal, a large number of cDNA libraries have been produced from *T. cacao*/pathogens or pest interactions, and an important set of unique transcripts homologous to genes known in other species involved in defense and resistance mechanisms have been identified in the whole EST collection using keywords and Gene Ontology tools. It provides a cDNA resource available for the broad scientific community and suitable for cDNA-based microarray analyses.

This collection of ESTs also provides a valuable framework for the discovery of candidate genes involved in chocolate quality traits. Tested for two distinct metabolic pathways, this collection displays a good representation of the *T. cacao* transcriptome involved in quality trait elaboration and will allow the comparative analysis of contrasting genotypes for *T. cacao* qualities to better understand the genetic basis of quality.

This EST collection also will provide a large number of genetic tools, such as SSR and SNP markers, which will be used to construct high density gene maps, facilitating the integration of genetic and genomic approaches to discover the genes that effect trait variations, and also facilitating the sequence assembly in further activities of whole *T. cacao* genome sequencing.

Finally, the assembly and annotation associated will also provide a valuable resource for future investigation of *T.*

Table 3: Distribution of dimers and trimers motifs in SSRs dataset

Group	Motif	Number of SSRs	Frenquency
AC	AC/CA/GT/TG	55	2.4
AG	AG/CT/GA/TC	754	33.5
AT	AT/TA	323	14.3
CG	CG/GC	0	
AAT	AAT/ATA/TAA/ATT/TTA/TAT	121	5.4
AAG	AAG/AGA/GAA/CTT/TTC/TCT	308	13.7
AAC	AAC/ACA/CAA/GTT/TTG/TGT	54	2.4
ATG	ATG/TGA/GAT/CAT/ATC/TCA	117	5.2
AGT	AGT/GTA/TAG/ACT/CTA/TAC	9	0.4
AGG	AGG/GGA/GAG/CCT/CTC/TCC	81	3.6
AGC	AGC/GCA/CAG/GCT/CTG/TGC	79	3.5
ACG	ACG/CGA/GAC/CGT/GTC/TCG	9	0.4
ACC	ACC/CCA/CAC/GGT/GTG/TGG	66	2.9
GGC	GGC/GCG/CGG/GCC/CCG/CGC	13	0.6

cacao evolutionary genomics with related species such as *Gossypium hirsutum* or *Arabidopsis thaliana*.

Methods

Material used for libraries construction

In total, 56 different libraries were constructed. The organs and *T. cacao* genotypes used for cDNA construction, and the treatments carried out on these organs are reported in Table 1.

Most of the libraries were constructed from 2 genotypes:

- Scavina 6 (SCA6) is a self incompatible Forastero genotype originating from the Upper Amazonian region of Peru. SCA6 is highly resistant to *Phytophthora* species and *Moniliophthora perniciosa* diseases. It has been widely used in the breeding programs.

- ICS1 is a self compatible Trinitario genotype, a hybrid involving Criollo, the first *T. cacao* variety domesticated in Central America, and a Forastero variety originated from the Lower Amazonia of Brazil; ICS1 is known for its large beans and good quality traits. This clone was used for RNA production during the different stages of development of the *T. cacao* seeds.

A post harvest treatment is generally applied to *T. cacao* seeds to develop chocolate, involving fermentation steps, drying and torrefaction. Tissues from ICS1 Seeds were collected during the first 2 days of fermentation to construct cDNA libraries.

Other genotypes were used more specifically to represent particular traits or genetic origins:

- Jaca is a Brazilian Forastero genotype from the Upper Amazonian region, and resistant to *Ceratocystis fimbriata*. Inoculation was done according to Silva *et al*. [52]

- B97 C-C-2 is a pure and homozygous Criollo genotype. This material was collected in Belize [53] by a mission conducted by the CRU (Cocoa Research Unit, Univ. West Indies, Trinidad) in conjunction with The Maya Mountain Archaeological Project (MMAP – Cleveland State Univ.) and is now grown in the international collection of CRU.

- GU255V is a genotype originated from French Guyana, resistant to *Phytophthora palmivora*. Inoculation was done according to Tahi *et al*. [54]

- PNG seedlings are from a progeny produced in Papua New Guinea from the cross of two hybrids: 17/3-1 × 36/3-1, and segregating for *Phytophthora resistance*. Inoculation was done according to Tahi *et al*. [54]

- UF676 is a Trinitario genotype tolerant to mirids. Insect attack was done using protocol described by Babin *et al*. [55].

- P7, IMC47, UPA134 are Forastero genotypes originated from the Upper Amazonian region of Peru, known for their resistance to *Phytophthora palmivora or P. megakarya*. Inoculation was done according to Tahi *et al*. [54]

- UF 273 is a Trinitario genotype resistant to *Moniliophthora* rorer. Inoculation was done according to Khun *et al*. [56]

- 33–49 and BE240 are Nacional genotypes from Ecuador known for their aromatic and floral taste.

SSH libraries or direct libraries were constructed from these genotypes. More information related to these genotypes is available through the International Cocoa Germplasm Database [57].

Drought Stress Libraries were constructed from total RNA isolated from leaves and roots of Scavina 6 plants that were initially grown under standard conditions in a greenhouse [58]. Rooted cuttings were generated and grown to about 6 months old, then were moved into a Conviron growth chamber and were not watered until leaves were visibly wilted (approx 36 hours) at which time tissues were flash frozen in liquid nitrogen.

RNA Extraction

Plant tissues were frozen in liquid nitrogen or placed in RNA stabilization reagent (RNA *later*™, Qiagen) and stored at -20°C before RNA extraction. Approximately 100 mg of plant tissues were crushed in liquid nitrogen with poly-vinyl-poly pyrrolidone. The powder was transferred in a tube containing 1 ml of extraction buffer " TE3D " (14.8 g EDTA, 84.4 g Tris, 20 g Nonidet P-40, 30 g lithium dodecyl sulfate, 20 g sodium deoxycholate, 95 ml H2O) [59]. After 15 min incubation at room temperature, 1 ml of sodium acetate (3 M) and one volume of chloroformisoamyl alcohol (24:1) were added. Purification of the aqueous phase was carried out following centrifugation by adding one volume of mixed alkyl tri-ethyl ammonium bromine solution (2% MATAB, 3 M NaCl) followed by 15 min at 74°C. The residual polysaccharides were then eliminated by addition of one volume of chloroformisoamyl alcohol (24:1) and centrifugation; the aqueous phase was precipitated by the addition of one volume of isopropyl alcohol. After centrifugation, the pellet was resuspended in 50 µl of ribonuclease free water containing 1 µl of ribonuclease inhibitor (RiboLock™, Fermentas).

RNA samples from cacao tissues were isolated following the procedure of Charbit *et al* [59] with modifications. Following DNase treatment (DNase I, Fermentas), RNA was then extracted with the phenolchloroformisoamyl alcohol (25:24:1) step and precipitated with one-tenth volume of 3 M sodium acetate, pH 5.3, and 2.5 volumes of 100% ethyl alcohol. An aliquot of RNA was then run by elecrophoresis on a 1.2% agarose gel and stained with ethidium bromide to confirm RNA integrity.

Construction of full-length enriched cDNA library

First strand cDNA were synthesized using the Clontech BD SMART PCR cDNA Synthesis KIT (cat No 634902) as recommended by the supplier. 0.5–1 µg of total RNA was incubated at 72°C for 2 min with 1 µl 3' BD SMART CDS Primer II A (12 µM) and 1 µl BD SMART II A Oligonucleotide (12 µM) in a total volume of 5 µl. Then 2 µl 5× First-Strand Buffer, 1 µl DTT (20 mM); 1 µl dNTP Mix (10 mM of each dNTP), 1 µl BD PowerScript Reverse Transcriptase were added and the mix was incubated at 42°C for 1 hour in an air incubator. According to Glen K Fu (2003) [60], 3 µl Biotin-dATP (Invitrogen), 3 µl Biotin-dCTP (Invitrogen), 1 µl 5'-NVVVVV-3' primer 30 µM (50 ng), 2 µl 5× First-Strand Buffer, 1 µl BD PowerScript Reverse Transcriptase were added, and the mix was kept at 42°C for 30 min. For capture of the unfinished strand, the reaction was mixed with 600 µl of Streptavidine MagneSphere Paramagnetic Particles (Promega) and eluted as recommended by the supplier.

A 2 µl aliquot from the first strand synthesis was used for the cDNA Amplification by LD PCR (Clontech). Each reaction was performed with 80 µl deionized water, 10 µl 10× BD Advantage 2 PCR Buffer, 2 µl 50× dNTP Mix (10 mM of each dNTP), 4 µl 5' PCR Primer II A (12 µM), 2 µl 50× BD Advantage 2 Polymerase Mix in a 98 µl total volume. The PCR reaction consisted of 18 to 25 PCR cycles at 95°C for 15 sec, 65°C for 30 sec, 68°C for 6 min, following with a final extension at 70°c for 10 min.

After comparison of fragment sizes with those of model species (rice and *Arabidopsis*), fragment sizes of some cDNA libraries were improved using cDNA size fractionation. These libraries were submitted to an "agarase step" [61] after 18 cycles PCR. Double-stranded cDNA was separated on 1% low-melting agarose gel and the DNA ladder "lane" was stained and photographed with a ruler. Two size fractions (< 1.2 kb and > 1.2 kb) were excised from the unstained cDNA "lane" based on the DNA ladder "lane". cDNAs were extracted from the gel slices with agarase (Fermentas) according to the supplier instructions. After a gelase digestion, the cDNA was precipitated with one volume of isopropanol. The pellets were dried and suspended in ribonuclease free water. Four to five additional PCR cycles were performed in order to improve the efficiency of ligation in pGEM®-T Easy Vector.

For SSH cDNA libraries: The procedure was performed with the PCR-Select cDNA Substraction kit (Clontech) according to the manufacturer's recommendations with slight modifications. The cDNA generated from the SMART procedure was restricted with 15 U of *Rsa*I (Fermentas) and the two aliquots of the tester cDNA were ligated to adaptors 1 and 2R, respectively, with 30 U of T4 DNA ligase (Fermentas). The PCR mixture enriched for differentially expressed sequences was cloned using pGEMT (Promega) as mentioned above.

One µl of the second strand product was cloned in pGEM®-T Easy Vector Systems (Promega) and transformed by electroporation in the DH10B T1 resistant strain of *Escherichia coli* (Invitrogen); transformation

products were plated on LB-ampicillin agar plates and incubated overnight at 37°C. White colonies were picked using a Qpix 2 XT biorobot (Genetix) and stored in 384 well plates at -80°C.

Sequencing

All clones were end-sequenced using either Forward or Reverse M13 primers. The sequencing reactions were performed with Applied Biosystems BigDye V3.1 kits, and were resolved on ABI3730xl DNA Analysers

Sequence processing

Sequences were managed and stored using our own tool called Expressed Sequence Tag Treatment and Investigation Kit (ESTtik) which is an information system that contains a pipeline for processing, a database and a web site for querying data (Figure 7). The ESTtik pipeline program is a set of Perl packages which contain a main program related to 9 modules in charge of completing different processings. The pipeline executes a series of programs to assess quality and nucleotides from chromatograms, then edits, and assembles the input DNA sequence information into a non-redundant data set. This unigene is then searched for microsatellites and SNPs. It is used as input

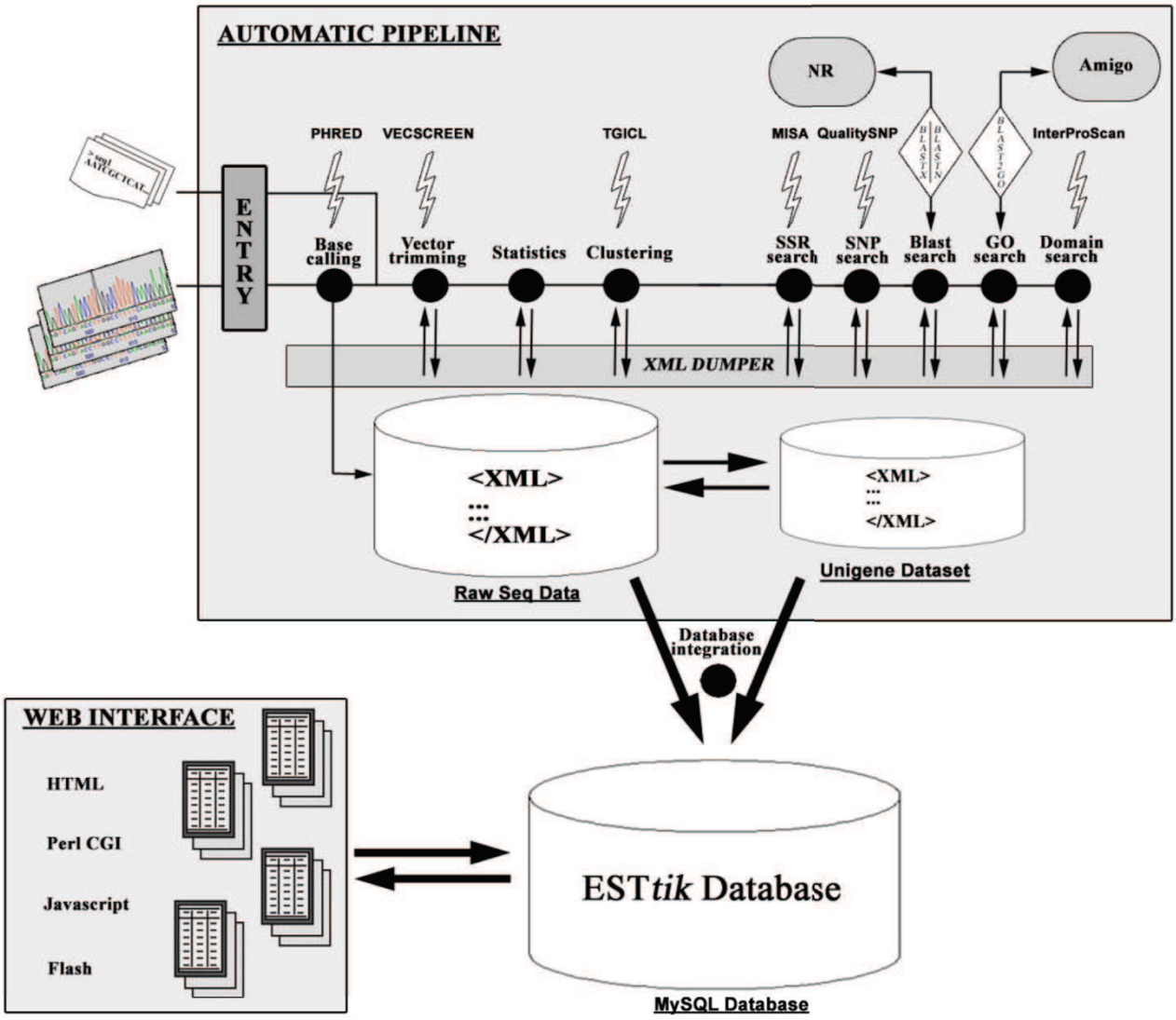


Figure 7
Schematic overview of the ESTtik information System.

for an annotation against public databases including an extraction of Gene Ontology terms [30]. All the results produced by automatic processing are finally stored into XML files. The information collected from individual program modules of the pipeline is stored into a MySQL database. The database model was specially designed using the UML technology to fit data. To visualize Blast [62] results, annotations and to search for sequences by gene keywords or GO terms, the ESTtik database records can be accessed using 7 query pages combining PerlCGI, HTML, Javascript and Flash technologies.

The software Phred [63] was used for base calling linked to Vecscreen [64] for vector and adapters trimming. Cleaning of sequences was performed with the standalone low complexity filter mdust and bioperl modules. Each forward and reverse ESTs were individually assembled with the CAP3 program, using an overlap percent identity cutoff of 65 (p) and an overlap length cutoff of 20 (o).

Special attention has been paid to the global assembly of ESTs, in order to obtain the most representative transcription units. The TGI Clustering tools (TGICL) were used because they provide an optimized protocol for the analysis of EST sequences [65]. This package performs a clustering phase (using megablast) without multiple alignments, and then creates contigs (consensus sequences) with the assembly program CAP3. Many parameters were tested and because we had clusters made of ESTs coming from several highly expressed genes, we increased the clustering and assembly stringency. For the clustering step, we used a minimum percent identity for overlaps (p) of 94, a minimum overlap length (l) of 30, a maximum length of unmatched overhangs (v) of 30. For the assembly, we used a specify overlap percent identity cutoff (p) of 93.

Annotation

Similarity searches were performed with the standalone version 2.2.16 of BLAST [62] against non redundant proteins and nucleotides. The XML Blast output was used and parsing of results was performed with the Bio::SearchIO module of Bioperl toolkit [66].

We built a local Blast2GO MySQL database and we first used the Blast2GO program [29] with default parameters to assign Gene Ontology (GO) terms to the unigenes based on the BLAST definitions. To best exploit GO annotations, results were integrated into a local AmiGO browser and database.

Molecular markers

SSRs searches were performed with MIcroSAtellite identification tool (MISA) [46] and primers designed with Primer3 software [47].

The QualitySNP pipeline [49] was used for detecting single nucleotide polymorphisms in the unigenes.

Data availability

Sequence data, molecular markers and high quality annotation will be integrated into CocoaGen DB [67], a Web portal developed for combining *T. cacao* molecular genetic and genomic information from TropgeneDB [68] and phenotypic data from The International Cocoa Germplasm Database [57]. The individual ESTs of the 56 libraries were deposited in the EMBL database under accession CU469588 to CU633156.

Authors' contributions

XA carried out bioinformatic tasks and drafted the manuscript. OF participated in construction of cDNA libraries. PW carried out EST sequencing. KG contributed to library construction and vegetal material preparation/inoculation. TL contributed to library construction. XS contributed to library construction and replication and sequence analyses. AMR contributed to library construction. CDS contributed to bioinformatic analyses. JC contributed to library construction and vegetal material preparation/inoculation. MA contributed to sequence analyses. DK contributed to library construction and vegetal material preparation/inoculation. JV contributed to library construction and vegetal material preparation/inoculation. BC contributed to bioinformatic analyses. GL contributed to vegetal material preparation/fermentation. RB contributed to vegetal material preparation/inoculation. OS contributed to vegetal material preparation/inoculation. MD contributed to vegetal material preparation/inoculation. MG contributed to library construction. MR contributed to bioinformatic analyses. LA contributed to vitroculture production. RM contributed to vitroculture production. WP contributed to vegetal material preparation/inoculation. RS contributed to library construction. MG participated in general discussion and management. ER participated in general discussion and management. SM contributed to library construction. CL coordinated the Project, drafted the manuscript and participated in library construction.

Acknowledgements

We thank USDA and MARS for their financial support in this project. We also gratefully acknowledge CNRG for having funded and carried out the sequencing work of the project. Finally we wish to thank Renaud Boulanger for critically reading the manuscript.

References

1. Figueira A, Janik J, Goldsbrough P: **Genome size and DNA polymorphism in Theobroma cacao.** *Journal of the American Society for Horticultural Science* 1992, **117:**673-677.
2. Lanaud C, Hamon P, Duperray C: **Estimation of nuclear DNA content of Theobroma cacao L. by flow cytometry.** *Café, Cacao, Thé* 1992, **36:**3-8.

3. Cheesman EE: **Notes on the nomenclature, classification possible and relationships of cocoa populations.** *Tropical Agriculture* 1944, **21:**144-159.
4. Bowers JH, Bailey BA, Hebbar PK, Sanogo S, Lumsden RD: **The impact of plant diseases on world chocolate production.** *Plant Health Progress* 2001.
5. Ampuero E: **Monilia pod rot of cocoa.** *Cocoa Grower's Bulletin* 1967, **9:**1518.
6. Enriquez GA, Brenes O, Delgado JC: **Development and impact of Monilia pod rot of cacao in Costa Rica.** *Proceedings of the 8th International Cocoa Research Conference, Cartagena, Colombia, 18–23 Oct, 1981* 1982:375-380.
7. Guiltinan MJ, Verica JA, Zhang D, Figueira A: **Genomics of Theobroma cacao, the chocolate tree.** *Genomics of Tropical Crop Plants* 2007 in press.
8. Chanliau S, Cros E: **Influence du traitement post-récolte et de la torréfaction sur le développement de l'arôme cacao.** *12th Alliance's Inter Cocoa Conf, Salvador de Bahia (Brazil)* 1996:959-964.
9. Clapperton JF, Yow STK, Chan J, Lim DHK: **Effects of planting materials on flavour.** *Cocoa Growers' Bulletin* 1994, **48:**47-59.
10. Pichersky E, Gang DR: **Genetics and biochemistry of secondary metabolites in plants: An evolutionary perspective.** *Trends Plant Sci* 2000, **205:**439-445.
11. Ziegleder G: **Linalol contents as characteristics of some flavour grade cocoas. Z Lebensm.** *Unters Forsch* 1990, **191:**306-309.
12. Counet C, Ouwerx C, Rosoux D, Collin S: **Relationship between Procyanidin and Flavor Contents of Cocoa Liquors from Different Origins.** *J Agric Food Chem* 2004, **52:**6243-6249.
13. Miller KB, Stuart DA, Smith NL, Lee CY, McHale NL, Flanagan JA, Ou B, Hurst WJ: **Antioxidant activity and polyphenol and procyanidin contents of selected commercially available cocoa-containing and chocolate products in the United States.** *J Agric Food Chem* 2006, **54:**4062-4068.
14. Gu L, House SE, Wu X, Ou B, Prior RL: **Procyanidin and catechin contents and antioxidant capacity of cocoa and chocolate products.** *J Agric Food Chem* 2006, **54:**4057-4061.
15. Stark T, Bareuther S, Hofmann T: **Sensory-guided decomposition of roasted cocoa nibs (Theobroma cacao) and structure determination of taste-active polyphenols.** *J Agric Food Chem* 2005, **53:**5407-5418.
16. Ewing RM, Ben Kahla A, Poirot O, Lopez F, Audic S, Claverie JM: **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression.** *Genome research* 1999, **9(10):**950-959.
17. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science (New York, NY)* 2006, **313(5793):**1596-1604.
18. Terol J, Conesa A, Colmenero JM, Cercos M, Tadeo F, Agusti J, Alos E, Andres F, Soler G, Brumos J, *et al.*: **Analysis of 13000 unique Citrus clusters associated with fruit quality, production and salinity tolerance.** *BMC genomics* 2007, **8:**31.
19. Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, *et al.*: **A global assembly of cotton ESTs.** *Genome research* 2006, **16(3):**441-450.
20. da Silva FG, Iandolino A, Al-Kayal F, Bohlmann MC, Cushman MA, Lim H, Ergul A, Figueroa R, Kabuloglu EK, Osborne C, *et al.*: **Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple Vitis species and development of a compendium of gene expression during berry development.** *Plant physiology* 2005, **139(2):**574-597.
21. Jones PG, Allaway D, Gilmour DM, Harris C, Rankin D, Retzel ER, Jones CA: **Gene discovery and microarray analysis of cacao (Theobroma cacao L.) varieties.** *Planta* 2002, **216(2):**255-264.
22. Leal GALJ, Albuquerque PSB, Figueira A: **Genes differentially expressed in Theobroma cacao associated with resistance to witches' broom disease caused by Crinipellis perniciosa.** *Molecular Plant Pathology* 2007, **8(3):**279-292.
23. Verica JA, Maximova SN, Strem MD, Carlson JE, Bailey BA, Guiltinan MJ: **Isolation of ESTs from cacao (Theobroma cacao L.) leaves treated with inducers of the defense response.** *Plant cell reports* 2004, **23(6):**404-413.
24. Gesteira AS, Micheli F, Carels N, Da Silva AC, Gramacho KP, Schuster I, Macedo JN, Pereira GA, Cascardo JC: **Comparative analysis of expressed genes from cacao meristems infected by Moniliophthora perniciosa.** *Annals of botany* 2007, **100(1):**129-140.
25. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, *et al.*: **EMBL Nucleotide Sequence Database in 2006.** *Nucleic acids research* 2007:D16-20.
26. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, *et al.*: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.** *Bioinformatics (Oxford, England)* 2003, **19(5):**651-652.
27. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic acids research* 2005:D71-74.
28. Zhu XY, Chase MW, Qiu YL, Kong HZ, Dilcher DL, Li JH, Chen ZD: **Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids.** *BMC evolutionary biology* 2007, **7:**217.
29. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21(18):**3674-3676.
30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25(1):**25-29.
31. **AmiGO browser** [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi]
32. Walters D, Newton A, Lyon G: **Induced resistance for plant defence.** Blackwell Publishing; 2007.
33. DeYoung BJ, Innes RW: **Plant NBS-LRR proteins in pathogen sensing and host defense.** *Nat Immunol* 2006, **7(12):**1243-1249.
34. Mishra NS, Tuteja R, Tuteja N: **Signaling through MAP kinase networks in plants.** *Arch Biochem Biophys* 2006, **452(1):**55-68.
35. Wróbel-Kwiatkowska M, Lorenc-Kukula K, Starzycki M, Oszmianski J, Kepczynska E, Szopa J: **Expression of [beta]-1,3-glucanase in flax causes increased resistance to fungi.** *Physiological and Molecular Plant Pathology* 2004, **65(5):**245-256.
36. Wollgast J, Anklam E: **Review on polyphenols in Theobroma cacao: changes in composition during the manufacture of chocolate and methodology for identification and quantification.** *Food Research International* 2000, **33:**423-447.
37. Dreosti IE: **Antioxydant Polyphenols in Tea, Cocoa, and Wine.** *Nutrition* 2000, **16(7/8):**692-694.
38. Othman A, Ismail A, Ghani NA, Adenan I: **Antioxydant capacity and phenolic content of cocoa beans.** *Food Chemistry* 2007, **100:**1523-1530.
39. Steinberg FM, Bearden MM, Keen CL: **Cocoa and chocolate flavonoids: implications for cardiovascular health.** *Journal of the American Dietetic Association* 2003, **103(2):**215-223.
40. Wollgast J, Anklam E: **Polyphenols in chocolate: is there a contribution to human health?** *Food Research International* 2000, **33:**449-459.
41. Winkel-Shirley B: **Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology.** *Plant physiology* 2001, **126(2):**485-493.
42. Djocgoue PF, Boudjeko T, Mbouobda HD, Nankeu DJ, El Hadrami I, Omokolo ND: **Heritability of phenols in the resistance of Theobroma cacao against Phytophthora megakarya, the causal agent of black pod disease.** *Journal of Phytopathology* 2007, **155:**519-525.
43. Chanliau S, Cros E: **Influence du traitement post-récolte et de la torréfaction sur le développement de l'arôme cacao.** *12th International Cocoa Research Conference, Salvador de Bahia (Brazil)* 1996:959-964.
44. Loor RG, Risterucci AM, Fouet O, Courtois B, Amores F, Suarez C, Rosenquist E, Vasco A, Madina M, Lanaud C: **Genetic diversity analysis of the Nacional cacao type from Equator.** *15th International Cocoa Research Conference, San José, Costa Rica* 2006.
45. Cros E: **Cocoa flavor development. Effects of post-harvest processing.** *Manufacturing Confectioner* 1999, **79:**70-77.
46. **MISA – MIcroSAtellite identification tool** [http://pgrc.ipk-gatersleben.de/misa/]

47. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods in molecular biology (Clifton, NJ)* 2000, **132:**365-386.
48. Buetow KH, Edmonson MN, Cassidy AB: **Reliable identification of large numbers of candidate SNPs from public EST data.** *Nature genetics* 1999, **21(3):**323-325.
49. Tang J, Vosman B, Voorrips RE, Linden CG van der, Leunissen JA: **QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species.** *BMC bioinformatics* 2006, **7:**438.
50. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D: **Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data.** *Plant physiology* 2003, **132(1):**84-91.
51. Dantec LL, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio JM, Chaumeil P, Leger P, Garcia V, *et al.*: **Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences.** *Plant molecular biology* 2004, **54(3):**461-470.
52. Silva SDVM, Mandarino EP, Damaceno VO, Santos Filho LP: **Reação de genótipos de cacaueiros a isolados de Ceratocystis cacaofunesta.** *Fitopatologia Brasileira* 2007, **32:**504-506.
53. Mooleedhar V: **A Study of the Morphological Variation in a Relic Criollo Cacao Population from Belize.** *Annual report CRU/ The University of West Indies* 1997:5-14.
54. Tahi M, Kebe I, Eskes AB, Ouattara S, Sangaré A, Mondeil F: **Rapid screening of cacao genotypes for field resistance to Phytophthora palmivora using leaves, twigs and roots.** *Eur J Plant Pathol* 2000, **106:**87-94.
55. Babin R, Sounigo O, Dibog L, Nyassé S: **Field tests for antixenosis and tolerance of cocoa towards mirids.** *Ingenic Newsletter* 2004, **9:**45-50.
56. Kuhn DN, MacArthur HC, Nakamura K, Borrone JW, Schnell RJ, Brown JS, Johnson ES, Phillips-Mora W: **Development of molecular genetic markers from a cDNA subtraction library of frosty pod inoculated cacao.** In *15th International Cocoa Research Conference: 2006* San Jose, Costa Rica; 2006:179-184.
57. **The International Cocoa Germplasm Database** [http://www.icgd.rdg.ac.uk/]
58. Maximova S, Miller C, Antunez de Mayolo G, Pishak S, Young A, Guiltinan MJ: **Stable transformation of Theobroma cacao L. and influence of matrix attachment regions on GFP expression.** *Plant cell reports* 2003, **21(9):**872-883.
59. Charbit E, Legavre T, Lardet L, Bourgeois E, Ferriere N, Carron M: **Identification of differentially expressed cDNA sequences and histological characteristics of Hevea brasiliensis calli in relation to their embryogenic and regenerative capacities.** *Plant cell reports* 2004, **8:**539-548.
60. Fu GK, Stuve LL: **Improved method for the construction of full-length enriched cDNA libraries.** *Biotechnique* 2003, **34:**954-957.
61. Wellenreuther R, Schupp I, The German cDNA Consortium, Poustka A, Wiemann S: **SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones.** *BMC genomics* 2004, **5:**36-44.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215(3):**403-410.
63. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome research* 1998, **8(3):**175-185.
64. **Vecscreen** [http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html]
65. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **An optimized protocol for analysis of EST sequences.** *Nucleic acids research* 2000, **28(18):**3657-3665.
66. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al.*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome research* 2002, **12(10):**1611-1618.
67. **CocoaGen DB** [http://cocoagendb.cirad.fr]
68. Ruiz M, Rouard M, Raboin LM, Lartaud M, Lagoda P, Courtois B: **TropGENE-DB, a multi-tropical crop information system.** *Nucleic acids research* 2004:D364-367.

3. Perspectives

L'analyse de cette importante collection d'ESTs du cacaoyer a permis l'identification de nombreux marqueurs moléculaires (SSR et SNP), qui ont été utilisés pour la construction de nouvelles cartes génétiques. Ces marqueurs moléculaires sont situés proches ou à l'intérieur des régions codantes des gènes où ils ont été identifiés, apportant une plus value importante pour les analyses génétiques des populations.

Fouet et al. ont publié en 2011 une nouvelle version de la carte génétique de référence UPA402 x UF676 (Pugh et al., 2004; Risterucci et al., 2000), comprenant 115 nouveaux marqueurs microsatellites positionnés dans des gènes candidats pour les caractères de résistance ou de qualité du cacaoyer. Cette version de la carté génétique de référence a porté le nombre de marqueurs moléculaires à 582 soit un intervalle entre 2 marqueurs de 1,3 cM.

Cette carte génétique de référence a été complétée par l'analyse du polymorphisme nucléotidique entre génotypes dans la collection d'ESTs (Allegre et al., 2012). Un sous ensemble de 1536 marqueurs SNPs, positionnés dans des gènes ayant une annotation fonctionnelle, a été sélectionné pour effecture le génotypage de plusieurs populations par la technologie Illumina® GoldenGate. L'étude du polymorphismes de ces marqueurs dans une collection de ressurces génétiques du cacaoyer et leur ségrégation dans la descendance de référence a permis de cartographier 681 nouveaux marqueurs moléculaires de type SNPs. Combinés avec 163 nouveaux marqueurs microsatellites identifiés dans les premières séquences génomiques de la variété Criollo, cette étude porte à 1,259 le nombre de marqueurs génétiques de la carte génétique de référence. La distance moyenne entre 2 markers a été réduite à 0,7 cM. Cette version de la carte génétique a été utilisée en 2010 pour l'ancrage des séquences des scaffolds sur les 10 chromosomes du cacaoyer lors de la production de la première séquence du génome du Criollo.

Enfin, les séquences de gènes exprimés ont été des outils précieux pour repérer les gènes le long de la séquence du génome. En effet, ces séquences d'ADNc révèlent la structure complète des gènes, notamment les régions non traduites mais transcrites et peuvent également montrer l'existence d'exons "alternatifs" conduisant par épissage alternatif à des protéines différentes à partir d'un même gène.

CHAPITRE 2 - LE SÉQUENÇAGE DU GENOME DU CACAOYER

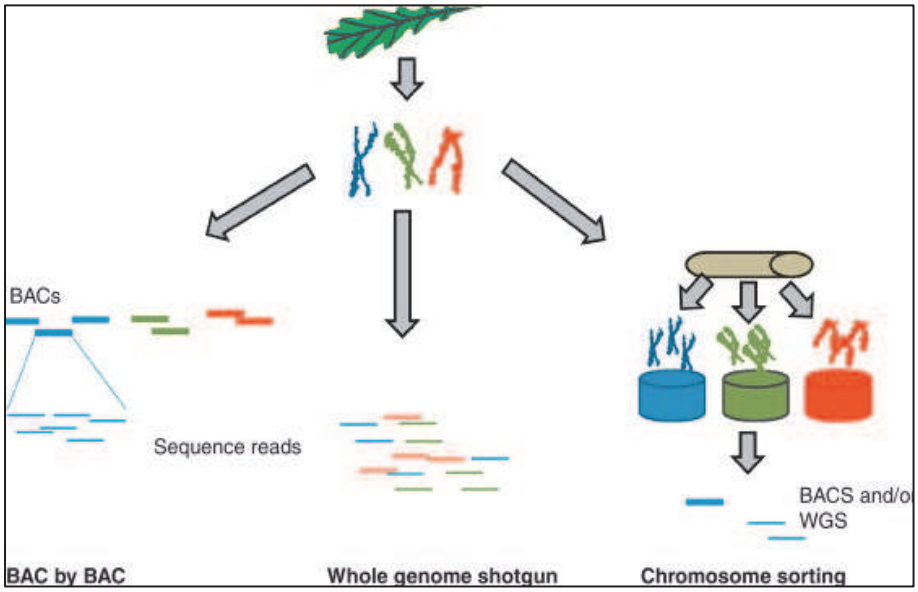


Figure 15. Stratégies de séquençage et d'assemblage (d'après Bolger et al., 2013).

Dans le cadre d'une stratégie BAC à BAC (à gauche sur la figure), le génome est scindé selon un chemin de recouvrement minimal représenté par des BACs qui sont séquencés. Dans une approche par WGS (au centre), le génome complet est fragmenté, séquencé et assemblé. Une nouvelle technique plus récente permet de séparer les chromosomes (à droite sur la figure), réduisant ainsi la complexité du génome à assembler.

1. Introduction

Le séquençage du génome des premiers organismes vivants a débuté il y a plus de 20 ans par la première bactérie (Fleischmann et al., 1995). Le premier génome eucaryote a été publié en 1996 (Goffeau et al.), le premier organisme eucaryote multicellulaire en 1998 (The *C. elegans* Sequencing Consortium) et enfin les première plantes avec *Arabidopsis thaliana* en 2000 (The *Arabidopsis* Genome Initiative) et le riz (*Oryza sativa*) en 2005 (International Rice Genome Sequencing Project). Ces premiers génomes ont été séquencés en utilisant des séquenceurs capillaires de première génération et avec une approche BAC à BAC (Figure 15), consistant à trouver un chemin de recouvrement minimal entre les BACs et à les séquencer complètement et les assembler. La réitération du processus permet de couvrir totalement les bras des chromosomes.

Une autre technique dite Whole Genome Shotgun (WGS), consiste à séquencer avec une grande couverture l'ADN génomique fragmenté, assembler les lectures chevauchantes en contigs, relier les contigs par des séquences pairées en Scaffolds et enfin grouper et ordonner les scaffolds en chromosomes à l'aide d'une carte génétique. Le séquençage de plusieurs plantes a été réalisé par ces techniques de WGS dont le peuplier (Tuskan et al., 2006) et la vigne (Jaillon et al., 2007; Velasco et al., 2007). Cette technique est plus simple à mettre en œuvre mais nécessite des algorithmes d'assemblage performant et conduit à de nombreux trous dans le génome, soit une "finition" généralement moins bonne que par les approches BAC à BAC.

En amélioration des plantes, la séquence du génome apporte un ensemble d'informations biologiques cruciales, depuis le catalogue des gènes (gènes codant des protéines ou petits ARNs, élément mobiles, etc...) jusqu'aux aspects évolutifs (duplications, synténie, phylogénie, etc...). Un des impacts majeurs que fourni la séquence du génome est la possibilité d'obtenir un marquage moléculaire très haute densité pouvant être utilisé pour cartographier des caractères agronomiques d'intérêt et identifier des gènes candidats dans les régions cartographiées. Les marqueurs moléculaires développés le long du génome du riz ont par exemple été rapidement utilisés après la publication du génome pour caractériser un gène majeur régulant la production des graines (Ashikari et al., 2005).

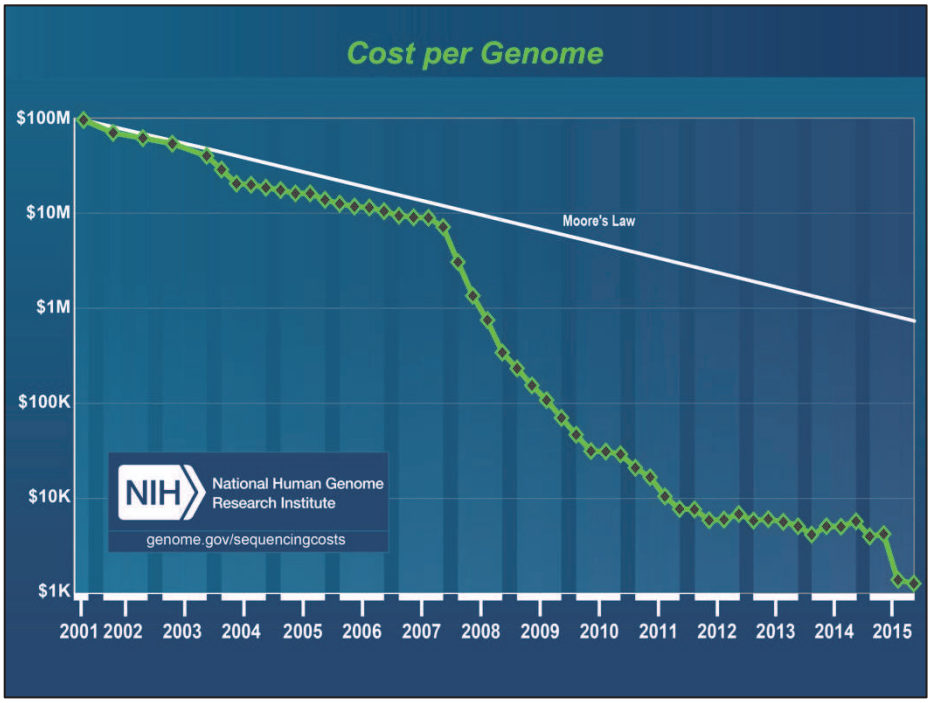


Figure 16 : Evolution du coût de séquençage d'un génome (source NCBI)

Ce type d'approche a été également utilisé chez le maïs pour cartographier des QTLs impliqués dans la biomasse et la bioénergie (Riedelsheimer et al., 2012). Chez la tomate, la séquence du génome de référence a été utilisée pour identifier une famille de gènes (estérases) impliquée dans la variation des composés aromatiques chez différents cultivars. L'analyse de la structure du génome a révélé des insertions de retrotransposons proches des estérases, favorisant leur expression et diminuant ainsi la présence d'esters dans les composés aromatiques (Goulet et al., 2012). De nombreux autres exemples d'identification de mécanismes moléculaires impliqués dans des caractères agronomiques d'intérêt chez diverses plantes sont à mettre au crédit de la séquence du génome. Parmi ceux-ci, nous pouvons citer l'identification de gènes régulant la floraison (Huang et al., 2012; Pin et al., 2010; Xia et al., 2012) ou la différentiation de certains organes comme le tubercule de la pomme de terre (Kloosterman et al., 2013).

Chez le cacaoyer, l'amélioration variétale doit relever plusieurs défis. Les producteurs sont confrontés à une menace croissante des maladies et attaques d'insectes qui compromettent la production de cacao. Dans des conditions climatiques perturbées, la production plus rapide de variétés adaptées aux nouvelles conditions environnementales est nécessaire. Par ailleurs, dans un contexte sociétal où les qualités nutritionnelles et organoleptiques des aliments sont de plus en plus centrales, l'acquisition de connaissances fondamentales sur les déterminants de la qualité sont stratégiques. Dans ces circonstances, la séquence du génome complet du cacaoyer est une aide qui doit permettre, d'une part, de mieux comprendre les déterminants génétiques impliqués dans les variations des caractères agronomiques d'intérêt et, d'autre part, de développer des outils exploitables dans les programmes de sélection .

En 2008, l'importante ressource de données moléculaires disponibles (cartes génétiques denses, transcriptome, banque BAC) et l'existence d'un génotype Criollo homozygote ont permis d'envisager le séquençage du génome du cacaoyer. Combiné à un large partenariat (International Cocoa Genome Sequencing consortium) et à une baisse des coûts apportée par les séquenceurs de 2ème génération (Figure 16), le projet de séquençage du génome complet du cacaoyer a débuté en 2009. Les résultats de ce travail sont présentés en détail dans l'article scientifique suivant, publié dans la revue Nature Genetics en 2011 et intitulé : "**The genome of *Theobroma cacao***".

ARTICLES

nature genetics

The genome of *Theobroma cacao*

Xavier Argout[1,24], Jerome Salse[2,24], Jean-Marc Aury[3–5,24], Mark J Guiltinan[6,7,24], Gaetan Droc[1], Jerome Gouzy[8], Mathilde Allegre[1], Cristian Chaparro[9], Thierry Legavre[1], Siela N Maximova[6], Michael Abrouk[2], Florent Murat[2], Olivier Fouet[1], Julie Poulain[3–5], Manuel Ruiz[1], Yolande Roguet[1], Maguy Rodier-Goud[1], Jose Fernandes Barbosa-Neto[9], Francois Sabot[9], Dave Kudrna[10], Jetty Siva S Ammiraju[10], Stephan C Schuster[11], John E Carlson[12,13], Erika Sallet[8], Thomas Schiex[14], Anne Dievart[1], Melissa Kramer[15], Laura Gelley[15], Zi Shi[7], Aurélie Bérard[16], Christopher Viot[1], Michel Boccara[1], Ange Marie Risterucci[1], Valentin Guignon[1], Xavier Sabau[1], Michael J Axtell[17], Zhaorong Ma[17], Yufan Zhang[15,7], Spencer Brown[18], Mickael Bourge[18], Wolfgang Golser[10], Xiang Song[10], Didier Clement[1], Ronan Rivallan[1], Mathias Tahi[19], Joseph Moroh Akaza[19], Bertrand Pitollat[1], Karina Gramacho[20], Angélique D'Hont[1], Dominique Brunel[16], Diogenes Infante[21], Ismael Kebe[18], Pierre Costet[22], Rod Wing[10], W Richard McCombie[15], Emmanuel Guiderdoni[1], Francis Quetier[23], Olivier Panaud[9], Patrick Wincker[3–5], Stephanie Bocs[1] & Claire Lanaud[1]

We sequenced and assembled the draft genome of *Theobroma cacao*, an economically important tropical-fruit tree crop that is the source of chocolate. This assembly corresponds to 76% of the estimated genome size and contains almost all previously described genes, with 82% of these genes anchored on the 10 *T. cacao* chromosomes. Analysis of this sequence information highlighted specific expansion of some gene families during evolution, for example, flavonoid-related genes. It also provides a major source of candidate genes for *T. cacao* improvement. Based on the inferred paleohistory of the *T. cacao* genome, we propose an evolutionary scenario whereby the ten *T. cacao* chromosomes were shaped from an ancestor through eleven chromosome fusions.

Theobroma cacao L. is a diploid tree fruit species ($2n = 2x = 20$ (ref. 1)) endemic to the South American rainforests. Cocoa was domesticated approximately 3,000 years ago[2] in Central America[3]. The Criollo cocoa variety, having a nearly unique and homozygous genotype, was among the first to be cultivated[4]. Criollo is now one of the two cocoa varieties providing fine flavor chocolate.

However, due to its poor agronomic performance and disease susceptibility, more vigorous hybrids created with foreign (Forastero) genotypes have been introduced. These hybrids, named Trinitario, are now widely cultivated[5]. Here we report the sequence of a Belizean Criollo plant[6].

Consumers have shown an increased interest for high-quality chocolate, and for dark chocolate, containing a higher percentage of cocoa[7]. Fine-cocoa production is nevertheless estimated to be less than 5% of the world cocoa production due to the low productivity and disease susceptibility of the traditional fine-flavor cocoa varieties. Therefore, breeding of improved Criollo varieties is important for sustainable production of fine-flavor cocoa.

3.7 million tons of cocoa are produced annually (see URLs). However, fungal, oomycete and viral diseases, as well as insect pests, are responsible for an estimated 30% of harvest losses (see URLs). Like many other tropical crops, knowledge of *T. cacao* genetics and genomics is limited. To accelerate progress in cocoa breeding and the understanding of its biochemistry, we sequenced and analyzed the genome of a Belizean Criollo genotype (B97-61/B2). This genotype is suitable

[1]Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)-Biological Systems Department-Unité Mixte de Recherche Développement et Amélioration des Plantes (UMR DAP) TA A 96/03-34398, Montpellier, France. [2]Institut National de la Recherché Agronomique UMR 1095, Clermont-Ferrand, France. [3]Commissariat à l'Energie Antomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France. [4]Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France. [5]Université d'Evry, Evry, France. [6]Penn State University, Department of Horticulture and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [7]Penn State University, Plant Biology Graduate Program and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [8]Institut National de la Recherche Agronomique (INRA)-CNRS Laboratoire des Interactions Plantes Micro-organismes (LIPM), Castanet Tolosan Cedex, France. [9]UMR 5096 CNRS-Institut de Recherche pour le Développement (IRD)-Université de Perpignan Via Domitia (UPVD), Laboratoire Génome et Développement des Plantes, Perpignan Cedex, France. [10]Arizona Genomics Institute and School of Plant Sciences, University of Arizona, Tucson, Arizona, USA. [11]Penn State University, Department of Biochemistry and Molecular Biology, University Park, Pennsylvania, USA. [12]Penn State University, the School of Forest Resources and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [13]The Department of Bioenergy Science and Technology (WCU), Chonnam National University, Buk-Gu, Gwangju, Korea. [14]Unité de Biométrie et d'Intelligence Artificielle (UBIA), UR875 INRA, Castanet Tolosan, France. [15]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [16]INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, CP5724, Evry, France. [17]Penn State University, Bioinformatics and Genomics PhD Program and Department of Biology, University Park, Pennsylvania, USA. [18]Institut des Sciences du Végétal, UPR 2355, CNRS, Gif-Sur-Ivette, France. [19]Centre National de la Recherche Agronomique (CNRA), Divo, Côte d'Ivoire. [20]Comissão Executiva de Planejamento da Lavoura Cacaueira (CEPLAC), Itabuna Bahia, Brazil. [21]Centro Nacional de Biotecnología Agrícola, Instituto de Estudios Avanzados (IDEA), Caracas, Venezuela. [22]Chocolaterie VALRHONA, Tain l'Hermitage, France. [23]Département de Biologie, Université d'Evry Val d'Essonne, Evry, France. [24]These authors contributed equally to this work. Correspondence should be addressed to X.A. (xavier.argout@cirad.fr).

Received 10 August; accepted 1 December; published online 26 December 2010; doi:10.1038/ng.736

Table 1 Global statistics of the genome assembly and annotation of *Theobroma cacao*

Assembly		Number	N50 (kb)	Longest (kb)	Size (Mb)	Percentage of the assembly
Contigs	All	25,912	19.8	190	291.4	–
	All	4,792	473.8	3,415	326.9	100
Scaffolds	Anchored on chromosomes	385		3,415	218.4	66.8
	Anchored on chromosomes and oriented	206		3,415	162.8	49.8
Annotation		Number				
	Protein coding	28,798			96.4	29.4
	rRNA	6[a]			<0.03	<0.01
Genes	tRNA	473			<0.03	<0.01
	miRNA	83			<0.03	<0.01
	Transposable Element	17,342			52.6	16.1
Transposable elements		67,575			84.0	25.7

[a]The rRNA number is greatly underestimated due to the sequencing method (**Supplementary Note**).

for a high-quality genome sequence assembly because it is highly homozygous as a result of the many generations of self-fertilization that occurred during the domestication process.

RESULTS

Sequencing and assembly

We used a genome-wide shotgun strategy incorporating Roche/454, Illumina and Sanger sequencing technologies. The International Cocoa Genome Sequencing consortium (ICGS) produced a total of 17.6 million 454 single reads, 8.8 million 454 paired end reads, 398.0 million Illumina paired end reads and about 88,000 Sanger BAC end reads, corresponding to 26 Gb of raw data (**Supplementary Note**, **Supplementary Tables 1** and **2** and **Supplementary Figs. 1** and **2**). We used the Roche/454 and Sanger raw data to produce the assembly. This represented ×16.7 coverage of the 430-Mb genome of B97-61/B2, whose size was estimated by flow cytometry (**Supplementary Note** and **Supplementary Table 3**).

This assembly, performed with Newbler software (Roche, Inc.), consists of 25,912 contigs and 4,792 scaffolds (**Table 1**, **Supplementary Note** and **Supplementary Table 4**). Eighty percent of the assembly is in 542 scaffolds, and the largest scaffold measures 3.4 Mb. We determined the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) to be 473.8 kb. The total length of the assembly was 326.9 Mb, which represents 76% of the estimated genome size of the *T. cacao* genotype B97-61/B2 (430 Mb). In addition, we used a high coverage of Illumina data (×44 coverage of the genome), which has a different error profile than 454 data, to improve accuracy of the *T. cacao* genome sequence (Online Methods).

The resulting assembly appears to cover a very large proportion of the euchromatin of the *T. cacao* genome. We confirmed the high genome coverage of this assembly by comparing it to the unigene resource (38,737 unigenes assembled from 715,457 expressed sequence tag (EST) sequences from the B97-61/B2 genotype). We recovered 97.8% of the unigene resource in the *T. cacao* genome assembly.

Using a set of 1,259 molecular markers from a consensus genetic linkage map established from two mapping populations, 94% of the markers (1,192) were unambiguously located on the assembly, allowing us to anchor 67% of the assembled 326 Mb along the ten *T. cacao* linkage groups. The remaining 33% of the genome assembly was in 2,207 scaffolds. Fifty percent of the assembly could be anchored and oriented (**Supplementary Note**, **Supplementary Table 5** and **Supplementary Fig. 3**). The average ratio of genetic-to-physical distance was 1 cM per 444 kb in centromeric regions and 1 cM per 146 kb in distal chromosomal regions.

Gene content and repeated sequences annotation

We performed the identification and annotation of transposable elements using a two step approach: the first approach was based on the *de novo* identification of transposable elements from the assembled scaffolds and the second one was based on the search for transposable elements from the unassembled reads (**Supplementary Note**).

This *de novo* search lead to the identification of a total of 67,575 transposable element–related sequences in the assembled cocoa sequences (**Table 1**). The second step led to the identification of three highly repeated transposable element families from the unassembled reads dataset. The most common transposable element was a long terminal repeat (LTR) retrotransposon that we named Gaucho. It is a Copia-like element 11,297 bp in length that is repeated approximately 1,100 times, based upon its occurrences in the 454 unassembled sequences. Fluorescence *in situ* hybridization (FISH) analysis revealed that Gaucho is distributed on all chromosome arms but is found mainly in their median region (as opposed to the centromeric and telomeric regions), a classical feature shared by many LTR retrotransposons in plants (**Fig. 1**). The other two highly repeated families were another Copia-like LTR retrotransposon and a Mu-type transposon.

Class I elements were the most abundant, representing 69% of the total transposable elements in the cocoa genome, with a total of 290 Gypsy-like and 159 Copia-like families. In addition, we identified 36 transposons and 1,353 miniature inverted-repeat transposon element (MITE) families (class II) (**Supplementary Table 6**). The most highly repeated transposon families were Mutator and Vandal (Mu type), with copy numbers of 994 and 1,978, respectively. Transposable elements were particularly abundant in centromeric regions, as illustrated in **Figure 2**; this feature was already observed in other sequenced genomes[8].

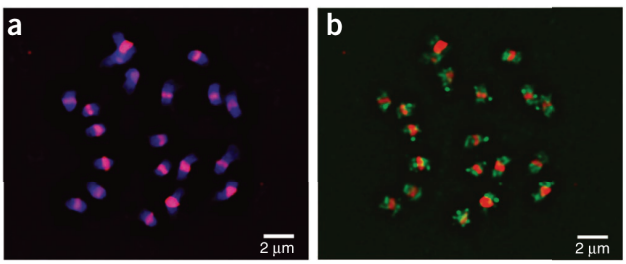


Figure 1 FISH analysis of *T. cacao* chromosomes. (**a**) *In situ* hybridization of *T. cacao* chromosomes stained with DAPI (blue) using a ThCen repeat probe (red). (**b**) *In situ* hybridization using Gaucho LTR retrotransposon (green) and ThCen repeat (red) probes.

Altogether, the transposable elements identified in both assembled (84 Mb) and un-assembled (20.3 Mb) reads represent about 24% of the *T. cacao* genome. This value is substantially lower than that for other sequenced genomes of similar size, for example, rice (35%, for 380 Mb)[9] and grape (41.4%, for 475 Mb)[10]. However, sequencing and assembling of highly repeated sequences can be expected to be the major limitation of *de novo* sequencing of a complex genome using next-generation sequencing. This is particularly true for transposon element families that have undergone very recent amplification, like in the case of Gaucho. Therefore, we conclude that the total contribution of repetitive elements to the whole cocoa genome may be underestimated.

In addition, we identified a tandem repeat sequence (that we named ThCen) from the 454 repeat reads. This tandem repeat is 212 bp long and, when used as probe in a FISH experiment, was found to be located in the centromeres of all cocoa chromosomes (**Fig. 1**). Tandem repeats and retrotransposons are the major components of plant centromeres[11]. The copy number of Gaucho and ThCen repeat sequences varied up to 2.5-fold among *T. cacao* genotypes from various genetic origins. We observed a positive Pearson correlation (r = 0.56) between genome size and ThCen repeat copy number, suggesting a possible contribution of ThCen repeats in genome size variation (**Supplementary Note**, **Supplementary Fig. 4** and **Supplementary Table 7**).

We annotated the genome sequence using the integrative gene prediction package EUGene[12] following specific training for *T. cacao* (**Supplementary Note**). Homology searching and functional annotation (**Supplementary Fig. 5**) led to the identification of 28,798 *T. cacao* protein-coding genes (**Table 1**), with an average gene size of 3,346 bp and a mean of 5.03 exons per gene (**Supplementary Table 8**). Compared to the smaller *Arabidopsis thaliana* genome, the *T. cacao* genome has a higher gene number, a similar exon number per gene and a lower mean gene density per 100 kb (**Supplementary Table 8**). The genes in *T. cacao* were more abundant in subtelomeric regions (**Fig. 2**), as previously observed in other sequenced plant genomes[13].

The comparison of cocoa, *A. thaliana*, grape, soybean and poplar proteomes revealed 6,362 clusters of genes (totaling 52,176 genes) distributed among all five eudicot genomes and 682 gene families (totaling 2,053 genes) specific to the cocoa genome (**Fig. 3**, **Supplementary Note** and **Supplementary Table 9**). Most of these 682 gene families encode hypothetical proteins, as supported at the transcript level. The functional analysis of these five proteomes using gene ontology terms revealed a similar pattern among them (**Supplementary Note** and **Supplementary Fig. 6**). A specific feature of the *T. cacao* clusters common to the other species was the relatively high level of metabolic and cellular processes (**Supplementary Note** and **Supplementary Figs. 6** and **7**). On the other hand, grape- and cocoa-specific clusters showed the highest unknown-function percentages (**Supplementary Note** and **Supplementary Fig. 7**).

MicroRNAs (miRNAs) are short noncoding RNAs that regulate target genes transcriptionally or post-transcriptionally. Many of them play important roles in development and stress responses[14]. A total of 83 *T. cacao* miRNAs from 25 families were computationally predicted based on sequence similarity to known plant miRNAs in

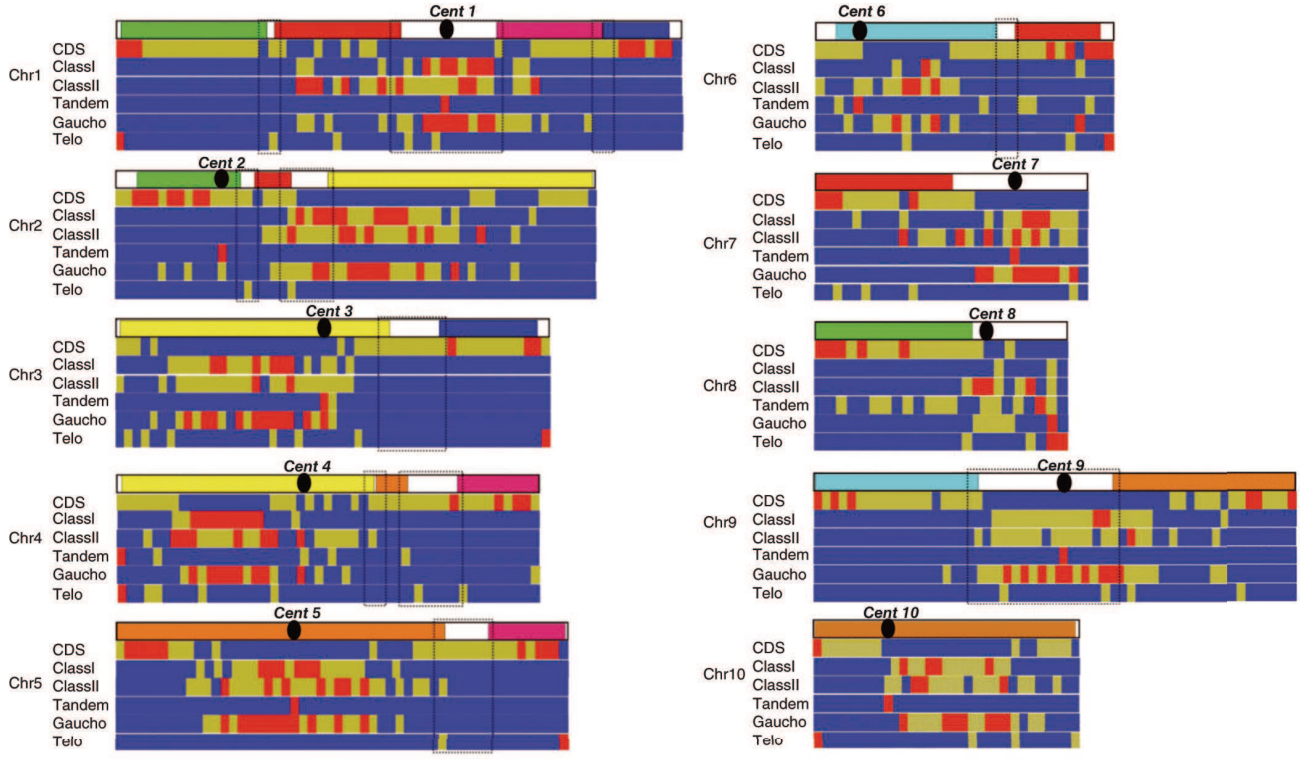


Figure 2 *T. cacao* genome heat map. The ten *T. cacao* chromosomes harboring 11 chromosome fusions (in black dotted boxes) identified in these genomes are illustrated according to their ancestral chromosomal origin (see paleo-chromosome color code in **Fig. 4**). Centromeres are marked 'Cent'. For the ten chromosomes, heat maps are provided for the CDS (blue <60%, yellow 60%–90% and red >90%), class I and II transposable elements (blue <80%, yellow >80% and red ~100%), ThCen and Gaucho elements (blue <50% of maximum, yellow ≥50% of maximum and red = maximum) and telomeric repeats (blue = 0, yellow <40% and red >40%). Only the elements present in the assembled part of the genome are represented. Therefore, the genome distribution of the repeated sequences represented in this figure could be biased due to the major limitations of *de novo* sequencing of complex genomes using next-generation sequencing (NGS), which is limited in its ability to assemble highly repeated sequences.

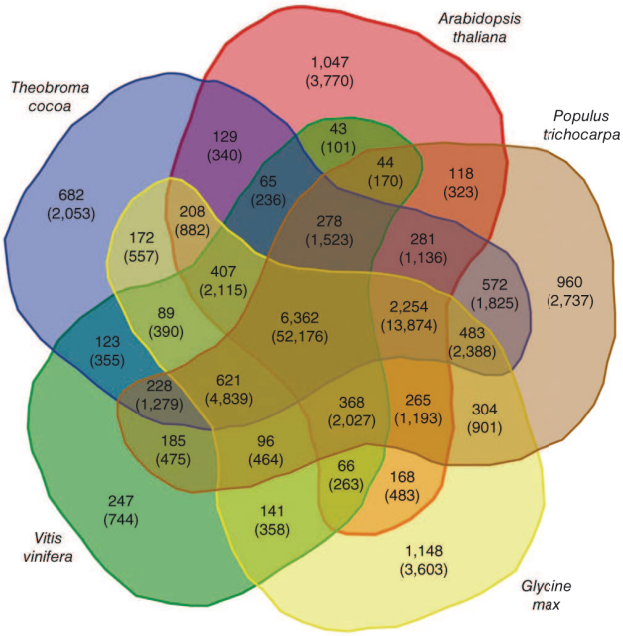


Figure 3 Venn diagram showing the distribution of shared gene families among *Theobroma cacao*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Glycine max* and *Vitis vinifera*. Numbers in parentheses indicate the number of genes in each cluster. The Venn diagram was created with web tools provided by the Bioinformatics and Systems Biology of Gent (see URLs).

miRBase 14 (ref. 15) (**Supplementary Note**, **Supplementary Table 10** and **Supplementary Figs. 8** and **9**). Ninety-one *T. cacao* miRNA targets were predicted. Most predicted targets were homologous to known miRNA targets in other plant species, but there was a profound bias toward putative transcription factors compared to the other species (**Supplementary Table 11**), suggesting that miRNAs are major regulators of gene expression in *T. cacao*.

Disease resistance–related genes

Fungal and oomycete diseases are a major constraint to world cocoa production, and the search for natural disease resistance is one of the main objectives of all *T. cacao* breeding programs. Resistance genes (*R* genes) are divided into 2 classes: *NBS-LRR*, the nucleotide-binding site leucine-rich repeat class of genes, and *RPK*, the receptor protein kinase class of genes[16].

Within the *RPK* class, one family of plant-specific transmembrane receptors was shown to also possess *LRR*-class genes in its extracellular domain and have important roles in defense responses or in plant development[17]. The *LRR-RLK* family consists of more than 200 members in *A. thaliana* and more than 400 members in poplar[18]. Here we show that the *T. cacao* genome contains at least 253 *LRR-RLK* genes orthologous to *Arabidopsis LRR-RLK* genes (**Supplementary Note**). Reports indicate that the *LRR-RLK* family is divided into 19 subfamilies[18]. In Viridiplantae, some of these subfamilies have expanded dramatically. As in *Populus trichocarpa*, the *LRR-XII* subfamily has greatly expanded in the *T. cacao* genome (**Supplementary Note**, **Supplementary Table 12** and **Supplementary Table 13**), with approximately 36% of the *LRR* genes belonging to this subfamily.

The *NBS* genes, encoding nucleotide-binding site proteins, also play an important role in resistance to pathogens and in the cell cycle[19]. The *NBS* gene family is rather abundant in plant genomes, ranging from 0.6% to approximately 2% of the total gene number (**Supplementary Note**).

We identified a total of 297 non-redundant *NBS*-encoding orthologous genes in the *T. cacao* genome (**Supplementary Note**, **Supplementary Table 14** and **Supplementary Figs. 10** and **11**). Among these genes, one class, characterized by the *TIR* (encoding the toll interleukin receptor) motif, is markedly underrepresented in the *T. cacao* genome compared to other eudicot plants, with only 4% of *NBS* orthologous genes containing *TIR* motifs, in contrast to grape, poplar, Medicago (20%) or *Arabidopsis* (65%) (**Supplementary Table 14**). The *TIR* motifs have been shown to be present in basal angiosperms and eudicots, but are nearly absent in monocots[20]. It has been suggested that the *TIR-NBS-LRR* resistance genes are more ancient than the divergence of angiosperm and gymnosperms and that they have been lost in the cereal genomes[21]. Their lower level in the cocoa genome compared to other eudicots, and the close relatedness of cocoa to a common eudicot ancestor as shown by paleohistory studies (see below), suggests a divergent evolution of *NBS-TIR* orthologous cocoa genes from an ancestral locus, leading to a lower expansion of this gene family in *T. cacao*.

Another gene family that plays a major role in plant defense is the *NPR* gene family. *NPR1* is an *Arabidopsis* BTB/POZ domain protein that acts as a central mediator of the plant defense signal transduction pathway[22]. We surveyed the *T. cacao* genome sequence and found four related *T. cacao* orthologous genes corresponding to each of the *NPR1* subfamilies found in *Arabidopsis* (**Supplementary Note** and **Supplementary Fig. 12**). Recently, we showed that one of these genes (located on chromosome 9) is a functional ortholog of *Arabidopsis NPR1* by transgenic complementation[23].

We mapped the *NBS*, *LRR-LRK* and *NPR1* orthologous genes along the *T. cacao* pseudomolecules (**Supplementary Note** and **Supplementary Fig. 13**). They were distributed across the ten chromosomes, with a large number being organized in clusters, as is classically observed for these classes of genes[24].

A meta-analysis of quantitative trait loci (QTL) related to disease resistance previously identified in *T. cacao* was recently done[25]. We compared the QTL genetic localizations found in this previous study with the distribution of *NBS-LRR*, *LRR-LRK* and *NPR* orthologous cocoa genes (**Supplementary Fig. 13**). Considering an average confidence interval of about 20 cM for the 76 QTLs identified[25], most of the QTLs are located in genome regions containing candidate resistance genes. However, due to the fact that a large number of QTLs and candidate resistance genes are widespread across the genome, many colocalizations may have occurred at random. Therefore, the candidate genes potentially underlying QTLs need to be further studied by functional genomics approaches to confirm their potential roles in disease resistance in cocoa.

Genes potentially involved in cocoa qualities

T. cacao seeds are fermented, dried and then processed into cocoa mass (ground, dehusked seeds), cocoa butter (triacylglycerol storage lipids from cotyledonary endosperm cells) and cocoa powder (defatted mass consisting primarily of cell walls, endosperm storage proteins, starch and proanthocyanins, as well as other flavonoids, aromatic terpenes, theobromine and many other metabolites). In order to characterize the gene families involved in cocoa quality traits, we used a translational approach to survey the *T. cacao* genome using molecular and biochemical knowledge from *Arabidopsis*, poplar, grape and other model plant species.

Oils, proteins, starch and various secondary metabolites such as flavonoids, alkaloids and terpenoids comprise the principal molecular components of cocoa affecting flavor and quality. Storage lipids (triacylglycerols) provide carbon and energy reserves for germinating cocoa embryos (**Supplementary Fig. 14**). *T. cacao* seed storage

lipids (cocoa butter), representing 50% of the dry seed weight, are exceptional in their very high level of stearate (30–37%), which gives cocoa butter its relatively high melting point (34–38 °C)[26]. The unique fatty acid profile of cocoa butter enhances the olfactory qualities of chocolate and confectionaries and makes it valuable for cosmetic and pharmaceutical products. We discovered a total of 84 orthologous *T. cacao* genes potentially involved in lipid biosynthesis, which is 13 more than those discovered in *Arabidopsis*[27] (**Supplementary Table 15**). Consistent with the large amount of storage lipids produced in *T. cacao* seeds, the genome contained five additional genes encoding acyl-ACP thioesterase fat B (*FATB*) and three additional genes encoding ketoacyl-ACP synthase, the two key workhorse enzymes leading to the synthesis of triacylglycerols.

Flavonoids are a diverse group of plant secondary metabolites that play many important roles during plant development[28]. They are involved in plant defense against insects, pathogens and microbes, in absorption of free radicals and ultraviolet light, and in attraction of beneficial symbionts and pollinators. Proanthocyanidins are flavonoid polymers that are present in large amounts in *T. cacao* seeds (**Supplementary Fig. 15**). Recent evidence suggests that proanthocyanidins may be beneficial to human health by improving cardiovascular health, providing cancer chemopreventative effects and also through neuroprotective activities[29,30]. We identified 96 *T. cacao* genes orthologous to *Arabidopsis* genes that are involved in the flavonoid biosynthetic pathway (**Supplementary Table 16**), which is 60 more than are present in *Arabidopsis*. Of these genes, we evaluated the function of *TcANS*, *TcANR* and *TcLAR* in transgenic *Arabidopsis* and Tobacco, demonstrating that they are functional orthologs of *Arabidopsis* genes[31]. Notably, although *Arabidopsis* has only one gene encoding dihydroflavonol-4-reductase (*DFR*), the *T. cacao* genome contains 18 orthologous *DFR* genes. DFR catalyzes the reaction that produces the flavan-3,4-diols, the immediate precursors of the flavonoids catechin and epicatachin. These compounds can accumulate to as much as 8% of the dry *T. cacao* seed, making *T. cacao* one of the richest known sources of this phytonutrient[32].

Terpenoids constitute a large family of natural compounds and play diverse roles in plants as hormones, pigments and in plant-environment interactions and defense. They are major components of resins, essential oils and aromas[33]. Among them, two subclasses of terpenoids are particularly involved in aromas: monoterpenes (C10), which represent aromatic compounds that are the basis of floral essences and essential oils, and sesquiterpenes (C15), which may also constitute a defense response of plants toward microorganisms or insect aggression. Compared to bulk cocoa, a higher level of monoterpenes (such as linalool, an acyclic monoterpene alcohol found in the floral scent of *Clarkia breweri* and of many other plants species) has been observed in fine-flavored cocoa varieties like Criollo and Nacional, which are characterized by fruity and floral notes[34,35].

We identified 57 *T. cacao* genes that are orthologs of *Arabidopsis* genes that encode terpene synthase (*TPS*), which catalyze terpenoid synthesis[33], and nine pseudogenes (**Supplementary Table 17** and **Supplementary Figs. 16** and **17**). This number is higher than in *Arabidopsis* and poplar[36], which have 30 and 40 genes, respectively, and lower than in grape[10], which has 89 functional genes. The classification of *TPS*s in the different subclasses revealed that 34% of them correspond to monoterpenes and 31% correspond to sesquiterpenes. Two gene families are particularly expanded in *T. cacao*: the linalool synthase family (monoterpenes), which is represented by 7 genes clustered in a region of chromosome 6, and the cadinene synthase family (sesquiterpenes), comprising 10 members, among which 7 are localized in a same region of chromosome 7. Cadinene synthase is one of the key enzymes involved in the synthesis of gossypol, a toxic terpenoid produced in the seeds of cotton, a species belonging to Malvaceae, the same family as *T. cacao*. In cocoa, cadinene synthase has been found to be expressed in pod tissues[37] in response to attacks by mirids (*Sahlbergella singularis*), a major insect pest of cocoa trees in Africa. Therefore, the cadinene synthase orthologous genes are candidates for elements of the cocoa insect resistance response.

Colocalization of quality related genes and QTL

Previous studies have reported QTLs associated with quality traits such as lipid and flavonoid content[38,39] (**Supplementary Note**). For most of these QTLs, genes encoding key enzymes of these biosynthetic pathways were found to be colocalized with most of these QTLs (**Supplementary Fig. 16**). For example, a major QTL for fat content is associated on chromosome 9 with a gene orthologous to *KCS*, encoding beta ketoacyl-CoA synthase, and is located very close to an ortholog of one member of the *FATB* gene group, which is specifically expanded in *T. cacao*. A strong QTL for cocoa butter hardness[39] was found localized in linkage group 7 near a gene orthologous to *FAD4*, which is involved in creating a bond between C2 and C3 of the lipid chain, resulting in lipids with a higher melting point, which, in terms of cocoa butter, represents greater hardness.

Similarly, we found each of the QTLs for astringent taste to be closely associated with genes potentially involved in the proanthocyanidins biosynthetic pathway (**Supplementary Fig. 15**). For example, two genes orthologous to that encoding flavonoid 3-hydroxylase (*F3H*) and one orthologous to that encoding dihydro-flavonol-4-reductase (*DFR*), which are specifically expanded in *T. cacao*, co-localize on linkage group 1 with a major astringency QTL.

T. cacao genome paleo-history

Angiosperms have been shown to evolve through rounds of paleo-polyploidy[10,40]. Two types of events have been reported in the literature for eudicots: an ancestral event (referenced as γ) and lineage-specific events (referenced as α and β). In order to investigate the paleo-history of the *T. cacao* genome, we characterized shared paleo-polyploidies based on the integration of orthologous relationships identified between *T. cacao* and five eudicot sequenced genomes (*Arabidopsis*[41], grape[10], poplar[36], soybean[42] and papaya[43]), as well as paralogous relationships identified between the ten *T. cacao* chromosomes.

Recently, we published a method for the identification of orthologous regions between plant genomes as well as for the detection of duplicated blocks within genomes based on integrative sequence alignment criteria combined with statistical validations[44]. This approach has been recently applied to available monocot and eudicot genomes and has allowed us to propose a common ancestor with five (core gene set of 9,138) and seven proto-chromosomes (core gene set of 9,731) for monocots and eudicots, respectively[45]. We have integrated the *T. cacao* genome sequence information (23,529 gene models anchored) into our previous paleo-genomics analysis in order to investigate the *T. cacao* evolutionary paleo-history (Online Methods).

Using the alignment parameters and statistical tests reported previously[44], 7,866 orthologous relationships covering 80% of the *T. cacao* genome were identified between the *T. cacao* and the *Arabidopsis*, poplar, grape, soybean and papaya genomes. The chromosome-to-chromosome orthologous relationships that were established between the *T. cacao* and the five sequenced eudicot genomes are illustrated in **Figure 4a** and are available in **Supplementary Table 18**.

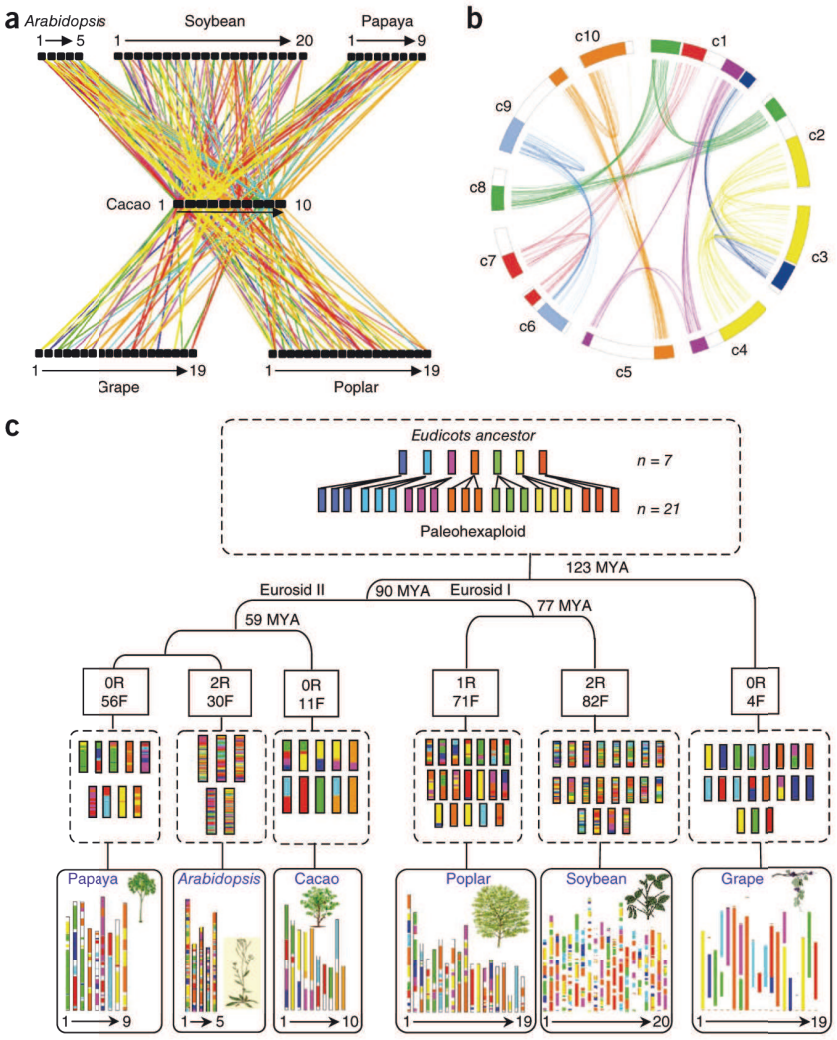


Figure 4 *T. cacao* genome paleohistory. (**a**) *T. cacao* genome synteny. A schematic representation of the orthologs identified between *cacao* chromosomes (c1 to c10) at the center and the grape (g1 to g19), *Arabidopsis* (a1 to a5), poplar (p1 to p19), soybean (s1 to s20) and papaya (p1 to p9) chromosomes. Each line represents an orthologous gene. The seven different colors used to represent the blocks reflect the origin from the seven ancestral eudicot linkage groups. (**b**) *T. cacao* genome duplication. The seven major triplicated chromosomes groups in *T. cacao* (c1 to c10) are illustrated (colored blocks) and related with paralogous gene pairs identified between the *T. cacao* chromosomes (colored lines). The seven different colors reflect the seven ancestral eudicot linkage groups. (**c**) *T. cacao* genome evolutionary model updated from Abrouk *et al.*[46]. The eudicot chromosomes are represented with a seven-color code to illustrate the evolution of segments from a common ancestor with seven protochromosomes (top). The different lineage-specific shuffling events that have shaped the structure of the six genomes during their evolution from the common paleo-hexaploid ancestor are indicated as R (for rounds of whole-genome duplication (WGD)) and F (for fusions of chromosomes). The current structure of the eudicot genomes is represented at the bottom of the figure.

Moreover, we aligned the 23,529 gene models from the *T. cacao* genome onto themselves. Seven blocks of duplicated genes (344 gene pairs) were identified and characterized in *T. cacao*, covering 64% of the genome and involving the following chromosome to chromosome (c) relationships: c2-c3-c4 (yellow), c1-c3 (blue), c1-c2-c8 (green), c6-c9 (light blue), c1-c4-c5 (purple), c5-c9-c10 (orange) and c1-c6-c7 (red) (**Fig. 4b** and **Supplementary Table 19**). We found this ancestral paleo-polyploidy event shared at orthologous positions between eudicot genomes on the following chromosome pair combinations in *T. cacao* compared to the seven ancestral triplicated chromosome groups reported in grape (g)[10]: g1-g14-g17/c2-c3-c4 (yellow), g2-g12-g15-g16/c1-c3 (blue), g3-g4-g7-g18/c1-c2-c8 (green), g4-g9-g11/c6-c9 (light blue), g5-g7-g14/c1-c4-c5 (purple), g6-g8-g13/c5-c9-c10 (orange), g10-g12-g19/c1-c6-c7 (red) (**Fig. 4c**). This result confirms the paleo-hexaploid origin of the eudicot species recently reported for the grape[10] and soybean[42] genomes. Moreover, we confirmed the known γ paleo-hexaploidization event in the *T. cacao* genome through classical Ks-based (synonymous substitution rate) data analysis between paralogous genes (**Supplementary Fig. 18**).

Based on the ancestral (γ) and lineage-specific duplications (α, β) reported for eudicots, it became possible to propose an evolutionary scenario that shaped the ten *T. cacao* chromosomes from the seven chromosomes of the eudicot ancestor and, more precisely, to the 21 chromosomes of the paleo-hexaploid ancestor (**Fig. 4c**). We suggest, from the 21 chromosomes intermediate ancestor, at least 11 major chromosome fusions (referenced as 'F' in **Fig. 4c**) to reach the actual ten-chromosome structure (compared to 30, 4, 71, 82 and 56 reported, respectively, for *Arabidopsis*, grape, poplar, soybean and papaya genomes)[46].

Finally, in order to gain insight into our understanding of the molecular mechanisms driving the chromosome number reduction from the 21 chromosome ancestor intermediate to the actual 10 chromosome structure of the *T. cacao* genome, we produced heat maps scoring particular features such as CoDing Sequences (CDS), transposable element repeats for class I and II, tandem-Gaucho elements and telomeric repeats (**Fig. 2**). We observed a classical distribution pattern of CDS and transposon elements that were more abundant at the telomeric and centomeric regions of the chromosomes, respectively. Moreover, we identified a clear correlation between the position of chromosome fusion points (dotted rectangles) and the occurrence of telomeric repeats (telomeric remnants collocating with eight out of the eleven sites) consistent with a telomere-telomere recombination process leading to the chromosome fusion events reported previously in eudicots[47].

DISCUSSION

We sequenced the genome of *T. cacao*, resulting in the assembly of 76% of its genome and identification of 28,798 protein-coding genes, among which 23,529 (82%) were anchored onto the ten cocoa chromosomes. A large proportion of the euchromatin of the *T. cacao* genome is likely covered by this assembly, allowing for the recovery of 97.8% of the *T. cacao* unigene resource. We found that 682 gene

families are specific to *T. cacao*, as compared to *A. thaliana*, grape, soybean and poplar proteomes. Only 24% of the *T. cacao* genome consists of transposable elements, a lower percentage than in other genomes of similar size. The analysis of specific gene families that are potentially linked to cocoa qualities and disease resistance showed that particular expansion or reduction of some gene families appears to have occurred. The mapping of these gene families along the cocoa chromosomes and comparison with the genome regions involved in trait variation (QTLs) constitutes an invaluable source of candidate genes for further functional studies that aim to discover the specific genes directly involved in trait variation. This draft genome sequence will facilitate a better understanding of trait variation and will accelerate the genetic improvement of *T. cacao* through efficient marker-assisted selection and exploitation of genetic resources. Using an updated version of the Newbler software (released by Roche on 8/17/2010), we performed a second-generation assembly of the cacao genome data (ICGS Assembly 1.2). The new assembly covers 84.3% of the *T. cacao* B97-61/B2 genome, with a N50 scaffold size of 5.624 Mb and the largest scaffold of 18.20 Mb. This enhanced assembly enabled us to improve its anchorage onto the genetic map, which now includes approximately 87% of the assembled sequences on ten pseudochromosomes. Additional details of this and further assembly improvements will be available on the ICGS website (see URLs).

This study has highlighted the close evolutionary relationship of the *T. cacao* genome to the eudicot putative ancestor, showing a limited number of recombinations between ancestral chromosomes, as has also been observed in grape[10]. *T. cacao*, which has only ten pairs of chromosomes, is easily propagated by both sexual and vegetative methods and can be transformed[48], and therefore, it represents a new and simple model to study the evolutionary processes, gene function, genetics and biochemistry of tree fruit crops.

URLs. Cocoa statistics, http://www.icco.org/economics/market.aspx, http://www.dropdata.org/cocoa/cocoa_prob.htm; ICGS website, http://cocoagendb.cirad.fr/gbrowse/cgi-bin/gbrowse/theobroma/; web tools provided by the Bioinformatics and Systems Biology of Gent, http://bioinformatics.psb.ugent.be/webtools/Venn/; MUST, http://csbl1.bmb.uga.edu/ffzhou/MUST/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The *Theobroma cacao* whole-genome sequences are deposited in the EMBL, GenBank and DDBJ databases under accession numbers CACC01000001–CACC01025912. A genome browser and further information on the project are available from http://cocoagendb.cirad.fr/gbrowse and http://cocoagendb.cirad.fr.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank CIRAD, the Agropolis foundation, the Région Languedoc Roussillon, Agence Nationale de la Recherche (ANR), Valrhona and the Venezuelan Ministry of Science, Technology and Industry for their financial contribution to this project. We thank the Toulouse Midi-Pyrénées bioinformatic platform for providing us with computational resources. Activities at Pennsylvania State University were supported by a gift from the Hershey Corp. and through support from the Schatz Center for Tree Molecular Genetics in the School of Forest Resources. Acquisition of an Illumina sequencer by the Cold Spring Harbor Laboratory was supported by the National Science Foundation grant DBI0923128 to W.R.M. We would like to thank F.C. Baurens, Y. Jiao, O. Garsmeur and C. dePamphilis for helpful advice and assistance with bioinformatics. We would like to express our special appreciation to V. Mooleedhar, who collected the cacao accession in Belize and made it available to us for our analysis.

AUTHOR CONTRIBUTIONS

X.A., J.S., J.-M.A., M.J.G., J.G., D.K., M.J.A., S. Brown, K.G., A. D'Hont, A. Dievart, D.B., D.I., P.C., R.W., W.R.M., E.G., F.Q., O.P., P.W., S. Bocs and C.L. designed the analyses.

X.A., J.S., J.-M.A., M.J.G., J.G., M.R., D.K., M.J.A., S. Brown, A. D'Hont, D.B., W.R.M., O.P., P.W., S. Bocs and C.L. managed the several components of the project.

X.A., M.A., O.F., Y.R., A.B., M. Bocca, D.C., R.R., M.T., J.M.A., K.G., I.K., J.-M.A. and C.L. performed material preparation and multiplication, DNA and RNA extractions, genotyping, genetic mapping and anchoring of the assembly.

D.K., J.S.S.A., W.G. and X.S. performed BAC libraries.

J.-M.A., J.P., S.C.S., J.E.C., M.K., L.G. and W.R.M. performed sequencing and assembly.

X.A., G.D., J.G., M. Allegre, T.L., S.N.M., E.S., T.S., Z.S., C.V., V.G., Y.Z., B.P. and S. Bocs performed automatic and manual gene annotations and database management.

C.C., J.F.B.-N., F.S., A.M.R., M.J.A., Z.M., O.P. and S. Brown performed repeated elements and miRNA analyses.

M.R.-G., M. Bourge, S. Brown and A. D'Hont performed in situ hybridizations and genome-size evaluations.

M.J.G., G.D., T.L., S.N.M., M.R., A. Dievart, Z.S., X.S. and Y.Z. performed gene family analyses.

J.S., M. Abrouk and F.M. performed evolution analyses.

X.A., J.S., J.-M.A., M.J.G., G.D., J.G., C.C., T.L., S.N.M., M.R., M.R.-G., D.K., S.C.S., A. D'Hont, A. Dievart, X.S., M.J.A., S. Brown, P.C., F.Q., O.P., S. Bocs and C.L. wrote and/or revised the paper.

C.L. initiated and coordinated the whole project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Davie, J.H. Chromosome studies in the Malvaceae and certain related families. II. *Genetica* **17**, 487–498 (1935).
2. Henderson, J.S., Joyce, R.A., Hall, G.R., Hurst, W.J. & McGovern, P.E. Chemical and archaeological evidence for the earliest cacao beverages. *Proc. Natl. Acad. Sci. USA* **104**, 18937–18940 (2007).
3. Coe, S.D. & Coe, M.D. *The True History of Chocolate.* (Thames and Hudson Ltd., London, England, 1996).
4. Motamayor, J.C. *et al.* Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* **89**, 380–386 (2002).
5. Motamayor, J.C., Risterucci, A.M., Heath, M. & Lanaud, C. Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. *Heredity* **91**, 322–330 (2003).
6. Mooleedhar, V., Maharaj, W. & O'Brien, H. The collection of Criollo cocoa germplasm in Belize. *Cocoa Grower's Bull.* **49**, 26–40 (1995).
7. Cocoa Resources in consuming Countries–ICCO Market Committee, 10th meeting. *EBRD Offices London, MC* **10**, 16 (2007).
8. Paterson, A.H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
9. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
10. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
11. Wolfgruber, T.K. *et al.* Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
12. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGène as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
13. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
14. Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell* **136**, 669–687 (2009).
15. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
16. Afzal, A.J., Wood, A.J. & Lightfoot, D.A. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol. Plant Microbe Interact.* **21**, 507–517 (2008).

17. Diévart, A. & Clark, S.E. LRR-containing receptors regulating plant development and defense. *Development* **131**, 251–261 (2004).
18. Lehti-Shiu, M.D., Zou, C., Hanada, K. & Shiu, S.H. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* **150**, 12–26 (2009).
19. DeYoung, B.J. & Innes, R.W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243–1249 (2006).
20. Tarr, D.E.K. & Alexander, H.M. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res. Notes* **2**, 197 (2009).
21. Pan, Q., Wendel, J. & Fluhr, R. Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.* **50**, 203–213 (2000).
22. Mukhtar, M.S., Nishimura, M.T. & Dangl, J. NPR1 in plant pefense: it's not over 'til it's turned over. *Cell* **137**, 804–806 (2009).
23. Shi, Z., Maximova, S., Lui, Y., Verica, J. & Guiltinan, M.J. Functional analysis of the *Theobroma cacao NPR1* Gene in *Arabidopsis*. *BMC Plant Biol.* **10**, 248 (2010).
24. Lehmann, P. Structure and evolution of plant disease resistance genes. *J. Appl. Genet.* **43**, 403–414 (2002).
25. Lanaud, C. *et al.* A meta–QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol. Breed.* **24**, 361–374 (2009).
26. Griffiths, G. & Harwood, J.L. The regulation of triacylglycerol biosynthesis in cocoa (*Theobroma cacao*) L. *Planta* **184**, 279–284 (1991).
27. Beisson, F. *et al. Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed secuence tags in organs, and a web-based database. *Plant Physiol.* **132**, 681 (2003).
28. Pourcel, L., Routaboul, J., Cheynier, V., Lepiniec, L. & Debeaujon, I. Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* **12**, 29–36 (2007).
29. Spencer, J.P. Flavonoids and brain health: multiple effects underpinned by common mechanisms. *Genes Nutr.* **4**, 243–250 (2009).
30. Rimbach, G., Melchin, M., Moehring, J. & Wagner, A.E. Polyphenols from cocoa and vascular health-a critical review. *Int. J. Mol. Sci.* **10**, 4290–4309 (2009).
31. Liu, Y. Molecular analysis of genes involved in the synthesis of proanthocyanidins in theobroma cacao. *Thesis* 1–146 (2010).
32. Tomas-Barberan, F.A. *et al.* A new process to develop a cocoa powder with higher flavonoid monomer content and enhanced bioavailability in healthy humans. *J. Agric. Food Chem.* **55**, 3926–3935 (2007).
33. Liu, Y., Wang, H., Ye, H. & Li, G. Advances in the plant isoprenoid biosynthesis pathway and its metabolic engineering. *J. Integr. Plant Biol.* **47**, 769–782 (2005).
34. Ziegleder, G. Linalol contents as characteristics of some flavour grade cocoas. *Z. Lebensm. Unters. Forsch.* **191**, 306–309 (1990).
35. Chanliau, S. & Cros, E. Influence du traitement post-récolte et de la torréfaction sur le développement de l'arôme cacao. *12th Int. Cocoa Res. Conf., Salvador de Bahia (Brazil)* 959–964 (1996).
36. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
37. Argout, X. *et al.* Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* generated from various tissues and under various conditions. *BMC Genomics* **9**, 512 (2008).
38. Lanaud, C. *et al.* Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma cacao* L. *Proc. 14th Int. Cocoa Res. Conf.* 13–18 (2003).
39. Araújo, I.S. *et al.* Mapping of quantitative trait loci for butter content and hardness in cocoa beans (*Theobroma cacao* L.). *Plant Mol. Bio. Rep.* **27**, 177–183 (2009).
40. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
41. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796–815 (2000).
42. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
43. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
44. Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. & Feuillet, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Briefings Bioinf.* **10**, 619–630 (2009).
45. Salse, J. *et al.* Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. USA* **106**, 14908–14913 (2009).
46. Abrouk, M. *et al.* Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487 (2010).
47. Murat, F. *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **11**, 1545–1547 (2010).
48. Maximova, S.N. *et al.* Over-expression of a cacao class I chitinase gene in *Theobroma cacao* L. enhances resistance against the pathogen, *Colletotrichum gloeosporioides*. *Planta* **224**, 740–749 (2006).

ONLINE METHODS

High molecular weight DNA preparation. High molecular weight DNA was prepared from nuclei of B97-61/B2 cocoa leaves according to previously described protocols[49], except that the steps of filtration were replaced by five successive filtrations with nylon filters (SEFAR NITEX) having a decreasing mesh diameter: 250 µM, 100 µM, 50 µM and two times 11 µM. (**Supplementary Note**).

Construction of BAC libraries. Two BAC libraries were constructed from cocoa leaves. DNA was isolated from nuclei collected in agarose plugs and DNA digestions were performed with HindIII or EcoRI, followed by ligation to the pAGIBAC1 vector (a modified pIndigoBAC536Blue with an additional *Swa*I site[49]). Ligation products were transformed into DH10B T1 phage-resistant *Escherichia coli* cells (Invitrogen) and plated on Lysogeny broth agar that contained chloramphenicol (12.5 µg ml^{-1}), X-gal (20 mg ml^{-1}) and Isopropyl β-D-1-thiogalactopyranoside (0.1 M). For characteristics, quality assessment and estimated genome coverage see the **Supplementary Note** and **Supplementary Table 1**.

Genome sequencing. The genome was sequenced using a genome-wide shotgun strategy. All data were generated using next-generation sequencers: Roche/454 GSFLX (×16.5 coverage) and Illumina GAIIx (×44 coverage), except for data from BAC ends (×0.2 coverage), which were produced by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers (**Supplementary Note** and **Supplementary Table 2**).

Genome assembly and automatic error corrections with Solexa/Illumina reads. Sanger and 454 reads were assembled with Newbler version 2.3. From the initial 26,519,827 reads, 80.65% (21,387,691) were assembled by Newbler. The 454 assembly was improved by automatic error corrections with Solexa/Illumina reads, which have a different bias in error type, as described previously[50]. Short-read sequences were aligned on the cocoa genome assembly using the SOAP software (with a seed size of 12 bp and a maximum gap size allowed on a read of 3 bp). Only uniquely mapped reads were retained (**Supplementary Note**).

Estimation of nuclear DNA content by flow cytometry. The genome sizes of B97 61/B2 and a panel of diverse cocoa clones were estimated by flow cytometry[51]. Leaves of studied samples and internal standards were chopped with a razor blade and then stained using propidium iodide. DNA content of 5,000–10,000 stained nuclei was determined using a CyFlow SL3 flow cytometer with a 532-nm green solid state laser (100 mWatt). The monoploid C-value (1C) was calculated and expressed in Mb using the conversion factor 1 pg DNA = 978 Mb (**Supplementary Note**).

Anchoring of the assembly in the genetic map. A consensus map was established from two progenies: an F1 progeny of 256 individuals (UPA 402 × UF676) and an F2 progeny of 136 individuals recently produced (Scavina 6 × ICS1). The F1 progeny was previously used to established the cocoa reference map, which includes 600 markers[52,53]. New SSR and SNP markers were mapped in these two progenies, and a consensus map including 1,259 markers was established[54]. BLAT software was used to align the markers of the genetic map with the scaffolds (**Supplementary Note**).

Prediction of transposable elements. The annotation of transposable elements in the cocoa genome was achieved in two stages. First, a combination of *de novo* analyses (for example, LTR_finder, LTRharvest, MUST (see URLs)) and extrinsic comparisons (BLAST) was conducted. Then, a *de novo* approach was used to construct highly repeated elements from the 3,220,522 unassembled reads. A total of 67,575 transposable elements were annotated (**Supplementary Note**).

***In situ* hybridization of transposable element probes.** FISH was performed on mitotic metaphase spreads prepared from meristem root tip cells. The probes were labeled with Alexa-488 dUTP and Alexa-594 dUTP by random priming (Fisher Bioblock Scientific) and the *in situ* hybridizations were performed according to D'Hont *et al.*[55] (**Supplementary Note**).

Prediction of protein-coding genes. Gene structures were predicted using EUGene[12]. Translation start sites and RNA splice sites were predicted by SpliceMachine[56]. Available *T. cacao* ESTs were aligned onto the scaffolds using GenomeThreader[57]. Similarities to proteins from Swiss-Prot, TAIR, Malvacae GenBank extraction, *Glycine max* high confidence gene models[58] and translated *T. cacao* EST contigs were searched using BLASTX. Similarities to *A. thaliana*, *Gossypium*, *V. vinifera*, *Citrus* and *T. cacao* ESTs were searched using TBLASTX. A total of 50,582 genes were predicted, giving a final count of 28,798 *T. cacao* genes after filtering (**Supplementary Note**).

Gene family analysis. Protein domains were searched using InterProscan against the InterPro database. Cocoa, *Arabidopsis*, grape, poplar and soybean Best Blast Mutual Hit (BBMH) were computed, and protein clustering was done using OrthoMCL[59].

Synteny and duplication analysis. Arabidopsis, grape, poplar, soybean and papaya proteomes were aligned using an approach based on BLASTP in order to identify accurate paralogous and orthologous relationships[44,45]. K_S divergence (million year ago (MYA) scale) for paralogous and orthologous gene pairs as well as speciation events dating were calculated with PAML[60].

49. Ammiraju, J.S.S. *et al.* The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res.* **16**, 140–147 (2006).
50. Aury, J.M. *et al.* High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
51. Marie, D. & Brown, S.C. A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biology of the Cell/Under the Auspices of the European Cell Biology Organization* **78**, 41–51 (1993).
52. Pugh, T. *et al.* A new *cacao* linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* **108**, 1151–1161 (2004).
53. Fouet, O. *et al.* Structural characterization and mapping of functional EST-SSR markers in *Theobroma cacao*, in the press.
54. Allegre, M. *et al.* A high-density consensus genetic map for *Theobroma cacao* L., in the press.
55. D'hont, A. *et al.* Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet.* **250**, 405–413 (1996).
56. Degroeve, S., Saeys, Y., De Baets, B., Rouzé, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332–1338 (2005).
57. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
58. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
59. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
60. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).

CHAPITRE 3 - AMÉLIORATION DE LA SÉQUENCE DU GÉNOME DU CRIOLLO

1. Introduction

Les nouvelles méthodes de séquençage haut débit (NGS) ont favorisé la réalisation du premier assemblage du génome du cacaoyer en 2010. Cependant, même si la première version de cet assemblage peut être considérée de bonne qualité (Chain et al., 2009), la technique de séquençage par WGS a eu pour conséquence de fragmenter le génome en de nombreuses séquences contiguës (contigs). Ainsi, cet assemblage a généré 25 912 contigs qui, grâce aux liaisons apportées par les séquences pairées, ont abouti à 4 795 scaffolds dont une partie a été ancrée sur les chromosomes. Un autre revers de cette fragmentation est la présence d'une quantité non négligeable de trous (gaps) dans les séquences des scaffolds. Ces trous sont des séquences non assemblées mais dont la position est connue par les liaisons pairées situées à leurs extrémités. On estime à 10,8% la part de nucléotides non déterminés (Ns) dans l'assemblage finale de la première version du génome du Criollo.

Le défi principal de l'algorithmique pour l'assemblage de génome à partir de la technique WGS - et donc à l'origine de la fragmentation - sont les éléments répétés (Alkan et al., 2011). En effet, leurs très nombreuses copies, presque identiques, sont difficiles à résoudre avec les courtes séquences produites par les techniques NGS et tendent à être assemblées en une seule région. De plus, les contigs représentant des régions à copie unique ne peuvent pas être étendus de manière non équivoque à la frontière des éléments répétés. Ce manque de continuité dans les assemblages NGS et la présence de régions nucléotidiques non élucidées représentent des obstacles majeurs pour conduire les analyses génétiques. Par exemple, les études d'association pangénomiques utilisent des marqueurs moléculaires ordonnés le long des chromosomes. Si l'ordonnancement de ces marqueurs n'est pas adéquat, des biais peuvent apparaitre dans les résultats.

Récemment, de nouvelles méthodes dérivant des séquençages NGS ont été introduites pour améliorer l'assemblage des génomes. Les plateformes de séquençage NGS proposent désormais des séquences de plus grande longueur ainsi qu'une information positionnelle produite à partir de fragments de grande taille permettant de traverser certains éléments répétés.

Ces outils ont été utilisés chez le bananier pour améliorer la séquence du génome de référence (Martin et al., 2016) et ont démontré leur bonne propriété pour augmenter la taille des séquences contigües et pour corriger les erreurs d'assemblage. De plus, les nouvelles méthodes de génotypage par séquençage ou GBS permettent la construction de cartes génétiques très haute densité pouvant être utilisées pour améliorer l'ancrage des contigs ou scaffolds sur les chromosomes (Glazer et al., 2015).

Dans ce chapitre, je présenterai sous forme d'article scientifique l'amélioration de la première version d'assemblage du génome du cacaoyer. Pour cela nous avons utilisé l'information apportée par le séquençage de 4 banques Illumina de grands fragments combinée à l'utilisation de lectures de grande longueur PacBio pour corriger les erreurs d'assemblage, réduire le nombre de scaffolds et améliorer la qualité globale du génome Criollo du cacaoyer. Cet article présentera également le génotypage par séquençage d'une descendance de Guyane utilisée pour augmenter la part de la séquence du génome ancrée sur les 10 chromosomes.
L'article a été présenté au Workshop congrès Plant & Animal Genome XXV Conference cacao workshop, San Diego, 2017 et sera prochainement soumis à la revue BMC Genomics.

The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies

X. ARGOUT, G. MARTIN, G. DROC, K. LABADIE, E. RIVALS, O.FOUET, J.M. AURY and C. LANAUD

Abstract

Background

Theobroma cacao L., native from the Amazonian basin of South America is an economically important fruit tree crop for tropical countries, source of chocolate. The first draft genome of the species, from a Criollo cultivar was published in 2011. Although a useful resource, some improvements can be made, including efforts to identify misassemblies, and reduction of the number of scaffolds, gaps, and un-anchored sequences to the ten chromosomes.

Results

In this work, we used a NGS-based approach to significantly improve the assembly of the Belizian Criollo B97-61/B2 genome. We combined 4 Illumina large insert size mate paired libraries with 52x of Pacific Biosciences long reads to correct misassembled regions, reduce the number of scaffolds to 554 (4,792 in assembly V1) with a N50 increased from 0.47 Mb to 6.5 Mb. 96.7% of the assembly was anchored to the 10 chromosomes compared to the previous 66.8%. Unknown sites (Ns) were reduced from 10.8% to 5.7%. Moreover, the NCBI Eukaryotic Genome Annotation Pipeline carried out a new RefSeq structural annotation based on RNAseq evidences

Conclusion

The release of the *Theobroma cacao* Criollo genome version 2 will be a valuable resource for investigating complex traits at the genomic level and is an important step for future comparative genomics and genetics studied on cocoa. New functional tools and annotations are available through the cacao genome hub (http://cocoa-genome-hub.southgreen.fr).

Background

Theobroma cacao L. is a tropical fruit tree endemic from the Amazonian basin of South America. It's a diploid species (2n=2x=20), with a relative small genome size, ranging among *T. cacao* genotypes from 411Mb to 494Mb [1]. In the last decade, the genetic diversity structure of the species has been decipher with SSR markers [2,3] and 10 genetic groups have been proposed: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Naconal, Nanay and Purús. Recently, to accelerate cocoa breeding and to better study molecular mechanisms of agronomic traits, several projects have been conducted in *Theobroma cacao* genomics [1,4,5].

In 2011, the first *T. cacao* genome of the B97-61/B2 genotype, a member of the Criollo genetic group was released, providing a major source of candidate genes for *T. cacao* improvement and highlighting a close evolutionary process with the putative Eudicot ancestor [1]. This draft genome sequence was performed with a Whole Genome Shotgun (WGS) strategy, comprising Roche/454 reads for contigs assembly, Sanger Bac Ends sequence for the scaffolding step and a genetic map of 1,259 markers to anchor the scaffolds to chromosomes. The final assembly covered 76% of the estimate size of B97-61/B2 into 4,792 scaffolds. Moreover, 97,8% of the unigenes (assembled from the transcriptome data) was recovered into the assembly.

Despite the fact that this genome assembly can be considered as high quality draft [6], some improvements can be made, including efforts to identify misassemblies, and reduction of the number of scaffolds, gaps, and un-anchored sequences to the ten chromosomes.

In 2013, another WGS project was released for the *Theobroma cacao* Matina1-6 genotype, covering 77% of the evaluated genome size of this member of the Amelonado genetic group [5].

Any WGS-based de novo sequence assembly, because they used short sequences from relatively short insert size libraries, suffers to deal with the redundancy due to common repeats such as transposable elements (TEs) and duplicated sequences [7]. As a result, WGS assembly algorithms collapse identical repeats in single regions resulting in reduce genomic complexity. Moreover, these collapse regions can link to multiple other genomic regions and the assembly process can either stops, resulting in a high number of genome fragments, or produce misassembled regions.

It has been found that 41,5% of the Matina1-6 assembled genome was covered by TEs while B97-61/B2 comprised 35,4% of TEs [5].

Nowadays, new methods derived from Next Generation Sequencing (NGS) data have been developed to improved genome assemblies. Current NGS platforms offer the possibility to produce long reads and positional information using mate-pair templates of large insert size libraries that are capable of spanning many repetitive or low complexity elements into the assembly process [8,9]. Combined with accurate gap closing procedure [10], the result is a significant augmentation of the size of contiguous genomic sequences, reduction of scaffolds number and provide no discernable misassemblies [11]. Moreover, NGS-based genotyping has also enable the discovery of sequence polymorphism segregating in mapping populations [12]. Recently, reports described the use of Genotyping by Sequencing (GBS) methods to construct dense linkage map to anchor assembly contigs or scaffolds into chromosomes [11,13,14].

In this work, we used a NGS-based approach to significantly improve the assembly of the Belizian Criollo B97-61/B2 genome. We combined 4 Illumina large insert size mate paired libraries with 52x of Pacific Biosciences long reads to correct misassembled regions, reduce the number of scaffolds and upgrade their quality. We also used a high coverage of SNP markers derived from a GBS assay of a UF676 x ICS95 mapping population of 434 individuals to greatly increase the size of the anchored genome sequences into chromosomes.

Materials and methods

Sequence data

For this work, we reused some of the dataset produced within the Theobroma cacao B97-61/B2 first draft genome project:

1- The 25,912 contigs sequences (Acc. number CACC01000000) generated from Roche/454 and Newbler assembler (Roche, Inc) as initial dataset

2- The 88,000 Sanger Bac Ends reads (available on the cocoa Genome Hub http://cocoa-genome-hub.southgreen.fr) for scaffolding step

3- The 398 million Illumina paired end reads as Short Reads (SR) for error correction (available on the cocoa Genome Hub http://cocoa-genome-hub.southgreen.fr). Cleaning of SR consisted in 1/ the removal of Hi-seq adapter sequences left in reads, 2/ a quality trimming of read extremities (Q>20), and last 3/ discarding reads shorter than 70 pb. All three improvements were performed using a single execution of CutAdapt [15]. The cleaning resulted in 336 million SR for a total of 32 gigabases.

We also generated 2 new datasets:

1- We created 4 large insert size mate paired libraries of *Theobroma cacao* B97-61/B2 genome with insert size of 3-5kb, 5-8kb, 8-11kb and 11-15kb using the Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA). These libraries were sequenced by Illumina HiSeq 2000 to respectively 40x, 35x, 19x and 10x genome coverage. The reads were trimmed with the following criteria: (i) sequences of the Illumina adapters and primers used during the library construction were removed from the whole reads; (ii) low quality nucleotides with quality value < 20 were removed from both ends (iii) the longest sequence without adapters and low quality bases was kept and sequences between the second unknown nucleotide (N) and the end of the read were also trimmed; (iv) reads shorter than 30 nucleotides after trimming were discarded; (v) finally, reads and their mates that mapped onto run quality control sequences (PhiX genome) were removed. These trimming steps were achieved using fastx_clean software (http://www.genoscope.cns.fr/fastxtend) based on the FASTX library (http://hannonlab.cshl.edu/fastx_toolkit/index.html).

2- We produced 78 SMRT Cells Pacific Biosciences sequencing data with C2 chemistry which corresponded to 52x genome coverage of Long Reads (LR) data.

The raw LR dataset was error corrected using a hybrid approach by the cleaned SR with LoRDEC [16]. We use a k-mer length of 23 and a solidity threshold of 3. Finally, uncorrected regions at the extremities of LR were trimmed. This yielded 3 millions LR with an average length of 2573 bp representing 21x genome coverage.

Contigs V1 trimming

Identification of T. *cacao* Criollo contigs misassemblies was done using the methods described by Martin et al., 2016. Briefly, large insert size paired reads (LPR) were aligned to contigs using bowtie2 [17] in -very-sensitive mode and misassemblies boundaries were identified based on the absence of overlap of read-pairs in the region. Misassembled contigs were then split and contigs smaller than 1000bp were discarded for further analysis.

Identification of chloroplast and mitochondrion contigs was carried out with BLAST [18]. Contigs were aligned to the T. *cacao* Criollo genotype chloroplast genome (acc. no. JQ228379.1) and contigs covered by 80% or more with chloroplast sequence were discarded. For mitochondrion identification, the T. *cacao* Criollo contigs were compared to the Gossypium *hirsutum* mitochondrion complete genome (acc. no. JX065074.1). T. *cacao* contigs with homology at evalue 1e-40 to the cotton mitochondrion genome were discarded.

Scaffolding

T. *cacao* Criollo contigs were scaffolded using SSPACE [9], with 4 Illumina large insert size paired reads (LPR) library and Sanger BAC ends sequences generated for the first T. *cacao* Criollo genome assembly [1]. The scaffolding process was done in a five steps, from the shortest inserts size library to the biggest. Between each steps, scaffold misassemblies were identified and resolved based on the absence of overlap of read-pairs as described previously (see section). To prevent scaffolding errors and because the sequencing depth of the first two LPR libraries was higher than the last two, more stringent parameters were used for the 5kb and 8kb LPR (-a 0.5, -k 50) than the 11kb and 15kb (-a 0.5, -k 30). The BAC ends sequences were mapped as single end-reads using bowtie2 in -very-sensitive mode and read-pairs were reconstruct for scaffolding with SSPACE (-a 0.5, -k 5).

Gap closing

Gaps in scaffolds were closed in 2 steps. GMcloser [10] was executed in --long_read mode and default option with the set of PacBio reads error-corrected and reads larger than 500bp. Then, GapCloser [19] was used with the 4 Illumina LPR libraries with a pair number cutoff for a reliable connection of 5 and a minimum aligned length to contigs for a reliable read location of 35bp.

Genetic markers

A total of 434 individuals from the cross between UF676 and ICS95 were sequenced by Diversity Arrays Technology compagny using Illumina Hiseq2000 instrument after DNA restriction with enzymes PstI and MseI. Sequencing fragments were analyzed using Tassel 5 GBS v2.2.24 pipeline [20], and parameter (-mnQS 20). Reads were aligned to scaffolds using Bowtie2 (end-to-end algorithm) and in -very-sensitive mode. Reads that aligned at different locations of the genome were discarded. SNPs were called and variant call data were filtered out with VCFtools [21]. First, indels and non-biallelic sites were excluded. Then, genotyped data with less than 10 reads were recoded as missing data and SNPs with more than 50% of missing data were excluded. Finally SNPs with minor allele frequency > 0.01, P-value > 1e-6 (X^2 test) and with a minimum distance of 64bp were selected for further analysis.

Scaffold anchoring

The method described by Martin et al., 2016 was used to assemble scaffolds into chromosomes. Marker location on scaffolds was computed using bowtie2 in very-sensitive mode. Then Pairwise linkage recombination frequencies were calculated between markers with JoinMap4.1 [22] and these data were used to compute order and orientation with an UPGMA like approach.

Gene Annotation

We first used Blastn to transfer the structural annotations from the previously annotated reference genome to the new assembly. For each gene, each exon, extended to 20 bp on both side, was aligned to the new assembly. We defined drastic criteria (no mismatch and full length alignment) to keep only the complete HSP.

Then, we performed some quality checks by comparing protein-coding sequences before and after the transfer as some discrepancies may occur. For the remaining non-transfered genes, we used Exonerate (cdna2genome model), with the same selection criteria.

A new RefSeq structural annotation has also been carried out by the NCBI Eukaryotic Genome Annotation Pipeline with the methodology described here https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/.

The functional annotation was performed with Blastp for each predicted coding sequence against the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases [23]. Based on three parameters: (1) Qcov (Query coverage = length high-scoring segment pair (HSP)/length query), (2) Scov (Subject coverage = length HSP/length subject) and (3) identity, we kept only the best result to assign a putative function to a polypeptide. InterProScan [24] was used for sequence comparison to the InterPro database [25] to obtain additional protein-signature information. KEGG pathways [26] have been reconstructed with the BlastKOALA annotation server [27]. The KEEG Orthology assignments have been done against the KEGG plant genes database at the genus level.

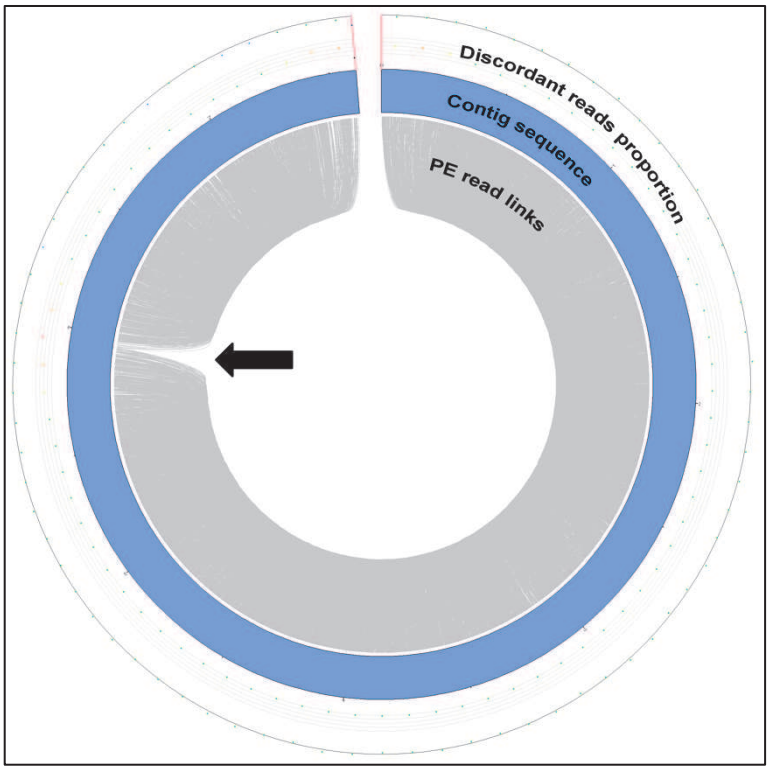


Figure 1: CIRCOS graphical representation of paired reads mapping a misassembled contig. The blue circle represents the contig sequence. In the inner circle, grey lines represent concordant links (orientation and insert size) between read pairs. In the external circle the scatter plot indicates with warm-cold colors the proportion of discordant reads on window size of one third of expected read pair insert size. The black arrow points the misassembled region.

	Sequence Number	Assembly length (Mb)	N50 (kb)	Unknown (Ns) (Mb)
Contigs	25527	290,5	19,8	-
Scaffolds (3-5kb)	4383	303,9	189,1	13,4
Scaffolds (5-8kb)	1906	312,3	439,4	21,8
Scaffolds (8-11kb)	1271	315,9	709,4	25,4
Scaffolds (11-15kb)	980	318,2	906,5	27,7
Scaffolds (BacEnds)	554	325,2	5324,1	34,6
Scaffolds (GapClosure)	554	324,7	6465,7	18,5

Table 1: Evolution of statistics during scaffold assembly.

Results

Assembly

Since the original *Theobroma cacao* Criollo draft genome was released in 2011 [1], reduction of the cost of sequencing, improved technologies and the publication of the complete *Theobroma cacao* chloroplastic genome [28] and the complete mitochondrial genome of the closely related specie *Gossypium hirsutum* [29] made it possible to update the quality of the *Theobroma cacao* Criollo genome assembly. To reach this goal, we re-scaffolded the original 25,912 contigs of the first version of the *Theobroma cacao* Criollo assembly with the methodology described in material and methods.

We first checked the consistency of the original contig dataset by searching for the absence of overlap of read-pairs in a region that may result of initial contig assembly errors with the Illumina large insert size libraries (Figure 1). We discovered that 53 contigs presented unoverlap read-pairs region. Consequently, identified misassembled regions were used to split the corresponding contigs and the new contigs dataset created was analyzed to detect organelle sequences.

A total of 37 contigs were detected as chloroplastic and 21 remains similar to mitochondrion. These contigs were discarded for the scaffolding step as well as short contigs (<1000bp). The final and clean contigs dataset comprised 25,527 sequences.

Scaffolding

The 25,527 contigs were scaffolded with SSPACE in a five step as described in the material and methods section. From the shortest insert size library (3-5kb) to the largest (Bac Ends), the number of scaffolded sequences decreases to 554 scaffolds while the assembly length increase to reach 325,2 Mb (Table 1). After gap closing procedure, the final assembly V2 comprises 554 scaffolds (4792 in assembly V1) and provides a total genome length of 324.7 Mb which represents 75.5% of the estimate size of the B97-61/B2 accession. Fifty percent of the assembly is in 17 scaffolds and the N50 is 6.5Mb. Gaps in the scaffolds represent only 5.7% of the total assembly.

We then used molecular markers to anchor the scaffolds into the 10 *Theobroma cacao* chromosomes. For that, we used Genotyping by Sequencing methods to genotype 434 individuals from the cross UF676 x ICS95.

Chromosome	SNP marker number	Scaffold number	Length (Mb)
Chr1	694	13	37.3
Chr2	581	11	41.2
Chr3	573	15	36.4
Chr4	530	7	31.9
Chr5	597	19	39.4
Chr6	369	13	26.3
Chr7	293	20	21.6
Chr8	321	15	19.6
Chr9	601	13	38.6
Chr10	298	8	21.8
Total	4857	134	314.2

Table 2: Statistics on *T. cacao* version 2 pseudo-molecule assembly

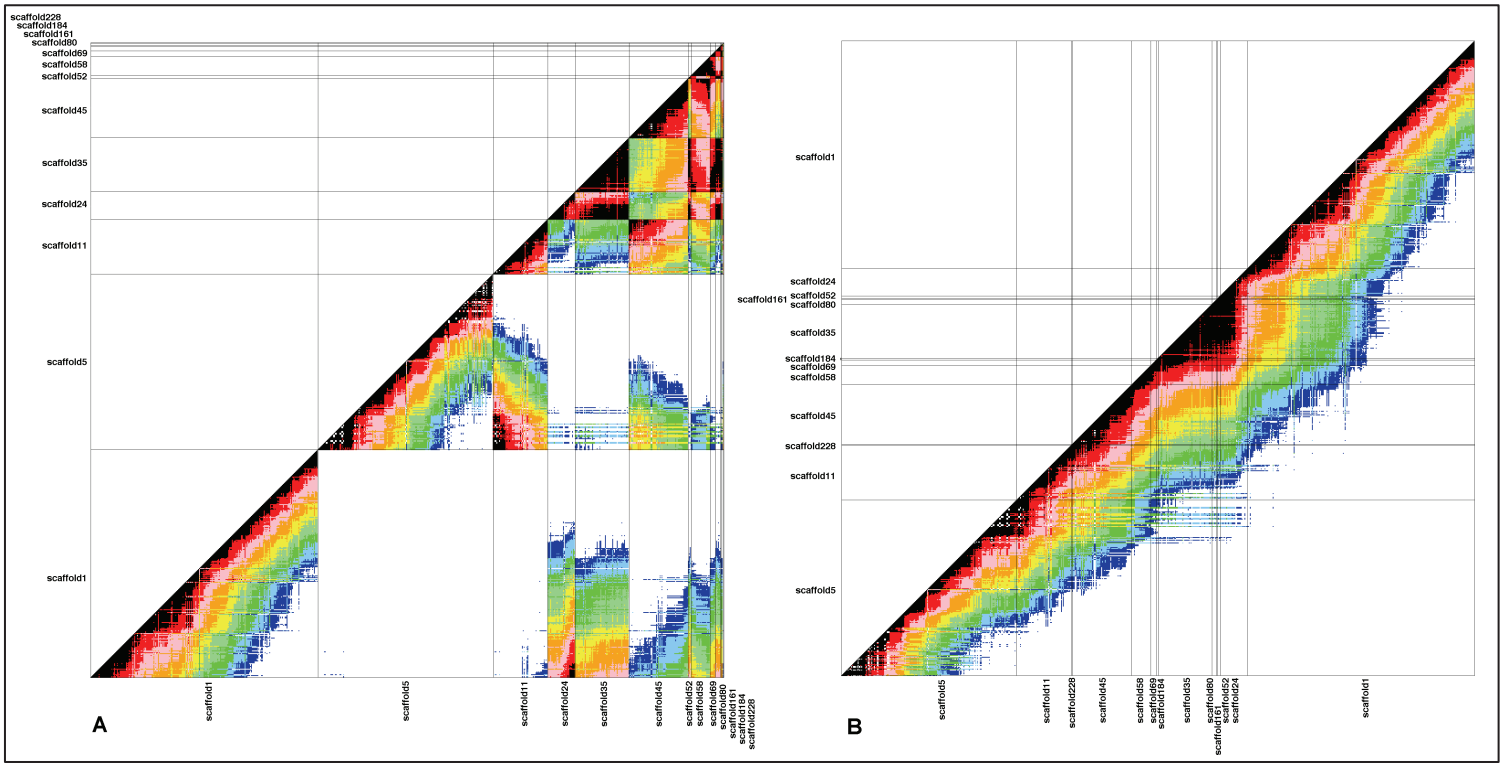


Figure 2 :Illustration of chromosome reconstruction

Linkage dot-plots between markers along scaffolds non-ordered (A) and ordered in chromosome 1. Each dot represents the recombination frequency between two markers. The intensity of the linkage between markers is color-coded. Warm colors indicate strong linkage and cold colors indicate weak linkage. Grey bars in dot plots divide markers belonging to a same scaffold.

Sequencing of the progeny generated a total of $2x10^9$ single end Illumina reads and each individual of the progeny was covered by a mean of 129.5 Mb of high quality sequence. The reads were then aligned to the 554 scaffolds dataset with Bowtie2 and SNP markers were called with the Tassel 5 GBS pipeline. After filtering out non-diallelic and indels markers, the 434 individuals were first genotyped with 39 408 SNPs. Genotype data with less than 10 reads per datapoint were recoded as "missing data". From this raw dataset, we selected a subset of 4857 SNP markers with a minor allele frequency of 0.01, a percentage of missing data < to 50%, distant of a minimum of 64bp and with a segregation distortion ratio (P > 1e-6) for genotyping the population. The molecular markers were then grouped with the JoinMap 4.1 software and a linkage group was assigned to each marker. The number of markers per linkage group ranged from 694 for linkage group 1 to 298 for linkage group 10 (Table 2).

The linkage group information assigned to each SNP marker was first used to verify the consistency of the scaffolds. None of the scaffolds contained molecular marker from different linkage group, indicating that the scaffolding step was done accurately.

Chromosome reconstruction and annotation

The pairwise recombination frequencies between each SNP markers were exported from JoinMap4.1 software and used to anchor and orientate the scaffolds into pseudochromosome with the methodology described by Martin et al., 2016. In Figure 2 is illustrated the process of chromosome reconstruction for chromosome 1. First (Figure 2A), blocks of already ordered markers based on their position on scaffolds are created. Then, the recombination frequencies are used to calculate a mean of divergence between scaffolds. At last, scaffolds were then grouped using an UPGMA like approach and scaffolds were oriented and positioned into the pseudochromosome with a round of optimization (Figure 2b).

The process was done for the 10 Theobroma cacao chromosomes and finally, a total of 134 scaffolds were anchored and oriented into chromosomes (Figure 3, Table 2). The number of scaffolds per chromosome range from 7 for chromosome 4 to 20 for chromosome 7.

The total length of the anchored genome sequence is 314.2Mb, representing 96.7% of the nuclear genome assembly.

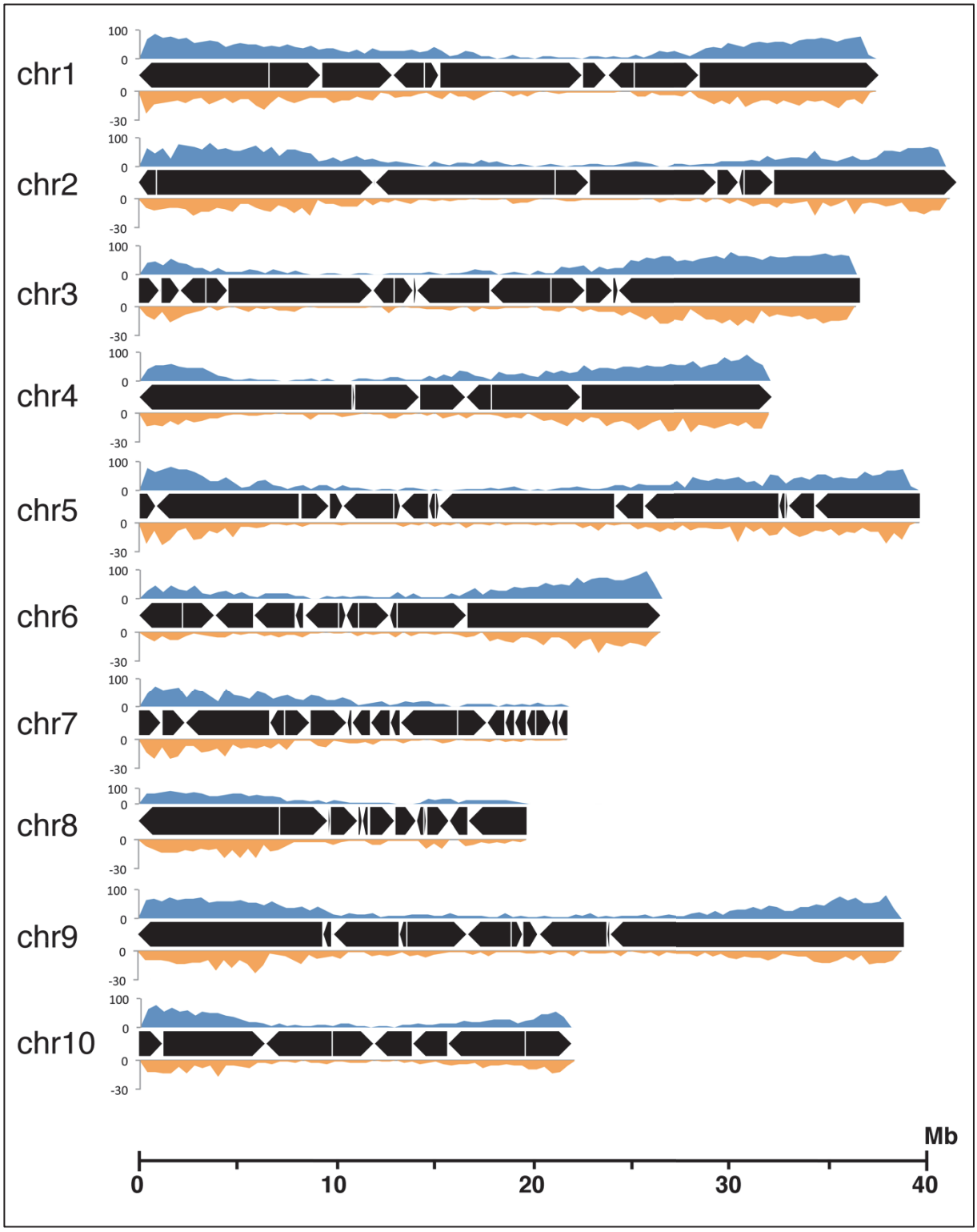


Figure 3: Anchored scaffolds into the 10 *Theobroma cacao* pseudochromosomes.

Black boxes represent scaffolds with orientation. Gene and SNP marker density are represented in blue and orange respectively and were computed with a window size of 400kb.

In comparison to the previous version of the Criollo genome assembly, we were able to anchor 95.8 Mb more DNA sequence into the 10 pseudochromosomes, leading to a significant reduction of the unknown chromosome (Tc00 in the genome v1) (Figure 4A). We also identified and corrected 45 misassemblies points distributed in the 10 pseudochromosomes of the first version of the assembly (Figure 4B). After gap closing, the proportion of unknown sites (Ns) was reduced from 10.8 % in the first version of the assembly to 5.7% in this new assembly (Table 3).

Annotation

The structural annotations of the protein-coding genes computed during the first draft genome project were transferred to the new assembly version. From the 28,798 predicted genes, 28,391 (98.6%) were relocated to the assembly version 2. A total of 91.6% of genes previously located in non-anchored scaffolds (version 1) were transferred to a known chromosome in version 2. Furthermore, 345 genes from the assembly version 1 were relocated to a different chromosome in assembly version 2. Another structural annotation, supporting evidence from RNA-Seq experiments, was carried out by the NCBI Refseq annotation system. The RefSeq annotation comprised 21,437 protein coding genes, 2,229 non coding genes and 1,165 pseudogenes.

The functional annotation was carried out for both structural annotations. The full annotations as well as a genome browser are available through the cacao genome hub (http://cacao-genome-hub.southgreen.fr).

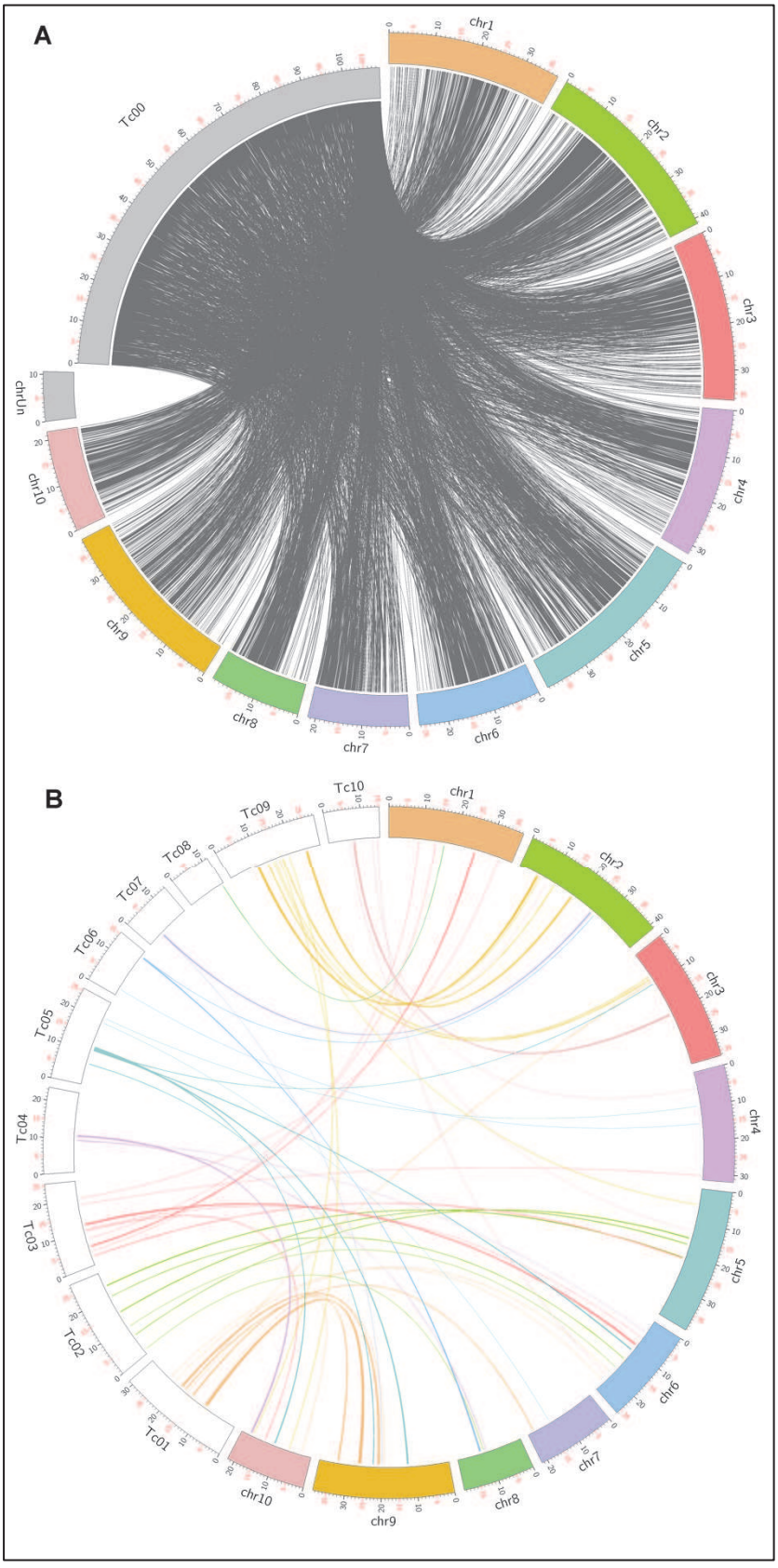


Figure 4: Comparison of *Theobroma cacao* Criollo assembly version 1 vs version 2.

A: graphical representation of insertions and reduction of the unknown chromosome version 1 (Tc00) into chromosomes version 2 (chr1-10). B: graphical representation of regions previously anchored to a different chromosome in the two version of assemblies. Chromosomes "Tc" refer to assembly version 1 and chromosomes "chr" to assembly version 2.

Discussion

The rapid evolution of NGS-based methods developed since the first draft genome sequence of *Theobroma cacao* Criollo B97-61/B2 published in 2011 [1] provided an opportunity to update the quality of the genome assembly.

During the first steps of the scaffolding process, the 4 mate-paired libraries, decreased the number of scaffold to 980 which is 80% less than the first published version of the genome (4,792) and increased 2 times the N50 size value (470kb vs 932kb). The addition of Bac End sequences lead us to reduce the number of scaffolds by almost 90% (4,792 vs 554) and to increase by 14 times the N50 size value (0,47 vs 6,5 Mb). Our result demonstrates the usefulness of mate-pair templates of large insert size to correct misassemblies and to reduce the number of scaffolds and consequently increase the size of contiguous sequences.
In the final assembly, we closed almost half of the gaps (34,6Mb vs 18,5Mb) with a combination of long reads sequences and large insert size mate paired libraries. Gaps closed by Pacific Bioscience reads, after error correction, represent 4,4Mb of sequence. The LR data would have had a bigger impact if they were generated with the P6-C4 chemistry which yields longer average read lengths compared to the C2 chemistry (average length of 2,573 nucleotides of the corrected LR dataset used in this study). The absence of un-overlapped read-pairs region inside gap filled scaffolds using the 3-5kb library highlighted the efficiency of the gap closure step.
The power of the Genotyping by Sequencing methods to produce a high number of SNP molecular markers was applied to increase the proportion of anchored assembly into chromosome to 96,7% (66,8% in version 1). Moreover, 99% of the genes are now anchored to chromosomes compared to 82% in the first assembly.

Sequences comparison of this new version of the B97-61/B2 genome with the Matina1-6 genome revealed a good colinearity between the two genomes assemblies (Figure 5). The main observed differences are scaffold inversions located in peri-centromeric regions of chromosome 2, 6, 7, 8 and 9.

	B97-61/B2 Version 1 (argout *et al.*, 2010)	B97-61/B2 Version 2	Mat 1-6 (Motamayor *et al.*, 2013)
Scaffold number	4792	554	814
Cumulated size(Mb)	326.9	324.7	346
N50(Mb)	0.47	6.5	4.3
Anchored on chromosomes (Mb)	218,4 (66,8%)	314.2 (96.7%)	330 (95.5%)
Unknown sites (Mb)	35.4 (10,8%)	18.5 (5,7%)	15.2 (4,4%)

Table 3: Comparative metrics of *T.cacao* assemblies

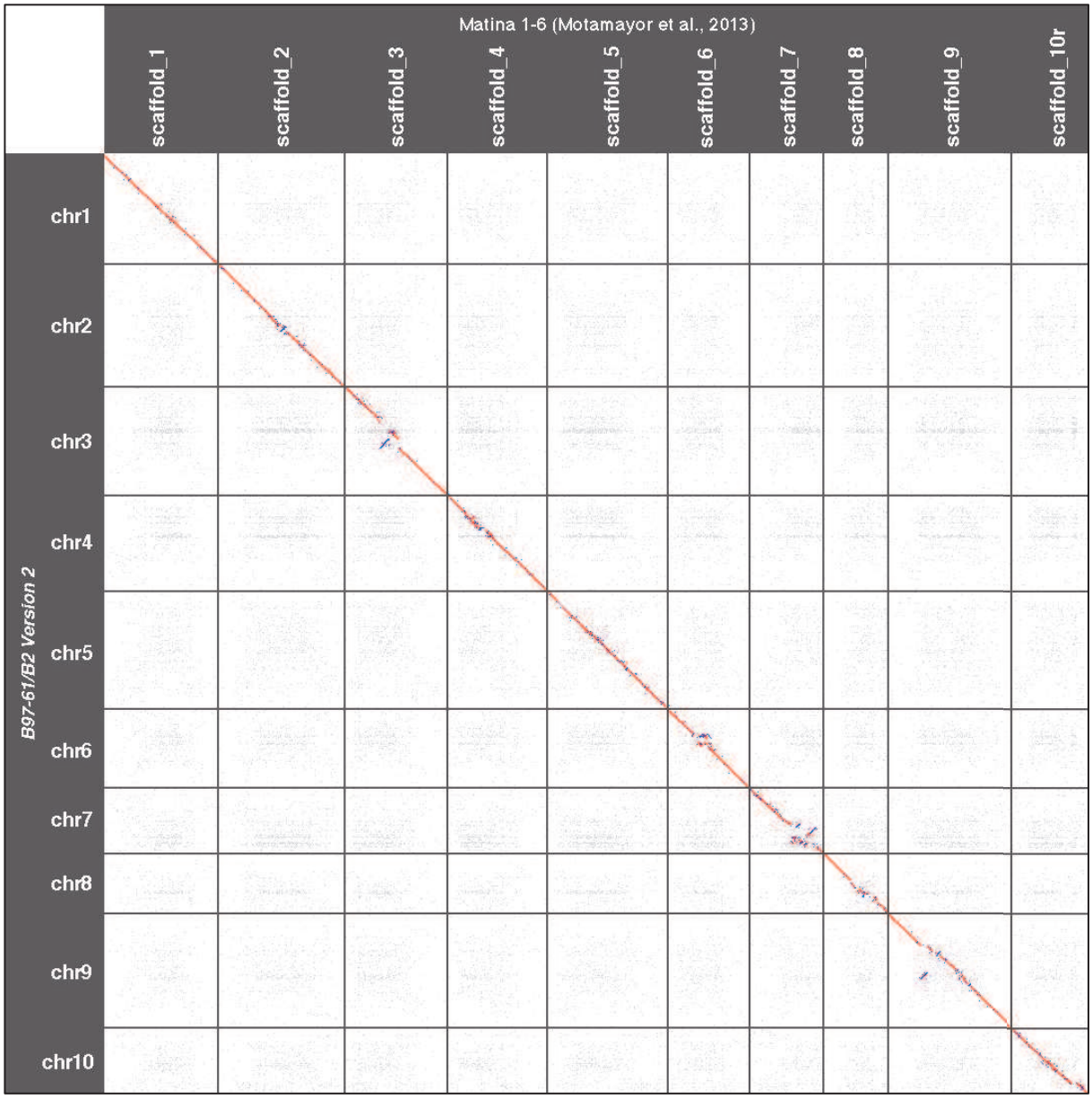
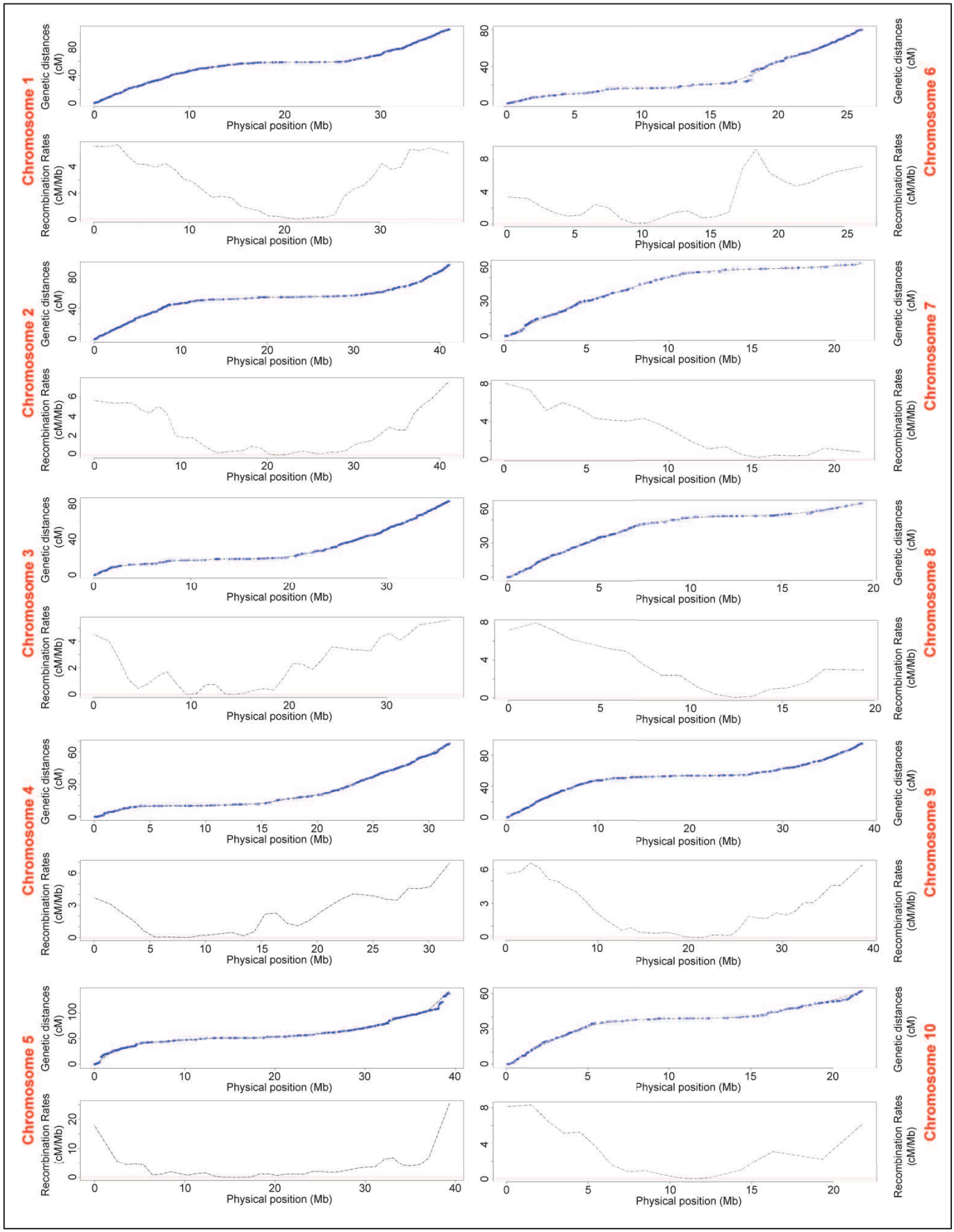


Figure 5: Dot plot comparison between Criollo B97-61/B2 version 2 and Amelonado Matina 1-6 genomes computed with Last (Kielbasa *et al.*, 2011).

Red and blue dots indicate forward and reverse alignments respectively.

Because both assemblies were anchored to chromosome using genetic information and low recombination rate observed in centromeric regions (Supplementary Figure 1), these differences can be explained by the difficulty to order and orient scaffolds in these regions were genotyping errors could be considered by the algorithm as recombination.

The corrections of misassemblies, resolutions of gaps, reductions of scaffold number and non-anchored regions and updates of the functional and structural annotations we report in this study for the first Theobroma cacao genome sequence published is an important step for future comparative genomics and genetics studied on cocoa.



Supplementary Figure 1: Genetic to Physical distance and Recombination rate for the 10 *T. cacao* chromosomes.

The estimation of recombination Rates (cM/Mb) was computed with they MareyMap R package (Rezvoy et al., 2007)

List of abbreviations

NGS: Next Generation Sequencing; WGS: Whole Genome shotgun Sequencing; TE: Transposable Element; GBS: Genotyping By Sequencing; SNP: Single Nucleotide Polymorphism; SR: Short Read; LR: Long Read; LPR: Large insert size Paired Read; BAC: Bacterial Artificial Chromosome.

Declarations

Availability of data and material

Datasets (contigs, scaffold assembly, Pseudo-molecules, makers matrix) are available through the cacao genome hub (http://cacao-genomehub.southgreen.fr/).

Competing interests

The authors declare that they have no competing interests

Authors' contributions

XA and GM conceived and designed the study. XA performed the analysis and wrote the manuscript. GM, GD and ER contributed to the analysis. KL and JMA produced the sequencing data. XA, GM, GD, KL, JMA and CL edited the manuscript. CL coordinated the study.

Acknowledgements

We would like to thank Valrhona for their financial contribution to this project.

References

1. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, et al. The genome of Theobroma cacao. Nat. Genet. 2011;43:101–8.
2. Solorzano RGL, Fouet O, Lemainque A, Pavek S, Boccara M, Argout X, et al. Insight into the Wild Origin, Migration and Domestication History of the Fine Flavour Nacional Theobroma cacao L. Variety from Ecuador. PLOS ONE. 2012;7:e48438.
3. Motamayor JC, Lachenaud P, Mota JW da S e, Loor R, Kuhn DN, Brown JS, et al. Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L). PLOS ONE. 2008;3:e3311.
4. Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, et al. Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of Theobroma cacao L. generated from various tissues and under various conditions. BMC Genomics. 2008;
5. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, III DL, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013;14:r53.
6. Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, et al. Genome Project Standards in a New Era of Sequencing. Science. 2009;326:236–7.
7. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat. Methods. 2011;8:61–5.
8. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLOS ONE. 2012;7:e47768.
9. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578–9.
10. Kosugi S, Hirakawa H, Tabata S. GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics. 2015;btv465.
11. Martin G, Baurens F-C, Droc G, Rouard M, Cenci A, Kilian A, et al. Improvement of the banana “Musa acuminata” reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics. 2016;17:243.
12. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19:1068–76.
13. Mascher M, Stein N. Genetic anchoring of whole-genome shotgun assemblies. Genomic Assay Technol. 2014;5:208.
14. Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT. Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Sticklebacks with Genotyping-by-Sequencing. G3 GenesGenomesGenetics. 2015;5:1463–72.
15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.
16. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;btu538.
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012;9:357–9.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215:403–10.
19. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18.
20. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLOS ONE. 2014;9:e90346.
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
22. Van Ooijen JW. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet. Res. 2011;93:343–349.
23. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol. Biol. Clifton NJ. 2007;406:89–112.
24. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17:847–8.
25. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D211-215.

26. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;gkw1092.
27. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J. Mol. Biol. 2016;428:726–31.
28. Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, et al. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am. J. Bot. 2012;99:320–9.
29. Liu G, Cao D, Li S, Su A, Geng J, Grover CE, et al. The Complete Mitochondrial Genome of Gossypium hirsutum and Evolutionary Analysis of Higher Plant Mitochondrial Genomes. Xu Y, editor. PLoS ONE. 2013;8:e69476.

2. Perspectives

La recherche en amélioration des plantes a grandement profité de l'avènement des séquençages de génomes entiers. Initialement, les génomes ont été séquencés avec la méthode Sanger, nécessitant des investissements considérables de temps, de main d'œuvre et de coût. Cependant, depuis l'achèvement de la séquence du génome d'*Arabidopsis thaliana* en 2000, les technologies de séquençage par les technologies de deuxième génération (NGS) ont augmenté la vitesse et l'échelle des capacité de séquençage, tout en réduisant très significativement les coûts.

Pour la recherche cacaoyère, cette évolution des technologies a permis dès 2013 de disposer de 2 séquences du génome de *Theobroma cacao* pour 2 variétés très contrastées génétiquement. Les informations apportées par ces génomes ont facilité la recherche biologique visant à caractériser la physiologie ou la biologie moléculaire de processus cellulaires d'intérêt agronomique, l'espace génique des chromosomes étant couvert à 98% (comparaison réalisée entre le génome et le transcriptome). Cependant ces deux initiatives de séquençage du génome du cacaoyer ont été largement basées sur des stratégies d'assemblage à partir de courtes lectures et ont conduit à des génomes relativement fragmentés. De plus, une grande partie des séquences génomiques du cacaoyer ne sont pas contenues dans ces deux assemblages. L'assemblage du génome Criollo couvre en effet 76% de la taille estimée du génotype B97-61/B2 et l'assemblage du génome Amelonado environ 72 %.

Comme nous venons de le voir dans le cadre de l'amélioration de la séquence du génome du Criollo, cette fragmentation peut en partie être résolue par l'apport de lectures de pairées issues de fragments génomiques de grandes tailles. Cependant, des bénéfices importants pour les futures études à mener sur le cacaoyer pourraient être apportés par la mise à disposition d'un assemblage chromosomique complet et continu. Par exemple, l'étude des gènes impliqués dans la résistance aux pathogènes chez le cacaoyer a démontré une organisation sous forme de clusters de gènes (Argout et al., 2011; Legavre et al., 2015).

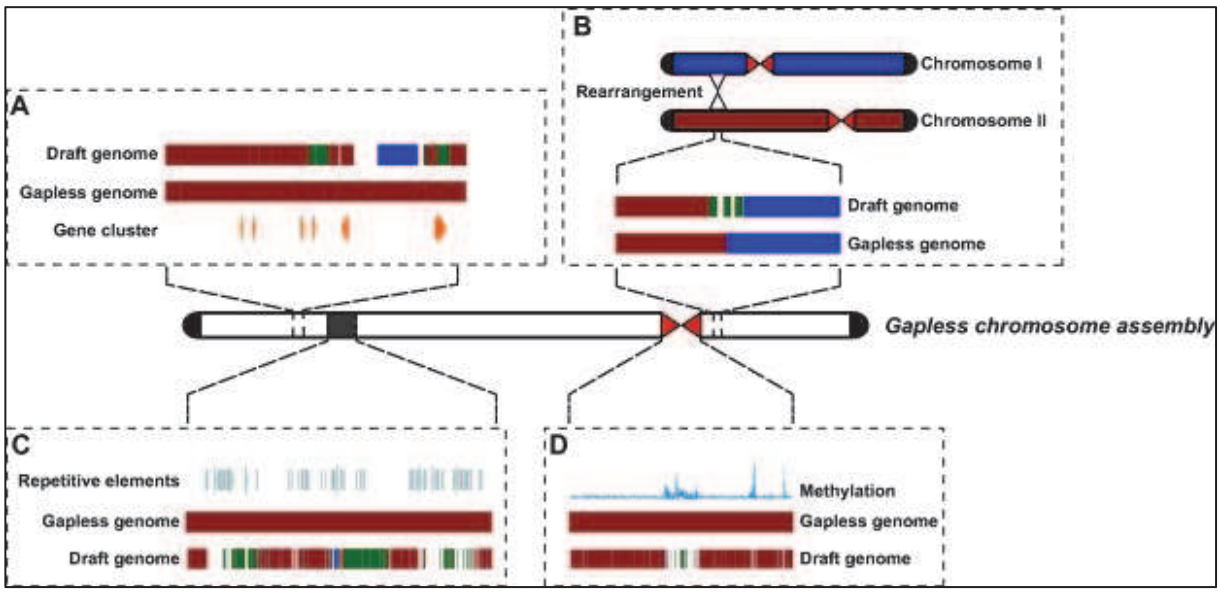


Figure 17 : Illustration schématique des avantages apportés par une séquence génomique complète (d'après Thomma et al., 2016). Les différentes couleurs des segments dans les 4 panels décrivent des chromosomes différents dans l'assemblage final. A. Exemple des clusters de gènes. B. Etude de synténie et réarrangements chromosomiques. C. Eléments répétés. D. Epigénétique.

Ces clusters de gènes sont souvent situés dans des régions riches en éléments répétés et dépendent par conséquent de la qualité de l'assemblage (Figure 17).
La résolution du génome dans les parties péri-centromériques reste encore un challenge car la reconstruction des pseudochromosomes, basée sur des marqueurs génétiques, souffre des faibles taux de recombinaisons observés dans ces régions. De même pour les régions télomériques qui contiennent beaucoup d'éléments répétés.

De nouvelles technologies dites de troisième génération, développées par des sociétés telles que Pacific Biosciences ou Oxford Nanopore produisent des lectures de grandes tailles (>10kb) et qui peuvent couvrir des régions répétées complètes (Huddleston et al., 2014). Cependant, l'assemblage des régions répétées comme celles localisées dans les régions centromériques restent difficile. La technique de cartographie optique (optical mapping), qui fournit des cartes de restriction à très haute résolution permet désormais de résoudre ces problèmes (Dong et al., 2013). L'utilisation de ces 2 technologies combinées pourrait permettre d'améliorer encore significativement l'assemblage du génome du cacaoyer.

CHAPITRE 4 - RECHERCHE DE GÈNES CANDIDATS IMPLIQUÉS DANS LA VOIE DE BIOSYNTHÈSE DES ANTHOCYANINES DES FÈVES DE CACAO

1. Introduction

Comme nous l'avons vu dans la révision bibliographique (Chapitre 1), les fèves de cacao sont exceptionnellement riches en polyphénols, et les anthocyanines représentent jusqu'à 4% des polyphénols totaux. La majorité des composés phénoliques sont stockés dans les cellules pigmentaires des cotylédons, conférant une couleur généralement violette aux fèves de cacao. Les bénéfices de ces molécules contenues dans le cacao sur la santé humaine ont été largement étudiés ces dernières années en médecine. Cependant, certaines variétés aromatiques naturelles, dont le Criollo, présentent des fèves complètement blanches et une teneur en polyphénols totaux beaucoup plus faible. Les fèves issues de ces variétés aromatiques sont en général vendues à un tarif plus avantageux pour les producteurs. Ainsi, il n 'est pas rare, dans des plantations hybrides issues de semis de type Trinitario (Criollo x Amelonado), comme à Madagascar, d'effectuer un tri sur la couleur des fèves après séchage.

Une descendance issue du croisement en 2 génotypes Trinitario (UF676 x ICS95) a été plantée dans la station expérimentale du CIRAD Paracou-Combi (Guyane) en 2010. Ces 2 clones Trinitario, fortement hétérozygotes, partagent des allèles en commun issus des 2 ancêtres Criollo et Amelonado. Les fèves issues de ce croisement contrôlé présentent une large gamme de couleurs, variant de complètement blanches à complètement violettes avec une gamme de couleurs intermédiaires rose pâle à rose foncé. Les cotylédons des fèves étant les deux premières feuilles du jeune plant, ils reflètent aussi le genotype de l'individu qui sera impliqué dans la formation de ses futures fèves. La couleur de chacune des fèves a été consignée avant semis et 434 individus de cette descendance ont été génotypés par la technique de Génotypage par Séquençage comme décrite au chapitre précédent.

Dans ce chapitre, une analyse QTL sera conduite pour localiser des régions génomiques impliquées dans la variation de la couleur des fèves de cacao. Puis, en s'appuyant sur la version améliorée du génome du Criollo, des gènes potentiellement impliqués dans la voie de biosynthèse des anthocyanines seront identifiés dans ces régions. Ces analyses sont présentées sous la forme d'un article scientifique.

Identification of candidate genes involved in anthocyanin biosynthesis in *Theobroma cacao* seeds

X. ARGOUT, B. GUITTON, F. DE LAMOTTE, P. LACHENAUD, O. FOUET, J.M. THEVENIN and C. LANAUD

Introduction

Theobroma cacao is a tropical perennial tree native to the Amazonian basin of South America that belongs to the *Malvaecae* family [1] and has a diploid chromosome number of 2n=2x=20. Molecular analyses have permitted to decipher the genetic diversity of *T. cacao* which includes 10 major genetic groups: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañón, Nanay, Nacional, and Purús [2,3].

Its beans (seeds) are used for chocolate, confectionary and cosmetic industries and are produced in humid tropical worldwide regions. Cocoa is mainly grown on small farms of 2 to 5 hectares and provides livelihoods for between 40 and 50 million farmers, rural workers and their families in the Global South (source: World Cocoa Foundation).

The genetic origin of the cultivated cocoa tree cultivated is very important for the quality of the cocoa product [4,5]. The market distinguishes between two broad categories of cocoa beans: "bulk or ordinary" beans and "fine or flavor" cocoa beans, mostly produced by Criollo and Nacional varieties.

Criollo cocoa seeds are morphologically different compare to other genetic groups: they are large, plump and white while in other varieties they are generally purple and flattened. In hybrid population like in Trinitario population (Criollo x Amelonado), colors of the seeds range from white to purple with a variety of intermediate tones.

The purple pigmentation of the seeds and of several other organs (leaves, staminodes, flowers) is due to anthocyanin pigments [6], one of the three main classes of flavonoids with flavonols and proanthocyanidins. Anthocyanins in cocoa comprise mainly cyanidin-3-*O*-galactoside and cyanidin-3-*O*-arabinoside [7], which may represent in fresh cocoa beans, 4% of the total phenolic compounds [5].

Anthocyanins and more generally flavonoids are plant secondary polyphenolic metabolites and their antioxidant properties, particularly those from chocolate, have gained attention in the last years for their potential to protect human health against cardiovascular disease and cancer [8,9]. Cocoa beans are particularly rich in phenolic compounds and flavonoids (up to 8% of dry seed weight) [10]. They participate in reactions of protein hydrolysis or polymerization during cocoa fermentation leading to precursors of aromatic compounds. However, in Criollo cocoa seeds, the phenolic compounds content is only 2/3 of the amount of these compounds found in other varieties [4,11–13] and Criollo seeds no contain anthocyanins in their composition [7].

The biosynthesis pathway of flavonoids and its regulation has been very well described in plants [14]. Three R3-R3-MYB proteins control flavonol biosynthesis via activating the early biosynthetic genes whereas the production of anthocyanins and proanthocyanidins requires a MYB-bHLH-WD40 (MBW) complex to activate the late biosynthetic genes. Additional regulators of flavonoid biosynthesis recently came to light which interact with R2RM-MYBs or bHLHs to organize or disrupt the formation of the MBW complex, leading to enhanced or compromised flavonoid production.

In *T. cacao*, the genetic control of seed color is poorly understood. In 1931, Wellensiek [15] suggested the role of one major dominant gene responsible of seed pigmentation in hybrid Trinitario populations.

Recently, Marcano et al., 2009 [16] identified 3 significant associations located in linkage group 1, 4 and 10 between molecular markers and seed color trait in cultivated population of Modern Criollo.

The recent availability of the two well annotated cacao genomes from two highly homozygous genotypes: ‘B97-61/B2’, a pure Criollo genotype [17] and ‘Matina 1–6’ a pure Amelonado genotype [18] gives an ideal opportunity to further investigate the genetic basis of anthocyanin production in cocoa seeds by comparative genomics. For example 96 *T. cacao* genes orthologous to *Arabidopsis* genes involved in the flavonoid biosynthetic pathway have been previously identified in the Criollo genome [17] and one R2R3 MYB transcription factor candidate gene for the regulation of pod color has been identified in the Amelonado genome [18].

Here, we performed a cross between two Trinitario genotypes that share common pure Criollo and Amelonado ancestors and studied the segregation of seed color in the progenies linked to marker segregation revealed by Genotyping By Sequencing (GBS). Using a large progeny, we were able to build a highly saturated genetic map and to detect QTLs associated with seed color traits and identified a candidate gene strongly associated with this trait.

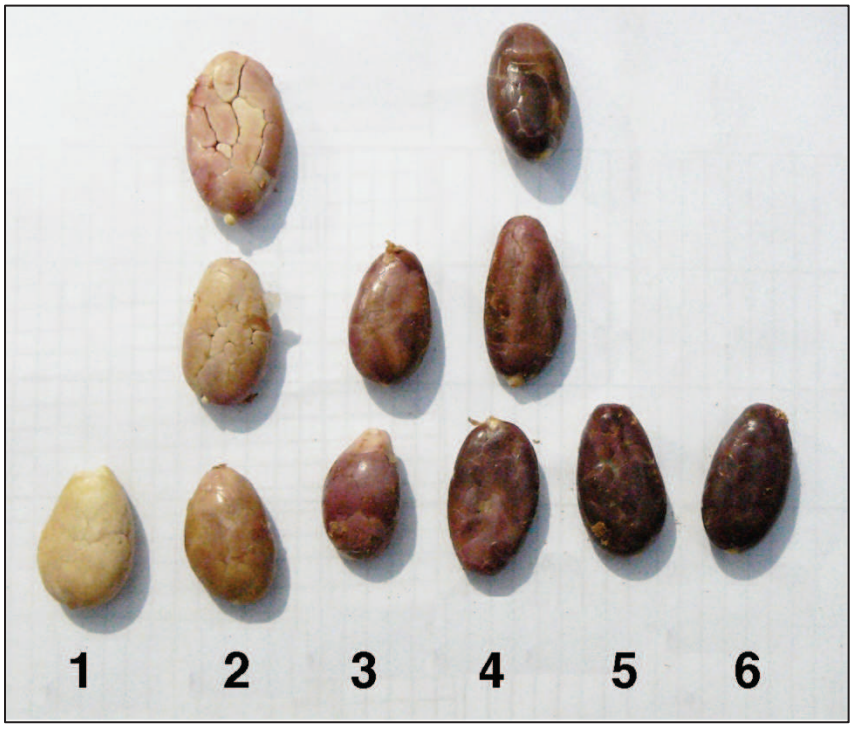


Figure 1: color scale used to score seed pigmentation.

Methods

Plant material and seed color observation

The mapping population, planted at CIRAD experimental station of Paracou-Combi (French Guiana), consisted of 434 individuals originated from the cross between UF676 x ICS95, two Trinitario clones hybrid between Amelonado and Criollo types. The color of the seeds was measured before seeding. The pigmentation was scored with the scale defined by Marcano et al., 2008 with ascending values from the lightest to the darkest tones. Six levels of tones were recorded (Figure 1). Levels 1 and 2 corresponded to white and light pink, level 3 to intermediate tone and gradually darker levels 4, 5 and 6 for fully purple seed.

Genotyping by Sequencing data

The 434 individuals from the cross between UF676 and ICS95 were sequenced by the Diversity Arrays Technology compagny, using Illumina Hiseq2000 instrument after DNA restriction with enzymes *Pst*I and *Mse*I. Sequencing fragments were analyzed using Tassel 5 GBS v2.2.24 pipeline [19], and parameter (-mnQS 20). Reads were aligned to the Criollo B97-61/B2 genome version 2 [20] using Bowtie2 (end-to-end algorithm) and in -very-sensitive mode. Reads that aligned at different locations of the genome were discarded. SNPs were called and variant call data were filtered out with VCFtools (ref). First, indels and non-biallelic sites were excluded. Then, genotyped data with less than 10 reads were recoded as missing data and SNPs with more than 50% of missing data were excluded. Finally SNPs with minor allele frequency > 0.01, P-value > 1e-6 (X^2 test) and with a minimum distance of 64bp were selected for further analysis.

Genetic linkage map

Sequencing errors can lead to a small percentage of falsely called genotypes and as a consequence, false recombination breakpoints may appeared in the genotyping data [21]. To address this issue, we developed a modified version of the sliding-window approach applied by Spindel et al., 2013.

We used the information of the 6 surrounding SNPs to detect and correct potential error breakpoints. Consecutive SNPs markers with the same genotype information in the 434 individuals were then binned together.

The genetic linkage map was build with the maximum likelihood mapping algorithm of JoinMap 4.1. A fixed marker order was set according to the physical order of markers along chromosomes.

QTL mapping for seed color

First, the non-parametric Kruskal–Wallis (KW) test implemented in MapQTL version 6.0 (Kyazma software) was employed to detect association between all markers and trait.

In a second step, to better estimate the parameters of the multiple putative QTLs detected with the non parametric method, we conducted Multiple QTL Mapping with R/qtl package. Because the current implementation of R/qtl-MQM is limited to experimental crosses F2, we first selected markers that were heterozygous in both parents. The Composite Interval Mapping method (CIM) was applied with a mapping step of 1 cM (function *calc.genoprob*). To identify the multiple QTL model with maximal LOD score while controlling false positive rates, the *stepwiseqtl* function was applied. The function performs forward/backward model selection using a penalized likelihood approach to compare different sizes models, with penalties on QTL and pairwise interactions [22]. Penalties for the penalized LOD scores were derived for each trait on the basis of permutation results (1000 permutations) from a two-dimensional genome scan with a two-QTL model allowing covariates (functions *scantwo* and *calc.penalties*). The model optimizing the penalized LOD score criterion was fitted with the *fitqtl* function to get QTL estimated effects. For each detected QTL, positions with maximum likelihood were identified (function *refineqtl*) before approximate 95% Bayesian credible intervals were calculated (function *bayesint*).

Identification of candidate genes in QTL regions

Using markers flanking the 95% Bayesian credible intervals of the QTLs, we located the corresponding physical position in the Criollo reference genome and extracted all genes within this interval.

A set of literature-curated plant genes supporting experimental evidences within anthocyanin biosynthesis, accumulation and regulation was used as a reference for finding candidate genes responsible to seed color variation. Sequence similarity between *T. cacao* genes and the reference dataset was computed by BlastP, with an Evalue threshold of 1e-80.

Three-dimensional structure modeling

Theoretical structures have been calculated using the @tome2 suite of programs to perform homology modeling [23]. A structural alignment was performed using PsiBalst (blast.ncbi.nlm.nih.gov) against the Protein Data Bank (rcsb.org) and structural models were generated using the following templates (pdb ID: 1CHW, 1CGK, 1U0V and 1i8B). All of these templates were structures of Chalcone or Stilbene synthase and presented a high similarity with both Criollo and Amelonado protein coding sequences (79 and 83% respectively). The quality of each final structure model has been evaluated by a set of several tools including Qmean [24].

Promoter sequence analysis

Sequence analysis of promoters was carried out with the Plant Promoter Analysis Navigator [25].

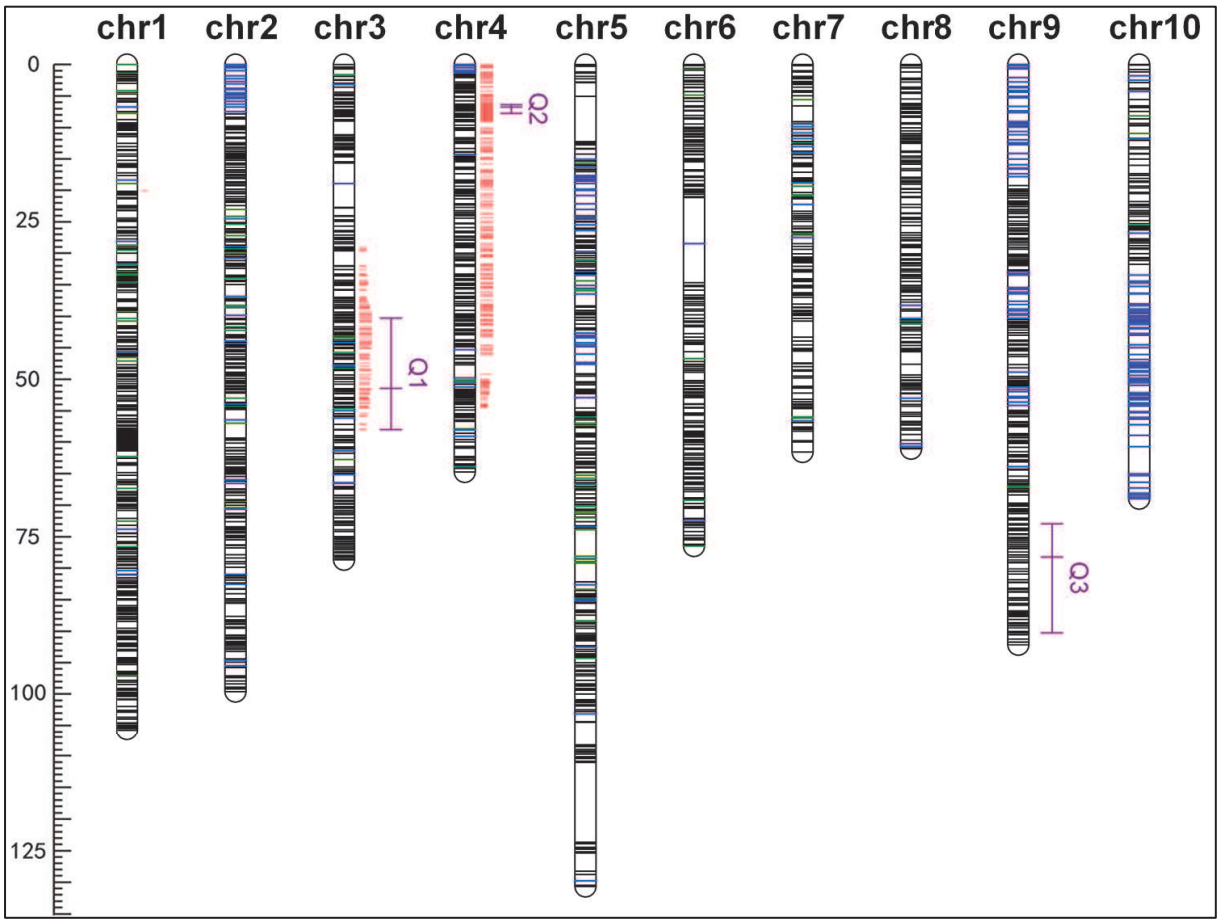


Figure 2: Linkage map of the UF676xICS95 mapping population and QTLs detected for seed color.
The vertical scale line indicates genetic distance in centimorgan (cM). The vertical bars display the 10 linkage groups. Lines in each linkage group indicate marker positions. Black lines represent markers segregating in both parents, green lines markers segregating only for ICS95 parent and blue lines markers segregating only for UF676 parent. Red bars indicate significant associations with Kruskal and Wallis method. Purple lines indicate MQM QTLs confidence intervals.

Linkage Group	Length (cM)	Total number of markers	Average distance between adjacent markers (cM)
chr1	106.2	435	0.27
chr2	100.0	388	0.28
chr3	79.0	305	0.30
chr4	65.0	297	0.26
chr5	131.3	396	0.43
chr6	76.8	232	0.37
chr7	61.8	192	0.39
chr8	61.3	190	0.36
chr9	92.6	336	0.34
chr10	69.3	197	0.44
Total	843,3	2968	0.33

Table 1: Distribution of SNP markers in the Linkage Groups

Results and discussion

SNP filtering and Linkage map construction

The 434 individuals from the cross UF676xICS95 were genotyped using 4,857 SNPs in the frame of the T.cacao Criollo B97-61/B2 genome assembly version 2 [20]. For each Individual and for each chromosome, we applied a sliding window to detect putative false recombination breakpoints due to sequencing error. Our algorithm identified 0.3% genotyping error in the whole dataset, which were recoded as missing data. From the 4,857 SNP markers, 1,889 SNP markers brought the same genotyping information and were binned with adjacent markers for map construction.

The linkage map presented in Figure 2 was computed from the 2,968 binned SNP markers. The ten linkage groups were clearly identifiable at minimal LOD score of 5 and were named according to their corresponding chromosomes (chr1-10). The map comprised a length of 843.3 cM (Table 1) with an average distance between adjacent markers of 0.33 cM. The linkage group lengths ranged from 61.3 cM to 131.3 cM for chr8 to chr5 respectively. The number of mapped loci varied substantially between linkage groups from 190 to 435 for chr8 and chr1 respectively.

Compared to previously published genetic maps, this map was 89.4 cM bigger than the composite linkage map generated by Allegre et al., 2012 [26] with 1,240 SNPs and 424.7 cM smaller than the CATIE type 2 linkage map [27] comprising 2,589 SNPs. Otherwise, the average distance between adjacent markers was 4 times shorter than the map of Allegre et al, 2012 and 6 times shorter than the CATIE type 2 map.

QTL analysis and candidate genes for seed color trait

A first QTL screening by non parametric mapping methods (Kruskal and Wallis) was carried out with the 2968 binned SNP. At significance threshold of 0.005, we identified 1, 116 and 261 positive marker associations with seed color trait on linkage group chr1, chr3 and chr4 respectively (Figure 2). We observed that significant marker/color trait associations detected by the Kruskal and Wallis method were located in large genetic regions segregating for both parents.

Chromosome	Significance threshold	LOD peak	Add.	Dom.	Variance explained (%)	Closest marker		Left flanking marker		Right flanking marker		Interval	
						Position (cM)	Position (bp)	Position (cM)	Position (bp)	Position (cM)	Position (bp)	cM	Mb
chr3	3.7	4.0	0.28	-0.06	2.3	51.74	30329741	40.51	27462813	59.24	32057260	18.73	4.59
chr4	3.7	54.2	1.15	0.56	41.4	6.95	3158436	6.49	2774003	7.87	5777434	1.39	3.00
ch9	3.7	5.2	0.31	-0.16	3.0	78.61	35893255	73.28	34637976	90.45	38184711	17.17	3.55

Table 2: QTLs detected by MQM for seed color

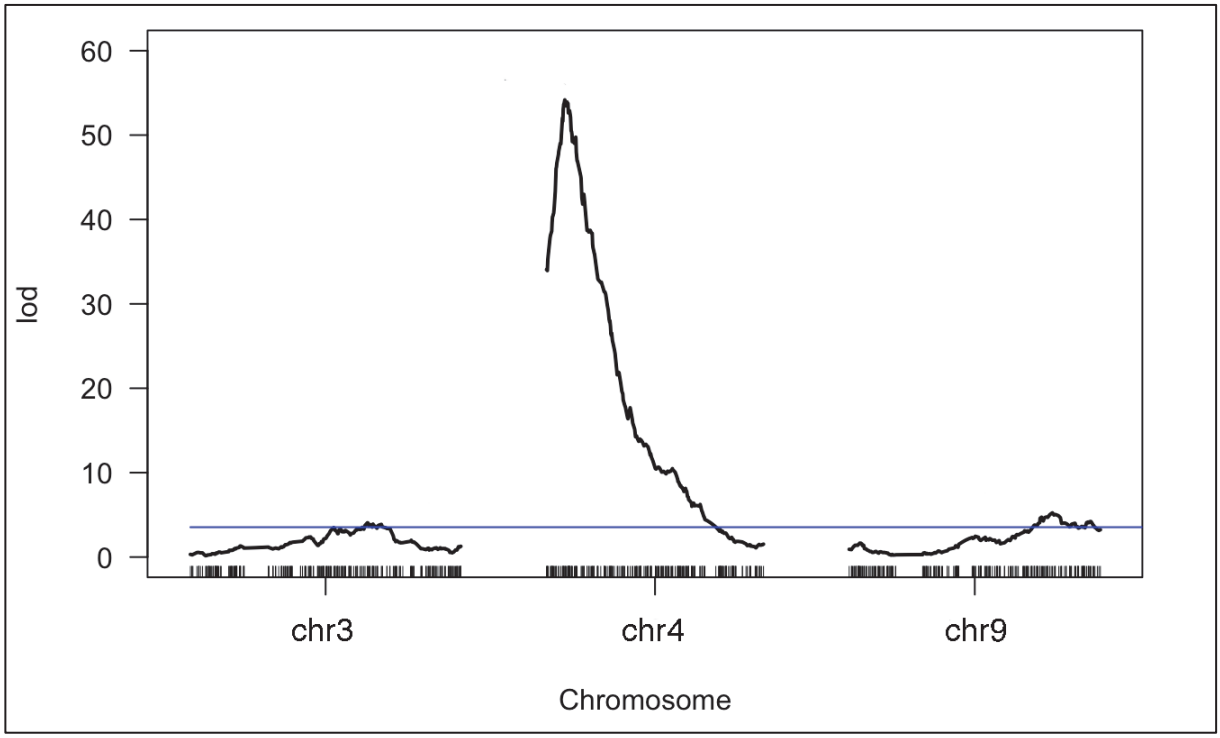


Figure 3: LOD profiles plot for each QTL. Blue line indicate significant LOD threshold.

To better estimate the parameters of the multiple putative QTLs detected with the non parametric method, we then selected from the whole dataset markers, 2138 markers that were heterozygous in both parents (*i.e.* AB/AB where A and B are alleles from the two ancestors Criollo and Amelonado) and performed Multiple QTL Mapping analysis.

The penalized likelihood approach we used lead to a model of 3 QTLs in 3 distinct linkage groups without interaction (Formula: y ~ Q1 + Q2 + Q3) explaining in total 48% of phenotypic variation (Table 2). LOD profiles for each QTL are presented in Figure 3.

A major QTL (Q2) located in the top of chr4 (6.9cM), was characterized by a very narrow confidence interval (1.4cM), high LOD score (54.2) and explained 41,4% of the phenotypic variation. We detected a positive dominance effect for this QTL, indicating that alleles promoting purple color were partially dominant to those promoting white color.

Two other significant QTLs were identified in chr3 (51.7 cM) and chr9 (78.6 cM) but with smaller LOD score (4.0 and 5.2 respectively). They were characterized by a larger confidence interval than Q2 (18.7cM and 17.2cM for Q1 and Q3) and explained 2.3% and 3.0% of the phenotypic variance respectively.

To further analyze the major effect of the QTL located in chr4, we then looked for putative candidate genes in the neighboring regions of the SNP loci associated with the QTL. The LOD peak surrounding the QTL on chr4 spanned a 95% Bayesian credible interval of 1.4 cM which corresponds to a region of 3Mb on chromosome chr4 (i.e. from 2.8 to 5.8 Mb). A total of 117 genes were identified within this physical interval. Their protein coding sequences were extracted and compared to a manually curated dataset of protein supporting experimental evidences within anthocyanin biosynthesis and accumulation.

In this region, two genes models, Tc04v2_g004230 (acc. number LOC18601149) and Tc04v2_g004950 (acc. number LOC18601254) showed significant similarity with genes involved in anthocyanin biosynthesis and accumulation.

The first gene, Tc04v2_g004230 is located 87,2 kb downstream from the QTL peak and encodes a chalcone synthase (CHS) a key enzyme that catalyze the initial step of the flavonoid biosynthesis [28].

The *Theobroma cacao* Criollo protein exhibited 77% of identity with the *Arabidopsis thaliana* CHS TT4 protein (AT5G13930), a protein identified in mutants deficient in flavonoid and required for the accumulation of purple anthocyanins [29].We also found a very good similarity (84% identity) with the *Matthiola incana* CHS protein (AJ427536), identified in white-flowering mutant lines [30].

The second gene, Tc04v2_g004950 encodes a Glutathione S-transferase (GST) and is located 2.2 Mb downstream the QTL peak. GST proteins have been previously described in mutants in maize *bronze 2* [31], petunia *AN9* [32] and *Arabidopsis thaliana* Transparent Testa 19 (TT19) to be required for efficient anthocyanin export from the site of synthesis in the cytoplasm into permanent storage in the vacuole. In mutant tissues, anthocyanin accumulates in the cytoplasm where it undergoes oxidation and polymerization reactions, the oxidized products appear brown instead of the bright colors typical of vacuolar anthocyanins. The *Theobroma cacao* Criollo protein exhibits 56% of identity with both petunia AN9 (CAA68993.1) and *Arabidopsis thaliana* TT19 (AT5G17220) proteins.

In *Theobroma cacao* Criollo white seeds, no anthocyanins have been found and the amount of polyphenols is approx. 2/3 of the *Theobroma cacao* Amelonado purple seeds [11]. In that context the hypothesis of a failure in the vacuolar storage of anthocyanins appears less plausible than the assumption of a breaking point in the flavonoid biosynthetic pathway. Therefore, due to its localization on the flavonoid biosynthesis pathway, the CHS gene, Tc04v2_g004230, seems to be a particularly promising candidate gene for seed color variation.

In the Amelonado genome [18], Tc04v2_g004230 is annotated as TCM_017370. Sequence comparison of the CHS coding sequences of two genotypes revealed several polymorphisms. From them, one SNP in the second exon is responsible of an amino acid substitution in position 390 of the traduced protein (Serine in Criollo vs Proline in Amelonado).

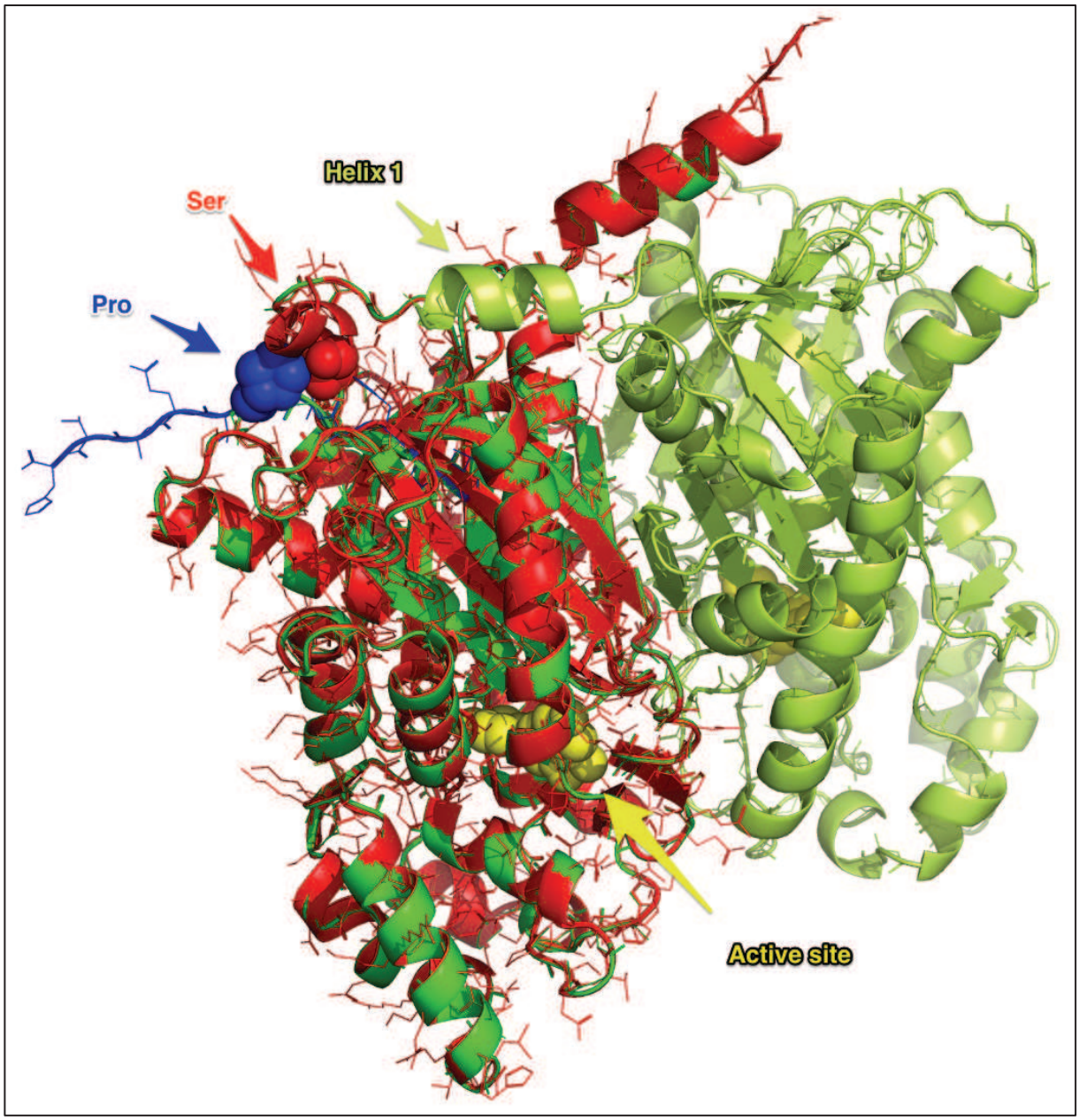


Figure 4: Structural representation of 1CGK CHS-nargenin complex (Ferrer et al., 1999) with superposition of both Amelonado and Criollo models.

The homodimer 1CGK is displayed in cartoon representation. The first subunit is colored in green and the other one in lime. *T.cacao* models are superposed in red. The ligand located in the active site is displayed with yellow spheres. Red spheres represent Criollo serine residue while blue spheres display Amelonado proline residue.

The CHS enzyme, also known as type III Polyketide synthase (PKS), function as homodimeric iterative PKS (monomer M_r≈42 – 45 kDa) with two independent active sites that catalyze a series of decarboxylation, condensation, and cyclization reactions [33]. To investigate the effect of this amino acid substitution in the structure of the CHS homodimer protein, we carried out a three-dimensional structure model of both Criollo and Amelonado proteins and studied the conformational changes (Figure 4). As anticipated, the 2 models were extremely similar (RMSD 0.175) and exhibited a very good Qmean value (0.78 and 0.80 respectively).

The main differences were located nearby the amino acid substitution area and were mainly due to the geometrical differences conferred by the Amelonado Proline 390 (cyclic side chain with limited movements) or Criollo Serine 390. We observed a different conformation of the backbone comprising the 6 amino acids after the substitution. This area was located opposite to the active site surrounding the ligand (illustrated with yellow spheres in Figure 4) and thus, no catalytic effect of this mutation might be anticipated. We observed that the amino acid substitution of one subunit is located in the axis of the first alpha helix (amino acids 1 to 12) of the second subunit and might be involved in the stabilization of this helix dipole [34]. Any change in this area may weaken the interaction between the two subunits and then impede the activity of the synthase. If this mutation has an effect on CHS activity, it might be linked to the dimerization required for CHS function.

Otherwise, it is also well known that the CHS gene is transcriptionally regulated by 3 closely related transcriptional R2R3-MYB proteins, MYB11, MYB12 and MYB111. Therefore, we carried out a comparative analysis of Amelonado and Criollo CHS promoters and searched for putative MYB homologous cis-regulation elements in the promoter region. The comparative analysis of 1000 bp upstream the ATG shows 96% of identity between the two genotypes sequences. We found one deletion of 12 nucleotides (-330 bp upstream of the ATG) in the Criollo CHS. The promoter sequences contain several consensus eukaryotic regulatory domains such as TATA-box-like sequence (-204 bp upstream of the ATG) and CAAT-box-like sequence (-177 bp upstream of the ATG). The analysis of the MYB transcription factor binding sites (TFBSs) identified several putative motifs into both Amelonado and Criollo sequences (Figure 5).

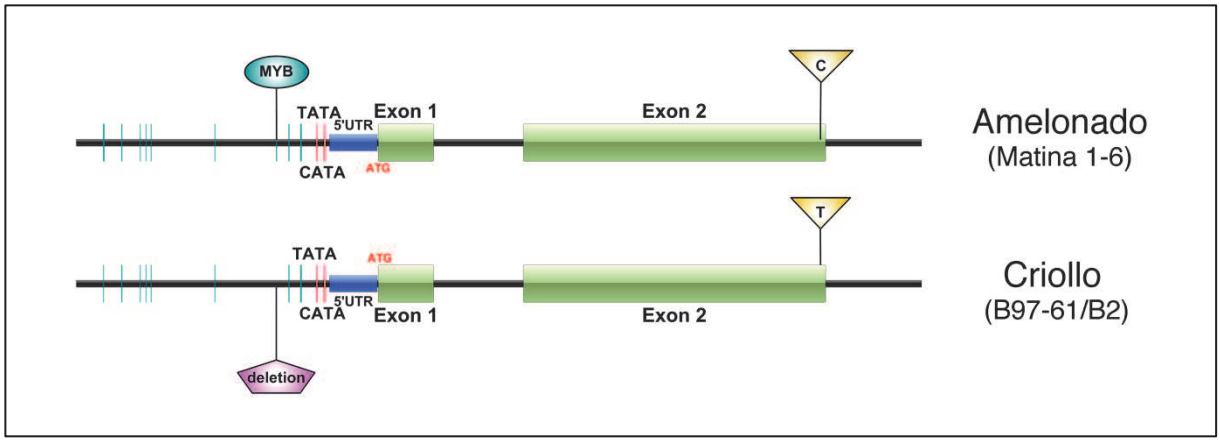


Figure 5: Structural annotation of CHS promoter and coding sequence in Matina 1-6 and B97-31/B2 genomes. All domains are drawn to the scale using IBS (Illustrator for Biological Sequences, Wenzhong Liu, 2015). In the promoter region, blue lines and red lines represent MYB TFBSs and TATA/CATA boxes respectively.

One of them is located in the deleted region identified after sequence comparision of both genotypes. This MYB TFBS, has been previously identify in CHS promoter of *Petunia hybrida* to be involved in the regulation of the flavonoid pathway [35] and in *Malus* crabapple to induce anthocyanin biosynthesis [36].

Conclusion

The high density genetic map built from the cross UF676 x ICS95 combined with QTL analyses for seed color provide a starting framework to decipher molecular mechanisms involved in anthocyanin biosynthesis in *Theobroma cacao* seeds. In the confidence interval of the major QTL located in chr4, we identified a putative candidate gene encoding a chalcone synthase. We observed two different structures between the two genotypes Criollo (white seeds) and Amelonado (purple seeds) located in CHS coding sequence and promoter. According to the analyses we have carried out, we can formulate the two following hypotheses to explain the trait variation:

Hypothesis 1: the amino acid substitution observed in the translated protein could be responsible of a conformational change that could prevent the dimerization of the protein or either modify the formation of transient multi-protein complexes in the Criollo flavonoid biosynthetic pathway

Hypothesis 2: the missing region in the promoter sequence of the Criollo genotype that contain a MYB TFBS could be responsible for a decrease of CHS gene expression in Criollo genotypes.

Both hypotheses lead to a significant reduction of chalcone product in Criollo genotype, mandatory for downstream enzymes to produce a number of biologically important compounds, including anthocyanins.

Further studies have to be conducted to validate the involvement of this CHS gene in the variation of anthocyanins observed between Amelonado and Criollo genotypes. Transcript profiling during developmental stage of the seed and transient expression assays after cloning the Criollo/Amelonado CHS genes and promoters would confirm the role of this CHS in the accumulation of anthocyanins.

References

1. Alverson WS, Whitlock BA, Nyffeler R, Bayer C, Baum DA. Phylogeny of the core Malvales: evidence from ndhF sequence data. Am. J. Bot. 1999;86:1474–86.

2. N'Goran J, Laurent V, Risterucci A, Lanaud C. Comparative genetic diversity studies of *Theobroma cacao* L. using RFLP and RAPD markers. Heredity. 1994;73:589–97.

3. Motamayor JC, Lachenaud P, Mota JW da S e, Loor R, Kuhn DN, Brown JS, et al. Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L). PLOS ONE. 2008;3:e3311.

4. Clapperton J. A review of research to identify the origins of cocoa flavour characteristics. Cocoa Grow. Bull. 1994;48:7–16.

5. Oracz J, Zyzelewicz D, Nebesny E. The Content of Polyphenolic Compounds in Cocoa Beans (*Theobroma cacao* L.), Depending on Variety, Growing Region, and Processing Operations: A Review. Crit. Rev. Food Sci. Nutr. 2015;55:1176–92.

6. Forsyth WGC, Quesnel VC. Cacao glycosidase and colour changes during fermentation. J. Sci. Food Agric. 1957;8:505–9.

7. Elwers S, Zambrano A, Rohsius C, Lieberei R. Differences between the content of phenolic compounds in Criollo, Forastero and Trinitario cocoa seed (*Theobroma cacao* L.). Eur. Food Res. Technol. 2009;229:937–48.

8. Spencer JP. Flavonoids and brain health: multiple effects underpinned by common mechanisms. Genes Nutr. 2009;4:243–250.

9. Rimbach G, Melchin M, Moehring J, Wagner AE. Polyphenols from cocoa and vascular health-a critical review. Int. J. Mol. Sci. 2009;10:4290–309.

10. Tomas-Barberán FA, Cienfuegos-Jovellanos E, Marín A, Muguerza B, Gil-Izquierdo A, Cerdá B, et al. A New Process To Develop a Cocoa Powder with Higher Flavonoid Monomer Content and Enhanced Bioavailability in Healthy Humans. J. Agric. Food Chem. 2007;55:3926–35.

11. Wollgast J, Anklam E. Polyphenols in chocolate: is there a contribution to human health? Food Res. Int. 2000;33:449–59.

12. Nazaruddin R, Seng LK, Hassan O, Said M. Effect of pulp preconditioning on the content of polyphenols in cocoa beans (*Theobroma cacao*) during fermentation. Ind. Crops Prod. 2006;24:87–94.

13. Jalil AMM, Ismail A. Polyphenols in Cocoa and Cocoa Products: Is There a Link between Antioxidant Properties and Health? Molecules. 2008;13:2190–219.

14. Li S. Transcriptional control of flavonoid biosynthesis: Fine-tuning of the MYB-bHLH-WD40 (MBW) complex. Plant Signal. Behav. 2014;9:e27522.

15. Wellensiek S. The genetics of cotyledon colour of cocoa as a basis for quality selection. Transl. H Toxopeus Arch. Voor Koffiecult. Ned.-Indië Buitenzorg Java. 1931;

16. Marcano M, Morales S, Hoyer MT, Courtois B, Risterucci AM, Fouet O, et al. A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (*Theobroma cacao* L.) population. Tree Genet. Genomes. 2009;5:329–337.

17. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, et al. The genome of *Theobroma cacao*. Nat. Genet. 2011;43:101–8.

18. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, III DL, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013;14:r53.

19. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLOS ONE. 2014;9:e90346.

20. Argout X, Martin G, Droc G, Labadie K, Rivals E, Aury J-M, et al. The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. Prep. 2017;

21. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19:1068–76.

22. Manichaikul A, Moon JY, Sen Ś, Yandell BS, Broman KW. A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses, Allowing Epistasis. Genetics. 2009;181:1077–86.

23. Pons J-L, Labesse G. @TOME-2: a new pipeline for comparative modeling of protein–ligand complexes. Nucleic Acids Res. 2009;37:W485–91.

24. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins Struct. Funct. Bioinforma. 2008;71:261–77.

25. Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, et al. PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. Nucleic Acids Res. 2015;gkv1035.

26. Allegre M, Argout X, Boccara M, Fouet O, Roguet Y, Bérard A, et al. Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L. DNA Res. 2012;19:23–35.

27. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, et al. Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. DNA Res. 2015;22:279–91.

28. Winkel-Shirley B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiol. 2001;126:485–93.

29. Shirley BW, Kubasek WL, Storz G, Bruggemann E, Koornneef M, Ausubel FM, et al. Analysis of Arabidopsis mutants deficient in flavonoid biosynthesis. Plant J. 1995;8:659–71.

30. Hemleben V, Dressel A, Epping B, Lukačin R, Martens S, Austin M. Characterization and structural features of a chalcone synthase mutation in a white-flowering line of Matthiola incana. Plant Mol. Biol. 55:455–65.

31. Marrs KA, Alfenito MR, Lloyd AM, Walbot V. A glutathione S-transferase involved in vacuolar transfer encoded by the maize gene Bronze-2. Nature. 1995;375:397–400.

32. Mueller LA, Goodman CD, Silady RA, Walbot V. AN9, a Petunia Glutathione S-Transferase Required for Anthocyanin Sequestration, Is a Flavonoid-Binding Protein. Plant Physiol. 2000;123:1561–70.

33. Jez JM, Ferrer J-L, Bowman ME, Austin MB, Schröder J, Dixon RA, et al. Structure and mechanism of chalcone synthase-like polyketide synthases. J. Ind. Microbiol. Biotechnol. 2001;27:393–8.

34. Hol WGJ, van Duijnen PT, Berendsen HJC. The α-helix dipole and the properties of proteins. Nature. 1978;273:443–6.

35. Solano R, Nieto C, Avila J, Cañas L, Diaz I, Paz-Ares J. Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrida. EMBO J. 1995;14:1773–84.

36. Tian J, Shen H, Zhang J, Song T, Yao Y. Characteristics of chalcone synthase promoters from different leaf-color malus crabapple cultivars. Sci. Hortic. 2011;129:449–58.

2. Perspectives

Dans ce chapitre, nous avons pu mettre en évidence 3 régions du génome localisées sur 3 chromosomes différents potentiellement impliquées dans la variation de la couleur des fèves dans une descendance issue du croisement entre deux Trinitario. Une région localisée sur le chromosome 4, déjà identifiée précédemment par une étude d'association réalisée sur une population de Criollo moderne (Marcano et al., 2009), semble expliquer une grande partie de la variation phénotypique observée. Dans l'intervalle de confiance du QTL, un gène codant pour une Chalcone synthase parait être un gène candidat particulièrement intéressant. Cependant, l'origine de la couleur des fèves ne semble pas être monogénique et des gènes localisés sur les chromosomes 3 et 9 semblent également participer à la variation du caractère.

Des analyses complémentaires sont actuellement en cours afin de préciser les mécanismes moléculaires intervenant dans la variation de la couleur des fèves de cacao. D'une part, il est nécessaire de préciser les intervalles de confiance de chacun des QTLs détectés, pour restreindre le nombre de gènes potentiellement impliqués dans la variation du caractère. Pour cela, une étude d'association est en cours à partir d'une population naturelle hybride Criollo/Amelonado localisée à Madagascar. Les fèves correspondant à 600 individus ont été phénotypés et de nouveaux marqueurs moléculaires ont été définis dans les 3 intervalles de confiances afin de réaliser une cartographie fine des régions génomiques identifiées.

L'expression des gènes alors identifiés dans les nouveaux intervalles de confiance devra être quantifiée dans les premiers stades de développement de la fève pour des génotypes purs Criollo et Amelonado. Des essais de transformation transitoire du tissu des cotylédons en culture pourraient également permettre de mettre en évidence l'effet de certains variants alléliques sur la voie de biosynthèse des anthocyanines. L'ensemble de ces analyses complètera l'article scientifique présenté dans ce chapitre.

CHAPITRE 5 - DISCUSSION GÉNÉRALE ET PERSPECTIVES

1. rappel des principaux résultats

En amélioration des plantes en général et chez le cacaoyer en particulier, l'évolution des technologies de séquençage a permis d'ouvrir de nombreuses perspectives pour étudier les mécanismes moléculaires impliqués dans les caractères agronomiques d'intérêt.

Ce travail de thèse a débuté par le séquençage complet du transcriptome du cacaoyer. Celui-ci a été réalisé à partir de 56 banques construites à partir de différents organes, différentes conditions environnementales et plusieurs génotypes et a abouti à l'identification et à la caractérisation de plus de 48 000 transcrits. L'analyse de cette collection a révélé une bonne représentativité générale des séquences de gènes exprimés chez le cacaoyer et a permis de caractériser des familles de gènes impliqués dans les mécanismes de défense et dans des voies métaboliques importantes pour la qualité du cacao. De nombreux marqueurs moléculaires ont également été développés à partir de cette étude et ont conduit à l'élaboration de cartes génétiques haute densité (Allegre et al., 2012; Fouet et al., 2011) qui ont notamment été utilisées pour ancrer les scaffolds du génome du cacaoyer sur les chromosomes.

Le séquençage du génome du cacaoyer et son analyse ont quant à eux été initiés en 2009 et ont été publiés sur le site internet de la revue Nature Genetics le 26 décembre 2010. Ce projet s'est appuyé sur un large partenariat rassemblant 23 partenaires internationaux et nationaux. L'assemblage du génome de ce génotype Criollo réalisée par la technique de fragmentation génomique WGS a couvert 97,8% de l'espace génique et a permis de caractériser 28 798 gènes codant pour des protéines et 67 575 éléments transposables. Des familles de gènes impliquées dans les mécanismes de défense et dans les caractères de qualité associés au cacao ont été étudiées et comparées aux études génétiques conduites par le passé. Enfin l'étude de la structure du génome a révélé une histoire évolutive mettant en évidence une relation évolutive étroite entre le génome du cacaoyer et le génome de l'ancêtre putatif commun aux dicotylédones.

Bien que la séquence du génome ait fourni une source majeure de gènes candidats pour les études génétiques du cacaoyer et son amélioration, la méthodologie et les technologies disponibles en 2010 ont conduit à un assemblage relativement fragmenté (4792 scaffolds pour un N50 de 473,8 kb). Par ailleurs, seul 66,8% de cet assemblage et 82% des gènes ont été ancrés sur les 10 chromosomes du cacaoyer. En conséquence, la conduite d'études génétiques pangénomiques ou la résolution de QTL par approche gènes candidats peuvent souffrir de ce morcellement du génome. L'évolution des technologies et méthodologies en 2015 a permis de corriger une partie de ces problèmes. Une nouvelle version d'assemblage a ainsi été produite en réassemblant les contigs de la première version du génome à l'aide de l'information positionnelle apportée par des lectures pairées issues de fragments de grande taille. Combinée à des lectures de grandes tailles apportées par les séquenceurs de 3ème génération, cette nouvelle version d'assemblage ne comprend plus que 554 scaffolds. La taille de ces séquences contigües a par ailleurs été multipliée par 14 (N50 de 6,5 Mb), si bien que certains bras de chromosomes ne sont plus couverts que par 1, 2 ou 3 séquences. Enfin, la quasi-totalité de l'assemblage (95,7%) et des gènes (98,5%) est désormais ancrée sur les 10 chromosomes du cacaoyer.

Cette nouvelle version de l'assemblage a été utilisée comme support de référence pour génotyper par séquençage une descendance issue d'un croisement entre deux clones hétérozygotes Trinitario (Criollo x Amelonado) phénotypées pour la couleur des fèves avant semis. Une étude QTL a permis d'identifier 3 régions génomiques localisées sur 3 chromosomes différents potentiellement impliquées dans la voie de biosynthèse des anthocyanines. Un des QTLs, localisé sur le chromosome 4, présente un fort LOD score, explique une grande partie de la variance observée (41,4%) et est caractérisé par un intervalle de confiance réduit (1,4 cM). Cette région correspond dans l'assemblage V2 à une région génomique de 3 Mb. Proche du marqueur localisé au pic du QTL, un gène codant pour une chalcone synthase, enzyme clé conduisant aux précurseurs de la voie de biosynthèse des anthocyanines, semble être un bon candidat pour expliquer une partie de la variation phénotypique observée. L'étude comparative de la structure du gène dans la variété Criollo (à fève blanche) et la variété Amelonado (à fève violette) révèle 2 différences majeures localisées dans la séquence codante et la région promotrice du gène.

Des études complémentaires sont initiées pour préciser les mécanismes moléculaires impliqués dans la variation de la couleur des fèves.

2. Le transcriptome du cacaoyer, une ressource clé pour la compréhension des mécanismes moléculaires impliqués dans les caractères agronomiques d'intérêt

Cette première ressource de séquences nucléotidiques à grande échelle chez le cacaoyer a permis d'accéder simultanément à des milliers de gènes et à leur expression. Elle a fournit une première indication sur la fonction potentielle des gènes du cacaoyer et leur implication dans des processus biologiques connus. L'annotation fonctionnelle et sa consultation via la mise en place d'un système automatique d'analyse et de recherche (http://esttik.cirad.fr) a facilité l'accès aux données aux chercheurs du domaine. Plusieurs publications scientifiques se sont appuyées sur cette collection pour caractériser finement des gènes impliqués dans des caractères agronomiques variés. Par exemple, en utilisant les ressources transcriptomiques du projet, Sabau et al., ont caractérisés dès 2006, l'expression de la linalool synthase (enzyme clé intervenant dans la voie de biosynthèse d'un composé volatile aromatique floral) dans des fèves de type Nacional et Trinitario. Par ailleurs, Shi et al. en 2010 ont caractérisé un gène fonctionnel du transcriptome du cacaoyer orthologue au gène NPR1 d'*Arabidopsis thaliana* jouant un rôle majeur dans les mécanismes de défense en réponse aux pathogènes. Dans une autre étude scientifique, Liu et al. en 2013 ont confirmé in vivo l'effet de 3 enzymes clés identifiées dans les ressources du transcriptome du cacaoyer dans la voie de biosynthèse des Proanthocyanidines. Enfin, Legavre et al. en 2015 ont identifié et caractérisé grâce aux ESTs issues du projet du transcriptome, des gènes de résistance différentiellement exprimés durant l'infection du cacaoyer par le pathogène *Phytophthora megakarya.*

Cette ressource transcriptomique a également été exploitée pour l'assemblage et l'annotation du génome Criollo du cacaoyer. Les marqueurs moléculaires définis dans cette ressource de données ont été utilisés pour construire la carte génétique nécessaire à l'ancrage de l'assemblage sur les chromosomes (Allegre et al., 2012) et l'ensemble des ESTs produites dans le cadre de ce projet ont été utilisées pour l'entrainement du modèle statistique du prédicteur de gène EuGene (Foissac et al., 2008) utilisé pour l'annotation structurale des gènes codant pour des protéines.

Enfin, la méthodologie et les outils développés dans le cadre de ce projet ont été utilisés pour caractériser les transcriptomes du Citrus (Luro et al., 2008; Terol et al., 2007), du Bananier (Hippolyte et al., 2010), de l'Hévéa (Duan et al., 2013; Putranto et al., 2012), du Palmier à huile (Tranbarger et al., 2011) et de *Gmelina arborea* (Rosero et al., 2011).

3. Le génome du cacaoyer, de nouvelles perspectives pour l'amélioration génétique

Comme chez beaucoup de plantes cultivées, la séquence du génome du cacaoyer a fourni une ressource sans précédent qui peut être exploité par des multiples approches intégrées. Elle a notamment le rôle de référence nécessaire pour

- Etudier l'expression des gènes, via les méthodes RNAseq par exemple.
- Reséquencer des régions du génome ou des génotypes complets et développer des marqueurs moléculaires.
- Conduire des analyses de génomique comparative.
- Conduire des études d'association pangénomique.

Ces approches combinées basées sur le génome du Criollo sont actuellement utilisées par plusieurs équipes de recherche pour résoudre des QTLs ou des régions génomiques impliquées dans des caractères agronomiques d'intérêt (résistances aux maladies, qualité, production, etc...).

L'identification de ces régions et des gènes candidats impliqués confèreront une meilleure compréhension des mécanismes moléculaires impliqués dans les caractères étudiés et permettront de développer des marqueurs diagnostiques utilisables dans les programmes de sélection.

Enfin, grâce au marquage moléculaire dense du génome du Criollo, nous conduisons actuellement des programmes de sélection génomique sur plusieurs types de populations de cacaoyers pour prédire la valeur génétique de candidats à la sélection.

RÉFÉRENCES BIBLIOGRAPHIQUES

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nat. Methods* 8, 61–65. doi:10.1038/nmeth.1527.

Allegre, M., Argout, X., Boccara, M., Fouet, O., Roguet, Y., Bérard, A., et al. (2012). Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of Theobroma cacao L. *DNA Res.* 19, 23–35. doi:10.1093/dnares/dsr039.

Allen, J., and Lass, R. (1983). London cocoa trade Amazon project: final report, phase 1. *Cocoa Grow. Bull* 34, 1–71.

Alverson, W. S., Whitlock, B. A., Nyffeler, R., Bayer, C., and Baum, D. A. (1999). Phylogeny of the core Malvales: evidence from ndhF sequence data. *Am. J. Bot.* 86, 1474–1486.

Alvim, P. de T. (1965). Ecophysiology of the cacao tree. in (Abidjan, Côte d'Ivoire).

Appiah, A. A., Flood, J., Bridge, P. D., and Archer, S. A. (2003). Inter- and intraspecific morphometric variation and characterization of Phytophthora isolates from cocoa. *Plant Pathol.* 52, 168–180. doi:10.1046/j.1365-3059.2003.00820.x.

Araújo, I. S., de Souza Filho, G. A., Pereira, M. G., Faleiro, F. G., de Queiroz, V. T., Guimarães, C. T., et al. (2009). Mapping of Quantitative Trait Loci for Butter Content and Hardness in Cocoa Beans (Theobroma cacao L.). *Plant Mol Bio Rep* 27, 177–183.

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of Theobroma cacao. *Nat. Genet.* 43, 101–108.

Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., et al. (2005). Cytokinin oxidase regulates rice grain production. *Science* 309, 741–745. doi:10.1126/science.1113373.

Bailey, B. A., and Meinhardt, L. W. eds. (2016). *Cacao Diseases*. Cham: Springer International Publishing Available at: http://link.springer.com/10.1007/978-3-319-24789-2 [Accessed September 20, 2016].

Barel, M. (2013). *Qualité du cacao: l'impact du traitement post-récolte*. Editions Quae.

Bastide, P. (1987). Evolution et metabolisme des composes phenoliques des feves de cacao Durant leur developpement au cours de la croissance et de la maturation du fruit de Theobroma cacao L.

Bonvehí, J. S. (2005). Investigation of aromatic compounds in roasted cocoa powder. *Eur. Food Res. Technol.* 221, 19–29. doi:10.1007/s00217-005-1147-y.

Boyer, J. (1970). Influence des régimes hydrique, radiatif et thermique du climat sur lactivité végétative et la floraison de cacaoyers cultivés au Cameroun. *Café Cacao Thé* 14, 189–201.

Brown, J. S., Motamayor, J. C., Lopes, U., Kuhn, D., and Borrone, J. W. (2005). Resistance Gene Mapping for Witches' Broom Disease in Theobroma cacao L. in an F2 Population using SSR Markers and Candidate Genes. *J. Am. Soc. Hortic. Sci.* 130, 366–373.

Brown, J. S., Phillips-Mora, W., Power, E. J., Krol, C., Cervantes-Martinez, C., Motamayor, J. C., et al. (2007). Mapping QTLs for Resistance to Frosty Pod and Black Pod Diseases and Horticultural Traits in L. *Crop Sci.* 47, 1851. doi:10.2135/cropsci2006.11.0753.

Cakirer, M. S., Ziegler, G. R., and Guiltinan, M. J. (2010). Seed color as an indicator of flavanol content in Theobroma cacao L. *Choc. Fast Foods Sweeten. Consum. Health Nova Sci. Pub Inc N. Y.*, 257–270.

Cakirer, M., Ziegler, G. R., Guiltinan, M. J., and Jones, A. D. (2003). Fresh bean colour as an indicator of chocolate flavour potential. *Flavour Res. Dawn Twenty First Century Proc. 10th Weurman Flavour Res. Symp. Beaune Fr. 25-28 June*. Available at: https://eurekamag.com/research/003/777/003777644.php [Accessed October 17, 2016].

Campos, V., and Villain, L. (2005). "Nematode parasites of coffee and cocoa.," in *lant parasitic nematodes in subtropical and tropical agriculture.*, 529–579.

Cervantes-Martinez, C., Brown, J. S., Schnell, R. J., Phillips-Mora, W., Takrama, J. F., and Motamayor, J. C. (2006). Combining Ability for Disease Resistance, Yield, and Horticultural Traits of Cacao (Theobroma cacao L.) Clones. *J. Am. Soc. Hortic. Sci.* 131, 231–241.

Chain, P. S. G., Grafham, D. V., Fulton, R. S., FitzGerald, M. G., Hostetler, J., Muzny, D., et al. (2009). Genome Project Standards in a New Era of Sequencing. *Science* 326, 236–237. doi:10.1126/science.1180614.

Chaiseri, S., and Dimick, P. S. (1989). Lipid and hardness characteristics of cocoa butters from different geographic regions. *J. Am. Oil Chem. Soc.* 66, 1771–1776. doi:10.1007/BF02660745.

Cheesman, E. E. (1944). Notes on the nomenclature, classification possible and relationships of cocoa populations. *Trop. Agric.* 21, 144–159.

Clapperton, J. (1994). A review of research to identify the origins of cocoa flavour characteristics. *Cocoa Grow. Bull* 48, 7–16.

Clement, D., Risterucci, A. M., Motamayor, J. C., N'Goran, J., and Lanaud, C. (2003). Mapping QTL for yield components, vigor, and resistance to Phytophthora palmivora in Theobroma cacao L. *Genome* 46, 204–212. doi:10.1139/g02-125.

Cope, F. W. (1958). Incompatibility in Theobroma cacao. *Nature* 181, 279. doi:10.1038/181279a0.

Crouzillat, D., Lerceteau, E., Petiard, V., Morera, J., Rodriguez, H., Walker, D., et al. (1996). Theobroma cacao L: A genetic linkage map and quantitative trait loci analysis. *Theor. Appl. Genet.* 93, 205–214.

Cuatrecasas, J. (1964). Cacao and its allies: a taxonomic revision of the genus Theobroma. *Bull. U. S. Natl. Mus. Smithson. Inst.* 35, 379–614.

da Silva, M. R., Clément, D., Gramacho, K. P., Monteiro, W. R., Argout, X., Lanaud, C., et al. (2016). Genome-wide association mapping of sexual incompatibility genes in cacao (Theobroma cacao L.). *Tree Genet. Genomes* 12. doi:10.1007/s11295-016-1012-0.

Denkyirah, E. K., Okoffo, E. D., Adu, D. T., Aziz, A. A., Ofori, A., and Denkyirah, E. K. (2016). Modeling Ghanaian cocoa farmers' decision to use pesticide and frequency of application: the case of Brong Ahafo Region. *SpringerPlus* 5, 1113. doi:10.1186/s40064-016-2779-z.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., et al. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* 31, 135–141. doi:10.1038/nbt.2478.

Duan, C., Argout, X., Gébelin, V., Summo, M., Dufayard, J.-F., Leclercq, J., et al. (2013). Identification of the Hevea brasiliensisAP2/ERF superfamily by RNA sequencing. *BMC Genomics* 14, 30. doi:10.1186/1471-2164-14-30.

Duthie, D. (1938). Observations on the biochemistry of the cacao kernel. Tannin and cateehin. *Seventh Annu. Rep. Cacao Res. 1937*, 47–51.

End, M. J., Daymond, A. J., and Hadley, P. (2014). Technical guidelines for the safe movement of cacao germplasm. Available at: http://www.bioversityinternational.org/e-library/publications/detail/technical-guidelines-for-the-safe-movement-of-cacao-germplasm/ [Accessed September 20, 2016].

Fademi, O., Orisajo, S., and Afolami, S. (2006). Impact of plant parasitic nematodes on cocoa production (in Nigeria) and outlook for future containment of the problem. in *Proceedings 15th International Cocoa Research Conference, San Jose, Costa Rica*.

Faleiro, F. G., Queiroz, V. T., Lopes, U. V., Guimarães, C. T., Pires, J. L., Yamada, M. M., et al. (2006). Mapping QTLs for Witches' Broom (Crinipellis Perniciosa) Resistance in Cacao (Theobroma Cacao L.). *Euphytica* 149, 227–235. doi:10.1007/s10681-005-9070-7.

Falque, M. (1994). Pod and seed development and phenotype of the M1 plants after pollination and fertilization with irradiated pollen in cacao (Theobroma cacao L.). *Euphytica* 75, 19–25. doi:10.1007/BF00024527.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512. doi:10.1126/science.7542800.

Feitosa Jucá Santos, F., Vanderlei Lopes, U., Pires, J. L., Melo, G. R. P., Peres Gramacho, K., and Clément, D. (2014). QTLs Detection under natural infection of Moniliophtora perniciosa in a cacao F2 progeny with Scavina-6 descendants. *AgroTropica* 26, 65–72.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., et al. (2008). Genome annotation in plants and fungi: EuGène as a model platform. *Curr. Bioinforma.* 3, 87–97.

Fouet, O., Allegre, M., Argout, X., Jeanneau, M., Lemainque, A., Pavek, S., et al. (2011). Structural characterization and mapping of functional EST-SSR markers in Theobroma cacao. *Tree Genet. Genomes* 7, 799–817. doi:10.1007/s11295-011-0375-5.

Gardella, D., Enriquez, G., Saunders, J., and others (1982). Inheritance of clonal resistance to Ceratocystis fimbriata in cacao hybrids. in *Proceedings 8th International Cocoa Research Conference, Cartagena, Colombia, 18 23 Oct 1981.* (Cocoa Producers' Alliance), 695–702.

Gesteira, A. S., Micheli, F., Carels, N., Da Silva, A. C., Gramacho, K. P., Schuster, I., et al. (2007). Comparative analysis of expressed genes from cacao meristems infected by Moniliophthora perniciosa. *Ann Bot Lond* 100, 129–40.

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., and Miller, C. T. (2015). Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Sticklebacks with Genotyping-by-Sequencing. *G3 GenesGenomesGenetics* 5, 1463–1472. doi:10.1534/g3.115.017905.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 Genes. *Science* 274, 546–567. doi:10.1126/science.274.5287.546.

Goulet, C., Mageroy, M. H., Lam, N. B., Floystad, A., Tieman, D. M., and Klee, H. J. (2012). Role of an esterase in flavor volatile variation within the tomato clade. *Proc. Natl. Acad. Sci.* 109, 19009–19014. doi:10.1073/pnas.1216515109.

Gutiérrez, O. A., Campbell, A. S., and Phillips-Mora, W. (2016). "Breeding for disease resistance in cacao," in *Cacao Diseases* (Springer), 567–609.

Heim, M. A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B., and Bailey, P. C. (2003). The Basic Helix–Loop–Helix Transcription Factor Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity. *Mol. Biol. Evol.* 20, 735–747. doi:10.1093/molbev/msg088.

Hippolyte, I., Bakry, F., Seguin, M., Gardes, L., Rivallan, R., Risterucci, A.-M., et al. (2010). A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. *BMC Plant Biol.* 10, 65. doi:10.1186/1471-2229-10-65.

Höfte, H., Desprez, T., Amselem, J., Chiapello, H., Caboche, M., Moisan, A., et al. (1993). An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana†. *Plant J.* 4, 1051–1061. doi:10.1046/j.1365-313X.1993.04061051.x.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39. doi:10.1038/ng.1018.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* 24, 688–696. doi:10.1101/gr.168450.113.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi:10.1038/nature03895.

Iwaro, A. D., Thévenin, J.-M., Butler, D. R., and Eskes, A. B. (2005). Usefulness of the Detached Pod Test for Assessment of Cacao Resistance to Phytophthora Pod Rot. *Eur. J. Plant Pathol.* 113, 173–182. doi:10.1007/s10658-005-2929-6.

Jaakola, L. (2013). New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant Sci.* 18, 477–483. doi:10.1016/j.tplants.2013.06.003.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi:10.1038/nature06148.

Jones, P. G., Allaway, D., Gilmour, D. M., Harris, C., Rankin, D., Retzel, E. R., et al. (2002). Gene discovery and microarray analysis of cacao (Theobroma cacao L.) varieties. *Planta* 216, 255–64.

Keith, C. S., Hoang, D. O., Barrett, B. M., Feigelman, B., Nelson, M. C., Thai, H., et al. (1993). Partial Sequence Analysis of 130 Randomly Selected Maize cDNA Clones. *Plant Physiol.* 101, 329–332. doi:10.1104/pp.101.1.329.

Kim, H., and Keeney, P. (1984). (-)-Epicatechin Content in Fermented and Unfermented Cocoa Beans. *J. Food Sci.* 49, 1090–1092.

Kirchhoff, P.-M., Biehl, B., and Crone, G. (1989). Peculiarity of the accumulation of free amino acids during cocoa fermentation. *Food Chem.* 31, 295–311.

Kloosterman, B., Abelenda, J. A., Gomez, M. del M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* 495, 246–250. doi:10.1038/nature11912.

Knight, R., and Rogers, H. (1955). Incompatibility in Theobroma cacao. *Heredity* 9, 69–77.

Kokutse, F. (2008). Swollen Shoot Disease Devastating Cocoa Trees. *Inter Press Serv. News Agency*. Available at: http://www.ipsnews.net/2008/11/trade-west-africa-swollen-shoot-disease-devastating-cocoa-trees/ [Accessed September 20, 2016].

Lachenaud, P. (1995). Variations in the number of beans per pod in Theobroma cacao L. in the Ivory Coast. III. Nutritional factors, cropping effects and the role of boron. *J. Hortic. Sci.* 70, 7–13. doi:10.1080/14620316.1995.11515267.

Lanaud, C. (1986). Utilisation des marqueurs enzymatiques pour l'étude génétique du cacaoyer : *Theobroma cacao* L. I. Contrôle génétique et "linkage" de neuf marqueurs génétiques. *Café Cacao Thé* 30, 259–270.

Lanaud, C., Boult, E., Clapperton, J., N'Goran, J. K. A., Cros, E., Chapelin, M., et al. (2003). Identification of QTLs related to fat content, seed size and sensorial traits in Theobroma cacao L. in *Proceedings of the 14th International Cocoa Research Conference, October*, 13–18.

Lanaud, C., Fouet, O., Clément, D., Boccara, M., Risterucci, A. M., Surujdeo-Maharaj, S., et al. (2009). A meta–QTL analysis of disease resistance traits of Theobroma cacao L. *Mol. Breed.* 24, 361–374.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756.

Lange, H., and Fincke, A. (1970). "Kakao und schokolade," in *Alkaloidhaltige Genussmittel, Gewürze, Kochsalz* (Springer), 210–309.

Laurent, V. (1993). Etude de la diversite genetique du cacaoyer (Theobroma cacao L.) basee sur le polymorphisme de la longueur des fragments de restriction (RFLP).

Leal, G. A. L. J., Albuquerque, P. S. B., and Figueira, A. (2007). Genes differentially expressed in Theobroma cacao associated with resistance to witches' broom disease caused by Crinipellis perniciosa. *Mol. Plant Pathol.* 8, 279–292.

Legavre, T., Ducamp, M., Sabau, X., Argout, X., Fouet, O., Dedieu, F., et al. (2015). Identification of Theobroma cacao genes differentially expressed during Phytophthora megakarya infection. *Physiol. Mol. Plant Pathol.* 92, 1–13. doi:10.1016/j.pmpp.2015.08.005.

Lepiniec, L., Debeaujon, I., Routaboul, J.-M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006). Genetics and biochemistry of seed flavonoids. *Annu. Rev. Plant Biol.* 57, 405–430. doi:10.1146/annurev.arplant.57.032905.105252.

Leppik, E. E. (1970). Gene centers of plants as sources of disease resistance. *Annu. Rev. Phytopathol.* 8, 323–344.

Li, S. (2014). Transcriptional control of flavonoid biosynthesis: Fine-tuning of the MYB-bHLH-WD40 (MBW) complex. *Plant Signal. Behav.* 9, e27522. doi:10.4161/psb.27522.

Liu, Y., Shi, Z., Maximova, S., Payne, M. J., and Guiltinan, M. J. (2013). Proanthocyanidin synthesis in Theobroma cacao: genes encoding anthocyanidin synthase, anthocyanidin reductase, and leucoanthocyanidin reductase. *BMC Plant Biol.* 13, 202. doi:10.1186/1471-2229-13-202.

Lopes, U. V., Monteiro, W. R., Pires, J. L., Clement, D., Yamada, M. M., and Gramacho, K. P. (2011). Cacao breeding in Bahia, Brazil - strategies and results. *Crop Breed. Appl. Biotechnol.* 1, 73–81.

Luro, F. L., Costantino, G., Terol, J., Argout, X., Allario, T., Wincker, P., et al. (2008). Transferability of the EST-SSRs developed on Nules clementine (Citrus clementina Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. *BMC Genomics* 9, 287.

Marcano, M., Morales, S., Hoyer, M. T., Courtois, B., Risterucci, A. M., Fouet, O., et al. (2009). A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (Theobroma cacao L.) population. *Tree Genet. Genomes* 5, 329–337.

Marelli, J.-P., Fernandez-Silva, I., Correa, F. M., Royaert, S., and Schnell, R. J. (2014). QTL mapping of Ceratocystis wilt resistance in Theobroma cacao. in *Plant and Animal Genome XXII Conference* (Plant and Animal Genome).

Martin, G., Baurens, F.-C., Droc, G., Rouard, M., Cenci, A., Kilian, A., et al. (2016). Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* 17, 243. doi:10.1186/s12864-016-2579-4.

Meinhardt, L. W., Rincones, J., Bailey, B. A., Aime, M. C., Griffith, G. W., Zhang, D., et al. (2008). Moniliophthora perniciosa, the causal agent of witches' broom disease of cacao: what's new from this old foe? *Mol. Plant Pathol.* 9, 577–588. doi:10.1111/j.1364-3703.2008.00496.x.

Mossu, G. P., and Reffye, D. (1981). Influence de la floraison et de la pollinisation sur les rendements du cacaoyer. *Café Cacao Thé V 25 3 P 155-168*.

Motamayor, J. C., Lachenaud, P., Mota, J. W. da S. e, Loor, R., Kuhn, D. N., Brown, J. S., et al. (2008). Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L). *PLOS ONE* 3, e3311. doi:10.1371/journal.pone.0003311.

Motamayor, J. C., Risterucci, A. M., Lopez, P. A., Ortiz, C. F., Moreno, A., and Lanaud, C. (2002). Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* 89, 380–386. doi:10.1038/sj.hdy.6800156.

N'Goran, J., Laurent, V., Risterucci, A., and Lanaud, C. (1994). Comparative genetic diversity studies of Theobroma cacao L. using RFLP and RAPD markers. *Heredity* 73, 589–597.

Paulin, D., Decazy, B., and Coulibaly, N. (1983). Etude des variations saisonnières des conditions de pollinisation et de fructification dans une cacaoyère. *Café Cacao Thé V 27 3 P 165-176*.

Paulin, D., Ducamp, M., and Lachenaud, P. (2008). New sources of resistance to Phytophthora megakarya identified in wild cocoa tree populations of French Guiana. *Crop Prot.* 27, 1143–1147. doi:10.1016/j.cropro.2008.01.004.

Phillips-Mora, W. (2003). Origin, biogeography, genetic diversity and taxonomic affinities of the cacao (Theobroma cacao L.) fungus Moniliophthora roreri (Cif.) Evans et al. as determined using molecular, phytopathological and morpho-physiological evidence.

Phillips-Mora, W., and Castillo, J. (1999). Artificial inoculations in cacao with the fungi Moniliophthora roreri (Cif. Par) Evans et al. and Phytophthora palmivora (Butl.) Butler. *CATIE Ed Actas IV Sem. CientíficaTurrialba Logros Investig. Para Un Nuevo Milen. Programa Investig. Turrialba Costa Rica CATIE*.

Pin, P. A., Benlloch, R., Bonnet, D., Wremerth-Weich, E., Kraft, T., Gielen, J. J. L., et al. (2010). An Antagonistic Pair of FT Homologs Mediates the Control of Flowering Time in Sugar Beet. *Science* 330, 1397–1400. doi:10.1126/science.1197004.

Pires, J. L., Marita, J. M., Lopes, U. V., Yamada, M. M., Atiken, W., Melo, G., et al. (2000). Diversity for phenotypic traits and molecular markers in CEPEC's germplasm collection in Bahia, Brazil. in *Proceedings of the International Workshop on New Technologies and Cocoa Breeding*, 16–17.

Ploetz, R. C. (2007). Cacao diseases: important threats to chocolate production worldwide. *Phytopathology* 97, 1634–1639.

Pound, F. J. (1938). Cacao and witches' broom disease (Marasmius perniciosus) of South America, with notes on other species of Theobroma. *Yuilles' Printery Port Spain Trinidad Tobago* 1, 21–64.

Pugh, T., Fouet, O., Risterucci, A. M., Brottier, P., Abouladze, M., Deletrez, C., et al. (2004). A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* 108, 1151–1161.

Putranto, R.-A., Sanier, C., Leclercq, J., Duan, C., Rio, M., Jourdan, C., et al. (2012). Differential gene expression in different types of Hevea brasiliensis roots. *Plant Sci.* 183, 149–158. doi:10.1016/j.plantsci.2011.08.005.

Queiroz, V. T., Guimarães, C. T., Anhert, D., Schuster, I., Daher, R. T., Pereira, M. G., et al. (2003). Identification of a major QTL in cocoa (Theobroma cacao L.) associated with resistance to witches' broom disease. *Plant Breed.* 122, 268–272. doi:10.1046/j.1439-0523.2003.00809.x.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44, 217–220. doi:10.1038/ng.1033.

Rimbach, G., Melchin, M., Moehring, J., and Wagner, A. E. (2009). Polyphenols from cocoa and vascular health-a critical review. *Int. J. Mol. Sci.* 10, 4290–4309. doi:10.3390/ijms10104290.

Risterucci, A. M., Grivet, L., N'Goran, J. A. K., Pieretti, I., Flament, M. H., and Lanaud, C. (2000). A high-density linkage map of Theobroma cacao L. *Theor. Appl. Genet.* 101, 948–955.

Rosero, C., Argout, X., Ruiz, M., and Teran, W. (2011). A drought stress transcriptome profiling as the first genomic resource for white teak - Gamhar - (Gmelina arborea Roxb) and related species. *BMC Proc.* 5, P178. doi:10.1186/1753-6561-5-S7-P178.

Sabau, X., Loor, G., Boccara, M., Fouet, O., Jeanneau, M., Argout, X., et al. (2006). Preliminary results on linalool synthase expression during seed development and fermentation of nacional and trinitario clones. in (Conference center, San Jose, Costa Rica).

Santos, R. M. F., Lopes, U. V., Silva, S. D. V. M., Micheli, F., Clement, D., and Gramacho, K. P. (2012). Identification of quantitative trait loci linked to Ceratocystis wilt resistance in cacao. *Mol. Breed.* 30, 1563–1571. doi:10.1007/s11032-012-9739-2.

Schnermann, P., and Schieberle, P. (1997). Evaluation of key odorants in milk chocolate and cocoa mass by aroma extract dilution analyses. *J. Agric. Food Chem.* 45, 867–872.

Schwan, R. F., and Wheals, A. E. (2004). The microbiology of cocoa fermentation and its role in chocolate quality. *Crit. Rev. Food Sci. Nutr.* 44, 205–221.

Shi, Z., Maximova, S. N., Liu, Y., Verica, J., and Guiltinan, M. J. (2010). Functional analysis of the theobroma cacao NPR1 gene in arabidopsis. *BMC Plant Biol.* 10, 248. doi:10.1186/1471-2229-10-248.

Solorzano, R. G. L., Fouet, O., Lemainque, A., Pavek, S., Boccara, M., Argout, X., et al. (2012). Insight into the Wild Origin, Migration and Domestication History of the Fine Flavour Nacional *Theobroma cacao* L. Variety from Ecuador. *PLOS ONE* 7, e48438. doi:10.1371/journal.pone.0048438.

Soria, V., and J Salazar, L. (1965). Pruebas preliminares de resistencia a Ceratocystis fimbriata en clones e híbridos de cacao. *Turrialba IICA V 15 4 P 290-295*.

Spencer, J. P. . (2009). Flavonoids and brain health: multiple effects underpinned by common mechanisms. *Genes Nutr* 4, 243–250.

Stuber, C. W. (North C. S. U. (1989). Marker-based selection for quantitative traits. in *Vortraege fuer Pflanzenzuechtung (Germany)* Available at: http://agris.fao.org/agris-search/search.do?recordID=DE94B1408 [Accessed October 21, 2016].

Terol, J., Conesa, A., Colmenero, J. M., Cercos, M., Tadeo, F., Agusti, J., et al. (2007). Analysis of 13000 unique Citrus clusters associated with fruit quality, production and salinity tolerance. *BMC Genomics* 8, 31.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–815. doi:10.1038/35048692.

The C. elegans Sequencing Consortium (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* 282, 2012–2018. doi:10.1126/science.282.5396.2012.

Thevenin, J.-M., Rossi, V., Ducamp, M., Doare, F., Condina, V., and Lachenaud, P. (2012). Numerous Clones Resistant to Phytophthora palmivora in the “Guiana” Genetic Group of Theobroma cacao L. *PLOS ONE* 7, e40915. doi:10.1371/journal.pone.0040915.

Tranbarger, T. J., Dussert, S., Joët, T., Argout, X., Summo, M., Champion, A., et al. (2011). Regulatory Mechanisms Underlying Oil Palm Fruit Mesocarp Maturation, Ripening, and Functional Specialization in Lipid and Carotenoid Metabolism. *Plant Physiol.* 156, 564–584. doi:10.1104/pp.111.175141.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313, 1596–604.

Uchimiya, H., Kidou, S., Shimazaki, T., Aotsuka, S., Takamatsu, S., Nishi, R., et al. (1992). Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (Oryza sativa L.). *Plant J.* 2, 1005–1009. doi:10.1111/j.1365-313X.1992.01005.x.

Udall, J. A., Swanson, J. M., Haller, K., Rapp, R. A., Sparks, M. E., Hatfield, J., et al. (2006). A global assembly of cotton ESTs. *Genome Res* 16, 441–50.

Van Hall, C. J. J. (1914). *MacMillan*. London, United Kingdom.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. *PLOS ONE* 2, e1326. doi:10.1371/journal.pone.0001326.

Verica, J. A., Maximova, S. N., Strem, M. D., Carlson, J. E., Bailey, B. A., and Guiltinan, M. J. (2004). Isolation of ESTs from cacao (Theobroma cacao L.) leaves treated with inducers of the defense response. *Plant Cell Rep* 23, 404–13.

Wellensiek, S. (1931). The genetics of cotyledon colour of cocoa as a basis for quality selection. *Transl. H Toxopeus Arch. Voor Koffiecult. Ned.-Indië Buitenzorg Java*.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol* 126, 485–93.

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., et al. (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci.* 109, E2155–E2164. doi:10.1073/pnas.1117982109.

Ziegleder, G. (1990). Linalol contents as characteristics of some flavour grade cocoas. *Z Leb. Unters Forsch* 191, 306–309.

Zumbé, A. (1998). Polyphenols in cocoa: are there health benefits? *Nutr. Bull.* 23, 94–102.

Abstract

For several years, cocoa research programs have focused on studying the genetic basis of agronomic traits of interest, especially for disease resistance and quality of cocoa beans, two important attributes for cocoa farmers and chocolate production. This work presents, by exploring the transcriptome and cocoa genome, the constitution of molecular resources and the analysis of biosynthesis pathways involved in several of these agronomical traits.

The transcriptome study allowed the identification of several tens of thousands of genes expressed in various organs and for different environmental conditions and provided numerous molecular markers, used to produce high-density genetic maps. The information provided by this work led to the genome sequencing of a cocoa Criollo variety. Its analysis and annotation have provided a set of crucial biological information, from the catalog of genes and transposable elements to evolutionary aspects, which revealed a close evolutionary relationship to the eudicot putative ancestor, showing a limited number of recombination between ancestral chromosomes. Subsequently, the work we carried out to improve the complete sequence led to a considerable reduction of the chromosomal fragmentation observed in the first version. In addition, 97% of the assembled sequence and 99% of the genes are now anchored on the cocoa chromosomes.

To exploit this new sequence of the genome, we conducted a QTL study from a progeny between Trinitario clones established in French Guiana, allowing to identify genomic regions involved in color trait variation observed in cocoa beans. Based on the improved version of the Criollo genome, we identified two genes potentially involved in the biosynthesis pathway of anthocyanins and flavonoids in the main genomic region concerned. One of the two genes is located nearby the QTL peak and encoding a chalcone synthase, appears to be a promising candidate gene. The comparative study of its structure in the Criollo genome (white bean) and in the Amelonado genome (purple bean) revealed several variations that could be responsible for functional modifications.

The results presented in this thesis provide a variety of knowledge and tools useful to conduct multiple integrated approaches to studying cocoa tree genetics.

Keywords : *Theobroma cacao* (L.), transcriptome, genome, genetic map, assembly, candidate genes

Résumé

Depuis plusieurs années, les programmes de recherche chez le cacaoyer ont mis l'accent sur l'étude des bases génétiques des caractères agronomiques d'intérêt, notamment concernant la résistance aux maladies et la qualité des fèves de cacao, qui représentent deux attributs importants pour la cacaoculture et la production de chocolat. Ce travail présente, par l'exploration du transcriptome et du génome du cacaoyer, la constitution de ressources moléculaires et l'analyse des voies de biosynthèse impliquées dans plusieurs de ces caractères agronomiques d'intérêt.

L'étude du transcriptome a permis l'identification de plusieurs dizaines de milliers de gènes exprimés dans divers organes et pour différentes conditions environnementales et ont fourni de nombreux marqueurs moléculaires qui ont été utilisés pour réaliser des cartes génétiques haute densité. Les informations apportées par ce travail ont permis d'engager le séquençage du génome de la variété Criollo du cacaoyer. Son analyse et son annotation ont apporté un ensemble d'informations biologiques cruciales, depuis le catalogue des gènes et éléments mobiles jusqu'aux aspects évolutifs qui a révélé une structure du génome peu remanié par rapport à l'ancêtre commun aux dicotylédones. Par la suite, les travaux que nous avons menés pour améliorer la séquence complète ont conduit à une réduction considérable de la fragmentation chromosomique observée dans la première version. Par ailleurs 97% de la séquence assemblée et 99% des gènes sont désormais ancrés sur les chromosomes du cacaoyer.

Pour commencer à exploiter cette nouvelle séquence du génome, nous avons conduit une étude QTL à partir d'une descendance entre Trinitario implantée en Guyane, permettant de localiser les régions génomiques impliquées dans la variation de la couleur des fèves de cacao. En s'appuyant sur la version améliorée du génome du Criollo, nous avons identifié deux gènes potentiellement impliqués dans la voie de biosynthèse des anthocyanines et flavonoïdes dans la principale région génomique concernée. Un des deux gènes, situé proche du marqueur situé au pic du QTL et codant pour une chalcone synthase, semble être un gène candidat prometteur. L'étude comparative de sa structure dans le génome du Criollo (à fève blanche) et du génome de l'Amelonado (à fève violette) a mis en évidence des différences structurales pouvant être à l'origine d'une modification fonctionnelle.

L'ensemble des résultats présentés dans ce travail de thèse apporte une connaissance et des outils variés qui peuvent être exploités par de multiples approches intégrées pour étudier la génétique du cacaoyer.

Mots clés: *Theobroma cacao* (L.), transcriptome, génome, cartographie génétique, assemblage, gènes candidats