Contribution à l'étude des déterminants génétiques et biochimiques des arômes des cacaos fins d'Equateur : Le Nacional, de la diversité aromatique de sa population d'origine jusqu'à sa domestication récente

Kelly Colonges

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En génétique et amélioration des plantes

École doctorale GAIA

Unités de recherche AGAP institut et QUALISUD

Contribution à l'étude des déterminants génétiques et biochimiques des arômes de cacaos fins d'Equateur :

Le Nacional, de la diversité aromatique de sa population d'origine jusqu'à sa domestication récente.

Présentée par Kelly Colonges
Le 1 décembre 2021

Sous la direction de Claire Lanaud
et de Renaud Boulanger

Devant le jury composé de

Mathilde Causse, directrice de recherches, HDR, INRAE, Avignon	Rapportrice
Philippe Hugueney, directeur de recherches, HDR, INRAE, Colmar	Rapporteur
Rey Gaston Loor Solorzano, directeur de recherches, INIAP, Equateur	Examinateur
Chantal Menut, Professeure émérite, HDR, Université de Montpellier	Examinatrice
Claire Lanaud, chercheure émérite, HDR, CIRAD	Directrice de thèse
Renaud Boulanger, chercheur, CIRAD	Co-encadrant de thèse

UNIVERSITÉ DE MONTPELLIER

Remerciements

Ces trois années de thèse ont été riches en émotions, en rencontres, en péripéties…

Je tiens tout d'abord à remercier ma directrice de thèse, Claire Lanaud, qui m'a donné l'envie de faire cette thèse. C'est en grande partie grâce à ses encouragements lors de mon stage de Licence 3 que j'ai eu l'envie et la motivation d'aller jusqu'ici. Un grand merci d'avoir accepté d'être ma directrice de thèse, d'avoir cru en moi, d'être restée présente jusqu'au bout et d'avoir partagé ta passion sur le cacao.

Je tiens également à remercier mon co-encadrant de thèse, Renaud Boulanger. Merci de m'avoir accueillie au sein de ton équipe à Qualisud et d'avoir partagé tes connaissances sur la biochimie des arômes, en particulier du cacao. Ce n'était pas ma spécialité au début mais je suis très contente d'avoir pu travailler sur cette thématique avec toi. Une thématique qui m'a inspiré aussi pour mes futurs projets. Un grand merci d'avoir été présent durant tous les moments de cette thèse.

J'aimerais aussi remercier tous les membres du Jury de m'avoir fait l'honneur d'accepter de juger ce travail. Un grand merci à Mme Mathilde Causse (Directrice de recherche à l'INRAE d'Avignon) et Mr Philippe Hugueney (Directeur de recherche à l'INRAE de Colmar) pour leur participation en qualité de rapporteur et pour le temps qu'ils ont consacré à la lecture critique de ce manuscrit. Un grand merci également à Mme Chantal Menut (Professeur émérite de la faculté de chimie de Montpellier) et Mr Rey Gaston Loor Solorzano (Directeur de recherche cacao et café à l'INIAP en Equateur) pour leur participation en qualité d'examinateur et leur avis critique et scientifique sur ce travail. Une mention spéciale pour Mr Rey Gaston Loor Solorzano, merci de nous permettre de travailler avec vous et de nous donner accès aux collections de cacaoyers en Equateur, sans quoi ce travail n'aurait pas été possible.

Je tiens à remercier également les membres de la direction de l'UMR AGAP et ceux de l'UMR Qualisud pour leur accueil au sein de leur UMR. Un grand merci également à l'équipe Génétique et Sélection de Pérennes de l'UMR AGAP et l'équipe 1 de l'UMR Qualisud pour leur accueil chaleureux.

J'aimerais également remercier l'université d'excellence MUSE et l'université de Montpellier pour leurs financements qui m'ont permis de mener à bien cette thèse.

Un grand merci également aux membres de mon comité de thèse. Merci à vous, Mme Pascale Chalier, Mr Patrice This, Mr Benjamin Brachi et Mme Nancy Terrier pour vos précieux conseils. Merci également à Pierre Costet de Valrhona d'avoir assisté à mon deuxième comité et de m'avoir apporté son expertise et ses conseils.

J'aimerais également remercier l'Université de Montpellier et Mr Bruno Touraine pour m'avoir donné l'opportunité d'enseigner à la faculté de sciences lors de Missions Complémentaires d'Enseignements. Un grand merci également aux responsables des UE de biologie cellulaire, de biochimie et de physiologie végétale de m'avoir accordé leur confiance lors de mes MCE. J'aimerais remercier tout particulièrement Nelly Godefroy et Laila Gannoun pour leur transmission passionnée de l'enseignement et de leurs précieux conseils.

J’aimerais remercier aussi les partenaires du projet avec qui j’ai pu échanger. Ce fut des échanges très riches et très intéressants pour moi, un grand merci à tous :

- Les membres de Valrhona : Pierre Costet, Clothilde Hue, Florent Coste, merci pour les différentes discussions que l’on a pu avoir ensemble et les chocolats !
- Les collègues de l’INIAP : Gaston Rey Loor Solorzano, Juan-Carlos Jimenez Barrangan, Cristian Subia, Dario Calderon, Fabian Fernandez, Ignacio Sotomayor Cantos. Muchas gracias por sus ayudas con los microfermentaciones !
- Edward Seguine, many thanks for your work in sensory analysis and your precious help!
- Les collègues de l’UMR IBMM de l’Université de Montpellier : Chantal Menut et Alain Morère

Un énorme merci également aux collègues du Cirad qui m’ont encadré et enseigné toutes les combines du laboratoire pour que je puisse mener à bien et avec succès mes expériences au laboratoire. Merci Olivier Fouet pour avoir partagé avec moi ton expertise sur la génétique du cacao et toutes les techniques nécessaires pour l’étudier. Merci à Marie-Christine Lahon de m’avoir montré les secrets de la GC-MS. Et surtout merci à Marie-Christine Lahon, à Karine Allary et également à Olivier Fouet de m’avoir aidé pour le décorticage et le broyage des 218 échantillons de cacaos, vraiment merci ce fut long et fastidieux mais nous y sommes arrivés. Merci également à Fabienne Ribeyre pour ces conseils et son expertise en analyses statistiques. Un grand merci également à Juan-Carlos Jimenez Barrangan, à Alejendra Saltos pour avoir réalisé les dosages GC-MS et NIRS de la population de Nacional et à Jérôme Minier, à Fabrice Davrieux pour avoir réalisé les dosages NIRS de la population native d’Amazonie.

J’aimerais remercier également l’ensemble des membres du groupe de travail GQMS2. Merci pour les discussions que nous avons eu et merci David Pot de m’avoir permis de présenter à plusieurs reprises mes travaux de recherches.

Je tiens à remercier également tous les collègues du bâtiment 1, 3 et 16 avec qui j’ai partagé des discussions, des pauses cafés, des repas, des blagues…

Évidemment, j’aimerais remercier les doctorants d’AGAP avec qui j’ai partagé cette aventure. Merci à vous Marion, Aurélien, Benjamin, Clara, Léo, Ian, Céline, Nicolas, Charlotte, Laurianne, Lison, Stella, Abdoulaye…

Et non je ne t’ai pas oublié, bien sur un grand merci à toi aussi mon petit palmier Aurélie !! Merci d’avoir été là pendant ces trois années et de m’avoir supportée dans ton bureau (enfin dans notre bureau ;)) pendant la dernière année. On s’en souviendra de cette thèse…

J’aimerais aussi remercier ma famille et mes ami.e.s qui ont été un soutien sans faille durant ces années pas toujours faciles… Un merci spécial pour : ma maman, un grand merci d’avoir toujours été présente ; merci Max’ de m’avoir supportée durant les deux dernières années 24h/24h (et là pour le coup c’était vraiment 24h/24 merci le coco), merci à Marie D. et Franz alias les voisins du C pour les apéros qui remontent toujours le moral et merci à Elodie pour ses petits gâteaux et son coaching oral. Merci également à Marie B. pour la relecture de ce manuscrit.

Merci à ceux que j’aurais oublié, j’en suis désolée… **Encore un grand merci à vous tous !!!**

A bientôt je l’espère, pour de nouvelles aventures !

Contribution à l'étude des déterminants génétiques et biochimiques des arômes des cacaos fins d'Equateur :

Le Nacional, de la diversité aromatique de sa population d'origine jusqu'à sa domestication récente.

Avant-propos

Les travaux de thèse présentés dans ce manuscrit s'inscrivent dans le projet MUSE Amazcacao et le projet conjoint Aromcacao qui en ont assuré les frais de fonctionnement et font partie intégrante du module de travail numéro 3 intitulé « Etude des déterminants génétiques et biochimiques des composants d'arôme ». Ces travaux ont été réalisés au sein de l'UMR AGAP institut et de l'UMR Qualisud et en partenariat entre le CIRAD, l'INIAP (Equateur), Valrhona (France) et Seguine Cacao/Guittard Chocolate (États-Unis).

La thèse présentée ci-après a été financée par l'ANR (Agence National de Recherche française) sous le programme "Investissement d'avenir" avec la référence ANR-16-IDEX-0006. Elle est rattachée à l'école doctorale GAIA, filière BIDAP.

Sommaire

Liste des abréviations

ADN	Acide Désoxyribonucléique
AFLP	Amplified Fragment Lenght Polymorphism
	Polymorphisme de la longueur des fragments amplifiés
AGAP	Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales
AIC	Auto-Incompatible
AMP	Adénosine monophosphate
ANR	Agence National de Recherche française
Asp	Acide aspartique
AAT	Alcool AcétylTransférases
CA	Acide cinnamique
CCAT	Centro de Cacao de Aroma de Tenguel
	Centre des Cacaos Aromatiques de Tenguel
COV	Composés Organiques Volatils
CSSV	Cocoa Swollen Shoot Virus
EET-P	Estación Experimental Tropical de Pichilingue
	Station Expérimentale Tropical de Pichilingue
GBS	Genotyping By Sequencing
	Génotypage par séquencage
GCO	Gaz-Chromatography-Olfactometry
	Chromatographie gazeuse couplée à l'olfactométrie
Glu	Acide glutamique
GPP	Géranyl Pyrophosphate
GWAS	Genome Wide Association Study
	Etude d'Association sur le Génome Entier
HE	Huiles essentielles
His	Histidine
HPPA	3-hydroxy-3-phenylpropionique
IMP	Inosine 5P-monophosphate
INAP	Instituto Nacional de Investigaciones Agropecurias
	Institut Nacional de Recherches en Agronomie
LIS	Linalol synthase
PAAS	Phenylacetaldehyde synthase

PAR	Phenylacetaldehyde reductase
QTL	Quantitative Trait Loci
	Locus d'un trait quantitatif
RAPD	Random Amplification of Polymorphic DNA
	Amplification Aléatoire d'ADN polymorphe
RFLP	Restriction Fragment Length Polymorphism
	Polymorphysme de Longueur des Fragments de Restriction
RhPAAS	Rosa hybrida Phenylacetaldehyde synthase
SAH	S-adenosyl-L-homocysteine
Sam	S-adenosyl-L-methionine
SAM	Sélection Assistée par Marqueurs
Ser	Sérine
SNP	Single Nucleotide polymorphism
	Polymorphisme d'un nucléotide
SPME	Solid-Phase Microextraction
	Micro-extraction en phase solide
SSR	Simple Sequence Repeats
	Répétition d'une séquence simple
UMR	Unité Mixte de Recherche

Liste des figures

Liste des tableaux

Introduction générale

Les travaux de cette thèse portent sur l'étude des bases génétiques, génomiques et biochimiques des arômes de cacaos d'Équateur. Ils s'inscrivent dans le cadre du projet de recherche Amazcacao dont l'ambition est de développer une approche pluridisciplinaire afin de comprendre, en utilisant les outils de la génétique/génomique, de la paléogénomique, de la biochimie et des sciences sociales, comment s'est effectuée la domestication des deux variétés aromatiques anciennes de cacaoyers originaires d'Amazonie (Criollo et Nacional). Une autre finalité de ce projet est de savoir comment exploiter la biodiversité actuelle de l'Amazonie équatorienne pour y développer la culture de nouvelles variétés de cacaos fins, avec l'aide des populations locales, ce qui pourra leur apporter des revenus substantiels et améliorer leur niveau de vie. Un des modules de travail de ce projet est l'étude des déterminants génétiques et biochimiques des composants d'arômes. C'est dans ce dernier volet que s'inscrivent les travaux de cette thèse.

Afin d'étudier ces déterminants, deux populations ont été analysées. La première est constituée d'individus appartenant à la variété de Nacional moderne mais également de quelques individus de la variété de Nacional ancestrale. La deuxième population est constituée de cacaoyers natifs de la région sud de l'Amazonie équatorienne, issus de prospections effectuées dans cette région qui fait partie de la zone d'origine de *Theobroma* cacao et en particulier de celle de la variété Nacional (Loor S. et al., 2012).

Ce manuscrit est réalisé sous forme d'articles. Il est organisé en cinq chapitres :

- Une synthèse bibliographique faisant l'état de l'art sur les connaissances acquises ces dernières années sur le cacaoyer ainsi que la description d'études sur les composés aromatiques connus et identifiés chez le cacaoyer ;
- Un article portant sur le déterminisme génétique et biochimique de l'arôme floral de la variété de Nacional moderne ;
- Un article portant sur le déterminisme génétique et biochimique de l'arôme fruité de la variété de Nacional moderne ;
- Un article portant sur l'impact du processus de la domestication de la variété de Nacional moderne sur les arômes de cette variété ;
- Un article sur la variabilité et le déterminisme génétique et biochimique des arômes de cacaoyers sauvages issus de la zone d'origine du Nacional.

- Un article portant sur le déterminisme génétique et biochimique de l'amertume et l'astringence des cacaoyers de la variété de Nacional moderne et des cacaoyers de la population issue de la prospection en Amazonie ;

Une partie « discussion et conclusion générale », en fin de manuscrit, présente un résumé des résultats obtenus ainsi qu'une discussion sur les limites et les résultats de ces études. Cette partie sera suivie des perspectives envisagées à ces études.

Chapitre 1 : Synthèse bibliographique

Chapitre 1 : Synthèse bibliographique

1-Origine des cacaoyers et domestication

1.1-Historique et domestication des cacaoyers

Theobroma cacao L., est un arbre fruitier originaire des forêts tropicales humides du nord de l'Amérique du Sud. Il appartient à la famille des *Malvaceae* (Bayer and Kubitzki, 2003). Il produit ses fruits sur son tronc et ses plus grosses branches. *Theobroma cacao L.* est diploïde (2n = 2X= 20) et possède un petit génome de 430 Mb pour la variété Criollo. Cette taille est semblable à la taille du génome du riz (Eckardt, 2000).

Le genre *Theobroma* est composé de 22 espèces mais seules deux d'entre elles sont cultivées : *T. cacao*, largement cultivée à travers le monde et *T. grandiflorum* (cupuaçu), cultivée en petite quantité au Brésil notamment pour l'utilisation de sa pulpe dans la fabrication de boissons ou sorbets (Ministério da Educação, 2007). À l'état sauvage *T. cacao* peut atteindre plus de 20 mètres de hauteur.

La culture du cacaoyer (cacaoculture) s'est développée il y a plus de trois mille ans avec les civilisations précolombiennes. Les découvertes faites sur la domestication du cacaoyer sont intimement liées à celles faites sur les « zones de vie » des populations Olmèques (1500-400 avant Jésus-Christ), des Mayas (1500 avant Jésus-Christ - 1500 après Jésus-Christ) et des Aztèques (800 après Jésus-Christ jusqu'au milieu du 16ème siècle) (Bond, 2011). Plus récemment, des études ont montré que les cacaoyers étaient déjà domestiqués dans le sud de l'Amazonie équatorienne depuis près de 5000 ans (Zarillo et al., 2018). Cette découverte est pour l'instant la trace la plus ancienne de cacao consommé.

C'est en 1585 que les espagnols ont apporté le cacao en Espagne et ont fait une première version d'une boisson chocolatée. Les français, eux, planteront des cacaoyers en Martinique et à Haïti en 1750. L'introduction de ces arbres en Afrique se fera plus tardivement. C'est en 1822 que furent plantés les premiers arbres sur ce continent, sur l'île de Principe, puis leur plantation s'étendit ensuite à Sao Tomé dès 1850, à partir de matériels originaires du Brésil (Bartley, 2004).

1.2-Diversité génétique des cacaoyers

La zone d'origine du cacaoyer a longtemps été controversée. Pour Van Hall, (1914), cette zone s'étendrait de la région d'Orénoque (recouvrant la partie nord des plaines orientales

de Colombie) au bassin amazonien alors que pour Cheesman, (1944), elle se situerait plutôt au niveau du bassin amazonien colombo-équatorien.

T. cacao a initialement été classé en 3 groupes morphogéographiques : le Criollo, le Forastero et le Trinitario. Ce dernier est un groupe hybride entre les variétés Criollo et Forastero (Cuatrecasas, 1964). Le groupe Forastero est quant à lui composé d'une grande diversité de cacaoyers de différentes origines géographiques (Laurent et al., 1993; Motamayor et al., 2002). Après des études génétiques faites à l'aide de marqueurs moléculaires, les cacaoyers ont été classés en 10 groupes génétiques majeurs (figure 1) : Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañón, Nanay, Nacional et Purùs (Motamayor et al., 2008).

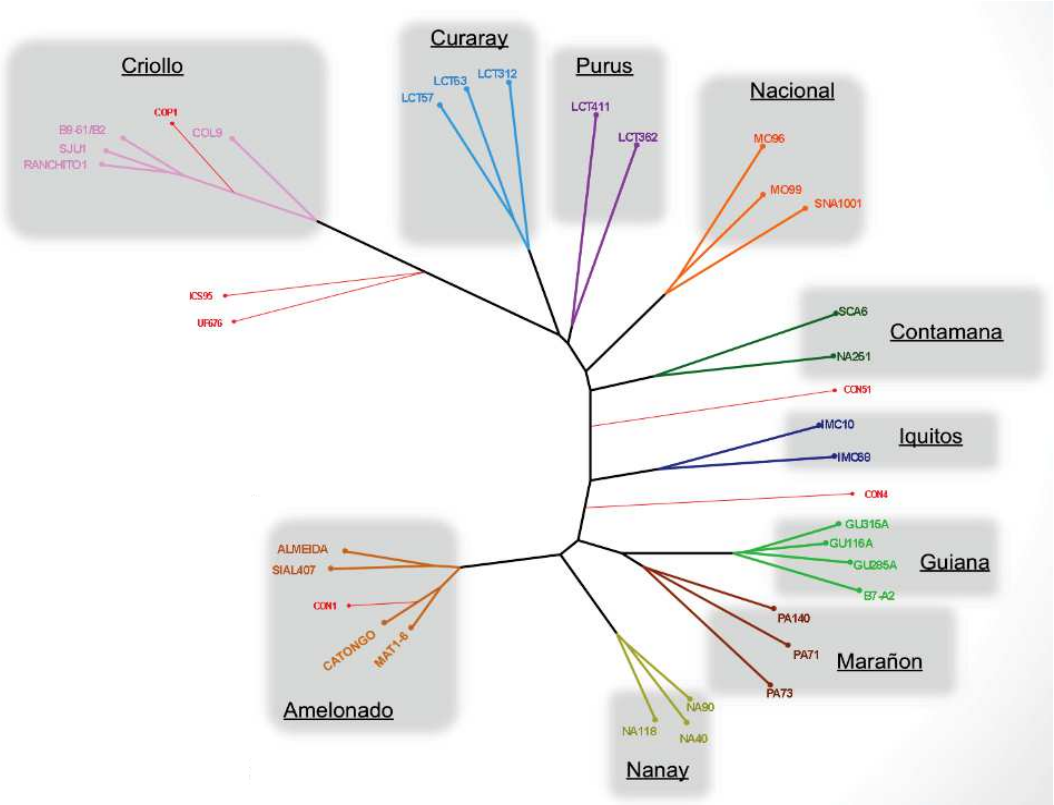


Figure 1: Arbre phylogénétique représentant les 10 groupes de cacaoyers. Défini par Motamayor et al., 2008.

De nouvelles prospections faites en Amazonie équatorienne ont montré une diversité accrue et au moins deux nouveaux groupes génétiques (Caqueta et Pangui) qui pourraient participer à la structuration de la diversité génétique de l'espèce *T. cacao* (Fouet et al, Unpublished data).

1.3-Cas particulier du cacaoyer équatorien : de la variété Nacional ancestrale au Nacional moderne

Les populations de cacaoyers de la côte pacifique de l'Équateur font partie des plus anciennes populations de cacaoyers d'Equateur à avoir été cultivées dans un but commercial.

Á la fin du XIXème siècle, le nom « Nacional » a été introduit pour définir la variété présente avant l'introduction de matériel génétique étranger qui a eu lieu à cette même époque (Bartley, 2005). Jusqu'en 1890, la variété Nacional fut la seule à être cultivée de façon prépondérante sur la côte équatorienne. Le terme « Nacional » est le plus souvent appliqué à des arbres portant des fruits allongés aux cabosses vertes avant leur maturité, à la surface rugueuse et au cortex épais (Pound, 1938). En 1938, Pound avait noté des similitudes entre les fruits rencontrés dans la haute vallée du fleuve Marañón et ceux du type Nacional. Selon cet auteur, la zone d'origine de la population de cacaoyers cultivée sur la zone littorale équatorienne serait les contreforts des Andes au niveau du fleuve Marañón. Á partir des ancêtres présumés identifiés et dans le but de déterminer la zone d'origine du Nacional, Loor S. *et al*., (2012) ont étudié, grâce à 80 marqueurs microsatellites, la ressemblance génétique entre ces ancêtres et des accessions de cacaoyers sauvages et de cacaoyers cultivés provenant de l'Amérique centrale et du sud. La plus forte similarité génétique avec les accessions de Nacional ancestraux a été observée avec les individus provenant du Sud de la région Amazonienne d'Equateur. Ces résultats suggèrent donc que le possible centre d'origine du Nacional se situe au sud de l'Amazonie équatorienne. Les récentes prospections et collectes effectuées au niveau de cette région renforcent ces observations. Elles ont révélé l'existence d'arbres dont la morphologie des fruits est très proche de ceux du type « Nacional » et également proches génétiquement (Loor S. et al., 2015).

Le Nacional a d'abord été associé au groupe Forastero par Cheesman (1944), puis il a été classé proche des Criollo par Enriquez (1992). Peu de temps après, grâce à des analyses génétiques avec des marqueurs moléculaires RAPD et RFLP, le Nacional a été classé comme un groupe à part entière (Lerceteau et al., 1997). Le Nacional ancestral était cultivé dans de petites plantations situées dans la région de la côte pacifique appelée « Arriba ». Ces arbres sont connus pour avoir un arôme très prononcé, caractérisé principalement par des notes florales connues sous le nom d'arôme « Arriba », certainement en relation avec le lieu de culture historiquement connu. Grâce à cette saveur spécifique, le cacao Nacional fut reconnu comme

« fino de aroma », c'est-à-dire fin aromatique. Le cacao fin est grandement apprécié des chocolatiers et est uniquement produit en Équateur (Loor S. et al., 2009).

D'après Bartley (2005), l'utilisation du nom « Nacional » pour définir ces arbres ancestraux sous-entend qu'il s'agissait d'une seule variété avec un génotype uniforme. Actuellement il existe des variétés populations comme c'est le cas pour Nacional moderne. Ce type de variété est composé de plusieurs individus proches génétiquement mais avec une variabilité génétique et phénotypique. Il existe également une variabilité au sein de la population de Nacional ancestrale, notamment en ce qui concerne la qualité aromatique. Les arômes présents peuvent varier d'un arbre à un autre.

Avant 1920, la diffusion des semences et la culture de cacaoyers étaient faites à partir de fèves issues de croisements incontrôlés. Après la découverte de *Moniliophthora perniciosa*, le champignon responsable de la maladie du balai de sorcière, et de ses ravages sur les cacaoyers équatoriens, des arbres résistants à la maladie ont été recherchés afin de réduire le niveau de susceptibilité de la population (Bartley, 2005).. Des cacaoyers appelés cultivars Vénézuéliens ont été introduits (Bartley, 2005; Loor S. et al., 2009; Rottiers et al., 2019). L'origine de ces arbres est incertaine. Selon Bartley, (2005) et (Loor S. et al., 2009), ils pourraient être issus de Trinidad ou du Vénézuéla. Ces variétés sont des arbres de type Trinitario (hybrides entre le Criollo et l'Amelonado) et ils sont apparus plus résistants aux maladies et plus producteurs que le Nacional ancestral.

La variété de Nacional moderne actuellement cultivée se compose donc d'hybrides entre les types Nacional ancestraux et des types Trinitario. Cette nature hybride a été démontrée par Loor S. et al., (2009) grâce aux marqueurs moléculaires. La culture généralisée des nouveaux matériels génétiques par les grandes propriétés a contribué à l'accélération du brassage génétique ainsi qu'à la dilution du Nacional ancestral. Un deuxième facteur participant à cette dilution est l'introduction dès 1940 d'autres génotypes étrangers (Bartley, 2005). En effet, du matériel résistant principalement à la maladie du balai de sorcière a également été collecté lors de plusieurs prospections effectuées dans la partie haute de l'Amazonie (Pound, 1938).

Ce mélange génétique a mené à une dilution de la saveur Arriba (Loor S. et al., 2009; Beckett et al., 2017). À partir de 1940, des prospections sur la côte équatorienne ont eu lieu afin de collecter des cacaoyers de type Nacional pour en préserver les ressources génétiques. Ces collectes ont été placées dans deux principales stations expérimentales appartenant à l'INIAP (Instituto Nacional de Investigaciones Agropecurias) : la station Expérimentale Tropical de

Pichilingue (EET-P) et à l'Université : le Centre des Arômes de Cacao de Tenguel (CCAT) (Loor S., 1998).

Jusqu'en 1994, l'Équateur est le premier pays producteur de cacao fin. Mais en juillet de cette même année, les problèmes sanitaires précédemment évoqués, le vieillissement des vergers de productions ainsi que la sélection et la diffusion dans tout le pays d'un clone haut producteur mais non aromatique, le CCN51, a conduit l'ICCO à déclasser le pourcentage de cacao fin produit dans le pays de 100% à 75% (Petithuguenin and Roche, 1995).

Grâce à une vaste étude de diversité génétique de populations natives de cacaoyers issues d'un grand nombre de pays d'Amérique du Sud, Loor S. et al., (2012) ont montré que c'est de la partie sud de l'Amazonie équatorienne qu'est originaire le Nacional ancestral. Suite à ces résultats, différentes prospections ont été effectuées afin de caractériser et sauver la diversité génétique existante des cacaoyers apparentés au Nacional (Loor S. et al., 2016). La parenté des arbres de cette région avec le Nacional ancestral a été confirmée. Une diversité génétique importante a en outre été observée (Fouet et al, Unpublished data).

1.4-Séquençage du génome du cacaoyer

La variété Criollo a été la première variété de cacaoyer à avoir été séquencée en 2011 (Argout et al., 2011). Le génome complet est disponible sur un « génome browser » (http://cocoa-genome-hub.southgreen.fr/gbrowse). Une deuxième version plus complète a été publiée en 2017 (Argout et al., 2017). Cette seconde version est également disponible en ligne (http://cocoa-genome-hub.southgreen.fr/). Elle a été produite à partir des nouvelles technologies de séquençage (NGS) qui ont permis d'ancrer 99% des gènes sur les chromosomes. En 2013, la variété Amelonado a été séquencée (Motamayor et al., 2013), ce qui constitue une source supplémentaire d'informations pour les études génétiques dédiées aux programmes d'amélioration des cacaoyers.

2-Biologie de la plante

2.1-Système de reproduction

T. cacao peut facilement être multiplié par reproduction sexuée ou végétative. Il est parfois également possible de réaliser de la culture in-vitro de cellules de cacaoyers pour certains génotypes, et de procéder à la régénération des cellules en plantules (Li et al., 1998).

2.1.1-Floraison

Selon les variétés et l'environnement de culture, le cacaoyer ne présente pas toujours les mêmes périodes de floraison. La floraison peut s'étendre sur toute l'année en présentant des pics à certaines époques (Glendinning, 1972). Souvent, deux périodes sont observées : une forte floraison donnant lieu à la période de récolte principale et une floraison plus faible donnant lieu à la récolte intermédiaire. La saison des pluies fait baisser le taux de pollinisation. La pollinisation est plutôt effectuée le matin (Glendinning, 1972) via des insectes pollinisateurs volants, dont le principal, *Forcipomyia midge*, frotte son thorax sur les anthères (Posnette, 1950).

La fleur du cacaoyer est petite, mesure environ 1cm (figure 2), est composée de cinq sépales rosés ou pourpres appelées staminodes qui alternent avec les pétales recourbées abritant les anthères (Bouharmont, 1960).

Figure 2: Photo d'une fleur de cacaoyer. Photo : Colonges Kelly, 2015.
a : pétales et sépales, b : cucule, c : staminode, d : style et stigmate, e : étamines.

Les ovaires peuvent contenir environ 40 ovules à nucelle dont le sac embryonnaire mesure 5µm (figure 3A). Il est constitué de :

- Deux synergides, de grandes cellules allongées s'étendant jusqu'au milieu du sac embryonnaire ;
- Une oosphère dont le noyau est le gamète femelle ;
- Deux noyaux polaires volumineux, toujours accolés voire même soudés, entourés de protoplasmes et de grains d'amidon (figure 3B). Ils fusionneront avec un des noyaux du gamétophyte mâle pour donner l'endosperme ;
- Deux cellules antipodales qui persistent exclusivement dans les sacs embryonnaires à structure aberrante (Bouharmont, 1960).

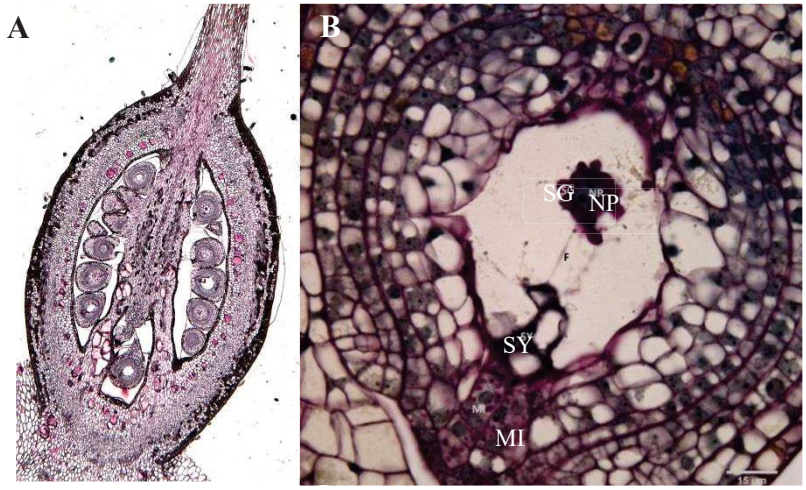


Figure 3: A : Coupe longitudinale d'un ovaire de cacaoyer. B : Coupe longitudinale d'un ovule de cacaoyer 15h après pollinisation incompatible colorée au Schiff et à l'hématoxyline de Johansen. Photos : Colonges Kelly, 2015.
SY : synergides, NP : noyaux polaires, SG : grains d'amidon, F : filaments, MI : zone micropylaire.

Les ovaires sont petits et surmontés d'un style mince avec cinq stigmates. L'ensemble style et stigmates mesure 3 mm. L'ovaire est entouré de cinq staminodes (étamines stériles ou avortées, souvent rudimentaires) (Glendinning, 1972). La fleur est aussi composée d'étamines aux filets courts et recouverts d'un pétale recourbé, le cucule. Les grains de pollen d'un diamètre d'environ 20 µm sont sphériques. L'exine est épaisse et montre des dessins réticulés avec trois pores germinatifs. Ces fleurs sont regroupées sur des coussinets, principalement sur les troncs et parfois sur les branches. Quand celles-ci ne sont pas pollinisées, elles tombent un jour après l'anthèse (Bouharmont, 1960).

Les tubes polliniques pénètrent dans la majorité des cas simultanément dans les ovules de la fleur. Le déversement de leur contenu s'effectue dans les synergides. Les deux gamètes mâles vont ensuite se placer à proximité de l'oosphère et des noyaux polaires. L'un des gamètes mâles va fusionner avec l'oosphère pour former l'embryon diploïde ou zygote et l'autre va fusionner avec les deux noyaux polaires pour former l'endosperme triploïde (Bouharmont, 1960).

Il a été démontré que les rendements élevés étaient corrélés à l'auto-compatibilité mais pas à la vigueur (Lachenaud et al., 2005).

2.1.2-Système d'auto-incompatibilité chez le cacaoyer

En fonction de leur origine génétique, les cacaoyers peuvent être autogames ou allogames. Ce dernier système est renforcé par le système d'auto-incompatibilité (AIC) gaméto-sporophytique qui a pu être observé et étudié sur cet arbre (Cope, 1939; Cope, 1940; Knight and Rogers, 1955; Cope, 1958; Bouharmont, 1960; Cope, 1962; Glendinning, 1967). Les plus

fortement touchés par l'AIC sont les cacaoyers de haute-Amazonie qui sont également fortement hétérozygotes. Ce niveau d'hétérozygotie pourrait refléter le brassage génétique entre les populations mais aussi le mode de reproduction majoritairement allogame (Hamon et al., 1999). Le contrôle génétique du système d'AIC de *T. cacao* a été étudié par plusieurs auteurs, qui ont émis l'hypothèse de l'existence d'un locus S et de plusieurs allèles avec des relations de dominance entre eux (Knight and Rogers, 1955; Cope, 1962; Glendinning, 1967). D'après Cope, (1962), deux autres gènes nommés A et B interviendraient également dans le mécanisme de l'AIC et l'un de ces deux gènes à l'état homozygote rendrait le cacaoyer auto-compatible.

Plus récemment, deux locus ont été identifiés sur les chromosomes 1 et 4 comme impliqués dans l'auto-incompatibilité du cacao par deux processus différents. Le locus du chromosome 1 agit avant l'étape de fusion des gamètes et indépendamment du locus du chromosome 4. Les deux locus sont responsables de la sélection gamétique, mais seul le locus sur le chromosome 4 est impliqué dans la chute principale des fruits (Lanaud et al., 2017).

Chez *Theobroma cacao*, le système d'AIC se manifeste par des troubles postérieurs au développement du tube pollinique (Enriquez and Alarcòn, 1977). Ils se manifestent principalement par une non fusion des gamètes mâles et femelles. Le mécanisme d'AIC impacte donc fortement le rendement des cacaoyers.

2.2-Physiologie du fruit et composition des fèves

Les fruits du cacaoyer sont de grosses baies allongées appelées cabosses (Bouharmont, 1960). Elles sont attachées au tronc ou aux branches maîtresses par un court et fort pédoncule. Les cabosses présentent une grande diversité morphologique de couleurs, de formes, de textures et de tailles dépendant principalement des génotypes des arbres (figure 4). La couleur peut

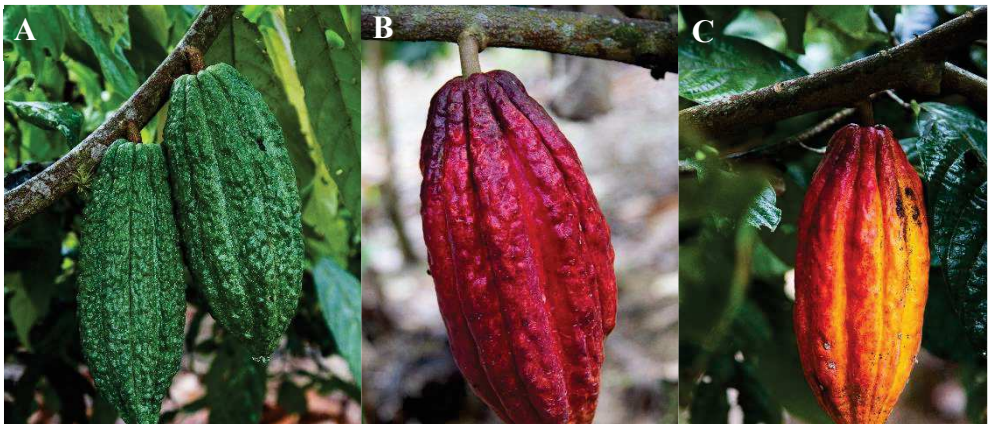


Figure 4: Cabosses de cacaoyer. A: cabosses immatures de cacaoyer, B: cabosse immature de cacaoyer de type Criollo, C: cabosse en maturation de cacaoyer de type Criollo. Photos : Lans Tom, Valrhona, 2019.

également dépendre du stade de maturité du fruit (figure 4 B et C) qui arrive à son terme 4 à 6 mois après floraison. La taille est généralement de 12 à 30 cm de long (Kongor et al., 2016).

Chaque cabosse contient une quarantaine de fèves. Elles correspondent aux embryons diploïdes et sont constituées de deux cotylédons intimement liés et recouverts d'une coque (ou tégument séminal), elle-même recouverte d'une couche de pulpe appelée mucilage (figure 5). La pulpe représente 40% du poids frais de la fève (Cros and Jeanjean, 1995).

Les cotylédons sont très riches en matières grasses (50 à 55% de la masse sèche des fèves). Ils contiennent également 17% de fibres, 12% de glucides, 10 à 12% de protéines, 7 à 15% de polyphénols et de tannins, 2% de théobromine (composé de la famille des méthylxanthines comme la caféine), 2% de sels minéraux et d'oligoéléments et 1% de caféine (Cirad, 1999; Kadow et al., 2013).

Les fèves de cacao sont riches en antioxydants par leur teneur en catéchines, épicatéchines et procyanidines. Ces molécules de la famille des polyphénols sont également présentes dans le vin, le thé et même certains légumes (Afoakwa et al., 2008). Les fèves fraîches

Figure 5 : Cabosse mature ouverte. Les fèves sont recouvertes d'un mucilage blanc. Photo: Colonges Kelly 2019.

de cacao sont connues pour avoir un goût astringent dû à la forte concentration en polyphénols, comme les flavan-3-ols (Wollgast and Anklam, 2000). Les fèves fraîches de cacao de type fins aromatiques n'ont pas toujours une astringence ou une amertume prononcées.

La pulpe aqueuse contient 12 % de mono et disaccharides, 2% d'acide citrique ainsi que d'autres acides organiques, mais encore des composés volatils comme des esters, aldéhydes, méthyl-cétones, alcool secondaires et terpènes (Kadow et al., 2013).

Le cacao marchand est obtenu à partir des fèves après plusieurs étapes de transformation. La qualité du cacao dépend de plusieurs critères. Premièrement, la taille des

fèves, l'homogénéité du grainage, la qualité et la teneur en beurre ainsi que la teneur en métaux lourds, sont des critères liés au génotype mais aussi à l'environnement et aux pratiques culturales. Enfin, la couleur ou les arômes, sont des critères liés à l'origine génétique des fèves, aux traitements post-récolte (fermentation et séchage) et à la transformation du produit, principalement lors de la torréfaction (Cros and Jeanjean, 1995).

3-Conditions et lieux de culture

Le cacaoyer est un arbre fruitier qui pousse préférentiellement dans les zones ombragées dans les étages inférieurs (1250 mètres d'altitude maximum) des forêts humides tropicales. À l'état sauvage, il peut atteindre jusqu'à 20 mètres de hauteur et ne pousse que dans les zones tropicales ayant une température moyenne de 24°C et ne descendant pas au-dessous de 10°C (Jumelle, 1900).

En culture, le cacaoyer atteint en général 4 à 5 mètres de hauteur (Edoh Adabe and Ngo-Samnick, 2014). La culture de cacaoyer requiert préférentiellement un climat tropical humide avec une pluviométrie régulière. Le cacaoyer préfère les sols profonds, riches en matières organiques et bien drainés (Jumelle, 1900). Un hectare de cacaoyer peut fournir entre 300 et 3000 kg de cacao par an. Il commence à produire à partir de sa quatrième année (Essola Etoa, 2014).

3.1-Production mondiale de cacao

Actuellement, la cacaoculture est une source de revenu importante pour beaucoup d'agriculteurs en Afrique, en Amérique latine et en Asie, seuls à pouvoir produire du cacao grâce à leur situation géographique. En 2018, la production mondiale de cacao s'élevait à plus de 4 000 mille tonnes. Les trois plus grands producteurs étant la Côte d'Ivoire, le Ghana et l'Équateur avec respectivement : 2 150, 900 et 298 mille tonnes (Shahbandeh, 2019). Les fèves de cacao sont la clef de l'industrie du chocolat qui se chiffre en millions de dollars (Pipitone, 2012).

Le cacao est classé en deux types de produits : les cacaos dits standards au goût de cacao prononcé et les cacaos dits fins aux notes florales et fruitées (Sukha et al., 2008). La production de cacao « standards » représente environ 95% de la production mondiale. La production de cacaos fins représente donc environ seulement 5% de la production mais n'en est pas moins importante. Certains pays d'Amérique latine produisent presque exclusivement du cacao fin qui est pour eux un revenu essentiel. À ce jour, les types de cacao dit fins les plus cultivés sont la variété Nacional, la variété Criollo et les arbres dits « Trinitario ». Les Trinitarios sont des

hybrides entre la variété Criollo et la variété Amelonado. La variété Criollo est peu cultivée à cause de sa faible vigueur et de sa sensibilité accrue aux maladies (Cheesman, 1944).

3.2-Les principaux problèmes phytopathologiques

3.2.1-Le balai de sorcière

La maladie du balai de sorcière est causée par un champignon hémibiotrophe, le *Moniliophthora perniciosa*, anciennement appelé *Crinipellis perniciosa*. Elle a été observée pour la première fois au Surinam. Elle s'est ensuite étendue à la Guyane, à l'Équateur et à Trinidad et Tobago où elle a eu un effet dévastateur sur le développement de l'agriculture cacaoyère. En 1973, l'Équateur était le pays le plus affecté par la maladie et c'est ainsi que des études de compréhension du champignon ont été lancées (Evans, 1978). Á la fin du XX^{ème} siècle, il est considéré comme le facteur limitant majeur de la production de cacao en Amérique du Sud et aux Caraïbes. L'infection des tissus méristématiques du cacaoyer (jeunes pousses de feuilles, fleurs et cabosses) s'effectue via les basidiospores. Elle entraîne la croissance de balais verts ramifiés et gonflés à l'intérieur desquels prolifère le mycélium biotrophe intercellulaire (figure 6A). La mort de ce balai vert intervient 4 à 6 semaines après l'infection et est suivie de l'invasion des tissus de l'hôte par un mycélium plus étroit (figure 6B et C). Le cycle de vie se termine par la production de basidiocarpes à la surface des balais 6 à 24 mois plus tard (Griffith and Hedger, 1993).

Figure 6: Jeunes pousses de cacaoyer infectées par la maladie du balai de sorcière. A : Stade précoce de l'infection, B : mort de l'organe foliaire et dispersion des spores du champignon, C : 1, jeune cabosse infectée par la maladie du balai de sorcière ; 2, mort du jeune fruit après infection et dispersion des spores du champignon. Photos : Colonges Kelly,2019.

3.2.2-La Moniliose

La Moniliose est une maladie causée par le champignon *Moniliophtora roreri*. Elle est responsable de la pourriture des cabosses, souvent de couleur blanche, d'où sa dénomination

anglaise « frosty pods » qui signifie cabosses gelées (figure 7). Ce champignon pathogène n'envahit que les cabosses de *T. cacao* dans leur phase de croissance active et celles des espèces apparentées aux genres *Theobroma* ou *Herrania* (Phillips-Mora and Wilkinson, 2007). Des analyses préliminaires de données de séquence de *M. roreri* ont montré que cet agent pathogène pourrait être étroitement lié au *Crinipellis perniciosa* (nouvellement : *Moniliophthora perniciosa*) malgré leurs différences morphologiques. Des similitudes entre les procédés d'infection ont été notées (Griffith et al., 2003). Jusque dans les années 1950, *M. roreri* était uniquement présent dans la partie nord-ouest de l'Amérique du Sud. En 2007, il a été retrouvé dans onze pays d'Amérique du Sud. Le champignon est dans une phase de dispersion active, probablement dûe à l'homme. La moniliose est plus destructrice que les espèces de *Phytophthora* (présentés dans le prochain paragraphe) et plus dangereuse et difficile à contrôler que la maladie du balai de sorcière. L'agressivité de *M. roreri*, sa capacité à survivre dans

Figure 7: Cabosse infectée par la moniliose. Photo : Colonges Kelly, 2019.

différentes conditions climatiques, sa dispersion naturelle rapide et la sensibilité de la plupart des génotypes commerciaux de cacaoyers font de ce champignon une véritable menace pour la culture du cacaoyer en Amérique et dans le monde entier (Phillips-Mora and Wilkinson, 2007).

3.2.3-Les espèces de *Phytophthora*

Différentes espèces de *Phytophthora* sont impliquées dans la pourriture des cabosses, le chancre des tiges et la brûlure des feuilles du cacaoyer. Le *Phytophthora palmivora* est la première espèce à avoir été identifiée comme responsable de la pourriture des cabosses. Il est présent dans toutes les régions productrices de cacao. Dans les années 1970-1980, plusieurs auteurs ont révisé la classification effectuée ultérieurement à partir de la taille et du nombre de chromosomes. Alors quatre espèces différentes de *Phytophthora* ont été discriminées et décrites : *P. palmivora, P. megakarya, P. capsici* and *P. citrophthora* (Brasier and Griffin,

1979; Babacauh, 1983; Weinert et al., 1999). Quand les conditions climatiques sont favorables à la croissance de *P. palmivora*, les pertes peuvent atteindre jusqu'à 30% de la récolte. La maladie se manifeste d'abord par une tache brune à noire sur le fruit, dont la taille augmente avec le développement du parasite. Plusieurs jours après l'infection, de nouveaux symptômes apparaissent : un mycélium blanc avec des sporanges se développe sur la surface du fruit malade puis les fruits se momifient. Dans le but de réduire la propagation de la maladie, les fruits malades sont récoltés et détruits (Flament et al., 2001). Plusieurs études sur le déterminisme génétique de la résistance au *Phytophthora* ont été menées dans le but d'inclure la résistance génétique dans les programmes de sélection (Nyassé et al., 1995; Flament et al., 2001; Risterucci et al., 2003; Lanaud et al., 2009; Akaza et al., 2016).

3.2.4-Le virus CSSV (Cocoa Swollen Shoot Virus)

Le virus CSSV est un badnavirus qui peut être transmis par au moins quatorze espèces de cochenilles de la famille des *Pseudococcidae*. Il touche principalement l'Afrique. Il est présent dans les principales zones de culture du Ghana (Thresh et al., 1988) et s'est répandu en Côte d'Ivoire ces dernières années. Plusieurs auteurs pensent que le CSSV était présent dans les régions forestières d'Afrique de l'Ouest avant l'introduction du cacaoyer (Posnette, 1951; Thresh, 1961; Thresh et al., 1988; Quainoo et al., 2008). Les arbres appartenant au groupe génétique des Trinitario (hybrides entre Criollo et Amelonado) possèdent une certaine tolérance (Thresh et al., 1988). Cependant, ces arbres sont moins représentés au Ghana où le groupe majoritairement représenté est l'Amelonado qui est lui-même sensible au CSSV. Une plus grande tolérance a été observée chez certains individus provenant de la haute-Amazonie. Des croisements ont ensuite été réalisés entre les Trinitario et les Amelonado présents en Afrique et des cacaoyers haut-amazoniens (Thresh et al., 1988) dans le but d'obtenir des arbres plus résilients.

4-Processus de transformation de la fève en chocolat

Après récolte, les fèves de cacao sont transformées en chocolat par un processus qui comporte plusieurs étapes : fermentation, séchage, torréfaction, broyage, conchage, tempérage. Au cours de celles-ci, les fèves vont libérer différents composés qui pourront évoluer et qui seront notamment responsables de la qualité aromatique du chocolat final. La somme de toutes ces transformations a un impact sur le potentiel aromatique du chocolat final (Afoakwa et al., 2008).

4.1-La fermentation et le séchage

L'étape de fermentation du cacao constitue le premier pas de sa transformation en

chocolat. Après récolte des cabosses et écabossage, les fèves sont séparées les unes des autres.

Figure 8: Caisse de fermentation en bois en remplissage de fèves avant mise en fermentation. Photo : Colonges Kelly, 2019. Des petits sacs remplis de fèves de différents génotypes sont répartis dans la caisse pour réaliser des micro-fermentations.

La mise en fermentation des fèves de cacao avec la pulpe peut se faire suivant différentes méthodes. La fermentation peut être réalisée dans des caisses en bois (figure 8), des sacs de jute, ou des caisses en plastique par exemple. La méthode la plus utilisée est la mise en fermentation dans des caisses en bois. Des feuilles de bananiers sur le haut des caisses limitent les pertes de chaleur. Dans ces conditions, le mélange de pulpe et de fèves est un milieu anaérobie, c'est-à-dire sans oxygène, où les levures vont pouvoir se développer (Ho et al., 2014).

Plusieurs espèces de levures ont été retrouvées dans les caisses de fermentation de cacao. Les genres les plus couramment retrouvés sont *Candida*, *Pichia* et *Saccharomyces*. Lors de cette phase, les levures transforment les sucres contenus dans la pulpe en alcool. Ensuite, les fèves à l'intérieur des caisses sont brassées afin de faire entrer de l'air pour favoriser la deuxième phase de la fermentation. Cette entrée d'air permet de passer d'un milieu anaérobie à un milieu aérobie, ce qui permet le développement de bactéries acétiques. Elles vont alors transformer l'éthanol en acide acétique (Matsushita et al., 1994). Les acides produits au niveau de la pulpe vont entrer dans les cotylédons.

L'acidification des fèves pendant la fermentation provoque plusieurs réactions biochimiques (Qin et al., 2017). Certaines, comme les réactions d'hydrolyses et/ou d'oxydations sont à l'origine de la synthèse de précurseurs d'arôme. Leurs concentrations vont donc dépendre de ces mécanismes enzymatiques mais vont également varier selon les fermentations (Biehl et al., 1982; Afoakwa et al., 2008). L'oxydation de certains composés phénoliques permet également la réduction de l'astringence et de l'amertume (Afoakwa et al.,

2008). D'autres réactions permettent la perte du pouvoir germinatif des fèves ce qui entraine la mort du cotylédon (Biehl et al., 1982; Schwan and Wheals, 2004; Afoakwa et al., 2008). C'est également pendant cette phase que les goûts floraux et fruités apparaissent (Afoakwa et al., 2008; Rodriguez-Campos et al., 2011).

L'étape de la fermentation est une étape clef car elle permet le développement d'un grand nombre de composés volatils aromatiques qui ne sont pas observés dans les fèves séchées non fermentées (Utrilla-Vázquez et al., 2020). Comme cela a été prouvé pour les grains de cafés, il est probable que les molécules produites lors de la fermentation à la surface de la pulpe des fèves de cacao pénètrent dans la fève et produisent des réactions chimiques et enzymatiques (Hadj Salem et al., 2020).

La durée de la fermentation varie en fonction des génotypes. Généralement, elle dure de 5 à 7 jours pour les fèves de cacao de type Forastero (Amelonado) alors que les fèves de cacao fines aromatiques, Criollo ou Nacional, sont fermentées pendant 3 à 4 jours (Cevallos-Cevallos et al., 2018).

L'étape du séchage des fèves permet l'arrêt de la fermentation. Le but de cette étape est de réduire le taux d'humidité à moins de 8% afin de permettre le stockage des fèves sans qu'elles évoluent et de réduire le risque de développement des moisissures (Afoakwa et al., 2008).

Durant le séchage, une initiation de la transformation des précurseurs d'arôme en pyrazines et en aldéhydes via la réaction de Maillard peut se réaliser. De plus, il permet l'évaporation d'acide volatils et de l'eau contribuant ainsi à la continuité du développement des arômes. Il a été observé que la concentration en acide acétique augmente pendant le séchage. Cette observation suggère que les bactéries acétiques continuent d'en produire lors de cette étape de séchage (Rodriguez-Campos et al., 2011).

4.2-Torréfaction

Les fèves fermentées séchées sont appelées « cacao marchand ». Ce sont elles qui vont être vendues et utilisées pour la fabrication de chocolats. La transformation de cacao marchand en chocolat commence par la torréfaction. C'est à ce moment que la plupart des goûts empyreumatiques (grillé, chocolat) apparaissent. La torréfaction qui est effectuée à des températures élevées permet la majorité des modifications dues à la réaction de Maillard (Afoakwa et al., 2008).

La réaction de Maillard joue un rôle important dans la qualité des aliments depuis que la cuisson existe. Elle est un défi majeur pour l'industrie alimentaire car elle joue un rôle important, principalement dans l'apparence et le goût des aliments cuits. Elle intervient par exemple lors de la torréfaction du café ou des fèves de cacao marchand, de la cuisson des viandes (grillades) ou encore lors de la cuisson du pain (Martins et al., 2000).

La réaction de Maillard est très complexe. En réalité ce n'est pas une seule réaction chimique mais plusieurs dizaines en parallèle ou qui se succèdent et qui permettent la synthèse d'une large gamme de produits de réactions. Elle est représentée simplement dans la figure 9 selon le schéma de synthèse réalisé par Starowicz and Zieliński, (2019). Ces réactions influent sur la valeur nutritionnelle des aliments à différents niveaux : la digestibilité, la formation de composés toxiques ou mutagènes et l'élimination d'antioxydants (Martins et al., 2000). Bien que la réaction de Maillard soit plutôt bien connue, il est très difficile de la contrôler.

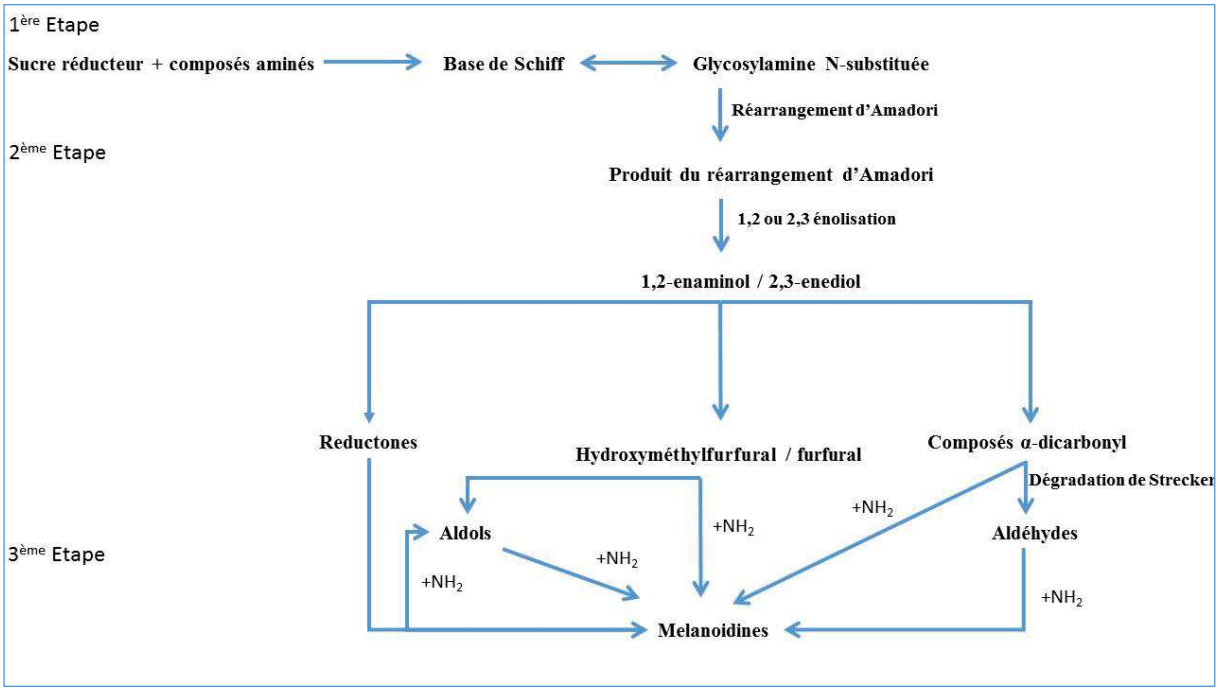


Figure 9: Schéma simplifié de la réaction de Maillard. Adaptée de Starowicz and Zieliński, (2019).

La première étape de la réaction de Maillard consiste en la condensation d'un sucre réducteur et d'un composé possédant un groupe amino libre (souvent un acide aminé libre mais aussi des polypeptides peu complexes). Cette première réaction permet la formation d'une base de Schiff puis d'une glycosylamine N-substituée qui va se réarranger et former ainsi ce que l'on appelle le produit du réarrangement d'Amadori ou de Heins. La dégradation de ces composés dépend du pH de la matrice dans laquelle ils se trouvent. D'un pH neutre à acide (pH≤7), lorsque des pentoses sont impliqués, il y a une formation plus spécifique de furfural ; lorsque des

hexoses sont impliqués, il y a formation d'hydroxyméthylfurfural. À un pH basique (pH>7), il y a formation de réductones et de plusieurs types de produits de fission tels que l'acétol, le pyruvaldéhyde et le diacétyle. Puis tous ces produits très réactifs participent à d'autres réactions (Martins et al., 2000). La formation de pyrazines a été décrite lors de divers processus agroalimentaires, principalement lors des processus de torréfaction à partir d'acides aminés et de sucres lors de la réaction de Maillard (Dickschat et al., 2010).

4.3-Raffinage, conchage, tempérage

Les dernières étapes de transformation (raffinage, tempérage, conchage) permettent d'obtenir une meilleure texture de chocolat. Le raffinage permet de réduire la taille des particules des fèves de cacao à quelques micromètres de diamètre (Beckett et al., 1994). Lorsque cette étape est bien réalisée, le chocolat ne présente pas de texture granuleuse en bouche.

Le conchage est une étape essentielle pour le développement des dernières saveurs. Il permet l'évaporation des derniers résidus d'acides et d'eau. Il permet également de parfaire la texture en modifiant les cristaux de sucre et la viscosité du chocolat et en lui apportant une onctuosité spécifique. Suivant la durée de cette étape, il y a également un changement de couleur dû à l'oxydation et à l'émulsion des tannins (Reineccius, 2006; Afoakwa et al., 2008).

Le tempérage du chocolat consiste en une succession d'augmentations puis de diminutions de la température afin de produire une cristallisation de la matière grasse du beurre de cacao sous sa forme β, qui confère sa brillance au chocolat (Dhonsi and Stapley, 2006).

5-Les différents arômes présents dans les fèves de cacao

Selon les variétés et leurs caractéristiques aromatiques, les fèves de cacao sont classées en deux lots : les cacaos standards au caractère amer et forts en cacao, et les cacaos fins caractérisés par des notes florales et fruitées (Cook and Meursing, 1982; Sukha et al., 2008). Afin de déterminer et de classer les cacaos dans ces différentes catégories, il a été proposé de caractériser les différences physico-chimiques et organoleptiques contenues dans les fèves (Amores et al., 2007).

Différents profils aromatiques ont été observés en fonction de l'origine du cacaoyer. Au Cameroun, les liqueurs de cacao sont connues pour leur amertume. En Équateur, le cacao est réputé pour ses notes florales et épicées. D'après Afoakwa et al., (2008), les variétés provenant de Trinidad présentent des arômes vineux. Les fèves d'Asie et d'Océanie présentent des profils

gustatifs variés, allant des notes de cacao et de noix (fèves de Java) aux notes acides et phénoliques intenses (fèves de Malaisie).

Les cacaos fins les plus cultivés sont le Nacional, le Criollo et le Trinitario (famille hybride entre le Criollo et l'Amelonado). La variété Nacional spécifique à l'Équateur est reconnaissable à son arôme floral, appelé Arriba, et présente également des notes épicées (Luna et al., 2002; Afoakwa et al., 2008) avec peu d'astringence et d'amertume (International Cocoa Organization, 2017). La variété Criollo est caractérisée par des notes fruitées et les fèves issues de Trinitario ont des notes fruitées et florales.

En plus du génotype, le potentiel aromatique des fèves de cacaos peut être affecté par l'environnement dans lequel il est cultivé (Kongor et al., 2016). Les arômes fins peuvent être produits pendant la fermentation (Rodriguez-Campos et al., 2011).

Il a été démontré que certaines molécules odorantes peuvent être spécifiques d'un génotype de cacao aromatique (Kadow et al., 2013). Cette étude compare les fèves fraîches et la pulpe de deux génotypes de cacaos fins et aromatiques SCA6 et EET62, au génotype de cacao standard CCN51. Les auteurs ont pu montrer que des molécules aromatiques étaient présentes uniquement chez SCA6 ou EET62 permettant ainsi de conclure que les arômes des cacaos fins et aromatiques dépendaient de plusieurs composés d'arôme. En conséquence, les arômes de ces cacaos pourraient être le résultat de la mise en place de différentes voies métaboliques dépendantes des génotypes (Kadow et al., 2013). Le génotype SCA6 contient des monoterpènes (β-myrcène, β-trans-ocimène et β-cis-ocimène) que les deux autres génotypes ne contiennent pas ou seulement à l'état de traces. En revanche, EET62 contient de l'acétate de 2-heptanol, 2-heptanone, 2-heptanol et 2-nonanone que les deux autres génotypes ne contiennent pas ou seulement à l'état de traces (Kadow et al., 2013).

Dans les cacaos fins, plusieurs molécules ont déjà été identifiées comme responsables des saveurs fruitées et florales. Une première famille de composés a été mise en évidence, avant et après fermentation : celle des terpènes (Ziegleder, 1990; Kadow et al., 2013; Qin et al., 2017; Cevallos-Cevallos et al., 2018). Un composé de la famille des monoterpènes semble particulièrement intéressant : le linalol. D'après Ziegleder (1990), la spécificité de l'arôme floral des cacaos fins d'Équateur (Nacional) s'expliquerait en partie par la présence accrue de linalol. Cette concentration varierait en fonction de l'origine génétique du cacao. Du linalol et de l'époxylinalol, connus tous deux pour avoir une note florale, ont été retrouvés dans les fèves de cacao fins fermentées et non dans les fèves de Forastero. Plus précisément, une forte

augmentation de la concentration de ces terpènes après trois jours de fermentation dans les fèves issues du Criollo (Cevallos-Cevallos et al., 2018) et une plus grande quantité dans les fèves non fermentées séchées de Criollo et de Trinitario (Qin et al., 2017) ont été avérées.

D'autres terpènes, comme le (Z)-ocimène ou le β-myrcène, à l'origine des notes épicées et florales des fèves de cacao (Arn and Acree, 1998), ont été identifiés dans les cacaoyers fins et aromatiques, dans cette même étude.

Utrilla-Vázquez et al., (2020) ont récemment comparé plusieurs types de Criollo et de Trinitario à différentes étapes du processus de transformation : des fèves non fermentées séchées, au début et à la fin de la fermentation ainsi qu'au début et à la fin du séchage. Dans cette étude, plusieurs molécules aux notes florales ont été détectées dans toutes les conditions examinées : l'alcool benzylique, l'acétophénone et le 2-phényléthanol, ce qui suggère que ces molécules ont été synthétisées par le cacaoyer.

6-Voies de biosynthèse des composés aromatiques aux notes florales

6.1-Voie de dégradation de la L-phénylalanine

Dans les années 1990, la voie de dégradation de la L-phénylalanine a commencé à être étudiée, caractérisée et décrite chez les organismes unicellulaires. Pour la première fois, deux voies de dégradation de la L-phénylalanine ont été décrites chez un organisme plus complexe : le champignon *Bjerkandera adusta* (figure 10), l'une étant une voie de dégradation non oxydative permettant la synthèse du 2-phényléthanol et l'autre étant la voie de β–oxydation permettant la synthèse d'acétophénone (Lapadatescu et al., 2000). Chez le camélia, une voie de biosynthèse de l'acétophénone, différente de celle décrite précédemment, a été caractérisée. Premièrement, la L-phénylalanine est transformée en acide cinnamique lui-même transformé ensuite en acide 3-hydroxy-3-phenylpropionique transformé à son tour en acide 3-phenylpropionique, pour finir lui-même transformé en acétophénone (Dong et al., 2012). L'acétophénone a été trouvé dans le jus de raisin muscadine et le camélia (Baek et al., 1997; Dong et al., 2012).

L'acétate de benzyle est connu pour sa note florale typique de jasmin (van Schie et al., 2006; Melo et al., 2017). Il est synthétisé à partir de l'alcool benzylique grâce à une réaction de transestérification (Melo et al., 2017). Il est généré grâce à l'action des alcools acétyle transférases (AATs) (Guterman et al., 2006) (figure 10).

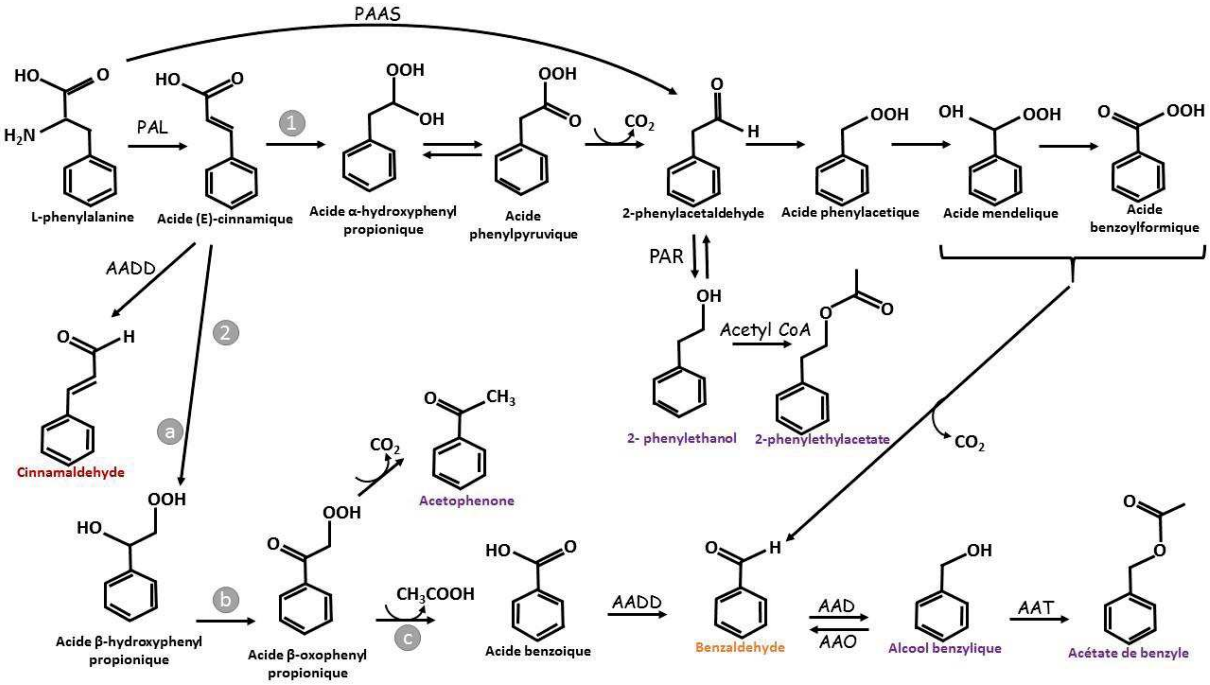


Figure 10: : Schéma des voies de dégradation de la L-phénylalanine par Bjerkandera adusta. Adapté de Lapadatescu et al., 2000.

(1), Voie non-oxydative. (2), Voie β-oxydative. a, b, c : Séquence de la β-oxydation AAD : Aryl-alcool déshydrogénase, AADD : Aryl-aldéhyde déshydrogénase, AAO : Aryl alcool oxydase, AAT : alcools acétyle transférases, PAAS : Phenylacetaldehyde Synthase, PAL : L-phénylalanine amonia lyase, PAR : Phenylacetaldehyde Reductase. La couleur des noms des composés indique leur note aromatique potentielle, violet : florale ; rouge : épicée ; orange : fruitée.

Le 2-phényléthanol ou l'alcool de phényléthyle a été retrouvé dans le jus de raisin muscadine, le vin et les roses (Baek et al., 1997; Helsper et al., 1998; Genovese et al., 2007). Il est issu du même précurseur que celui de l'acétophénone et de l'acétate de benzyle mais ne possède pas la même voie de biosynthèse. La voie de biosynthèse du 2- phényléthanol a été notamment mise en évidence chez la rose (*Rosa* spp.) et *Enterobacter* sp (Liu et al., 2018; Roccia et al., 2019). Le géraniol ainsi que des composés phénylpropanoïdes tels que le 2-phényléthanol contribuent largement à l'odeur que l'on connait de la rose. La voie de biosynthèse du 2- phényléthanol a largement été étudiée et il a été mis en évidence que les enzymes clefs de cette voie sont la Phenylacetaldehyde Synthase (RhPAAS; Kaminaga et al., 2006; Sakai et al., 2007; Farhi et al., 2010) responsable de la transformation de la L-phénylalanine en 2-phenylacetaldéhyde et la Phenylacetaldehyde Reductase (PAR; Chen et al., 2011) responsable de la réduction du 2-phenylacetaldéhyde en 2- phényléthanol (figure 10). D'autres transformations peuvent avoir lieu ensuite comme le 2- phényléthanol qui est acétylé par l'acetyl-Coenzyme A (Geraniol/Citronellol Acetyl Transferase) en acétate de 2-phényléthyle (figure 10), connu pour avoir des notes florales (Guterman et al., 2006; Carvalho et al., 2017; Roccia et al., 2019).

6.2-Voie de biosynthèse des monoterpènes

Les terpènes sont des molécules couramment présentes chez les végétaux et peuvent être impliqués ou non dans des fonctions métaboliques essentielles (Lamarti et al., 1994). Les huiles essentielles (HE) sont souvent composées de monoterpènes et de sesquiterpènes. Elles sont synthétisées dans différents organes des plantes, dépendant fréquemment de leurs fonctions. Par exemple la synthèse d'HE dans les fleurs permet l'attraction des pollinisateurs et celle dans les feuilles permet la répulsion des herbivores.

La synthèse des terpènes s'effectue dans différents compartiments cellulaires. La synthèse des sesquiterpènes et des triterpènes s'effectue dans le cytoplasme alors que la synthèse des monoterpènes, diterpènes et tetraterpènes s'effectue dans les plastes (Bohlmann et al., 1998).

Le linalol a été très étudié et a été retrouvé dans beaucoup de matrices : les fleurs comme *Clarkia breweri*, *Syringa vulgaris* (le lilas), *Antirrhinum majus* (le muflier) ou la rose ; le thé vert jasmin chinois, le vin, *Actinidia arguta* (le kiwi), les huiles essentielles d'orange ou de pêche (Pichersky et al., 1995; Dudareva et al., 1996; Ito et al., 2002; Högnadóttir and Rouseff, 2003; Kreck et al., 2003; Oswald, 2006; Genovese et al., 2007; Nagegowda et al., 2008; Chen et al., 2010; Eduardo et al., 2013; Feng et al., 2014). Sa voie de biosynthèse est également très étudiée. Diverses études ont démontré la transformation du Geranyl Pyrophosphate (GPP) en linalol grâce à la Linalol Synthase (LIS) (figure 11), du linalol en 6,7-époxylinalol grâce à l'activité des Cytochromes P450, enfin du 6,7-époxylinalol soit en linalol oxyde pyranoïde soit en linalol oxyde furanoïde par l'action des Cyclases (Pichersky et al., 1994; Kreck et al., 2003; Chen et al., 2010).

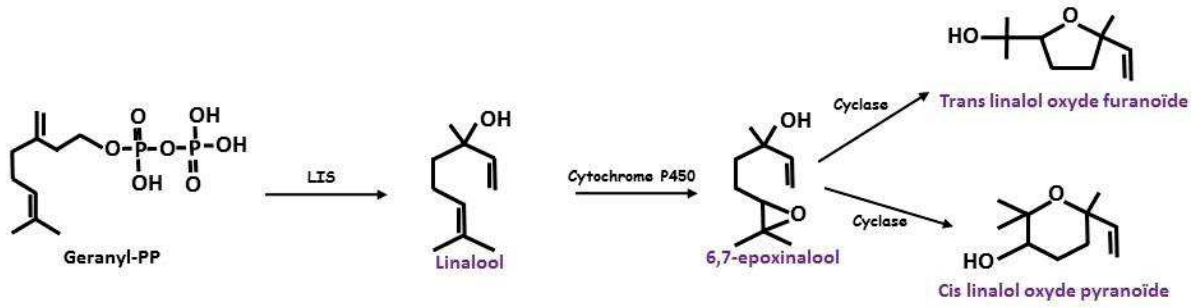


Figure 11: Schéma de la voie de biosynthèse du linalol et ses dérivés. Adapté de Chen et al, 2010.
Les composés dont le nom est écrit en violet son connus pour avoir une note florale.

7-Voie de biosynthèse des composés aromatiques aux notes fruitées

7.1-Synthèse des composés responsables des arômes fruits frais

Les arômes de fruits frais sont représentés par différentes familles de composés volatils comme les alcools, les aldéhydes, les esters et les cétones (Aprotosoaie et al., 2016). D'après Rodriguez-Campos et al. (2011) les alcools, les aldéhydes et les cétones sont les groupes de molécules majoritairement présents dans les fèves de cacaos crues et/ou au début de la fermentation. Les alcools peuvent être synthétisés à partir de sucres via la fermentation alcoolique. Ces alcools peuvent également être synthétisés à partir d'une molécule de pyruvate, présente dans tous les organismes (Sun et al., 2015). Cette transformation peut être faite par les levures lors de la fermentation des fèves. Les aldéhydes sont formés grâce à l'oxydation d'un alcool primaire. C'est le cas par exemple pour la fleur de Lilas (*Syringa vulgaris L.*) où les « alcools de lilas » sont oxydés pour donner les « aldéhydes de lilas » (Kreck et al., 2003). Les esters sont produits lors de la fermentation par les fèves de cacao et/ou par les levures lors de la phase anaérobique de fermentation (Peddie, 1990). Les esters sont le résultat de la réaction d'estérification, réaction chimique réversible. L'estérification permet la synthèse d'une molécule d'ester et d'une molécule d'eau à partir d'une molécule d'alcool et d'une molécule d'acide (Aranda et al., 2008). Les cétones peuvent être le résultat de l'oxydation d'un alcool secondaire.

7.2-Synthèse des composés responsables des arômes fruits secs

Les arômes de fruits secs des fèves de cacao sont principalement dus aux pyrazines et aux furannes qui apparaissent majoritairement après la torréfaction (Aprotosoaie et al., 2016). Lors de cette étape de transformation, le chauffage induit un grand nombre de réactions chimiques dont la réaction de Maillard. Cette réaction est à l'origine de la synthèse de pyrazines à partir d'acides aminées et de sucres réducteurs (Arnoldi et al., 1988). Toutefois, il est possible que la réaction de Maillard commence lors du séchage (Rodriguez-Campos et al., 2011).

Jinap et al., (2008) ont caractérisé les précurseurs d'arôme ainsi que les méthylpyrazines dans des fèves sous-fermentées. Ces précurseurs d'arôme se trouvent également dans les fèves bien fermentées (Kongor et al., 2016). Lors de cette étude, ils ont pu observer que les fèves sous-fermentées (1 à 3 jours de fermentation) possédaient des peptides hydrophiles et des acides aminées libres hydrophobes en quantité plus élevée que les fèves non-fermentées. Ces deux précurseurs de la réaction de Maillard sont produits grâce à la dégradation des protéines et des polypeptides par des carboxypeptidases endogènes pendant la fermentation du cacao.

Certains aldéhydes tels que le 2-methylbutanal ou le 2-phenylbut-2-enal, certaines cétones tel que le 3-methyl-2-cyclohexen-1-one et certains esters tel que l'acétate de 1-methylbutyle peuvent également contribuer aux notes de fruits secs (The good scent compagny, 2021).

8-Voie de biosynthèse des composés responsables de l'amertume et de l'astringence

L'astringence est une sensation orale déclenchant une sécheresse dans la bouche et le froncement des lèvres alors que l'amertume est une sensation gustative reconnue par des signaux nerveux sur la langue (Ma et al., 2014). Les caractéristiques de la structure chimique des proanthocyanidines influent sur l'intensité de les sensations d'astringence et d'amertume (Lesschaeve and Noble, 2005; Ma et al., 2014). Dans les produits cacaotés, il a été observé que la torréfaction permet de diminuer la capacité des polyphénols à interagir avec les protéines salivaires entrainant ainsi une diminution de l'astringence (Misnawi et al., 2005). Dans les boissons, d'autres paramètres influencent les sensations d'amertume et d'astringence tels que : le pH, la concentration en éthanol, le goût sucré et la viscosité (Lesschaeve and Noble, 2005).

Selon la définition de la société Américaine du goût et des matériaux (ASTM), l'astringence désigne l'ensemble des sensations dues au rétrécissement, à l'étirement ou au plissement de l'épithélium à la suite de l'exposition à des substances telles que les aluns ou les tanins. L'astringence peut être perçue à la suite de l'ingestion de certaines classes de composés comme les sels de cations métalliques multivalents, les agents déshydratants tels que l'éthanol et l'acétone, les acides minéraux et organiques, les tanins et les petits polyphénols. Dans le vin, l'intensité de l'astringence a montré une forte corrélation positive avec la concentration en tannins. En effet une réduction de la concentration en tanins est considérée comme la raison principale du déclin de la sensation d'astringence (Ma et al., 2014). Misnawi et al., (2003) ont réalisé des tests sur des fèves de cacao avec un degré de fermentation différent. Ils ont observé une diminution de la teneur en procyanidines (des monomères aux pentamères) dans les fèves de cacao simultanée à la diminution de la perception de l'astringence. Plus récemment, une accumulation de preuves a permis d'affirmer l'implication des anthocyanidines et du changement de degré moyen de leur polymérisation dans la perception de l'astringence (Ma et al., 2014).

Les composés amers sont présents dans toutes les plantes, initialement pour se défendre contre les insectes et les champignons qui les attaquent. Ces composés peuvent être des acides

gras, des peptides, des acides aminés, des amines, des azacycloalcanes, des composés N-hétérocycliques, des amides, des carbamides, des thiourées (organosulfure de formule $SC(NH_2)_2$), des urées, des esters et des phénols (Ma et al., 2014). Dans le cas des vins contenant une grande quantité de polyphénols, l'amertume est principalement déclenchée par les flavan-3-ol et ses polymères. Elle peut également être déclenchée par certains flavonols. La perception de l'amertume est une reconnaissance gustative médiée par des bourgeons gustatifs situés dans les papilles de la langue. Les cellules réceptrices du goût amer sont principalement situées à l'arrière de la langue près de la gorge (Ma et al., 2014). Soares et al., (2013), ont montré que l'épicatéchine, le dimère B3 de procyanidine et le trimère C2, tanins condensés, étaient responsables de la stimulation de certains récepteurs de l'amertume. La présence de polyphénols, de tanins et de (-)-épicatéchine sont des facteurs critiques pour définir la qualité du cacao utilisé dans la fabrication du chocolat (Serra Bonvehi and Ventura Coll, 1997).

8.1-Voie de biosynthèse de la caféine

La caféine et la théobromine sont également des composés faisant réagir les récepteurs de l'amertume. La biosynthèse de la caféine est initiée par la dégradation des nucléotides de purines. Il a été montré que la théobromine était le précurseur direct de la caféine dans les feuilles de café (Ashihara et al., 1996). Comme illustrée dans la figure 12, la synthèse de la caféine à partir de la théobromine s'effectue grâce à l'action de la théobromine 1-N-methyltransferase aussi appelée caféine synthase (Koshiishi et al., 2001; Zheng et al., 2004).

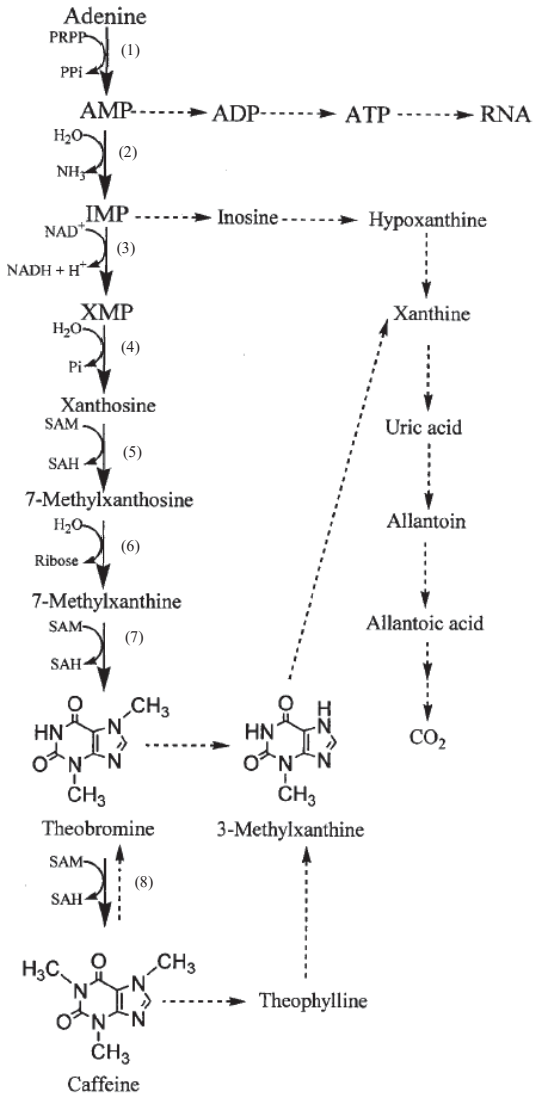


Figure 12: Schéma de biosynthèse de la caféine dans les cabosses de cacaoyers, proposé par Zheng et al., (2004)
Enzymes d'après Koshiishi et al., (2001): (1) adénine phosphoribosyltransférase ; (2) Adénosine monophosphate (AMP) désaminase ; (3) inosine 5P-monophosphate (IMP) déshydrogénase ; (4) 5P-nucléotidase ; (5) xanthosine 7-N-méthyltransférase ; (6) 7-méthylxanthosine nucléosidase ; (7) 7-méthylxanthine 3-N-méthyltransférase (caféine synthase) ; (8) théobromine 1-N-méthyltransférase (caféine synthase).

En 1975, Pickenhagen *et al.* identifient la théobromine comme la principale molécule responsable de l'amertume des fèves de cacao. Contrairement aux grains de café, les fèves de

cacao contiennent plus de théobromine que de caféine (Martínez-Pinilla et al., 2015), ce qui suggère que la caféine synthase est peu active chez le cacaoyer.

8.2-Voie de biosynthèse des polyphénols

Les fèves de cacao sont connues pour être riches en polyphénols tels que la (-)-épicatéchine, les procyanidines ou les proanthocyanidines. Les polyphénols présents dans les fèves de cacao sont stockés dans les cellules des cotylédons. En fonction de la quantité d'anthocyanes, la couleur de ces cellules varie de blanc à pourpre foncé (Wollgast and Anklam, 2000). Le degré de polymérisation des procyanidines présentes dans les fèves de cacao varie d'un monomère à des polymères à longues chaînes (Ioannone et al., 2015). Couramment on retrouve des dimères d'épicatéchine ((-)-épicatéchine-(4β→8)-(-)épicatéchine, procyanidine B2 ou (-)-épicatéchine-(4β→6)-(-)-épicatéchine, procyanidine B5) ou des trimères d'épicatéchine (épicatéchine-(4β→8)-épicatéchine-(4β→8)-épicatéchine , procyanidine C1).

Dans la pomme et dans les feuilles de thé, il a été montré que la (+)-catéchine, la (-)-épicatéchine, ainsi que les proanthocyanidines sont des dérivés de la dégradation du L-phénylalanine (Punyasiri et al., 2004; Henry-Kirk et al., 2012). Comme illustré dans la figure 13, la L-phénylalanine est dégradée en *p*-coumarate, lui-même transformé en chalcones ensuite transformées en leucoanthocyanidines et en anthocyanidines. Ces derniers vont respectivement être transformés en (+)-catéchine et (-)-épicatéchine qui eux-mêmes seront transformés en proanthocyanidines (Henry-Kirk et al., 2012).

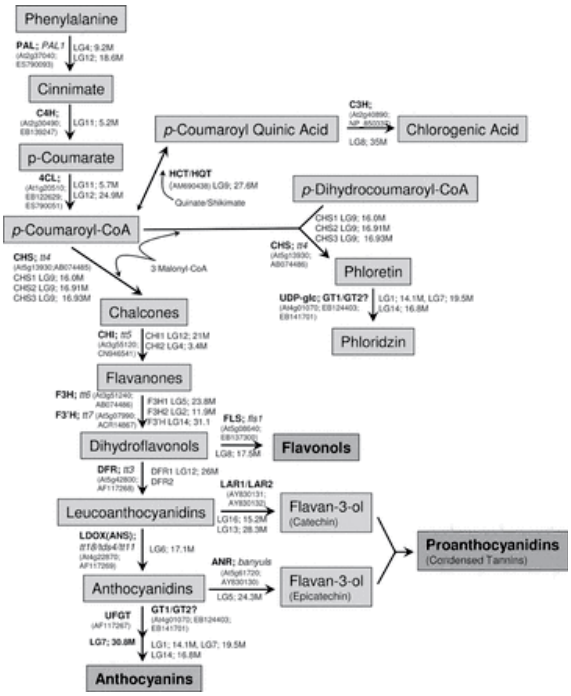


Figure 13: Schéma de la voie de biosynthèse des composés polyphénoliques chez la pomme (Henry-Kirk et al., 2012). PAL, phenylalanine ammonia lyase; C4H, cinnamate-4-hydroxymate; 4CL, 4-coumarate:coenzyme A ligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase, F3'H, flavanone 3'-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; UFGT, UDP-glucose flavonoid 3-O-glucosyl transferase; FLS, flavonol synthase; LAR1/2, leucoanthocyandin reductase; ANR, anthocyanidin reductase; GT1/2, glycosyltransferases; HQT/HCT, quinate hydroxycinnamoyl/hydrxycinnamoyl CoA shikimate; C3H, p-coumarate 3-hydroxylase

9- Méthodes d'analyse génétique des caractères d'intérêt agronomique

9.1- Les différents marqueurs génétiques

Les marqueurs génétiques sont des caractères pas ou peu influencés par l'environnement et qui permettent donc une observation directe du génotype des individus d'organismes vivants. Les lois de Mendel de 1865 (Mendel, 1865; van Dijk and Ellis, 2016) permettent d'interpréter la ségrégation des allèles de ces marqueurs lors de la transmission des parents aux descendants. Un marqueur génétique fiable doit avoir les qualités suivantes :

- Etre non influencé par l'environnement ;
- Etre monolocus, soit présent à un seul endroit du génome ;
- Etre polymorphe, donc représenté par au minimum deux allèles différents ;
- Etre co-dominant, c'est-à-dire que les deux allèles d'un locus doivent pouvoir être observés sans que l'un ne masque la présence de l'autre. Dans le cas d'une dominance allélique, il n'est pas possible de différencier les phénotypes des individus homozygotes portant l'allèle dominant des individus hétérozygotes ;
- Etre non épistatique, c'est-à-dire qu'il doit être indépendant d'autres marqueurs ou gènes (Gallais, 2013).

Il existe différents types de marqueurs moléculaires : les marqueurs morphologiques, les marqueurs biochimiques (isozymes) et les marqueurs moléculaires (éléments constitutifs de l'ADN). Les marqueurs moléculaires ont évolué au fur et à mesure des avancées technologiques. Les premiers sont basés sur la taille des fragments de restriction de la molécule d'ADN (RFLP), puis sur la taille des fragments amplifiés aléatoirement ou non (RAPD, AFLP). Ces marqueurs sont actuellement très peu, voire plus du tout utilisés. Á suivi la découverte des marqueurs microsatellites, très courts motifs de séquences d'ADN, souvent de deux à quatre bases et qui sont répétés un certain nombre de fois (Gallais, 2013). Le nombre de répétitions détermine les différents allèles. Dans le cas du cacaoyer, ces marqueurs sont régulièrement utilisés pour déterminer la structure génétique des accessions ainsi que leur appartenance aux dix groupes génétiques actuellement connus (Motamayor et al., 2008; Loor S. et al., 2015; Fouet et al, Unpublished data). Plus réce mment, grâce aux progrès réalisés dans les techniques de séquençage, des nouveaux marqueurs, les SNP, montrent un polymorphisme d'une seule paire de base (Gallais, 2013). Chez le cacaoyer, ils sont actuellement très utilisés dans les études de diversité génétique, de cartographie génétique, la détermination de QTL ou des études GWAS et ont presque totalement remplacé les marqueurs microsatellites dans ces types d'études.

9.2- Déterminisme génétique des caractères d'intérêt agronomique

9.2.1- Détection de QTL (Quantitative Trait Loci)

La détection de QTL (Quantitative Trait Loci) est une méthode permettant de mettre en lumière les zones du génome impliquées dans la variation génétique d'un caractère quantitatif (ou qualitatif). Cette méthode d'analyse génétique consiste en l'étude de la liaison génétique entre un marqueur moléculaire et le caractère d'intérêt. Pour réaliser cette méthode, il est essentiel de posséder la cartographie génétique du croisement utilisé (Gallais, 2013; Gallais, 2015).

9.2.2- Génétique d'association (GWAS, Genome - wide association study)

Le principe de la génétique d'association est basé sur la diversité naturelle d'une population et repose sur la même logique que la détection de QTL (liaison génétique entre un marqueur moléculaire et le caractère d'intérêt). Elle consiste à évaluer l'association (ou la liaison) entre chaque marqueur polymorphe (SSR ou SNP) et le phénotype d'intérêt dans un panel composé d'un grand nombre d'individus. Contrairement aux populations utilisées pour la détection de QTL, les individus des populations utilisées pour les études de génétique

d'association sont généralement éloignées génétiquement et obtenues après un grand nombre de recombinaisons à partir des ancêtres fondateurs, ce qui induit un faible déséquilibre de liaison marqueur/caractère qui lui-même permet une localisation plus précise des QTLs. De ce fait, cette technique nécessite un grand nombre de marqueurs afin de couvrir le plus de zones possible du génome, d'où l'intérêt d'utiliser des SNP.

La génétique d'association a permis notamment d'identifier le déterminisme génétique des terpénols aromatiques chez la vigne. Ces composés sont à l'origine des arômes floraux de la vigne (et d'autres espèces) avec laquelle le cacaoyer semble posséder des similitudes au niveau des caractères aromatiques, des étapes de fermentation similaires au cours de leur transformation, de la présence de polyphénols et de terpènes (Oswald, 2006). Récemment, la méthode de génétique d'association, couplée aux nouvelles technologies de séquençage, a également permis de mettre en évidence un grand nombre de régions génomiques associées à des traits quantitatifs (QTL) liés aux arômes chez diverses espèces comme par exemple la tomate, la pêche, la rose ou encore le café (Causse et al., 2002; Mathieu et al., 2009; Eduardo et al., 2013; Sauvage et al., 2014; Magnard et al., 2015; Sant'Ana et al., 2018).

10- Déterminisme génétique des caractères d'intérêt agronomique chez le cacaoyer

Chez le cacaoyer, des études de détection de QTL ainsi que de GWAS ont été menées et ont permis de mettre en lumière un grand nombre de QTL associés à divers caractères phénotypiques tels que des caractères associés au rendement, à la résistance aux maladies ou encore aux critères de qualité des fèves.

10.1- Physiologie de la plante

Chez le cacaoyer, des études GWAS ont été menées et ont permis de mettre en lumière des régions génomiques associées au nombre de cabosses par arbre, à la taille des fèves ou encore aux caractères de pigmentation (Lanaud et al., 2003; Marcano et al., 2009). Même si ces différents caractères (le nombre de cabosses par arbre, la taille des fèves et les caractères de pigmentation) sont influencés par l'environnement, l'identification de régions génomiques associées à ces différents caractères montrent qu'ils sont également influencés par la génétique.

10.2- Auto-Incompatibilité

Plusieurs études génétiques ont porté sur le déterminisme génétique de l'auto-incompatibilité du cacaoyer. Royaert et al., (2011) ont mis en évidence des marqueurs associés aux traits d'auto-incompatibilité dans une population de ségrégation de *Theobroma cacao L*.

Deux loci putativement impliqués dans l'auto-incompatibilité observée par le % de nouaison, ont été localisés par analyse QTL : l'un était localisé au sommet du chromosome 4(Crouzillat et al., 1996; Yamada et al., 2010; Royaert et al., 2011), l'autre dans le chromosome 7 (Yamada et al., 2010).

Plus récemment, deux locus ont été identifiés, sur le chromosome 1 et sur le chromosome 4, comme impliqués dans l'auto-incompatibilité du cacaoyer par deux processus différents. Les deux locus sont responsables de la sélection gamétique. Seul le locus sur le chromosome 4 est impliqué dans la chute principale des fruits (Lanaud et al., 2017).

10.3- Résistance aux maladies

Un grand nombre d'analyses génétiques ont été menées dans le but d'étudier le déterminisme génétique de la résistance de *Theobroma cacao L.* aux diverses maladies dont il est victime. Des analyses de détection de QTL et de génétique d'association ont été menées pour étudier la résistance de manière générale (Motilal et al., 2016), aux espèces de *Phytophtora* (Flament et al., 2001; Risterucci et al., 2003; Efombagn et al., 2011; Akaza et al., 2016), à la maladie du balais de sorcière (Almeida et al., 2017) et à la moniliose (Romero Navarro et al., 2017; Gutiérrez et al., 2021). Une méta-analyse regroupant l'ensemble des études précédemment faites a également été réalisée (Lanaud et al., 2009). Lors de cette étude, des QTL consensus ont pu être identifiés.

10.4- Critères de qualités.

Peu d'analyses génétiques ont été faites sur les critères de qualité du cacaoyer. Quelques études ont tout de même été réalisées sur les déterminismes génétiques des critères de qualité des fèves de cacao, comme la teneur en matière grasse, les notes sensorielles ou encore la taille des fèves (Lanaud et al., 2003; Lanaud et al., 2012). Une autre étude réalisée par Mustiga et al., (2019) portant sur la composition en acides gras et la teneur en matière grasse a permis de mettre en lumière plusieurs QTL dont un majeur sur le chromosome 4. Dans cette même étude et indépendemment des analyses QTL, ils ont pu montrer l'impact des facteurs climatiques sur la composition en acides gras.

Chapitre 2: Deux principales voies de biosynthèse impliquées dans la synthèse de l'arôme floral de la variété de cacao Nacional

Chapitre 2: Deux principales voies de biosynthèse impliquées dans la synthèse de l'arôme floral de la variété de cacao Nacional

Two main biosynthesis pathways involved in the synthesis of the floral aroma of the Nacional cocoa variety

Kelly Colonges[1,2,3,4*], Juan-Carlos Jimenez[5], Alejandra Saltos[5], Edward Seguine[6], Rey Gastón Loor Solorzano[5], Olivier Fouet[1,2], Xavier Argout[1,2], Sophie Assemat[3,4], Fabrice Davrieux[3,4], Emile Cros[3,4], Renaud Boulanger[3,4*], Claire Lanaud[1,2*]

1 Cirad, UMR AGAP, F-34398 Montpellier, France.

2 AGAP Institut, Univ Montpellier, Cirad, INRAE, Institut Agro, Montpellier, France.

3 Cirad, UMR Qualisud, F-34398 Montpellier, France.

4 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

5 Instituto Nacional de Investigacion Agropecurias, INIAP, Ecuador.

6 Guittard, United-States

*** Correspondence:**

Corresponding author

Kelly.colonges@cirad.fr (KC)

Renaud.boulanger@cirad.fr (RB)

Claire.lanaud@cirad.fr (CL)

Keywords: GWAS, Cocoa aroma, floral, monoterpenes, phenolic compounds

1-Abstract

T. cacao is the only source that allows the production of chocolate. It is of major economic importance for producing countries such as Ecuador, which is the third-largest cocoa producer in the world. Cocoa is classified into two groups: bulk cocoa and aromatic fine flavour cocoa. In contrast to bulk cocoa, fine flavour cocoa is characterized by fruity and floral notes. One of the characteristics of Nacional cocoa, the emblematic cocoa of Ecuador, is its aromatic ARRIBA flavour. This aroma is mainly composed of floral notes whose genetic and biochemical origin is not well known. This research objective is to study the genetic and biochemical determinism of the floral aroma of modern Nacional cocoa variety from Ecuador. Genome-Wide Association Study (GWAS) was conducted on a population of 152 genotypes of cocoa trees belonging to the population variety of modern Nacional. GWAS was conducted by combining SSR and SNP genotyping, assaying biochemical compounds (in roasted and unroasted beans), and sensory evaluations from various tastings. This analysis highlighted different areas of association for all types of traits. In a second step, a search for candidate genes in these association zones was undertaken, which made it possible to find genes potentially involved in the biosynthesis pathway of the biochemical compound identified in associations. Our results show that two biosynthesis pathways seem to be mainly related to the floral note of Nacional cocoa: the monoterpene biosynthesis pathway and the L-phenylalanine degradation pathway. As already suggested, the genetic background would therefore appear as largely explaining the floral note of cocoa.

2-Introduction

Theobroma cacao L. is native to the tropical rainforests of northern South America and is a member of the family *Malvaceae* (Bayer and Kubitzki, 2003). The cocoa tree is a diploid (2n = 20) with a small genome that is now sequenced and of which 96.7 % of the assembly is anchored on all 10 chromosomes (Argout et al., 2011; Motamayor et al., 2013; Argout et al., 2017).

Cocoa farming represents an important economic issue for many tropical countries because it is the only source of chocolate supply. In 2018/2019, cocoa production represented more than 4 780 thousand tonnes worldwide. The three largest producers are Ivory Coast, Ghana, and Ecuador with respectively 1964, 905 and 287 thousand tons produced (ICCO, 2020). Even if Africa remains the leading producer, America maintains its reputation thanks to the aromatic quality of its cocoa. Cocoa is classified into two types of products: bulk cocoa and

fine flavour cocoa. Fine flavour cocoa is characterized by fruity and floral notes unlike bulk cocoa (Sukha et al., 2008). Bulk cocoa accounts for around 95 % of world production compared to 5 % for fine flavour cocoa. *Theobroma* cacao L. is highly diverse and has been classified into ten genetic groups: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañón, Nanay, Nacional, and Purùs (Motamayor et al., 2008).

Nowadays, three varieties are mainly capable to produce fine flavour cocoa: Criollo, Nacional, and Trinitario (hybrids between Criollo and Amelonado). Criollo is not widely cultivated because of its high susceptibility to diseases and low vigour (Cheesman, 1944). Nacional is native to Ecuador and is well known for its Arriba floral flavour. It is for this reason that it is sought after by chocolate makers. It is characterized by floral and woody notes (Luna et al., 2002). Also, Nacional is known for its low astringency and bitterness (International Cocoa Organization, 2017). The first hypothesis explaining floral notes of Arriba flavour was suggested by Ziegleder, (1990) who observed that linalool, a volatile compound belonging to monoterpenes, was observed in higher concentration in Nacional cocoa.

Overall, fine flavours are often produced during the fermentation process (Rodriguez-Campos et al., 2011). The cocoa fermentation takes place in two stages: first, the alcoholic fermentation made by yeast thanks to the presence of sugar in the cocoa pulp, then, there is an acetic fermentation carried out by bacteria (Ho et al., 2014). Fermentations produce aroma precursors but also volatile organic compounds (VOC). An adaptation of fermentation conditions is required to improve cocoa beans fine flavour. Fermentation time has an important effect on the concentration of different volatile compounds, as for some alcohol concentrations, which decreases from 2 to 8 days of fermentation (Rodriguez-Campos et al., 2012; Hamdouche et al., 2019). The drying process occurs after fermentation, which allows stopping it. This step is very important for cocoa bean conservation. It allows moisture decrease from 80 % to under 8 % (Cros and Jeanjean, 1995; Afoakwa et al., 2008). The artificial drying temperature can also influence the aromatic fraction with a decrease in isobutyric acid and an increase in tri and tetramethylpyrazine at lower drying temperature (70 °C versus 80 °C) (Rodriguez-Campos et al., 2012).

Cocoa beans have been studied to understand how their specific flavour is synthesized. A study on unfermented dry cocoa beans showed that terpenes are already present and important for fruity and floral aromas, even without fermentation (Qin et al., 2017). Other scientists have also proven the importance of terpenes such as linalool or epoxylinalool in cocoa fine flavour

after fermentation (Kadow et al., 2013; Cevallos-Cevallos et al., 2018). Kadow et al., (2013) demonstrated that the aroma specificity depends on the presence of volatile compounds and can be different depending on the genotype. The most important volatile compounds for the floral aroma of cocoa have been identified: they include terpenes mainly linalool, 2-phenylethanol (or phenylethyl alcohol), 2-phenylethyl acetate, and acetophenone (Ziegleder, 1990; Afoakwa et al., 2008; Ziegleder, 2009; Kadow et al., 2013; Cevallos-Cevallos et al., 2018; Utrilla-Vázquez et al., 2020). Rottiers et al., (2019) also compared the compounds contained in cocoa beans from the modern Nacional (EET varieties) and a standard cocoa variety CCN51. They were able to identify 14 compounds known to have a floral taste by GC-MS. Only five of them were found during the analysis with an electronic nose: 2-phenylacetaldehyde, 2-phenylethyl acetate, 2-phenylethanol, acetophenone, and linalool. However, other volatile compounds could be responsible for floral aroma (Schwab et al., 2008).

The biosynthesis pathway of aromatic compounds has been studied. Linalool is a volatile floral compound present in various flowers as *Clarkia breweri*, rose, Chinese jasmine green tea and wine (Dudareva et al., 1996; Ito et al., 2002; Genovese et al., 2007; Feng et al., 2014). Its biosynthesis pathway is very well studied. Pichersky et al., (1994) highlighted the linalool biosynthesis pathway in *C. breweri* flowers. They observed the transformation of geranyl pyrophosphate (GPP) to linalool by linalool synthase (LIS). Subsequently, the linalool was transformed into 6,7-epoxylinalool. The 6,7-epoxylialool was then converted to pyranoid linalool oxide or furanoid linalool oxide. Other studies showed that cytochrome P450 is responsible for the transformation of linalool to 6,7-epoxylinalool and cyclases for the transformation of 6,7-epoxylinalool to pyranoid linalool oxide or furanoid linalool oxide (Kreck et al., 2003; Meesters et al., 2007; Chen et al., 2010).

2-phenylethanol (or phenylethyl alcohol) has been found in muscadine grape juice, wine, and roses (Baek et al., 1997; Helsper et al., 1998; Genovese et al., 2007). 2-phenetylethanol and 2-phenylethyl acetate were observed in the same biosynthesis pathway in roses (Roccia et al., 2019). L-phenylalanine is converted to 2-phenylacetaldehyde by phenyl acetaldehyde synthase (PAAS). Subsequently, 2-phenylacetaldehyde is reduced to 2-phenylethanol by phenyl acetaldehyde reductase (PAR). Next, 2-phenylethanol is acetylated to 2-phenylethyl acetate by acetyl-coenzyme a: geraniol/citronellol acetyl transferase (AAT) (Roccia et al., 2019).

Acetophenone has been found in muscadine grape juice and Camellia (Baek et al., 1997; Dong et al., 2012). It has the same precursor as 2-phenylethanol but has a parallel biosynthesis pathway identified in the fungus *Bjerkandera adusta*. The transformation of L-phenylalanine to 2-phenylethanol is due to the non-oxidative degradation pathway of L-phenylalanine, while L-phenylalanine transformation to acetophenone belongs to β–oxidation pathway (Lapadatescu et al., 2000). In Camellia, the acetophenone biosynthesis pathway has been characterized (Dong et al., 2012). First, L-phenylalanine (L-phe) is converted to cinnamic acid (CA). Next, CA is transformed into 3-hydroxy-3-phenylpropionic acid (HPPA). HPPA is converted to 3-phenylpropionic acid (PPA) and PPA is transformed into acetophenone. The enzymes involved in these reactions have not yet been identified.

Few studies were carried out on the genetic determinants of cocoa qualities. The first were based on QTL analyses of some sensory traits and fat content (Lanaud et al., 2003) and also showed hotspots of volatile compounds co-located on the genome (Lanaud et al., 2012).

This study aims to contribute to the deciphering of the genetic and biochemical determinism of Nacional cocoa floral notes. To this end, we conducted a genome-wide association study (GWAS) on a modern cultivated Nacional population, composed of trees resulting from hybridizations between three contrasting main ancestors: Criollo, Amelonado, and the ancestral Nacional variety. This population was characterized by volatile compounds and sensory analyses and presented a high degree of variability. Thanks to the availability of the genome sequence and high-density SNP genotyping, candidate genes involved in key traits could be proposed.

3-Materials and Methods

3.1-Vegetal material

The plant material used for these experiments was composed of a collection of 152 cocoa trees from Ecuador conserved in the Pichilingue experimental station of the “Instituto Nacional de Investigaciones Agropecurias” (INIAP) and the “Colecion de Cacao de Aroma Tenguel” (CCAT) of Tenguel. This population represents the Nacional variety currently grown in Ecuador and has been described by Loor S., (2007).

3.2-Fermentation processes

Micro-fermentations of cocoa beans were carried out in a wooden box in the most homogeneous way possible with a homogeneous cocoa mass. The process lasted 4 days with

two turns at 24 and 72 hours after the beginning of the fermentation. Each clone sample (152) was placed in a protective laundry bag and micro-fermented in a cocoa mass. After fermentation, the samples were put in a dry place. They were considered dried when their moisture content was below 8 %.

3.3-Sensorial Analysis

146 individuals were characterized by sensory analyses based on blind tastings carried out on 3 repetitions per sample. The tastings were carried out on cocoa liquor. The cocoa liquor corresponds to merchant cocoa (dried fermented beans) which have been roasted and crushed. Sixteen floral notes were judged with a score ranging from zero (no floral notes detected) to ten. We used the average of the three replicates for the phenotype of the GWAS analysis (ISCQF, 2020).

Mr. Edward Seguine, whose work consists of conducting sensory analyses of chocolate samples (see-attached documents), managed this study. This study does not require the approval of an ethics committee.

3.4-Volatile compound analysis by GC-MS

3.4.1-Preparation of cocoa samples

The analysis of volatile compounds was carried out on dried fermented beans and roasted beans. For each sample, 50 g of beans were taken. The beans were deshelled and crushed to obtain nibs. Then, nibs were put in liquid nitrogen and ground with a blender (SEB, France), to obtain cocoa powder, which was stored at -80 °C until analysis. In a 10 mL vial, 2.85 g of powder, 1 mL of standard internal solution (butan-1-ol at a concentration of about 600 µg/ml) and 2 mL of distilled water were added.

3.4.2-Compounds extraction

The volatile compounds of cocoa samples were extracted using the technique of solid-phase microextraction in the headspace (SPME-HS) using a 50/30-µm divinylbenzene/carboxene/polydimethylsiloxane (DVB/CAR/PDMS) fiber provided by Supelco to extract volatiles. The fiber was previously conditioned at 250 °C for 3 min and then exposed to the sample headspace at 50 °C for 45 min. Extracted aroma volatile compounds were analysed using an Agilent 6890 N gas chromatography–mass spectrometer (GC–MS) equipped with a Hewlett Packard capillary column DBWAX, 30 m length × 0.25 mm internal diameter × 0.25 µm film thickness (Palo Alto, CA, USA). The GC oven temperature was initially set at 40 °C for 5 min, increased to 140 °C at a rate of 2 °C/min and then increased at

a rate of 10 °C/min to 250 °C for 66 min. The carrier gas was high-purity helium at 1 ml min^{-1}. Injection mode was split less at 250 °C for 2 min. The selective mass detector was a quadrupole (Hewlett Packard, Model 5973), with an electronic impact ionization system at 70 eV and at 230 °C.

3.4.3-Compounds identification

The identification was done by comparing the mass spectra with the commercial NIST Wiley 275L database. No deconvolution was applied. Co-eluted volatile compounds were excluded from this study.

3.5-DNA extraction protocol

DNA extraction was conducted according to Risterucci et al., (2000) protocol.

3.6-Genotyping by SSR

This population was genotyped using SSR markers by Loor S., (2007). SSR loci were scored individually and alleles were recorded by the presence of polymorphic DNA fragments (alleles) among the individuals of each population. Only those alleles that showed consistent amplification were used in the analysis of results and smeared or weak bands were ignored.

3.7-Genotyping by sequencing

DNA samples were genotyped by sequencing (GBS) using DArTseq (Diversity Arrays Technology Sequencing) technology (Kilian et al., 2012). This method is based on enzymatic restriction of coding regions of the genome by the restriction enzymes: Pst1 and Mse1. The restriction generated many short fragments, with each locus represented more than ten times. Then, illumina Hiseq2000 machine sequenced all the fragments and the result was analysed. Reads were aligned with the V2 sequence of the Criollo genome (Argout et al., 2017). Reads that have more than one location were discarded. Markers with unknown locations were discarded for analysis.

3.8-Population's structure analysis

The phylogenetic tree was generated using DARwin software (Perrier and Jacquemoud-Collet, 2006). The genetic distances were calculated using the Dice coefficient and the Neighbour-Joining method (Dice, 1945; Saitou and Nei, 1987).

3.9-Association mapping

The graphic representation of the markers along the ten chromosomes was made with the R package "CMplot"(Yin, 2020). Several analyses of associations with SNP or SSR markers have been performed:

3.9.1-SNP GWAS

First, we performed a GWAS analysis with SNP markers associated with biochemical (146 accessions x 5195 markers) and sensory (144 accessions x 5195 markers) traits using TASSEL v5.

For all the traits, we used a mixed model (MLM) on the one hand. The MLM was carried out with a structure matrix, determined by running a PCA (principal component analyses integrated with TASSEL v5 software), considered as a fixed effect, and also with a kinship matrix considered as a random effect as covariates to control the false-positive rate. The option of not compressing and re-evaluating the components of variance for each marker was chosen. The kinship matrix using the identity by state (IBS) pairwise method proposed by Tassel v5 was established.

On the other hand, we used a fixed-effect model (GLM) with a structure matrix, determined by running a PCA. The option of 500 permutations was chosen.

For both methods, quantile-quantile plots were used to graphically evaluate the false-positive numbers observed in the selected model, based on deviations from the uniform law. The threshold was determined using the Bonferonni correction formula as proposed by Gao et al., (2008) with the effective number of independent tests (Meff) used as the denominator and calculated by SimpleM R package (Gao et al., 2010). Meff was 2796, which corresponds to a P-value of approximately $1.79e^{-05}$. The significance of all markers was plotted using Manhattan plots with the R QQman package.

3.9.2-SSR GWAS

We performed an analysis with SSR markers associated with biochemical (180 accessions x 180 markers) and sensory (197 accessions x 180 markers) traits using TASSEL v3. We used a fixed-effect model (GLM) with a structure matrix; the option of 500 permutations was chosen. The threshold was determined using the Bonferonni correction corresponding to a p-value about $2.78e^{-04}$.

The borders of the association zones were calculated using Haploview (Barrett et al., 2005). The haplotypic blocks were calculated with SNP data using Haploview with the association test, Family trio data, Standard TDT, and ignore pairwise comparison of markers above to 10 000 kb calculation parameters. The haplotypic block information was used to determine the confidence intervals of association areas.

The physical maps with the QTL representation were created using SpiderMap v1.7.1 software (Rami, 2007 unpublished). The size of the dots is correlated to the R2.

The identification of candidate genes was performed using the *Theobroma cacao* genome sequence (Argout et al., 2017).

3.10-Statistical analysis

PCA analysis and visualization were made with the "mixOmics" R package. Calculation of correlation was made with "agricolae" R package and visualization of correlation matrix with "corrplot" R package.

4-Results

4.1-Genetic diversity and population structure

The population studied represents the modern population of the Nacional variety cultivated in Ecuador. It is the result of various crosses between three main ancestors: the Criollo, the Amelonado, and the ancient Nacional varieties (Loor S., 2007). Using SNP markers, the structure of the genetic diversity of the population was studied. There was a continuous distribution of population trees between the three ancestors (Criollo, Nacional, and Amelonado varieties) as shown in Figure 14. Loor S., (2007) had also shown this distribution using microsatellite markers.

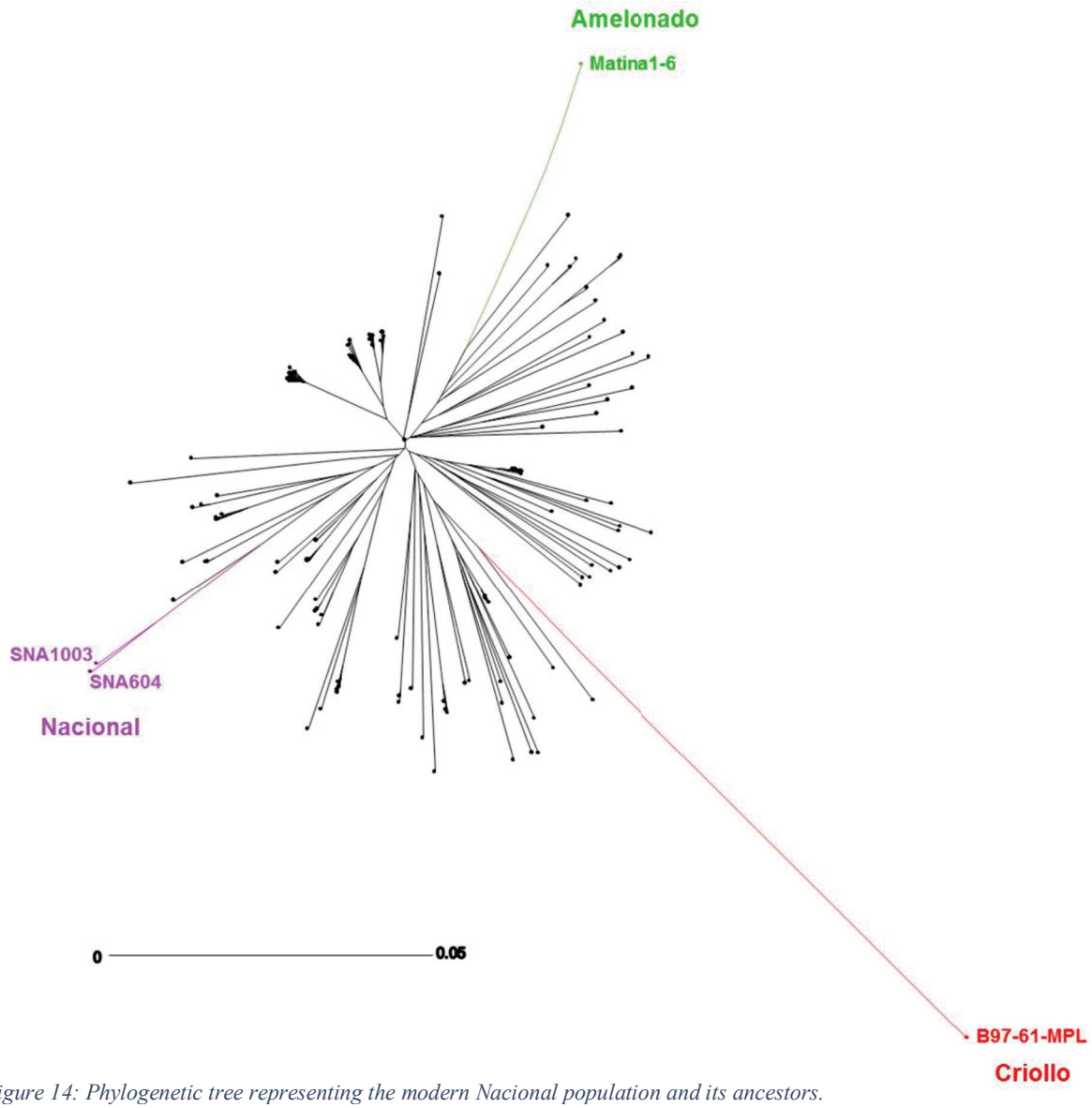


Figure 14: Phylogenetic tree representing the modern Nacional population and its ancestors.
Phylogenetic tree of the individuals of the studied population made with 4130 SNPs and including the ancestor controls of the population: in red, the Criollo variety (B97-61-B2); in purple, the Nacional variety (SNA604, SNA1003); in green, the Amelonado variety (Matina 1-6); in black, the individuals of the studied population. The graph's scale represents the edge lengths which are proportional to the genetic distance.

4.2-Characterization of the studied traits

To identify the areas of *T. cacao* genome involved in the synthesis of typical Nacional floral aromas, a genome-wide association study (GWAS) was conducted with two types of traits: the volatile compounds present in cocoa beans (before and after roasting) and sensory analysis data.

4.2.1-Sensorial traits analysis

Sixteen floral notes were determined by sensory analyses performed on cocoa liquor. A total of 16 sensorial traits were therefore used for this study (Appendix 1).

Principal component analysis (PCA) for sensory traits showed continuous variation in the population (Appendix 2). Axis 1 is mainly defined by the aromatic notes: browned flavour, floral bark woody and smoky. Axis 2 is mainly defined by the aromatic notes: floral tobacco, fruity acidity, and astringency. Correlation analyses between sensory traits showed strong positive and negative correlations (Figure 15A). These strong correlations suggest either that the correlated sensory notes are produced by the same compounds or that an interaction exists between the perceptions of the two sensory traits.

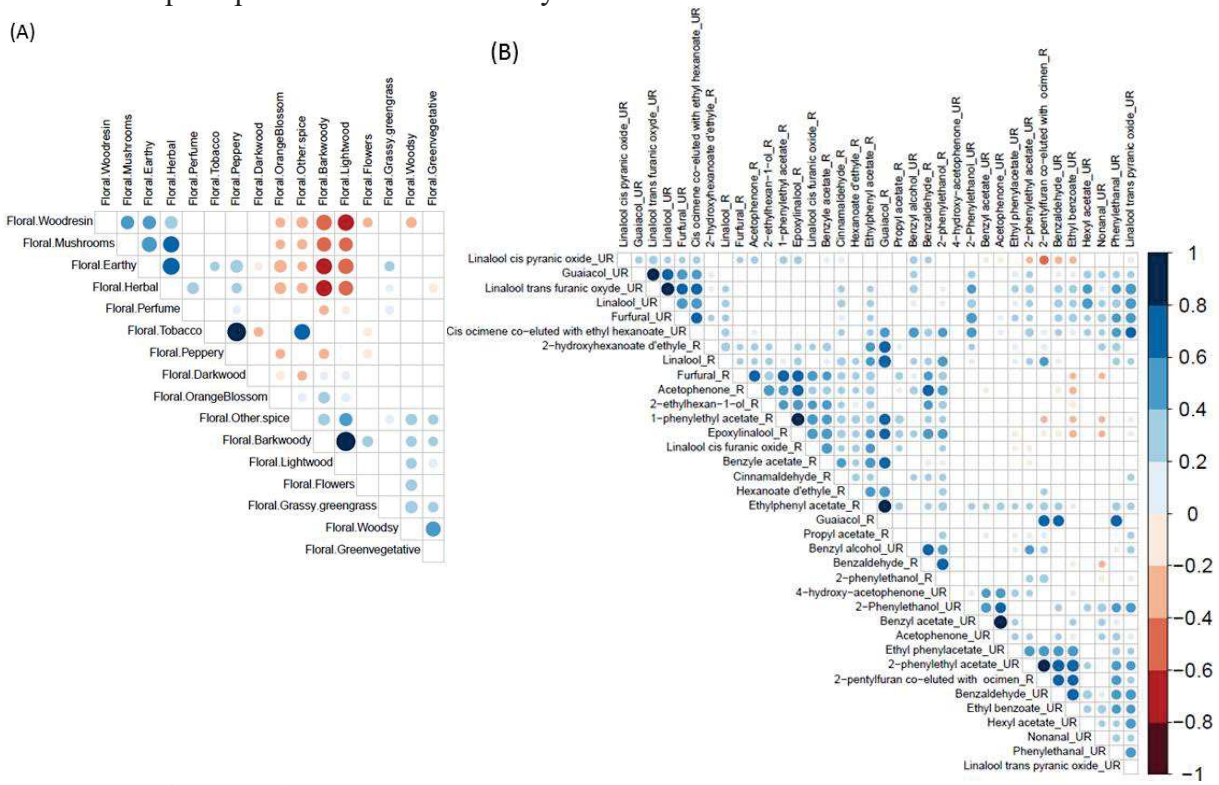


Figure 15: Significant correlation matrix

(A)Correlation matrix between the sensorial profiles determined in cocoa liquor. (B) Correlation matrix between the biochemical compounds measured in unroasted (UR) and roasted beans (R). The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient. The areas of circles correspond to a p-value of correlation coefficients. The p-value threshold for a significant correlation is 0.05. The different shades of blue represent a positive correlation coefficient while the different shades of red represent a negative correlation coefficient. The intensity of the colour depends on the strength of the R2 correlation coefficient. The scale on the right indicates the interpretations of different colours.

4.2.2-Analysis of aroma volatile compounds

The biochemical characterization was done on unroasted and roasted beans. Among one hundred and sixty volatile compounds (VOC) identified, twenty-seven volatile compounds are known to have a floral taste or are involved in biosynthetic pathways of known floral compounds (Table 1). Eighteen of them were detected in unroasted beans and seventeen in roasted beans such as linalool, acetophenone, or 2-phenylthanol. These VOC were used to conduct a GWAS analysis (Table 1).

Table 1: List of biochemical related to floral traits used for the GWAS analysis of unroasted (UR) and roasted (R) beans.

UR	R	Biochemical traits	Compound family	Aroma
	X	1-phenylethyl acetate	Ester	Fruity (Garg et al., 2018)
	X	2-ethylhexan-1-ol	Alcohol	Leafy, rose (Garg et al., 2018)
X	X	2-phenylethanol	Alcohol	Rose, honey (Jezussek et al., 2002; Genovese et al., 2007)
X		2-phenylethyl acetate	Ester	Floral, underwood, rose (Guichard et al., 2003; Genovese et al., 2007; Wang et al., 2014)
X		4-hydroxy-acetophenone	Ketone	
X	X	Acetophenone	Ketone	Acacia honey, floral and fruity (Genovese et al., 2007; Wang et al., 2014)
X	X	Benzaldehyde	Aldehyde	Bitter, cherry, almond, fruity (Perestrelo et al., 2006; Pham et al., 2008; Wang et al., 2014)
X	X	Benzyl acetate	Ester	Floral, Jasmin (Ito et al., 2002)
X		Benzyl alcohol	Alcohol	Fruity (Ito et al., 2002)
	X	Cinnamaldehyde	Aldehyde	Spicy, cinnamon (Garg et al., 2018)
	X	Epoxylinalool	Terpene	Floral (Arn and Acree, 1998)
	X	Ethyl 2-hydroxyhexanoate	Ester	Floral (Wang et al., 2014)
X		Ethyl benzoate	Ester	Fruity, violet, candy (Ferreira et al., 1997)
	X	Ethyl dodecanoate	Ester	Floral, fruity, leafy (Garg et al., 2018)
	X	Ethyl hexanoate	Ester	Fruity (strawberry, green apple) and floral (Larsen and Poll, 1992; Ferreira et al., 1997; Genovese et al., 2007; Wang et al., 2014)
X	X	Ethylphenyl acetate	Ester	Rose, floral (Perestrelo et al., 2006)
X	X	Furfural	Furan	Incense, fruity, floral, toasted, sweet and almond (Ferreira et al., 1997; Colahan-Sederstrom and Peterson, 2005; Wang et al., 2014)
X	X	Guaiacol	Aromatic hydrocarbon	Phenolic, floral, smoky, sweet, medicament (Ferreira et al., 1997; Arn and Acree, 1998; Genovese et al., 2007)
X		Hexyl acetate	Ester	Fruity (pear) and floral(Guichard et al., 2003; Wang et al., 2014)
X	X	Linalool	Terpene	Floral, citrus peel, orange flower (Ferreira et al., 1997; Genovese et al., 2007)
	X	Linalool cis furanic oxide	Terpene	Floral, woody (Arn and Acree, 1998)
X		Linalool cis pyranic oxide	Terpene	Fruity, citrus, green (Arn and Acree, 1998; Ito et al., 2002)
X		Linalool trans furanic oxide	Terpene	Citrus, leafy, floral (Arn and Acree, 1998; Ito et al., 2002)
X		Linalool trans pyranic oxide	Terpene	
X		Nonanal	Aldehyde	Orange-like, floral, soapy (Kumazawa and Masuda, 2002; Mahajan et al., 2004; Karagül-Yüceer et al., 2006)
X		Phenylethanal	Aldehyde	Floral, rose, honey (Perestrelo et al., 2006)
	X	Propyl acetate	Ester	Celery, floral, pear, red fruit (Garg et al., 2018)

UR: unroasted beans; R: roasted beans; *: biochemical compounds known for floral notes.

PCA of aroma volatile compounds was made (Appendix 3 and Appendix 4). Axis 1 of the PCA from analyses of biochemical compounds in unroasted beans is mainly defined by the linalool trans furanic oxide, meso-2,3-butan-di-yl diacetate, and linalool trans pyranic oxide. Axis 2 is mainly defined by ethyl acetate, ethyl-(2-methyl)-propionate, and benzaldehyde. Axis 1 of the PCA from analyses of biochemical compounds in roasted beans is mainly defined by epoxylinalool, 2-acetylpyrrole, and ethylphenyl acetate. Axis 2 is mainly defined by pentan-2-ol, pentan-2-one, and 1.2.5-trimethylbenzene. As with sensory traits, PCA of aroma volatile compounds showed that the distribution of traits showed a continuous variation within the

population which can be explained by the great genetic diversity present in this group of individuals deriving from several generations of crosses.

Correlation analyses between the different traits showed positive correlations between several biochemical compounds in roasted and unroasted beans (Figure 15B). The highest correlations (greater than 0.8) were observed in unroasted beans: between benzyl acetate and acetophenone; between 2-phenylethyl acetate and 2-pentylfuran co-eluted with ocimene; between guaiacol and trans furanic oxide linalool; between trans furanic oxide linalool and linalool. High correlations were also observed in roasted beans: between 1-phenylethyl acetate and epoxylinalool; between ethylphenyl acetate and guaiacol. A negative correlation between -0.4 and -0.6 was observed between linalool cis pyranic oxide and 2-pentylfuran co-eluted with ocimene in unroasted beans.

These various correlations between compounds can be partly explained by the fact that they belong to the same biosynthesis pathway. This is the case for the different terpenes which are strongly correlated or compounds resulting from the degradation of L-phenylalanine (acetophenone, 2-phenylethanol and benzaldehyde). On the other hand, no strong correlation between biochemical and sensory traits was detected (Appendix 5).

4.3-Genome-Wide Association Study

The linkage disequilibrium observed in this population amounts to 15 cM (Loor S., 2007). GWAS analyses were performed by different methods (GLM and MLM) and with different types of markers (SSR and SNP).

4.3.1-Marker Sorting

To limit the biases due to rare alleles, sorting by the frequencies of the minor alleles (MAF) was done at 5 % (MAF5). The population being very heterozygous, the sorting by MAF allowed to eliminate the alleles with a total frequency lower than 5 % but left homozygous genotypes very poorly represented (1 individual per class). The hypothesis was that the low representation of genotypic classes could induce a bias in the analyses, in the same way as a minor allele. It was therefore undertaken to do a further sorting of markers by discarding markers for which genotype classes had less than 5 % representation of the total population (Minor genotype frequencies, MGF). We conserved markers that had at least 7 individuals per genotype class (G7). Several tests were performed such as the comparison of Q-Q plots or the comparison of p-values (Zhang et al., 2019) to determine which of the two sorting methods had the least bias (Appendix 6). None of the tests could determine which of the two was the most

biased. The results differed in some respects, so both marker-sorting methods were retained for the GWAS studies.

4.3.2-SNP marker distribution

For the GWAS, SNPs were selected without missing data and with a genotype frequency above 5% or a minor allele frequency above 5%. The final data set consisted of 5195 SNP markers for the G7 data set and 6541 SNP markers for the MAF5 data set (Ruiz et al., 2017). The SNP markers are well spread over all ten chromosomes of T. cocoa. However, a decrease in marker density is observed in the centromeric and peri-centromeric areas (Figure 16).

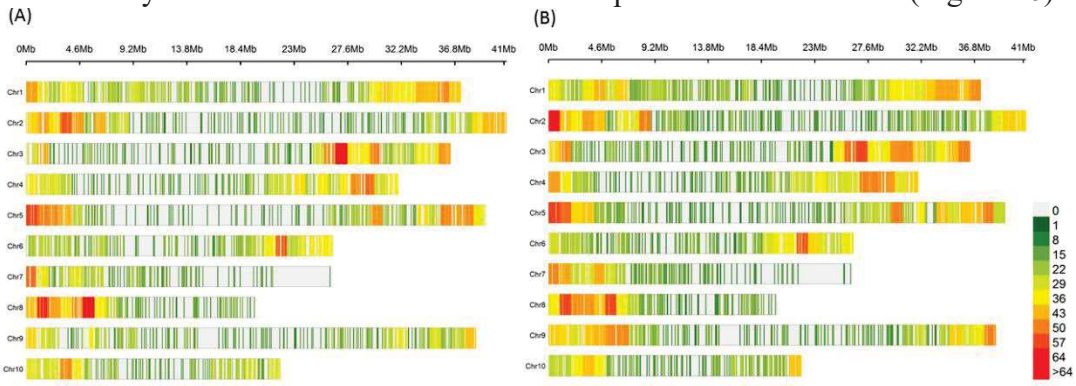


Figure 16: Distribution of markers along the ten chromosomes of T.cacao.
(A) Distribution of markers from the G7 dataset along the ten chromosomes of T.cacao. The graph shows the distribution of markers along the ten chromosomes. The density is calculated on a 1Mb window. The areas without markers are shown in grey. The weakly marked areas are in green and the strongly marked areas are in red. A colour gradient between green and red represents the marking gradient. (B) Distribution of markers from the MAF5 dataset along the ten chromosomes of T.cacao.

4.3.3-Determination of confidence intervals of associations based on haplotypes

Haplotypes were calculated based on the known linkage disequilibrium of the population which is 15 cM, corresponding to 10000 kb. A total of 681 haplotypic blocks were thus determined with a minimum of 42 haplotypic blocks present on chromosome 8 and a maximum of 96 haplotypic blocks present on chromosome 1. Confidence intervals were defined based on these haplotypic blocks. In this paper, each association zone, thus corresponding to a haplotypic block, is represented by its association peak. The association peak corresponds to the marker for which the association is the most significant.

4.3.4-Comparison of the four different methods used for SNP association studies

The GLM method has made it possible to highlight more areas of association than the MLM method. In both cases, the use of the set of markers sorted according to a 5 % MAF (MAF5) also made it possible to highlight more association zones: 333 against 295 for the GLM

method and 152 against 94 for the MLM method. The MLM method, therefore, appears to be more stringent.

Some areas of the association are common for different methods. For example, in the case of terpene relatives' traits, sixty-three co-locations between positive associations for different methods for the same trait was found on all chromosomes except chromosome 4 and 8. A co-localization between GLM_MAF5 and GLM_G7 methods for linalool cis pyranic oxide (UR) was observed on chromosome 2 as shown in Figure 17A. In the case of L-phenylalanine relatives' traits, co-locations of the association zones between the different methods for the same trait was observed on all chromosomes. This is the case for example on chromosome 5 where co-localization of associations for GLM_MAF5, MLM_G7, and MLM_MAF5 linked to 4-hydroxyacetophenone (UR) was observed (Appendix 7).

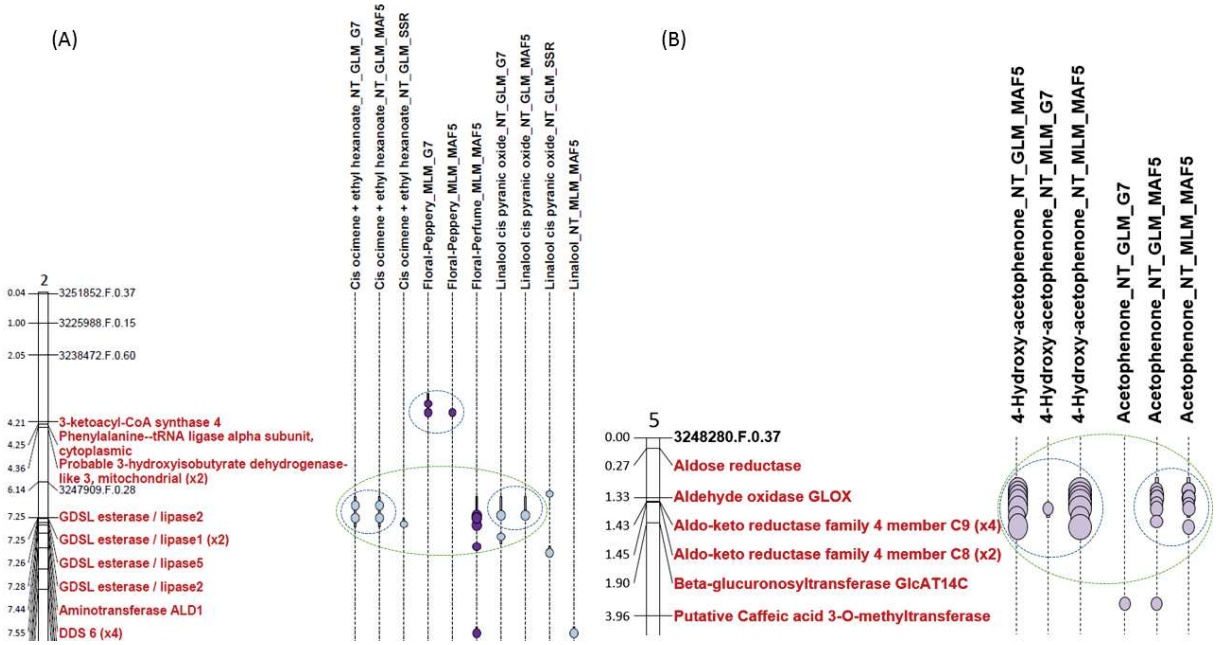


Figure 17: Extract of chromosome 2 map and chromosome 5 map.

(A) Extract from the chromosome 2 map representing the associations detected for compounds involved in the monoterpene biosynthetic pathway. (B) Extract from the chromosome 5 map representing the associations detected for compounds involved in the L-phenylalanine degradation pathway. The light blue dots represent the peaks of associations in relation to traits whose beans have not been roasted. The dark purple dots represent the peaks of association in relation to traits whose beans have been roasted. The bars around these points correspond to the confidence intervals of the association zone. Co-locations are represented by a blue circle for the associations co-localized for the same trait and identified by different methods. Co-locations are represented with a green circle for the co-locations between different biochemical compounds. Candidate genes are written in red. One scale unit on the chromosome corresponds to 1Mb.

4.4-Identification of significant associations for sensorial traits

Among all the associations, only 38 are related to the sensory data with floral notes. Out of a total of sixteen floral perceptions, significant associations were detected for eleven of them, on all chromosomes except chromosome 5 and chromosome 7. Only one area of association

was revealed for each of the six floral notes: the floral notes bark woody, dark wood, mushrooms, orange blossom, other spice, and tobacco (Table 2). Four association zones were also detected for the floral note Lightwood on chromosome 1. The area of strongest association detected for the light wood floral note and the tobacco floral note is in the same haplotypic block. The floral note that allowed detecting the most areas of association is the floral perfume where thirteen areas were highlighted. The variation in the floral perfume note is the one that seems to be the most explained by the genetic variation observed, with an explanation rate for variation in the trait of 24 %.

Table 2: Most significant association detected for each of the sensory floral traits

CH	Position of the association peak	N° hap. bloc	Floral note detected	*p-value* of the strongest association	Explanation rate of the trait of the strongest association	Total number of associations for the character
1	4 079 457 bp	NA	Floral-Other spice	1,45E-05	14%	1
1	4 129 759 bp	10	Floral-Lightwood	2,56E-06	16%	4
1	4 131 970 bp	10	Floral-Tobacco	4,52E-06	16%	1
2	3 606 270 bp	NA	Floral-Peppery	9,42E-07	15%	5
2	7 476 546 bp	36	Floral-Perfume	1,87E-09	24%	13
6	21 137 437 bp	45	Floral-wood resin	8,74E-07	15%	6
6	26 160 073 bp	NA	Floral-Dark wood	5,78E-06	13%	1
8	15 196 137 bp	37	Floral-Orange Blossom	0,000132	12%	1
9	38 188 583 bp	58	Floral-Mushrooms	1,57E-08	22%	1
9	4 248 470 bp	17	Floral-Bark woody	6,56E-06	15%	1
9	6 245 108 bp	21	Floral-Green vegetative	7,45E-09	23%	4

CH: chromosome; hap: haplotypic; bp: base pair

4.5-Identification of significant associations for biochemical traits

The GWAS analyses brought to light 393 association zones. Some of them were detected with several volatile compounds. All the associations found can be consulted in the Appendix 8.

Significant associations for eighteen volatile compounds in unroasted beans and seventeen volatile compounds in roasted beans were identified (Appendix 8). No association zones were detected for five volatile compounds, four of which were assayed in roasted beans: ethylphenyl acetate (UR), ethyl 2-hydroxyhexanoate (R), ethyl hexanoate (R), guaiacol (R), and cis linalool oxide (R).

Two major pathways for the biosynthesis of compounds known to have a floral taste, among those compounds for which a significant association was detected, seem to be particularly represented: the monoterpene biosynthesis pathway and, the L-phenylalanine degradation pathway that allows the synthesis of, among others, acetophenone and 2-phenylethanol.

The results obtained were mapped to visualize the areas of significant associations, their locations, as well as possible co-locations between them. Two maps were made. A map with the results of significant associations related to the compounds involved in the terpene biosynthesis pathway and the floral traits from the sensorial evaluation. A second map includes the results of the significant associations of floral tastes and of compounds involved in the degradation pathway of L-phenylalanine which allows, the synthesis of acetophenone and 2-phenylethanol known to have a floral taste. Some results differ between the different methods (GLM and MLM) or the sorting of SNP markers (MAF5 or G7) or between the type of SNP and SSR markers. All results are shown on the maps in Appendix 7 and Appendix 9. Results that are repeatable between methods appear to be the most conclusive.

4.6-Significant associations identified for the biochemical compounds involved in terpene biosynthetic pathway

Among the 27 compounds related to the floral note, six volatile compounds derived from the terpene biosynthesis pathway: linalool (UR and R), trans furanic oxide linalool (UR), cis pyranic oxide linalool (UR), epoxylinalool (R), and cis ocimene co-eluted with ethyl hexanoate (UR) Figure 18. Eighteen zones of association were revealed for the linalool in unroasted beans (UR) against two zones for linalool in roasted beans (R). The most significant association of linalool (UR) was found on chromosome 7 while that of linalool (R) was found on chromosome 6. Twenty-nine association zones were highlighted for the linalool trans furanic oxide (UR). The most significant association linked to linalool trans furanic oxide (UR) was detected on chromosome 7 which is in the same haplotypic bloc of the most significant association of cis ocimene co-eluted with ethyl hexanoate (UR). Twenty-seven associations were observed for linalool cis pyranic oxide (UR). Finally, thirty-eight areas of associations were revealed for the epoxylinalool (R) (table 3, Appendix 8).

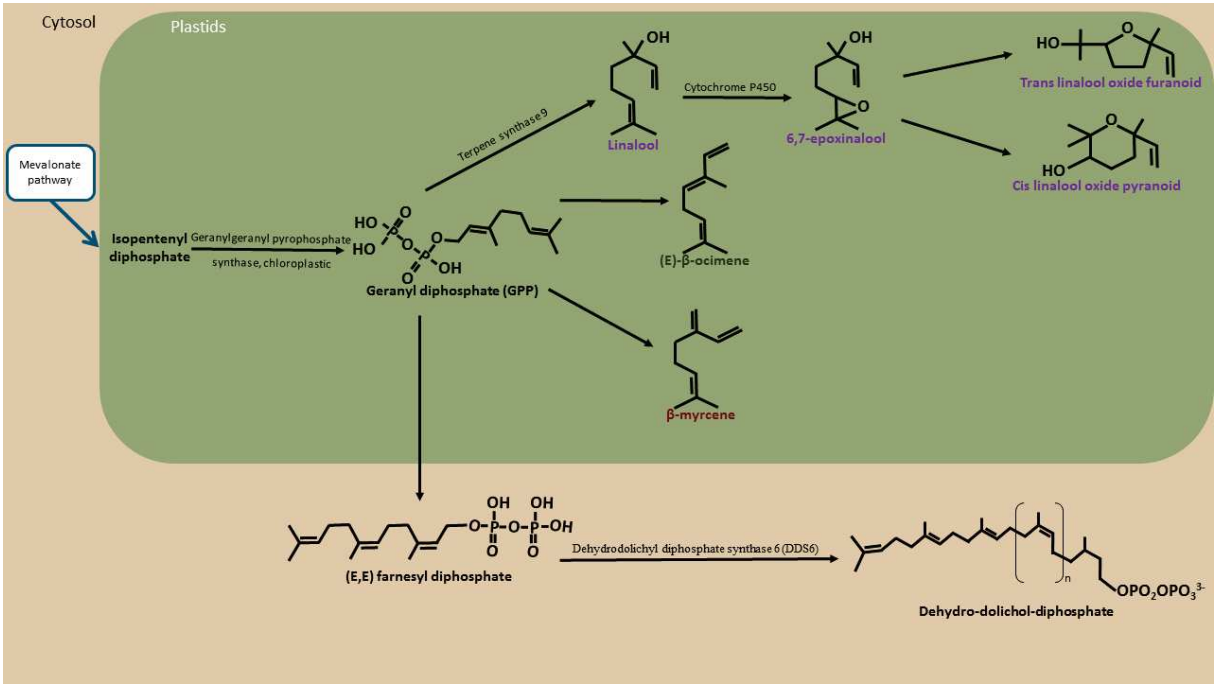


Figure 18: Terpene biosynthesis pathway.

The schema illustrates the different biosynthesis pathways of compounds belonging to the terpene biosynthesis pathway identified in cocoa. Compounds known to have a floral taste are noted in purple. The blue arrows represent the bridges between the terpene biosynthesis pathway and mevalonate pathways. The names of these other biosynthetic pathways are framed in blue. The black arrows represent the enzymatic actions. The names of the enzymes are indicated (when identified) around these arrows. In green are represented the limit of the plastids. In light brown are represented the limit of cytosol.

Table 3:Most significant associations for biochemical compounds related to terpene pathway

CH	Position of the association peak	N° hap. bloc	Traits	*p-value* of the strongest association	Explanation rate of the trait of the strongest association	Associations detected
1	6 448 063 bp	23	Epoxylinalool_R *	3,23E-11	32%	38
6	5 543 124 bp	17	Linalool_ R *	5,57E-06	16%	2
7	5 607 833 bp	24	Linalool trans furanic oxyde_UR *	5,51E-10	29%	29
7	5 607 833 bp	24	Cis ocimene + ethyl hexanoate_UR	3,73E-12	35%	42
7	10 459 413 bp	31	Linalool_ UR *	3,53E-13	37%	18
10	6 167 221 bp	27	Linalool cis pyranic oxide_ UR *	3,27E-07	23%	27

The most significant association detected for each compound is reported. CH: chromosome; hap.: haplotypic; UR: unroasted beans; R: roasted beans; *: biochemical compounds known for floral notes, bp: base pair.

The map with the results for terpenes (Appendix 9) shows several interesting results. Among a large number of associations, several co-locations can be observed between different biochemical compounds involved in the terpene pathway. For example, a co-localization between the Linalool (UR), the Linalool cis-pyranic oxide (UR) and the Linalool trans-furanic oxide (UR) was observed in chromosome 6 (Appendix 9). This suggests the greater likelihood that most of these compounds already known for their floral notes are well involved in floral notes of Nacional cocoa.

4.6.1-Co-locations between biochemical compounds

Sixteen co-locations between different biochemical compounds were also observed on chromosomes 2, 4, 5, 7, 9 and 10, for example on chromosome 2 between the linalool cis pyranic oxide (UR) and cis-ocimene co-eluted with ethyl hexanoate (Figure 17A). Various numbers of co-locations could be observed according to chromosomes. Only one co-location are observed on chromosome 9 and chromosome 10 and five co-locations were highlighted on chromosome 7 (Appendix 9). Co-localizations between association zones identified for different volatile compounds can be explained by their belonging to the same biosynthesis pathway such as for linalool trans furanic oxide (UR) and linalool (UR) on chromosome 3, or for cis pyranic oxide (UR) and epoxylinalool (R) on chromosome 4 (Appendix 9). It can then be thought that this zone of associations is due to the presence of a gene coding for an enzyme that is part of this biosynthetic pathway. To verify this hypothesis, we have begun to search for candidate genes at the level of the association zones.

4.6.2-Co-locations between biochemical compounds and sensorial traits

Seven co-locations between at least one biochemical compound and a floral note were detected on chromosomes 1 and 2. On chromosome 1, two co-locations were observed between epoxylinalool (R) and the floral note lightwood and one between epoxylinalool (R), floral notes lightwood and floral notes tobacco (Appendix 9). On chromosome 2, a co-localization exists between cis ocimene co-eluted with ethyl hexanoate (UR), cis pyranic oxide linalool (UR) and floral scent (Figure 17A). A co-localization is also observable between cis ocimene co-eluted with ethyl hexanoate (UR) and floral perfume. A co-localization is also observable between linalool (UR) and the floral perfume note (Appendix 9).

4.7-Significant associations identified for the biochemical compounds involved in the degradation of L-phenylalanine pathway.

Eighteen compounds for which significant associations have been identified appear to be involved in the degradation pathway of L-phenylalanine to either 2-phenylethanol or acetophenone (Table 4, Figure 19). Among these compounds for two of them, ethylphenyl acetate (R) and phenylethanal (UR), only one zone of the association was identified. The most significant association for phenylethanal (UR) co-localizes with the strongest association detected for linalool (R) on chromosome 6. Thirty-six association zones were showed for acetophenone (UR) compared to forty for acetophenone (R). The most significant association of acetophenone (UR) is on chromosome 2 while that of acetophenone (R) is on chromosome

6. Two hundred and six association zones were detected for cinnamaldehyde (R). Twelve zones of associations were revealed for 2-phenylethanol (UR) and three for 2-phenylethanol (R). The most significant association zones for 2-phenylethanol (UR) and (R) are located on chromosome 4 but at a different position. Two association zones were highlighted for ethyl benzoate (UR). Three areas of association were revealed for 2-phenylethyl acetate (UR). Two zones of associations were revealed for benzaldehyde (UR) against seventy-two with benzaldehyde (R). Benzaldehyde (UR) presents its most significant association on chromosome 7, while that of benzaldehyde (R) is located on chromosome 6. Thirty-eight association zones were revealed for benzyl acetate (UR) against two for benzyl acetate (R). Twenty-nine association zones were highlighted for 4-hydroxy acetophenone (UR). Seven regions of associations were revealed 2-ethylhexan-1-ol (R). Seventy-three association areas were highlighted for 1-phenylethyl acetate (R). The last two compounds involved in these biosynthetic pathways, benzyl acetate (R) and 1-phenylethyl acetate (R), have their most significant area of association co-locating and forming part of the same haplotypic block number 26 on chromosome 10. The variation of two biochemical compounds seems to be explained mainly by genetic variation. Indeed, the variation in the concentration of 4-hydroxy-acetophenone is explained at 79 % by the strongest association zone as well as the variation in cinnamaldehyde which is explained at 65 % by the association zone.

Table 4:Most significant associations for biochemical compounds related to L-phenylalanine degradation pathway

CH	Position of the association peak	N° hap. bloc	Traits	*p-value of the strongest association*	Explanation rate of the trait of the strongest association	Associations detected
10	5 308 832 bp	26	1-phenylethyl acetate_ R	1,17E-10	31%	73
9	1 099 704 bp	5	2-ethylhexan-1-ol_ R *	2,61E-08	29%	7
4	23 646 147 bp	40	2-phenylethanol_ R *	2,12E-06	16%	3
4	17 349 904 bp	NA	2-phenylethanol_ UR *	9,89E-07	18%	12
5	29 926 884 bp	53	2-phenylethyl acetate_ UR *	6,92E-06	18%	3
7	2 815 797 bp	11	4-hydroxy-acetophenone_ UR	4,86E-43	79%	29
6	23 097 785 bp	58	Acetophenone_ R *	1,59E-08	25%	40
2	8 389 914 bp	39	Acetophenone_ UR *	2,53E-12	31%	36
6	25 213 164 bp	65	Benzaldehyde_ R	8,53E-09	26%	72
7	10 459 413 bp	31	Benzaldehyde_ UR	1,32E-05	17%	2
7	2 092 063 bp	9	Benzyl acetate_ UR *	2,36E-15	41%	38
10	5 228 191 bp	26	Benzyl acetate_ R *	8,15E-06	18%	2
2	7 448 797 bp	36	Benzylic alcohol_ UR	2,12E-08	25%	42
3	31 503 427 bp	58	Cinnamaldehyde_ R	1,39E-28	65%	206
5	27 513 744 bp	45	Ethyl benzoate_ UR *	2,33E-06	18%	2
1	36 447 062 bp	91	Ethylphenyl acetate_ R *	1,35E-05	17%	1
6	5 543 124 bp	17	Phenylethanal_ UR *	1,77E-05	14%	1

The most significant association detected for each compound is reported. CH: chromosome; hap.: haplotypic; UR: unroasted beans; R: roasted beans; *: biochemical compounds known for floral notes, bp: base pair.

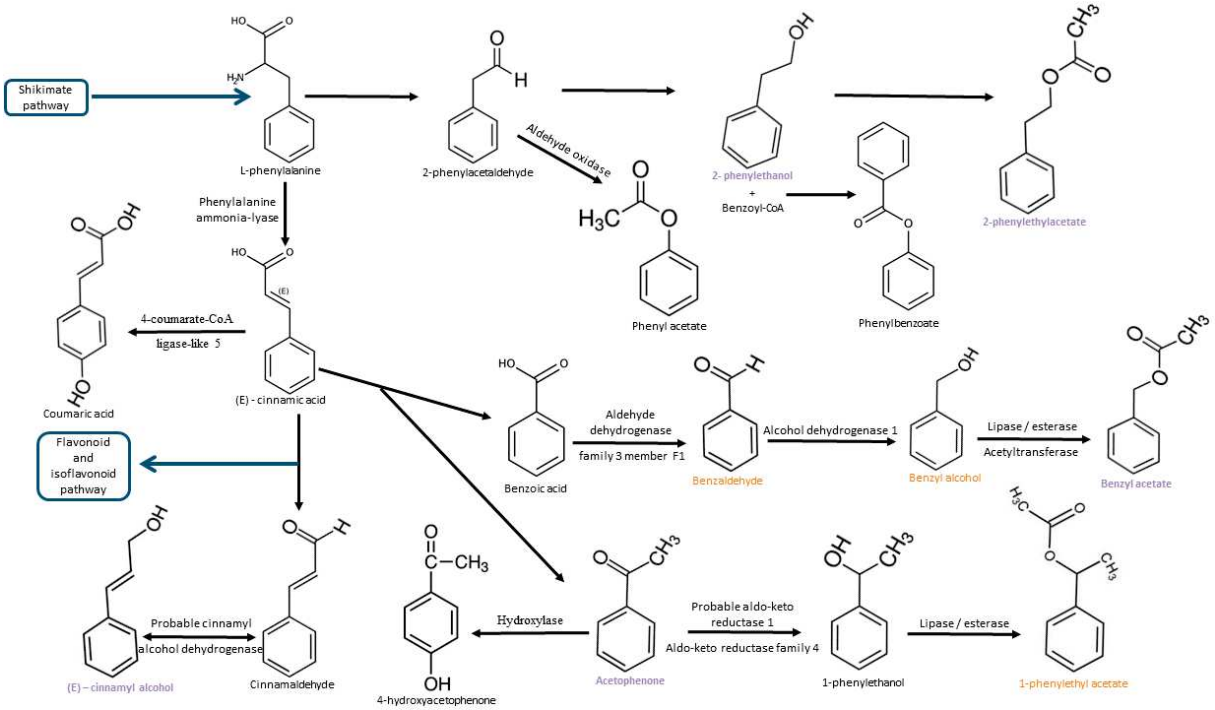


Figure 19: Degradation pathway of L-phenylalanine adapted from Lapadatescu et al., (2000)

The schema illustrates the different biosynthesis pathways of compounds belonging to the L-phenylalanine degradation pathway identified in cocoa. Compounds known to have a floral taste are noted in purple. Compounds known to have a fruity taste are noted in orange. The blue arrows represent the bridges between the L-phenylalanine degradation pathway and other biosynthetic pathways. The names of these other biosynthetic pathways are framed in blue. Black arrows represent the enzymatic actions. The names of the enzymes are indicated (when identified) around these arrows.

The map showing the results for compounds of the L-phenylalanine degradation pathway (Appendix 7) shows several interesting results.

One hundred and eleven co-locations between different volatile compounds were also observed on all chromosomes. An example of co-localization was observed between 4-hydroxyacetophenone (UR) and acetophenone (UR) on chromosome 5 (Figure 17B).

Thirteen co-locations between at least one biochemical trait and one sensory trait were observed on chromosomes 1, 2, 8 and 9 (Appendix 7).

4.8-Significant associations were identified for the biochemical compounds involved in other pathways.

Several areas of association were highlighted for seven other compounds known also to have a floral taste: ethyl dodecanoate (R), guaiacol (UR and R), hexyl acetate (UR), furfural (UR and R), propyl acetate (R), and nonanal (UR). One hundred and seventeen association

zones were detected for guaiacol (UR) against zero for guaiacol (R). Twelve association zones were observed for furfural (UR) compared to thirty for furfural (R) (Table 5). The variation in hexyl acetate concentration is very high compared to other compounds. On the other hand, the genetic explanation for the variation in the concentration of propyl acetate is very weak compared to the other characteristics of this study (4 %).

Table 5: Most significant associations for biochemical compounds related to other pathways

CH	Position of the association peak	N° hap bloc	Traits	*p-value of the strongest association*	Explanation rate of the trait of the strongest association	Associations detected
3	31 273 182 bp	57	Ethyl dodecanoate _ R *	6,60E-11	36%	50
10	5 308 832 bp	26	Furfural_ R *	3,15E-07	22%	13
7	4 776 442 bp	NA	Furfural_ UR *	3,70E-08	24%	5
4	737310 bp	NA	Guaiacol_ UR *	1,60E-09	28%	54
7	10 459 413 bp	31	Hexyl acetate_ UR *	1,01E-138	67%	94
9	37 557 289 bp	54	Nonanal_ UR *	2,45E-07	22%	9
9	3 653 985 bp	14	Propyl acetate _ R *	2,42E-14	4%	41

The most significant association detected for each compound is reported. CH: chromosome; hap: haplotypic; UR: unroasted beans; R: roasted beans; *: biochemical compounds known for floral notes, bp: base pair.

4.9-Candidate genes potentially involved in the formation of the floral aroma

Of the 393 association zones exposed, twenty-seven with candidate genes with predicted functions were identified.

4.9.1-Candidate genes linked to the terpene biosynthesis pathway

Candidate genes related to the terpene biosynthetic pathway were found on chromosomes 1, 2, 5, 7, 9 and 10. The association zone number and candidate genes are reported in Appendix 9, Appendix 10 and Table 6.

Table 6: Candidate genes identified for terpene biosynthesis pathway

N° asso	CH	Position of candidate gene (bp)	Position of the pic of association (bp)	Candidate gene function	Trait in association
1	1	1 883 959	2 368 915	*Geranylgeranyl pyrophosphate synthase, chloroplastic*	Epoxylinalool-R
2	1	3 173 783	3 379 361	*Cytochrome P450 81E8*	Epoxylinalool (R), floral note lightwood
2	1	3 179 883	3 379 361	*Cytochrome P450 81E8*	Epoxylinalool (R), floral note lightwood
3	1	6 026 043	6 130 253	*Cytochrome P450 78A7*	Epoxilinalool (R)
4	2	7 549 475	7 324 500	*Dehydrodolichyl diphosphate synthase 6*	Cis ocimene co-eluted with ethyl hexanoate (UR), floral perfume
4	2	7 551 259	7 324 500	*Dehydrodolichyl diphosphate synthase 6*	Cis ocimene co-eluted with ethyl hexanoate (UR), floral perfume
4	2	7 553 795	7 324 500	*Dehydrodolichyl diphosphate synthase 6*	Cis ocimene co-eluted with ethyl hexanoate (UR), floral perfume
4	2	7 572 073	7 324 500	*Putative Dehydrodolichyl diphosphate synthase 6*	Cis ocimene co-eluted with ethyl hexanoate (UR), floral perfume
5	2	8 257 841	8 389 914	*Probable 3-hydroxyisobutyryl-CoA hydrolase 2*	Linalool cis pyranic oxide (UR)
6	5	32 749 861	33 303 465	*Cytochrome P450 89A2*	Linalool (UR) and linalool trans furanic oxide (UR)
6	5	33 057 632	33 303 465	*Cytochrome P450 89A9*	Linalool (UR) and linalool trans furanic oxide (UR)
6	5	33 064 477	33 303 465	*Cytochrome P450 89A2*	Linalool (UR) and linalool trans furanic oxide (UR)
6	5	33 073 996	33 303 465	*Cytochrome P450 89A2*	Linalool (UR) and linalool trans furanic oxide (UR)
6	5	33 094 200	33 303 465	*Cytochrome P450 89A2*	Linalool (UR) and linalool trans furanic oxide (UR)
6	5	33 099 009	33 303 465	*Cytochrome P450 89A2*	Linalool (UR) and linalool trans furanic oxide (UR)
7	7	6 346 577	6 181 185	*Putative Probable terpene synthase 9*	Linalool cis pyranic oxide (UR)
7	7	6 365 007	6 181 185	*Probable terpene synthase 9*	Linalool cis pyranic oxide (UR)
7	7	6 380 614	6 181 185	*Putative Probable terpene synthase 9*	Linalool cis pyranic oxide (UR)
8	9	791 968	749 365	*3-hydroxyisobutyryl-CoA hydrolase-like protein 2, mitochondrial*	Epoxylinalool (R)
9	10	6 317 543	6 167 221	*Probable terpene synthase 9*	Linalool cis pyranic oxide (UR)

asso.: Associations, CH: chromosome, R: Roasted beans, UR: unroasted beans

On chromosome 1, three association zones contain candidate genes. Association zone 1 (805,132 – 2,445,782 bp) linked to epoxylinalool (R) contains a gene coding for a "*Geranylgeranyl pyrophosphate synthase, chloroplastic*". This enzyme allows the synthesis of geranylgeranyl pyrophosphate in chloroplasts. This compound is a precursor of terpenes. As the monoterpene biosynthesis pathway is located in the plastids, the indication of chloroplastic synthesis seems to confirm the correspondence to another compound derived from linalool also synthesized in Chloroplast (Ying and Qingping, 2006; Feng et al., 2014). Association zone 2 (3,083,032 - 3,398,183 bp) linked to epoxylinalool (R) and the floral note lightwood contains two candidate genes encoding a "*Cytochrome P450 81E8*". Cytochrome P450 has been identified to be responsible for the synthesis of epoxylinalool from linalool in kiwifruit (Chen et al., 2010). Association zone 3 (5 940 526 - 6 204 028 bp) linked to epoxylinalool (R) contains a candidate gene encoding a "*Cytochrome P450 78A7*".

On chromosome 2 (Appendix 9), two association zones contain candidate genes. Association zone 4 (7,324,500 - 7,617,242 bp) linked to cis ocimene co-eluted with ethyl hexanoate (UR) and floral perfume contains four genes encoding a "*Dehydrodolichyl diphosphate synthase 6*" (DDS 6) in Figure 17A. Dehydrodolichyl diphosphate synthase 6

allows the synthesis of dehydrodolichyl diphosphate, one of the precursors of which is geranyl diphosphate, the main precursor of the monoterpene biosynthesis pathway. The synthesis of dehydrodolichyl diphosphate could thus compete with the synthesis of cis-ocimene and explain the association with this compound as well as with the floral perfume, which is a taste attributed to several monoterpenes (linalool, epoxylinalool, ocimene). Association zone 5 (8,239,972 - 8,416,672 bp) linked to linalool cis pyranic oxide (UR) contains a gene encoding a "*Probable 3-hydroxyisobutyryl-CoA hydrolase 2*". The enzyme 3-hydroxyisobutyryl-CoA hydrolase 2 can enable the production of acetyl-CoA by releasing a CoA. Acetyl-CoA is a precursor of the mevalonate biosynthetic pathway that allows the production of geranyl diphosphate (Kreck et al., 2003; Miziorko, 2011).

On chromosome 5, only association region 6 (32,660,102 - 33,718,239bp) contains candidate genes. It is linked to linalool (UR) and linalool trans furanic oxide (UR) and contains six candidate genes, five of which are known to code for "*Cytochrome P450 89A2*" and one for "*Cytochrome P450 89A9*" (Figure 20, Appendix 9). The presence of cytochrome P450 could explain the associations with linalool and trans furanic oxide linalool as they would allow the transformation of linalool into epoxylinalool (Chen et al., 2010).

On chromosome 7 (Appendix 9), only association zone 7 (6,128,106 - 6,410,151bp) contains candidate genes. It is linked to linalool cis pyranic oxide (UR) and contains three genes encoding "*Probable terpene synthase 9*". Terpene synthases 9 are known to be involved in the synthesize of linalool, one of the precursors of linalool cis pyranic oxide (Cseke et al., 1998).

On chromosome 9 (Appendix 9), only association zone 8 (713,588 - 857,818bp) contains a candidate gene. It is linked to epoxylinalool (R) and contains a gene encoding a "*3-hydroxyisobutyryl-CoA hydrolase-like protein 2, mitochondrial*". This enzyme is involved in the mevalonate biosynthetic pathway, one of the biosynthetic pathways leading to the formation of geranyl diphosphate, a key compound in the monoterpene biosynthetic pathway (Lamarti et al., 1994).

On chromosome 10 (Appendix 9), the association zone 9 (6,023,982 - 6,718,126bp) linked to linalool cis pyranic oxide (UR) contains a gene coding for "*Probable terpene synthase 9*". This enzyme is known to synthesize linalool, which could enable the synthesis of linalool cis pyranic oxide.

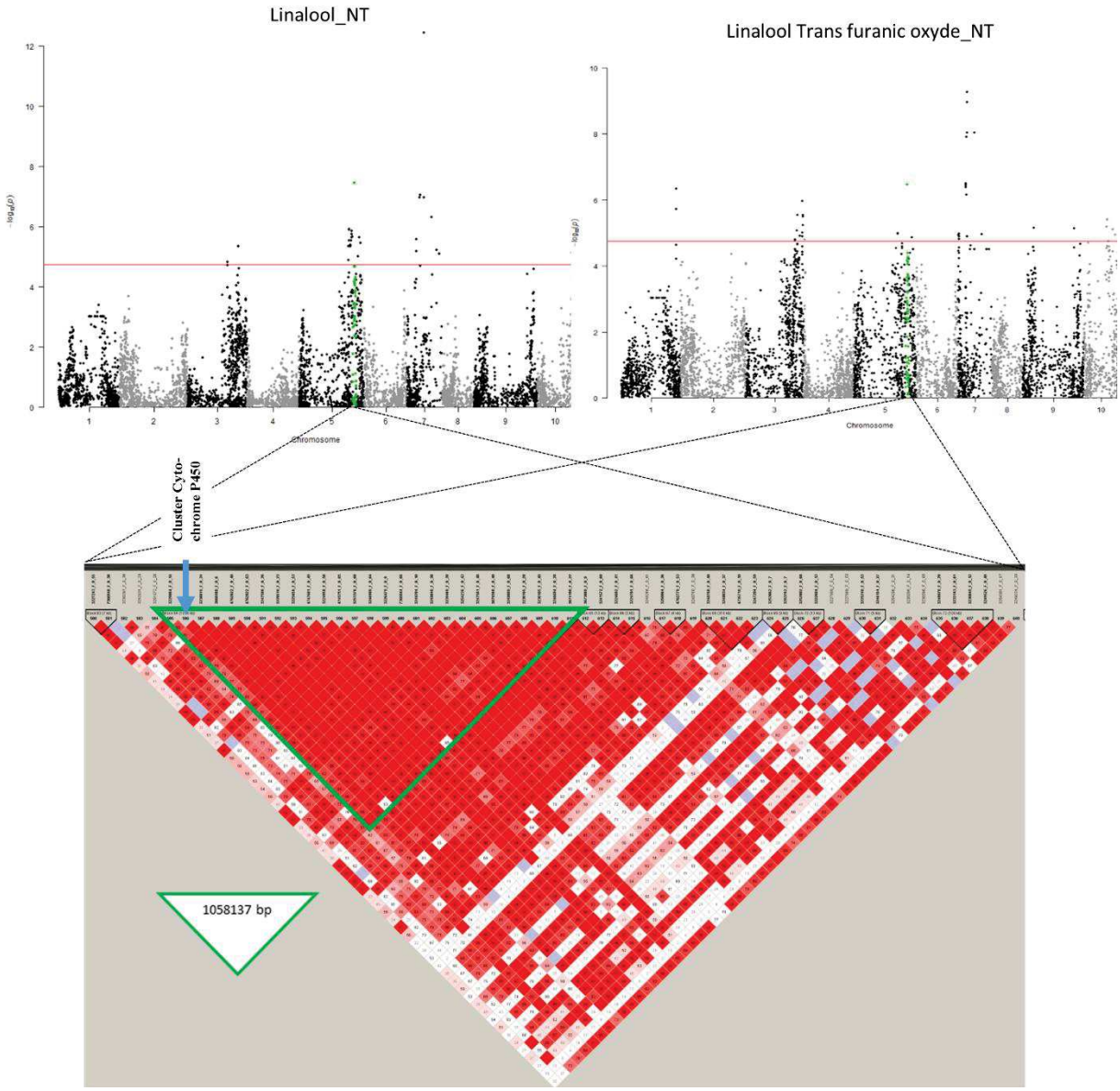


Figure 20: Co-localization between linalool (UR) and trans furanic oxide (UR) with candidate genes.

(A) Manhattan plot representing association results for the Linalool (UR) trait, revealed by GLM-MAF5 method. (B) Manhattan plot representing association results for linalool trans furanic oxide (UR) trait, revealed by GLM-MAF5 method. (C) Heat map of a part of chromosome 5. The common region of association is represented by a green triangle.

4.9.2-Candidate genes linked to the L-phenylalanine degradation pathway

In a second step, candidate genes linked to the L-phenylalanine degradation pathway were found on chromosomes 1, 2, 4, 5, 7, 8, 9 and 10. The association zone number and candidate genes are reported in Appendix 7, Appendix 10 and Table 7.

Table 7: Candidate genes identified for L-phenylalanine degradation pathway

N° asso	CH	Position CG (bp)	Position PA (bp)	Gene function	Trait in association
10	1	1 102 658	2 430 002	*Aldehyde dehydrogenase family 3 member F1*	1-phenylethyl acetate (R), benzaldehyde (R) and cinnamaldehyde (R)
11	1	3 107 999	3 379 361	*Probable cinnamyl alcohol dehydrogenase 7/8*	1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R), cinnamaldehyde (R) and the floral note lightwood
11	1	3 112 047	3 379 361	*Probable cinnamyl alcohol dehydrogenase*	1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R), cinnamaldehyde (R) and the floral note lightwood
12	1	6 039 217	5 940 526	*Shikimate kinase 1, chloroplastic*	1-phenylethyl acetate (R) and cinnamaldehyde (R)
13	1	7 124 110	6 855 567	*Alcohol dehydrogenase 1*	1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R) and cinnamaldehyde (R),
13	1	7 131 266	6 855 567	*Alcohol dehydrogenase 1*	1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R) and cinnamaldehyde (R),
14	2	7 436 835	7 453 377	*Aminotransferase ALD1*	Acetophenone (NT and R), benzaldehyde (R), benzyl alcohol (UR), cinnamaldehyde (R) and the floral perfume note
15	4	22 566 348	22 503 297	*Acetyltransferase NSI*	1-phenylethyl acetate and cinnamaldehyde (R)
16	4	26 741 963	26 876 494	*3-ketoacyl-CoA thiolase 2, peroxisomal*	1-phenylethyl acetate (R)
16	4	26 715 852	26 876 494	*Chalcone synthase 2*	1-phenylethyl acetate (R)
17	4	27 604 955	27 507 597	*2-hydroxyisoflavanone dehydratase*	Floral perfume
17	4	27 608 704	27 507 597	*2-hydroxyisoflavanone dehydratase*	Floral perfume
18	4	28 270 492	28 285 175	*Probable aldo-keto reductase 1*	1-phenylethyl acetate (R)
19	5	1 328 453	1 353 636	*Aldehyde oxidase GLOX*	4-hydroxy acetophenone (UR), acetophenone (UR) and benzyl acetate (UR)
20	5	1 431 497	1 380 802	*Aldo-keto reductase family 4 member C9*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
20	5	1 435 043	1 380 802	*Aldo-keto reductase family 4 member C9*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
20	5	1 438 324	1 380 802	*Aldo-keto reductase family 4 member C9*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
20	5	1 441 067	1 380 802	*Aldo-keto reductase family 4 member C9*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
20	5	1 444 916	1 380 802	*Aldo-keto reductase family 4 member C8*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
20	5	1 450 213	1 380 802	*Aldo-keto reductase family 4 member C8*	4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R)
21	5	2 978 188	2 732 709	*Phenylalanine ammonia-lyase*	Cinnamaldehyde (R)
22	5	30 446 215	30 471 918	*Alcohol dehydrogenase-like 6*	Benzaldehyde (R)
23	7	2 055 963	2 092 063	*GDSL esterase/lipase At1g28570*	4-hydroxy acetophenone (UR), acetophenone (UR) and benzyl acetate (UR)
24	8	1 199 801	1 148 435	*3-ketoacyl-CoA synthase 4*	Floral note wood resin
25	8	2 170 173	2 251 806	*GDSL esterase/lipase EXL3*	Cinnamaldehyde (R)
26	8	6 559 785	6 751 843	*Acetyltransferase At1g77540*	Acetophenone (UR) and benzyl acetate (UR)
26	8	6 570 129	6 751 843	*Caffeic acid 3-O-methyltransferase*	Acetophenone (UR) and benzyl acetate (UR)
26	8	6 581 002	6 751 843	*Caffeic acid 3-O-methyltransferase*	Acetophenone (UR) and benzyl acetate (UR)
27	8	15 370 133	14 498 544	*Putative O-acyltransferase WSD1*	Benzaldehyde (R), benzyl acetate (UR), cinnamaldehyde (R) and orange blossom note
27	8	15 402 378	14 498 544	*Putative O-acyltransferase WSD1*	Benzaldehyde (R), benzyl acetate (UR), cinnamaldehyde (R) and orange blossom note
28	8	18 679 641	19 249 315	*Putative GDSL esterase/lipase At1g29670*	Benzyl acetate (UR)
29	9	5 605 457	6 010 658	*GDSL esterase/lipase EXL3, putative*	Benzyl alcohol (UR) and the floral note green vegetative
29	9	6 069 175	6 010 658	*3-hydroxyisobutyryl-CoA hydrolase-like protein 3, mitochondrial*	Benzyl alcohol (UR) and the floral note green vegetative
30	9	23 325 443	23 302 911	*Feruloyl CoA ortho-hydroxylase 2*	Acetophenone (R), benzaldehyde (R) and benzyl acetate (UR)
31	10	5 303 984	5 308 832	*Putative 4-coumarate--CoA ligase-like 5*	1-phenylethyl acetate (R), benzyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), to acetophenone (R), benzaldehyde (R) and cinnamaldehyde (R)
6	5	32 749 861	33 303 465	*Cytochrome P450 89A2*	2-phenylethanol (UR)
6	5	33 057 632	33 303 465	*Cytochrome P450 89A9*	2-phenylethanol (UR)
6	5	33 064 477	33 303 465	*Cytochrome P450 89A2*	2-phenylethanol (UR)
6	5	33 073 996	33 303 465	*Cytochrome P450 89A2*	2-phenylethanol (UR)
6	5	33 094 200	33 303 465	*Cytochrome P450 89A2*	2-phenylethanol (UR)
6	5	33 099 009	33 303 465	*Cytochrome P450 89A2*	2-phenylethanol (UR)

asso. : Associations, CH : chromosome,CG: candidate gene, PA: pic of the association, R: Roasted beans, UR: unroasted beans

On chromosome 1, four association zones contain candidate genes. Association zone 10 (805 132 - 2 445 782 bp) linked to 1-phenylethyl acetate (R), benzaldehyde (R) and cinnamaldehyde (R) contains a gene coding for an "*Aldehyde dehydrogenase family 3 member F1*". This enzyme could be responsible for the transformation of benzoic acid into benzaldehyde. The presence of this enzyme could compete with the production of cinnamaldehyde or 1-phenylethyl acetate (Figure 19;Lapadatescu et al., 2000). Association zone 11 (3,083,032 - 3,398,183 bp) linked to 1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R), cinnamaldehyde (R) and the floral note lightwood, contains two candidate genes encoding a "*Probable cinnamyl alcohol dehydrogenase*". These enzymes are known to transform cinnamaldehyde into (E)-cinnamyl alcohol (Wyrambik and Grisebach, 1975). According to another study, "*Probable cinnamyl alcohol dehydrogenase*" has the ability to remove hydrogen from cinnamyl alcohol to convert it to cinnamaldehyde. Cinnamyl alcohol is known to have a floral, cinnamon and balsamic taste (Steinhaus et al., 2009), which may be associated with the floral note lightwood. The association zone 12 (5 940 526 - 6 204 028 bp) linked to 1-phenylethyl acetate (R) and cinnamaldehyde (R) contains a gene encoding a "*Shikimate kinase 1, chloroplastic*". The shikimate biosynthesis pathway allows the synthesis of phenylalanine, a precursor of 1-phenylethyl acetate and cinnamaldehyde (Tohge et al., 2013). The association zone 13 (6 834 165 - 7 942 921), linked to 1-phenylethyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), acetophenone (R), benzaldehyde (R) and cinnamaldehyde (R), contains two genes coding for an "*Alcohol dehydrogenase 1*". Alcohol dehydrogenase is necessary for the degradation of benzaldehyde to benzyl alcohol, which are both compounds with a fruity taste. The other compounds in association in this area are upstream of this degradation reaction, which could explain their associations (Lapadatescu et al., 2000).

On chromosome 2 (Appendix 7), only association region 14 (7,324,500 - 7,617,242bp) contains candidate genes. It is linked to acetophenone (UR and R), benzaldehyde (R), benzyl alcohol (UR), cinnamaldehyde (R) and the floral perfume note and contains a candidate gene coding for an "*ALD1 Aminotransferase*". Several aminotransferases have been identified in the shikimate biosynthesis pathway that allows the synthesis of L-phenylalanine (Tohge et al., 2013).

On chromosome 4 (Appendix 7), four association zones contain candidate genes. Association region 15 (22,435,678 - 22,617,119bp) linked to 1-phenylethyl acetate and cinnamaldehyde (R) contains a gene encoding an "*NSI acetyltransferase*". The acetyl

transferase NSI has the function of acetylating histones. It is likely to play a role in regulating the expression of genes for the synthesis of 1-phenylethyl acetate or cinnamaldehyde. Association zone 16 (26,703,951 - 27,146,370bp) linked to 1-phenylethyl acetate (R) contains two candidate genes coding for: a "*Chalcone synthase 2*" and a "*3-ketoacyl-CoA thiolase 2, peroxisomal*". Chalcone synthases participate in the flavonoid and isoflavonoid biosynthesis pathway that follows the degradation of phenylalanine to cinnamic acid (Pyrzynska and Biesaga, 2009). A ketoacyl-Coa thiolase is required for the synthesis of benzoyl-CoA (Amano et al., 2018). The association zone 17 (27,507,597 - 27,608,727bp) linked to the floral perfume contains two genes encoding a "*2-hydroxyisoflavanone dehydratase*". 2-hydroxyisoflavanone is part of the isoflavonoid biosynthesis pathway. Its transformation could compete with the synthesis of compounds known to have a floral taste such as acetophenone or 2-phenylethanol (Pyrzynska and Biesaga, 2009). The association zone 18 (28,257,730 - 28,352,788bp) linked to 1-phenylethyl acetate (R) contains a gene coding for a "*Probable aldo-keto reductase 1*". An acetaldehyde reductase may be required for the synthesis of 1-phenylethanol from acetophenone, the probable precursor of 1-phenylethyl acetate (Dong et al., 2012).

On chromosome 5, five association zones contain candidate genes. Association region 19 (1,326,444 - 1,374,494bp) linked to 4-hydroxy acetophenone (UR), acetophenone (UR) and benzyl acetate (UR) contains a candidate gene encoding a "*GLOX Aldehyde oxidase*". An aldehyde oxidase is in some cases responsible for the oxidation of phenylacetaldehyde to phenylacetate, both of which are part of the L-phenylalanine degradation pathway (Kücükgöze and Leimkühler, 2018). Association zone 20 (1,380,802 - 1,510,054bp) linked to 4-hydroxy acetophenone (UR), acetophenone (UR), benzyl acetate (UR) and cinnamaldehyde (R) contains six candidate genes, four of which code for an *Aldo-keto reductase family 4 member C9* and two for an *Aldo-keto reductase family 4 member C8* (Figure 21 and Appendix 7). An acetaldehyde reductase may be required for the synthesis of 1-phenylethanol from acetophenone, a probable precursor of 1-phenylethyl acetate (Dong et al., 2012). The association zone 21 (2,674,400 - 3,039,540bp) linked to cinnamaldehyde (R) contains a gene coding for a *Phenylalanine ammonia-lyase*. This enzyme is known to transform L-phenylalanine into cinnamic acid, which is the precursor of cinnamaldehyde (Lapadatescu et al., 2000). The association zone 22 (30,407,214 - 30,473,075bp) linked to benzaldehyde (R) contains a gene coding for an *Alcohol dehydrogenase-like 6*. This enzyme could degrade benzaldehyde to benzyl alcohol. Association zone 6 (32,660,102 - 33,718,239bp) is linked to 2-phenylethanol (UR) (the same to terpene association zone 6). It contains six genes, five of

which code for *Cytochrome P450 89A2* and one for *Cytochrome P450 89A9*. Cytochrome P450 has redox activities. Several of these reactions are involved in the synthesis of 2-phenylethanol (Lapadatescu et al., 2000).

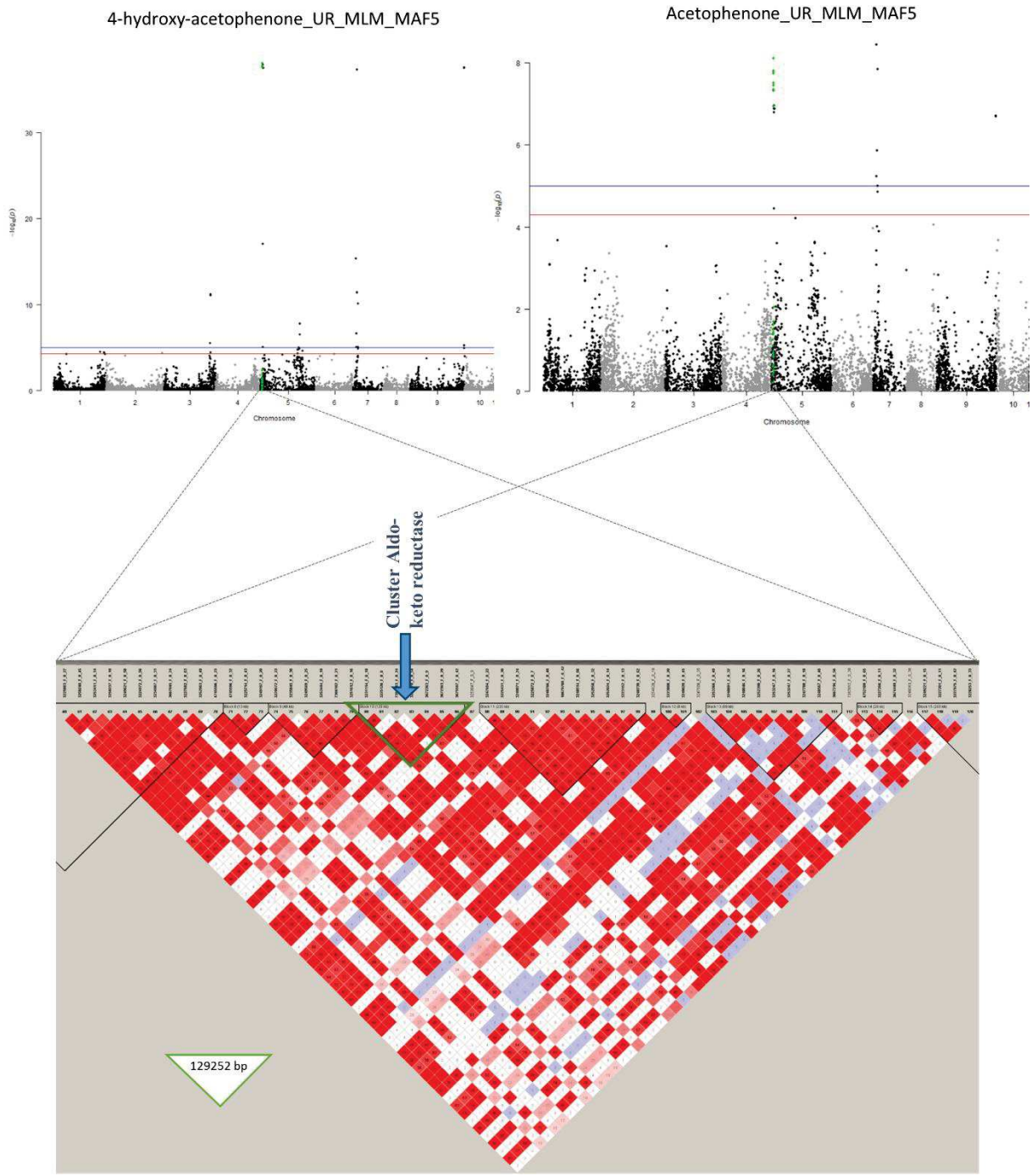


Figure 21: Co-localization between 4-hydroxy-acetophenone (UR) and acetophenone (UR) with candidate genes. (A) Manhattan plot representing association results for the trait 4-hydroxy-acetophenone (UR). (B) Manhattan plot representing association results for acetophenone (UR). (C) Heat map of a part of chromosome 5. The common region of association is represented by a green triangle.

On chromosome 7 (Appendix 7), only association zone 23 (1,894,664 - 2,092,063bp) contains a candidate gene. It is linked to 4-hydroxy acetophenone (UR), acetophenone (UR) and benzyl acetate (UR) and contains a gene encoding a GDSL esterase/lipase At1g28570. A lipase/esterase may be required for the formation of benzyl acetate from benzyl alcohol or the synthesis of 1-phenyl acetate from 1-phenyl ethanol (Mäki-Arvela et al., 2008; Melo et al., 2017).

On chromosome 8 (Appendix 7), five association zones contain candidate genes. Association zone 24 (1,121,979 - 1,520,555bp) linked to the floral note wood resin contains a candidate gene encoding a *3-ketoacyl-CoA synthase 4*. This enzyme is involved in the transformation of a very long chain of acyl-CoA into acetyl-CoA which can itself be transformed into ketones (Tong et al., 2006). Since this zone of associations is linked to the floral note wood resin, this gene can perhaps lead to the synthesis of ketones known to have a floral taste like acetophenone. Association zone 25 (2,021,946 - 2,268,116bp) linked to cinnamaldehyde (R) contains a candidate gene encoding a *GDSL esterase/lipase EXL3*. An esterase/lipase may be required as previously discussed for the formation of benzyl acetate from benzyl alcohol or the synthesis of 1-phenylehtyl acetate (Mäki-Arvela et al., 2008; Melo et al., 2017). The synthesis of these compounds could compete with the synthesis of cinnamaldehyde. The association zone 26 (6,533,242 - 6,978,549bp) linked to acetophenone (UR) and benzyl acetate (UR) is linked to three genes, two of which code for *Caffeic acid 3-O-methyltransferase* and one for *Acetyltransferase At1g77540*. Caffeic acid 3-O-methyltransferase has the role of transforming caffeic acid into ferulic acid and can thus compete with the synthesis of acetophenone or benzyl acetate (Tu et al., 2010). An acetyltransferase is required to convert benzyl alcohol to benzyl acetate (Hao et al., 2014). This function may explain the associations with acetophenone, which requires a common benzyl alcohol precursor for synthesis. Association zone 27 (14 444 953 - 15 439 624bp) linked to benzaldehyde (R), benzyl acetate (UR), cinnamaldehyde (R) and orange blossom note contains two genes encoding a *Putative O-acyltransferase WSD1*. This enzyme allows the synthesis of a "wax ester" from long-chain fatty alcohol. It could allow the synthesis of a "wax ester" with a floral taste of orange blossom type or contribute to this aromatic note. The association zone 28 (17 816 898 - 19 249 315bp) linked to benzyl acetate (UR) contains a candidate gene coding for a *Putative GDSL esterase/lipase At1g29670* that may play a role in the degradation of benzyl acetate (Mäki-Arvela et al., 2008; Melo et al., 2017).

On chromosome 9 (Appendix 7), two association zones contain candidate genes. Association zone 29 (5,327,028 - 6,165,415bp) linked to benzyl alcohol (UR) and the floral note green vegetative contains two genes: one coding for *3-hydroxyisobutyryl-CoA hydrolase-like protein 3, mitochondrial* and one for *GDSL esterase/lipase EXL3, putative*. The 3-hydroxyisobutyryl-CoA hydrolase-like enzyme could lead to the synthesis of terpenes with floral tastes as described above. It could thus explain the association with the floral green vegetative taste. Lipase may be required for the formation of benzyl acetate from benzyl alcohol (Melo et al., 2017). The enzyme encoded by the GDSL esterase/lipase gene EXL3, putative could compete with the synthesis of benzyl alcohol. Association zone 30 (23 101 222 - 23 892 356bp) linked to acetophenone (R), benzaldehyde (R) and benzyl acetate (UR) contains a gene encoding a *Feruloyl CoA ortho-hydroxylase 2*. Ferulic acid has cinnamic acid as a precursor, as do acetophenone, benzaldehyde and benzyl acetate. The activity of this enzyme could therefore compete with the synthesis of these compounds.

On chromosome 10 (Appendix 7), one association zone contains candidate genes. Association zone 31 (5,153,882 - 5,419,006bp) linked to 1-phenylethyl acetate (R), benzyl acetate (R), phenylethyl acetate co-eluted with 2-ethylphenol (R), to acetophenone (R), benzaldehyde (R) and cinnamaldehyde (R) contains a candidate gene encoding a *Putative 4-coumarate-CoA ligase-like 5*. The activity of this enzyme could compete with the synthesis of compounds associated with this region as it could induce a transformation of cinnamic acid to coumaric acid.

5-Discussion

This study contributes to highlighting the importance of cocoa genetic background in the aroma composition of cacao products. The GWAS analyses revealed a large number of associations. Several are related to volatile compounds known for their floral aromas, others are related to compounds, without floral aroma, but involved in the biosynthesis of these aromatic compounds, and others are related to the perception of sensory notes.

5.1-Determination of associations area

The confidence interval of the association zones was determined using haplotypic blocks. This method gives an idea of the size of the association zone as a function of the linkage disequilibrium of the population, which seems biologically logical. However, in some cases, this limit may underestimate the true size of the association, as it is certainly the case on chromosome 1 for the epoxylinalool (R) trait (Appendix 9) where we see hot spots of

associations extending over the first seven megabases. In cases where there is a cluster of very close association zones, it is legitimate to ask whether the method of determining the association zones is not too stringent.

5.2-Insights into the genetic architecture of floral aromas in cocoa

GWAS analysis, two main biosynthesis pathways of compounds known for their floral notes seem to be involved in cocoa floral aromas: the monoterpene synthesis pathway and the L-phenylalanine degradation pathway. These biosynthesis pathways have already been identified in other such as grapes or its derivative wine as important contributors to their floral aromas (Ferreira et al., 1997; Mateo and Jiménez, 2000). Some of the association zones contain candidate genes directly involved in the synthesis of the associated compound, or candidate genes involved upstream in the biosynthetic pathway. The presence of these genes increases the probability that the detected association is not a false positive. The GWAS analyses revealed several genes that appear to be involved in the synthesis of compounds known to have a floral taste and could thus be involved in the variation of floral tastes. Candidate genes coding for enzymes are the most obvious, but other types of genes may be involved in cocoa floral taste such as certain transcriptional factors that could activate or repress several biosynthetic pathways at the same time.

Some associations linked to compounds from the same biosynthesis pathway have been co-localized. Roasting has been suggested to play a role in the transformation of these compounds (Jinap et al., 1998). This could explain some of the co-localization observed in this study, for example, in the terpene biosynthesis pathway the degradation of linalool to epoxylinalool or vice versa (co-localization on chromosome 5), the transformation of cis pyranic oxide linalool to epoxylinalool or the opposite (co-localization on chromosomes 4 and 10). Roasting may also play a role in the transformation of compounds in the L-phenylalanine degradation pathway as, for example: 4-hydroxy acetophenone to acetophenone or vice versa (co-localization on chromosomes 7 and 10), the transformation of benzyl acetate into benzaldehyde or the opposite (co-locations on chromosomes 2, 5, 7, 8, 9 and 10) and the transformation of benzyl alcohol into benzaldehyde or vice versa (co-locations on chromosomes 2, 3, 4, 5, 6, 8 and 10).

Other associations give information on a balance between the presence of aromatic and non-aromatic compounds of the same biosynthetic pathway: suggesting that an enzyme could be responsible for the transformation of one of these compounds into another and thus influence

the flavour as observed in roses by Farhi et al., (2010). The presence of certain odours would thus depend on the activation or repression of the enzyme responsible for the synthesis of the compound with the floral aroma. This is the case, for example, for an area on chromosome 1 associated with cinnamaldehyde and the floral note lightwood containing a gene coding for a "Probable cinnamyl alcohol dehydrogenase". When this enzyme is active, it would allow the transformation of cinnamaldehyde into cinnamyl alcohol. There would then be a possible accumulation of cinnamyl alcohol known to have a floral note. When this enzyme is not active, cinnamaldehyde, which has a spicy (cinnamon) taste, would accumulate. Other areas of association suggest that a similar system has been put in place: this is the case for the co-locations between 1-phenylethyl acetate and acetophenone on chromosomes 1, 6, 9 and 10 where a gene coding for an esterase/lipase has been detected in nearby location for association zones in chromosome 1, 6 and 9 (Appendix 10). If that gene would be active, an accumulation of 1-phenylethyl acetate known to have a fruity odour would be possible. Otherwise, a possible accumulation of acetophenone, also known to have a floral note would be obtained. This is also the case for the co-localization between benzyl acetate and benzyl alcohol on chromosome 2. A cluster of genes coding for an esterase/lipase and a gene with an acetyltransferase function was detected close to co-location (Appendix 10). In this case, if the enzyme is active, an accumulation of benzyl alcohol known to have a sweet taste could be observed. If the enzyme is inactive, a possible accumulation of benzyl acetate known to have a jasmine note could be observed. In the case of co-locations between 4-hydroxy acetophenone and acetophenone on chromosomes 5, 7 and 9 the enzyme transforming 4-hydroxy acetophenone into acetophenone has not been characterized. The candidate gene must have a hydroxylase function that allows the addition of the hydroxyl function on carbon number 4. Two genes (*2-nonaprenyl-3-methyl-6-methoxy-1, 4-benzoquinol hydroxylase* and *Abscisic acid 8'-hydroxylase 2*) with this function been identified close to the association zones on chromosomes 7 and 9 (Appendix 10).

The position of the most significant association zones for the same compound may be different if this compound has been detected in roasted or unroasted beans. This is the case for benzyl acetate, acetophenone, benzaldehyde, furfural and linalool (Table 3-5). This difference can be explained by the response to two different phenomena: during fermentation, the enzymes responsible for the synthesis of compounds would be activated. A "classical" synthesis would then be carried out in the bean. Whereas during roasting, the thickness of the shell or the size of the bean could play a role in the chemical conditions of the bean such as temperature or pH and thus influence the degradation of certain aromatic compounds. In that case, the detection

of association would depend also on the location of genes involved in the bean structure and size. It is also possible that the difference is due to the presence of precursors that allow the genesis of aromatic compounds during roasting.

This is not the case for all compounds. On the contrary, 2-phenylethanol dosed in roasted and unroasted beans has peaks of very close associations and there are also co-locations between acetophenone related associations dosed in roasted and unroasted beans on chromosomes 2, 6 and 9 confirming the importance of these areas in the genesis of these compounds.

The formation of an aroma as well as its perception depends on a large number of conditions. An aromatic note is generally composed of a combination of several volatile compounds at different concentrations (Pérez-Silva et al., 2006). Aromatic traits, therefore, have a high probability of being polygenic, which is consistent with the large number of associations that have been found in this study. The expression of an aromatic note also depends on the matrix in which volatile compounds are contained (Afoakwa et al., 2008). The production of these compounds by plants also depends on their environment (Baldwin, 2010). These factors therefore partly explain why large number of associations was found.

The synthesis of a flavour is therefore due to many external parameters but also the genetic background of the *T. cacao* trees (Luna et al., 2002; Afoakwa et al., 2008). Due to its multigenic determinism, the total variance of a compound is the result of many small associations, each of which would explain, a small part of the genetic variance. Once these small associations are combined, they could explain a large part of the genetic variance. In this case, some associations may contain only one associated marker, as is the case for linalool on chromosome 2. It is also possible that some associations do not cross the significance threshold and are therefore not identified. This hypothesis suggests that some associations with certain volatile compounds have not been revealed, explaining why the analysis of some compounds known to have a floral taste does not reveal an association zone as for guaiacol (R).

5.3-Role of fermentative micro-organisms in cocoa flavour synthesis

The analysis of three other compounds known to have a floral taste belonging to the family of esters did not detect zones of associations: ethyl 2-hydroxyhexanoate (R), ethylphenyl acetate (UR) and ethyl hexanoate (UR). These compounds present after fermentation and before roasting could also be synthesized by yeasts during fermentation (Soles et al., 1982). In this case, no area of association can be found as this would depend on the micro-organisms

population and not on the cocoa seeds. The non-detection of association zones can also be due partially to the pollination of the mother tree made by a mix of progenitors. While genotyping is done on the mother tree, phenotyping (volatile compound assay and sensory analysis) is done on the beans, hybrids between the mother tree and male pollinators, which could lead to a partial discrepancy between genetic and phenotypic data. Currently, it is not possible to genotype and phenotype individually each bean.

Volatile organic compounds (VOCs) produced by plants are involved in various processes and often released for defence, signalling or pollinator attraction purposes (Baldwin, 2010). VOCs belong to different biochemical families such as terpenes. They are notably involved in direct and indirect defence against insects (Martin et al., 2002) and micro-organisms (Pichersky et al., 1995). Compounds of the terpene family are recognized as a molecular signal in many interactions between plants and various other species, particularly in competition reactions, in the presence of herbivores or pathogenic microorganisms, but also the presence of beneficial insects (Langenheim, 1994; Bohlmann et al., 1998). The same is true for certain phenolic compounds such as acetophenone or 4-hydroxyacetophenone that could be involved in defence mechanisms (Parent et al., 2018), which has also been observed for furfural (Palmqvist et al., 1999; Miller et al., 2009).

During fermentation, the change in environment and chemical composition of the medium induced by yeasts and bacteria can be taken as a threat and cause the seed to react. Then, they could release VOCs to defend themselves and would be responsible for the synthesis of volatile compounds involved in fine flavour, as suggested by Sabau et al., (2006) who observed an increase in the expression of the gene coding for linalool synthase during fermentation. Also, a strong increase in the concentration of linalool, epoxylinalool and 2-phenylethanol has also been observed during fermentation in aromatic fine cocoa beans by other authors (Cevallos-Cevallos et al., 2018).

If cocoa beans use VOCs as a defence mechanism against external microorganisms such as fermentative yeasts, lactic bacteria, or acetic bacteria, some questions remain unanswered: by which mechanisms do they detect such microorganisms? Knowing that different types of yeast have been identified according to the place of fermentation (Schwan and Wheals, 2004), we can also ask ourselves whether certain types of yeast or microorganisms are more favourable to this activation. Another hypothesis is that the presence of microorganisms and the transformations they induce (change in pH, synthesis of unknown compounds in the seed, etc.)

induce the synthesis of VOCs. In this case, VOCs could be triggered in the absence of microorganisms.

5.4-Conclusions and perspectives

The perception of an aroma and the sensorial analyses is a difficult task. They, therefore, depends on a large number of conditions, including the perception threshold of aromatic molecules. The presence of a molecule is therefore not synonymous with the perception of its taste. Similarly, regions of the genome identified as being associated with the content of biochemical compounds do not mean that these compounds are involved in the flavour of cocoa. Additional analyses are necessary to validate the involvement of these molecules in the formation of taste such as gas chromatography coupled to olfactometry (GCO) analyses for example. Knowing the main molecules responsible for the floral taste as well as the mechanisms of synthesis and degradation of the compounds during fermentation and roasting could also, in the long term, allow the adaptation of the roasting process (temperatures and roasting time) to preserve the most fragile aromatic compounds. Knowledge of the biosynthesis pathway of cocoa aromatic compounds could provide a better mastering of the parameters of fermentations allowing the synthesis of these molecules.

The identification of these molecules and their biosynthetic pathway within the cocoa tree is complex. A genomic selection approach could allow early prediction of aroma traits for the search of cocoa trees having good aroma potential, especially as certain genetic variation could explain a large extend of biochemical compounds in the beans. In this case, a marker-assisted selection could be envisaged in the selected programmes to make it easier for the selection of the cocoa trees aromatic quality.

Conflict of Interest
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions
EC., CL, RGLS, conceived the experiment; JCJ, AS conducted biochemical analyses; ES carried out sensorial analyses; OF carried out DNA experiments; KC, JCJ, AS, RB, CL, FD, SA, XA analyzed data; KC, RB, CL wrote the manuscript.

Funding
The study was funded by the United States Department of State (U.S. Foreign Ministry); the U.S. Embassy, Quito; the U.S. Department of Agriculture (USDA-ARS); the MUSE Amazcacao project with the reference ANR-16-IDEX-0006.

Acknowledgement
We thank the USDA and the I-Site MUSE for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" program with the reference ANR-16-IDEX-0006.

Comme présenté ci-dessus, des analyses GWAS sur un lot d'individus appartenant à la variété de Nacional moderne ont permis de mettre en évidence un grand nombre de zones d'association en lien avec les notes florales, issues de la présence de composés volatils ou issues de notes sensorielles. L'ensemble des zones d'association en lien avec la présence de composés volatils a permis de mettre en évidence deux voies principales de biosynthèse, responsables de la synthèse de l'arôme floral chez la variété de Nacional moderne. Des gènes candidats potentiellement impliqués dans ces voies de biosynthèse ont également été identifiés.

La variété Nacional moderne fait partie des cacaoyers reconnus comme fins (riches en arômes floraux et fruités). Même si la variété Nacional moderne est principalement connue pour ses arômes floraux et épicés, elle contient également des notes fruitées (Luna et al., 2002; Rottiers et al., 2019).

Dans le chapitre suivant, est présentée l'étude du déterminisme génétique et biochimique des arômes fruités des cacaoyers de type Nacional moderne. En effet, un grand nombre de composés volatils connus pour avoir un arôme fruité (fruits secs ou fruits frais) ont été identifiés dans cette population. Grâce à des analyses GWAS, des zones d'association ont également pu être mises en évidence. Des gènes candidats ont été identifiés dans ces zones.

L'ensemble de ces résultats est présenté dans la première partie du Chapitre 3. La deuxième partie du chapitre 3 présente des compléments d'analyses préliminaires d'expression de gènes candidats, identifiés dans les analyses GWAS, présentés dans le chapitre 2 et dans le chapitre 3 partie 1. Ces résultats ont fait l'objet d'une présentation sur le déterminisme des arômes fruités du Nacional moderne, à l'occasion du congrès international du 16ème WEURMAN.

Chapitre 3: Déterminisme génétique et biochimique de l'arôme fruité de la variété de Nacional moderne

Chapitre 3: Déterminisme génétique et biochimique de l'arôme fruité de la variété de Nacional moderne

Partie 1: Révélation de nouvelles voies métaboliques impliquées dans l'arôme fruité du cacao grâce à une analyse intégrative utilisant des analyses sensorielles, de métabolomique et de GWAS.

Integration of GWAS, metabolomics, and sensorial analyses to reveal novel metabolic pathways involved in cocoa fruity aroma

Kelly Colonges[1,2,3,4*], Juan-Carlos Jimenez[5], Alejandra Saltos[5], Edward Seguine[6], Rey Gastón Loor Solorzano[5], Olivier Fouet[1,2], Xavier Argout[1,2], Sophie Assemat[3,4], Fabrice Davrieux[3,4], Emile Cros[3,4], Claire Lanaud[1,2*], Renaud Boulanger[3,4*]

1 Cirad, UMR AGAP, F-34398 Montpellier, France.

2 AGAP Institut, Univ Montpellier, Cirad, INRAE, Institut Agro, Montpellier, France.

3 Cirad, UMR Qualisud, F-34398 Montpellier, France.

4 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

5 Instituto Nacional de Investigacion Agropecurias, INIAP, Ecuador.

6 Guittard Chocolate Co./Seguine Cacao, United-States

Keywords: GWAS, cocoa, aroma, fruity note, genetic bases, biochemical bases

1-Abstract

Nacional is a variety of cocoa tree known for its "Arriba" aroma characterised mainly by fruity, floral, and spicy aromatic notes. In this study, the genetic basis of the fruity aroma of modern Nacional cocoa will be investigated. GWAS studies have been conducted on biochemical and sensorial fruity traits and allowed to identify a large number of association zones. These areas are linked to both the volatile compounds known to provide fruity flavours and present in the beans before and after roasting, and to the fruity notes detected by sensorial analysis. Five main metabolic pathways were identified as involved in the fruity traits of the Nacional population: the protein degradation pathway, the sugar degradation pathway, the fatty acid degradation pathway, the monoterpene pathway, and the L-phenylalanine pathway. Candidate genes involved in the biosynthetic pathways of volatile compounds identified in association areas were detected for a large number of associations.

2-Introduction

Theobroma cacao is a tree species belonging to the Malvaceae family (Bayer and Kubitzki, 2003). Native to the tropical rainforests of northern South America, it is the world's only source of cocoa products obtained after fermentation, drying, and roasting of cocoa beans. *T. cacao* is a diploid plant (2n=2x=20). Its small genome has been sequenced with 96.7% of the assembly anchored on all 10 chromosomes (Argout et al., 2011; Motamayor et al., 2013; Argout et al., 2017).

Cocoa is classified into two types of products: so-called standard or bulk cocoa, which has a pronounced cocoa taste, and so-called fine aromatic cocoa, which is characterised by floral and fruity notes (Sukha et al., 2008). The production of fine aromatic cocoa, therefore, represents about 5% of the global production but is no less important. Some Latin American countries produce almost exclusively fine cocoa, which is a significant source of income for them.

The *T. cacao* species showed a high genetic diversity. Currently, ten genetic groups have been identified (Motamayor et al., 2008). The most widely cultivated varieties providing fine aromatic cocoa are the Nacional, Criollo, and Trinitario. Trinitarios are hybrids between the Criollo and the Amelonado. The Amelonado variety is a variety producing mainly "standard" cocoa. The Criollo variety produces cocoa beans with mainly fruity aromas (Lachenaud and

Motamayor, 2017) but is not widely cultivated because of its low vigour and increased susceptibility to disease (Cheesman, 1944).

The Nacional variety is native to Ecuador. The Nacional variety trees currently cultivated (called modern Nacional in this paper) are the result of several generations of crossbreeding between the ancestral Nacional and the Trinitarios (Loor S. et al., 2009).

The Nacional variety is well known for its floral and spicy flavour, denominated "Arriba" flavour. It is for this reason that it is sought after by chocolate makers. It is characterised by floral and woody notes (Luna et al., 2002). In addition, Nacional is known for its low astringency and low bitterness (International Cocoa Organization, 2017). The floral aroma of Nacional has been studied and two main biosynthesis pathways have been highlighted as being mainly responsible for this floral aroma: the terpene biosynthesis pathway and the L-phenylalanine degradation pathway (Ziegleder, 1990; Colonges et al., 2021b). The aroma of the modern Nacional probably contains floral and fruity aromas that could be the legacy of crossbreeding with Trinitarios (hybrid trees between Criollo and Amelonado) (Loor S. et al., 2009; Rottiers et al., 2019). Loor S. et al., (2009), have demonstrated this hybrid nature using molecular markers. This genetic mixing has led to a dilution of the Arriba flavour (Loor S. et al., 2009; Boza et al., 2014; Beckett et al., 2017). From 1940 onwards, surveys were carried out on the Ecuadorian coast to collect Nacional-type cocoa trees to preserve their genetic resources. These collections were placed in two main experimental stations: the Experimental Tropical Station of Pichilingue (EET-P) of INAP (Instituto Nacional de Investigaciones Agropecurias) and the Cocoa Flavour Centre of Tenguel (CCAT). The fine aromas of the modern Nacional are therefore a blend of the aromas of the ancestral Nacional and the aromas of the Criollo and Amelonado ancestors. Here, this study will be focused on fruity aromas. The fruity aromas are composed of two main classes of aromas: dried fruit aromas and fresh fruit aromas. Dried fruit aromas are mainly due to pyrazines, which mainly appear after roasting. During this stage of transformation, heating induces a large number of chemical reactions, including the Maillard reaction. This reaction leads to the synthesis of pyrazines from amino acids and reducing sugars (Arnoldi et al., 1988).

Different families of volatile compounds such as alcohols, esters, and ketones represent fresh fruit aromas. During alcoholic fermentation, yeasts synthesize short-chain alcohols (e.g., propan-1-ol; butan-1-ol, 2-methyl-propan-1-ol, pentan-1-ol, hexan-1-ol, heptan-1-ol, octan-1-

ol) from simple sugars. These alcohols can also be synthesised from a pyruvate molecule, present in all organisms (Sun et al., 2015). Aldehydes are formed through the oxidation of alcohol. This is the case, for example, in the Lilac flower (Syringa vulgaris L.) where the "lilac alcohols" are oxidised to give "lilac aldehydes". The esters are the result of the esterification reaction that is a reversible chemical reaction. Esterification allows the synthesis of an ester molecule and a water molecule from an alcohol molecule and an acid molecule (Aranda et al., 2008).

The objectives of this study are to decipher the genetic, genomic, and biochemical bases of the fruity aromas of the Nacional cocoa variety. A GWAS (Genome-Wide Association Study) study was carried out to discover the areas of the genome responsible for the fruity flavour. It involved phenotyping data related to volatile compounds, potentially linked to the fruity taste, as well as sensory analysis data. To refine the determinants of the fruity aroma of the modern Nacional variety, candidate genes potentially involved in the biosynthesis pathways of fruity compounds were sought.

3-Material and methods

3.1-Vegetal material

The plant material used for these experiments were composed of a collection of 152 cocoa trees from Ecuador conserved in the Pichilingue experimental station of the “Instituto Nacional de Investigaciones Agropecurias” (INIAP) and the “Coleccion de Cacao de Aroma Tenguel” (CCAT) of Tenguel. This population represents the Nacional variety currently grown in Ecuador and has been described by Loor S. (2007).

3.2-Fermentation processes

Micro-fermentations of cocoa beans were carried out in a wooden box in the most homogeneous way possible with a homogeneous cocoa mass. The process lasted 4 days with two turns at 24 and 72 hours after the beginning of the fermentation. Each clone sample (152) was placed in a protective laundry bag and micro-fermented in a cocoa mass. After fermentation, the samples were put in a dry place. They were considered dried when their moisture content was below 8%.

Actual roasting conditions were 120°C x 22 minutes. Each bean sample was adjusted from this basis roast using the validated ISCQF, (2020) moisture and bean size (bean weight) from that source. Times measured from -2C of set point.

3.3-Sensorial Analysis

146 individuals were characterized by sensory analyses based on blind tastings carried out on three repetitions per sample by Edward Seguine. The tastings were carried out on cocoa liquor. The cocoa liquor corresponds to merchant cocoa (dried fermented beans) which have been roasted and crushed. Thirteen fruity notes were judged with a score ranging from zero (no fruity notes detected) to ten (intense note detected) according to ISCQF, (2020) protocol (Appendix 11).

3.4-Volatile compound analysis by GC-MS

GC/MS analysis were conducted according to the condition described by Assi-Clair et al., (2019).

3.5-Statistical analysis

PCA analysis and visualization were made with “mixOmics” R package. Calculation of correlation was made with “agricolae” R package and visualization of correlation matrix with “corrplot” R package.

3.6-DNA extraction protocol

DNA extraction was conducted according to Risterucci et al. (2000) protocol.

3.7-Genotyping by SSR

This population was genotyped using SSR markers by Loor S. (2007).

3.8-Genotyping by sequencing (GBS)

DNA samples were genotyped by sequencing (GBS) using DArTseq (Diversity Arrays Technology Sequencing) technology (Kilian et al., 2012). Reads were aligned with the V2 sequence of the Criollo genome (Argout et al., 2017). Markers with unknown locations were discarded for analysis.

3.9-Association mapping

GWAS using SNP and SSR was conducted according to Colonges et al., (2021) protocol.

4-Results

4.1-Characterisation of the traits studied

4.1.1-Sensory trait analysis.

Thirteen fruity notes and notes associated with pyrazine notes (roasted degree, cocoa, browned flavour, Carmal browned sugar) were evaluated during the sensory analysis carried

out on the cocoa liquors. All of these thirteen sensory traits were used to carry out this study (Appendix 11).

The results of a PCA for fruity notes (sensory analysis traits) showed a continuous variation within the population (figure 22). The aromatic notes of brown flavour, fruity browned dried fruit, and nutty mainly define axis 1. Aromatic notes fruity acidity, Fruity-Citrus, and Fruity-Dark tree fruit mainly define axis 2.

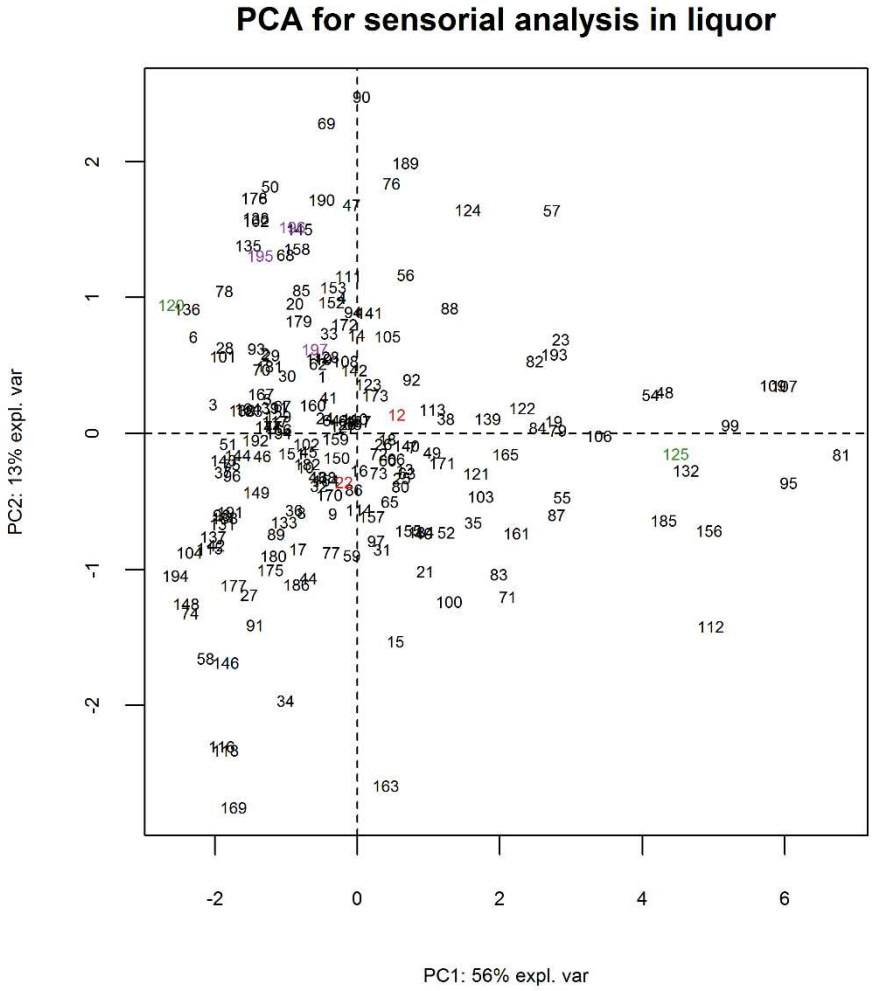


Figure 22: PCA of sensorial analysis results related to fruity notes detected in liquor.
Black numbers represent the population of modern Nacional variety. Green numbers represent the two individuals closest to Amelonado's ancestor. Red numbers represent the two individuals closest to the Criollo ancestor. Purple numbers represent SNA604 (Nacional ancestor) and two individuals closest to the Nacional ancestor.

Analyses of correlations between the different sensory traits related to the fruity taste did not show strong negative correlations. Three strong positive correlations were detected; a correlation of 0.8 to 1 between the hazelnut note and browned flavour characteristics, two correlations of 0.6 to 0.8 between the browned dried fruit note and the nutty note on the one hand and the browned flavour note, on the other hand (Figure 23).

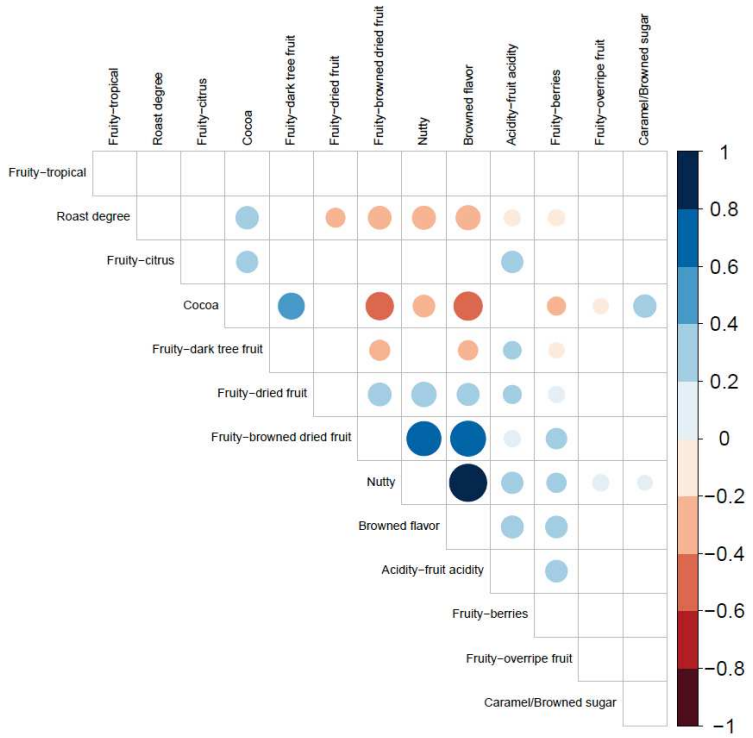


Figure 23: Significant correlation matrix of sensorial analysis.
Correlation matrix between the sensorial profiles determined in cocoa liquor. The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient. The areas of circles correspond to a p-value of correlation coefficients. The p-value threshold for a significant correlation is 0.05. The different shades of blue represent a positive correlation coefficient while the different shades of red represent a negative correlation coefficient. The intensity of the colour depends on the strength of the R2 correlation coefficient. The scale on the right indicates the interpretations of different colours.

4.1.2-Biochemical traits analysis

The identification of volatile compounds was carried out in cocoa beans before and after roasting. A total of one hundred and sixty-one volatile compounds were identified. Thirty-five of them are known to have fresh or dried fruity notes (Table 8). These compounds are mainly synthesised through five biosynthesis pathways: proteins degradation pathways (required for the production of pyrazines via the Maillard reaction), the fatty acid degradation pathway, the simple sugar degradation pathway, the L-phenylalanine degradation pathway, and the monoterpene biosynthesis pathway.

Table 8: List of biochemical compounds related to fruity traits and used for the GWAS analysis of unroasted (UR) and roasted (R) beans.

UR	R	Biochemical traits	Compound family	Path	Aroma
X	X	Acetic acid	Acid	FA	Vinegar (Ferreira et al., 1997; Mahajan et al., 2004)
X	X	Pentanoic acid	Acid	FA	Unpleasant, acre, cheese, sweat (Jezussek et al., 2002; Mahajan et al., 2004; Genovese et al., 2007; Kalua et al., 2007)
	X	Hexanoic acid	Acid	FA	Fatty acid, cheese, sweat, rancid, sweet and vinegar (Ferreira et al., 1997; Jezussek et al., 2002; Mahajan et al., 2004; Colahan-Sederstrom and Peterson, 2005; Perestrelo et al., 2006; Genovese et al., 2007)
X	X	Octanoic acid	Acid	FA	Fatty acid, rancid, cheese, waxy (Ferreira et al., 1997; Karagül-Yüceer et al., 2003; Colahan-Sederstrom and Peterson, 2005; Perestrelo et al., 2006; Genovese et al., 2007)
X	X	Nonanoic acid	Acid	FA	Acid (Colahan-Sederstrom and Peterson, 2005)
X	X	Ethanol	Alcohol	SS	Alcohol (Kalua et al., 2007)
	X	Pentan-1-ol	Alcohol	FA / SS	Fruity (Kalua et al., 2007)
X	X	Pentan-2-ol	Alcohol	FA / SS	
	X	Hexan-2-ol	Alcohol	SS	Green grass (Guichard et al., 2003)
X	X	Heptan-2-ol	Alcohol	SS	
X		Nonan-2-ol	Alcohol	SS	Cucumber (Arn and Acree, 1998)
X		2-methylbutan-1-ol	Alcohol	SS	
X	X	3-methylbutan-1-ol	Alcohol	SS	Plastic, Fuel oil, whiskey characteristic pungent (Guichard et al., 2003)
X		2-methylbutan-2-ol	Alcohol	SS	
	X	2-methyl-but-3-en-2-ol	Alcohol	SS	
	X	2-ethylhexan-1-ol	Alcohol	SS	
	X	Butan-2,3-diol	Alcohol	FA / SS	Fruit, onions (Arn and Acree, 1998)
	X	Butane-1,3-diol	Alcohol	FA / SS	
	X	Cis-hept-4-en-1-ol	Alcohol	SS	
X	X	Benzyl alcohol	Alcohol	L-phe	Fruity (Ito et al., 2002)
X	X	Acetaldehyde	Aldehyde	SS	Acre, ether (Arn and Acree, 1998; Wang et al., 2014)
	X	2-methylpropanal	Aldehyde	SS	Acre, malted, green (Arn and Acree, 1998)
X	X	2-methylbutanal	Aldehyde	SS	Malted (Fickert and Schieberle, 1998)
X	X	3-methylbutanal	Aldehyde	SS	Stimulant, malted (Kumazawa and Masuda, 2002)
X		Nonanal	Aldehyde	Other	Orange-like, floral, soapy (Kumazawa and Masuda, 2002; Karagül-Yüceer et al., 2003; Mahajan et al., 2004)
	X	2-phenylbut-2-enal	Aldehyde	Other	
	X	5-methyl-2-phenylhex-2-enal	Aldehyde	SS	Cacao (Garg et al., 2018)
X	X	Benzaldehyde	Aldehyde	L-phe	Bitter, cherry, almond, fruity (Perestrelo et al., 2006; Pham et al., 2008; Wang et al., 2014)
X	X	Methyl acetate	Ester	SS	Ester, green (Garg et al., 2018)
	X	Methyl octanoate	Ester	FA	Fruity, orange, wax, wine (Garg et al., 2018)
X	X	Ethyl acetate	Ester	FA / SS	Fruity, solvent, apple, sweet (Ferreira et al., 1997; Karagül-Yüceer et al., 2003; Genovese et al., 2007)
X	X	Ethyl propanoate	Ester	FA / SS	Fruity, strong, strawberry (Kalua et al., 2007)
	X	Ethyl butanoate	Ester	FA	Strawberry, fruity, sweet, kiwi, pineapple (Larsen and Poll, 1992; Ferreira et al., 1997; Genovese et al., 2007; Wang et al., 2014)
	X	Ethyl hexanoate	Ester	Other	Strawberry, green apple, fruity and floral (Larsen and Poll, 1992; Ferreira et al., 1997; Genovese et al., 2007; Wang et al., 2014)
X		Ethyl 3-hydroxyhexanoate	Ester	SS	Strawberry (Genovese et al., 2007)
X	X	Ethyl octanoate	Ester	FA / SS	Sweet, fruity, fresh, pineapples (Perestrelo et al., 2006; Genovese et al., 2007)
	X	Ethyl dodecanoate	Ester	FA	Floral, fruity, leaf (Garg et al., 2018)
X		Ethyl hexadecanoate	Ester	SS	
X		Ethyl benzoate	Ester	L-phe	Ripe fruit, fruity, violets, candy (Ferreira et al., 1997)
X		Ethyl 2-methylpropanate	Ester	FA	Fruity, sweet, rubber (Schieberle et al., 1990; Arn and Acree, 1998)
	X	Ethyl 2-methylbutanoate	Ester	FA	Green apple, fruity, red fruit (Baek et al., 1997; Genovese et al., 2007)
X		Butyl benzoate	Ester	FA	
	X	Propyl acetate	Ester	Other	Celery, floral, pear, red fruit (Garg et al., 2018)
X		Hexyl acetate	Ester	Other	Fruity, pear, sweet and floral (Guichard et al., 2003; Wang et al., 2014)
X		2-methylpropyl acetate	Ester	FA	Strawberry (Aznar et al., 2001)
X	X	1-methylbutyle acetate	Ester	SS	
X	X	3-methylbutyle acetate	Ester	FA / SS	Banana (Guichard et al., 2003; Genovese et al., 2007)
X		Butan-2,3-diyl diacetate	Ester	SS	
X		Meso butan-2,3-di-yl diacetate	Ester	SS	
X	X	3-methyl-but-2-enyl acetate	Ester	SS	Jasmin, banana, fresh, ripe, heliotrope, sweet, fruity, balsam, lavender (Garg et al., 2018)
X	X	1-methylhexyle acetate	Ester	SS	
	X	1-phenylethyl acetate	Ester	L-phe	Fruity (Garg et al., 2018)
X	X	Furfural	Furan	Other	Incense, fruity, flowery, roasted, sweet and almond (Ferreira et al., 1997; Colahan-Sederstrom and Peterson, 2005; Wang et al., 2014)
	X	Acetone	Ketone	?	Acre (Garg et al., 2018)
X		Butan-2-one	Ketone	SS	Fruity, ethereal (Kalua et al., 2007)
X	X	3-hydroxy-butan-2-one	Ketone	FA / SS	Butter, cream (Arn and Acree, 1998)
X	X	Pentan-2-one	Ketone	FA / SS	Ether, fruity (Arn and Acree, 1998)
	X	Pentan-2,3-dione	Ketone	FA / SS	Cream, butter (Arn and Acree, 1998)
X		Hexan-2-one	Ketone	FA / SS	Ether (Arn and Acree, 1998)
X	X	Heptan-2-one	Ketone	FA / SS	Sweet, fruity (Kalua et al., 2007)
	X	Octan-2-one	Ketone	FA / SS	Green, mouldy (Kalua et al., 2007)
	X	Nonan-2-one	Ketone	FA / SS	Warm milk, syrup, green (Arn and Acree, 1998)
X	X	Acetophenone	Ketone	L-phe	Acacia honey, floral and fruity (Genovese et al., 2007; Wang et al., 2014)
X	X	Gamma–butyrolactone	Lacton	FA?	Caramel, fat, sweet, creamy, oily (Garg et al., 2018)
	X	Methylpyrazine	Pyrazine	P	Popcorn (Arn and Acree, 1998)
	X	2,3-dimethylpyrazine	Pyrazine	P	Hazelnut, cooked (Mahajan et al., 2004; Wang et al., 2014)
	X	2,6-dimethylpyrazine	Pyrazine	P	Cooked meat, cooked rice, hazelnut (Mahajan et al., 2004; Pham et al., 2008; Wang et al., 2014)
	X	2,3,5-trimethylpyrazine	Pyrazine	P	Hazelnut, roasted (Mahajan et al., 2004; Wang et al., 2014)
X	X	2,3,5,6-tetramethylpyrazine	Pyrazine	P	Cooked (Wang et al., 2014)
	X	2,3-dimethyl-5-ethylpyrazine	Pyrazine	P	
	X	2-ethyl-5-methylpyrazine	Pyrazine	P	Fruity, green, sweet (Arn and Acree, 1998; Garg et al., 2018)
	X	2-ethyl-6-methylpyrazine	Pyrazine	P	Green, hazelnut, grilled (Garg et al., 2018)
	X	Trimethylpyridine	Pyridine	P	
X	X	Linalool	Terpene	M	Floral, lemon peel, orange blossom (Ferreira et al., 1997; Genovese et al., 2007)
X		Linalool cis pyranic oxide	Terpene	M	Fruity, citrus, green (Arn and Acree, 1998; Ito et al., 2002)
X		Linalool trans furanic oxide	Terpene	M	Citrus, leafy, floral (Arn and Acree, 1998; Ito et al., 2002)

UR: Unroasted beans; R: Roasted beans; FA: Fatty acid degradation pathway; SS: Simple Sugar degradation pathway; M: Monoterpene biosynthesis ; P: Pyrazine biosynthesis; L-phe: L-phenylalanine degradation pathway; **X**: detected by GC-MS and association was detected in link with this compound; **X:** detected by GC-MS but any association was detected in link with this compound

The result of a PCA for volatile compounds in unroasted beans involved in fruity notes shows a continuous variation within the population (figure 24). The compounds 1-methylbutyl acetate, meso-butan-2,3-diyl acetate, and ethyl acetate mainly define axis 1. The compounds 3-methylbutyl acetate, heptan-2-one, and ethanol mainly define axis 2.

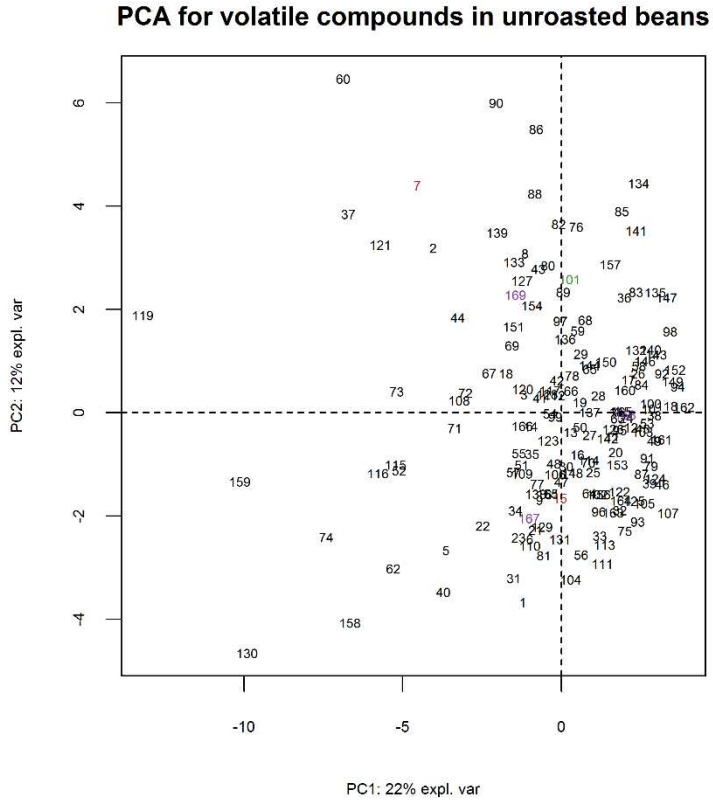


Figure 24: PCA of biochemical compounds detected related to fruity notes detected in unroasted beans.
Black numbers represent the population of modern Nacional variety. Green numbers represent the two individuals closest to Amelonado's ancestor. Red numbers represent the two individuals closest to the Criollo ancestor. Purple numbers represent SNA604 (Nacional ancestor) and two individuals closest to the Nacional ancestor.

Correlation analyses were carried out between the different volatile compounds related to fruity notes. Twenty-seven strong positive correlations (greater than or equal to 0.8) were detected between several volatile compounds from unroasted or roasted beans. Sixteen strong positive correlations were between compounds involved in the fatty acid and simple sugar degradations. Only one strong positive correlation was between the two pyrazines. Four strong positives correlations were between compounds involved in fatty acid and simple sugar degradation pathway and pyrazines. Seven strong negative correlations were also detected (Figure 25) of which five were detected between compounds involved in fatty acid and simple sugar degradation. The two other negative correlations are between pyrazine and pentan-2-ol (R) or 1, 2, 5-trimethylbenzene (R). No strong correlation was detected between biochemical compounds and sensory analysis data (Appendix 12).

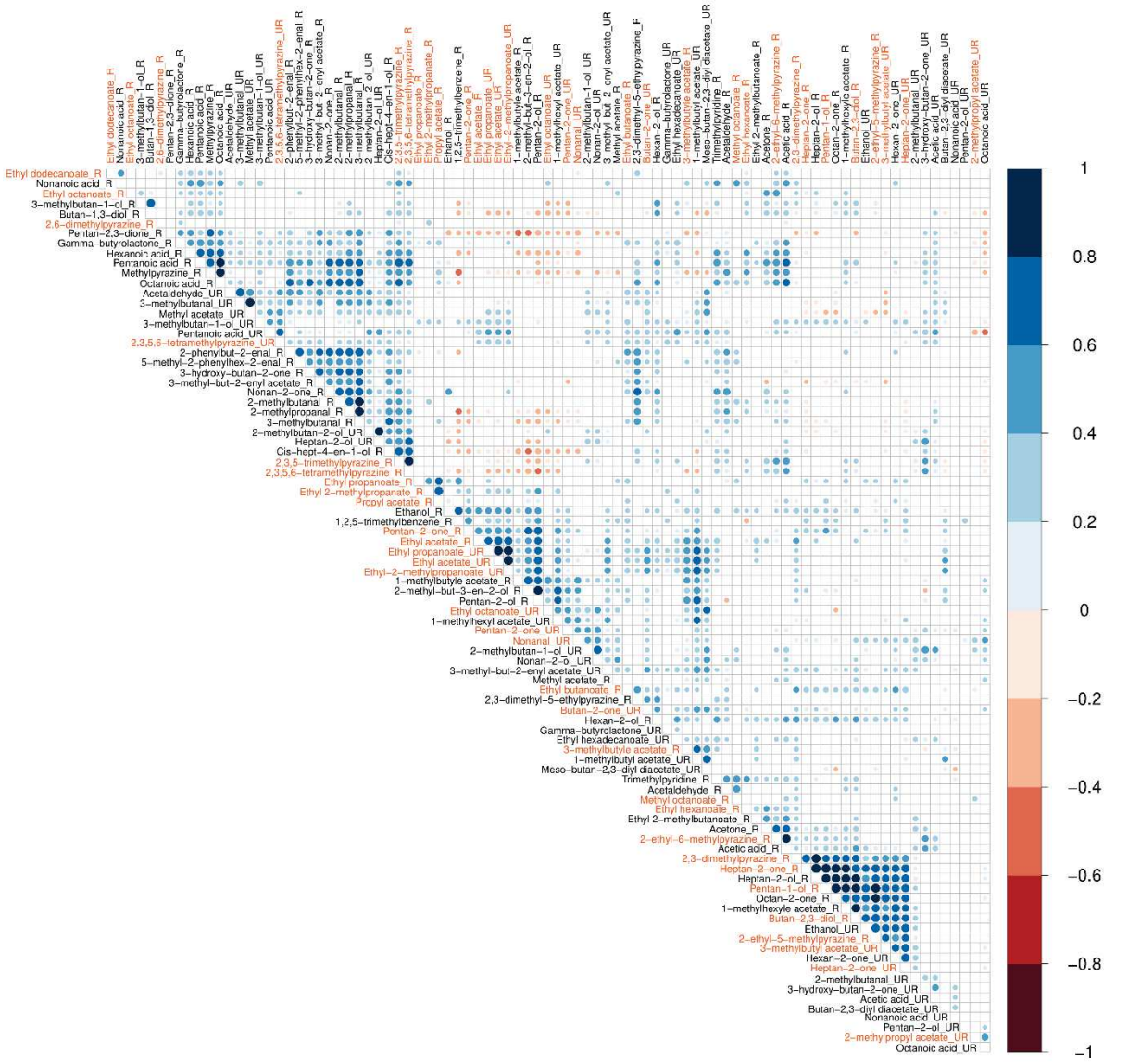


Figure 25: Significant correlation matrix of biochemical compounds involved in the synthesis of a fruity note.
Correlation matrix between the biochemical compounds measured in unroasted (UR) and roasted beans (R). The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient. The areas of circles correspond to a p-value of correlation coefficients. The p-value threshold for a significant correlation is 0.05. The different shades of blue represent a positive correlation coefficient while the different shades of red represent a negative correlation coefficient. The intensity of the colour depends on the strength of the R2 correlation coefficient. The scale on the right indicates the interpretations of different colours.

4.2-Genome Wide Association Study (GWAS)

The genetic diversity of the population and its structure has been described by Loor S. et al., (2009) using SSR markers and by Colonges et al., (2021), using SNP markers. The population has a high rate of heterozygosity.

The choice of markers for the genotyping data and the determination of the confidence interval was made using the same method as described by Colonges et al., (2021).

In this study two sets of genotyping data were selected. An analysis was carried out with a panel of markers with no missing data and with a minor allele frequency (MAF) greater than 5%, the results of which are annotated MAF5. This dataset contains 6541 SNP markers. A second analysis was performed with a panel of markers having a representation of each genotype higher than 5%, the results of which are annotated G7. This dataset contains 5195 SNP markers. The confidence intervals of the association zones were calculated based on the haplotype blocks calculated by Haploview. All detected association areas are available in Appendix 13.

4.2.1-Identification of significant associations for sensorial traits

Out of all the associations, only 22 relate to fruity sensory data. With thirteen sensory perceptions related to fruity traits, associations could be detected for only three of them: "Fruity-Dark Tree Fruit", "Fruity-Dried fruit" and "Fruity-Berries". Only one association could be detected for three of the four characters (Table 9). Nineteen areas of associations could be detected for the note "Fruity-Dark Tree Fruit". These associations were detected on chromosomes 1, 2, 4, 7, 8, and 10. The strongest association detected for the note "Fruity-Dark Tree Fruit" is also the strongest association for all fruit notes and explains 24% of the variation in this trait.

Table 9: Most significant association detected for each sensory fruity note.

CH	Position of the association peak	N° hap. bloc	Fruity note detected	*p-value* of the strongest association	Explanation rate of the trait of the strongest association	Total number of associations for the character
1	2 430 002 bp	4	Fruity-Dark Tree Fruit	6,33E-09	24%	1/19
2	32 194 678 bp	61	Fruity-Dried fruit	3,19E-06	16%	1
7	4 613 061bp	19	Fruity-Berries	1,18E-06	15%	1

CH: chromosome; hap. : haplotypic; R: roasted beans; UR: Unroasted beans

4.2.2-Identification of significant associations for biochemical traits

The GWAS analysis revealed 480 areas of association for biochemical compounds related to the fruity taste. All the associations found are reported in Appendix 13.

For all the compounds for which significant associations have been detected, three major biosynthesis pathways seem to emerge. The protein degradation pathway (required for the production of pyrazines via the Maillard reaction), the sugar degradation pathway, and the fatty

acid degradation pathway. The two degradation pathways (sugars and fatty acids) are often linked (figure 26). They will therefore be presented jointly.

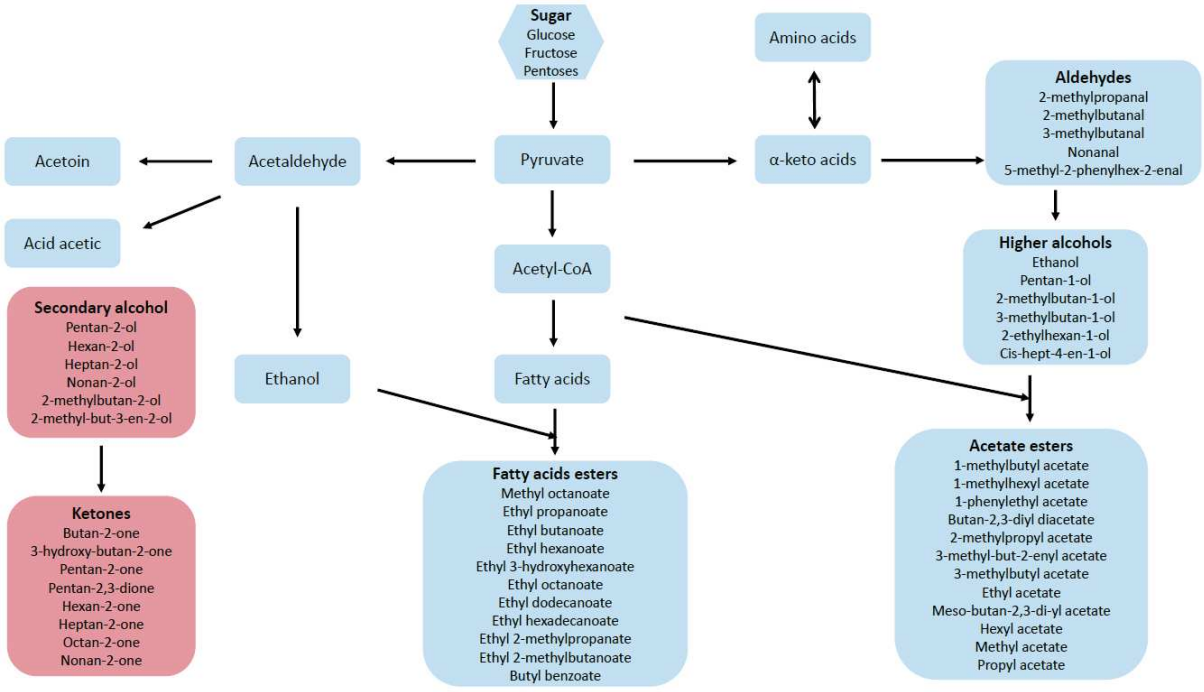


Figure 26: Scheme representation of fatty acid and sugar degradations adapted to Swiegers et al., (2005); Dzialo et al., (2017).)

The areas of significant associations have been mapped to visualize their locations and co-locations. Two maps have been produced. A map showing the associations of traits related to pyrazine production (appendix 14). A second map showing the areas of associations related to the compounds known to be involved in the degradation pathway of sugars or fatty acids (appendix 15). These biosynthesis pathways often have common compounds. Some results are different according to the methods (GLM and MLM) or according to the sorting of markers (MAF5 or G7) or even between the type of markers (SNP or SSR).

Repeatable results between methods appear to be the most conclusive.

4.2.3-Significant associations were identified for the biochemical compounds involved in the pyrazine production pathway.

Of the 74 volatile compounds related to the fruity notes, eight pyrazines were identified by GC-MS in the roasted beans and only one in the unroasted beans (2, 3, 4, 5-tetramethylpyrazine). Two hundred and sixty-eight association zones have been detected in relation to pyrazines (Table 10). Association zones were detected on chromosomes 1, 2, 3, 9, and 10. The most significant association zone for 2, 3, 5-trimethylpyrazine (R) co-locates with the most significant association zone detected for methylpyrazine (R). Six of the most

significant association zones for these compounds are located outside the calculated haplotypic blocks.

Table 10: Most significant associations for biochemical compounds related to pyrazine pathway.

CH	Position of the association peak	N° hap. bloc	Traits	*p-value* of the strongest association	Explanation rate of the trait for the strongest association	Associations detected
1	7,081,063	26	2,3,5-trimethylpyrazine_R	2.15E-08	25%	17
1	7,081,063	26	Methylpyrazine_R	2.57E-08	25%	38
2	8,389,914	NA	2,3-dimethylpyrazine_R	1.99E-11	29%	37
2	11,231,613	NA	2-ethyl-5-methylpyrazine_R	7.80E-16	49%	94
3	31,136,388	NA	2,6-dimethylpyrazine_R	4.67E-07	25%	6
4	23,199,282	39	2,3,5,6-tetramethylpyrazine_R	5.19E-09	24%	18
4	30,166,293	70	2,3-dimethyl-5-ethylpyrazine_R	1.35E-08	55%	50
9	29,324,700	NA	2,3,5,6-tetramethylpyrazine_UR	9.14E-07	18%	2
10	3,974,481	19	2-ethyl-6-methylpyrazine_R	4.45E-07	21%	6

CH: chromosome; hap. : haplotypic; R: roasted beans; UR: Unroasted beans

The most significant association found for pyrazines is the association zone associated with 2-ethyl-5-methylpyrazine (R). The variation of this trait is explained at 49% by the genetic variation located in this association zone.

The map showing the significant associations for the traits linked to the pyrazine production pathway (appendix 14) shows interesting results because, in a large number of associations represented, co-locations can be observed between the associations detected for different pyrazines. For example, on chromosome 1 between 2, 3-dimethyl-5-ethyl pyrazine (R) and methylpyrazine (R) or between 2, 3, 5, 6-tetramethylpyrazine (R); 2, 3, 5-trimethylpyrazine (R) and 2, 3-dimethyl-5-ethylpyrazine (R).

4.2.4-Co-locations between biochemical compounds and sensorial traits

Six co-locations between pyrazine-related associations and those related to sensory notes could be observed. Four are present on chromosome 1: two co-locations between the note "Fruity- Dark Tree fruit", methylpyrazine (R) and 2,3-dimethyl-5-ethylpyrazine (R); one between the note "Fruity- Dark Tree fruit" and methylpyrazine (R); one between the note "Fruity- Dark Tree fruit" and 2,3-dimethyl-5-ethylpyrazine (R). A co-location is present on chromosome 4 between the note "Fruity- Dark Tree fruit", 2,3,5,6-tetramethylpyrazine (R), and 2,3,5-trimethylpyrazine (R). A co-location is present on chromosome 10 between the note "Fruity- Dark Tree fruit" and methylpyrazine (R). Only 2,3,5,6-tetramethylpyrazine (R) and 2, 3, 5-trimethylpyrazine (R) are known to have fruity notes. It is, therefore, possible that other compounds explain the “Fruity- Dark Tree fruit” note.

4.2.5-Significant associations were identified for the biochemical compounds involved in the degradation of fatty acid and simple sugar pathway.

Four hundred and eighty associations were detected in connection with the compounds involved in the degradation pathways of sugars and fatty acids (appendix 16). Some of the most significant associations are located in the same haplotypic block but with a different position of the peak of the associations (table 11). Some peaks of the most significant associations co-locate (table 11).

Table 11: Co-localisations between compounds involved in the fatty acid and sugar degradations pathways or/and in the L-phenylalanine degradation pathway.

Chromosome	Position	N° of Haplotypic bloc	Traits in co-localization
1	4 498 932 4 522 166	14	2-methylbutanal (R) 5-methyl-2-phenylhex-2-enal (R)
1	6 281 788 6 448 063 6 448 063	23	hexan-2-ol (R) 2-phenylbut-2-enal (R) phenyla-R
1	6 855 567 7 081 063	26	pentatonic acid (R) octanoic acid (R)
1	36 447 062 36 447 686	91	3-methylbutan-1-ol (R) acetone (R)
2	8 389 914	NA	Heptan-2-ol-R Heptan-2-one-R Octan-2-one-R
2	11 231 613	NA	3-Methyl-butyl acetate-UR 1-methylhexyl acetate -R Ethanol-UR Pentan-1-ol-R
5	27 513 744	45	1-methylhexyl acetate (UR) ethyl benzoate (UR)
5	38 411 373	87	ethyl butanoate (R) hexan-2-one (UR)
6	25 201 654 25 355 204	65	pentan-2,3-dione (R) 3-hydroxy-butan-2-one (UR)
7	5 596 523 5 607 833 5 607 833 5 607 833	24	nonan-2-ol (UR) 2-methylbutan-1-ol (UR) meso-2,3-butan-di-yl diacetate (UR) methyl acetate (UR)
7	10 459 413	31	hexyl acetate (UR) pentanoic acid (UR)
9	1 099 704	5	2-ethylhexan-1-ol (R) 2-methylpropanal (R)
9	37 554 266 37 557 289	54	methyl acetate (R) nonanal (UR)

R: roasted beans; UR: Unroasted beans

Some associations outside the calculated haplotypic blocks present co-localizations, this is the case for the most significant associations detected for heptan-2-ol (R), heptan-2-one (R), octan-2-one (R); the most significant associations detected for 2 and 3-Methyl-butyl acetate (UR), 1-methylhexyl acetate (R), ethanol (UR) and pentan-1-ol (R).

4.2.6-Co-locations between biochemical compounds and sensorial traits

Nine co-locations between associations linked to fruity notes and associations linked to volatile compounds were detected (Table 12) of which seven are between the fruit note "Dark tree fruit", one with the dried fruit note and one with the berries note.

Table 12: Table of co-localization between sensory trait and biochemical compounds involved in fatty acid and simple sugar degradation

Chr	Positions	N° hap. Bloc	Sensory trait	Biochemical Trait
1	2 368 915 2 445 782	4	Dark Tree fruit	2-methylbutan-2-ol (UR), 2-phenylbut-2-enal (R), 3-methylbutanal (R), octanoic acid (R) and ethyl butanoate (R)
1	2 694 498 2 793 273	6	Dark Tree fruit	2-phenylbut-2-enal (R), 3-methylbutanal (R) and ethyl butanoate (R)
1	3 083 032 3 398169	7	Dark Tree fruit	2-phenylbut-2-enal (R), 3-methylbutanal (R), octanoic acid (R), pentanoic acid (R), ethyl butanoate (R), 3-hydroxy-butan-2-one (R) and gamma-butyrolactone (R)
1	3 721 132 4 022 340	9	Dark Tree fruit	2-phenylbut-2-enal (R), 3-methylbutanal (R), 1-methylhexyl acetate (R), propyl acetate (R), octanoic acid (R), pentanoic acid (R), ethyl butanoate (R), gamma-butyrolactone (R), heptan-2-ol (UR), heptan-2-one (R) and pentan-1-ol (R)
1	4 114 978 4 134 570	10	Dark Tree fruit	octanoic acid (R), pentanoic acid (R), 2-phenylbut-2-enal (R), 3-methylbutanal (R) and ethyl dodecanoate (R)
2	31 957 983 32 194 678	61	Dried fruit	Methyl acetate (UR) and pentan-2-ol (UR)
4	22 748 126 23 333 424	39	Dark Tree fruit	2 and 3-methylbutyl acetate (UR), 2-methylbutan-2-ol (UR), 2-phenylbut-2-enal (R), 3-hydroxy-butan-2-one (UR), pentanoic acid (R), butan-2,3-diol (R), heptan-2-ol (R), heptan-2-one (UR), octan-2-one (R), pentan-1-ol (R) and pentan-2,3-dione (R)
7	4 613 061	19	Berries	Ethyl octanoate (R)
10	6 167 221	NA	Dark tree fruit	3-methylbutanal (R) and pentanoic acid (R)

N° co-loc.: number of co-localization; Chr: Chromosome; N° hap. Bloc: number of the haplotypic block; R: roasted beans; UR: Unroasted beans

4.3-Involvement of monoterpenes and L-phenylalanine pathway compounds in the fruity notes of cocoa

Two volatile compounds with fruity notes belonging to the monoterpene biosynthesis have been identified in this population of cocoa: the linalool cis pyranic oxide, known for its citrus note, and the linalool Trans furanic oxide known for its floral and citrus notes. Respectively, twenty-seven and twenty-nine association area was detected related to these two compounds.

Five volatile compounds known for fruity notes belonging to the L-phenylalanine degradation pathway have been identified: benzaldehyde, benzyl alcohol, 1-phenylethyl acetate, ethyl benzoate, and acetophenone. Two associations' zones were identified for benzaldehyde (UR) and seventy-two for benzaldehyde (R), known for its cherry and bitter almond notes. Forty-two associations' zones were identified related to benzyl alcohol (UR), seventy-three for 1-phenylethyl acetate (R), two for ethyl benzoate (UR), thirty-six for acetophenone (UR), and forty for acetophenone (R), known for their fruity notes.

4.4-Candidate genes are potentially involved in the formation of the dried fruity aroma

Across the pyrazine association zones, 100 candidate genes were identified (appendix 17). 94 genes are identified as being involved in the synthesis of precursors of the Maillard reaction (amino acids and reducing sugars). These genes are therefore involved in either protein degradation or sugar degradation (appendix 14).

Of the 100 genes, 30 could be involved in the degradation of proteins, sugar, and fatty acid (table 13), including 25 genes coding for the Alpha/Beta hydrolase family of proteins known to have various functions, including peptidase function (Holmquist, 2000; Mindrebo et al., 2016). In addition, five genes coding for enzymes with oxidase/reductase or hydrolase functions.

Table 13: Synthesis of gene functions found in association zones linked to pyrazine compounds required for the production of Maillard precursor.

Number of genes	Gene function	General function
25	Alpha/Beta hydrolase family	Proteins, Fatty acid, Sugars degradations
17	Protease function	Proteins degradation
16	Peptidase function	Proteins degradation
11	Glutathione S-transferase	Proteins degradation
5	Oxidase/reductase or hydrolase function	Proteins, Fatty acid, Sugars degradations
4	Proteinase function	Proteins degradation
3	Aminotransferase function	Proteins degradation
3	Amino acid decarboxylase function	Proteins degradation
1	Delta-1-pyrroline-5-carboxylate	Proteins degradation
1	Anthranilate synthase	Proteins degradation
1	Bifunctional aspartokinase/homoserine dehydrogenase	Proteins degradation
1	Lactoylglutatione lyase	Proteins degradation
1	Peptidyl-prolyl cis-trans isomerase	Proteins degradation
1	Protein S-acyltrasferase 16	Proteins degradation
1	Peptide-N4-(N-acetyl-beta-glucosaminyl) asparagine amidase	Proteins degradation
1	Endoglucanase	Sugars degradation
1	Xyloglucan endotransglucosylase/hydrolase	Sugars degradation
1	Inositol oxygenase 1	Sugars degradation
4	Carboxylesterase	Fatty acid degradation
1	Isoprenylcysteine alpha-carbonyl methylesterase	Fatty acid degradation
1	Phospholipase	Fatty acid degradation

Of the 100 genes, 61 are involved in protein degradation (table 13). These genes could therefore be at the origin of the synthesis of amino acids, necessary for the Maillard reaction, by degrading existing proteins.

Among the 100 genes, three are involved in the degradation of sugars, and six are involved in fatty acid degradation (table 13). These genes could be at the origin of the synthesis of simple sugars also necessary for the Maillard reaction.

4.5-Candidate genes potentially involved in the formation of the fresh fruity aroma

4.5.1-Fatty acid and sugar degradation pathways

In the set of association zones linked to volatile compounds (acids, alcohols, aldehydes, esters, ketones) from the degradation pathways of sugars and fatty acids, 227 candidate genes were identified (appendix 15, table 14).

Table 14: Synthesis of gene functions found in association zones linked to compounds involved in Fatty acid or simple sugar degradation

Number of genes	Gene function	General function
1	Branched-chain-amino-acid aminotransferase	Amino acid degradation
1	N-acetyltransferase	Amino acid degradation
1	Peptidyl prolyl cis-trans isomerase	Amino acid degradation
1	Arabinosyl transferase	Amino acid degradation
2	Methyltransferase	Amino acid degradation
1	Malonate CoA ligase	Carboxylic acids degradation
1	Methylmalonate semialdehyde dehydrogenase	Carboxylic acids degradation
2	Malate dehydrogenase	Carboxylic acids degradation
30	Alpha-beta hydrolase	Fatty acid and sugars degradation
8	Hydrolase	Fatty acid and sugars degradations
1	3-ketoacyl-coA thiolase	Fatty acid degradation
1	Plastidial glycolate/glycerate translocator	Fatty acid degradation
1	Enoyl CoA hydratase	Fatty acid degradation
1	Non-specific lipid transfer	Fatty acid degradation
1	Malonyl CoA decarboxylase enzyme	Fatty acid degradation
4	Desaturase	Fatty acid degradation
4	Linoleate lipoxygenase	Fatty acid degradation
6	Dehydrogenase, dehydratase or oxygenase	Fatty acid degradation
9	Oxidation-reduction of fatty acids	Fatty acid degradation
9	Esterase	Fatty acid degradation
19	Lipase or phospholipase functions	Fatty acid degradation
21	Acyltransferase or transferase	Fatty acid degradation
44	GDSL esterase/lipase enzymes	Fatty acid degradation
1	Acetyl-coenzyme A synthetase	Fatty acid synthesis
1	Lipoyl synthase	Fatty acid synthesis
1	Lipid-A-disaccharide synthase	Fatty acid synthesis
1	Long-chain acyl-CoA synthetase 1	Fatty acid synthesis
2	3-ketoacyl-CoA synthase	Fatty acid synthesis
5	3-oxoacyl-[acyl-carrier-protein] synthase	Fatty acid synthesis
1	Amylase	Sugars degradation
1	Fructofuranosidase	Sugars degradation
1	Simple sugar dehydrogenase	Sugars degradation
1	Glucose-6-phosphate-1-epimerase	Sugars degradation
1	Polygalacturonase	Sugars degradation
2	Phosphatase	Sugars degradation
6	Xylodase	Sugars degradation
7	Glucosidase	Sugars degradation
8	Transferase function on sugar molecules	Sugars degradation
8	Galactosidase	Sugars degradation
11	Glycosyltransferase	Sugars degradation

Of the 227 genes identified, 30 encode an "alpha-beta hydrolase" with a probable lipase function (Holmquist, 2000; Mindrebo et al., 2016) and 8 for enzymes with hydrolase functions. As their name indicates, the enzymes encoded by these genes would participate in lipids or sugars degradations.

Of the 227 genes identified, 121 have functions for fatty acid degradation (table 14), 11 genes were identified as being involved in the fatty acid synthesis, 47 involved in sugars degradation, six involved in amino acid degradation, four involved in carboxylic acids degradation. The degradation of sugars, amino acids, and carboxylic acids are necessary steps for the synthesis of compounds such as acids, alcohols, or esters.

4.5.2-Monoterpene biosynthesis pathway

Some genes involved in their biosynthesis were identified in the precedent study (Colonges et al., 2021b). For the linalool cis pyranic oxide four genes coding for "*Probable terpene synthase 9*" was identified (three on chromosome 7 and one on chromosome 10). This gene is known to be responsible for the transformation of the geranyl diphosphate into linalool, the precursor of linalool cis-pyranic oxide (Cseke et al., 1998). For the linalool Trans furanic oxide six candidate's genes were identified on chromosome 5: five coding for *Cytochrome P450 89A2*" and one for "*Cytochrome P450 89A9*". At least, two studies have shown the implication of Cytochrome P450 for the transformation of linalool into 6,7-epoxylinalool (Meesters et al., 2007; Chen et al., 2010).

4.5.3-L-phenylalanine degradation pathway

Some genes involved in their biosynthesis were identified in the precedent study (Colonges et al., 2021b). In total, twenty-nine candidate genes were identified in these associations' zones involved in their biosynthesis. Any candidate gene was detected in associations linked to ethyl benzoate (UR). Forty-six candidate genes were identified in the areas of association with the other compounds. They were chosen of their predictive functions. They indicate possible involvement in the degradation pathway of L-phenylalanine (sup table 4).

4.6-Candidates genes potentially involved in the general plant defence identified in association areas linked with biochemical compounds

As suggested earlier (Sabau et al., 2006), the detection of fermentative micro-organisms seems to induce defensive responses that may be responsible for the synthesis of certain aromatic compounds.

In the association zones linked to compounds involved in fatty acid and sugar degradation, twenty-three genes are involved in the general defences of the plant (Table 14). Jasmonic acid and salicylic acid are two phytohormones with an important role in the defence mechanisms.

In the association zones linked to pyrazines compounds, sixteen genes coding for enzymes with functions involved in general plant defences were identifying (Table 14). The functioning of the proteasome plays an essential role in the activation of the jasmonic acid response pathway (Song et al., 2014). The fumaryl-acetoacetate hydrolase was identified in Arabidopsis Thaliana as affecting cell death in response to jasmonic acid (Zhou et al., 2020). Ethylene is also a phytohormone that plays a role in the response to biotic stresses (Song et al., 2014).

5-Discussion

It is the first time that integrative analysis of genetic, biochemical and sensorial fruity traits was conducted, leading to the identification of a few metabolomic pathways involved in fruity traits.

A large number of associations could be detected in relation to the different analysed traits related to fruity taste. Out of the thirteen sensory descriptors relating to fruity notes, associations were detected for only three of them. This is far less than what was detected for floral notes. Colonges et al., (2021), detected associations for eleven of the sixteen floral descriptors obtained by sensory analysis. These results show that there is greater variability in the presence of floral notes than fruity notes in this population of modern Nacional.

However, a greater number of associations with volatile compounds known to have a fruity taste were detected. Possibly, these compounds are not in sufficient concentration to be statistically detected in the sensory analysis. Each compound would thus have a weak effect on the fruity note and in this case, they are not detectable by the GWAS method. In addition, some volatile compounds with a fruity taste can be synthesised by the micro-organisms present during fermentation (Schwan and Wheals, 2004). This is the case, for example, of esters, which are largely synthesized by fermentative yeasts (Soles et al., 1982; Ho et al., 2014). In our case, it is possible that these microorganisms played a role in the production of volatile aromatic compounds, but they are not the only ones responsible, since a large number of associations related to volatile compounds known to have a fruity taste have been detected. This means that the cocoa tree is also involved in the synthesis of these compounds in associations.

Of the 74 volatile compounds related to the fruity notes, association zones could not be detected for 13 of them. These compounds belong to different chemical families (two acids, three alcohols, one aldehyde, five esters, and two ketones). As these different compounds do not appear to be genetically related to the cocoa trees in this population, the microorganisms present during fermentation likely synthesized them (Soles et al., 1982; Abbas, 2006; Ho et al., 2014).

Five biosynthesis pathways appear to be involved in the synthesis of fruity compounds in cocoa: the monoterpene biosynthesis pathway, the L-phenylalanine degradation pathway, the pyrazine production pathway, the sugar degradation pathway, and the fatty acid degradation pathway.

The pyrazine production seems to be achieved by the Maillard reaction during roasting. Indeed, ten of the eleven pyrazines were detected in cocoa beans after roasting. The eleventh pyrazine, tetramethylpyrazine, was certainly synthesised during drying also through the Maillard reaction (Starowicz and Zieliński, 2019). For this reaction to take place, free amino acids as well as reducing sugars are needed. This is why genes coding for proteinase or sugar degradation enzymes were found in the pyrazine association zones. As was observed for floral notes, co-locations between pyrazines known to have a fruity note and other pyrazines with no known notes suggest that a gene coding for a key enzyme involved in their biosynthesis could be responsible for the presence or absence of the fruity note. This is the case, for example, of the co-locations between methylpyrazine, known to have a nutty note, and 2,3-dimethyl-5-ethylpyrazine, known to have burnt notes, on chromosomes 1, 4, and 6. It is possible that in these areas of the genome, the same enzyme is at the origin of their precursors. In the case where the enzyme is active, the precursors for pyrazine A would be synthesized and in the case where the enzyme is inactive, the precursors for pyrazine B would be present. If pyrazine A is the one bringing a dry fruit note then when the enzyme is activated, a dry fruit note is present; otherwise it is absent.

The pathways of sugar degradation and fatty acid degradation are linked and have several biochemical compounds in common. A large number of areas of the association have been detected for these many traits. With the large number of results and the multiple possibilities of compound synthesis, it is difficult to establish a hypothetical biosynthetic pathway in cocoa from these results. Furthermore, some intermediate compounds may be synthesised by the microorganisms present during fermentation. The biosynthesis of these

compounds could be a synergy between the enzymatic actions of the microorganisms and the cocoa beans. Further studies on the presence of these volatile compounds with or without some micro-organisms (yeast or bacteria) could identify more precisely the biosynthetic mechanisms. However, the co-localisations observed between certain compounds suggest enzymatic actions initiated in the cocoa beans. This is the case, for example, on chromosome 1 where a co-location between the association linked to pentan-1-ol (R) and pentanoic acid (R) is observed in the haplotypic block number 18 (5 356 899pb – 5 820 933pb). In this case, a cocoa enzyme could be responsible for the oxidation of the alcohol pentan-1-ol leading to the carboxylic acid: pentanoic acid. A candidate gene "Acyl-coenzyme A oxidase 2, peroxisomal" encoding an enzyme with an oxidative function was found at position 5 520 690 on chromosome 1 next to the haplotypic block of the associations (appendix 15).

The fruity aroma of Nacional could also include compounds belonging to the L-phenylalanine degradation pathway such as benzaldehyde, benzyl alcohol and 1-phenylethyl acetate, known to have fruity flavours. Associations have been detected for these three compounds and exposed in a previous study (Colonges et al., 2021b). These compounds are also precursors of compounds known to have floral tastes. The enzymatic activity allowing the degradation of benzaldehyde or benzyl alcohol is one of the keys to the increased presence of compounds with fruity or floral tastes.

The study of the determinism of aromas is complex and involves several disciplines. No biosynthesis pathway produces only one of the compounds with the same type of taste. It is the balance between all the compounds that allow the synthesis of an aromatic profile and that of Nacional cocoa is complex. The presence of an aromatic compound is not necessarily synonymous with the perception of its aromatic note. Identification of the volatile compounds with a perceived fruity note could be carried out thanks to a gas chromatography analysis coupled with olfactometry. With this approach, it would be possible to identify the molecules whose fruity notes are perceived and thus select the candidate genes favourable alleles associated with these compounds, providing new tools for marker-assisted selection.

Conflict of Interest
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions
EC., CL, RGLS, conceived the experiment; JCJ, AS conducted biochemical analyses; ES carried out sensorial analyses; OF carried out DNA experiments; KC, JCJ, AS, RB, CL, FD, SA, XA analyzed data; KC, RB, CL wrote the manuscript.

Funding

The study was funded by the United States Department of State (U.S. Foreign Ministry); the U.S. Embassy, Quito; the U.S. Department of Agriculture (USDA-ARS); the MUSE Amazcacao project with the reference ANR-16-IDEX-0006.

Acknowledgement

We thank the USDA and the I-Site MUSE for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" programme with the reference ANR-16-IDEX-0006.

Partie 2 : Bases génétiques des notes fruitées (fraîches et séchées) de la variété de cacao Nacional.

Cette partie reprend la présentation sur le déterminisme des arômes fruités du Nacional moderne, faite à l'occasion du congrès international du 16ème WEURMAN.

Genetic bases of fruity notes (fresh and dried) of the Nacional cocoa variety

Kelly Colonges[1, 2,3], Juan-Carlos Jimenez[4], Alejandra Saltos[4], Edward Seguin[5], Rey Gastón Loor Solorzano[4], Olivier Fouet[1], Xavier Argout[1], Sophie Assemat[2,3], Fabrice Davrieux[2,3] Eduardo Morillo[4], Renaud Boulanger[2,3], Emile Cros[2,3], Claire Lanaud[1]

1 CIRAD, UMR AGAP, F-34398 Montpellier, France. AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France. Kelly.colonges@cirad.fr

2 CIRAD, UMR QUALISUD, Montpellier, France. QualiSud, Univ Montpellier, CIRAD, Montpellier SupAgro, Univ Avignon, Univ Réunion, Montpellier, France.

3 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

4 Instituo Nacional de Investigacion Agropecurias, INIAP, Ecuador

5 Guittard, United-States

Key words: Cocoa, fruity aroma, genetics

1-Abstract

Theobroma cacao is the only source of cocoa. Cocoa is classified into two types of products: bulk cocoa and fine flavour cocoa. Contrary to bulk cocoa, fine aromatic cocoa is characterized by its floral and fruity aromas (Sukha et al., 2008). In order to understand the genetic determinism of the formation of these aromas in cocoa beans, a genetic study using the Genetic Wide Association Study (GWAS) method was undertaken. It was carried out on 158 clones belonging to a population of Nacional tree type cultivated in Ecuador, whose volatile compound concentrations and sensory profiles were characterized for its diversity. This study revealed areas of correlation between, on the one hand, the genetic diversity of this population, revealed by molecular marker alleles and the volatile compounds detected in the different clones, and on the other hand, between this same genetic diversity and their sensory profiles.

These correlation zones, also called associations, are therefore linked to one of these traits, but also in some cases to both types of traits. Thanks to these associations, which correspond to a restricted area of the cocoa genome, and the knowledge of its complete sequence (Argout et al., 2017), candidate genes have been brought to light. Some of them are known and identified in biosynthesis pathways of volatile compounds, which are themselves known to have a fruity note. In a preliminary study, a difference in the expression of these genes was identified between four genotypes (two floral and two fruity genotypes) during different stages of development and fermentation of the beans. The results showed that the candidate genes tended to be activated during fermentation and not during the maturation stages of the pods.

2-Introduction

There are two types of cocoa: "Standard or bulk" cocoa, which has a strong cocoa taste, and aromatic fine cocoa, which is characterised by floral and fruity notes (Sukha et al., 2008). The Nacional variety of cocoa is classified as a fine variety, characterised by floral and spicy notes (Luna et al., 2002) known as the "ARRIBA" flavour. At present, the trees grown as Nacional (called Nacional modern in this study) are the result of several generations of crosses between the ancestral Nacional and Trinitarios (themselves hybrids between the Criollo and Amelonado varieties)(Loor S. et al., 2009). Criollo is also a fine aromatic cocoa variety characterised by fruity notes (Lachenaud and Motamayor, 2017). While Amelonado is known for its strong cocoa aroma. The floral aroma of Nacional has been studied and two main biosynthetic pathways have been identified as responsible for this aroma: the terpene biosynthetic pathway and the L-phenylalanine degradation pathway. [6, 7]. In this study, part of the deciphering of the fruity flavour in cocoa from trees of the Nacional modern variety will be presented. A GWAS (Genome Wide Association Study) was conducted to find out which areas of the genome are responsible for this fruity flavour. The GWAS study was carried out using phenotyping data including the determination of volatile compounds related to the fruity taste as well as sensory analyses. To further investigate the genomic determinants of the fruity aroma of Nacional modern, candidate genes in the biosynthetic pathways of the identified fruity compounds were searched for in the identified association areas**.**

3-Experimental

3.1-Plant material

The plant material used for these experiments was composed of a collection of 151 cocoa trees from Ecuador conserved in the Pichilingue experimental station of the “Instituto

Nacional de Investigaciones Agropecurias" (INIAP) and the "Colecion de Cacao de Aroma Tenguel" (CCAT) of Tenguel. This population represents the Nacional variety currently grown in Ecuador and has been described by Loor et al (Loor S., 2007).

3.2-Biochemical analysis

Cocoa beans samples were all fermented and dried with the same method and in the same place of Pichilingue. A part of cocoa beans samples was also roasted. Volatile compounds analysis was carried out on dried fermented beans and roasted beans. The SPME extraction fiber and GC/MS analysis were conducted according to the conditions described by Assi Clair et al (Assi-Clair et al., 2019) with Agilent 6890N gas chromatography-mass spectrometer (GC–MS) equipped with a Hewlett Packard capillary column DBWAX, 60 m length × 0.25 mm internal diameter × 0.25 µm film thickness (Palo Alto, CA, USA).

3.3-Sensory analysis

146 individuals were characterized by sensory analyses based on blind tastings carried out on 3 repetitions per sample. The tastings were carried out on cocoa liquor. Thirteen fruity notes were judged with a score ranging from zero (no fruity notes detected) to ten. We used the average of the three replicates for the phenotype used to carry out GWAS.

3.4-Genetic analysis

GWAS was performed with SNP and SSR markers associated with biochemical (146 accessions x 5195 markers) and sensory (144 accessions x 5195 markers) traits using TASSEL v5. Two models were used: mixed linear model (MLM) and a fixed effect model (GLM). In both cases, a structure matrix was determined by running a PCA. Candidate genes were identified through association zones and their annotated function in the cocoa genome.

3.5-Expression analysis

The expression of four genes (two coding for alpha-beta hydrolases, one for carboxylesterase and one for GDSL esterase/lipase) was studied during pod ripening at 18, 20 and 22 weeks of development as well as 24 hours after the start of fermentation. These studies were conducted on four different clones, known for their fine flavour, characterised by a fruity and floral aroma: EET103, EET19, EET575 and EET62.

4-Results and discussion

GWAS analyses are used to statistically determine which areas of the genome are linked to the traits being tested.

4.1-Sensory traits

Thanks to this genetic analysis method, 22 areas of associations related to sensory data were detected. Of the 13 fruit sensory categories, associations were detected for 3 of them: "Fruity-Dark Tree Fruit", "Fruity-Dried fruit" and "Fruity-Berries". The strongest areas of association were detected for the note "Fruity-Dark Tree Fruit" on chromosome 1 (Figure 27).

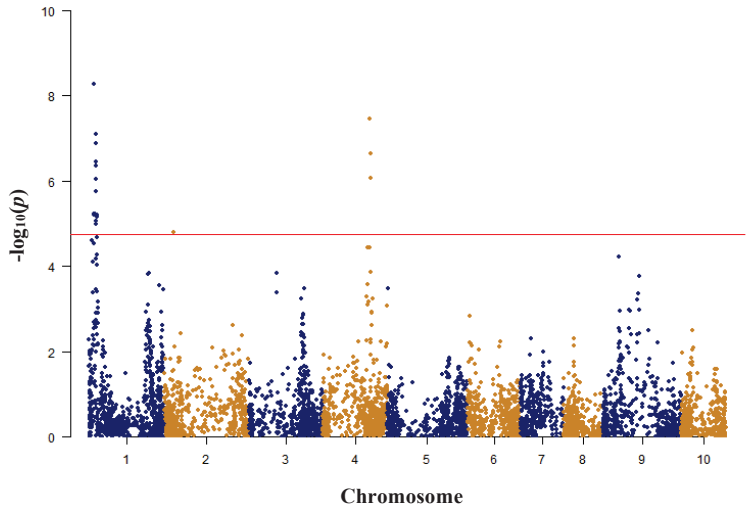


Figure 27: Manhattan plot representing all markers tested along the 10 cocoa chromosomes for associations with the Fruity Dark tree fruit note.

4.2-Biochemical traits

The GWAS analysis allowed the detection of 480 areas of associations related to the concentrations of volatile compounds. These association zones were detected in relation to 3 acids, 12 alcohols, 7 aldehydes, 22 esters, 8 ketones, 1 lactone, 11 pyrazines and 5 terpenes. According to these results, five biosynthetic pathways appear to play a role in the synthesis of fruity aroma in cocoa: the monoterpene biosynthetic pathway, the L-phenylalanine degradation pathway, the simple sugar degradation pathway, the fatty acid degradation pathway and the pyrazine biosynthetic pathway. The study of associations related to the monoterpene biosynthetic pathway and the L-phenylalanine degradation pathway have been extensively studied previously (Colonges et al., 2021b).

Areas of common associations between sensory and biochemical traits were first sought. These areas have a higher probability to explain the fruity aroma of cocoa. Six co-locations between associations related to pyrazines and those related to sensory traits were identified. Four of them are present on chromosome 1, one on chromosome 4 and one on chromosome 10. Eight areas of co-locations between associations related to sensory traits and compounds involved in the degradation pathways of fatty acids and simple sugars were observed: four co-

locations are on chromosome 1, two on chromosome 2, one on chromosome 4 and one on chromosome 10.

4.3-Candidate genes

In all the areas of association detected, candidate genes were sought. A first search was carried out by looking for genes coding for proteins with essential enzymatic functions for the different biosynthesis pathways identified.

In cocoa, it is well known that pyrazines are synthesized through the Maillard reaction (Afoakwa et al., 2008). It is a set of chemical reactions that take place during the process. In cocoa processing, the Maillard reaction occurs mainly during roasting, but it can also occur during drying. The Maillard reaction allows the synthesis of pyrazine by combining free amino acids and reducing sugars. In this study, we therefore looked for genes coding for enzymes involved in protein degradation or in the synthesis of reducing sugars.

In the 277 association areas related to pyrazines, 213 candidate genes involved in the production of precursors of the Maillard reaction were identified. These genes have mainly peptidase or protease functions.

Preliminary expression studies of some of these genes were conducted. The expression of two genes coding for alpha-beta hydrolases was studied. The two genotypes EET103 and EET19 express alpha-beta hydrolase 1 equally or more during fermentation. The EET575 genotype expresses it more at 18 weeks of maturity and the EET62 genotype expresses it more at 18 and 20 weeks of maturity. Alpha-beta hydrolase 2 is most expressed at 20 weeks of pod development for genotype EET103, at 22 weeks of development for genotypes EET19 and EET575, at 18 and 20 weeks of development for genotype EET62 (Figure 28). Some genotypes appear to express more alpha-beta hydrolase 1 during fermentation than during pod maturity. It is at this time that alpha-beta hydrolases would break down proteins into amino acids through their peptidase functions. Cocoa beans would then be richer in amino acids, which would allow more pyrazine synthesis through the Maillard reaction during drying and roasting. Reineccius (2006) observed a higher concentration of pyrazine in well-fermented beans (Reineccius, 2006). The expression of alpha-beta hydrolase 2 seems to be more intense during pod maturation than during the fermentation. Pod ripening also seems to play a role in aroma synthesis.

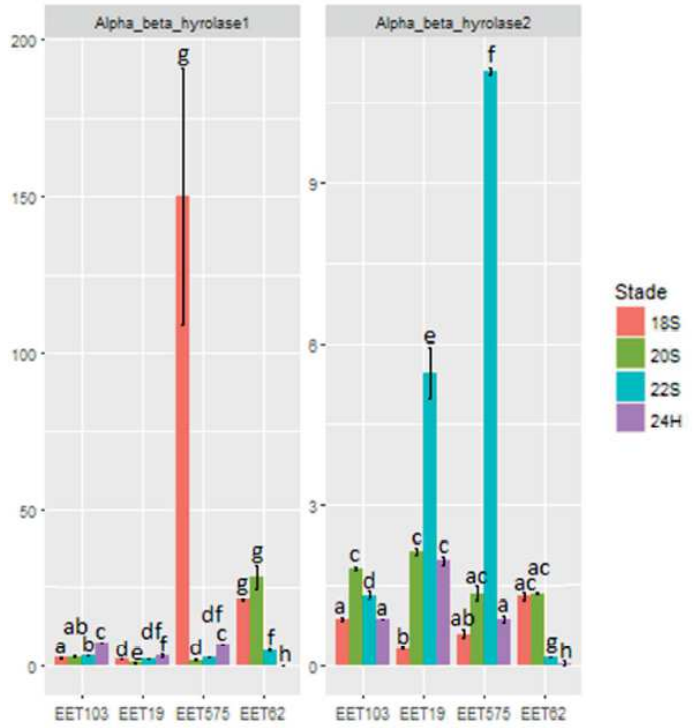


Figure 28: Histogram of the expression profile of the genes coding for Alpha-Beta Hydrolase at different stages of bean maturity and during fermentation.

Histograms with the same letter are not significantly different at the 1% threshold.

Acids, alcohols, ketones and esters are mainly synthesised as a result of the degradation of fatty acids and/or simple sugars. Candidate genes encoding enzymes involved in these degradation pathways were searched for in all association zones of all volatile compounds.

In the 480 association areas related to compounds involved in simple sugars degradation, 125 candidate genes were identified. These genes have mainly hydrolase or esterase functions. In the association areas related to compounds involved in fatty acid degradation, 217 candidate genes were identified. These genes have mainly lipase or esterase functions. There are common candidate genes between these simple sugars degradation pathway and fatty acid degradation pathway.

Preliminary expression studies of some of these genes were conducted. The first one coding for a carboxylesterase and the second for a GDSL esterase/lipase. The carboxylesterase gene is more expressed during fermentation in genotypes EET19 and EET575, more expressed after 20 weeks of seed development in genotype EET103 and similarly expressed at 18 and 22 weeks of seed development in genotype EET62. The GDSL esterase lipase gene is more

expressed during fermentation in all genotypes except EET103. In EET103, it is more expressed at 20 weeks of development (Figure 29).

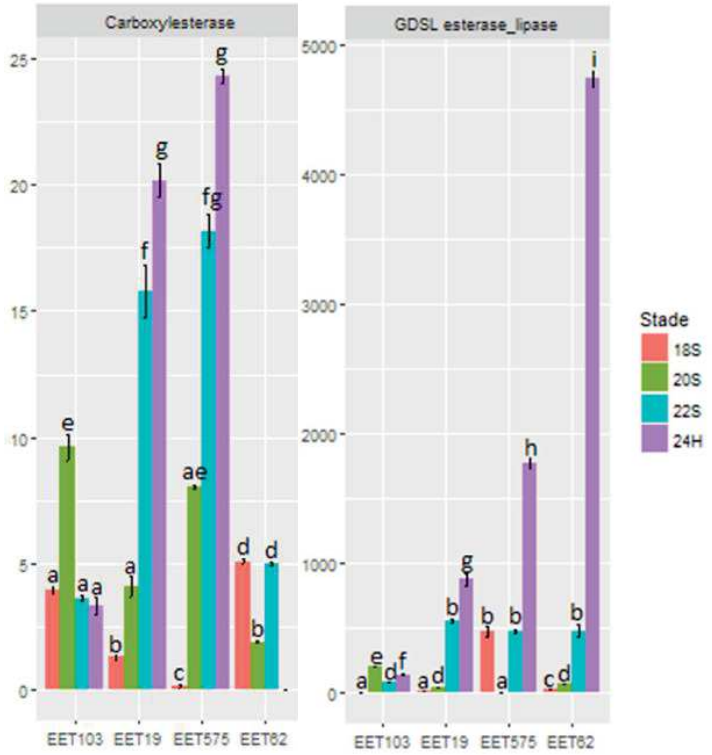


Figure 29: Histogram of the expression profile of the genes coding for Carboxylesterase and GDSL esterase/lipase at different stages of bean maturity and during fermentation.
Histograms with the same letter are not significantly different at the 1% threshold.

In most cases, the genes seem to be more strongly expressed during fermentation. During this processing stage, cocoa beans are exposed to fermenting micro-organisms. These exponentially growing micro-organisms are detected by the beans, which then set up a defence system. A large number of volatile compounds have been identified as being used by plants to defend themselves against micro-organism attacks (Pichersky et al., 1995; Palmqvist et al., 1999; Parent et al., 2018).

Of the 67 volatile compounds, association zones could not be detected for 13 of them. These compounds belong to different chemical families (two acids, three alcohols, one aldehyde, five esters and two ketones). As most of these different compounds do not seem to be genetically related to the cocoa trees in this population, it is very likely that they were synthesised by the microorganisms present during the fermentation (Soles et al., 1982; Abbas, 2006; Ho et al., 2014). In addition, it is possible that the microorganisms present during fermentation synthesise some intermediate compounds. The biosynthesis of these compounds could be a synergy between the enzymatic actions of the fermenting microorganisms and the cocoa tree.

The Nacional fruity aroma could also include compounds belonging to the L-phenylalanine degradation pathway, such as benzaldehyde, benzyl alcohol and 1-phenylethyl acetate. Associations have been detected for these three compounds and reported in a previous study (Colonges et al., 2021b). These compounds are also precursors of compounds known to have floral tastes. The enzymatic activity allowing the degradation of benzoic acid into benzaldehyde or benzyl alcohol is one of the keys to the increased presence of compounds with fruity tastes (figure 30).

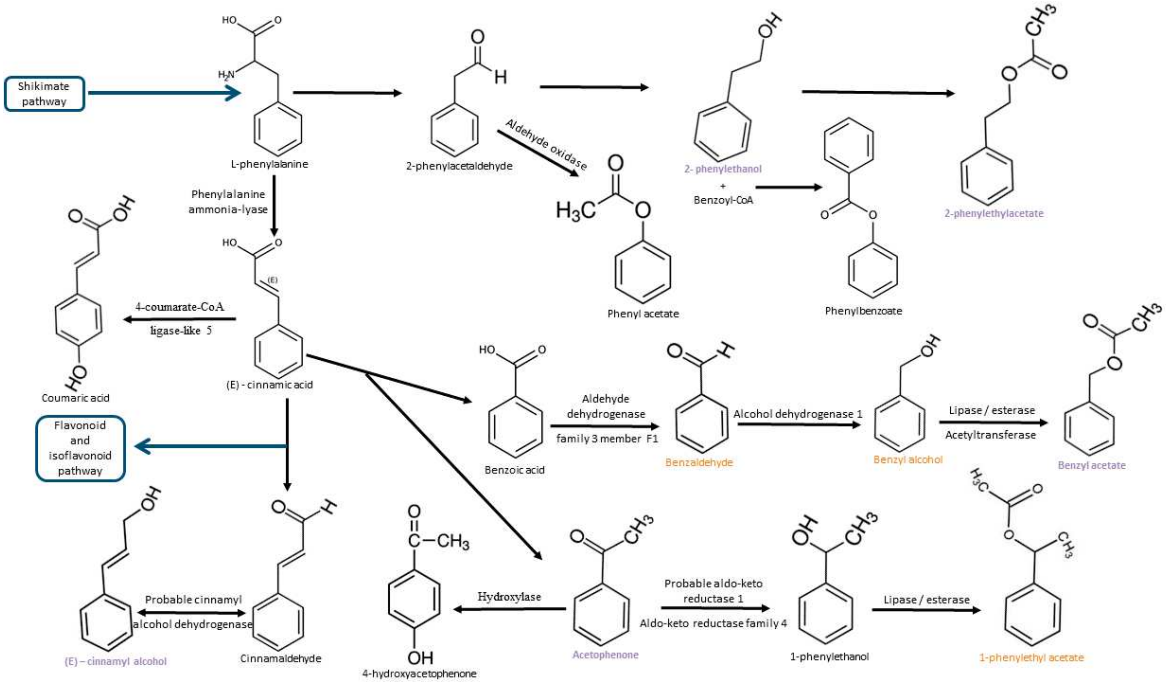


Figure 30: Degradation pathway of L-phenylalanine identified in cocoa.
Compounds known to have a floral taste are noted in purple and compounds known for have a fruity taste are noted in orange.

5-Conclusion

The study of aroma determinism is complex. Flavours depend on different factors such as genetics, growing environment and processing. In this study, the objective was to investigate the genetic part of the fruity notes of cocoa (fresh and dry). Using GWAS, we were able to begin to determine how the fruit notes were synthesised by the cocoa beans. Two types of fruity notes were observed: fresh fruity notes, which are mainly, composed of esters and terpene compounds. In this case, three metabolic pathways seem to be involved: the degradation of fatty acids, the degradation of simple sugars and the monoterpene biosynthetic pathway. Similarly, dried fruit aromas were observed, mainly composed of pyrazines. Pyrazines were synthesised by the Maillard reaction during the drying or roasting process. The concentration of pyrazine depends on the synthesis of the Maillard precursors: amino acids and redacting sugars.

A large number of genes potentially involved in the synthesis of fruity aromas have been identified. Preliminary study of the expression of candidate genes shows that the synthesis of enzymes responsible for the production of certain volatile compounds takes place during bean development but also during fermentation, before the death of the embryo. The hypothesis is that cocoa beans detect fermenting microorganisms and trigger defence mechanisms.

The exact role of these candidate genes in this synthesis is still difficult to determine. A study of the enzymatic activity of the different enzymes encoded by these candidate genes could complement this study. This could also help to identify more precisely the roles of each enzyme.

A complementary analysis, made by Gas-Chromatography-Olfactometry (GCO), comparing fruity, floral and standard genotypes of cocoa, would allow identifying key compounds of the fruity aroma of cocoa.

L'ensemble des résultats portant sur le déterminisme des notes florales et fruitées et également sur les sensations d'amertume et d'astringence (qui seront plus amplement décrite dans le chapitre 6) du cacaoyer Nacional moderne ont permis de déterminer différentes zones du génome associées à ces arômes. Des gènes candidats directement impliqués dans la voie de biosynthèse des composés volatils ont également pu être identifiés dans ces zones.

La variabilité génétique présente dans la variété population de Nacional moderne provient de l'histoire de sa domestication. En effet, cette variété est issue de diverses générations de croisements entre des arbres appartenant à la variété de Nacional ancestrale et des cultivars Vénézuéliens de type Trinitario (Bartley, 2005). Ces cultivars Trinitario sont eux-mêmes issus de diverses générations de croisements entre les groupes génétiques Criollo et Amelonado. Ces croisements résultent de croisements naturels entre arbres ancestraux (Nacional) et arbres introduits en Equateur (Trinitario). Cette domestication effectuée par l'homme a-t-elle influencé la qualité aromatique de la variété de Nacional Moderne que nous connaissons actuellement ? La réponse est oui. Dans ce cas, comment les différentes variétés ancêtres ont-elles contribué à façonner l'arôme de la variété Nacional moderne ? C'est à cette question que le prochain chapitre ébauchera une réponse.

Grâce aux analyses GWAS présentées dans les deux chapitres précédents, il a été possible de connaitre les zones importantes pour la synthèse des arômes chez le cacaoyer appartenant à la variété de Nacional moderne. Á partir de ces zones et grâce aux génotypages de trois individus par groupe génétique ancestral (Nacional, Criollo, Amelonado), il a été possible de retracer l'origine des allèles aux marqueurs importants pour la synthèse des arômes. Ainsi il a pu être montré que les trois ancêtres ont joué un rôle positif dans l'arôme que nous connaissons actuellement du Nacional moderne avec un enrichissement de zones favorables à la synthèse d'arôme fruités et/ou floraux par rapport au Nacional ancestral. Il a aussi été mis en évidence une augmentation des allèles favorables à la synthèse de polyphénols et à l'apparition de l'amertume et de l'astringence dans cette population de Nacional moderne. L'ensemble des résultats vous sont présentés dans le chapitre suivant.

Chapitre 4: Les arômes de la variété moderne Nacional sont façonnés par l'histoire de sa domestication

Chapitre 4: Les arômes de la variété moderne Nacional sont façonnés par l'histoire de sa domestication

Flavours of the modern Nacional variety are shaped by cocoa domestication history

Kelly Colonges[1,2,3,4*], Rey Gastón Loor Solorzano[5], Juan-Carlos Jimenez[5], Alejandra Saltos[5], Edward Seguine[6], Olivier Fouet[1,2], Xavier Argout[1,2], Sophie Assemat[3,4], Fabrice Davrieux[3,4], Emile Cros[3,4], Renaud Boulanger[3,4*], Claire Lanaud[1,2*]

1 Cirad, UMR AGAP, F-34398 Montpellier, France.

2 AGAP Institut, Univ Montpellier, Cirad, INRAE, Institut Agro, Montpellier, France.

3 Cirad, UMR Qualisud, F-34398 Montpellier, France.

4 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

5 Instituto Nacional de Investigacion Agropecurias, INIAP, Ecuador.

6 Guittard, United-States

Keywords: domestication, *Theobroma cacao*, cocoa aroma

1-Abstract

Nacional variety is a cocoa variety known for its specific flavours, called "Arriba" flavour. In the 19th century, this Nacional variety, cultivated in Ecuador for several centuries, were hybridised over several generations with “Venezuelans” cultivars of Trinitario type, corresponding to hybrids between Amelonado and Criollo, so involving three main contrasted ancestors. Several analyses of modern Nacional had shown that this new variety was still an aromatic fine cocoa variety but that its present aromas had evolved. Effect of this recent domestication step on the aromas of the modern Nacional variety was studied. Using genotyping data from the 3 reference ancestors and GWAS results for all quality traits (volatile and non-volatile compounds, sensory analysis); it was possible to trace the origin of the alleles with a positive effect on aroma in the different association areas. This study showed that all founders contributed alleles favourable to the synthesis of quality aromas (floral, fruity, ...) but also to the synthesis of defects. This study was able to show that the association zones linked to favourable aromas and those linked to defects were not genetically linked. It is therefore possible to select the areas of interest for the flavours while counter-selecting the areas that bring defects. It was thus possible to show that the recent domestication of Nacional, which happened one century ago, has shaped the aromas profile of this variety, and refine the contribution of each ancestor at the genome level.

2-Introduction

Theobroma cacao is the only source of chocolate in the world. *T. cacao* whose genome has been sequenced (Argout et al., 2011; Motamayor et al., 2013; Argout et al., 2017) displays a great genetic diversity. To the present day, ten distinct genetic groups have been identified: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañón, Nanay, Nacional et Purùs (Motamayor et al., 2008). Currently, the cocoa market is divided into two groups. Firstly, the "bulk cocoa" is characterised by a strong cocoa taste and secondly, the "fine aromatic cocoa" is characterised by fruity and floral aromas. The latter represents about 5% of world production. Today, the most widely cultivated varieties of fine aromatic cocoa are Criollo, Nacional and Trinitario trees. Trinitario is a hybrid between Criollo and Amelonado varieties. The Criollo variety was the first to be cultivated on a large scale (Motamayor et al., 2002) in Central America by the Maya populations. The Criollo variety is known for its fine, mainly fruity flavours. Currently, the Criollo variety is not widely grown due to its low vigour and increased susceptibility to disease (Cheesman, 1944).

Nacional is known for its aroma, characterised by floral notes, known as the "Arriba" aroma. The ancestral Nacional was cultivated in small Ecuadorian plantations in the region known as "ARRIBA", which gives its name to the Nacional flavour. Thanks to this specific flavour, Nacional cocoa was recognised as "fino de aroma", i.e. fine aromatic cocoa. Fine aromatic cocoa was highly valued by chocolate makers and was only produced in Ecuador (Loor S. et al., 2009). Nacional's production represents around 75% of Ecuador's total production.

Nacional was first associated with the Forastero group by Cheesman, (1944) then it was classified as being close to the Criollo by Enriquez, (1992). Shortly afterwards, thanks to genetic analyses, the Nacional was classified as an independent group by Lerceteau et al., (1997).

In 1890, Ecuadorian cocoa trees experienced a serious health crisis. Diseases such as witches' broom and "frosty pod" had a severe impact on cocoa crops. To produce trees resistant to these diseases, foreign cocoa trees were introduced. The name "Nacional" was then used to define the variety introduced before the introduction of foreign genetic material (Bartley, 2005). Nacional' is most often applied to trees with elongated fruits whose pods are green before they ripen, with a rough surface and a spiky shell (Pound, 1938). Despite the narrow genetic base of the ancestral Nacional variety, there is some variability within the ancestral population (Loor S. et al., 2009).

Before 1920, seed dissemination and cultivation in Ecuador was based on beans from uncontrolled crosses. After the discovery of Moniliophthora perniciosa (the fungus responsible for witches' broom disease), the search for trees resistant to the disease was undertaken to reduce the susceptibility of the population. The modern Nacional variety currently grown is a population variety resulting from hybridizations between the "Nacional" types and Trinitario cocoa trees imported from Venezuela, and called "Venezuelan cultivars", or from Trinidad (Bartley, 2005; Loor S. et al., 2009; Rottiers et al., 2019). This hybrid nature was demonstrated by Loor S. et al., (2009). Several authors report that these hybridizations would have led to a dilution of the Arriba flavour (Loor S. et al., 2009; Boza et al., 2014; Beckett et al., 2017). Nacional trees are more resistant to disease and produce more than the ancestral Nacional. Populations of modern Nacional display a wide variation in their genetic composition and their phenotypic characteristics (Bartley, 2005; Loor S., 2007).

From 1940, surveys were carried out on the Ecuadorian coast to collect Nacional type to preserve their genetic resources. These collections were placed in two main experimental stations of INIAP (Instituto Nacional de Investigaciones Agropecurias) which are the Tropical Experimental Station of Pichilingue (EET-P) and the Cocoa Flavour Centre of Tenguel (CCAT).

The fine aromas of the modern Nacional are the result of the genetic mixing between the ancestral Nacional (known for its floral aroma) and the Trinitario trees. The genetic and biochemical determinism of the floral and fruity aromas of the modern Nacional have recently been studied (Colonges et al., 2021b; Colonges et al., 2021c). Based on the genetic analyses from these previous studies, the objectives of this study are to determine, at the genome and allele level, which ancestors, among ancestral Nacional, Criollo and Amelonado, have positively influenced the floral and fruity aromas of the modern Nacional variety, as well as the presence of bitterness or vegetal notes.

3-Materials and methods

3.1-Vegetal material

The plant material used for these experiments were composed of a collection of 152 cocoa trees from Ecuador conserved in the Pichilingue experimental station of the “Instituto Nacional de Investigaciones Agropecurias” (INIAP) and the “Collecion de Cacao de Aroma Tenguel” (CCAT) of Tenguel. This population represents the Nacional variety currently grown in Ecuador and has been described by Loor S. (2007). For the part studying the origins of the alleles, representative trees from each of the three ancestors of the modern Nacional variety were used: B_97_1, B_97_2 and B_97_61 representing Criollo population; Catongo, Matina1-6-A1 and Matina1-6-A2 representing Amelonado population and SNA1003, SNA405 and SNA604 representing ancestral Nacional (Loor S. et al., 2012). The three varieties have a very narrow genetic base. The three selected individuals represent them well even if there are only three representatives per variety.

3.2-Fermentation processes

Micro-fermentations of cocoa beans were carried out in a wooden box. The process lasted 4 days with two turns at 24 and 72 hours after the beginning of the fermentation. Each clone sample (152) was placed in a protective laundry bag and micro-fermented in a cocoa mass. After fermentation, the samples were put in a dry place. They were considered dried when their moisture content was below 8%.

3.3-Sensorial Analysis

146 individuals were characterized by sensory analyses based on blind tastings carried out on 3 repetitions per sample. The tastings were carried out on cocoa liquor. The cocoa liquor corresponds to merchant cocoa (dried fermented beans) which have been roasted and crushed. Sixteen floral notes were judged with a score ranging from zero (no notes detected) to ten according to ISCQF, (2020) protocol. We used the average of the three replicates for the phenotype of the GWAS analysis.

3.4-Sensory traits analysis

The sensory analysis of the cocoa liquors, as described in Colonges et al., 2021a; Colonges et al., 2021b revealed 38 sensory descriptors: cocoa taste, acidity, bitterness, astringency, 16 different floral notes, 7 different fruity notes, hazelnut taste, other positive notes, smoky taste, caramel taste, mouldy taste, brown flavours and degree of roasting.

3.5-Volatile compound analysis by GC-MS

GC/MS analyses were conducted according to the condition described by Assi-Clair et al., (2019).

3.6-Non-volatile compound analysis by NIRS

NIRS acquisitions and treatment were conducted according to the Davrieux et al., (2007) and Hue et al., (2014) protocol.

3.7-Biochemical traits analysis

The identification of volatile compounds was conducted on roasted and unroasted cocoa beans. In total, 160 volatile compounds were identified, of which 33 are known to have fruity notes (Colonges et al., 2021c) and 20 are known to have floral notes (Colonges et al., 2021b). These compounds are mainly synthesised through four metabolic pathways: the degradation of fatty acids, the degradation of sugars, the degradation of L-phenylalanine and the monoterpene biosynthetic pathway. The pyrazines involved in the presence of dried fruit notes are mainly synthesised through the Maillard reaction that requires amino acids and reducing sugars. The biosynthetic pathways of floral compounds in Nacional cocoa are mainly the L-phenylalanine degradation pathway and the monoterpene biosynthetic pathway (Colonges et al., 2021b). The synthesis of compounds responsible for fruity aromas involves all five biosynthetic pathways (Colonges et al., 2021c). The identification of non-volatile compounds was carried out only on roasted cocoa beans. Six compounds were identified: epicatechin, theobromine, caffeine and

three polyphenols (B2, B5 and C1). These compounds are known to provide bitterness in a lot of matrix like wine, tea and cocoa (Pickenhagen et al., 1975; Kallithraka et al., 1997; Misnawi et al., 2004; Ma et al., 2018).

3.8-Statistical analysis

PCA analysis and visualization were made with the "mixOmics" R package. Calculation of correlation was made with "agricolae" R package and visualization of correlation matrix with "corrplot" R package.

3.9-DNA extraction protocol

DNA extraction was conducted according to the Risterucci et al. (2000) protocol.

3.10-Genotyping by SSR

This population was genotyped, using SSR markers, by Loor S. et al., (2009).

3.11-Genotyping by sequencing (GBS)

DNA samples were genotyped by sequencing (GBS) using DArTseq (Diversity Arrays Technology Sequencing) technology (Kilian et al., 2012). Reads were aligned with the V2 sequence of the Criollo genome (Argout et al., 2017). Markers with unknown locations were discarded for analysis.

3.12-Association mapping

SNP GWAS and SSR GWAS was conducted according to the methodology described in (Colonges et al., 2021b).

3.13-Determination of associations zones

In this study, the areas of association were defined differently than in studies from which the GWAS results are drawn (Colonges et al., 2021b; Colonges et al., 2021c). The association areas were decided based on the population linkage imbalance which is 7.5Mb according to (Loor S., 2007). When a marker is significantly associated with a trait, the confidence interval is, therefore, 3.75Mb on either side of the marker. Markers with intersecting confidence intervals are reported as belonging to the same association area. The size of the association area is a maximum of 7.5Mb and its position depends on the marker with the lowest p-value or the most significant association.

3.14-Determination of allele origins

The study covers 5195 markers used for the GWAS analysis. For each significantly associated marker and each trait, the effects of the three possible genotypes present (AA, CC,

AC) were calculated by TASSEL 5. Only markers that were significantly associated with a trait with at least two GWAS methods were selected. The markers with genotypes with the superior effect differing according to the GWAS methods were eliminated for this study.

Then, the genotypes with the superior effects on each trait were compared to the genotypes of the nine individuals representing the three ancestors.

An allele was considered to belong to one of the ancestral varieties when at least two of the three individuals of the same variety had a heterozygous genotype with this allele or when at least one of the three individuals was homozygous for this allele. A given allele can be present in several ancestors.

3.15-Comparison of individual sensory scores and biochemical concentrations

The comparison of all the traits studied among the individuals belonging to the Modern Nacional variety with those close to the ancestral Nacional variety (here represented by SNA405, SNA418, SNA 604 and Fidencio) is represented in the form of a graph in parallel coordinates (Figure 31 to Figure34). These graphs were produced using the R package "GGally". The data on the concentrations of biochemical compounds were scaled: for each biochemical compound, the maximum concentration is equivalent to a score of one and a score of zero is equivalent to the minimum concentration observed.

4-Results

4.1-Description of the control used

The individuals used as a control to represent each of the ancestor varieties are very close to those previously identified by Motamayor et al., (2003) and Loor S. et al., (2012) as the most probable ancestors of these varieties using parentage analyses. Out of 52,635 SNP markers (GBS), B97_1 has 0.46% heterozygous markers, B97_2 has 0.55%, B97_61 has 0.30%, Catongo has 0.75%, Matina1.6-A1 has 0.22%, Matina1.6-A2 has 0.12%, SNA1003 has 1.69%, SNA405 has 6.47%, and SNA604 has 1.62%.

4.2-Distribution of the traits studied among individuals belonging to the ancestral Nacional variety and those belonging to the modern Nacional variety

4.2.1-Notes perceived during sensory analysis

The four individuals chosen as representatives of the ancestral Nacional variety have a cocoa note, floral notes 'Bark woody' and 'Lightwood' and smoky notes that are more important than most individuals belonging to the modern Nacional variety (Figure 31).

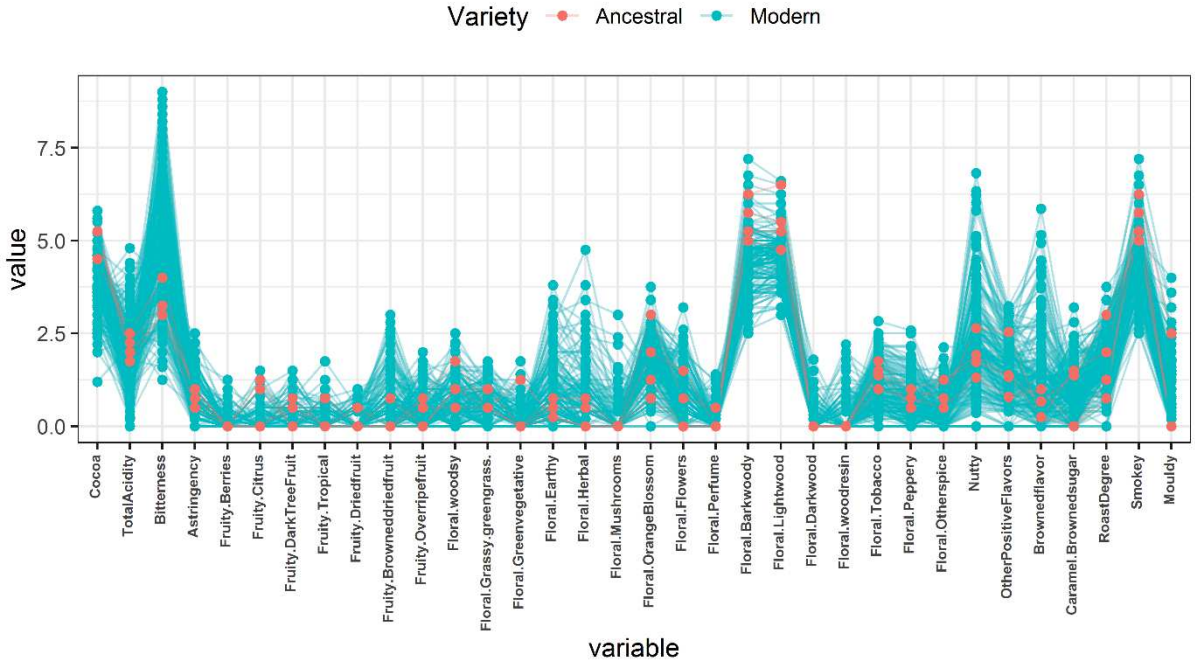


Figure 31: Parallel coordinate plot for the Sensory Data.

The values displayed correspond to the scores attributed to each genotype during the sensory analysis. In blue, the individuals belonging to the modern Nacional variety are shown and in red, the individuals close to the ancestral variety.

Individuals belonging to the modern Nacional variety show higher lactic acidity, fruity notes of berries, floral notes of mushrooms, dark wood and wood resin than those detected in the four individuals chosen as representatives of the ancestral Nacional variety (Figure 31).

A large proportion of the individuals belonging to the modern Nacional variety have a greater bitterness, astringency, earthy, herbal and browned flavour than those detected in the four individuals chosen as representatives of the ancestral Nacional variety (Figure 31).

4.2.2-Concentration of compounds known to have a floral note

A large proportion of the individuals belonging to the modern Nacional variety have a higher content of phenylethanal (UR), guaiacol (UR and R), cis-ocimene co-eluted with ethyl hexanoate (UR) and linalool (UR) than those detected in the four individuals chosen as representatives of the ancestral Nacional variety (Figure 32).

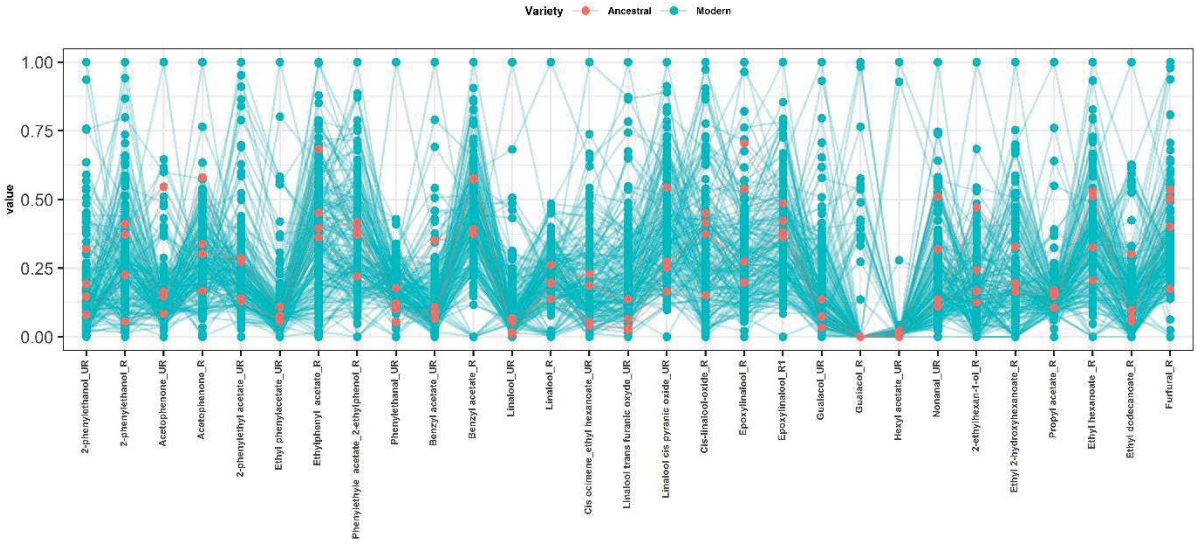


Figure 32: Parallel coordinate plot for the biochemical compounds known to have a floral note.

The values displayed correspond to the scores attributed to normalize the concentrations. In blue, the individuals belonging to the modern Nacional variety are shown and in red, the individuals close to the ancestral variety.

4.2.3-Concentration of compounds known to have a fruity note

A large proportion of individuals belonging to the modern Nacional variety have content of compounds known to have a fruity note, as ethanol (UR), benzaldehyde (UR), cis-ocimene co-eluted with ethyl hexanoate (UR), ethyl propanoate (UR and R), ethyl 2-methylpropanoate (UR), 2-methylpropylacetate (UR) ethyl 2-methylbutanoate (R), ethyl acetate (UR), 2,6-dimethylpyrazine (R), 2-ethyl-5-methylpyrazine (R) and linalool trans furanic oxide (UR) higher, than those detected in the four individuals chosen as representatives of the ancestral Nacional variety (Figure 33).

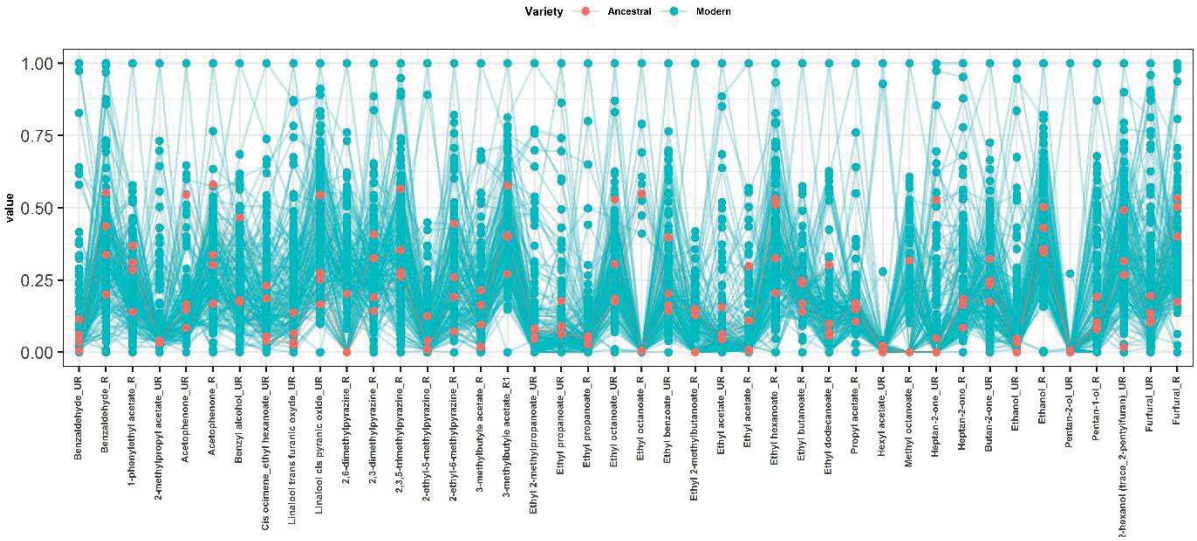


Figure 33: Parallel coordinate plot for the biochemical compounds known to have fruity notes.

The values displayed correspond to the scores attributed to normalize the concentrations. In blue, the individuals belonging to the modern Nacional variety are shown and in red, the individuals close to the ancestral variety.

4.2.4-Concentration of compounds known to have a vegetal or bitter note

A large proportion of the individuals belonging to the modern Nacional variety have a higher content of compounds known to have a vegetal note or bitterness, as Nonan-2-ol (UR), 2-pentylfuran co-eluted with ocimene (R), methanethiol (R), 2-acetylpyrrole (UR), caffeine (R) and theobromine (R), than those detected in the four individuals chosen as representatives of the ancestral Nacional variety (Figure 34).

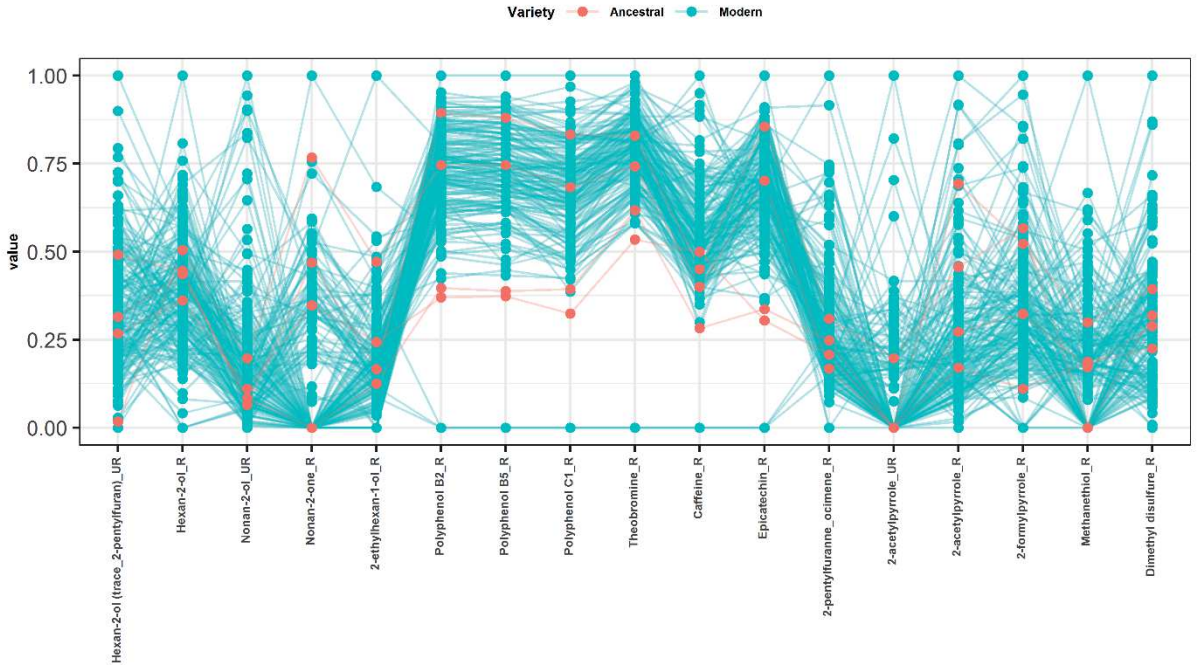


Figure 34: Parallel coordinate plot for the biochemical compounds known to have green notes or are involved in the bitterness. The values displayed correspond to the scores attributed to normalize the concentrations. In blue, the individuals belonging to the modern Nacional variety are shown and in red, the individuals close to the ancestral variety.

4.3-Determination of the origin of the alleles detected in the association zones that affect the fruity aroma.

For each significant association, the effect of each genotype was calculated. Using the three representatives of the three ancestral groups, an identification of the origin of the favourable genotypes for the fruity traits was made. The results for sensory scores and biochemical compounds involved in the fruity aroma was reported in Table 15.

Table 15: Synthesis of the specific origin of the alleles detected in the association zones linked to the fruity aroma.

Trait	AZ tot	Ame+	Cri+	Nac+	Ame-	Cri-	Nac-	Ame Cri Nac	Unknown origin
Fruity notes (2 notes)	9	**2**	**2**	0	0	0	**1**	4	0
Maillard reaction (6 compounds)	57	**15**	10	3	2	2	**9**	10	5
FA / S degradation (14 compounds)	238	49	**51**	21	8	12	**34**	42	21
L-phe degradation (5 compounds)	93	14	14	**20**	**15**	7	7	12	3
Monoterpene synt. (2 compounds)	26	**4**	3	**4**	1	**9**	2	62	65
Total alleles related to fruity notes	423	**84**	80	48	26	30	**53**	130	94

AZ: number of association zones tot, Ame+: number of the most favourable allele-specific to Amelonado, Cri+: number of the most favourable allele-specific to Criollo, Nac+: number of the most favourable allele-specific to Nacional, Ame-: number of worst-case allele-specific to Amelonado, Cri-: number of worst-case allele-specific to Criollo, Nac- : number of worst-case allele-specific to Nacional, Ame Cri Nac: number of most favourable alleles that can be traced back to the three ancestors.

The repartition among the genome of the origin of the most favourable genotype for each marker and significant association are represented in the map of Appendix 18 for fruity notes, pyrazines and furans compounds, in Appendix 19 for compounds involved in sugar and fatty acid degradation pathways, and Appendix 20 for compounds involved in the L-phenylalanine degradation pathway.

Taking into account all the association zones, Amelonado appears to be the ancestor that gives the highest number of favourable alleles for the presence of fruity notes (sensorial notes and aromatic compounds) and Nacional appears to be the ancestor that gives the highest number of unfavourable alleles to the presence of fruity notes.

4.4-Determination of the origin of the alleles detected in the association zones that affect the floral aroma.

For these significant associations, the same procedure as for the fruity association zones was used. The results for sensory scores and biochemical compounds involved in the floral aroma was reported in Table 16.

Table 16: Synthesis of the specific origin of the alleles detected in the association zones linked to the floral aroma.

Trait	AZ tot	Ame+	Cri+	Nac+	Ame-	Cri-	Nac-	Ame Cri Nac	Unknown origin
Floral notes (7 notes)	17	1	**7**	4	0	**1**	**1**	3	0
Monoterpene synt. (5 compounds)	45	**8**	3	7	**5**	2	4	11	2
L-phe degradation (6 compounds)	60	**13**	9	7	6	3	**8**	12	2
FA/ S degradation (4 compounds)	58	10	12	**13**	4	2	**9**	3	16
Maillard reaction (1 compound)	26	4	**5**	1	2	0	**6**	5	3
Total alleles related to floral notes	206	**36**	**36**	32	17	8	**28**	34	23

AZ: number of association zones tot, Ame+: number of the most favourable allele-specific to Amelonado, Cri+ : number of the most favourable allele-specific to Criollo, Nac+: number of the most favourable allele-specific to Nacional, Ame-: number of worst-case allele-specific to Amelonado, Cri-: number of worst-case allele-specific to Criollo, Nac-: number of worst-case allele-specific to Nacional, Ame Cri Nac: number of most favourable alleles that can be traced back to the three ancestors.

The repartition among the genome of the origin of the most favourable genotype for each marker and significant association are represented in the map of Appendix 21 for floral notes and monoterpene compounds, in Appendix 18 for hydrocarbon compounds, in Appendix 19 for compounds involved in sugar and fatty acid degradation pathways and in Appendix 20 for compounds involved in the L-phenylalanine degradation pathway.

Taking into account all the association's zones, Amelonado and Criollo appear to be the ancestors that gave the highest number of favourable alleles for the presence of floral notes (sensorial notes and aromatic compounds) and Nacional appears to be the ancestor that gives the highest number of alleles unfavourable for the presence of floral notes.

4.5-Determination of the origin of the alleles detected in the association zones that affect the spicy and cacao notes.

For these significant associations, the same procedure as for the fruity association zones was used. The results for sensory scores and biochemical compounds involved in the spicy and cocoa notes was reported in Table 17.

Table 17: Synthesis of the specific origin of the alleles detected in the association zones linked to spicy and cacao aromas.

Trait	AZ tot	Ame+	Cri+	Nac+	Ame-	Cri-	Nac-	Ame Cri Nac	Unknown origin
L-phe degradation (1 compound with spicy notes)	56	**19**	7	1	3	2	**10**	10	4
FA/ S degradation (1 compound with cacao notes)	11	1	**4**	**4**	0	**1**	**1**	0	0

AZ: number of association zones tot, Ame+: number of the most favourable allele-specific to Amelonado, Cri+ : number of the most favourable allele-specific to Criollo, Nac+: number of the most favourable allele-specific to Nacional, Ame-: number of worst-case allele-specific to Amelonado, Cri-: number of worst-case allele-specific to Criollo, Nac- : number of worst-case allele-specific to Nacional, Ame Cri Nac: number of most favourable alleles that can be traced back to the three ancestors.

The repartition among the genome of the origin of the most favourable genotype for each marker and significant association are represented in the map of Appendix 20 for compounds involved in the L-phenylalanine degradation pathway and Appendix 19 for compounds involved in sugar and fatty acid degradation pathways.

Amelonado appears to be the ancestors that gave a highest number of favourable alleles for the presence of spicy notes. Criollo and Nacional appeared to be the ancestors that gave a higher number of favourable alleles for the presence of cacao notes.

4.6-Determination of the origin of the alleles detected in the association zones that affect the presence of bitterness and astringency.

For these significant associations, the same procedure as for the fruity association's zones was used. The results for sensory scores and biochemical compounds involved in bitterness and astringency was reported in Table 18.

Table 18: Synthesis of the specific origin of the alleles detected in the association zones linked to bitterness, astringency and vegetal aroma.

Trait	AZ tot	Ame+	Cri+	Nac+	Ame-	Cri-	Nac-	Ame Cri Nac	Unknown origin
Bitterness (sensorial results)	11	**2**	0	0	0	**1**	**1**	7	0
Caffeine synthesis (2 compounds)	31	**7**	0	0	0	5	**6**	11	2
Polyphenols synthesis (4 compound)	53	**14**	1	0	0	8	**9**	21	0
FA/ S degradation (2 compounds - vegetal notes)	11	**2**	**2**	**2**	**1**	0	0	0	4
Maillard reaction (2 compounds - vegetal notes)	21	**6**	2	3	**3**	1	2	2	2
Sulfides (2 compounds - vegetal notes)	23	3	1	**6**	**6**	2	1	2	2
Astringency (Sensorial results)	2	0	**1**	0	0	0	0	1	0
Total alleles related to bitterness and vegetal notes	152	**34**	7	11	10	17	**19**	44	10

AZ: number of association zones tot, Ame+: number of the most favourable allele-specific to Amelonado, Cri+: number of the most favourable allele-specific to Criollo, Nac+: number of the most favourable allele-specific to Nacional, Ame-: number of worst-case allele-specific to Amelonado, Cri-: number of worst-case allele-specific to Criollo, Nac-: number of worst-case allele-specific to Nacional, Ame Cri Nac: number of most favourable alleles that can be traced back to the three ancestors.

The repartition among the genome of the origin of the most favourable genotype for each marker and significant association are represented in the map Appendix 22 for compounds involved in bitterness, astringency and vegetal notes.

Taking into account all the associations zones, Amelonado appears to be the ancestors that gave the highest number of alleles favourable to the presence of bitterness and vegetal notes (sensorial notes and non-volatile compounds) and Nacional appears to be the ancestor that gives the highest number of alleles unfavourable to the presence of bitterness and vegetal notes.

4.7-Common marker/biochemical traits associations with the most favourable genotype differing according to the considered trait

A total of 350 markers/biochemical trait associations have alleles with superior effects differing according to the associated traits (Appendix 23). These areas could be close to key genes involved in the same biosynthetic pathways leading to the synthesis and degradation of the co-localized compounds: the degradation of one of the compounds would allow the synthesis of the other, which would explain the alleles with different superior effects between these compounds.

For common marker/biochemical traits associations related to the fruity aroma, different alleles (and thus different origins) have been identified as having a superior effect on different compounds. This is the case for 101 markers in association with biochemical compounds belonging to sugar and fatty acid degradation pathways as well as for 29 markers in association with pyrazines. It was observed, for example, on chromosome 1 at position 4 066 816 pb where the homozygous A allele (originated from Criollo or Amelonado) has a superior effect on the presence of ethyl butanoate (R), synthesized through the breakdown of fatty acids and sugars, and on Fruity Dark tree fruit notes while the homozygous C allele (originated from Nacional) has a superior effect on the presence of 2-ethyl-5-methylfuran (UR), synthesized through Maillard reaction.

The same situation was observed for common marker/biochemical traits associations related to floral aroma: Forty-five markers are in this case. It was observed, for example, on chromosome 6 at position 3 359 215 bp that A homozygotes (originated from Nacional or Criollo) at this marker have a superior effect on the presence of 2-phenylethyl acetate (UR), whereas heterozygotes at this marker have a superior effect on the presence of cinnamaldehyde (R), both synthesized through the degradation of L-phenylalanine.

A similar situation is also observed for common marker/biochemical traits associations related to astringency and bitterness. One hundred and forty-one markers introduce this situation. On chromosome 2 at position 38 458 608bp heterozygotes (originate from Amelonado or Nacional) have a superior effect on the concentration of polyphenols B2 (R), while homozygotes for the A allele (originated from Amelonado) have a superior effect on the presence of 3-methylbutanal (R).

4.8-Heterosis effect

Some markers in associations show a higher effect of heterozygous genotype compared to homozygous genotypes, thus with a higher effect on the presence of biochemical compounds and sensorial notes provided by the heterozygous genotype, which is similar to what is called a heterosis effect in plant breeding. The results for biochemical compounds and sensorial notes are reported in Appendix 24.

The presence of the non-volatile compounds and the perception, by sensorial analysis, floral notes, of bitterness and astringency, are the traits most affected by the heterosis effect. Indeed, 32 percent of markers in association with floral notes and 31 percent in association with bitterness and astringency have a favourable effect when the locus is heterozygous (Table 19).

Table 19: Synthesis of markers with a genotype whose most favourable effect is due to the heterozygous genotype.

Trait	Total number of marker in association	Total number of markers with heterosis effect	% markers with heterosis effect
Bitterness / Astringency (sensory)	339	104	31%
Compounds with green notes	253	50	20%
Non-volatiles compounds (bitterness)	84	78	93%
Fruity notes (sensory)	39	7	18%
Compounds with fresh fruit notes	1111	74	7%
Compounds with dried fruit notes	208	26	12.5%
Floral notes (sensory)	37	12	32%
Compounds with floral notes	152	27	18%
Compounds with floral and fruit notes	528	15	3%

5-Discussion

In this work, it was possible to show that the aromatic qualities of the modern Nacional variety have been shaped throughout its domestication history over the last century. A large number of associations related to aromatic traits were previously detected by GWAS in a population representing the modern Nacional variety, a hybrid population with three main ancestors: Nacional, Amelonado and Criollo. Thanks to the genotyping of representatives of these ancestors, it was possible, for each of the association zones identified, to trace the origin of the alleles that have a superior effect on the presence of volatile aromatic compounds or the presence of sensorial notes detected by sensory analysis. During the domestication steps of the Nacional variety, Ecuadorian people introduced foreign *T. cacao* accessions of Trinitario types (hybrids between Criollo and Amelonado) from Venezuela and/or Trinidad one century ago. Natural hybridisations happened between the introduced Trinitario and the ancestral Nacional variety and trees resistant to witches broom disease were search among this population. This genetic mixing allowed the introgression in the ancestral Nacional, of a large number of alleles from the Trinitario trees favourable for the aromatic traits studied. Even though a dilution of the 'Arriba' aroma was reported ((Loor S. et al., 2009), this domestication process has also enriched the aroma of Nacional with fruity and new floral notes (tables 15 and 16).

The Criollo and the Nacional varieties are the two varieties known to be aromatic fine cocoa varieties (Luna et al., 2002; Ascrizzi et al., 2017; Rottiers et al., 2019), the Amelonado is known as standard cocoa with strong cocoa aromas (Assemat et al., 2005). Therefore, the assumption was that Criollo and Nacional would contribute the most favourable alleles for

floral and fruity notes. This was true for Criollo. However, Amelonado seems to also provide, many favourable alleles for the studied aromatic notes but involving pathways different from those of Nacional (as compounds belonging to the L-phenylalanine degradation pathway known to have a fruity note and compounds belonging to fatty acid and sugar degradations known to have a floral or a cacao note). This study was able to highlight the aromatic potential of Amelonado, which is certainly hidden by its bitterness.

During the study, we could observe that some favourable alleles did not come from any of the three ancestors. These markers represent 8% of the associated markers. This may be due to several reasons. The first hypothesis is that genotyping errors happened as classically observed using GBS. A second hypothesis is that we do not have the complete genotypes of the individuals at the origin of each genetic group (ancestral Nacional, Criollo and Amelonado). The individuals used to represent ancestors were identified by Motamayor et al., (2003) and Loor S. et al., (2009), using parentage analyses and these individuals have few heterozygous loci. Another hypothesis is that the ancestral Nacional variety may have been hybridised with another genetic group either by natural pollinators or human introduction without this being reported in the literature as suggested by Bartley (2005).

Some markers are associated with several traits and have the same allele with a superior effect on these different traits. These markers may be close to a key gene in the biosynthetic pathway of several of these compounds as is the case on chromosome 2 at position 8 193 772 bp where 2,3-dimethylpyrazine and 2-ethyl-5-methylpyrazine have the same allele at the marker giving superior effects. Two genes coding for enzymes involved in protein degradation are located at positions 8 043 343 bp and 8 135 517 bp of Chromosome 2. The second hypothesis is that the marker is near a transcription factor allowing the activation of the different biosynthetic pathways of the different compounds (Appendix 23).

It has also been observed that some markers are associated with several traits, this time with different alleles, favourable or unfavourable, depending on the traits in the association. It is possible that these compounds are part of the same biosynthetic pathway but for which the degradation of one of the two compounds is necessary for the synthesis of the other or that these compounds have the same precursor. This is the case on chromosome 6 at position 23 951 425 bp where the A allele from Amelonado has a superior effect on cinnamaldehyde content while the C allele from Nacional or Criollo has a superior effect on acetophenone, benzaldehyde and benzyl alcohol content (Appendix 23). They all have the same precursor which is (E)-cinnamic

acid. At position 23,996,474 bp there are two genes coding for alpha-beta hydrolases, which could explain this area of the association. Indeed, an enzyme with a hydroxylase function is required for the degradation of (E)-cinnamic acid and thus produce compounds such as cinnamaldehyde or acetophenone (Monisha et al., 2018).

The results of the origins of the alleles and the comparison between the individuals belonging to the Modern Nacional variety and those belonging to the ancestral Nacional variety seem to be consistent. The fruity "berries" note seems to be accentuated by the contribution of alleles from Criollo and Amelonado. Indeed, a higher concentration of various volatile compounds known to have a fruity taste, such as ethyl propanoate (UR and R), seems to be present in higher concentrations in individuals of the Nacional Modern variety. Associations related to fruity compounds derived from the degradation pathway of fatty acids and sugars, such as ethyl propanoate, have 49 favourable alleles coming from Amelonado and 51 from Criollo, compared with 21 from ancestral Nacional. It would therefore seem that these hybridisations have given Nacional more fruity notes.

Amelonado was also identified as the most important contributor to the presence of bitterness (non-volatile compounds and sensory analysis) and Nacional was identified as the most important ancestor contributing to the absence of bitterness. More bitterness was also observed in individuals from the modern Nacional variety, as well as a higher concentration of theobromine and caffeine in these individuals. It would therefore seem that the hybridisation of the ancestral Nacional with the Venezuelan Trinitario (hybrids between Criollo and Amelonado) has added bitterness. This increase probably conceals new fruity and floral notes. It also explains the decrease in aromatic value observed previously by par Loor S. et al., (2009).

Although Amelonado appears to be the source of the vast majority of associations for bitterness (34/152 of the alleles can come from this one), it is also the source of many favourable alleles for the presence of fruity (84/423) and floral notes (36/206). The areas of the genome responsible for the presence of bitterness and the presence of fruity or floral notes are not linked. It is therefore possible to improve each trait by selecting the most favourable set of alleles. It is thus possible to select trees with low bitterness and with more pronounced and/or different fruity and floral notes.

Conclusion

Domestication of the Nacional variety has made it possible to enrich its aromatic palette by natural hybridisations that happened between the ancestral Nacional variety and the

introduced Venezuelan Trinitario. This study highlighted the interest to cross cocoa genotypes from different genetic origins to improve varieties' aroma, taking into account the presence of volatile aroma compounds and sensory evaluations during the breeding steps. It is therefore important that breeders consider these parameters at early stages in their breeding schemes, at the same time as other criteria of yield and disease resistance, to increase the aroma value of cocoa and avoid the erosion of the aroma potential during the breeding process. Introgression of the Amelonado genome into an aromatic variety can have a positive impact on aromatic traits if controlled. Genomic prediction, as well as marker-assisted selection, could be tools allowing predicting the aroma value of cocoa trees and thus could help to select cocoa trees with the best aromatic profile.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

EC, CL, RGLS, conceived the experiment; JCJ, AS conducted biochemical analyses; ES carried out sensorial analyses; OF carried out DNA experiments; KC, JCJ, AS, RB, CL, FD, SA, XA analysed data; KC, RB, CL wrote the manuscript.

Funding

The study was funded by the United States Department of State (U.S. Foreign Ministry); the U.S. Embassy, Quito; the U.S. Department of Agriculture (USDA-ARS); the MUSE Amazcacao project with the reference ANR-16-IDEX-0006.

Acknowledgement

We thank the USDA and the I-Site MUSE for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" program with the reference ANR-16-IDEX-0006.

L'ensemble des études sur les déterminismes génétiques et biochimiques des arômes du Nacional moderne a montré que le génotype avait une influence sur les notes aromatiques perçues par le consommateur. De plus, l'origine génétique ainsi que la domestication et la sélection ont également joué un rôle sur l'élaboration du profil aromatique final de la variété de Nacional moderne. Dans le quatrième chapitre, il a été montré que chaque groupe d'ancêtres avait apporté des allèles favorables au développement des arômes fins du cacao (floraux et fruités) mais également des allèles favorables au développement de l'astringence ou de l'amertume. Cependant, les arômes fins et les notes non appréciées ne semblent pas liées génétiquement et il est donc possible de les séparer lors des processus de sélection.

Dans un but de sauvegarde de la diversité génétique des cacaoyers aromatiques, des prospections en Amazonie Équatorienne ont été effectuées dans la zone d'origine du Nacional préalablement identifiée par l'équipe. Ces arbres ont été mis en collections afin de constituer de nouvelles ressources génétiques permettant de diversifier la base génétique utilisée dans les programmes de sélection actuels. Dans le but de les utiliser, il a été entrepris de caractériser génétiquement et biochimiquement le potentiel et la diversité aromatique de ces arbres natifs d'Amazonie, et d'étudier l'architecture génétique de leurs arômes.

Dans le chapitre 5 est présentée une étude de GWAS portant sur l'ensemble des composés volatils identifiés par GC-MS dans cette nouvelle population, ainsi que sur quelques composés non-volatils identifiés par NIRS dans les fèves marchandes. Cette étude GWAS porte aussi sur l'ensemble des résultats d'analyses sensorielles effectuées sur les liqueurs de cacao. Des résultats communs avec les précédentes analyses ont été observés. Cependant, de nouveaux composés ainsi que de nouvelles notes aromatiques ont aussi été identifiés. Ces résultats sont présentés dans le chapitre 5.

Chapitre 5: Variabilité et déterminants génétiques des arômes des cacaoyers natifs du sud de l'Amazonie équatorienne

Chapitre 5: Variabilité et déterminants génétiques des arômes des cacaoyers natifs du sud de l'Amazonie équatorienne

Variability and genetic determinants of cocoa aromas in trees native to South Ecuadorian Amazonia

Kelly Colonges[1,2,3,4*†], Rey Gastón Loor Solorzano[5†], Juan-Carlos Jimenez[5], Marie-Christine Lahon[3,4], Edward Seguine[6], Darío Calderon[5], Cristian Subia[5], Ignacio Sotomayor[5], Fabián Fernández[5], Marc Lebrun[3,4], Olivier Fouet[1,2], Bénédicte Rhoné[1,2], Xavier Argout[1,2], Pierre Costet[8], Claire Lanaud[1,2*], Renaud Boulanger[3,4*]

1 Cirad, UMR AGAP, F-34398 Montpellier, France.

2 AGAP Institut, Univ Montpellier, Cirad, INRAE, Institut Agro, Montpellier, France.

3 Cirad, UMR Qualisud, F-34398 Montpellier, France.

4 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

5 Instituto Nacional de Investigacion Agropecurias, INIAP, Ecuador.

6 Seguine Cacao/Guittard Chocolate Co, Arroyo Grande, CA, United States

† these authors participate equally in this work

Keywords : *T. cacao*, GWAS, aromatic compound, aroma, Ecuadorian Amazonia, genetic resources

1-Abstract

Ecuador is known worldwide for its fine cocoa from the Nacional variety. Currently, cocoa trees belonging to the modern Nacional variety (hybrids between the ancestral Nacional and Trinitario) are grown. To date, the ancestral Nacional variety is no longer cultivated, and little information is available about it. To enlarge the genetic resources related to this ancestral variety, several surveys were carried out in its area of origin. The 202 trees resulting from these surveys in the Ecuadorian Amazon were characterised for their aromatic and genetic traits. Based on genotyping data, phenotypic traits related to volatile compounds and sensory analyses established for each tree of the population, a GWAS study was carried out to study the genetic and biochemical bases of the aroma traits of this population and to better exploit them in breeding programs. Many areas of marker/aroma traits associations were identified as well as candidate genes, thanks to the available Criollo cocoa genome sequence V2. Comparisons with previous GWAS analyses of trees belonging to the modern Nacional variety were also undertaken.

2-Introduction

Theobroma cacao belongs to the Malvaceae family (Bayer and Kubitzki, 2003) and is originated from the humid tropical regions of the north of South America, and mainly from the Amazonian basins.

T. cacao is a tree of great agronomic and economic interest. Indeed, it is the only worldwide source of chocolate, whose consumption is constantly increasing and will increase by 20% in 2025 (Cilas, 2020). It provides a large panel of various aromas (Andújar et al., 2012; Tuenter et al., 2018) highly appreciated, but also offers benefits for human health thanks to its richness in polyphenols. Cacao can be classified into two main classes: standard cocoa characterized by strong notes of cocoa, and fine flavour aromatic cocoa characterized by fruity and/or floral aromas (Cook and Meursing, 1982; Sukha et al., 2008). Fine aromatic cocoa represents about 5% of the current world production. However, its consumption is constantly increasing and sought after by chocolate-makers looking for new flavours, and it represents important economic niches for tropical countries producing this type of cocoa.

The varieties of fine aromatic cocoa currently mostly cultivated are Criollo, known for its fruity aromas but which is less cultivated due to its low vigour and resistance to disease (Cheesman, 1944), the Trinitario trees, which are hybrids between the Amelonado and Criollo

genetic groups, and finally, the Nacional (endemic to the equator) known for its floral notes, which are called Arriba (Luna et al., 2002; Loor S. et al., 2009).

The latter is only grown in Ecuador. The trees currently grown are part of the Nacional modern population variety, which are hybrids between the ancestral Nacional variety and Venezuelan cultivars corresponding to Trinitario genotypes (Bartley, 2005). Using molecular markers, Loor et al., (2009) confirmed this hybrid type.

Despite its main cultivation along the pacific coast, the Nacional variety originated from the Southern part of the Ecuadorian Amazonia (Loor et al., 2012)

Surveys were undertaken in the domestication centre of the Nacional variety, located in the South part of the Ecuadorian Amazonia, to search for cocoa trees related to the ancestral Nacional variety (Loor S. et al., 2009; Loor S. et al., 2015). These expeditions aimed to collect and save genotypes related to the ancestral Nacional, potentially more variable and offering new genetic resources to improve the Nacional modern variety or create new aromatic varieties.

Recently, studies on the genetic determinism of fine flavours of cocoa trees from the modern Nacional variety have been carried out. A first study focused on floral aromas (Colonges et al., 2021b). This study was able to identify two main biosynthetic pathways involved in the synthesis of floral aromas: the monoterpene biosynthetic pathway and the L-phenylalanine degradation pathway. The second study investigated the genetic determinism of fruity aromas (Colonges et al., 2021d) and allowed us to identify five main biosynthetic pathways involved: the monoterpene biosynthetic pathway, the L-phenylalanine, the fatty and sugar degradation pathways and the synthesis of Maillard precursors (for the production of pyrazines).

To study the variability and genomic determinants of the cocoa aromas of the Amazonian population, a complete characterisation of the genetic and biochemical traits of the cocoa trees collected on Amazonia was undertaken. To this end, volatile compounds of the fermented and dried beans of each Amazonian tree were evaluated by GC-MS. Sensory analyses of the liqueurs were also carried out. The molecular characterisation of this population, obtained by GBS, was used to conduct a GWAS (Genome-Wide Association Study) on all these traits to decipher their genetic determinants.

T. cacao has a small genome now fully sequenced from representatives of two varieties: Criollo (Argout et al., 2011; Argout et al., 2017) and Amelonado (Motamayor et al., 2013).

Then, the available genome sequences allowed to identify candidate genes in the association regions, potentially involved in these aromas.

3-Materials and methods

3.1-Vegetal material

Two hundred and two cocoa trees were used for this study. They came from surveys carried out in the South part of Ecuadorian Amazonia corresponding to the Zamora Chinchipe province and the presumed domestication centre of the ancestral Nacional variety in Ecuador (Loor S. et al., 2012; Loor S. et al., 2015). The collected trees were planted in two INIAP experimental centres: in Pichilingue (EET-P) and Domono; and in an experimental plot of an agricultural college in Pangui.

3.2-Harvest and micro fermentations

The pods were harvested at maturity in the different growing locations. Micro-fermentations took place in a unique place: Domono within 24 hours of harvest. Micro-fermentations were carried out under the most homogeneous conditions possible. The cocoa beans of each genotype were placed in delicate laundry nets. They were then distributed over four floors in the middle of the mass of modern Nacional cocoa beans. At 24 and 72 hours of fermentation, stirring was performed. At each stirring, the bags of beans at the bottom were placed at the top and those in the middle-low position were placed in the middle-high and vice versa. After 4.5 days, the beans were taken out of the net and dried separately in a greenhouse. During the entire fermentation process, button pills recorded the temperature to check the homogeneity of the transitions of the different stages of fermentation in the different cases. When the moisture content was less than or equal to 8% the beans were considered dry and were placed under vacuum until biochemical analyses.

3.3-Biochemical analysis of volatile compounds

GC/MS analyses were conducted according to the condition described by Assi-Clair et al., (2019).

3.4-Sensorial analysis

One hundred and fifty-nine genotypes were characterized by sensory analyses based on blind tastings carried out on three repetitions per sample. The tastings were carried out on cocoa liquor. The cocoa liquor corresponds to merchant cocoa (dried fermented beans) which have been roasted and crushed. Sensorial notes were judged with a score ranging from zero (no notes

detected) to ten according to ISCQF, (2020) protocol. We used the average of the three replicates for the phenotype of the GWAS analysis.

3.5-DNA extraction and SNP genotyping

DNA extraction was conducted according to Risterucci et al. (2000) protocol. DNA samples were genotyped by sequencing (GBS) using DArTseq (Diversity Arrays Technology Sequencing) technology (Kilian et al., 2012) and carried out by the DArT company. Reads were aligned with the V2 sequence of the Criollo genome (Argout et al., 2017). SNP Markers with unknown locations were discarded for analysis.

3.6-Diversity and structure analysis

For the calculation of diversity and structure of the population, 51 trees taken as a control was used (table 20).

Table 20 : Complete list of the control used for the diversity and structure population calculation

Amelonado	Criollo	Nacional	Iquitos	Guiana	Nanay	Maranon	Purus	Contanmana	Caqueta	Curaray
Catengo	B97M	LCTEEN91	IMC107	GU119	NA127	PA169	LCTEEN220	SCA12	EBC30	LCTEEN312
MAT16	COL10	SNA1001	IMC65	GU134	NA184	PA126	LCTEEN368	SCA5	EBC48	LCTEEN227
VEN20	HE4	SNA1003	IMC76	GU156	NA191	PA303		SCA6	EBC91	LCTEEN327
SIC840	LAN28b	SNA604	IMC31	GU3	NA672	PA296		U49	RB39	LCTEEN333
SIC23B	SJU1		IMC48	GU29	NA697	PA293		SCA11	RB46	LCTEEN57

The phylogenetic tree was generated using DARwin software (Perrier and Jacquemoud-Collet, 2006). The genetic distances were calculated using the Dice coefficient and the Neighbour-Joining method. The bar graph of structure was generated with the R package LEA (Frichot et al., 2014; Frichot and François, 2015).

3.7-Linkage disequilibrium calculation

Linkage disequilibrium (LD) was calculated with Haploview 4.2 (Barrett et al., 2005) according to Sardos et al., (2016) protocol. LD decay graphical representation was made with the "ggplot2" R package according to Sardos et al., (2016) protocol.

3.8-GWAS analysis

SNPs with no missing data and a MAF greater than 5% were selected for the GWAS. The final dataset is composed of 5337 SNP markers. The data set is available in the database tropgene (http://tropgenedb.cirad.fr/tropgene/JSP/interface.jsp?module=COCOA) at the study named "Cocoa_Amazonie_Ecuador_aroma".

A GWAS analysis was performed using SNP markers and biochemical (202 genotypes x 5337 markers) and sensory (159 accessions x 5337 markers) traits using TASSEL v5.

For all the traits, a mixed model (MLM) was carried out with a structure matrix, determined by running a PCA (principal component analyses integrated with TASSEL v5 software), considered as a fixed effect, and also with a kinship matrix considered as a random effect as covariates to control the false-positive rate. The option of not compressing and re-evaluating the components of variance for each marker was chosen. The kinship matrix using the Identity by State (IBS) pairwise method proposed by Tassel v5 was established.

The threshold was determined using the R package Simple M based on the Bonferonni correction (Gao et al., 2008; Gao et al., 2010). The threshold corresponds to a p-value of $1{,}676 \times 10^{-5}$.

The physical maps with the representation of association areas were created using SpiderMap v1.7.1 software (Rami, 2017). The size of the dots is correlated to the R2.

The identification of candidate genes was performed in a region of 300Kbp (on either side of the associated marker) using the *Theobroma cacao* genome sequence V2 (Argout et al., 2017).

4-Results

4.1-Phenotypic traits studied

4.1.1-Biochemical traits

To study the biochemical determinism of cocoa flavours, the volatile compounds contained in the fermented and dried beans were determined by Solid Phase MicroExtraction (SPME) coupled to GC-MS. All the compounds identified and their already known aromas are shown in appendix 25.

Strong positive correlations (between 0.8 and 1) were identified between several volatile compounds. In contrast, no strong negative correlations (between -0.8 and 1) were identified (Figure 35).

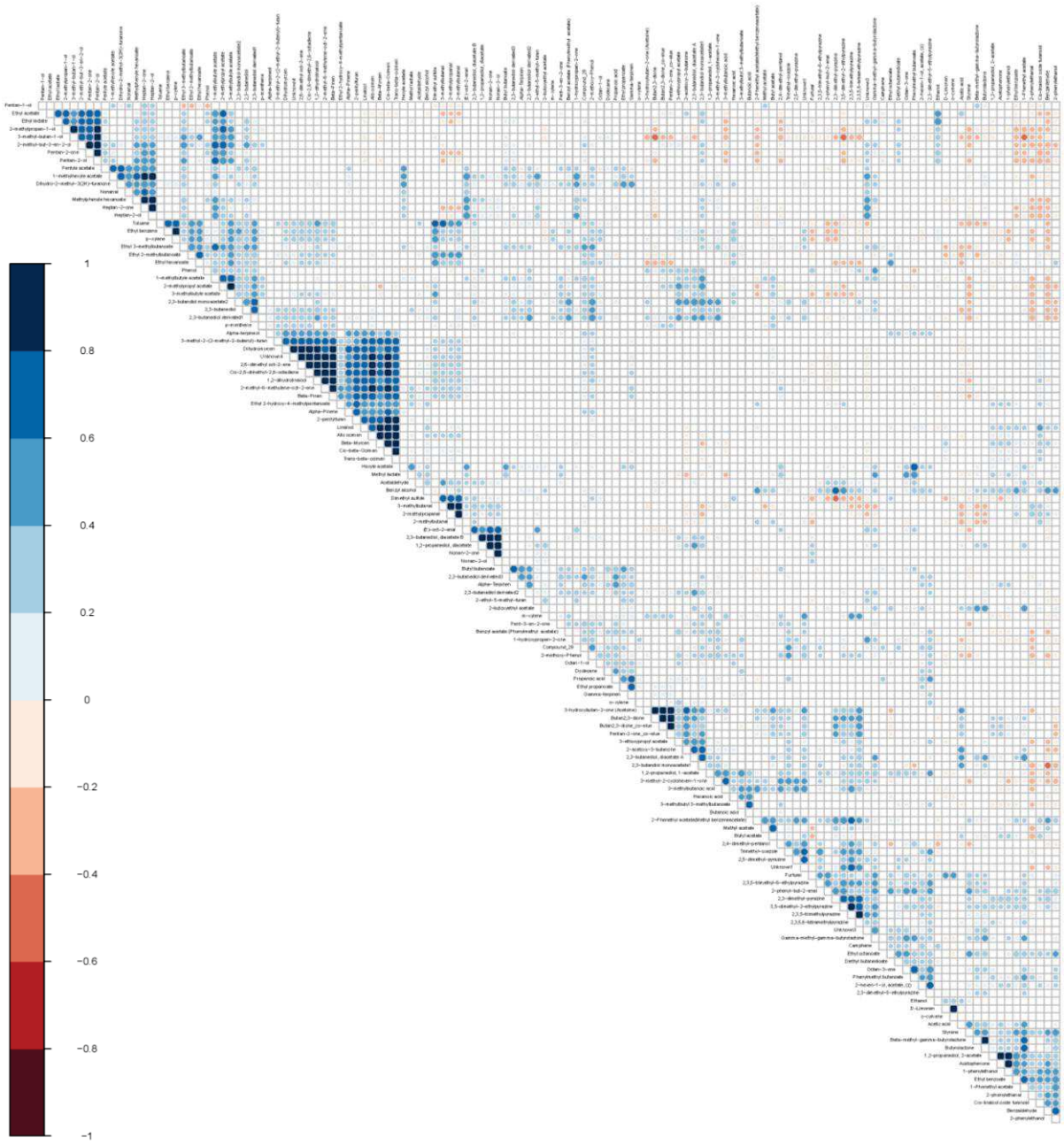


Figure 35: Significant correlation matrix for the biochemical compounds
Correlation matrix between the biochemical compounds observed in fermented and dried beans. The correlations were calculated by the Pearson method. They are organised by hierarchical clustering order. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient. The areas of circles correspond to a p-value of correlation coefficients. The p-value threshold for a significant correlation is 0.05. The different shades of blue represent a positive correlation coefficient while the different shades of red represent a negative correlation coefficient. The intensity of the colour depends on the strength of the R2 correlation coefficient. The scale on the right indicates the interpretations of different colours.

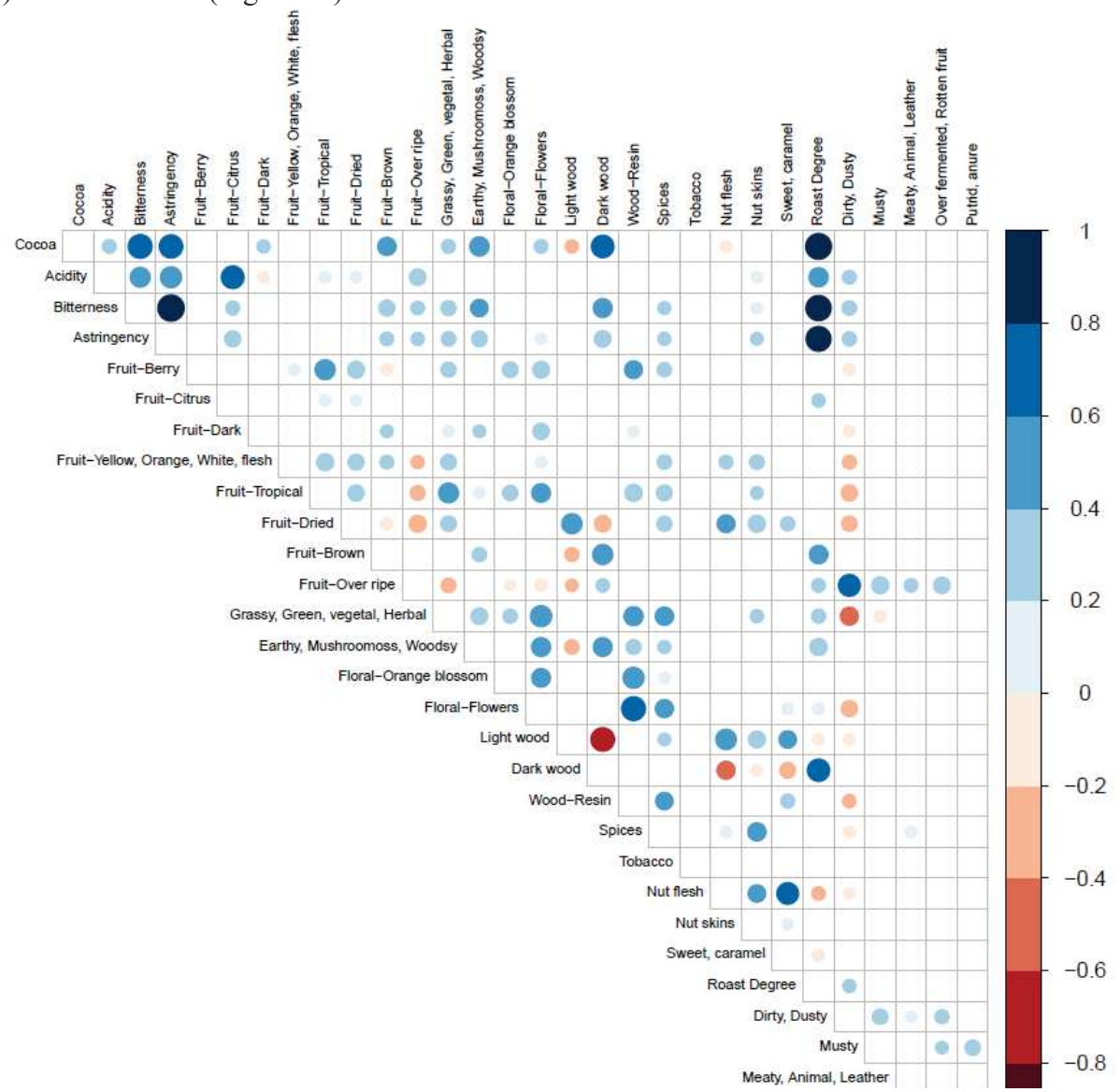
Principal component analyses (PCA) were performed. These analyses show a continuous variation within the population for the concentrations of volatile compounds (appendix 26A). Axis 1 is mainly influenced by the concentration of ethyl lactate, butan-2,3-

dione and 2-phenethyl acetate. Axis 2 is mainly influenced by the concentration of 2-butoxyethyl acetate, decane and β-methyl-γ-butyrolactone.

4.1.2-Sensorial traits

A sensory study of the liquors of each cocoa genotype was conducted in which 33 criteria were evaluated (appendix 27). The three notes: savoury/umami, mouldy and smoky, were not detected in any of the genotypes.

Three strong positive correlations (between 0.8 and 1) could be observed between the different sensory characteristics. In contrast, no strong negative correlations (between -0.8 and 1) were identified (Figure 36).



Correlation matrix between the sensorial profiles determined in cocoa liquor. The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient. The areas of circles correspond to a p-value of correlation coefficients. The p-value threshold for a significant correlation is 0.05. The different shades of blue represent a positive correlation coefficient while the different shades of red represent a negative correlation coefficient. The intensity of the colour depends on the strength of the R2 correlation coefficient. The scale on the right indicates the interpretations of different colours.

The results of the PCA performed on the sensory data also show a continuous variation within the population (appendix 26B). Axis 1 is mostly influenced by the spice-tobacco score, the wood-dark wood score and the meaty-animal leather score. Axis 2 is mostly influenced by the note "floral - flowers", the note "wood - resin" and the note "fruit - dark".

4.2-Genetic diversity and population structure

Using SNP markers, the genetic diversity and structure of the population could be studied. The cocoa population native from the Zamora Chinchipe province studied is mostly close to the Curaray and Nacional genetic groups. Some individuals are also close to the Caqueta, Contamana and Iquitos genetic groups (Figure 37). A subgroup composed of different trees from the Pangui region (PAN and PGI) seems to be more distant from the other individuals (Figure 37). These trees all originate from the Pangui region and were collected during the 2010 survey at the north of this region in contrast to the trees named 'PANGUI' (PGI) which were also collected in the Pangui region during the 2016 survey but more around the city of El Pangui (Loor S. et al., 2015).

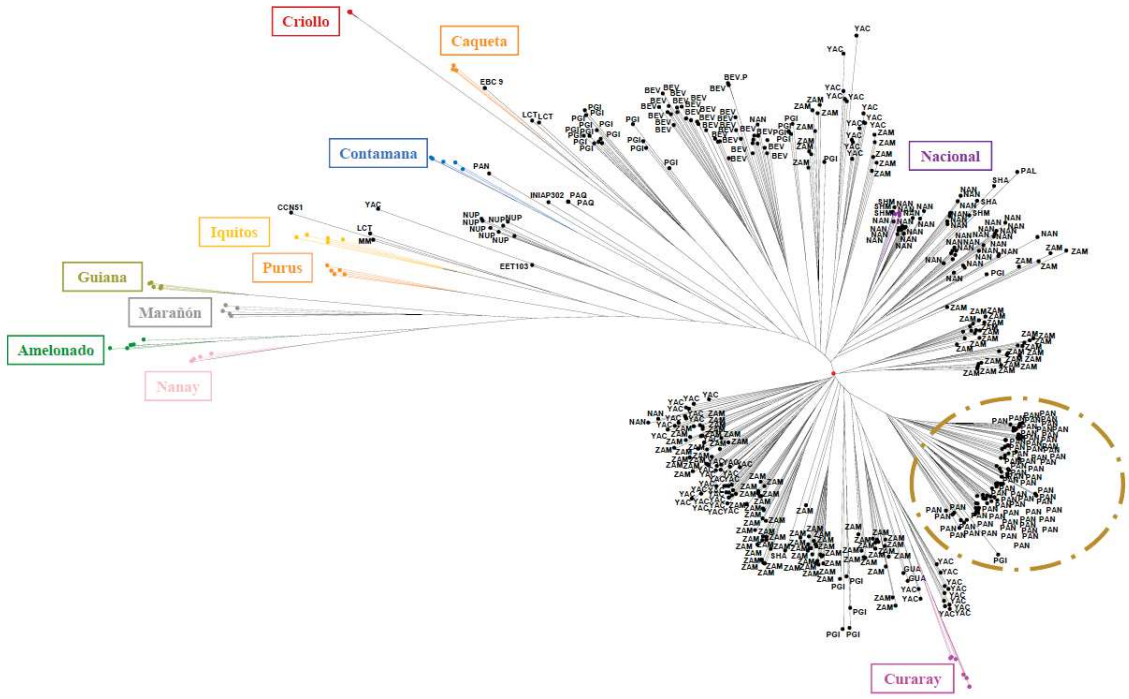


Figure 37: Phylogenetic tree representing the genetic diversity of the studied population.

The phylogenetic tree, established with the Darwin software, represents the genetic diversity of the individuals in the studied population (shown in black), with the known cocoa genetic groups (in colour) taken as witness. The name of each Amazonian location of collect is indicated as follow, and according to Loor S. et al., (2015) BEV: BEVI, GUA: GUAI, LCT: LCTEEN, NAN: NANK, NUP: NUPA, PGI: PANGUI 2, PAL:PALANDA, PAN: PANGUI 1, SHA: SHAI, SHM: SHAM, YAC:YACU, ZAM: ZAMORA. Abbreviation : P :Pichilingue.

The LD of each chromosome is represented in appendix 28. For each chromosome, we can observe that the r^2 value decreases by half around 600kb (represented by the red dotted

line). Given that 1 MB corresponds to around 2 cM (Loor S., 2007), the LD of this population is around 1,2 cM. We decided to use this 600 kb limit to determine the confidence interval of associations. For each positive marker we reported an association area of plus or minus 300 kb, i.e. an association area of 600 kb. If two or more markers have overlapping confidence intervals, they are grouped in a single association zone. The lowest and highest number of bp of the grouped markers represents the confidence intervals of this zone.

4.3-Genome-Wide Association Study

SNPs with no missing data and a MAF greater than 5% were selected for the GWAS. The SNPs are relatively well distributed along the 10 chromosomes of *T. cacao*. However, a decrease in marker density is observed in the centromeric and pericentromeric regions (Figure 38).

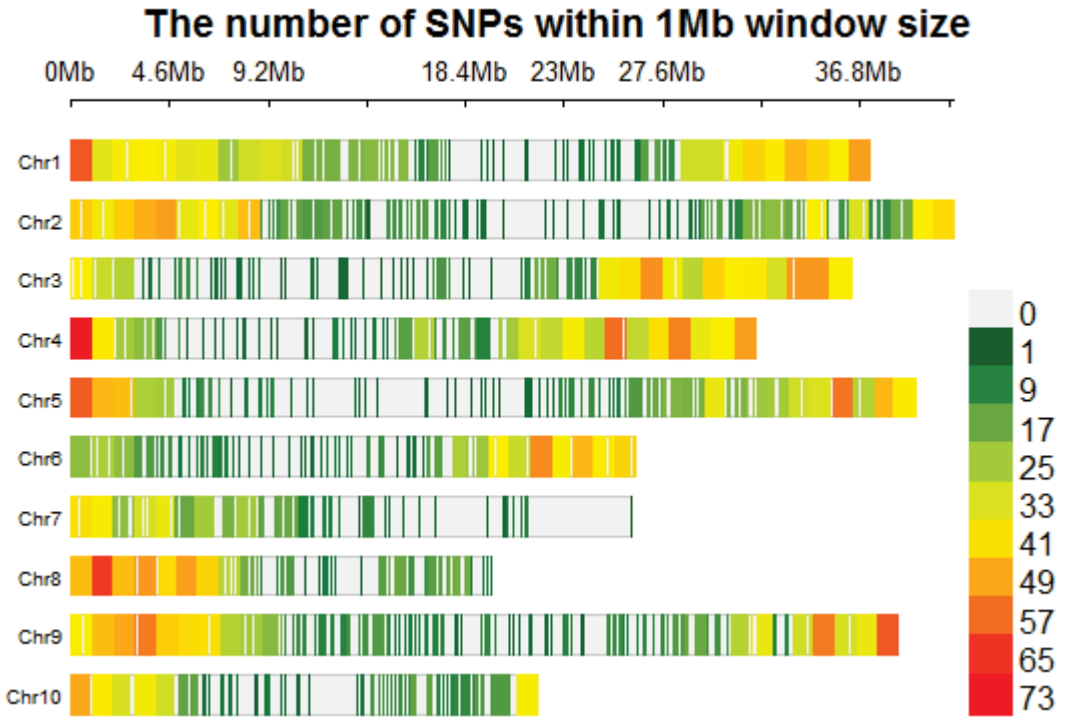


Figure 38 : Distribution of markers along the ten chromosomes of T. cacao.
Distribution of markers along the ten chromosomes of T. cacao. The graph shows the distribution of markers along the ten chromosomes. The density is calculated on a 1Mb window. The areas without markers are shown in grey. The weakly marked areas are in green and the strongly marked areas are in red. A colour gradient between green and red represents the marking gradient.

GWAS analyses were performed with different methods. The method with the fewest false positives was selected. This was the MLM method combined with the PCA data from the genetic data and the kinship matrix as a co-variate. All areas of significant associations are

shown in appendix 29. The number of markers associated and the total number of associations for each trait are reported in the appendix 30.

4.3.1-Identification of significant association areas in relation to floral notes

A total of 39 association areas were detected to compounds known to have a floral note and 9 other associations were detected linked to the floral notes, detected by sensory analysis (appendix 29 and 30). Associations were detected on all chromosomes except chromosome 4 (appendix 29). Seven areas of co-locations between different biochemical compounds as well as between biochemical compounds and sensory notes were observed (table 21). Two volatile compounds seem to be more related to the presence of orange blossom notes: the cis-beta-ocimene and the o-xylene.

Table 21:Co-locations between traits related to floral notes

Chromosome	Position (bp)	Traits
1	30 025 697 – 30 625 697 30 129 690 – 30 729 690	Cis-beta-ocimene Dihydromyrcen
5	0-799 633 0 – 659 385 0 – 574 697	Dihydromyrcen 1,2-dihydrolinalool Allo ocimen
5	33 897 871 – 35 986 881 34 315 551 – 34 915 551	Floral Orange blossom note Cis-beta-Ocimen
6	21,655,681 – 22 484 229 22,391,993 – 22 991 993	Dihydromyrcen Benzyl acetate (Phenylmethyl acetate)
7	3,129,126 - 3,729,126 3,722,709 - 4,322,709	1,2-dihydrolinalool Benzyl acetate (Phenylmethyl acetate)
9	93,947 - 693,947	Floral Orange blossom note o-xylene
9	29,047,112 - 29,647,112 29,194,885 - 29,794,885	Allo ocimen o-xylene

4.3.2-Areas of association detected in relation to fruity notes

A total of 194 association zones were detected in relation to volatile compounds known to have a fruity note and 39 linked to fruity notes detected by sensory analysis (appendix 31 - 33). Areas of association were detected on all chromosomes (appendix 29). Among the 56 areas of co-locations between the different traits, 26 of them display co-locations between volatile compounds and sensory notes (table 22).

Table 22: Co-location between traits related to fruity notes

Chromosome	Position	Traits
1	227,662 - 827,662 227,699 - 827,699	Heptan-2-ol Pentyl acetate
1	10,403,809 - 11,003,809	2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
1	26,058,831 - 26,658,831 26,551,903 - 27,151,903	Diethyl butanedioate Trimethyl-oxazole
1	28,523,447 - 29,123,447	2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
1	30,025,697 - 30,625,697 30,129,690 - 30,729,690 30,143,772 - 30,743,772	Heptan-2-ol 2-pentylfuran Heptan-2-ol
1	35,179,334 - 35,779,334 35,664,409 - 36,264,409	Ethyl 2-hydroxy-4-methylpentanoate Diethyl butanedioate
1	36,759,829 - 37,359,829 36,806,538 - 37,406,538	Over fermented Rotten fruit Diethyl butanedioate
2	507,611 - 1,107,611 703,576 - 1,303,576	Nonan-2-one Fruit Berry
2	3,329,347 - 3,929,347	Diethyl butanedioate and limonen
2	9,227,909 - 9,827,909 9,688,881 - 10,376,829	Ethyl 2-hydroxy-4-methylpentanoate 2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
2	30,661,408 - 31,261,408 31,061,092 - 31,661,092	Diethyl butanedioate Fruit Dark and Hexyl acetate
2	32,066,719 - 34,886,549 32,066,719 - 33,281,856	Fruit Dark Hexyl acetate
2	33,157,986 - 33,757,986 33,254,963 - 33,854,963	Diethyl butanedioate 1,2-propanediol diacetate
2	36,057,073 - 36,999,677 36,286,693 - 36,886,693	Limonen Fruit Dark
2	36,286,693 - 36,886,693 36,987,390 - 37,587,390 37,493,639 - 38,093,639 37,618,909 - 38,574,817 38,134,117 - 38,734,117 38,212,257 - 38,812,257 38,649,264 - 39,798,805 39,198,805 - 39,802,081 39,202,081 - 39,802,081 39,500,916 - 40,100,916 39,555,481 - 41,295,090 40,591,750 - 41,191,750	Hexyl acetate 1,2-propanediol diacetate Pentyl acetate Hexyl acetate Nonan-2-one Diethyl butanedioate Hexyl acetate Phenylmethyl butanoate Hexyl acetate Butyl butanoate Diethyl butanedioate Heptan-2-ol
3	1,367,283 - 1,967,283 1,536,416 - 2,136,416	2,3,5,6-tetramethylpyrazine Pentyl acetate
3	21,404,737 - 22,004,737	2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
3	25,587,254 - 26,187,254 25,686,659 - 26,286,659	Nutty Nut flesh Nonan-2-one
3	28,711,521 - 29,311,521 28,876,174 - 29,476,174	Diethyl butanedioate Nutty Nut flesh
3	31,255,106 - 31,855,106 31,784,376 - 32,384,376	Ethyl 2-hydroxy-4-methylpentanoate 3-methylbutyl 3-methylbutanoate
3	33,124,726 - 33,724,726 33,459,519 - 34,059,519	Ethyl 2-hydroxy-4-methylpentanoate Heptan-2-ol
4	1,523 - 601,523	2-pentylfuran and Nutty Nut flesh
4	22,858 - 1,000,690 400,690 - 1,813,500	Diethyl butanedioate 2,3,5-trimethylpyrazine
4	2,474,103 - 3,074,103 2,858,463 - 3,458,463	2,3,5-trimethylpyrazine 1,2-propanediol diacetate
4	18,763,963 - 19,363,963 18,887,669 - 19,487,669	Nutty Nut flesh 1,2-propanediol diacetate
4	21,039,263 - 21,639,263 21,043,930 - 22,066,066	Trimethyl-oxazole 2,3,5-trimethylpyrazine
4	22,371,100 - 23,469,866 23,017,920 - 24,748,366 23,512,495 - 24,112,495	2,3,5-trimethylpyrazine Hexyl acetate 2,3,5-trimethylpyrazine
4	24,903,361 - 25,503,361 25,346,030 - 25,946,030 25,429,071 - 26,029,071 25,571,632 - 26,615,620 26,026,282 - 27,175,785 26,687,962 - 27,716,239 26,744,417 - 27,908,710 27,366,259 - 28,407,791 27,459,054 - 28,503,851	2,3,5-trimethylpyrazine Trimethyl-oxazole Fruit Berry Pentyl acetate Trimethyl-oxazole 2,3,5-trimethylpyrazine 2,3,5,6-tetramethylpyrazine Hexyl acetate 1,2-propanediol diacetate
4	31,091,855 - 31,691,855 31,555,304 - 32,155,304	1,2-propanediol diacetate 2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
5	0 - 703,249 0 - 354,149	Ethyl 2-hydroxy-4-methylpentanoate Ethyl lactate
5	2,004,453 - 2,604,453 2,117,256 - 3,198,438	2,3,5,6-tetramethylpyrazine Diethyl butanedioate

	2,134,166 - 3,049,302 2,176,544 - 2,776,544	2,3,5-trimethylpyrazine and trimethyl-oxazole Heptan-2-ol
5	18,471,512 - 19,071,512	2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
5	26,290,212 - 26,890,212 26,366,869 - 26,966,869	Diethyl butanedioate 3-methylbutyl 3-methylbutanoate
5	27,922,159 - 28,522,159 28,319,215 - 28,919,215 28,986,415 - 29,586,415 29,310,859 - 30,773,075	1,2-propanediol diacetate Trimethyl-oxazole 1,2-propanediol diacetate 2,3,5-trimethylpyrazine
5	33,394,588 - 33,994,588 33,394,588 - 34,288,858 33,897,871 - 35,143,004 34,315,551 - 34,915,551	2-pentylfuran Ethyl 2-hydroxy-4-methylpentanoate Fruit Berry 2-pentylfuran and ethyl 2-hydroxy-4-methylpentanoate
5	36,031,917 - 36,631,917 36,215,073 - 36,836,724 36,215,073 - 36,815,073	Nutty Nut flesh 2,3,5-trimethylpyrazine 3-methyl-2-cyclohexen-1-one
6	6,068,335 - 6,668,335 6,680,133 - 7,280,133	Fruit Dark 1,2-propanediol diacetate
6	19,581,377 - 20,181,377 19,969,328 - 20,569,328	Ethyl 2-hydroxy-4-methylpentanoate 2,3,5,6-tetramethylpyrazine
6	22,931,122 - 23,866,379 23,806,262 - 24,406,262	2,3,5-trimethylpyrazine Nutty Nut flesh
7	0 - 406,706 0 - 363,832	2-pentylfuran Over fermented Rotten fruit
7	1,249,187 - 1,849,187 1,713,420 - 2,313,420 2,216,312 - 2,816,312	1,2-propanediol diacetate Nonan-2-one Fruit Dark
8	1,550,327 - 2,150,327 1,594,562 - 2,194,562 1,827,824 - 2,463,216	D-Limonen 2-pentylfuran Heptan-2-ol
8	2,903,311 - 3,503,311 2,943,415 - 3,543,415 2,982,287 - 3,582,287	2-pentylfuran Pentyl acetate Nutty Nut flesh
8	4,073,313 - 4,673,313 4,671,183 - 5,271,183	1,2-propanediol diacetate Heptan-2-ol
8	5,430,993 - 6,030,993 5,600,260 - 6,200,260	1,2-propanediol diacetate Nutty Nut flesh
8	7,259,551 - 7,859,551 7,853,804 - 8,453,804	Nutty Nut flesh 1,2-propanediol diacetate
9	809,358 - 1,620,668 1,158,899 - 1,758,899 1,237,448 - 2,322,672 1,607,334 - 2,207,334 1,641,967 - 2,241,967	Ethyl lactate Trimethyl-oxazole Diethyl butanedioate Over fermented Rotten fruit Heptan-2-ol
9	3,371,584 - 3,971,584 3,652,480 - 4,252,480 3,995,514 - 4,595,514 4,034,196 - 4,634,196 4,254,626 - 4,854,626 4,520,440 - 5,120,440 4,646,739 - 5,246,739 4,690,282 - 5,290,282 5,184,281 - 6,071,002 5,298,651 - 6,284,926 5,316,557 - 5,916,557 5,985,636 - 6,585,644 6,094,380 - 6,694,380 6,414,969 - 7,014,974 6,840,479 - 7,440,479	Heptan-2-ol Diethyl butanedioate Pentyl acetate Trimethyl-oxazole 2,3,5-trimethylpyrazine Diethyl butanedioate Nutty Nut flesh Nonan-2-one Diethyl butanedioate Trimethyl-oxazole Over fermented Rotten fruit Ethyl 2-hydroxy-4-methylpentanoate Diethyl butanedioate Nutty Nut flesh Diethyl butanedioate
9	8,038,805 - 8,638,805 8,603,894 - 9,961,696 9,660,571 - 10,260,571	Nutty Nut flesh Trimethyl-oxazole 1,2-propanediol diacetate
9	29,047,112 - 29,647,112	2-pentylfuran and nutty Nut flesh
9	33,599,131 - 34,199,131 34,157,432 - 34,757,432	3-methylbutyl 3-methylbutanoate Trimethyl-oxazole
9	37,706,972 - 38,306,972 37,734,657 - 38,334,657 37,850,408 - 38,890,668 38,213,689 - 38,813,689	Pentyl acetate 1,2-propanediol diacetate Pentyl acetate Nutty Nut flesh
10	0 - 426,218 0 - 426,235	Nutty Nut flesh Total Nutty
10	769,535 - 1,417,085 769,535 - 1,369,535	2-pentylfuran and Ethyl 2-hydroxy-4-methylpentanoate Over fermented Rotten fruit
10	17,784,779 - 18,384,779 18,251,843 - 18,851,843	Trimethyl-oxazole Ethyl 2-hydroxy-4-methylpentanoate
10	18,945,761 - 19,545,761 19,225,961 - 19,825,961 19,405,234 - 20,005,234 19,652,912 - 20,252,912	Trimethyl-oxazole Butyl butanoate Diethyl butanedioate Fruit Berry

Diethyl butanedioate, nonan-2-one, 2-pentylfuran, ethyl lactate, trimethyl-oxazole, pentyl acetate, heptan-2-ol, 2,3,5-trimethylpyrazine and ethyl 2-hydroxy-4-methylpentanoate seems to be more related to over fermented/rotten fruit notes.

Nonan-2-one, 2,3,5-trimethylpyrazine, trimethyl-oxazole, pentyl acetate, 2,3,5,6-tetramethylpyrazine, hexyl acetate, 2-pentylfuran, butyl butanoate, diethyl butanedioate ,ethyl 2-hydroxy-4-methylpentanoate (known to have a fresh blackberry note) and 1,2-propanediol diacetate (known to have a fruity acetic note) seems to be more related to fruit berry notes.

Diethyl butanedioate, hexyl acetate, limonen, 1,2-propanediol diacetate and nonan-2-one seems to be more related to fruit dark notes.

The note nutty nut flesh seems to be perceived, thanks to the mixture of three types of compounds: compounds having fresh fruit notes (1,2-propanediol diacetate, ethyl 2-hydroxy-4-methylpentanoate, heptan-2-ol), compounds having a fruity and green note (nonan-2-one, 2-pentylfuran, pentyl acetate, diethyl butanedioate) and compounds having a nutty note (2,3,5-trimethylpyrazine, 3-methyl-2-cyclohexen-1-one, trimethyl oxazole).

4.3.3-Areas of association detected in relation to vegetal notes

A total of 30 association areas were detected for the volatile compounds involved in vegetal notes (appendix 32-34). Areas of association were detected on all chromosomes (appendix 29). Two areas of co-localization were detected between different volatile compounds (Table 23).

Table 23: Co-location between traits related to green notes

Chromosome	Position	Traits
2	40,286,532 - 0,886,532 40,871,868 - 1,471,868	2-methyl-but-3-en-2-ol Trans-beta-ocimen
3	28,198,827 - 29,615,250 28,876,174 - 9,476,174	2-methyl-but-3-en-2-ol (Z)-2-hexen-1-ol acetate

4.3.4-Areas of association detected in relation to woody and spicy notes

A total of 76 association areas were detected in relation to volatile compounds known to have a woody and/or spicy note and 10 linked to the woody and/or spicy notes detected by

sensory analysis (appendix 31-34). Areas of association were detected on all chromosomes (appendix 29). Eleven areas of co-location were detected between different volatile compounds and between volatile compounds and sensorial notes (table 24). Camphene and alpha-terpinene seem to be more related to the presence of spice tobacco notes.

Table 24: Co-location between traits related to spicy and woody notes

Chromosome	Position	Traits
1	10,403,809 - 11,003,809 10,635,523 - 11,235,523	Beta-Myrcen Camphene
2	34,930,849 - 35,530,849 34,930,857 - 35,530,857	Spice Tobacco Camphene
2	37,724,506 - 38,324,511 38,143,664 - 39,990,777	Alpha-Terpinen Camphene
4	0 - 588,739 22,858 - 654,737	Alpha-Terpinen Camphene
5	29,291,725 - 29,891,725 29,626,890 - 30,226,890	Spice Tobacco Camphene
7	2,857,538 - 4,851,587 3,178,085 - 3,778,085	Camphene Spice Tobacco
8	5,024,136 - 6,200,257 5,600,244 - 6,857,883	Camphene Spice Tobacco
9	4,254,626 - 5,120,440 4,498,201 - 5,098,201	Camphene Spice Tobacco
9	26,640,576 - 27,240,576 27,120,722 - 27,810,582 27,120,724 - 27,720,724	Camphene Spice Tobacco Alpha-Terpinen
10	754,940 - 1,354,940 842,194 - 1,442,194	Spice Tobacco Camphene
10	3,348,574 - 4,171,361 3,679,144 - 4,438,496	Camphene Total Spice

4.3.5-Areas of association detected in relation to empyreumatic notes

We could identify 66 association zones related to volatile compounds known to have an empyreumatic note (caramel, brown sugar, roasted...) and 43 linked to empyreumatic notes detected by sensory analysis (appendix 31-34). Areas of association were detected on all chromosomes (appendix 29). Twenty areas of co-locations were observed (Table 25). The 2,3-dimethyl-5-ethylpyrazine, 2,5-dimethyl-pyrazine and 1-methylhexyl acetate seem to be more related to the presence of sweet, caramel notes. The 3-methyl-2-(2-methyl-2-butenyl)-furan and 1-methylhexyl acetate seems to be more related to the roast degree.

Table 25: Co-location between traits related to empyreumatic notes

Chromosome	Position	Traits
1	5,251,864 - 5,851,864 5,398,258 - 6,019,937	Roast Degree 3-methyl-2-(2-methyl-2-butenyl)-furan
1	7,817,873 - 8,477,387 7,817,873 - 7,748,797 8,205,326 - 8,805,326	Cocoa Roast Degree Sweet, caramel
1	14,322,116 - 14,922,116 14,325,744 - 14,925,744	Sweet, caramel 2,3-dimethyl-5-ethylpyrazine
1	26,056,034 - 26,656,034 26,551,903 - 27,152,249	Sweet, caramel 2,3-dimethyl-5-ethylpyrazine
1	27,313,384 - 27,913,384 27,622,563 - 28,222,563	2,3-dimethyl-5-ethylpyrazine 2,5-dimethyl-pyrazine
1	28,594,210 - 29,194,210 28,865,580 - 29,465,580	1-methylhexyl acetate 2,5-dimethyl-pyrazine
3	2,502,549 - 3,145,200 2,551,310 - 3,151,310	1-methylhexyl acetate 2,3-dimethyl-5-ethylpyrazine
3	12,353,462 - 12,953,564 12,353,462 - 12,953,462	1-methylhexyl acetate Roast Degree
3	26,244,291 - 26,844,291 26,406,759 - 27,006,759 26,656,014 - 27,256,014 26,748,027 - 27,348,027 26,883,974 - 27,483,974	2,3-dimethyl-5-ethylpyrazine Sweet, caramel 1-methylhexyl acetate 1-hydroxypropan-2-one 2,3-dimethyl-5-ethylpyrazine
3	28,145,339 - 28,745,339 28,145,339 - 28,745,339	Cocoa Roast Degree
4	18,763,963 - 19,471,045 18,871,045 - 19,471,045	Roast Degree Cocoa
5	515,972 815,972 1,115,972 1,281,629 1,581,629 1,881,629	1-methylhexyl acetate Sweet, caramel
5	2,569,575 2,869,575 3,296,902 2,569,575 2,869,575 3,296,902	Cocoa Roast Degree
5	34,094,673 - 34,694,673 34,094,673 - 34,694,673	Cocoa Roast Degree
6	19,179,147 - 19,926,195 19,581,377 - 20,181,377	Roast Degree 3-methyl-2-(2-methyl-2-butenyl)-furan
6	23,199,939 - 23,799,955 23,454,660 - 24,406,262 23,928,498 - 24,659,115	Roast Degree Sweet, caramel 1-methylhexyl acetate
8	3,449,438 - 4,049,438 4,010,117 - 4,610,117 4,040,269 - 4,640,329	Sweet, caramel 2,5-dimethyl-pyrazine 1-methylhexyl acetate
8	4,671,607 - 5,638,343 5,523,513 - 6,123,513 5,963,281 - 6,563,290	1-methylhexyl acetate 2,5-dimethyl-pyrazine 1-methylhexyl acetate
10	4,311,469 - 5,528,191 4,928,081 - 5,528,191	Roast Degree Cocoa
10	13,049,399 - 13,649,399 13,049,399 - 13,649,399	Roast Degree Cocoa

4.4-Identification of candidate genes involved in the different biosynthetic pathways of aromatic compounds detected in the Amazonian population

In each association zone, a search was undertaken for candidate genes involved in the synthesis of the compounds in the association. Candidate genes involved in general plant defences were also observed. The biosynthetic pathways previously described in other plants appear to be valid in cocoa. Schemes summarising these biosynthetic pathways that may be present in cocoa and the involvement of candidate genes in these pathways have been produced. The 1824 candidate genes identified are listed in appendix 35.

Of the candidate genes detected, 53 are in common with previous GWAS studies on the aroma traits of the Nacional modern population (appendix 35) (Colonges et al., 2021b; Colonges et al., 2021c). The repetition of detection of these candidate genes between two studies on two different populations brings additional weight to their role.

4.4.1-Candidate genes involved in the monoterpene biosynthetic pathway

In the areas of the genome associated with monoterpenes, candidate genes encoding enzymes involved in their biosynthesis have been identified. The role of each candidate gene identified as being involved in terpene biosynthesis is illustrated in Figure 39A. Co-locations between the association zones of different terpenes can be due to two possibilities either the compounds have the same precursor or they follow each other in the biosynthetic pathway (one is necessary to produce the other). Thanks to these possibilities, the hypothesis of a common precursor between camphene, o-xylene, m-xylene, α-terpinene, γ-terpinene, p-menthene, d-limonene, β-myrcene, trans-β-ocimene, allo-ocimene, 1,2-dihydrolinalol and α-terpineol were made (Figure 39B). The other co-locations made it possible to confirm certain transformations of terpenes into another (such as the transformation of α-pinene into camphene or of o-xylene into m-xylene) which was already known in the literature (Figure 39B).

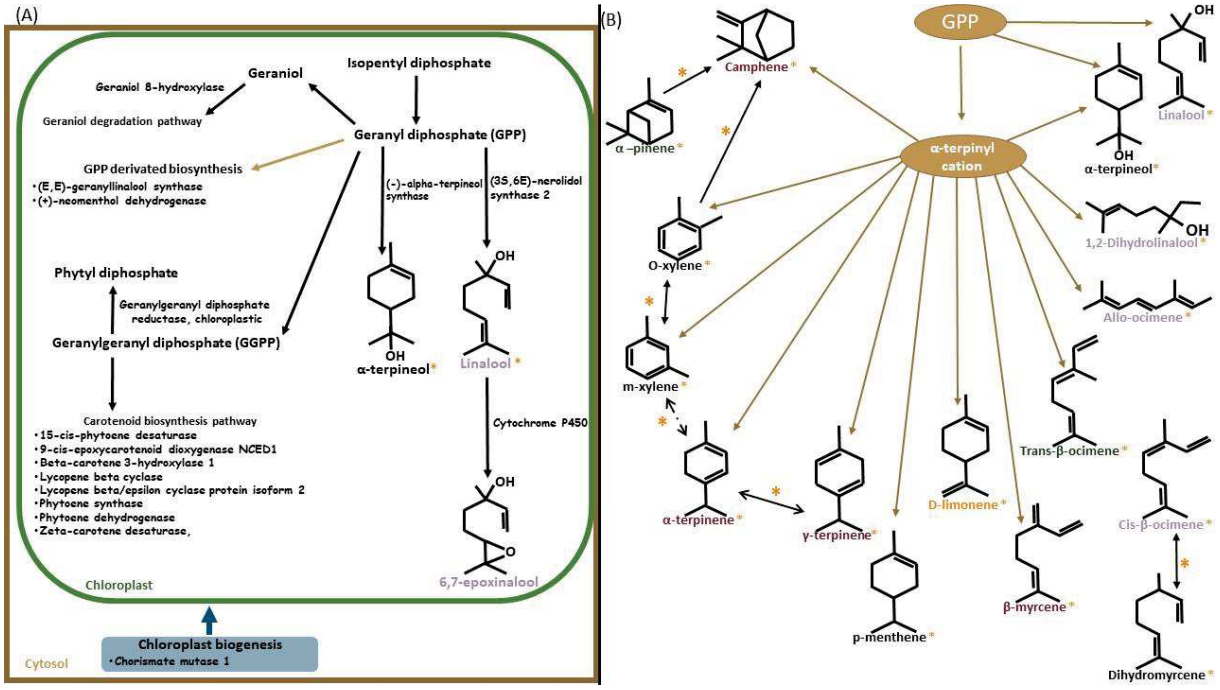


Figure 39: Schematic representation of the monoterpene biosynthetic pathway adapted to Bohlmann et al., (1998)

All compounds marked with an orange star were identified in this study.The volatile compounds are indicated in colour according to their taste : in purple for floral note, in orange for fruity note, in red for spicy or woody note, in green for vegetal and/or herbaceous note, in black are not known to be aromatic. The black arrows represent chemical transformations identified in other organisms. (A) Candidate genes, identified in areas of association with monoterpenes, are shown in black according to their known level of involvement in the monoterpene biosynthetic pathway in other crops plants. (B) Hypothetical pathway in the cocoa tree according the co-location of association zones. The brown arrows show the hypothetical part of the biosynthetic pathway according to the identified co-locations taking into account that the co-locations are due to a common precursor. The black arrows with an orange star show the hypothetical part of the biosynthetic pathway according to the identified co-locations taking into account that the co-locations between the two compounds are because one is a precursor of the other.

4.4.2-Candidate genes involved in the L-phenylalanine degradation pathway

In the areas of the genome associated with the compounds of the L-phenylalanine degradation pathway, candidate genes encoding enzymes involved in their biosynthesis have been identified. The role of each candidate gene identified as being involved in the L-phenylalanine degradation pathway is illustrated in Figure 40.

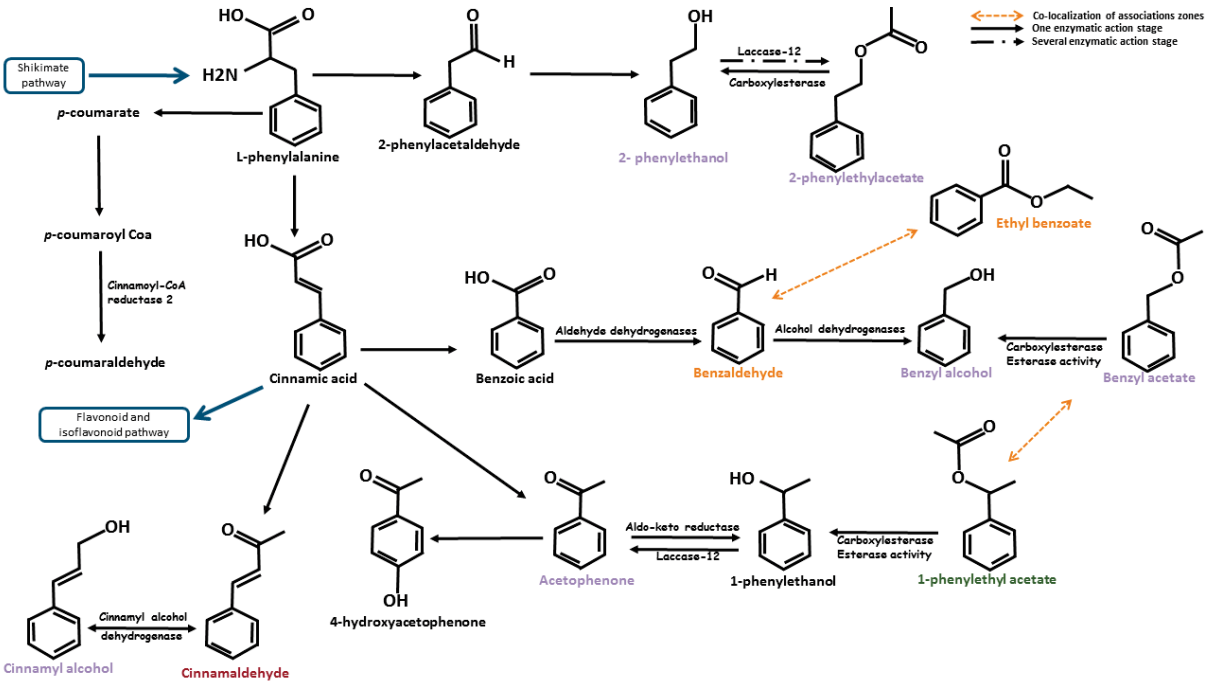


Figure 40: Schematic representing the L-phenylalanine degradation pathway according to Lapadatescu et al., (2000) and the hypothetical L-phenylalanine degradation pathway in cocoa.
The volatile compounds are indicated in colour according to their already known taste : in purple for floral note, in orange for fruity note, in red for spicy or woody note, in green for vegetal and/or herbaceous note, in black are not known to be aromatic. The black arrows represent chemical transformations identified in other organisms. The orange dotted arrows represent co-locations between the compounds and thus a probable link in the biosynthetic pathway. Candidate genes, identified in the association zones related to compounds included in the L-phenylalanine degradation pathway, are shown in black at their known level of involvement in the monoterpene biosynthetic pathway according to Lapadatescu et al., (2000). In blue, the nearby biosynthetic pathways are shown.

4.4.3-Candidate genes involved in the biosynthesis of Maillard reaction precursors

In the areas of the genome associated with pyrazines and furans, which are compounds primarily derived from the Maillard reaction, candidate genes encoding enzymes involved in the biosynthesis of their precursors as well as their biosynthesis have been identified. The role of each candidate gene identified is illustrated in Figure 41.

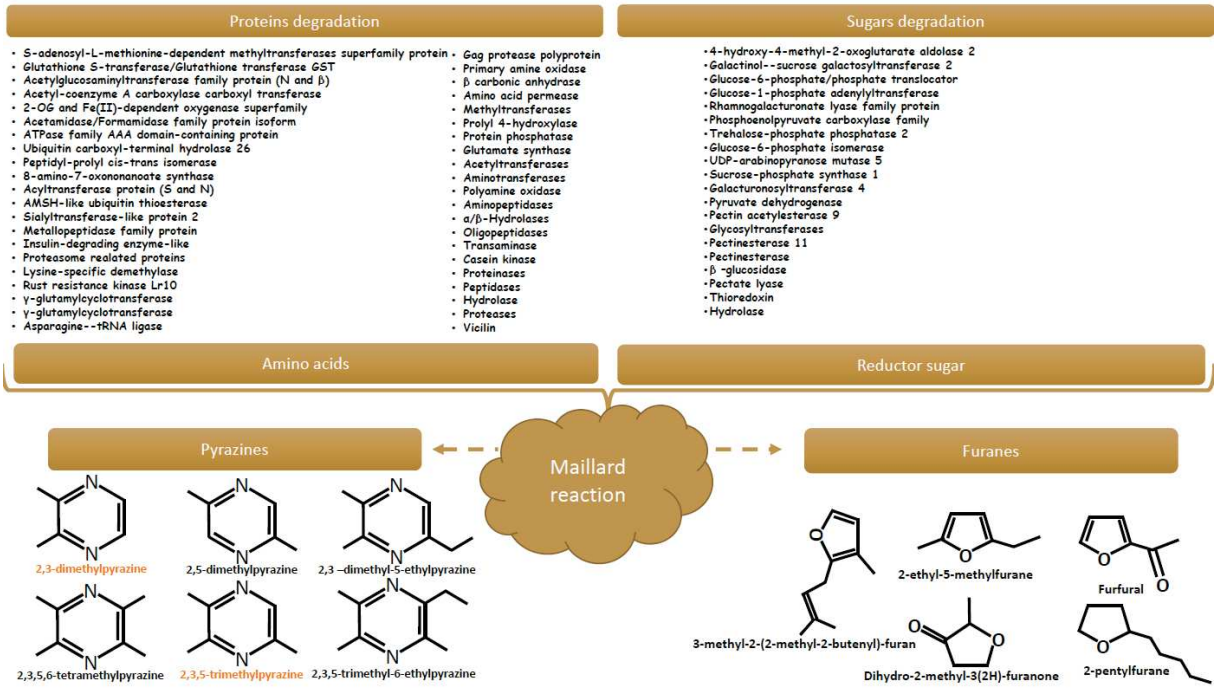


Figure 41: Diagram representing the hypothetical biosynthetic pathway of pyrazines and furans in the cocoa tree.

The volatile compounds are indicated in colour according to their taste: in orange for fruity note and in black are not known to be aromatic. The orange dotted arrows represent co-locations between the compounds and thus a probable link in the biosynthetic pathway. Candidate genes, identified in the association areas with pyrazines and furans, are shown in black in the left of the figure. 2-OG: 2-oxoglutarate.

4.4.4-Candidate genes involved in the degradation pathways of fatty acids and sugars

In the areas of the genome associated with compounds involved in the fatty acid and/or sugar degradation pathway, candidate genes encoding enzymes involved in their biosynthesis were identified. The role of each identified candidate gene is illustrated in Figure 42.

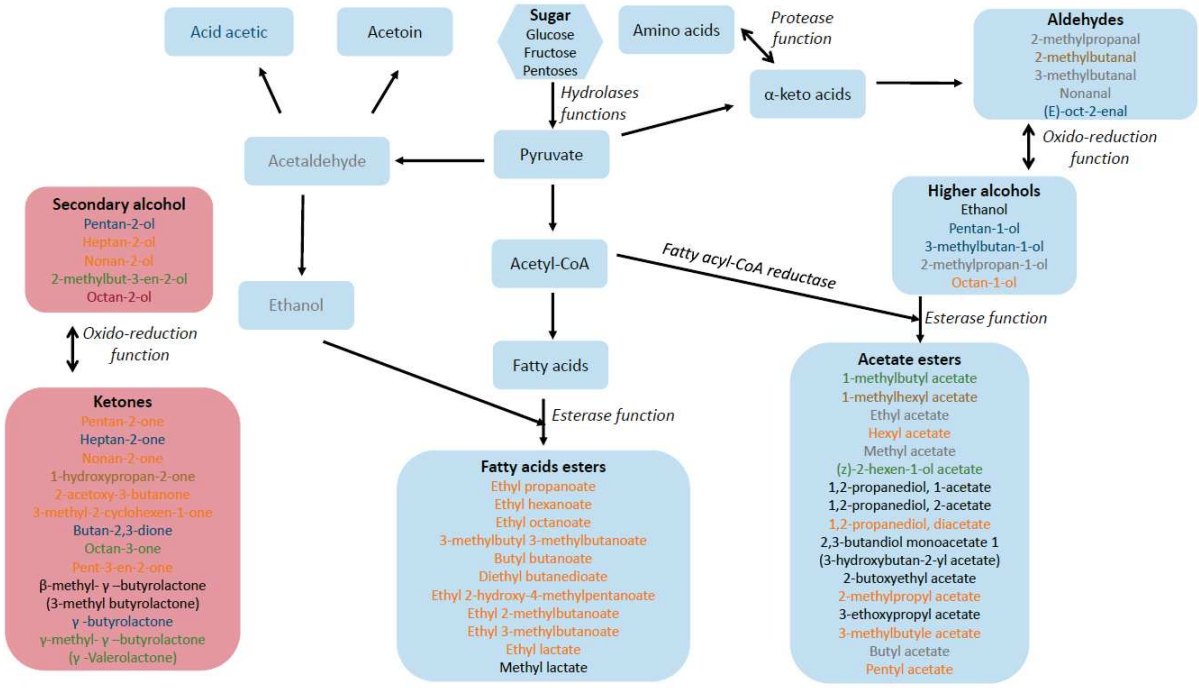


Figure 42: Schematic representation of the fatty acid and sugar degradation pathway according to Swiegers et al., (2005) and Dzialo et al., (2017) and hypothetical pathway in the cocoa tree.

The volatile compounds are indicated in colour according to their taste: in orange for fruity note, in red for spicy or woody note, in green for vegetal and/or herbaceous note, in blue for cheesy notes, in grey for chemical notes, in brown for chocolate and brown notes, in black are not known to be aromatic. The black arrows represent chemical transformations identified in other organisms. The function of candidate genes, identified in areas of association with compounds involved in the fatty acid and sugar degradation are shown in black italics.

4.4.5-Candidate genes involved in general plant defences

Plants synthesise volatile compounds for various purposes including defence (Ponzio et al., 2013). During fermentation, cocoa beans may react to biotic stress, as previously suggested (Sabau et al., 2006; Colonges et al., 2021b), and synthesise volatile compounds for defence against micro-organisms involved in fermentation. During the in-silico study of the genes present in the association zones, a large number of genes involved in the synthesis and/or signalling pathways of hormones involved in plant defence pathways were found in the areas of associations linked to volatile compounds. These genes could be at the origin of the synthesis of volatile aromatic compounds (Figure 43). Some candidate genes present in the various association zones are related to phytohormones important for plant growth such as auxins, abscisic acid (ABA) or gibberellins. Plant hormones interact in complex networks to balance the response to environmental and developmental signals and thus limit the adaptation costs associated with defence. The molecular mechanisms governing these hormonal networks are largely unknown but recent studies have shown their implications in defence responses (Denancé et al., 2013), which is why we have selected, as candidate genes, the genes listed in Table 26.

During fermentation, cocoa beans can also be subjected to abiotic stresses such as increased temperature or increased acidity and thus a decrease in pH (Afoakwa et al., 2008). It has also been reported in the literature that abiotic stresses can cause plants to produce volatile compounds (Baldwin, 2010). It is during the increase in heat that genes coding for heat shock proteins can intervene as well as heat stress transcription factors which can be at the origin of the activation of biosynthetic pathways of aromatic volatile compounds.

Table 26: Candidate genes related to growth phytohormones

Auxine related	ABA related	Gibberellin related
Protein AUXIN SIGNALING F-BOX 2		
Auxin-induced protein	Abscisic acid receptor PYL6	Gibberellin 2-beta-dioxygenase 8
Auxin transport protein BIG	Abscisic acid receptor PYL2	Gibberellin 3-beta-dioxygenase 4
Auxin response factor	Major allergen Pru ar 1	Gibberellin 2-beta-dioxygenase 2
Auxin-responsive protein	Abscisic acid 8'-hydroxylase 4	Dihydroflavonol-4-reductase
Indole-3-acetic acid-amido synthetase GH3.1	Protein phosphatase 2C 16	DELLA protein GAI
Indole-3-acetic acid-induced protein ARG7		
Protein-tyrosine-phosphatase IBR5		

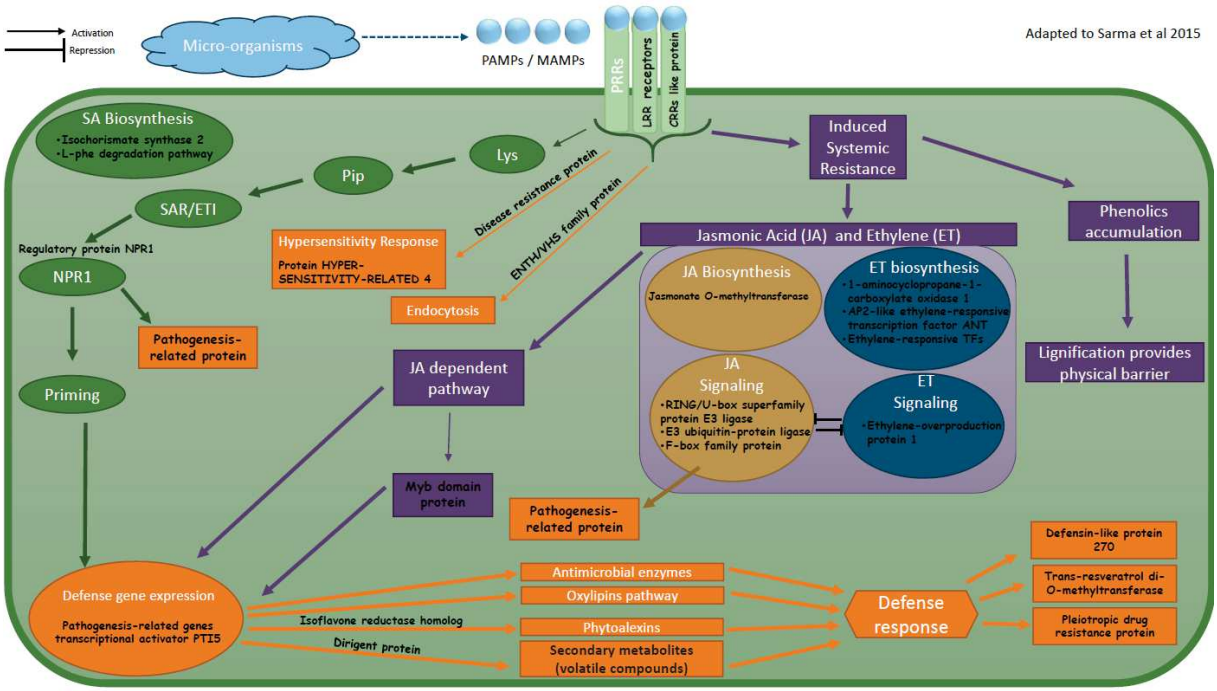


Figure 43: Diagram of general plant defences adapted from Sarma et al., (2015).

In green are represented the mechanisms associated with the systemic acquired resistance response. In purple are the mechanisms associated with the induced systemic resistance response. In blue are represented the mechanisms associated with ethylene. In brown are represented mechanisms related to jasmonic acid. In orange are represented the general mechanisms. The candidate genes, identified in the association areas linked to volatile compounds, are shown in black at the location of their action. PAMPs: Pathogens-Associated Molecular Patterns, MAMPs: Microbe-Associated Molecular Patterns, PRR: Pattern Recognition Receptors, LRR: Leucine Rich Repeat, CRR: Cysteine-Rich Receptors, GA: Giberellic acid (gibberellin), TFs: Transcription Factors, SA: Salicylic Acid, SAR: Systemic Acquired Resistance.

5-Discussion

The studied population was collected in the area of origin of the Nacional variety. The collection of these new genotypes was done to safeguard the genetic diversity of native cocoa trees from this Amazonian region and to enlarge the genetic resources available to improve or create new aromatic varieties. Indeed, until now, Ecuadorian aromatic cocoa trees are mainly represented by the modern Nacional variety. This population is a hybrid population with a narrow genetic base involving a limited number of ancestral Nacional genotypes hybridized with introduced and closely related Trinitario genotypes.

Our results showed that, as we could expect, a larger variability can be observed at the genetic level and at aromatic level, with common as well as new aromatic volatile compounds identified and segregating in this population.

The cocoa trees studied in this study, widespread in the South of Ecuadorian Amazonia, have a different genetic background than the Nacional variety. They are the result of natural evolution from thousands of years giving to this population a low LD favourable to a better

precision of localization of associations than the study carried out on the modern Nacional population (for which the hybridizations between the three contrasting ancestors are recent) (Bartley, 2005; Loor S., 2007; Colonges et al., 2021b). The associations detected are therefore more precise and reliable. Moreover, this population has a higher genetic diversity, increasing the number of markers or genes segregations.

Biosynthetic pathways involved in the synthesis of floral notes are the same as those previously described in the modern Nacional population: the monoterpene biosynthetic pathway and the L-phenylalanine degradation pathway (Colonges et al., 2021b). Common and new compounds have been identified as being involved in the variety of floral notes. Associations with five new monoterpenes known to have floral note were identified in this study ((-)-dihydromyrcen, 1,2-dihydrolinalool, allo-ocimene, cis-β-ocimene, dihydromyrcene). A new association with ethylbenzene acetate (compound involved in L-phenylalanine degradation) was identified. These new results complement the likely biosynthetic pathways used in cocoa for the synthesis of floral notes.

Associations with new compounds known to have a fruity note were also detected. Two compounds seem to be involved in fresh fruit notes: D-limonene (a monoterpene) and 2-acetoxy-3-butanone (a ketone probably synthetized during fatty acid or sugar degradation). Two other compounds seem to be involved in dried fruit notes: 2-pentylfuran and trimethyl-oxazole (probably synthetized during the Maillard reaction).

Areas of association were identified with five new compounds known to have a vegetal note. There are two monoterpenes: α-pinene and trans-β-ocimene, three compounds probably synthetized during fatty acid or sugar degradation: (Z)-2-hexen-1-ol acetate, 2-methyl-but-3-en-2-ol and octan-3-one. These compounds belong to the same biosynthetic pathways as the key compounds previously identified in the synthesis of fruity and floral aromas in the Nacional variety (Colonges et al., 2021b; Colonges et al., 2021d).

The biosynthesis of all these compounds could therefore be at the origin of new aroma notes in this population. Confirmation of the functional expression of candidate genes involved in the biosynthesis of these compounds would make it possible to identify diagnostic markers in these genes. These markers could then be used by breeders according to the expectations of chocolate makers or consumers.

The results concerning woody and spicy notes are very different from those observed in the study of the modern Nacional population. This population possesses biochemical compounds that give woody and spicy notes that the modern Nacional does not have, such as the presence of β-myrcene, α-terpinene, or camphene. A combination of the favourable alleles identified in the modern Nacional and those of this wild population for the synthesis of compounds involved in woody and spicy notes could give a unique aroma.

Associations with new compounds known to have an empyreumatic note were also detected. This is the case for the compounds: 1-hydroxypropan-2-one, 2,3-dimethyl-5-ethylpyrazine (not detected in fermented and dried beans), 2,5-dimethyl-pyrazine, 3-methyl-2-(2-methyl-2-butenyl)-furan.

Fifty-three candidate genes detected are common between both modern Nacional and Amazonian populations, which strongly increases the probability that these genes are involved in cocoa flavour synthesis. Furthermore, as the population in this study is a native cocoa population from Amazonia, resulting in low linkage disequilibrium, the areas of association detected in this study are smaller and therefore more accurate.

All these new results allow us to broaden our knowledge on the palette of aromatic notes synthesised by the cocoa tree itself and not by the micro-organisms present during fermentation. Even if the aromatic profile of the final chocolate is influenced by the fermentation environment (methods, micro-organisms present...) as well as by the processing protocols such as roasting (Owusu et al., 2012; Kongor et al., 2016; Assi-Clair et al., 2019), we were able to highlight the part of aroma variations depending on the genotype of the cocoa tree. The volatile compounds as well as the sensory notes identified, for which associations could not be detected, certainly constitute the part brought by the “global” terroir of the cocoa tree (place of cultivation, post-harvest protocols, microflora present throughout the processing ...).

This population shows a large number of aroma traits favourable to the search of flavours and economic niches sought by chocolate makers and potentially appreciated by consumers. It, therefore, constitutes germplasm of greatest interest for plant breeders.

However, this population also shows several unfavourable flavours, and as with all wild plants, domestication of the favourable traits will be necessary to select the desired parts of the genome. The compounds involved in the desired aromas do not appear to be related to the compounds responsible for unpleasant aromas. Indeed, no or very low correlations have been

observed between these compounds and few areas of association co-localise. The selection of the compounds responsible for the desired aromas combined with a counter-selection of the compounds responsible for the unpleasant aromas seems therefore possible.

This population offers to breeders a wider range of aromas to select for new aromatic niches. The combination of these genotypes between them or with other aromatic varieties could be also promising for the production of new original aromatic varieties adapted to the different Ecuadorian climates.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

CL, RGLS, conceived the experiment; KC, MCL, ML conducted biochemical analyses; KC, JCJ, DC, CS, IS, FF conducted microfermentation, ES carried out sensorial analyses; KC, OF carried out DNA experiments; KC, BR, RB, CL, analysed data; KC, RB, CL wrote the manuscript.

Funding

The study was funded by the United States Department of State (U.S. Foreign Ministry); the U.S. Embassy, Quito; the U.S. Department of Agriculture (USDA-ARS); the MUSE Amazcacao project with the reference ANR-16-IDEX-0006.

Acknowledgement

We thank the I-Site MUSE, Valrhona and the USDA, for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" programme with the reference ANR-16-IDEX-0006.

Après avoir contribué à l'étude des déterminants génétiques et biochimiques des arômes fins du cacao (notes florales et fruitées), les travaux de cette thèse ont également porté sur les déterminismes génétiques et biochimiques des sensations d'amertume et d'astringence pouvant être plus ou moins appréciées par les consommateurs mais pouvant également être à l'origine du déclassement de certains cacaos.

Les polyphénols ainsi que la théobromine et la caféine sont connus pour être impliqués dans les différentes sensations d'amertume et d'astringence (Serra Bonvehi and Ventura Coll, 1997). Grâce à des analyses NIRS sur les fèves de cacaos appartenant à la population de Nacional moderne et sur des fèves de cacaos appartenant à la population de cacaoyers issue des prospections en Amazonie, ainsi que des analyses sensorielles sur les liqueurs de cacaos fabriquées à partir de ces mêmes échantillons de fèves, des analyses GWAS ont pu être réalisées sur les deux populations de cacaoyers présentées dans les chapitres précédents.

Soixante-treize zones d'association en lien avec ces composés non volatiles et avec les données d'analyses sensorielles ont pu être mises en évidence. Quatre-vingt-un gènes candidats ont également pu être identifiés dans certaines des zones d'association détectées par GWAS. L'ensemble de ces résultats vous sont présentés dans le chapitre suivant.

Chapitre 6 : Diversité et déterminants de l'amertume, l'astringence et la teneur en acides gras des cacaoyers équatoriens cultivés ou natifs d'Amazonie

Chapitre 6 : Diversité et déterminants de l'amertume, l'astringence et la teneur en acides gras des cacaoyers équatoriens cultivés ou natifs d'Amazonie

Diversity and determinants of bitterness, astringency and fat content in cultivated Nacional and native Amazonian T. cacao accessions from Ecuador

Kelly Colonges[1,2,3,4], Edward Seguine[5], Alejandra Saltos[6], Fabrice Davrieux[4,7], Jérôme Minier[4,7], Juan-Carlos Jimenez[6], Marie-Christine Lahon[3,4], Darío Calderon[6], Cristian Subia[6], Ignacio Sotomayor[6], Fabián Fernández[6], Marc Lebrun[3,4], Olivier Fouet[1,2], Bénédicte Rhoné[1,2], Xavier Argout[1,2], Pierre Costet[8], Claire Lanaud[1,2*], Renaud Boulanger[3,4*], Rey Gastón Loor Solorzano[6]

1 Cirad, UMR AGAP, F-34398 Montpellier, France.

2 AGAP Institut, Univ Montpellier, Cirad, INRAE, Institut Agro, Montpellier, France.

3 Cirad, UMR Qualisud, F-34398 Montpellier, France.

4 Qualisud, Univ Montpellier, Avignon Université, Cirad, Institut Agro, IRD, Université de La Réunion, Montpellier, France.

5 Seguine Cacao/Guittard Chocolate Co, Arroyo Grande, CA, United States

6 Instituto Nacional de Investigacion Agropecurias, INIAP, Ecuador.

7 Cirad, UMR Qualisud, F-97400 Réunion, France.

8 Valrhona, France

Keywords: *T. cacao*, GWAS, bitterness, astringency, fat content, protein content

1-Abstract

Theobroma cacao is the only tree that can produce cocoa. Cocoa beans are highly sought after by chocolate makers to produce chocolate. Cocoa can be fine aromatic, characterised by floral and/or fruity notes, or it can be described as standard cocoa with a more pronounced cocoa aroma and bitterness. In this study, the genetic and biochemical determinants of non volatile compounds related to bitterness, astringency, fat content and protein content will be investigated in two populations: a cultivated modern Nacional population and a population of *T. cacao* accessions collected recently in the Ecuadorian South Amazonia. For this purpose, a GWAS study was carried out on two Ecuadorian cocoa populations, with results of biochemical compounds evaluated by NIRS assays and with sensory evaluations. Areas of association were detected with both types of data. Candidate genes could be identified in the areas of association.

2-Introduction

Theobroma cacao (cocoa tree) belongs to the family Malvaceae (Bayer and Kubitzki, 2003). *T. cacao* is a tree of great agronomic and economic interest. Indeed, it is the only source worldwide that allows the manufacture of chocolate. Worldwide consumption of chocolate is constantly increasing and is expected to rise by a further 20% by 2025 (Cilas, 2020).

T. cacao is a diploid plant (2n=2x=20). The cocoa tree has a relatively small genome equivalent in size to that of rice (Lanaud et al., 1992; IRGSP and Sasaki, 2005; Argout et al., 2011). The cocoa genome of two varieties has been fully sequenced and are available: Criollo (Argout et al., 2011; Argout et al., 2017) and Amelonado (Motamayor et al., 2013). *T. cacao* shows a great genetic diversity. Currently, ten genetic groups have been identified in *T. cacao* species through genetic analysis. These ten genetic groups are: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañón, Nanay, Nacional and Purus (Motamayor et al., 2008). Cacao can be classified into two types of products: bulk cocoa, which has a strong cocoa taste, and aromatic fine cocoa, which is characterised by floral and fruity notes (Sukha et al., 2008). The most widely grown varieties of fine aromatic cocoa are Nacional, Criollo and Trinitario trees. Trinitarios are hybrids between Criollo and Amelonado. The Amelonado is a variety that produces mostly bulk cocoa. The Criollo variety, on the other hand, produces cocoa beans with a predominantly fruity aroma (Lachenaud and Motamayor, 2017). The Criollo variety is not widely cultivated because of its low vigour and increased susceptibility to disease (Cheesman, 1944).

The Nacional variety originated in Ecuador. The trees of the Nacional variety currently cultivated belong to the modern Nacional variety. They are the result of several generations of crosses between the ancestral Nacional and Trinitario introduced in Ecuador in the last century (Bartley, 2005; Loor S. et al., 2009). The resulting hybrids constitute what is currently known as the modern Nacional variety. The latter is only grown in Ecuador. Surveys were undertaken in the presumed domestication centre of Nacional to search for native cocoa trees related to the ancestral Nacional variety (Loor S. et al., 2012; Loor S. et al., 2015) to enlarge the genetic resources for fine cocoa breeding.

The fine (floral and fruity) flavours of modern Nacional have started to be studied for their volatile compound composition (Ziegleder, 1990; Luna et al., 2002; Cevallos-Cevallos et al., 2018; Rottiers et al., 2019; Colonges et al., 2021b; Colonges et al., 2021d). Nacional cocoa, like all cocoa, also contains non-volatile compounds, such as polyphenols, caffeine or theobromine (Wollgast and Anklam, 2000; Zheng et al., 2004), which are known to provide bitterness and astringency to cocoa products (Lesschaeve and Noble, 2005). A high concentration of these compounds can therefore mask the fine flavours of cocoa. However, these compounds do not only bring defects. Thanks to its richness in polyphenols, cocoa contributes to good mental health and cardiovascular protect (Andújar et al., 2012; Tuenter et al., 2018).

In order to study the genetic and biochemical determinants of bitterness and astringency of the Ecuadorian cocoa trees, as well as their fat and protein contents, two important factors interacting with flavours, the non-volatile compounds contained in fermented roasted and non roasted beans were characterised by NIRS. Sensorial analyses on liquors were also carried out. All these data were used to conduct a GWAS (Genome Wide Association Study) on all these traits using molecular genotyping data obtained by genotyping by sequencing (GBS).

3-Material and methods

3.1-Vegetal material

Two populations of cocoa trees were used for this study.

The first population is a population of one hundred and fifty-two cocoa trees belonging to the modern Nacional variety as previously described (Colonges et al., 2021b; Colonges et al., 2021d).

The second population used is composed of two hundred and two cocoa trees. They belong to surveys carried out in the domestication centre of the ancestral Nacional variety in Ecuador previously identified (Loor S. et al., 2012; Loor S. et al., 2015). The collected trees were put into a germplasm collection located at an agricultural college in Pangui and in two INIAP (Instituto Nacional de Investigacion Agropecurias) experimental centres: in Pichilingue (EET-P) and in Domono (appendix 36).

3.2-Micro-fermentation

In both cases, the pods were harvested at maturity in the different growing locations. The micro-fermentations took place at Pichilingue for the Nacional population and at Domono within 24 hours of harvest. In both cases, the micro-fermentations were carried out under the most homogeneous conditions possible. The cocoa beans of each genotype were placed in delicate linen bag nets. They were then distributed over four floors in the middle of the mass of Nacional modern cocoa beans. At 24 and 72 hours of fermentation, stirring was performed. At each stirring, the bags of beans at the bottom were placed at the top and those in the middle-low position were placed in the middle-high and vice versa. After 4.5 days, the beans were taken out of the net and dried separately in a greenhouse. When the moisture content was less than or equal to 8% the beans were considered dry and were placed under vacuum.

3.3-Sensorial analysis

For the modern Nacional population, one hundred and forty-four individuals were characterised by sensory analysis based on blind tastings carried out on three replicates per sample. For the Amazonian native cocoa population, one hundred and fifty-nine genotypes were characterised. The tastings were conducted on cocoa liquor. The cocoa liquor corresponds to merchantable cocoa (dried fermented beans) that has been roasted and ground. The sensory notes (bitterness and astringency) were judged with a score ranging from zero (no note detected) to ten according to the ISCQF, protocol (2020). We used the average of the three replicates for the GWAS phenotype.

3.4-Non-volatile compounds analysis

NIRS acquisitions and processing were carried out according to the protocol of Álvarez et al., (2012). For the modern Nacional population these acquisitions were done on fermented, dried and roasted beans while for the native Amazon population these acquisitions were done on fermented and dried beans. These acquisitions made it possible to calculate the concentrations of fat content, caffeine, theobromine, procyanidins B2, procyanidins B5,

procyanidins C1, epicatechin and NH3, and proteins, which made it possible to deduce the ratio of theobromine to caffeine and the total procyanidin concentration.

3.5-DNA extraction and genotyping

DNA extraction was performed according to the protocol of (Risterucci et al., 2000). DNA samples were genotyped by sequencing (GBS) using Diversity Arrays Technology Sequencing (DArTseq) made by the Diversity Array Technology (DArT) company (Kilian et al., 2012). The reads were aligned with the V2 sequence of the Criollo genome (Argout et al., 2017). Markers with unknown location were discarded from the analysis.

3.6-Genetic analysis

3.6.1-Linkage disequilibrium calculation

For the modern Nacional population, linkage disequilibrium calculations were performed by (Loor S., 2007).

For the Amazon native cocoa population, the linkage disequilibrium (LD) was calculated with Haploview 4.2 (Barrett et al., 2005) following the protocol of Sardos et al., (2016). The graphical representation of the LD decay was done with the R package "ggplot2" following the protocol of Sardos et al., (2016).

3.6.2-Genome-Wide Association Study (GWAS)

For the modern Nacional population, GWAS analyses were performed according to the protocol of (Colonges et al., 2021b).

For the Amazon native cocoa population, through GBS genotyping, about 50,000 SNP markers were detected. Markers with missing data or with a frequency of presence of the minor allele lower than 5% were discarded for this study. After these different filters, 5337 SNP markers were selected. A GWAS analysis was performed on SNP markers associated with biochemical (202 genotypes x 5337 markers) and sensory (159 accessions x 5337 markers) traits using the TASSEL v5 software.

For all traits, the choice of the mixed model (MLM) was the most relevant.

After comparison of the QQ-plot two methods were selected:

- The use of a MLM model with a kinship matrix considered as a random effect, added as co-variables to control the false positive rate was chosen for the association analyses of biochemical compounds.

- The use of a MLM model with a structure matrix, determined by performing a PCA (principal component analyses integrated with TASSEL v5 software), considered as a fixed effect, and with a relatedness matrix considered as a random effect added as co-variables to control the false positive rate.

In both cases, the relatedness matrix was constructed using the Identity by State (IBS) pairwise method proposed by Tassel v5. The option of not compressing and re-evaluating the variance components for each marker was chosen.

The threshold was determined using the R Simple M package based on the Bonferonni correction (Gao et al., 2008; Gao et al., 2010). For the modern Nacional population, the threshold corresponded to a p-value of about $1.79x10^{-5}$. For the Amazon native cocoa population, the threshold corresponded to a p-value of $1.68x10^{-5}$.

The physical maps with the representation of the association zones were created using SpiderMap v1.7.1 software (Rami, 2017). The size of the points correlates with the R2.

Candidate gene identification was performed in a 300kb region (on either side of the associated marker) using the *Theobroma cacao* genome sequence V2 (Argout et al., 2017).

3.7-Statistical analysis

The PCA analysis on the traits of interest were performed with the R package "Mixomics" and the graphical representations were performed with the R package "factoextra". Box plots were performed with the R package ggplot2. Student's t tests to check the significance of the differences in the box plots were carried out using the R package "stats".

4-Results

4.1-Characterisation of biochemical non-volatile compounds and sensorial traits related to bitterness and astringency in the modern Nacional population

The NIRS analyses revealed the nine biochemical contents for each tree of this population. Strong positive correlations could be observed for these different traits. The presence of all types of detected polyphenols seems to be correlated between them (Figure 44). No strong correlation could be observed between the different compounds identified by NIRS and the results of the sensory analyses (Figure 44).

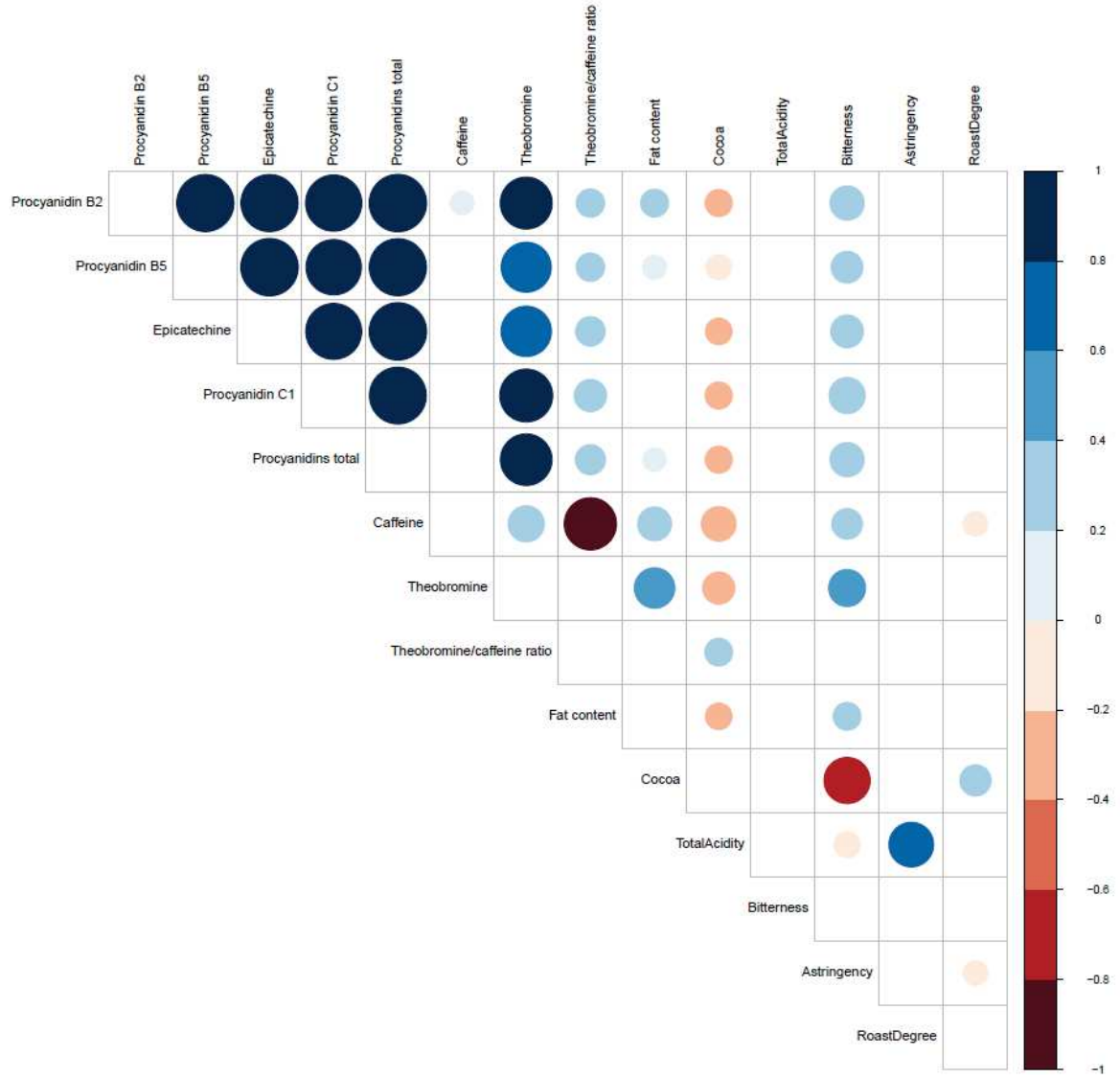


Figure 44: Correlation matrix of the results of the determination of non-volatile compounds by NIRS (in roasted beans) and sensory analysis (in liquors) from the modern Nacional population.
Non-volatile compounds are shown in black and sensory traits are shown in bold and brown. The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient (R2 correlation coefficient). The scale on the right indicates the interpretations of different colours (blue for positive correlation and red for negative correlation). The size of the circles correspond to the p-value corresponding to the calculation of each correlation coefficient. The p-value threshold for a significant correlation is 0.05.

PCA results from the NIRS assays gathering all the traits studied show a continuous variation within the modern Nacional population (Figure 45A). Axis 1 is mainly influenced by the concentrations of total procyanidins, procyanidins B2 and epicatechin. Axis 2 is mostly influenced by the amounts of caffeine, theobromine/caffeine ratio and fat content.

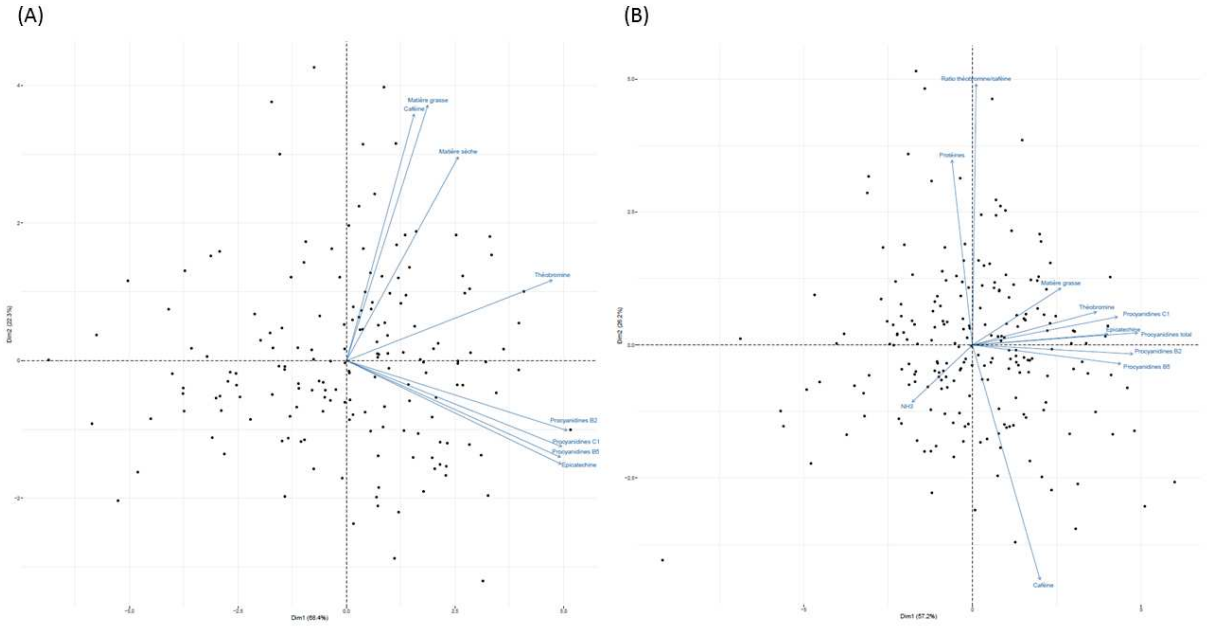


Figure 45: Graphical representation of PCA results.
(A) Results corresponding to NIRS determinations performed on cocoa beans from the modern Nacional population. (B) Results corresponding to the NIRS assays performed on cocoa beans from the Amazonian native cocoa tree population.

4.2-Characterisation of biochemical non-volatile compounds and sensorial traits related to bitterness and astringency in the native Amazonian cocoa population

The NIRS analyses revealed nine different traits. Strong positive and negative correlations could be observed between the different traits measured by NIRS. As in the case of the modern Nacional population, the presence of one type of polyphenol seems to be linked to the presence of all the other detected polyphenols (Figure 46). No strong correlation could be observed between the different compounds identified by NIRS and the results of the sensory analyses (Figure 46).

The PCA results from the NIRS results also show a continuous variation within the population (Figure 45B). Axis 1 is predominantly influenced by the concentrations of total procyanidins, B2 procyanidins and B5 procyanidins. Axis 2 is mainly influenced by the theobromine caffeine ratio, caffeine content and protein content.

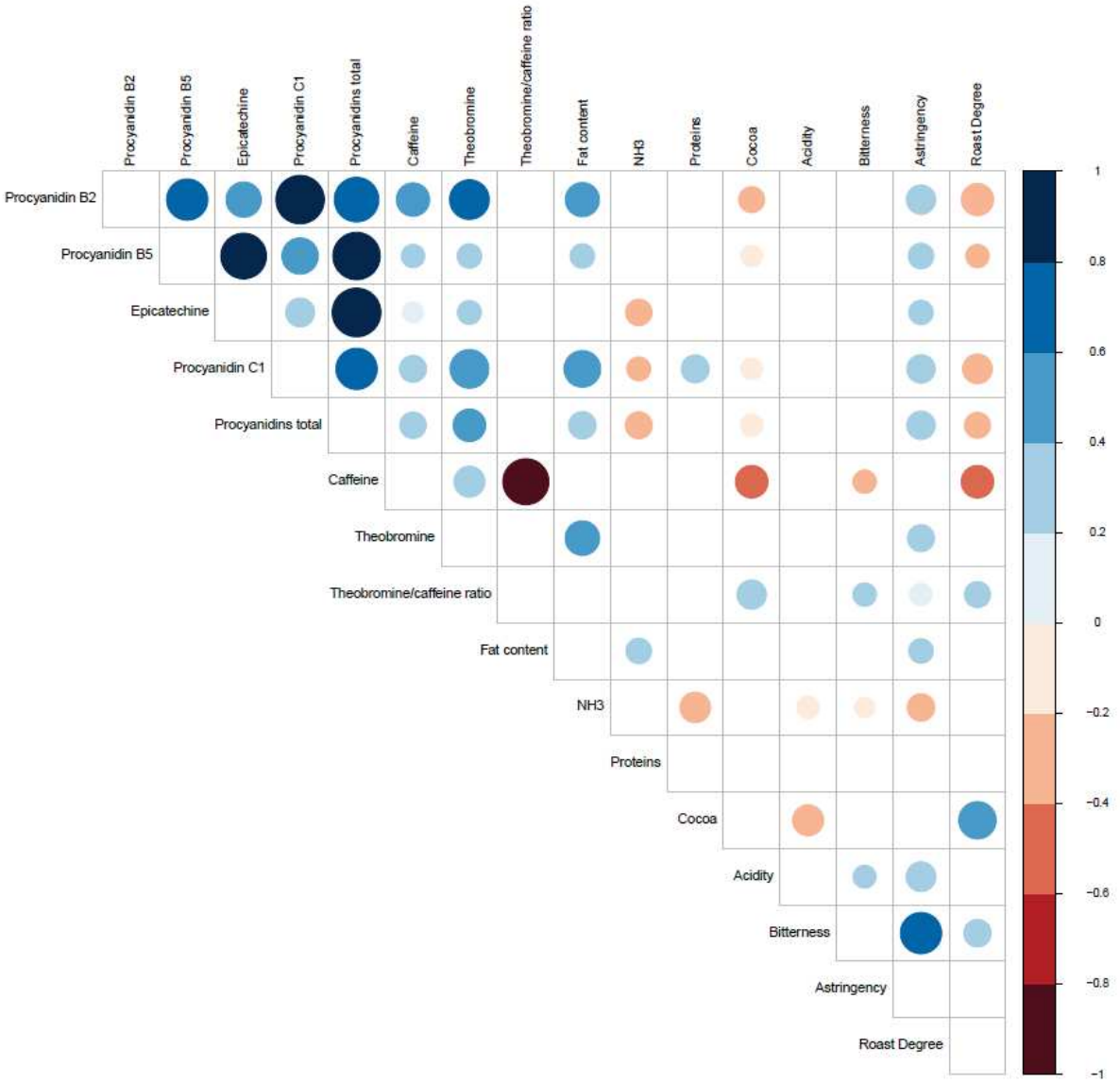


Figure 46: Correlation matrix of the results of NIRS determination of non-volatile compounds (in unroasted beans) and sensory analysis (in liquors) belonging to the native Amazonian cocoa population.

The correlations were calculated by the Pearson method. The white boxes represent no significant correlations. The colour of the circles corresponds to Pearson's correlation coefficient (R2 correlation coefficient). The scale on the right indicates the interpretations of different colours (blue for positive correlation and red for negative correlation). The areas of the circles correspond to the p-value corresponding to the calculation of each correlation coefficient. The p-value threshold for a significant correlation is 0.05.

4.3- Nacional modern vs native Amazonian cocoa populations

Significant differences could be observed between the concentrations and their variations among the nine traits measured by NIRS, depending on the cocoa tree population and on the bean treatment (Figure 47). Cocoa beans from the modern Nacional population (roasted beans) thus appeared to be richer in epicatechin, procyanidins B2, procyanidins B5, procyanidins C1, total procyanidins, theobromine, and had a higher theobromine/caffeine ratio (Figure 47). Cocoa beans from the native Amazonian population (unroasted beans) seem to have more fat content (Figure 47). As roasting is known to lower the polyphenol content

(Ioannone et al., 2015; Priftis et al., 2015), it seems that the Nacional modern population contains much more polyphenols than the native Amazon population.

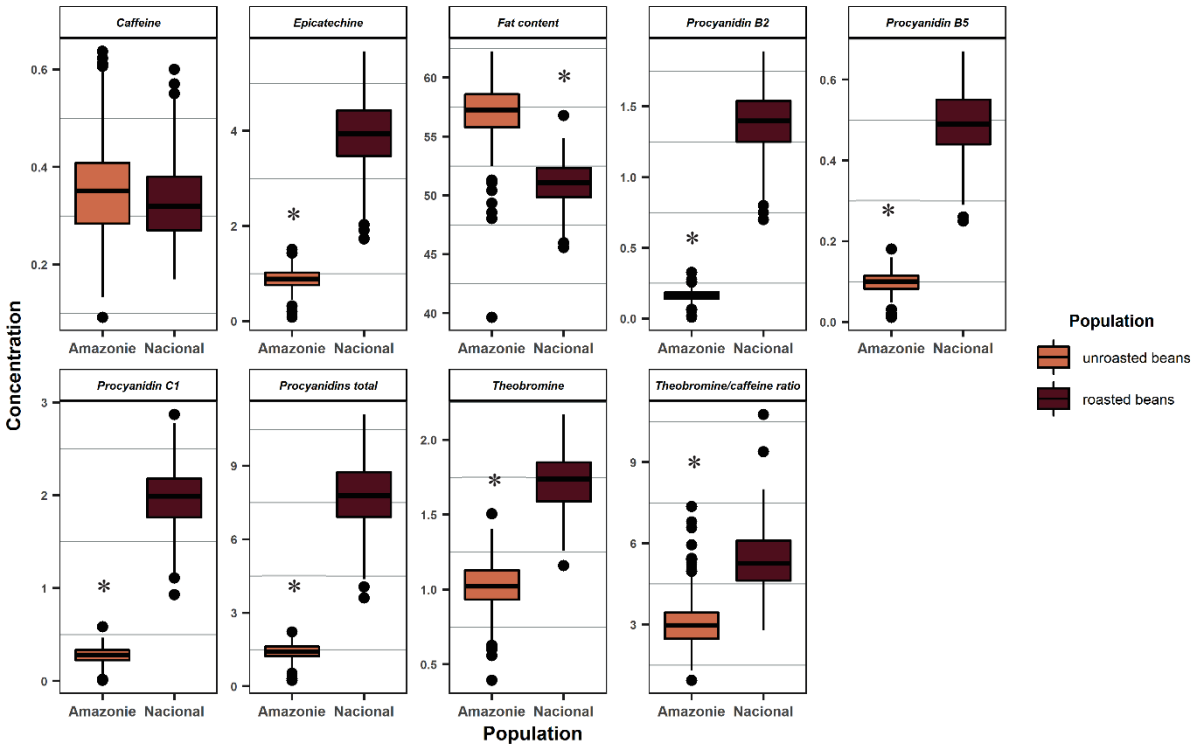


Figure 47: Boxplots representing the distribution of concentrations for each trait as a function of the cocoa tree population. A Student's t test was performed with a confidence level of 5%. Significantly different whisker boxes were annotated with a star. Unroasted beans from Amazonian population (in orange), and roasted beans from Nacional population (in brown).

The results of analyses of cocoa liquors from Amazonian trees and from trees of the modern Nacional variety, all made from roasted beans, show that Nacional is less astringent with a less pronounced cocoa taste and a lower taste of degree of roast (Figure 48).

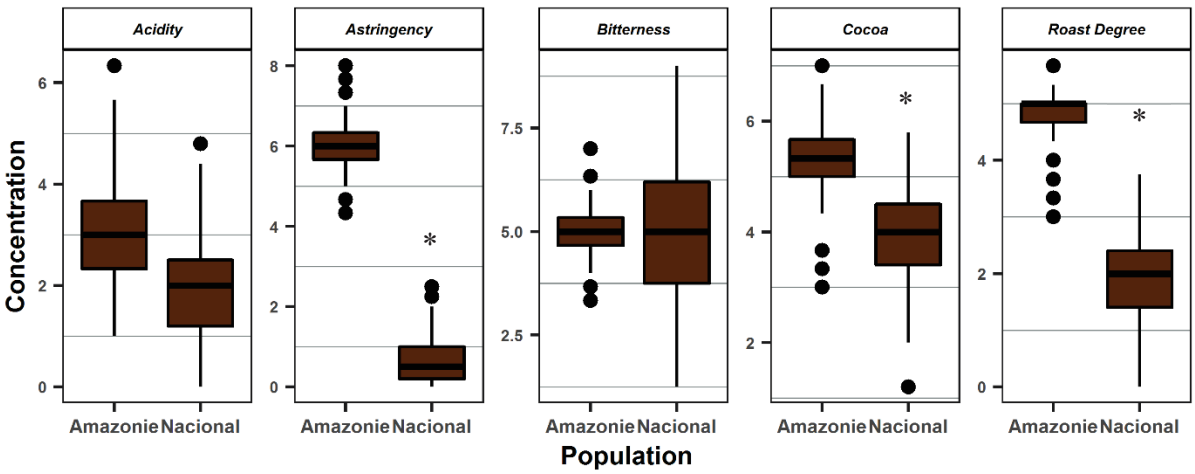


Figure 48: Boxplots representing the distribution of sensorial notes (made in liquors) for each trait as a function of the cocoa tree population.
A Student's t test was performed with a confidence level of 5%. Significantly different whisker boxes were annotated with a star

4.4-Genome-wide association genetics analyses

4.4.1-Analysis methods

For the native Amazonian population, the final dataset is composed of 5337 SNP markers. GWAS analyses were performed with different methods. The method with the least number of false positives was selected. This was the MLM method associated with the PCA data from the genetic data and the co-variate matching matrix for sensory traits, and the MLM method associated with the co-variate matching matrix for biochemical traits. All the significant association areas can be found in appendix 37.

4.4.2-Determination of the confidence intervals of the associations

The linkage disequilibrium of the native Amazonian population is approximately 1.2 cM (600kb) (Colonges et al., 2021a). This 600 kb limit was used to determine the confidence interval of associations. For each positive marker, we report an association zone of plus or minus 300kb, i.e. an association zone of 600 kb. If two or more markers have overlapping confidence intervals, they are grouped into a single association zone. The lowest and highest position of the grouped markers represent the confidence intervals of this zone.

4.5-Identification of significant associations for biochemical compounds

Fifty-three areas of significant associations were detected in relation to the biochemical compounds evaluated by NIRS analyses (two in the modern Nacional population and fifty-one in the native Amazon population). All the association zones are shown in appendix 38.

4.5.1-Identification of significant associations for biochemical compounds involved in the polyphenol biosynthetic pathway

No association zones were detected for polyphenol content in the modern Nacional population (Figure 49A). Of the fifty-three association zones detected in the Amazonian population, fifteen were detected in relation to the concentration of polyphenols, determined by NIRS in the population, on chromosomes 4, 6 and 8 (Figure 49B). Two co-locations are present on chromosome 4 and one on chromosome 6 in relation to epicatechin and total procyanidin concentration. One co-location on chromosome 8 is present in relation to epicatechin, procyanidin B5 and total procyanidin concentration (appendix 37 and 38).

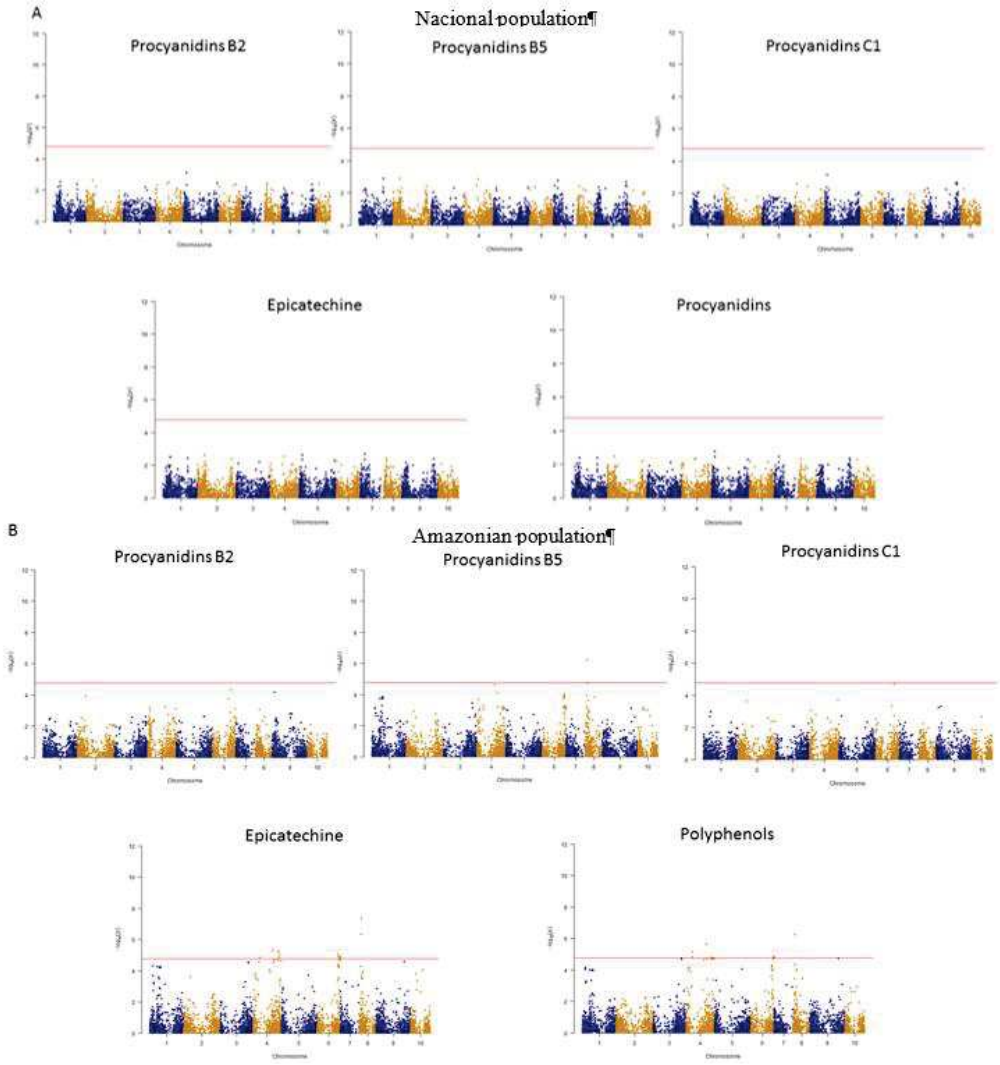


Figure 49: Manhattan plot representing the marker associations rate linked to polyphenols traits.
(A) Manhattan plot linked to polyphenols traits in Nacional population. (B) Manhattan plot linked to polyphenols traits in Amazonian population. The red line represent the threshold of significant association.

4.5.2-Identification of significant associations for biochemical compounds involved in the caffeine biosynthetic pathway

Of the two areas of association detected with the modern Nacional population, two were detected in relation to caffeine concentration on chromosomes 1 and 6 (Figure 50A).

Of the fifty-three areas of association detected with the Amazon population, six were detected in relation to the concentration of caffeine or the ratio of theobromine concentration to caffeine concentration, determined by NIRS in the population, on chromosomes 3, 4, 5, 7 and 10 (Figure 50B).

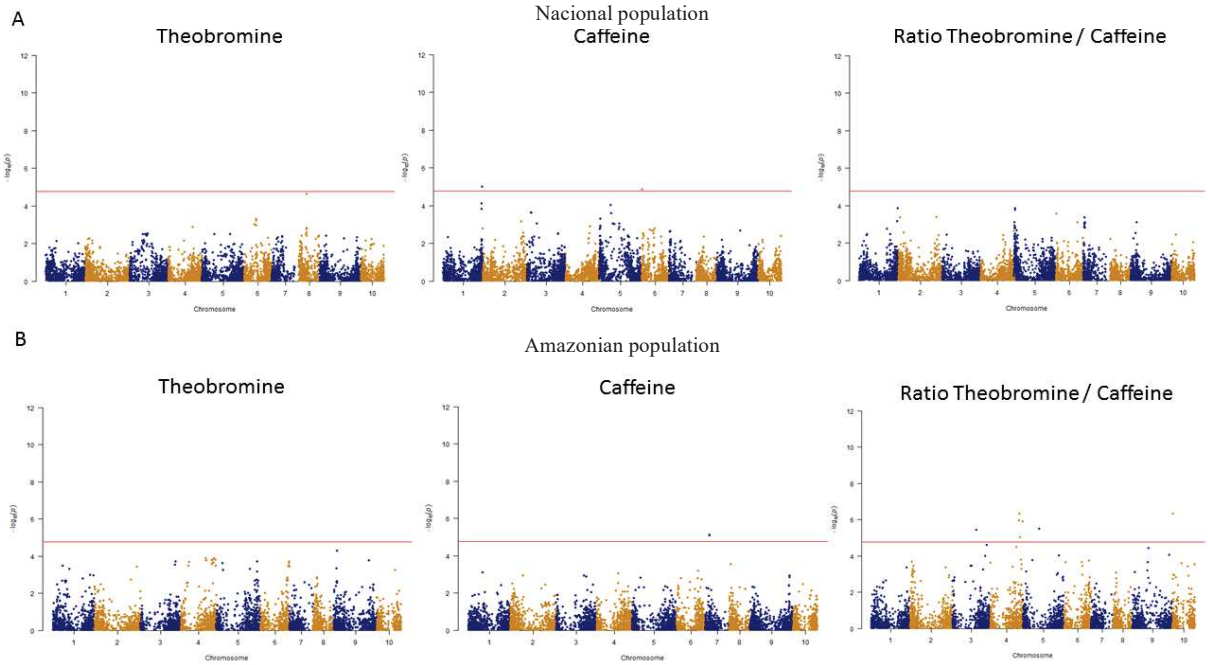


Figure 50: Manhattan plot representing the marker associations rate linked to caffeine and theobromine traits.
(A) Manhattan plot linked to caffeine and theobromine traits in Nacional population. (B) Manhattan plot linked to caffeine and theobromine traits in Amazonian population. The red line represent the threshold of significant association.

No co-locations between significant associations of the two populations was observed (appendix 37 and 38).

4.5.3-Identification of significant associations for traits related fat and proteins content

No significant association was identified for the Nacional population (Figure 51A). Twenty-nine significant association areas were detected in relation to fat content in the Amazonian population. They are located on all chromosomes except chromosome 2. One association zone was detected in relation to proteins content. It is located in chromosome 4 (Figure 51B).

No co-locations between significant associations of the two populations was observed.

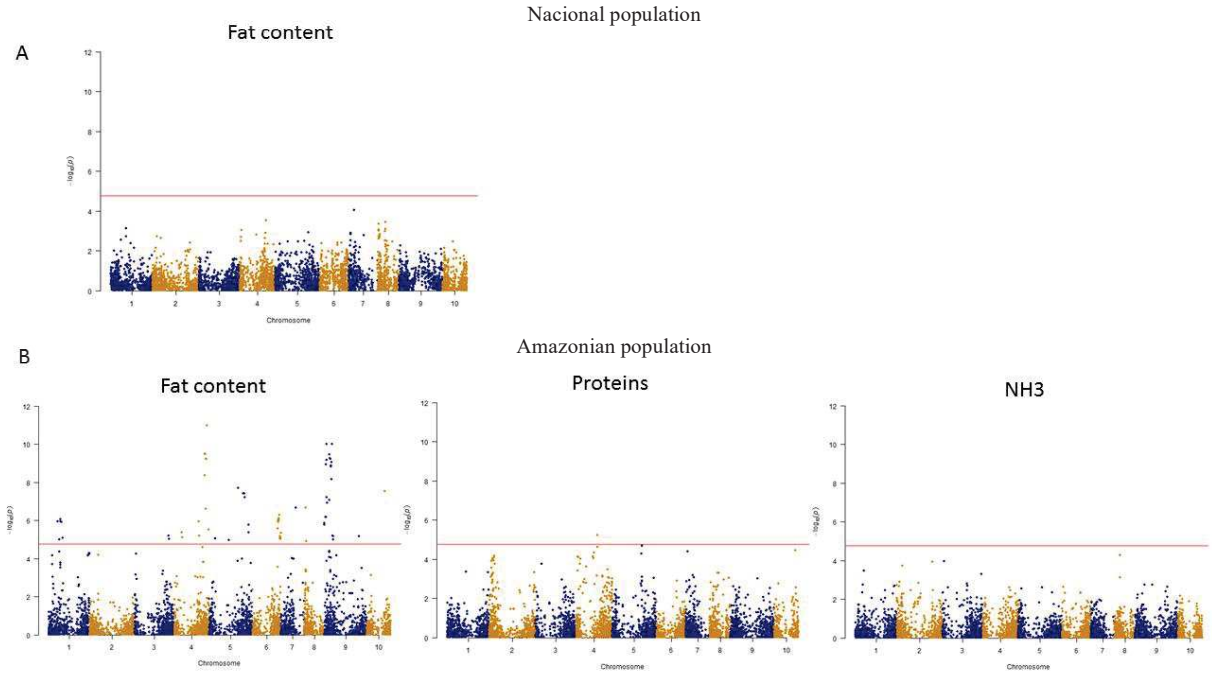


Figure 51: Manhattan plot representing the marker associations rate linked to fat and proteins content traits. (A) Manhattan plot linked to fat and proteins content traits in Nacional population. (B) Manhattan plot linked to fat and proteins content traits in Amazonian population. The red line represent the threshold of significant association.

4.6-Identification of significant associations for sensory traits

Twenty areas of association in relation to the scores established by sensory analysis were detected (three in the modern Nacional population and seventeen in the native Amazon population).

In the modern Nacional population, the three associations are related to astringency and are located on chromosome 2 (Figure 52A).

In the Amazon population, seventeen associations are related to bitterness and astringency (appendix 36). The areas of interest were detected on chromosomes 1, 3, 4, 5, 6, 9 and 10 (figure 52B and appendix 37). The perception of astringency and bitterness seems to be linked: eight co-locations were detected between these two sensorial traits.

No co-locations between the results of the two populations was observed.

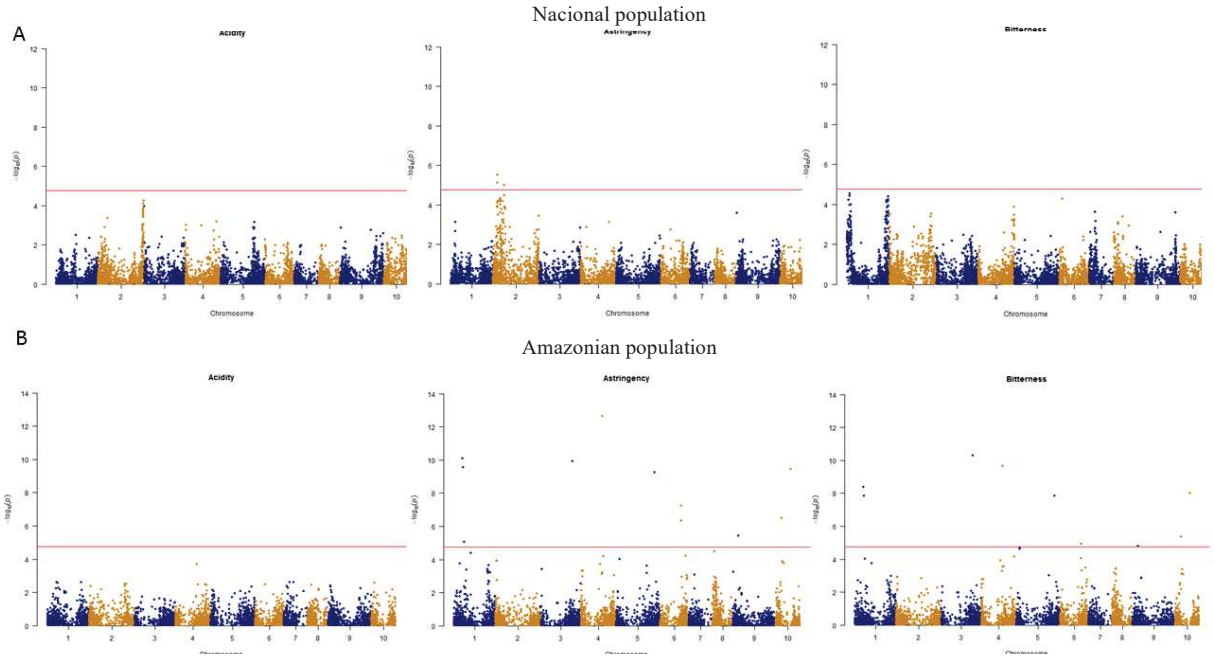


Figure 52: Manhattan plot representing the marker associations rate linked to astringency and bitterness traits.

(A) Manhattan plot linked to astringency and bitterness traits in Nacional population. (B) Manhattan plot linked to astringency and bitterness traits in Amazonian population. The red line represent the threshold of significant association.

4.7-Identification of candidate genes involved in the formation of biochemical compounds involved in bitterness

The set of association zones allowed the detection of 81 candidate genes potentially involved in the synthesis or degradation of the biochemical compounds identified by NIRS.

4.7.1-Candidate genes potentially involved in the polyphenol biosynthetic pathway

In the polyphenol association zones (epicatechin, procyanidin B5 and total procyanidins), ten candidate genes were identified (appendix 39). Their putative action is shown in appendix 39 and illustrated in Figure 53.

Figure 53: Diagram of the polyphenol biosynthetic pathway. Adapted from (Wollgast and Anklam, 2000; Chouhan et al., 2017). Biochemical compounds are shown in bold. Candidate genes identified in the association zones are shown in blue in this diagram and arrows indicate their putative functions in the biosynthetic pathway. The purple stars show the compounds and candidate genes identified in the Nacional modern population. The orange stars show the compounds and candidate genes identified in the native Amazonian population.

4.7.2-Candidate genes potentially involved in the caffeine biosynthetic pathway

One candidate gene has been identified in the purine compounds significant associations areas (Figure 54 and appendix 39).

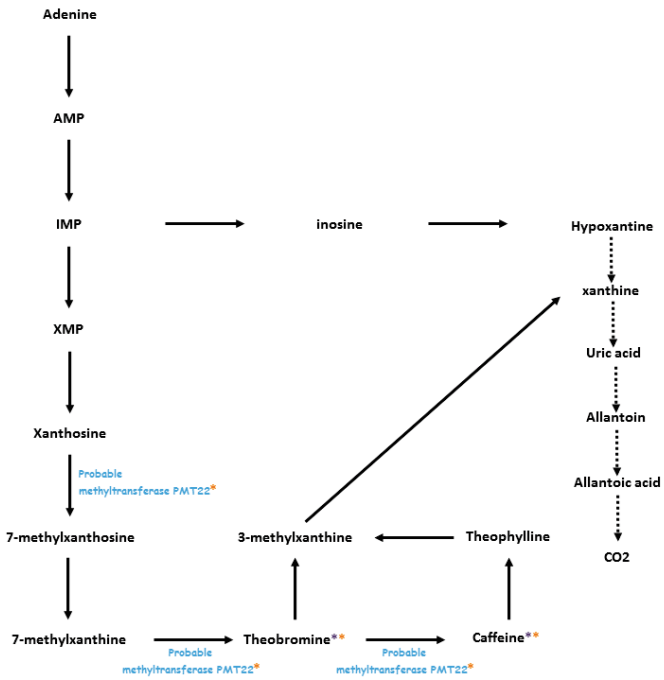


Figure 54: Scheme of caffeine biosynthesis. Adapted from Zheng et al., (2004).
Biochemical compounds are shown in bold. Candidate genes (in blue) located in the association zones are indicated at the side of arrows according to their putative functions in the biosynthetic pathway. The purple stars show the compounds and candidate genes identified in the Nacional modern population. The orange stars show the compounds and candidate genes identified in the native Amazonian population.

4.7.3-Candidate genes potentially involved in the fat biosynthetic and/or degradation pathway

Sixty-four candidate genes were identified in the areas of association with fat content identified in the Amazon population. Of these candidate genes, thirty appear to be involved in the synthesis of fatty acids or their precursors, twenty nine in lipid catabolism and five in transport of fatty acids (appendix 39).

4.7.4-Candidate genes involved in protein biosynthesis

Six candidate genes were identified in the significant association areas linked to protein content. Five of them have a function involved in protein transport and one gene has a hydrolase activity that could be responsible for the degradation of certain proteins (appendix 39).

5-Discussion

Two different cocoa populations were analysed in this work. Distinct results was observed between them. Sixty-eight association zones linked to non-volatile compounds and sensory analysis were detected for the Amazonian population and five for the Nacional population. Within these association zones, 81 candidate genes could be identified of which one in purine biosynthesis, 64 in fatty acid synthesis, degradation or transport, 10 in polyphenol biosynthesis and six in protein biosynthesis.

Other QTL studies have already been carried out using SSR markers, in relation to fat content but also to polyphenol content and to the presence of bitterness and astringency revealed by sensory analyses (Lanaud et al., 2003; Araújo et al., 2009; Argout et al., 2011; Mustiga et al., 2019). Some of the results found in this study are in common with previous studies. Three areas of association linked with fat content found in this new study co-locate with the associations reported by Argout et al., (2011) on chromosomes 3, 7, 9. Three others areas of association with fat content found in this new study co-locate with the associations found by Mustiga et al., (2019) on chromosomes 4, 5 and 9. One association zone linked with fat content found by Araújo et al., (2009) on chromosome 9 co-locate with an association zone found in this study. One area of association linked with astringency found in this new study co-locate with the association found by Lanaud et al., (2003) on chromosome 1.

The native Amazonian population showed more association areas (sixty-eight) than the modern Nacional population (three). This difference can have several causes:

The two populations are genetically different. The Nacional population has a narrow genetic basis, explained by only three main highly homozygous ancestors, contrary to the Amazon population, not selected, and which include native plants from Amazonia, with a higher allele richness (Loor S. et al., 2015). Therefore, the allele diversity is reduced in the Nacional population, limiting the number of segregations and associations revealed.

It can be also partly explained by the different treatment that the beans underwent before the NIRS analyses. Indeed, the beans from the modern Nacional were roasted in contrast to those from the Amazon population. Roasting is known to have an impact on polyphenol content (Jinap et al., 1998; Misnawi et al., 2005; Ioannone et al., 2015; Priftis et al., 2015). In some cases, roasting is responsible for the decrease in polyphenol content (Ioannone et al., 2015; Priftis et al., 2015), in others it is responsible for the increase in polyphenol content (Muzykiewicz-Szymańska et al., 2021). Another study has shown that roasting protocol can influence also the capacity of polyphenols to interact with protein and decrease the potential of astringency (Misnawi et al., 2005). These observations could also explain why the genetic component is more difficult to detect for modern Nacional individuals.

It can be concluded that trees belonging to the modern Nacional variety give beans with less astringency and a less strong chocolate flavour. No significant differences were observed for acidity and bitterness. However, the Amazon population seems to have a higher probability of bringing acidity than trees from the Nacional modern variety. The difference in bitterness

between the two populations has not been demonstrated, but the native Amazonian population shows a medium bitterness with little variation. The Nacional population, involving ancestors contrasted for this trait (Amelonado/Criollo/Nacional) could explain its larger variability for this trait.

Only one candidate gene involved in the biosynthetic pathway of purine biosynthesis or proteins biosynthesis have been identified. Further annotation of the cocoa genome could allow the identification of new genes. Furthermore, our method of searching for candidate genes based on annotations can be complemented with other methods without preconceptions to find genes whose function is not necessarily known.

At the side of its interaction on aroma, non-volatile cocoa compounds such as polyphenols are also useful compounds for human health (Cooper et al., 2008; Andújar et al., 2012). Characteristics related to bitterness and astringency are important to take into account when selecting clones to create new varieties, depending on the breeding objectives.

The results of our study have shown the polygenic nature of some traits as caffeine and theobromine content, fat content and polyphenol content. These results could provide useful information to define breeding strategies adapted to these traits, as a genomic selection strategy adapted to highly polygenic traits.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

CL, RGLS, conceived the experiment; AS, FD, JM, KC, MCL, ML conducted biochemical analyses; KC, JCJ, DC, CS, IS, FF conducted microfermentation, ES carried out sensorial analyses; KC, OF, XA carried out DNA experiments; KC, BR, RB, CL, analysed data; KC, RB, CL wrote the manuscript.

Funding

The study was funded by the United States Department of State (U.S. Foreign Ministry); the U.S. Embassy, Quito; the U.S. Department of Agriculture (USDA-ARS); the MUSE Amazcacao project with the reference ANR-16-IDEX-0006.

Acknowledgement

We thank the I-Site MUSE, Valrhona and the USDA, for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" programme with the reference ANR-16-IDEX-0006.

Discussion et conclusion

Discussion et conclusion

Les travaux de cette thèse, portant sur l'étude des bases génétiques et de la diversité des arômes des cacaoyers équatoriens, ont été réalisés à partir de l'analyse de deux populations de cacaoyers : une population cultivée de Nacional moderne et une population native d'Amazonie, directement collectée en 2010 dans la zone d'origine du Nacional. Les analyses de GWAS ont mis en lumière un grand nombre de zones d'association en lien avec des composés volatils et des notes sensorielles responsables des différents arômes de qualité (notes florales, fruitées, …) et également en lien avec les possibles défauts (notes végétales, amertume et astringence). Dans l'ensemble de ces zones d'association, 1831 gènes candidats ont été détectés dans les deux populations (354 dans la première et 1477 dans la deuxième) dont 97 communs aux deux populations. Au total, sept voies de biosynthèses ont pu être mises en lumière chez le cacaoyer. Elles sont impliquées dans la synthèse des composés volatils et non volatils participant à la flaveur du cacao. Parmi elles, deux sont communes aux arômes floraux et fruités :

- La voie de biosynthèse des monoterpènes,
- La voie de dégradation du L-phénylalanine,

Trois autres voies de biosynthèse concernent les arômes fruités :

- La voie de dégradation des acides gras,
- La voie de dégradation des sucres,
- La voie de dégradation des protéines,

Enfin, deux autres voies métaboliques concernent l'amertume et l'astringence :

- La voie de biosynthèse de la caféine,
- La voie de biosynthèse des flavonoïdes.

La voie de biosynthèse des monoterpènes

La voie de biosynthèse des monoterpènes a depuis longtemps été identifiée comme jouant un rôle clef dans la synthèse de composés aromatiques chez le cacaoyer. Ziegleder (1990) avait montré que la teneur en linalol, un monoterpène connu pour son arôme floral, était un élément pour discriminer le Nacional des autres variétés. Cevallos-Cevallos et al., (2018) ont également montré que la teneur en linalol et époxylinalol, un autre monoterpène connu pour son arôme floral, était plus forte chez les cacaoyers de type Criollo que chez ceux du Forastero et du Nacional.

Les différents résultats des analyses GWAS menées lors de cette thèse et les données déjà disponibles sur cette voie de biosynthèse chez d'autres espèces végétales (Pichersky et al., 1994; Chen et al., 2010; Gao et al., 2018), ont permis d'élaborer un schéma de biosynthèse hypothétique des monoterpènes chez le cacaoyer (Figure 55). Des gènes candidats identifiés dans les zones d'association en lien avec les monoterpènes, dont la fonction a été décrite comme impliquée dans cette même voie de biosynthèse, sont également des indices supplémentaires dans la recherche des déterminismes génétiques et biochimiques des arômes du cacao. La répétition de la présence des gènes candidats dans les zones d'association des différentes populations d'études augmente la probabilité de leur implication dans la synthèse des arômes du cacao. Au total, seize gènes candidats communs aux deux populations, sont impliqués dans cette voie de biosynthèse.

Les résultats de nos études montrent que la voie de biosynthèse des monoterpènes pourrait jouer un rôle important dans la synthèse des composés aux notes florales, (linalol, époxylinalol, trans linalol oxyde furanique, cis-linalol oxyde pyranoique, 1,2-dihydrolinalol, cis-β-ocimène et allo-ocimène), fruitées (D-limonène), épicées (α-terpinène, β-myrcène, camphène et γ-terpinène) et végétales (α-pinène, β-pinène et trans-β-ocimène) (Figure 55).

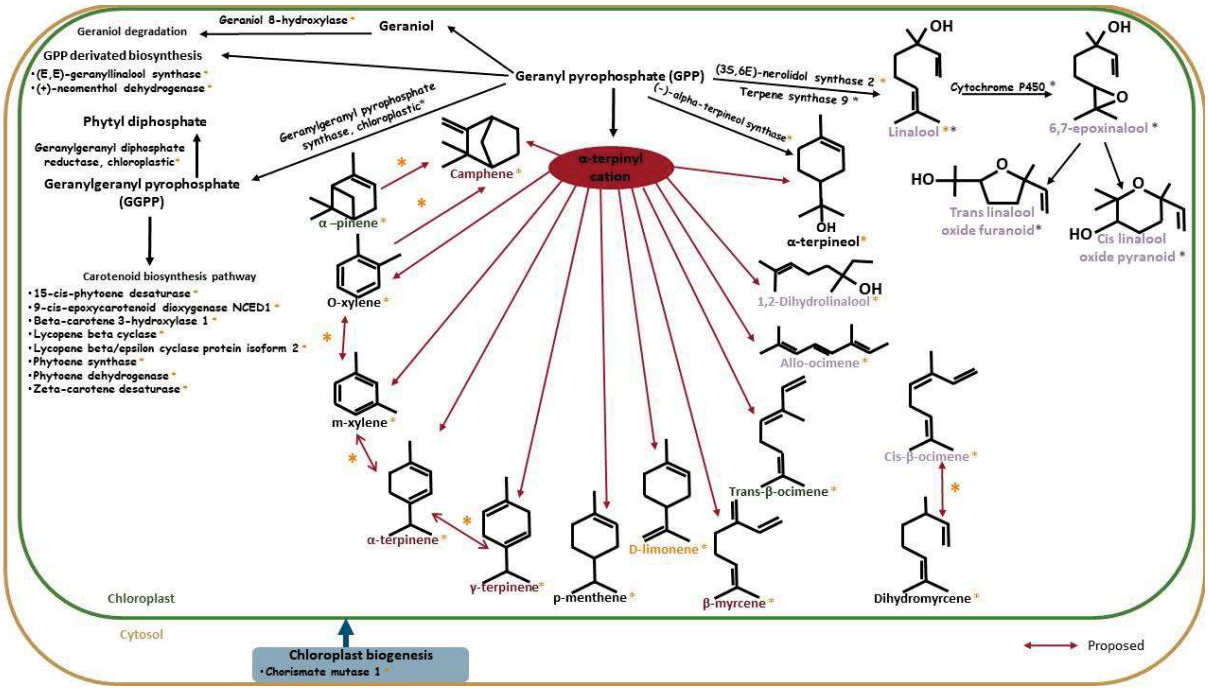


Figure 55: Schéma de la voie de biosynthèse des monoterpènes adapté de Bohlmann et al., (1998).

La couleur des noms des composés indique leur note aromatique potentielle, violet : florale ; rouge : épicée ; orange : fruitée ; vert : végétale. Les gènes candidats sont indiqués en bleu. Les étoiles oranges représentent les données détectés pour la population de cacaoyers natifs d'Amazonie. Les étoiles violettes représentent les données détectés pour la population de Nacional moderne.

La voie de dégradation du L-phénylalanine et la voie de biosynthèse des flavonoïdes

La voie de dégradation du L-phénylalanine peut produire un grand nombre de composés biochimiques. Elle permet d'une part la synthèse de composés volatils avec un cycle benzénique comme l'acétophénone, le benzaldéhyde ou le 2-phényléthanol, et d'autre part, elle peut aboutir à la synthèse des flavonoïdes et donc des polyphénols, polymères de flavonoïdes (Punyasiri et al., 2004; Sakai et al., 2007; Dong et al., 2012; Roccia et al., 2019).

Tout comme pour la voie de biosynthèse des monoterpènes, des résultats d'analyses GWAS communs entre les deux populations d'études ont été révélés et ont permis l'identification de quatorze gènes candidats pour la voie de dégradation du L-phénylalanine. Ces résultats ainsi que les informations disponibles sur cette voie de dégradation chez d'autres espèces végétales (Dong et al., 2012; Roccia et al., 2019), ont également permis de schématiser une voie de dégradation hypothétique du L-phénylalanine chez le cacaoyer qui pourrait être responsable de la synthèse de ces arômes (Figure 56). Cette voie pourrait permettre la synthèse de composés connus pour avoir une note florale (2-phényléthanol, acétate de 2-phényléthyle, acétate de benzyle, alcool benzylique, acétophénone, trans alcool cinnamique), fruitée (benzoate d'éthyle, benzaldéhyde), épicée (cinnamaldéhyde) ou végétale (acétate de 1-phényléthyle).

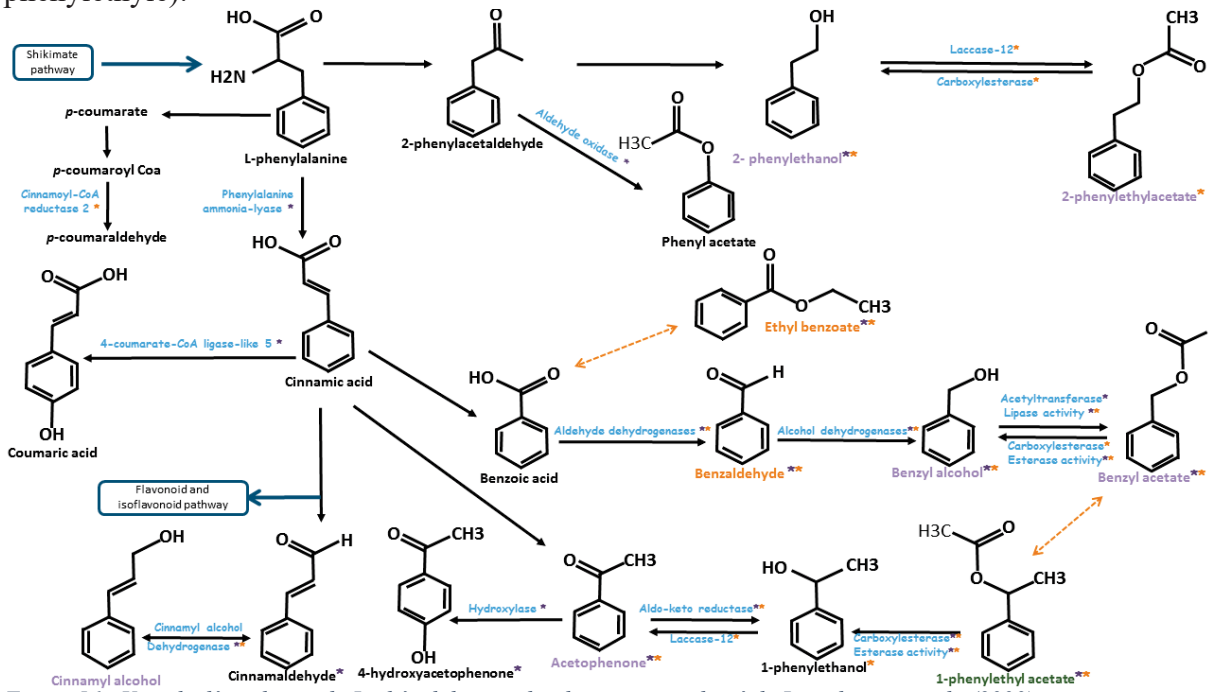


Figure 56 : Voie de dégradation du L-phénylalanine chez le cacaoyer adapté de Lapadatescu et al., (2000)
La couleur des noms des composés indique leur note aromatique potentielle, violet : florale ; rouge : épicée ; orange : fruitée ; vert : végétale. Les gènes candidats sont indiqués en bleu. Les étoiles oranges représentent les données détectés pour la population de cacaoyers natifs d'Amazonie. Les étoiles violettes représentent les données détectés pour la population de Nacional moderne.

La voie de dégradation du L-phénylalanine mène également à la synthèse de polyphénols via la synthèse de *p*-coumarate comme cela a été montré chez le thé ou la pomme (Punyasiri et al., 2004; Henry-Kirk et al., 2012) ou chez *Theobroma cacao* comme décrit dans la revue de Wollgast and Anklam, (2000). La production de dihydroflavonoles permet la synthèse de catéchine et d'épicatéchine. Enfin, la catéchine et l'épicatéchine permettent la synthèse de proanthocyanidines ou tanins condensés (polyphénols B2, B5, C1). Grâce aux résultats de GWAS obtenus dans les deux populations d'études, des gènes candidats ont également été détectés et une voie de biosynthèse hypothétique chez le cacaoyer a pu être schématisée (Figure 57), adaptée de Wollgast and Anklam, (2000) et Chouhan et al., (2017).

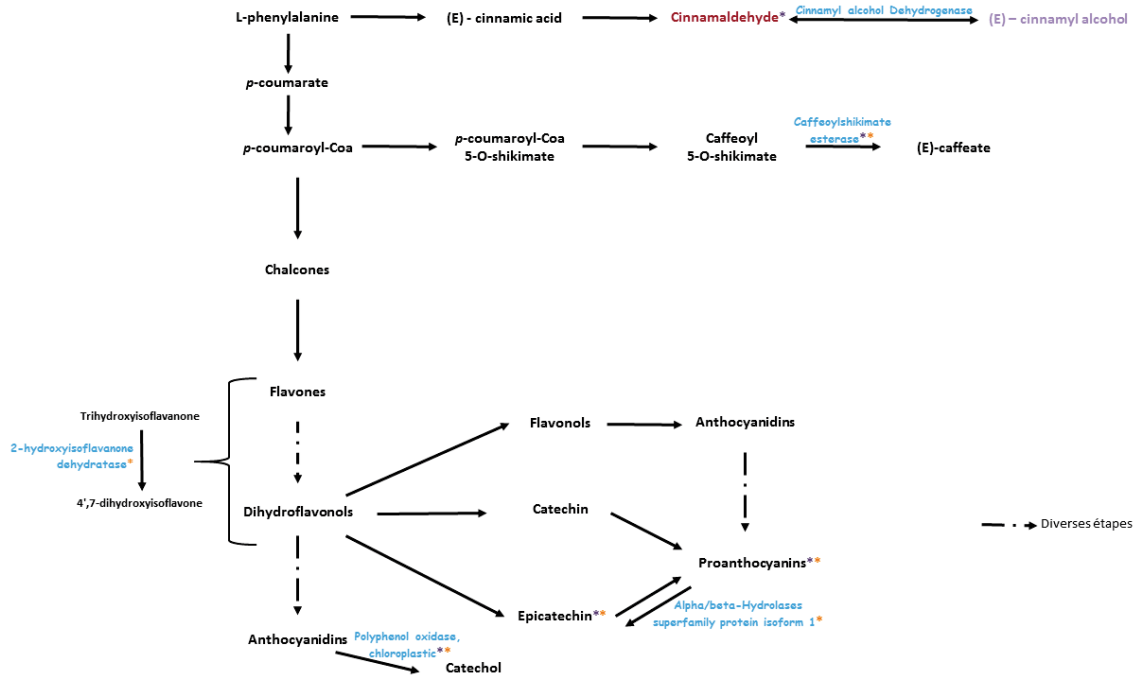


Figure 57: Schéma de la voie de biosynthèse des polyphénols chez le cacaoyer.
La couleur des noms des composés indique leur note aromatique potentielle, violet : florale et rouge : épicée. Les gènes candidats sont indiqués en bleu. Les étoiles oranges représentent les données détectés pour la population de cacaoyers natifs d'Amazonie. Les étoiles violettes représentent les données détectés pour la population de Nacional moderne.

La voie de dégradation des acides gras, des sucres et des protéines

Les voies de dégradations des acides gras, des sucres et des protéines sont des voies métaboliques liées qui interagissent à plusieurs niveaux. Ces différentes voies de biosynthèse aboutissent à la production d'alcools, d'aldéhydes, de cétones, d'esters (Figure 58) ou encore à celle de précurseurs de la réaction de Maillard qui permettent la synthèse de composés appartenant par exemple à la famille des pyrazines et des furannes.

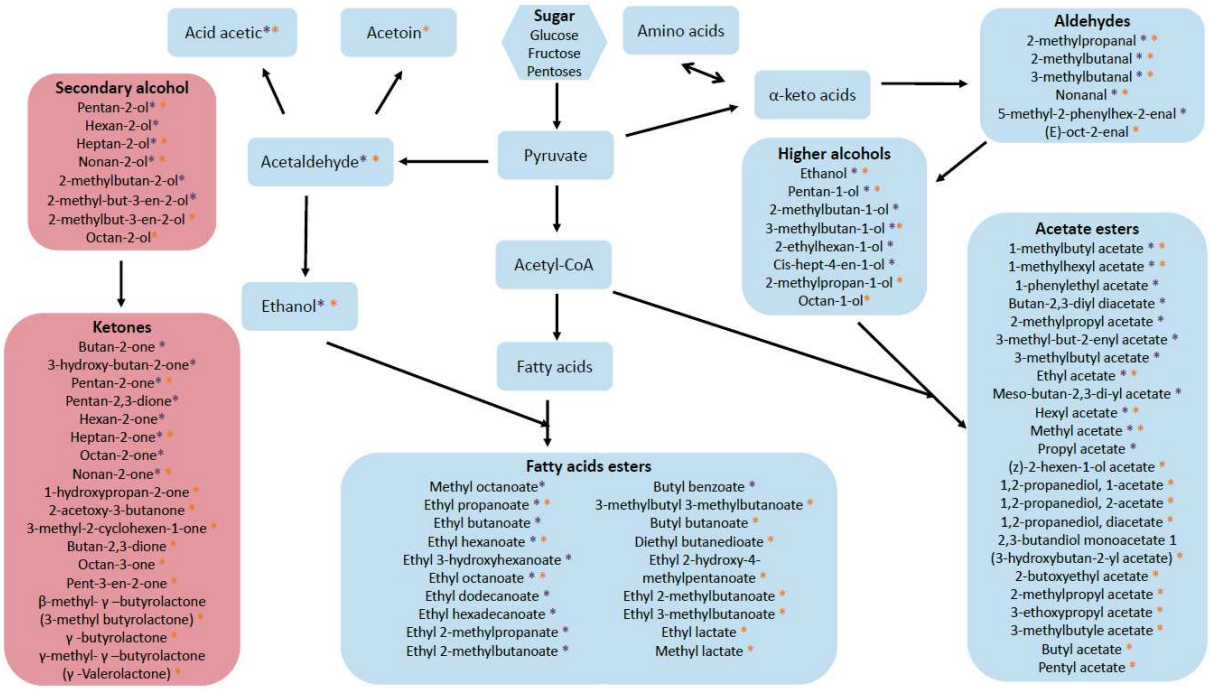


Figure 58: Schéma de la voie de biosynthèse des composés issus de la dégradation de lipides et des sucres chez le cacaoyer
Les étoiles à côté des noms des composés désignent les composés identifiés par GC-MS dans les fèves de cacaos étudiées. Les étoiles oranges représentent les données détectés pour la population de cacaoyers sauvages d'Amazonie. Les étoiles violettes représentent les données détectés pour la population de Nacional moderne.

La réaction de Maillard, comme cela a été décrit dans le premier chapitre, est un ensemble de réactions chimiques qui ont lieu lors d'un processus de transfert de chaleur tels que le séchage et/ou la torréfaction pour le cacao. Ces réactions permettent la synthèse de composés aromatiques intéressants souvent à l'origine de notes empyreumatiques (grillé, caramel, chocolat) mais également à des notes de fruits secs (noix, noisette, amande). Plusieurs composés de la famille des pyrazines connus pour avoir une note de fruits secs ont depuis longtemps été identifiés chez le cacaoyer comme la tetraméthylpyrazine ou la triméthylpyrazine (Arnoldi et al., 1988; Misnawi et al., 2002; Jinap et al., 2008). Des associations ont pu être détectées par GWAS en lien avec des composés de la famille des pyrazines ou celle des furannes. Certaines associations sont communes à différents composés volatils (des co-localisations ont été observées par exemple entre les zones d'association liées à la 2,3,5-trimethylpyrazine et la 2,3,5-trimethyl-6-ethylpyrazine ou encore à la 2,3-dimethylpyrazine-5-ethylpyrazine et le furfural), ce qui laisse penser que ces différents composés pourraient avoir le ou les mêmes précurseurs. Dans ces zones d'association, des gènes candidats impliqués dans la dégradation de protéines ou de sucres (principaux précurseurs de la réaction de Maillard) ont pu être détectés, dont trente gènes candidats communs aux deux populations.

La voie de biosynthèse de la caféine

La voie de biosynthèse de la caféine est possible grâce à la production de Xanthosine qui sera transformé en 7-methylxantiosine, lui-même transformé en 7-methylxanthine. Le 7-methylxanthine est le précurseur de la théobromine, elle-même précurseur de la caféine (Zheng et al., 2004; Tu et al., 2010).

Grâce aux résultats de GWAS obtenus dans les deux populations d'études, un gène candidat pouvant agir à plusieurs endroits de la voie de biosynthèse a également été détecté et une voie de biosynthèse hypothétique chez le cacaoyer a pu être schématisée (Figure 59), adaptée de Zheng et al., (2004).

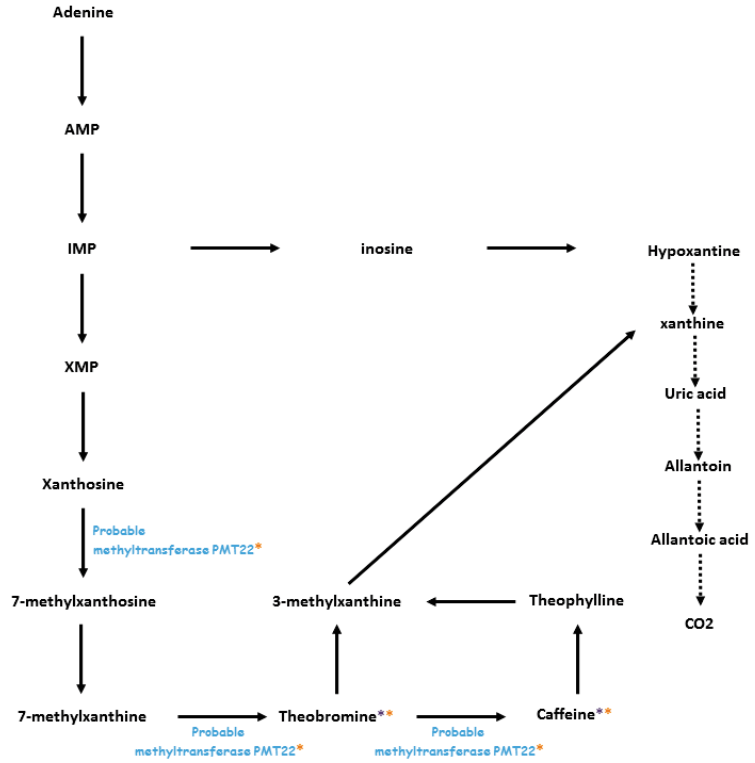


Figure 59: Schéma de la voie de biosynthèse de la caféine chez le cacaoyer.

Les étoiles à côté des noms des composés désignent les composés identifiés par NIRS dans les fèves de cacaos étudiées. Les étoiles oranges représentent les données détectés pour la population de cacaoyers sauvages d'Amazonie. Les étoiles violettes représentent les données détectés pour la population de Nacional moderne.

En conclusion, parmi les résultats obtenus, certaines zones d'association n'ont pas permis de détecter des gènes candidats. Cela peut venir de plusieurs raisons. Premièrement, la recherche de gènes candidats s'est faite par une méthode avec à priori se basant sur la connaissance des fonctions des gènes, il est donc possible que certains gènes soient mal annotés (connaissances en perpétuelle évolution) ou absents dans le fond génétique Criollo. Deuxièmement, il est possible que les intervalles de confiance choisis soient trop stringents et que le gène causal de l'association soit au-delà des limites. Il est également possible que ce soit

un changement dans un promoteur ou une région non codante qui provoque l'association et dans ce cas nous n'avons pas pu l'observer avec les données dont nous disposions. Enfin, la possibilité que l'association détectée soit un faux positif ne peut être complétement écartée. Des recherches de gènes candidats sans à priori dans ces zones d'association, ainsi que dans les autres zones, pourraient permettre de compléter ces résultats.

Les résultats des études GWAS permettent d'avoir une première idée des voies de biosynthèse utilisées par la plante pour synthétiser les composés aromatiques et des gènes responsables de leur variation. Cependant, ces résultats ne permettent pas de déterminer les voies de biosynthèse complètes, d'autant plus que certaines molécules intermédiaires peuvent être synthétisées par les micro-organismes fermentaires. Les molécules intermédiaires ont également pu être consommées par la plante pour la synthèse de composés nécessaires à son fonctionnement. Ils se retrouvent alors en trop petite quantité pour être détectés par SPME GC-MS, la méthode de dosage des composés volatils que nous avons utilisée. D'autres méthodes d'extraction des composés volatils, comme la méthode SAFE, qui permet l'extraction intégrale des composés volatils, pourraient compléter ces voies de biosynthèse hypothétiques. L'intérêt d'utiliser les résultats de GWAS est que les composés pour lesquels des associations ont été détectées sont des composés qui dépendent de la génétique de la plante et sont donc synthétisés par elle.

Une étude de chromatographie-gazeuse couplé à l'olfactométrie (GCO) permettrait également d'enrichir les connaissances sur les molécules aromatiques impactant les arômes du cacao des populations que nous avons étudié. Cette méthode d'analyse permet de mettre en lumière dans un échantillon de cacao donné les molécules aromatiques contenues en assez grande quantité pour être perçues par nos récepteurs olfactifs. En effet, la présence d'une molécule ainsi que son déterminisme génétique chez *T. cacao* sont de bons indices sur la possibilité de son implication dans l'arôme final du cacao, mais cela ne donne aucune information sur l'interaction avec nos récepteurs. Grâce à la GCO, il serait donc possible de déterminer si les molécules en associations sont également percues. Leurs notes aromatiques et la variation de cette perception en fonction de leur concentration pourraient donc être décrites. Une étude de GCO sur le chocolat noir à d'ailleurs pu permettre l'identification de composés odorants clefs dans la flaveur de ce chocolat (Counet et al., 2002) dont certains sont communs avce ceux identifiés par nos analyses génétiques tels que le 2-phényléthanol (note florale), le 3-methylbutanal (note chocolat), la méthylpyrazine (notes de noisettes et végétales), la 2,3-

dimethylpyrazine (notes de noisettes et grillées) et la tetramethylpyrazine (notes café au lait, moccha, grillée et végétales).

Les mécanismes de défense générale des plantes contre les stress biotiques et abiotiques potentiellement impliqués dans la production des arômes.

Grâce aux résultats préliminaires d'études d'expressions présentés dans le chapitre 3.2 (acte de congrès du Weurman), il a été montré que certains gènes candidats s'exprimaient plus fortement ou uniquement lors des premières vingt-quatre heures de la fermentation. Ces résultats ainsi que ceux présentés par Sabau et al., (2006), suggèrent que les fèves (qui sont encore vivantes à cette étape), interagissent ou réagissent aux nouvelles conditions du milieu produites par les levures et les bactéries et déclenchent différents mécanismes de défense. Cette hypothèse est d'autant plus probable que de nombreux gènes impliqués dans ces mécanismes ont été identifiés dans les zones d'association en lien avec les composés biochimiques et les analyses sensorielles.

Les fèves déclencheraient donc leurs différents mécanismes de défenses tels que la résistance systémique induite (ISR) et la résistance systémique acquise (SAR) impliquant plusieurs cascades de réactions. Ces systèmes de résistance donnent lieu à l'expression de gènes de défense qui eux-mêmes déclenchent la synthèse de métabolites secondaires pouvant être aromatiques (Sarma et al., 2015), tels les terpènes.

Le déclenchement des mécanismes de défense en réponse à la présence de micro-organismes fermentaires n'est cependant pas la seule raison à la synthèse de composés aromatiques chez le cacaoyer. En effet, les résultats préliminaires d'études d'expressions présentés dans le chapitre 3.2, montrent également que certains gènes candidats sont plus exprimés dans les fèves avant la fermentation et pendant leur développement dans les cabosses. D'autres mécanismes pourraient donc être mis en route pour la production de composés aromatiques comme la perception d'un stress hydrique pendant la croissance des fruits ou le stress provoqué par la récolte.

L'apport des différents ancêtres dans les arômes du Nacional

Les accessions appartenant à la variété de Nacional moderne actuellement cultivée en Equateur sont des hybrides entre des individus appartenant à la variété de Nacional ancestrale et des individus de type Trinitario (Bartley, 2005; Loor S. et al., 2009). Plusieurs publications ont rapporté une dilution de la saveur Arriba (saveur typique du Nacional ancestrale, connue

pour ses notes florales) dans les fèves de cacao issues des individus de Nacional moderne (Loor S. et al., 2009; Beckett et al., 2017).

L'une des études de cette thèse a eu pour objectif de déterminer quel(s) ancêtre(s) apportai(en)t les allèles favorables au développement des notes florales et/ou fruitées, de l'amertume et de l'astringence des fèves de cacao issues des arbres appartenant à la variété de Nacional moderne. Les résultats des analyses ont montré que les trois principaux ancêtres (le Nacional ancestral, le Criollo et l'Amelonado) apportaient divers allèles favorables pour le développement des notes florales et fruitées recherchées. Il a aussi été observé que les trois ancêtres apportaient divers allèles favorables au développement de l'amertume et de l'astringence. Les zones d'association liées aux caractères floraux et fruités et celles liées à l'amertume et à l'astringence ne semblent pas liées génétiquement.

Lors de croisements, il est possible d'observer un phénomène de supériorité chez les descendants par rapport aux parents, c'est ce que l'on appelle un effet d'hétérosis. Si les parents sont des homozygotes (ce qui est le cas ici), le génotype descendant peut-être hétérozygote et avoir un effet supérieur par rapport aux parents homozygotes pour des allèles différents. Dans notre étude, nous avons été capables de mettre en évidence des effets d'hétérosis. Nous avons donc des génotypes hétérozygotes qui sont plus favorables au développement des notes aromatiques (florales et fruitées) que les génotypes homozygotes, c'est également vrai pour l'amertume.

Ces dernières observations suggèrent que les croisements entre des variétés de cacaoyers très différentes peuvent être positifs pour accroître la valeur aromatique si l'on se base sur la présence de composés volatils aromatiques ainsi que sur les évaluations sensorielles. Même si une dilution de l'arôme floral a été perçue par plusieurs études, la domestication de la variété Nacional, qui s'est produite au cours du dernier siècle par l'hybridation naturelle d'arbres issus de la variété Nacional ancestrale et de Trinitario vénézuéliens, a permis d'enrichir génétiquement sa palette aromatique. L'augmentation de l'amertume a pu masquer les différentes notes aromatiques recherchées telles que les notes florales ou fruitées.

Ainsi, l'introgression du génome de l'Amelonado (variété non aromatique) dans une variété aromatique peut avoir un impact positif sur les caractères aromatiques, s'ils sont contrôlés. Dans le but d'améliorer une variété de cacao fin, toutes les variétés sont à prendre en considération, si l'on prête une attention particulière à la sélection des zones du génome favorable aux notes recherchées ainsi qu'à la contre-sélection des zones favorables au

développement de l'amertume et de l'astringence. La prédiction génomique ainsi que la sélection assistée par marqueurs pourraient être des outils permettant de prédire la valeur aromatique des cacaoyers et ainsi sélectionner les cacaoyers avec le meilleur profil.

La diversité génétique et aromatique des cacaoyers natifs d'Amazonie

Dans le dernier chapitre, la diversité et le déterminisme génétique et biochimique des arômes de cacaoyers issus de prospections faites au Sud de l'Amazonie Equatorienne dans l'aire d'origine du Nacional (Loor S. et al., 2012) ont été étudiés. En comparaison avec la population de Nacional moderne, la population de cacaoyers natifs d'Amazonie montre un grand nombre de nouveaux caractères favorables à la synthèse d'arômes recherchés par les chocolatiers et appréciés par les consommateurs. Elle constitue donc une nouvelle collection de ressources génétiques riche pour les sélectionneurs. Cependant, cette population montre également un nombre de caractères défavorables (arômes déplaisants). Comme pour toutes les plantes sauvages ou peu domestiquées, une domestication des caractères favorables sera nécessaire pour sélectionner les parties du génome voulues.

Les composés participant aux arômes recherchés ne semblent pas liés aux composés responsables des arômes déplaisants. En effet, aucune ou de très faibles corrélations ont été mesurées entre ces différents composés et peu de zones d'association co-localisent. La sélection des composés responsables des arômes recherchés combinée à une contre-sélection des composés responsables des arômes déplaisants semble donc possible. Cette population présente une palette aromatique plus grande à disposition des sélectionneurs. La combinaison de ces génotypes entre eux ou avec d'autres variétés aromatiques, comme le Nacional moderne, semble être prometteuse pour la production de nouvelles variétés aromatiques originales adaptées aux différents climats équatoriens.

Perspectives

Ces travaux ont permis d'initier ou d'enrichir les connaissances sur les déterminismes génétiques et biochimiques des arômes de cacao. Pour élargir nos connaissances dans ce domaine beaucoup de recherches restent encore à réaliser. La réalisation d'autres analyses GWAS sur de grandes populations très contrastées génétiquement, ou très contrastées géographiquement (Afrique, Asie et autres pays d'Amérique latine) permettraient d'affiner les connaissances et d'apporter de nouveaux indices sur l'implication des gènes candidats détectés. Afin de réduire les intervalles de confiance et de préciser les zones d'association, une méta-

analyse des deux populations étudiées dans cette thèse, mais également d'autres populations, pourrait être faite.

Dans le but de récolter de nouveaux indices sur l'implication des gènes candidats détectés par GWAS, une analyse de l'expression des gènes pourrait être faite dans des fèves issues de croisements contrôlés avec des parents aux arômes contrastés. Une étude de cette expression à différents stades de maturité de la cabosse, ainsi qu'à différents stades de fermentation des fèves, pourrait apporter de nouveaux indices pour valider l'hypothèse de l'activation des gènes via les mécanismes de défense de la plante. Une analyse d'expression par RNAseq pourrait également permettre l'identification sans à priori de gènes impliqués dans la synthèse des composés d'arômes, comme des facteurs de transcription ou d'autres, influençant la transcription des gènes. En étudiant les profils d'expression de l'ensemble des gènes, il serait également possible d'observer ceux qui s'expriment en synergie.

Une étude de la diversité fonctionnelle des gènes des différents clones de cacaoyers pourrait également être réalisée, montrant ainsi quels sont les gènes candidats dont l'expression varie en fonction des génotypes. Une étude de détection de SNP dans les gènes candidats pourrait également être réalisée afin de voir si des mutations non silencieuses sont présentes, engendrant ainsi une différence fonctionnelle au niveau de la protéine. Enfin, une étude de la diversité allélique des gènes candidats dans les collections de cacaoyers pourrait être faite. Cette étude permettrait alors de caractériser cette variation allélique afin de pouvoir la valoriser dans les différents programmes d'amélioration du cacaoyer.

Comme cela a été montré dans cette thèse, la domestication a joué un rôle important dans l'évolution des notes aromatiques du cacao Equatorien. Elle a donc probablement joué un rôle dans l'évolution des notes aromatiques de l'ensemble des génotypes actuellement cultivés dans le monde. Toujours dans le but de récolter de nouveaux indices sur les gènes impliqués dans le processus de synthèse des arômes du cacao, une étude des traces de sélection pourrait être envisagée grâce aux différentes séquences de cacaoyer actuellement disponibles. Cela permettrait de mettre en lumière les gènes candidats ayant eu un impact sur la qualité aromatique et/ou sur les défenses de la plante.

Les deux populations étudiées lors de ces travaux pourraient constituer une population de référence pour des essais de sélection génomique. Ces essais permettraient ainsi de prédire un index de sélection génomique pour les arômes, pour d'autres individus qui ne possèdent pas de valeurs phénotypiques connues.

Grâce à des méthodes associant prédictions ciblées sur certains gènes candidats et sélection génomique, il serait alors possible de prédire les notes aromatiques des cacaoyers et ainsi d'intégrer plus facilement des critères de qualité aux schémas de sélection.

La compréhension des processus de synthèse des arômes du cacao et de leur diversité génétique peut ainsi permettre d'optimiser les processus de sélection afin de produire de nouvelles variétés de cacaoyers d'une qualité aromatique unique tout en sélectionnant des arbres productifs et résistants, et de développer ainsi des marchés de niche.

Valorisation des travaux de thèse

➢ Publication dans des revues à facteurs d'impacts

Publié et disponible en ligne :

Colonges K, Jimenez JC, Saltos A, Seguine E, Loor Solorzano RG, Fouet O, Argout X, Assemat S, Davrieux F, Cros E, Boulanger R, Lanaud C. (2021) Two main biosynthesis pathways involved in the synthesis of the floral aroma of the Nacional cocoa variety. Front Plant Sci 12: 2064. https://www.frontiersin.org/article/10.3389/fpls.2021.681979 (Annexe 40)

Colonges K., Jimenez J.-C., Saltos A., Seguin E., Loor Solorzano R. G., Fouet O., Argout X., Assemat S., Davrieux F., Cros E., Lanaud C., Boulanger R. (2021). Genetic determinism of fruity aroma in Nacional cocoa variety. Plant physiology and biochemistry (Annexe 41).

Soumis :

Colonges K., Loor Solorzano R. G., Jimenez J.-C., Lahon M-C., Seguin E., Caledron D., Subia C., Sotomayor I., Fernandez F., Lebrun M., Fouet O., Rhoné B., Argout X., Costet P., Lanaud C., Boulanger R. (2021). Variability and genetic determinants of cocoa aromas in trees native to South Ecuadorian Amazonia. Plants, People, Planet.

Colonges K., Seguin E., Saltos A., Davrieux F., Minier J., Jimenez J.-C., Lahon M-C., Caledron D., Subia C., Sotomayor I., Fernandez F., Lebrun M., Fouet O., Rhoné B., Argout X., Costet P., Lanaud C., Boulanger R., Loor Solorzano R. G. (2021) Diversity and determinants of bitterness, astringency and fat content in cultivated Nacional and native Amazonian *T. cacao* accessions from Ecuador. The plant genome.

En attente de soumission :

Colonges K., Loor Solorzano R. G., Jimenez J.-C., Saltos A., Seguin E., Fouet O., Argout X., Assemat S., Davrieux F., Cros E., Boulanger R., Lanaud C. (2021). Flavours of the modern Nacional variety are shaped by cocoa domestication history. Nature communication.

➢ Présentation orale lors d'un congrès international :

Colonges K., Jimenez JC, Saltos A, Seguine E, Solorzano RL, Fouet O, Argout X, Assemat S, Davrieux F, Morillo E, Boulanger R., Cros E., Lanaud C. (2021) Genetic bases of fruity notes (fresh and dried) of the Nacional cocoa variety. E Guichard JL Quéré Eds Proc 16th Weurman Flavour Res Symp 2021. doi: 10.5281/zenodo.5046157(Annexe 42)

➢ Présentations orales lors de séminaires :

Colonges K., Jimenez J.-C., Saltos A., Seguin E., Loor Solorzano R. G., Fouet O., Argout X., Assemat S., Morillo E., Cros E., Boulanger R., Lanaud C. (2018). Étude génétique et biochimique des arômes de cacaos fins de la variété Nacional. *1ère édition des Agapiades*. *UMR AGAP*, Montpellier, France.

Colonges K., Jimenez J.-C., Saltos A., Seguin E., Loor Solorzano R. G., Fouet O., Argout X., Assemat S., Davrieux F., Cros E., Boulanger R., Lanaud C. (2018). Genetic and genomic determinants of quality traits. *First Workshop AMAZCACAO*, CIRAD, Montpellier, France.

Colonges K., Lanaud C., Boulanger R. (2018). Gaz - Chromatography Olfactometry (GCO). *First Workshop AMAZCACAO*, CIRAD, Montpellier, France.

Colonges K., Jimenez J.-C., Saltos A., Seguin E., Loor Solorzano R. G., Fouet O., Argout X., Assemat S., Cros E, Boulanger R., Lanaud C. (2019) Étude génétique et biochimique des arômes de cacaos fins d'Équateur. *2ème édition des Agapiades. UMR AGAP*, Montpellier, France.

Colonges K., Lanaud C., Boulanger R. (2019) Étude génétique et biochimique des arômes de cacaos fins d'Équateur. *PhD PUB* (vulgarisation des résultats de thèse). Montpellier, France.

Colonges K., Jimenez J.-C., Calderon D., Soubia C., Lanaud C., Boulanger R. (2019). Micro-fermentations' review. *2sd Workshop AMAZCACAO*. CIRAD, Montpellier, France.

Colonges K., Lahon M.C., Lebrun M., Alary K., Lanaud C., Boulanger R. (2020) Results of biochemical analyses carried out and diversity of volatile compounds in the Amazonian population. *Third Workshop AMAZCACAO*. CIRAD, Montpellier, France.

Bibliographie

Bibliographie

Abbas CA (2006) Production of Antioxidants, Aromas, Colours, Flavours, and Vitamins by Yeasts. Yeasts Food Beverages. Springer, Berlin, Heidelberg, pp 285–334

Afoakwa EO, Paterson A, Fowler M, Ryan A (2008) Flavor Formation and Character in Cocoa and Chocolate: A Critical Review. Crit Rev Food Sci Nutr **48**: 840–857

Akaza JM, Kouassi AB, Akaffou DS, Fouet O, N'guetta AS-P, Lanaud C (2016) Mapping QTLs for Black pod (Phytophthora palmivora) resistance in three hybrid progenies of cocoa (Theobroma cacao L.) using SSR markers. Int. J. Sci. Res. Publ. Volume 6, Issue 1, January 2016 Edition:

Almeida DSM de, Amaral DOJ do, Del-Bem L-E, Santos EB dos, Silva RJS, Gramacho KP, Vincentz M, Micheli F (2017) Genome-wide identification and characterization of cacao WRKY transcription factors and analysis of their expression in response to witches' broom disease. PLOS ONE **12**: e0187346

Álvarez C, Pérez E, Cros E, Lares M, Assemat S, Boulanger R, Davrieux F (2012) The Use of near Infrared Spectroscopy to Determine the Fat, Caffeine, Theobromine and (−)-Epicatechin Contents in Unfermented and Sun-Dried Beans of Criollo Cocoa. J Infrared Spectrosc **20**: 307–315

Amano I, Kitajima S, Suzuki H, Koeduka T, Shitan N (2018) Transcriptome analysis of Petunia axillaris flowers reveals genes involved in morphological differentiation and metabolite transport. PLOS ONE **13**: e0198936

Amores F, Butler D, Ramos G, Sukha D, Espin S, Gomez A, Zambrano A, Hollywood N, Van Loo R, Seguine E (2007) Study of the Chemical, Physical and Organoleptic Parameters to Establish the Difference Between Fine and Bulk Cocoa. 18

Andújar I, Recio MC, Giner RM, Ríos JL (2012) Cocoa Polyphenols and Their Potential Benefits for Human Health. Oxid Med Cell Longev **2012**: e906252

Aprotosoaie AC, Luca SV, Miron A (2016) Flavor Chemistry of Cocoa and Cocoa Products—An Overview. Compr Rev Food Sci Food Saf **15**: 73–91

Aranda DAG, Santos RTP, Tapanes NCO, Ramos ALD, Antunes OAC (2008) Acid-Catalyzed Homogeneous Esterification Reaction for Biodiesel Production from Palm Fatty Acids. Catal Lett **122**: 20–25

Araújo IS, de Souza Filho GA, Pereira MG, Faleiro FG, de Queiroz VT, Guimarães CT, Moreira MA, de Barros EG, Machado RCR, Pires JL, et al (2009) Mapping of Quantitative Trait Loci for Butter Content and Hardness in Cocoa Beans (Theobroma cacao L.). Plant Mol Biol Report **27**: 177–183

Argout X, Martin G, Droc G, Fouet O, Labadie K, Rivals E, Aury JM, Lanaud C (2017) The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. BMC Genomics **18**: 730

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al (2011) The genome of *Theobroma cacao*. Nat Genet **43**: 101

Arn H, Acree TE (1998) Flavornet: A database of aroma compounds based on odor potency in natural products. Dev Food Sci **40**: 27

Arnoldi A, Arnoldi C, Baldi O, Griffini A (1988) Flavor components in the Maillard reaction of different amino acids with fructose in cocoa butter-water Qualitative and quantitative analysis of pyrazines. J Agric Food Chem **36**: 988--992

Ascrizzi R, Flamini G, Tessieri C, Pistelli L (2017) From the raw seed to chocolate: Volatile profile of Blanco de Criollo in different phases of the processing chain. Microchem J **C**: 474–479

Ashihara H, Monteiro AM, Gillies FM, Crozier A (1996) Biosynthesis of Caffeine in Leaves of Coffee. Plant Physiol **111**: 747–753

Assemat S, Lachenaud P, Ribeyre F, Davrieux F, Pradon JL, Cros E (2005) Bean quality traits and sensory evaluation of wild Guianan cocoa populations (Theobroma cacao L.). Genet Resour Crop Evol **52**: 911–917

Assi-Clair BJ, Koné MK, Kouamé K, Lahon MC, Berthiot L, Durand N, Lebrun M, Julien-Ortiz A, Maraval I, Boulanger R, et al (2019) Effect of aroma potential of Saccharomyces cerevisiae fermentation on the volatile profile of raw cocoa and sensory attributes of chocolate produced thereof. Eur Food Res Technol **245**: 1459–1471

Aznar M, López R, Cacho JF, Ferreira V (2001) Identification and Quantification of Impact Odorants of Aged Red Wines from Rioja. GC−Olfactometry, Quantitative GC-MS, and Odor Evaluation of HPLC Fractions. J Agric Food Chem **49**: 2924–2929

Babacauh KD (1983) Structure des populations de *Phytophthora palmivora* (Butl.) Butl. emend. Bras. et Griff. parasite du Cacaoyer (*Theobroma cacao* L.). Bull Société Bot Fr Lett Bot **130**: 15–25

Baek HH, Cadwallader KR, Marroquin E, Silva JL (1997) Identification of Predominant Aroma Compounds in Muscadine Grape Juice. J Food Sci **62**: 249–252

Baldwin IT (2010) Plant volatiles. Curr Biol **20**: R392–R397

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**: 263–265

Bartley BGD (2005) The genetic diversity of cacao and its utilization. CABI Publishing

Bayer C, Kubitzki K (2003) Malvaceae. Fam. Genera Vasc. Plants Dicotyledons Malvales Capparales Non-Betalain Caryophyllales. Berlin: Springer, p .225-311

Beckett ST, Fowler M, Ziegler GR (2017) Beckett's Industrial Chocolate Manufacture and Use, 5th Edition, Wiley-Blackwell.

Beckett ST, Hancock BL, Krüger Ch, Reimerdes EH, Kleinert J (1994) Industrial Chocolate Manufacture and Use, Second Edition. doi: 10.1007/978-1-4615-2111-2

Biehl B, Passern D, Sagemann W (1982) Effect of acetic acid on subcellular structures of cocoa bean cotyledons. J Sci Food Agric **33**: 1101–1109

Bohlmann J, Meyer-Gauen G, Croteau R (1998) Plant terpenoid synthases: Molecular biology and phylogenetic analysis. Proc Natl Acad Sci **95**: 4126–4133

Bond TJ (2011) The Origins of Tea, Coffee and Cocoa as Beverages. Teas Cocoa Coffee. John Wiley & Sons, Ltd, pp 1–24

Bouharmont J (1960) Recherches cytologiques sur la fructification et l'incompatibilité chez Theobroma cacao L. pp 20–89

Boza EJ, Motamayor JC, Amores FM, Cedeño-Amador S, Tondo CL, Livingstone DS, Schnell RJ, Gutiérrez OA (2014) Genetic Characterization of the Cacao Cultivar CCN 51: Its Impact and Significance on Global Cacao Improvement and Production. J Am Soc Hortic Sci **139**: 219–229

Brasier CM, Griffin MJ (1979) Taxonomy of *Phytophthora palmivora* on cocoa. Trans Br Mycol Soc **72**: 111–143

Carvalho BT de, Holt S, Souffriau B, Brandão RL, Foulquié-Moreno MR, Thevelein JM (2017) Identification of Novel Alleles Conferring Superior Production of Rose Flavor Phenylethyl Acetate Using Polygenic Analysis in Yeast. mBio. doi: 10.1128/mBio.01173-17

Causse M, Saliba-Colombani V, Lecomte L, Duffé P, Rousselle P, Buret M (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J Exp Bot **53**: 2089–2098

Cevallos-Cevallos JM, Gysel L, Maridueña-Zavala MG, Molina-Miranda MJ (2018) Time-Related Changes in Volatile Compounds during Fermentation of Bulk and Fine-Flavor Cocoa (*Theobroma cacao*) Beans. J Food Qual **2018**: 1758381

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

Chen X, Yauk Y-K, Nieuwenhuizen NJ, Matich AJ, Wang MY, Perez RL, Atkinson RG, Beuning LL (2010) Characterisation of an (S)-linalool synthase from kiwifruit (*Actinidia arguta*) that catalyses the first committed step in the production of floral lilac compounds. Funct Plant Biol **37**: 232–243

Chen X-M, Kobayashi H, Sakai M, Hirata H, Asai T, Ohnishi T, Baldermann S, Watanabe N (2011) Functional characterization of rose phenylacetaldehyde reductase (PAR), an enzyme involved in the biosynthesis of the scent compound 2-phenylethanol. J Plant Physiol **168**: 88–95

Chouhan S, Sharma K, Zha J, Guleria S, Koffas MAG (2017) Recent Advances in the Recombinant Biosynthesis of Polyphenols. Front Microbiol **8**: 2259

Cilas C (2020) Cacao - Contexte et enjeux | Cirad. https://www.cirad.fr/nos-activites-notre-impact/filieres-agricoles-tropicales/cacao/contexte-et-enjeux

Cirad (1999) Les mondes du cacao.

Colahan-Sederstrom PM, Peterson DG (2005) Inhibition of Key Aroma Compound Generated during Ultrahigh-Temperature Processing of Bovine Milk via Epicatechin Addition. J Agric Food Chem **53**: 398–402

Colonges K, Jimenez JC, Lahon M-C, Seguine E, Calderon D, Subia C, Sotomayor I, Fernández F, Loor Solorzano RG, Lebrun M, et al (2021a) Variability and genetic determinants of native cocoa trees aromas from South Ecuadorian Amazonia. Draft - New Phytol.

Colonges K, Jimenez JC, Saltos A, Seguine E, Loor Solorzano RG, Fouet O, Argout X, Assemat S, Davrieux F, Cros E, et al (2021b) Two main biosynthesis pathways involved in the synthesis of the floral aroma of the Nacional cocoa variety. Front Plant Sci **12**: 2064

Colonges K, Jimenez JC, Saltos A, Seguine E, Loor Solorzano RG, Fouet O, Argout X, Assemat S, Davrieux F, Cros E, et al (2021c) Integration of GWAS, metabolomics, and sensorial analyses to reveal novel metabolic pathways involved in cocoa fruity aroma. Plant Physiol. Biochem. Accepted:

Colonges K, Jimenez JC, Saltos A, Seguine E, Solorzano RL, Fouet O, Argout X, Assemat S, Davrieux F, Morillo E, et al (2021d) Genetic bases of fruity notes (fresh and dried) of the Nacional cocoa variety. E Guichard JL Quéré Eds Proc 16th Weurman Flavour Res Symp 2021. doi: 10.5281/zenodo.5046157

Cook LR, Meursing EH (1982) Chocolate production and use. Harcourt Brace Jovanovich, New York

Cooper KA, Donovan JL, Waterhouse AL, Williamson G (2008) Cocoa and health: a decade of research. Br J Nutr **99**: 1–11

Cope FW (1939) Studies in the mechanism of self-incompatibility in cacao I. 8th Annu Rep Cocoa Res **Trinidad**: 20–21

Cope FW (1940) Studies in the mechanism of self-incompatibility in cacao II. 9th Annu Rep Cocoa Res **Trinidad**: 19–23

Cope FW (1958) Incompatibility in Theobroma cacao. Nature

Cope FW (1962) The mechanism of pollen incompatibility in *Theobroma cacao* L. Heredity **17**: 157

Counet C, Callemien D, Ouwerx C, Collin S (2002) Use of gas chromatography-olfactometry to identify key odorant compounds in dark chocolate. Comparison of samples before and after conching. J Agric Food Chem **50**: 2385–91

Cros E, Jeanjean N (1995) Qualité du cacao : influence de la fermentation et du séchage. Plant Rech Dev **2**: 21–27

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

Cseke L, Dudareva N, Pichersky E (1998) Structure and evolution of linalool synthase. Mol Biol Evol **15**: 1491–1498

Cuatrecasas J (1964) Cacao and Its Allies, a Taxonomic Revision of the Genus *Theobroma*. Contr NAt Herb **35**: 379–614

Davrieux F, Boulanger R, Assemat S, Portillo E, Alvarez C, Sukha D, Cros E (2007) Determination of biochemistry composition of cocoa powder using near infrared spectroscopy. Sfc Ed Proc Euro Food Chem XIV Food Qual Issue Mol Based Sci Paris 29-31 August 2007 463–466

Denancé N, Sánchez-Vallet A, Goffner D, Molina A (2013) Disease resistance or growth: the role of plant hormones in balancing immune responses and fitness costs. Front Plant Sci. doi: 10.3389/fpls.2013.00155

Dhonsi D, Stapley AGF (2006) The effect of shear rate, temperature, sugar and emulsifier on the tempering of cocoa butter. J Food Eng **77**: 936–942

Dice LR (1945) Measures of the Amount of Ecologic Association Between Species. Ecology **26**: 297–302

Dickschat JS, Wickel S, Bolten CJ, Nawrath T, Schulz S, Wittmann C (2010) Pyrazine Biosynthesis in Corynebacterium glutamicum. Eur J Org Chem **2010**: 2687–2695

van Dijk PJ, Ellis THN (2016) The Full Breadth of Mendel's Genetics. Genetics **204**: 1327–1336

Dong F, Yang Z, Baldermann S, Kajitani Y, Ota S, Kasuga H, Imazeki Y, Ohnishi T, Watanabe N (2012) Characterization of l-phenylalanine metabolism to acetophenone and 1-phenylethanol in the flowers of *Camellia sinensis*using stable isotope labeling. J Plant Physiol **169**: 217–225

Dudareva N, Cseke L, Blanc VM, Pichersky E (1996) Evolution of floral scent in Clarkia: novel patterns of S-linalool synthase gene expression in the *C. breweri* flower. Plant Cell **8**: 1137–1148

Eckardt NA (2000) Sequencing the Rice Genome. Plant Cell **12**: 2011–2017

Edoh Adabe K, Ngo-Samnick EL (2014) Production et transformation du cacao. Collect. - AGRO

Eduardo I, Chietera G, Pirona R, Pacheco I, Troggio M, Banchi E, Bassi D, Rossini L, Vecchietti A, Pozzi C (2013) Genetic dissection of aroma volatile compounds from the essential oil of peach fruit: QTL analysis and identification of candidate genes using dense SNP maps. Tree Genet Genomes **9**: 189–204

Efombagn MIB, Bieysse D, Nyassé S, Eskes AB (2011) Selection for resistance to Phytophthora pod rot of cocoa (*Theobroma cacao L.*) in Cameroon: Repeatability and reliability of screening tests and field observations. Crop Prot **30**: 105–110

Enriquez G (1992) Characteristics of cacao "Nacional" of Ecuador. International Workshop on Conservation, Characterisation and Utilisation of Cocoa Genetic Resources in the 21st century. Port of Spain, Trinidad 13–17th September. The Cocoa Research Unit, The University of the West Indies. 269–278

Enriquez G, Alarcòn E (1977) The nature of self-incompatibility. Rev CATIE 1–22

Essola Etoa LC (2014) Evaluation des rendements potentiels en cacao (*Theobroma cacao L.*) dans les systèmes agroforestiers complexes en zones forestière à pluviométrie bimodale du centre du Cameroun. Mémoire présenté en requis partiel pour l'obtention du diplôme d'Ingénieur Agronome option Productions Végétales. Univ. Dschang

Evans HC (1978) Witches' broom disease of cocoa (*Crinipellis perniciosa*) in Ecuador. Ann Appl Biol **89**: 185–192

Farhi M, Lavie O, Masci T, Hendel-Rahmanim K, Weiss D, Abeliovich H, Vainstein A (2010) Identification of rose phenylacetaldehyde synthase by functional complementation in yeast. Plant Mol Biol **72**: 235–245

Feng L, Chen C, Li T, Wang M, Tao J, Zhao D, Sheng L (2014) Flowery odor formation revealed by differential expression of monoterpene biosynthetic genes and monoterpene accumulation in rose (*Rosa rugosa Thunb.*). Plant Physiol Biochem **75**: 80–88

Ferreira V, López R, Escudero A, Cacho JF (1997) The aroma of Grenache red wine: hierarchy and nature of its main odorants. J Sci Food Agric **77**: 259–267

Fickert B, Schieberle P (1998) Identification of the key odorants in barley malt (caramalt) using GC/MS techniques and odour dilution analyses. Food Nahr **42**: 371–375

Flament M-H, Kebe I, Clément D, Pieretti I, Risterucci A-M, N'Goran J-A-K, Cilas C, Despréaux D, Lanaud C (2001) Genetic mapping of resistance factors to *Phytophthora palmivora* in cocoa. Genome **44**: 79–85

Fouet et al (Unpublished data) Etude de la diversité génétique des cacaoyers issus de prospections en Amazonie Equatorienne.

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol Evol **6**: 925–929

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics **196**: 973–983

Gallais A (2013) Chapitre 5 : Cartographie génétique et sélection assistée par marqueurs. Domest. À Transgénèse Evol. Outils Pour Amélioration Plantes, Quae. pp 113–128

Gallais A (2015) Comprendre l'amélioration des plantes. Enjeux, méthodes, objectifs et critères de sélection., Quae.

Gao F, Liu B, Li M, Gao X, Fang Q, Liu C, Ding H, Wang L, Gao X (2018) Identification and characterization of terpene synthase genes accounting for volatile terpene emissions in flowers of *Freesia x hybrida* . J Exp Bot **69**: 4249–4265

Gao X, Becker LC, Becker DM, Starmer JD, Province MA (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol **34**: 100–105

Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol **32**: 361–369

Garg N, Sethupathy A, Tuwani R, NK R, Dokania S, Iyer A, Gupta A, Agrawal S, Singh N, Shukla S, et al (2018) FlavorDB: a database of flavor molecules. Nucleic Acids Res **46**: D1210–D1216

Genovese A, Gambuti A, Piombino P, Moio L (2007) Sensory properties and aroma compounds of sweet Fiano wine. Food Chem **103**: 1228–1236

Glendinning DR (1972) Natural pollination of cocoa. New Phytol **71**: 719–729

Glendinning DR (1967) Incompatibility alleles of cocoa. Nature 306

Griffith GW, Hedger JN (1993) The breeding biology of biotypes of the witches' broom pathogen of cocoa, *Crinipellis perniciosa*. Heredity **72**: 278–289

Griffith GW, Nicholson J, Nenninger A, Birch RN, Hedger JN (2003) Witches' brooms and frosty pods: Two major pathogens of cacao. N Z J Bot **41**: 423–435

Guichard H, Lemesle S, Ledauphin J, Barillier D, Picoche B (2003) Chemical and Sensorial Aroma Characterization of Freshly Distilled Calvados. 1. Evaluation of Quality and Defects on the Basis of Key Odorants by Olfactometry and Sensory Analysis. J Agric Food Chem **51**: 424–432

Guterman I, Masci T, Chen X, Negre F, Pichersky E, Dudareva N, Weiss D, Vainstein A (2006) Generation of phenylpropanoid pathway-derived volatiles in transgenic plants: rose alcohol acetyltransferase produces phenylethyl acetate and benzyl acetate in petunia flowers. Plant Mol Biol **60**: 555–563

Gutiérrez OA, Puig AS, Phillips-Mora W, Bailey BA, Ali SS, Mockaitis K, Schnell RJ, Livingstone D, Mustiga G, Royaert S, et al (2021) SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of *Theobroma cacao L.* Tree Genet Genomes **17**: 28

Hadj Salem F, Lebrun M, Mestres C, Sieczkowski N, Boulanger R, Collignan A (2020) Transfer kinetics of labeled aroma compounds from liquid media into coffee beans during simulated wet processing conditions. Food Chem **322**: 126779

Hamdouche Y, Meile JC, Lebrun M, Guehi T, Boulanger R, Teyssier C, Montet D (2019) Impact of turning, pod storage and fermentation time on microbial ecology and volatile composition of cocoa beans. Food Res Int Ott Ont **119**: 477–491

Hamon P, Seguin M, Perrier X, Glaszmann J-C (1999) Diversité génétique des plantes tropicales cultivées. Cirad

Hao R, du D, Wang T, Yang W, Wang J, Zhang Q (2014) A comparative analysis of characteristic floral scent compounds in Prunus mume and related species. Biosci Biotechnol Biochem **78**: 1640–1647

Helsper JPFG, Davies JA, Bouwmeester HJ, Krol AF, Kampen MH van (1998) Circadian rhythmicity in emission of volatile compounds by flowers of *Rosa hybrida* L. cv. Honesty. Planta **207**: 88–95

Henry-Kirk RA, McGhie TK, Andre CM, Hellens RP, Allan AC (2012) Transcriptional analysis of apple fruit proanthocyanidin biosynthesis. J Exp Bot **63**: 5437–5450

Ho VTT, Zhao J, Fleet G (2014) Yeasts are essential for cocoa bean fermentation. Int J Food Microbiol **174**: 72–87

Högnadóttir Á, Rouseff RL (2003) Identification of aroma active compounds in orange essence oil using gas chromatography–olfactometry and gas chromatography–mass spectrometry. J Chromatogr A **998**: 201–211

Holmquist M (2000) Alpha Beta-Hydrolase Fold Enzymes Structures, Functions and Mechanisms. Curr Protein Pept Sci **1**: 209–235

Hue C, Gunata Z, Bergounhou A, Assemat S, Boulanger R, Sauvage FX, Davrieux F (2014) Near infrared spectroscopy as a new tool to determine cocoa fermentation levels through ammonia nitrogen quantification. Food Chem **148**: 240–245

ICCO (2020) Production of cocoa beans (thousand tonnes) year 2019/2020. Q. Bull. Cocoa Stat. XLVI:

International Cocoa Organization (2017) Fine or flavour cocoa- What is Fine or Flavour Cocoa?

Ioannone F, Di Mattia CD, De Gregorio M, Sergi M, Serafini M, Sacchetti G (2015) Flavanols, proanthocyanidins and antioxidant activity changes during cocoa (*Theobroma cacao* L.) roasting as affected by temperature and time of processing. Food Chem **174**: 256–262

IRGSP IRGS, Sasaki T (2005) The map-based sequence of the rice genome. Nature **436**: 793–800

ISCQF (2020) First draft of the Protocol for Cocoa Liquor Sensory Evaluation: part of the International Standards for the Assessment of Cocoa Quality and Flavour (ISCQF). Compiled by Bioversity International, in collaboration with the members of the ISCQF Working Group. doi: ISBN: 978-92-9255-158-2

Ito Y, Sugimoto A, Kakuda T, Kubota K (2002) Identification of Potent Odorants in Chinese Jasmine Green Tea Scented with Flowers of *Jasminum sambac*. J Agric Food Chem **50**: 4878–4884

Jezussek M, Juliano BO, Schieberle P (2002) Comparison of Key Aroma Compounds in Cooked Brown Rice Varieties Based on Aroma Extract Dilution Analyses. J Agric Food Chem **50**: 1101–1105

Jinap S, Ikrawan Y, Bakar J, Saari N, Lioe HN (2008) Aroma precursors and methylpyrazines in underfermented cocoa beans induced by endogenous carboxypeptidase. J Food Sci **73**: H141-7

Jinap S, Rosli WIW, Russly AR, Nordin LM (1998) Effect of roasting time and temperature on volatile component profiles during nib roasting of cocoa beans (Theobroma cacao). J---Sci--Food--Agric **77 (4)**: 441–448

Jumelle H (1900) Le cacaoyer sa culture et son exploitation dans tous les pays de production. Bibl. Numér. Manioc SCD Univ. Antill.

Kadow D, Bohlmann J, Phillips W, Lieberei R (2013) Identification of main fine flavour components in two genotypes of the chocolate tree (Theobroma cacao L). J Appl Bot Food Qual Bot **86**: 90--98

Kallithraka S, Bakker J, Clifford MN (1997) Evaluation of Bitterness and Astringency of (+)-Catechin and (-)-Epicatechin in Red Wine and in Model Solution. J Sens Stud **12**: 25–37

Kalua CM, Allen MS, Bedgood DR, Bishop AG, Prenzler PD, Robards K (2007) Olive oil volatile compounds, flavour development and quality: A critical review. Food Chem **100**: 273–286

Kaminaga Y, Schnepp J, Peel G, Kish CM, Ben-Nissan G, Weiss D, Orlova I, Lavie O, Rhodes D, Wood K, et al (2006) Plant phenylacetaldehyde synthase is a bifunctional homotetrameric enzyme that catalyzes phenylalanine decarboxylation and oxidation. J Biol Chem **281**: 23357–23366

Karagül-Yüceer Y, Drake MA, Cadwallader KR (2006) Aroma-active Components of Liquid Cheddar Whey. J Food Sci **68**: 1215–1219

Karagül-Yüceer Y, Vlahovich KN, Drake M, Cadwallader KR (2003) Characteristic Aroma Components of Rennet Casein. J Agric Food Chem **51**: 6797–6801

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, et al (2012) Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms. Data Prod Anal Popul Genomics Methods Protoc 67–89

Knight R, Rogers HH (1955) Incompatibility in *Theobroma cacao*. Heredity **9**: 69

Kongor JE, Hinneh M, de Walle DV, Afoakwa EO, Boeckx P, Dewettinck K (2016) Factors influencing quality variation in cocoa (*Theobroma cacao*) bean flavour profile — A review. Food Res Int **82**: 44–52

Koshiishi C, Kato A, Yama S, Crozier A, Ashihara H (2001) A new caffeine biosynthetic pathway in tea leaves: utilisation of adenosine released from the S-adenosyl-L-methionine cycle. FEBS Lett **499**: 50–54

Kreck M, Püschel S, Wüst M, Mosandl A (2003) Biogenetic Studies in *Syringa vulgaris* L.: Synthesis and Bioconversion of Deuterium-Labeled Precursors into Lilac Aldehydes and Lilac Alcohols. J Agric Food Chem **51**: 463–469

Kücükgöze G, Leimkühler S (2018) Direct comparison of the four aldehyde oxidase enzymes present in mouse gives insight into their substrate specificities. PLoS ONE. doi: 10.1371/journal.pone.0191819

Kumazawa K, Masuda H (2002) Identification of Potent Odorants in Different Green Tea Varieties Using Flavor Dilution Technique. J Agric Food Chem **50**: 5660–5663

Lachenaud P, Motamayor JC (2017) The Criollo cacao tree (*Theobroma cacao* L.): a review. Genet Resour Crop Evol **64**: 1807–1820

Lachenaud P, Sounigo O, Clément D (2005) The compatibility-yield efficiency relationship. 13–16

Lamarti A, Badoc A, Deffieux G, Carde J-P (1994) Biogénèse des monoterpènes. II - La chaîne isoprénique.

Lanaud C, Boult E, Clapperton J, N'Goran J, Cros E, Chapelin M, Petithuguenin P (2003) Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma cacao* L. 14th Int Cocoa Res Conf 1119–1126

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

Lanaud C, Fouet O, Legavre T, Lopes U, Sounigo O, Eyango MC, Mermaz B, Da Silva MR, Loor Solorzano RG, Argout X, et al (2017) Deciphering the *Theobroma cacao* self-incompatibility system: from genomics to diagnostic markers for self-compatibility. J Exp Bot **68**: 4775–4790

Lanaud C, Hamon P, Duperray C (1992) Estimation of nuclear DNA content of Theobrama cacao L. by flow cytometry. Café Cacao Thé 36 3–8

Lanaud C, Saltos A, Jimenez JC, Lemainque A, Pavek S, Argout X, Fouet O, Seguine E, Assemat S, Davrieux F, et al (2012) Adding value to *T. cacao* germplasm collections combining GWAS and genome sequence analysis. Plant Anim Genome XX Conf W118.

Langenheim JH (1994) Higher plant terpenoids: A phytocentric overview of their ecological roles. J Chem Ecol **20**: 1223–1280

Lapadatescu C, Giniès C, Le Quéré JL, Bonnarme P (2000) Novel scheme for biosynthesis of aryl metabolites from L-phenylalanine in the fungus *Bjerkandera adusta*. Appl Environ Microbiol **66**: 1517–1522

Larsen M, Poll L (1992) Odour thresholds of some important aroma compounds in strawberries. Z Für Lebensm-Unters Forsch **195**: 120–123

Laurent V, Risterucci AM, Lanaud C (1993) Chloroplast and mitoehondrial DNA diversity in *Theobroma cacao*. Theor Appl Genet 8

Lerceteau E, Robert T, Pétiard V, Crouzillat D (1997) Evaluation of the extent of genetic variability among *Theobroma cacao* accessions using RAPD and RFLP markers. Theor Appl Genet **95**: 10–19

Lesschaeve I, Noble AC (2005) Polyphenols: factors influencing their sensory properties and their effects on food and beverage preferences. Am J Clin Nutr **81**: 330S-335S

Li Z, Traore A, Maximova S, Guiltinan MJ (1998) Somatic embryogenesis and plant regeneration from floral explants of cacao (*Theobroma cacao* L.) using thidiazuron. Vitro Cell Dev Biol - Plant **34**: 293–299

Liu C, Zhang K, Cao W, Zhang G, Chen G, Yang H, Wang Q, Liu H, Xian M, Zhang H (2018) Genome mining of 2-phenylethanol biosynthetic genes from Enterobacter sp. CGMCC 5087 and heterologous overproduction in *Escherichia coli*. Biotechnol Biofuels **11**: 305

Loor S. RG (1998) Obtención de híbridos de cacao tipo nacional provenientes de materiales de alta productividad y resistente a enfermedades.

Loor S. RG (2007) Contribution à l'étude de la domestication de la variété de cacaoyer Nacional d'Equateur : recherche de la variété native et de ses ancêtres sauvages. doi: Corpus ID: 130056754

Loor S. RG, Lachenaud P, Fouet O, Argout X, Peña G, Castro Macias J, Amores Puyutaxi FM, Valdez F, Hurtado J, Lanaud C (2016) Rescue of cacao genetic resources related to the nacional variety : surveys in the ecuadorian Amazon (2010-2013). Rev ESPAMCiencia **6**: 12–14

Loor S. RG, Risterucci AM, Courtois B, Fouet O, Jeanneau M, Rosenquist E, Amores F, Vasco A, Medina M, Lanaud C (2009) Tracing the native ancestors of the modern *Theobroma cacao* L. population in Ecuador. Tree Genet Genomes **5**: 421–433

Loor S. RGL, Fouet O, Lemainque A, Pavek S, Boccara M, Argout X, Amores F, Courtois B, Risterucci AM, Lanaud C (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

Loor S. RGL, Lachenaud P, Fouet O, Argout X, Peña G, Macias JC, Puyutaxi FMA, Valdez F, Hurtado J, Lanaud C (2015) Rescue of cacao genetic resources related to the nacional variety: surveys in the Ecuadorian Amazon (2010-2013). Rev ESPAMCIENCIA ISSN 1390-8103 **6**: 7–15

Luna F, Crouzillat D, Cirou L, Bucheli P (2002) Chemical composition and flavor of Ecuadorian cocoa liquor. J Agric Food Chem **50**: 3527–3532

Ma J-Q, Jin J-Q, Yao M-Z, Ma C-L, Xu Y-X, Hao W-J, Chen L (2018) Quantitative Trait Loci Mapping for Theobromine and Caffeine Contents in Tea Plant (Camellia sinensis). J Agric Food Chem **66**: 13321–13327

Ma W, Guo A, Zhang Y, Wang H, Liu Y, Li H (2014) A review on astringency and bitterness perception of tannins in wine. Trends Food Sci Technol **40**: 6–19

Magnard J-L, Roccia A, Caissard J-C, Vergne P, Sun P, Hecquet R, Dubois A, Hibrand-Saint Oyant L, Jullien F, Nicolè F, et al (2015) Biosynthesis of monoterpene scent compounds in roses. Science **349**: 81–83

Mahajan SS, Goddik L, Qian MC (2004) Aroma Compounds in Sweet Whey Powder. J Dairy Sci **87**: 4057–4063

Mäki-Arvela P, Sahin S, Kumar N, Heikkilä T, Lehto V-P, Salmi T, Murzin DYu (2008) Cascade approach for synthesis of R-1-phenyl ethyl acetate from acetophenone: Effect of support. J Mol Catal Chem **285**: 132–141

Marcano M, Morales S, Hoyer MT, Courtois B, Risterucci AM, Fouet O, Pugh T, Cros E, Gonzalez V, Dagert M, 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

Martin D, Tholl D, Gershenzon J, Bohlmann J (2002) Methyl Jasmonate Induces Traumatic Resin Ducts, Terpenoid Resin Biosynthesis, and Terpenoid Accumulation in Developing Xylem of Norway Spruce Stems. Plant Physiol **129**: 1003–1018

Martínez-Pinilla E, Oñatibia-Astibia A, Franco R (2015) The relevance of theobromine for the beneficial effects of cocoa consumption. Front Pharmacol **6**: 30

Martins SIFS, Jongen WMF, van Boekel MAJS (2000) A review of Maillard reaction in food and implications to kinetic modelling. Trends Food Sci Technol **11**: 364–373

Mateo JJ, Jiménez M (2000) Monoterpenes in grape juice and wines. J Chromatogr A **881**: 557–567

Mathieu S, Cin VD, Fei Z, Li H, Bliss P, Taylor MG, Klee HJ, Tieman DM (2009) Flavour compounds in tomato fruits: identification of loci and potential pathways affecting volatile composition. J Exp Bot **60**: 325–337

Matsushita K, Toyama H, Adachi O (1994) Respiratory Chains and Bioenergetics of Acetic Acid Bacteria. *In* AH Rose, DW Tempest, eds, Adv. Microb. Physiol. Academic Press, pp 247–301

Meesters RJW, Duisken M, Hollender J (2007) Study on the cytochrome P450-mediated oxidative metabolism of the terpene alcohol linalool: indication of biological epoxidation. Xenobiotica Fate Foreign Compd Biol Syst **37**: 604–617

Melo ADQ, Silva FFM, Dos Santos JCS, Fernández-Lafuente R, Lemos TLG, Dias Filho FA (2017) Synthesis of Benzyl Acetate Catalyzed by Lipase Immobilized in Nontoxic Chitosan-Polyphosphate Beads. Molecules **22**: 2165

Mendel G (1865) Versuche über plflanzenhybriden. verhandlungen des naturforschenden vereines in brünn. Bd IV Für Jahr

Miller EN, Jarboe LR, Turner PC, Pharkya P, Yomano LP, York SW, Nunn D, Shanmugam KT, Ingram LO (2009) Furfural Inhibits Growth by Limiting Sulfur Assimilation in Ethanologenic Escherichia coli Strain LY180. Appl Environ Microbiol **75**: 6132–6141

Mindrebo JT, Nartey CM, Seto Y, Burkart MD, Noel JP (2016) Unveiling the functional diversity of the Alpha-Beta hydrolase fold in plants. Curr Opin Struct Biol **41**: 233–246

Misnawi, Jinap S, Jamilah B, Nazamid S (2005) Changes in polyphenol ability to produce astringency during roasting of cocoa liquor. J Sci Food Agric **85**: 917–924

Misnawi, Jinap S, Jamilah B, Nazamid S (2003) Effects of incubation and polyphenol oxidase enrichment on colour, fermentation index, procyanidins and astringency of unfermented and partly fermented cocoa beans. Int J Food Sci Technol **38**: 285–295

Misnawi, Jinap S, Jamilah B, Nazamid S (2004) Sensory properties of cocoa liquor as affected by polyphenol concentration and duration of roasting. Food Qual Prefer **15**: 403–409

Misnawi, Jinap S, Nazamid S, Jamilah B (2002) Activation of remaining key enzymes in dried under-fermented cocoa beans and its effect on aroma precursor formation. Food Chem **78**: 407–417

Miziorko HM (2011) Enzymes of the mevalonate pathway of isoprenoid biosynthesis. Arch Biochem Biophys **505**: 131–143

Monisha TR, Ismailsab M, Masarbo R, Nayak AS, Karegoudar TB (2018) Degradation of cinnamic acid by a newly isolated bacterium Stenotrophomonas sp. TRMK2. 3 Biotech. doi: 10.1007/s13205-018-1390-0

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

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, III DL, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, et al (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol **14**: r53

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. Heredity **91**: 322–330

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity **89**: 380–386

Motilal LA, Zhang D, Mischke S, Meinhardt LW, Boccara M, Fouet O, Lanaud C, Umaharan P (2016) Association mapping of seed and disease resistance traits in Theobroma cacao L. Planta **244**: 1265–1276

Mustiga GM, Morrissey J, Stack JC, DuVal A, Royaert S, Jansen J, Bizzotto C, Villela-Dias C, Mei L, Cahoon EB, et al (2019) Identification of Climate and Genetic Factors That Control Fat Content and Fatty Acid Composition of *Theobroma cacao* L. Beans. Front Plant Sci. doi: 10.3389/fpls.2019.01159

Muzykiewicz-Szymańska A, Nowak A, Wira D, Klimowicz A (2021) The Effect of Brewing Process Parameters on Antioxidant Activity and Caffeine Content in Infusions of Roasted and Unroasted Arabica Coffee Beans Originated from Different Countries. Molecules **26**: 3681

Nagegowda DA, Gutensohn M, Wilkerson CG, Dudareva N (2008) Two nearly identical terpene synthases catalyze the formation of nerolidol and linalool in snapdragon flowers. Plant J Cell Mol Biol **55**: 224–239

Nyassé S, Cilas C, Herail C, Blaha G (1995) Leaf inoculation as an early screening test for cocoa (*Theobroma cacao* L.) resistance to *Phytophthora* black pod disease. Crop Prot **14**: 657–663

Oswald M (2006) Déterminisme génétique de la biosynthèse des terpénols aromatiques chez la vigne. Strasbourg 1

Owusu M, Petersen MA, Heimdal H (2012) Effect of fermentation method, roasting and conching conditions on the aroma volatiles of dark chocolate. J Food Process Preserv **36**: 446--456

Palmqvist E, Almeida JS, Hahn-Hägerdal B (1999) Influence of furfural on anaerobic glycolytic kinetics of Saccharomyces cerevisiae in batch culture. Biotechnol Bioeng **62**: 447–454

Parent GJ, Giguère I, Mageroy M, Bohlmann J, MacKay JJ (2018) Evolution of the biosynthesis of two hydroxyacetophenones in plants. Plant Cell Environ **41**: 620–629

Peddie HAB (1990) Ester Formation in Brewery Fermentations. J Inst Brew **96**: 327–331

Perestrelo R, Fernandes A, Albuquerque FF, Marques JC, Câmara JS (2006) Analytical characterization of the aroma of Tinta Negra Mole red wine: Identification of the main odorants compounds. Anal Chim Acta **563**: 154–164

Pérez-Silva A, Odoux E, Brat P, Ribeyre F, Rodriguez-Jimenes G, Robles-Olvera V, García-Alvarado MA, Günata Z (2006) GC–MS and GC–olfactometry analysis of aroma compounds in a representative organic aroma extract from cured vanilla (Vanilla planifolia G. Jackson) beans. Food Chem **99**: 728–735

Perrier X, Jacquemoud-Collet JP (2006) DARwin software.

Petithuguenin P (Centre de CI en RA pour le D, Roche G (1995) Ecuador: the cocoa sector, results and prospects. Plant. Rech. Dev. Fr.

Pham AJ, Schilling MW, Yoon Y, Kamadia VV, Marshall DL (2008) Characterization of Fish Sauce Aroma-Impact Compounds Using GC-MS, SPME-Osme-GCO, and Stevens' Power Law Exponents. J Food Sci **73**: C268–C274

Phillips-Mora W, Wilkinson MJ (2007) Frosty pod of cacao: a disease with a limited geographic range but unlimited potential for damage. Phytopathology **97**: 1644–1647

Pichersky E, Lewinsohn E, Croteau R (1995) Purification and Characterization of S-Linalool Synthase, an Enzyme Involved in the Production of Floral Scent in *Clarkia breweri*. Arch Biochem Biophys **316**: 803–807

Pichersky E, Raguso RA, Lewinsohn E, Croteau R (1994) Floral Scent Production in *Clarkia* (Onagraceae) (I. Localization and Developmental Modulation of Monoterpene Emission and Linalool Synthase Activity). Plant Physiol **106**: 1533–1540

Pickenhagen W, Dietrich P, Keil B, Polonsky J, Nouaille F, Lederer E (1975) Identification of the Bitter Principle of Cocoa. Helv Chim Acta **58**: 1078–1086

Pipitone L (2012) Situation and prospects for cocoa supply and demand. ICCO- World Cocoa Conf. 19-23 November2012

Ponzio C, Gols R, Pieterse CMJ, Dicke M (2013) Ecological and phytohormonal aspects of plant volatile emission in response to single and dual infestations with herbivores and phytopathogens. Funct Ecol **27**: 587–598

Posnette AF (1950) The Pollination of Cacao in the Gold Coast. J Hortic Sci **25**: 155–163

Posnette AF (1951) Virus research at the West African Cacao Research Institute, Tafo, Gold Coast. Trop. Agric. Trinidad Tobago 28:

Pound FJ (1938) Cacao and witchbroom disease (*Marasmius perniciosus*) of South America. With notes on other species of Theobroma. Report by Dr. F. J. Pound on a visit to Ecuador, the Amazon Valley, and Colombia. April 1937-April 1938. Cacao Witch. Dis. Marasmius Perniciosus S. Am. Notes Species Theobroma Rep. Dr F J Pound Visit Ecuad. Amaz. Val. Colomb. April 1937-April 1938

Priftis A, Stagos D, Konstantinos K, Tsitsimpikou C, Spandidos DA, Tsatsakis AM, Tzatzarakis MN, Kouretas D (2015) Comparison of antioxidant activity between green and roasted coffee beans using molecular methods. Mol Med Rep **12**: 7293–7302

Punyasiri PAN, Abeysinghe ISB, Kumar V, Treutter D, Duy D, Gosch C, Martens S, Forkmann G, Fischer TC (2004) Flavonoid biosynthesis in the tea plant *Camellia sinensis*: properties of enzymes of the prominent epicatechin and catechin pathways. Arch Biochem Biophys **431**: 22–30

Pyrzynska K, Biesaga M (2009) Analysis of phenolic acids and flavonoids in honey. TrAC Trends Anal Chem **28**: 893–902

Qin X-W, Lai J-X, Tan L-H, Hao C-Y, Li F-P, He S-Z, Song Y-H (2017) Characterization of volatile compounds in Criollo, Forastero, and Trinitario cocoa seeds (*Theobroma cacao* L.) in China. Int J Food Prop **20**: 2261–2275

Quainoo AK, Wetten AC, Allainguillaume J (2008) Transmission of cocoa swollen shoot virus by seeds. J Virol Methods **150**: 45–49

Rami J-F (2017) Spidermap v1.7.1, a free software. Unpublished

Reineccius GA (2006) Flavor Chemistry and Technology. doi: 10.1201/9780203485347

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet **101**: 948–955

Risterucci AM, Paulin D, Ducamp M, N'Goran JAK, Lanaud C (2003) Identification of QTLs related to cocoa resistance to three species of *Phytophthora*. Theor Appl Genet **108**: 168–174

Roccia A, Oyant LH-S, Cavel E, Caissard J-C, Machenaud J, Thouroude T, Jeauffre J, Bony A, Dubois A, Vergne P, et al (2019) Biosynthesis of 2-Phenylethanol in Rose Petals Is Linked to the Expression of One Allele of RhPAAS. Plant Physiol **179**: 1064–1079

Rodriguez-Campos J, Escalona-Buendia HB, Contreras-Ramos SM, Orozco-Avila I, Jaramillo-Flores E, Lugo-Cervantes E (2012) Effect of fermentation time and drying temperature on volatile compounds in cocoa. Food Chem **132**: 277–288

Rodriguez-Campos J, Escalona-Buendia HB, Orozco-Avila I, Lugo-Cervantes E, Jaramillo-Flores ME (2011) Dynamics of volatile and non-volatile compounds in cocoa (*Theobroma cacao* L.) during fermentation and drying processes using principal components analysis. Food Res Int **44**: 250--258

Romero Navarro JA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Haiminen N, Mustiga G, Livingstone III D, van Bakel H, Kuhn DN, Parida L, et al (2017) Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases. Front Plant Sci **8**: 1905

Rottiers H, Tzompa Sosa DA, Lemarcq V, De Winne A, De Wever J, Everaert H, Bonilla Jaime JA, Dewettinck K, Messens K (2019) A multipronged flavor comparison of Ecuadorian CCN51 and Nacional cocoa cultivars. Eur Food Res Technol **245**: 2459–2478

Royaert S, Phillips-Mora W, Leal AMA, Cariaga K, Brown JS, Kuhn DN, Schnell RJ, Motamayor JC (2011) Identification of marker-trait associations for self-compatibility in a segregating mapping population of*Theobroma cacao* L. Tree Genet Genomes **7**: 1159–1168

Ruiz M, Sempéré G, Hamelin C (2017) Using TropGeneDB: A Database Containing Data on Molecular Markers, QTLs, Maps, Genotypes, and Phenotypes for Tropical Crops. *In* ADJ van Dijk, ed, Plant Genomics Databases Methods Protoc. Springer, New York, NY, pp 161–172

Sabau X, Loor RG, Boccara M, Fouet O, Jeanneau M, Argout X, Legavre T, Risterucci A-M, Wincker P, Da Silva C, et al (2006) Preliminary results on linalool synthase expression during seed development and fermentation of Nacional and Trinitario clones. 15th Int Cocoa Res Conf Cocoa Product Qual Profitab Hum Health Environ. doi: Corpus ID: 89720924

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**: 406–425

Sakai M, Hirata H, Sayama H, Sekiguchi K, Itano H, Asai T, Dohra H, Hara M, Watanabe N (2007) Production of 2-phenylethanol in roses as the dominant floral scent compound from L-phenylalanine by two key enzymes, a PLP-dependent decarboxylase and a phenylacetaldehyde reductase. Biosci Biotechnol Biochem **71**: 2408–2419

Sant'Ana GC, Pereira LFP, Pot D, Ivamoto ST, Domingues DS, Ferreira RV, Pagiatto NF, da Silva BSR, Nogueira LM, Kitzberger CSG, et al (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. Sci Rep **8**: 465

Sardos J, Rouard M, Hueber Y, Cenci A, Hyma KE, van den Houwe I, Hribova E, Courtois B, Roux N (2016) A Genome-Wide Association Study on the Seedless Phenotype in Banana (Musa spp.) Reveals the Potential of a Selected Panel to Detect Candidate Genes in a Vegetatively Propagated Crop. PloS One **11**: e0154448

Sarma BK, Yadav SK, Singh S, Singh HB (2015) Microbial consortium-mediated plant defense against phytopathogens: Readdressing for enhancing efficacy. Soil Biol Biochem **87**: 25–33

Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M (2014) Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. Plant Physiol **165**: 1120–1132

van Schie CC, Haring MA, Schuurink RC (2006) Regulation of terpenoid and benzenoid production in flowers. Curr Opin Plant Biol **9**: 203–208

Schieberle P, Ofner S, Grosch W (1990) Evaluation of Potent Odorants in Cucumbers (Cucumis sativus) and Muskmelons (Cucumis melo) by Aroma Extract Dilution Analysis. J Food Sci **55**: 193–195

Schwab W, Davidovich-Rikanati R, Lewinsohn E (2008) Biosynthesis of plant-derived flavor compounds. Plant J **54**: 712–732

Schwan RF, Wheals AE (2004) The Microbiology of Cocoa Fermentation and its Role in Chocolate Quality. Crit Rev Food Sci Nutr **44**: 205–221

Serra Bonvehi J, Ventura Coll F (1997) Evaluation of bitterness and astringency of polyphenolic compounds in cocoa powder. Food Chem **60**: 365–370

Shahbandeh M (2019) Cocoa production by country 2012/2013-2018/2019. Statista, https://www.statista.com/statistics/263855/cocoa-bean-production-worldwide-by-region/

Soares S, Kohl S, Thalmann S, Mateus N, Meyerhof W, De Freitas V (2013) Different Phenolic Compounds Activate Distinct Human Bitter Taste Receptors. J Agric Food Chem **61**: 1525–1533

Soles RM, Ough CS, Kunkee RE (1982) Ester concentration differences in wine fermented by various species and strains of yeasts. Am J Enol Vitic **33**: 94–98

Song S, Qi T, Wasternack C, Xie D (2014) Jasmonate signaling and crosstalk with gibberellin and ethylene. Curr Opin Plant Biol **21**: 112–119

Starowicz M, Zieliński H (2019) How Maillard Reaction Influences Sensorial Properties (Color, Flavor and Texture) of Food Products? Food Rev Int **35**: 707–725

Steinhaus M, Sinuco D, Polster J, Osorio C, Schieberle P (2009) Characterization of the Key Aroma Compounds in Pink Guava (Psidium guajava L.) by Means of Aroma Re-engineering Experiments and Omission Tests. J Agric Food Chem **57**: 2882–2888

Sukha DA, Butler DR, Umaharan P, Boult E (2008) The use of an optimised organoleptic assessment protocol to describe and quantify different flavour attributes of cocoa liquors made from Ghana and Trinitario beans. Eur Food Res Technol **226**: 405–413

Sun X, Shen X, Jain R, Lin Y, Wang J, Sun J, Wang J, Yan Y, Yuan Q (2015) Synthesis of chemicals by metabolic engineering of microbes. Chem Soc Rev **44**: 3760–3785

The good scent compagny (2021) http://www.thegoodscentscompany.com/.

Thresh JM (1961) Some isolates of virus causing swollen-shoot disease of cacao in Nigeria and their interrelationships. Ann Appl Biol **49**: 340–346

Thresh JM, Owusu GK, Boamah A, Lockwood G (1988) Ghanaian cocoa varieties and swollen shoot virus. Crop Prot **7**: 219–231

Tohge T, Watanabe M, Hoefgen R, Fernie AR (2013) Shikimate and Phenylalanine Biosynthesis in the Green Lineage. Front Plant Sci. doi: 10.3389/fpls.2013.00062

Tong MKH, Lam C-S, Mak TWL, Fu MYP, Ng S-H, Wanders RJA, Tang NLS (2006) Very long-chain acyl-CoA dehydrogenase deficiency presenting as acute hypercapnic respiratory failure. Eur Respir J **28**: 447–450

Tu Y, Rochfort S, Liu Z, Ran Y, Griffith M, Badenhorst P, Louie GV, Bowman ME, Smith KF, Noel JP, et al (2010) Functional Analyses of Caffeic Acid O-Methyltransferase and Cinnamoyl-CoA-Reductase Genes from Perennial Ryegrass (Lolium perenne). Plant Cell **22**: 3357–3373

Tuenter E, Foubert K, Pieters L (2018) Mood Components in Cocoa and Chocolate: The Mood Pyramid. Planta Med. doi: 10.1055/a-0588-5534

Utrilla-Vázquez M, Rodríguez-Campos J, Avendaño-Arazate CH, Gschaedler A, Lugo-Cervantes E (2020) Analysis of volatile compounds of five varieties of Maya cocoa during fermentation and drying processes by Venn diagram and PCA. Food Res Int Ott Ont **129**: 108834

Van Hall CJJ (1914) Cocoa. Macmillan, London

Wang X, Fan W, Xu Y (2014) Comparison on aroma compounds in Chinese soy sauce and strong aroma type liquors by gas chromatography–olfactometry, chemical quantitative and odor activity values analysis. Eur Food Res Technol **239**: 813–825

Weinert MP, Smith BN, Wagels G, Hutton D, Drenth A (1999) First record of *Phytophthora capsici* from Queensland. Australas Plant Pathol **28**: 93–93

Wollgast J, Anklam E (2000) Review on polyphenols in *Theobroma cacao*: changes in composition during the manufacture of chocolate and methodology for identification and quantification. Food Res Int **33**: 423–447

Wyrambik D, Grisebach H (1975) Purification and Properties of Isoenzymes of Cinnamyl-Alcohol Dehydrogenase from Soybean-Cell-Suspension Cultures. Eur J Biochem **59**: 9–15

Yamada MM, Faleiro FG, Clément D, Lopes U, Pires JL, Melo GRP (2010) Relationship between molecular markers and incompatibility in Theobroma cacao. Agrotropica 71–74

Yin L (2020) CMplot: Circle Manhattan Plot. https://github.com/YinLiLin/CMplot.

Ying H, Qingping Z (2006) Genetic manipulation on biosynthesis of terpenoids. Zhongguo Sheng Wu Gong Cheng Za Zhi J Chin Biotechnol **26**: 60—64

Zhang Y-M, Jia Z, Dunwell JM (2019) Editorial: The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits. Front Plant Sci. doi: 10.3389/fpls.2019.00100

Zheng X-Q, Koyama Y, Nagai C, Ashihara H (2004) Biosynthesis, accumulation and degradation of theobromine in developing *Theobroma cacao* fruits. J Plant Physiol **161**: 363–369

Zhou Z, Zhi T, Han C, Peng Z, Wang R, Tong J, Zhu Q, Ren C (2020) Cell death resulted from loss of fumarylacetoacetate hydrolase in Arabidopsis is related to phytohormone jasmonate but not salicylic acid. Sci Rep **10**: 13714

Ziegleder G (2009) Flavour Development in Cocoa and Chocolate. Ind. Choc. Manuf. Use. Wiley-Blackwell, pp 169–191

Ziegleder G (1990) Linalool contents as characteristic of some flavor grade cocoas. Z Für Lebensm-Unters Forsch **191**: 306–309

Annexes

Liste des annexes

Annexe 21 : Carte physique de la position des marqueurs des associations liées aux composés impliqués dans la biosynthèse des monoterpènes (Appendix 21)

Annexe 22 : Carte physique de la position des marqueurs des associations liées aux composés impliqués dans l'amertume et l'astringence (Appendix 22)

Annexe 23 : Carte physique de la position des marqueurs ayant des allèles dont l'effet supérieur change en fonction du caractère (Appendix 23)

Annexe 24 : Liste des composés biochimques et des notes sensorielles pour lequelles des marqueurs en associations avec un génotype hétérozygote ayant un effet plus favorable (Appendix 24)

Annexe 25: Liste complète des composés volatils identifiés par GC-MS dans les fèves de cacao (Appendix 25)

Annexe 26: Représentation graphique de l'analyse PCA des composés volatils (Appendix 26)

Annexe 27: Liste complète des notes sensorielles détectées (Appendix 27)

Annexe 28: Diagramme de dispersion du DL (r^2) pour les 263661 paires de marqueurs, en fonction de la distance physique (pb), calculé pour tous les individus de la population (Appendix 28)

Annexe 29: Liste complète des zones de associations (Appendix 29)

Annexe 30: Liste du nombre de marqueurs et des zones d'association détectées pour chaque caractère (Appendix 30)

Annexe 31: Carte physique montrant toutes les associations avec les composés de la voie de dégradation de la L-phénylalanine (Appendix 31)

Annexe 32: Carte physique représentant toutes les associations liées aux composés de la voie de biosynthèse des monoterpènes (Appendix 32)

Annexe 33: Carte physique représentant toutes les associations liées aux composés de la voie de dégradation des acides gras et des sucres (Appendix 33)

Annexe 34: Carte physique représentant toutes les associations liées aux composés biochimiques synthétisés grâce à la réaction de Maillard (Appendix 34)

Annexe 35: Liste complète des gènes candidats (Appendix 35)

Annexe 36 : Liste des accessions utilisées (Appendix 36)

Annexe 37 : Liste complète des zones d'association identifiées (Appendix 37)

Annexe 38 : Carte physique représentant les zones d'association identifiées en lien avec les sensations d'amertume et d'astringence (Appendix 38)

Annexe 39 : Liste complète des gènes candidats identifiés dans les zones d'association en lien avec les sensations d'amertume et d'astringence (Appendix 39)

Annexe 40 : Version publiée de l'article : "Two main biosynthesis pathways involved in the synthesis of the floral aroma of the Nacional cocoa variety"

Annexe 41 : Version publiée de l'article : "Integration of GWAS, metabolomics, and sensorial analyses to reveal novel metabolic pathways involved in cocoa fruity aroma"

Annexe 42 : Version publiée de l'acte du 16ème symposium de recherche sur les arômes Weurman