UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT
DROIT D'AUTEUR

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document. Toute contrefaçon, plagiat ou reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

Ecole Doctorale 607 SIReNA - Laboratoire Agronomie et Environnement (LAE)

PhD Thesis under joint supervision

Presented and publicly defended for the degree of

DOCTOR OF THE UNIVERSITY OF LORRAINE AND UNIVERSITY OF ABOMEY-CALAVI

Mention: Agronomic Sciences

by

Guillaume Jesugnon Bodjrenou

Targeted selection of shea trees (*Vitellaria paradoxa* C.F. Gaertn.) rich in fatty acids and specific unsaponifiable substances for the chocolate and cosmetics industries in Benin

Defended in public on May 6, 2026

EXAMINING BOARD:

President	Dr Fanou Nadia Fogny, Full Professor, Faculty of Agricultural Sciences, University of Abomey-Calavi, Republic of Benin
Co-supervisor	Dr Alexandre Olry, Research Engineer, Agronomy and Environment Laboratory, University of Lorraine, France
Supervisor	Dr Enoch G. Achigan-Dako, Full Professor, Faculty of Agricultural Sciences, University of Abomey-Calavi, Benin
Supervisor	Dr Alain Hehn, Full Professor, Agronomy and Environment Laboratory, University of Lorraine, France

Academic year : 2025-2026

PhD Thesis under joint supervision

Presented and publicly defended for the degree of

DOCTOR OF THE UNIVERSITY OF LORRAINE AND UNIVERSITY OF ABOMEY-CALAVI

Mention: Agronomic Sciences

by

Guillaume Jesugnon Bodjrenou

Targeted selection of shea trees (*Vitellaria paradoxa* C.F. Gaertn.) rich in fatty acids and specific unsaponifiable substances for the chocolate and cosmetics industries in Benin

Defended in public on May 6, 2026

EXAMINING BOARD:

President	Dr Fanou Nadia Fogny, Full Professor, Faculty of Agricultural Sciences, University of Abomey-Calavi, Republic of Benin
Reviewer 1	Dr Christine Ouinsavi, Full Professor, University of Parakou, Republic of Benin
Reviewer 2	Dr Kévin Kouamé Koffi, Full Professor, University of Nangui Abrogoua, Côte d'Ivoire
Examiner 1	Dr Nafan Diarrassouba, Full Professor, University of Peleforo Gon Coulibaly, Côte d'Ivoire
Examiner 2	Dr Rose-Monde Megnanou, Full Professor, University of Félix Houphouët-Boigny, Côte d'Ivoire
Co-supervisor	Dr Alexandre Olry, Research Engineer, Agronomy and Environment Laboratory, University of Lorraine, France
Supervisor	Dr Alain Hehn, Full Professor, Agronomy and Environment Laboratory, University of Lorraine, France
Supervisor	Dr Enoch G. Achigan-Dako, Full Professor, Faculty of Agricultural Sciences, University of Abomey-Calavi, Republic of Benin

Academic year : 2025-2026

Thèse de doctorat en cotutelle

Présentée pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITÉ DE LORRAINE ET DE L'UNIVERSITÉ D'ABOMEY-CALAVI

Spécialité : Sciences Agronomiques

par

Guillaume Jesugnon BODJRENOU

Sélection ciblée d'arbres à karité (*Vitellaria paradoxa* C.F. Gaertn.) riches en acides gras et en matières insaponifiables spécifiques pour l'industrie de la chocolaterie et de la cosmétique au Bénin

Soutenue publiquement le 06 mai 2026

COMPOSITION DU JURY :

Présidente du jury	Dr Fanou Nadia Fogny, Professeure titulaire, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, Bénin
Rapporteur 1	Dr Christine Ouinsavi, Professeure titulaire, Université de Parakou, Bénin
Rapporteur 2	Dr Kévin Kouamé Koffi, Professeur titulaire, Université Nangui Abrogoua, Côte d'Ivoire
Examinateur 1	Dr Nafan Diarrassouba, Professeur titulaire, Université Peleforo Gon Coulibaly, Côte d'Ivoire
Examinateur 2	Dr Rose-Monde Megnanou, Professeure titulaire, Université Félix Houphouët-Boigny, Côte d'Ivoire
Co-directeur de thèse	Dr Alexandre Olry, Ingénieur de Recherches, Laboratoire Agronomie et Environnement, Université de Lorraine, France
Directeur de thèse	Dr Alain Hehn, Professeur titulaire, Laboratoire Agronomie et Environnement, Université de Lorraine, France
Directeur de thèse	Dr Enoch G. Achigan-Dako, Professeur titulaire, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, Benin

Année académique: 2025-2026

Targeted selection of shea trees (*Vitellaria paradoxa* C.F. Gaertn.) rich in fatty acids and specific unsaponifiable substances for the chocolate and cosmetics industries in Benin

Guillaume J. Bodjrenou

Thesis Committee

Promoters

Prof. Enoch G. Achigan-Dako
Professor of Genetics and Plant Breeding
Faculty of Agricultural Sciences
University of Abomey-Calavi, Benin

Prof. Alain Hehn
Professor of Molecular Biology and Plant Metabolism
University of Lorraine, France

Co-promoter

Dr. Alexandre Olry
Doctor in Molecular Biology and Protein Biochemistry
University of Lorraine, France

Other members

Dr Fernande Honfo Zannou, Associate Professor, Laboratoire des Sciences et Technologie des Aliments (LaSTA), Faculty of Agricultural Sciences, University of Abomey-Calavi, Cotonou, Benin

Dr Paul-Césaire Gnanglè, Directeur de recherche, Institut National des Recherches Agricoles du Bénin (INRAB)

Targeted selection of shea trees (Vitellaria paradoxa C.F. Gaertn.) rich in fatty acids and specific unsaponifiable substances for the chocolate and cosmetics industries in Benin

Guillaume J. Bodjrenou

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at University of Abomey-Calavi and University of Lorraine
by the authorities of the Rectors,
Prof. Charlemagne Babatoundé Igue (UAC, Benin) and Prof. Hélène Boulanger (UL, France)
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 06 May 2026 at 1 pm in Ahanchede conference room of GBioS-UAC

Guillaume J. Bodjrenou

Targeted selection of shea trees (*Vitellaria paradoxa* C.F. Gaertn.) rich in fatty acids and specific unsaponifiable substances for the chocolate and cosmetics industries in Benin

209 pages

PhD thesis, University of Abomey-Calavi, Cotonou, Benin (2026)
PhD thesis, University of Lorraine, Vandoeuvre-lès-nancy, France (2026)

With references, with summaries in English and French

ISBN: 978-99982-3-865-7

Contents

List of acronyms and abbreviations..... 9

CHAPTER 1 11

General introduction..... 11

CHAPTER 2 9

Toward omics-based breeding strategies for the Shea tree (*Vitellaria paradoxa* C.F. Gaertn.) 9

CHAPTER 3 47

Genetic structure of the shea tree (*Vitellaria paradoxa* C.F. Gaertn.) population and fatty acid-based selection of elite genotypes in Benin 47

CHAPTER 4 78

Metabolomic-Guided Diversity Analysis Reveals Chemotypic Groups and Cosmetic-Relevant Biomarkers in Shea Trees (*Vitellaria paradoxa* C.F. Gaertn.) for Industrial Applications 78

CHAPTER 5 104

Multi-models genome-wide association study revealed candidate genomic regions linked with bioactive metabolites in shea tree (*Vitellaria paradoxa* C.F. Gaertn.) 104

CHAPTER 6 169

General Discussion 169

References 175

Summary 201

Résumé 203

Acknowledgements..... 205

List of publications..... 207

Biography..... 208

Academic Statement..... 209

List of acronyms and abbreviations

AMOVA	Analysis of Molecular Variance
ANOSIM	Analysis of Similarities
ANOVA	Analysis of Variance
AU	Arbitrary Units
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
BLINK	Bayesian Information and Linkage Disequilibrium Iteratively Nested Keyway
bp	Base pair
CBE	Cocoa Butter Equivalent
CID	Collision-Induced Dissociation
CIRAD	French Agricultural Research Centre for International Development
COGENT	International Coconut Genetic Resources Network
CSI:FingerID	Computational Structure Identification: FingerID
DAG	Diacylglycerol
DAPC	Discriminant Analysis of Principal Components
DArTseq™	Diversity Arrays Technology sequencing
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
ESI	Electrospray Ionization
FA	Fatty Acid
FAO	Food and Agriculture Organization
FAOSTAT	Food and Agriculture Organization Statistical Database
FarmCPU	Fixed and Random Model Circulating Probability Unification
FDR	False Discovery Rate
FWHM	Full Width at Half Maximum
G3P	Glycerol-3-Phosphate
GAPIT	Genomic Association and Prediction Integrated Tool
GBS	Genotyping-by-Sequencing
GC	Gas Chromatography
GDP	Gross Domestic Product
GE	Genome Editing
GFF3	General Feature Format version 3
GLM	General Linear Model
GNPS	Global Natural Products Social Molecular Networking
GO	Gene Ontology
GPS	Global Positioning System
GSA	Global Shea Alliance
GWAS	Genome-Wide Association Study
GxE	Genotype × Environment
HCD	Higher-energy Collisional Dissociation
He	Expected Heterozygosity
Ho	Observed Heterozygosity
HPLC	High Performance Liquid Chromatography
HRMS	High-Resolution Mass Spectrometry
IBA	Indole-3-Butyric Acid

INRAE	French National Research Institute for Agriculture, Food and Environment
kb	Kilobase (1,000 bp)
LAE	Laboratory of Agronomy and Environment
LaSTA	Laboratory of Food Science and Technology
LD	Linkage Disequilibrium
m/z	Mass-to-charge ratio
MAF	Minor Allele Frequency
Mb	Megabase (1,000,000 bp)
MCMC	Markov Chain Monte Carlo
MLM	Mixed Linear Model
MLMM	Multiple-Locus Mixed Linear Model
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MTA	Marker–Trait Association
PA	Phosphatidic Acid
PAG	Programme d'Actions du Gouvernement
PC	Phosphatidylcholine
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PGPB	Plant Growth Promoting Bacteria
PIC	Polymorphism Information Content
PLS-DA	Partial Least Squares Discriminant Analysis
ppm	Parts per million
QQ plot	Quantile–Quantile plot
QTL	Quantitative Trait Locus
QTN	Quantitative Trait Nucleotide
RAPD	Random Amplified Polymorphic DNA
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SDG	Sustainable Development Goal
SDGs	Sustainable Development Goals
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
TAG	Triacylglycerol
TEWL	Transepidermal Water Loss
UAC	University of Abomey-Calavi
UHPLC	Ultra-High-Performance Liquid Chromatography
UNDP	United Nations Development Programme
USD	United States Dollar
UV	Ultraviolet
$\mu mol \cdot L^{-1}$	Micromoles per liter
v/v	Volume per volume ratio
VIP	Variable Importance in Projection
ZAE	Agroecological Zone

CHAPTER 1

General introduction

Background and context of the study

Poverty, unemployment, and food insecurity remain major constraints to economic and social development in Africa, particularly in West Africa (UNDP, 2022; World Bank, 2020). In West Africa, agriculture plays a central role in addressing these challenges, as it employs between 60% and 70% of the labor force and contributes substantially to national economies through its share of Gross Domestic Product. Despite this importance, agricultural systems in the region continue to face structural constraints, including land insecurity, inadequate infrastructure, low productivity, climate variability, limited access to financing, and political and institutional instability (Egbule and Okonta, 2024; FAO, 2018; Rincón Barajas, 2024). These constraints hinder the capacity of agriculture to fully perform its role as a driver of inclusive growth and sustainable development (Hien and Kaboré, 2025). Moreover, it undermines long-term projection and weakens the foundations for structuring actions, particularly in financing, training, and strategic investment.

In Benin, these challenges are particularly evident through persistently low agricultural productivity, precarious employment conditions, especially among women and youth, and the underexploitation of plant genetic resources. In response, the Beninese government has, since 2016, implemented an ambitious program aimed at promoting strategic agricultural value chains, among which the shea sector (*Vitellaria paradoxa* C.F. Gaertn.) occupies a prominent position (PAG, 2016). This policy orientation recognizes the potential of shea not only as an economic commodity but also as a lever for social transformation and environmental sustainability (FAO et al., 2022; UNDP, 2019).

Shea is a perennial tree species native to the Sudano–Sahelian and Sudanian zones of Sub-Saharan Africa, where it constitutes a key component of traditional agroforestry parkland systems (FAO and GSA, 2020). It is well adapted to semi-arid environments characterized by poor soils and erratic rainfall, and it provides multiple ecosystem services, including soil fertility maintenance, biodiversity conservation, and carbon sequestration (Boffa, 1999; Hall et al., 1996). Through its integration into agroforestry systems, shea contributes to climate change mitigation and adaptation, while its renewable nature and low-carbon processing pathways further enhance its environmental value (Jepsen et al., 2024).

Socio-economic, environmental, and spatial importance of shea tree

Beyond its ecological functions, shea plays a major socio-economic role in rural livelihoods across its natural distribution range, which extends from Senegal in the west to Uganda in the east, forming what is commonly referred to as the "shea belt" of Africa (Hall et al., 1996). Within this belt, shea parklands are unevenly distributed, reflecting gradients in climate, soil characteristics, land-use systems, and human management practices. In Benin, shea trees are predominantly found in the central and northern regions, where they constitute a structural component of agroforestry landscapes and coexist with annual crops such as maize, sorghum, millet, and cotton (Gnanglè, 2005). The spatial distribution of shea populations is not random; rather, it reflects both ecological adaptation and historical land-use decisions (Allal et al., 2011). Variations in rainfall regimes, temperature, soil fertility, and topography contribute to differences in tree density, productivity, and phenotypic expression across regions (Allal et al., 2011; Gnanglè, 2005). At the same time, farmers' selective retention of trees during land clearing, combined with customary land tenure systems, has shaped the current spatial structure of shea parklands (Kuusaana, 2022). These spatial patterns have important implications for genetic diversity, gene flow, and local adaptation.

From a socio-economic perspective, the collection, processing, and marketing of shea nuts and butter provide income to millions of households, with women accounting for more than 80% of the workforce involved in post-harvest processing and commercialization (Global Shea Alliance, 2022). As such, the shea value chain represents a powerful instrument for women's empowerment, income generation, and the strengthening of rural economies (Opoku – Mensah, 2023a). Shea-related activities often serve as a safety net during periods of food scarcity and agricultural downturns, reinforcing their importance in livelihood resilience (Ayelazuno and Yaro, 2024). At national and regional levels, shea contributes to export earnings and supports the development of local value chains through cooperatives, small and medium-sized enterprises, and semi-industrial processing units (Ayelazuno and Yaro, 2024; Opoku – Mensah, 2023a). Its contribution aligns directly with several Sustainable Development Goals (SDGs), notably SDG 1 (No Poverty), SDG 2 (Zero Hunger), SDG 5 (Gender Equality), and SDG 8 (Decent Work and Economic Growth). In addition, the conservation of shea parklands supports SDG 12 (Responsible Consumption and Production), SDG 13 (Climate Action), and SDG 15 (Life on Land) (FAO et al., 2022; UNDP, 2019). Despite this strategic positioning, strong spatial heterogeneity exists in both productivity and product quality, reflecting differences in environmental conditions, genetic background, and management intensity. This heterogeneity

represents both a challenge for standardization and an opportunity for identifying region-specific or ecotype-specific elite genotypes (Choungo Nguekeng et al., 2021a).

Constraints to shea tree development and spatial implications

The development of the shea sector is constrained by a combination of ecological, agronomic, socio-economic, technological, and institutional factors. Among these, ecological and agronomic constraints are particularly decisive, as they directly affect the spatial distribution, density, and performance of shea trees across landscapes (Seghieri, 2019).

From an agronomic perspective, shea production is characterized by low yields and strong interannual variability, which differ markedly between regions and even between neighboring parklands (Di Vincenzo et al., 2005a). The absence of improved planting material, limited availability of selected germplasm, and difficulties associated with vegetative propagation techniques such as grafting, cutting, and *in vitro* culture significantly hinder domestication efforts (Seghieri, 2019). In addition, the lack of adapted agronomic practices related to fertilization, water management, fire control, and tree rejuvenation further limits productivity and contributes to spatially heterogeneous performance. Ecological constraints include the long biological cycle of the species, with a juvenile phase lasting 15 to 20 years before full production is achieved. Shea is also increasingly exposed to climate change, particularly rising temperatures, prolonged dry spells, and rainfall irregularity (Seghieri, 2019). These pressures are not uniform across its range and tend to exacerbate existing spatial disparities in growth, survival, and yield. Biotic stresses, including pests, diseases, and infestations by parasitic plants such as Loranthaceae, further reduce productivity and may vary in intensity depending on local environmental conditions (Gbemavo et al., 2022).

Together, these constraints highlight the need for spatially informed improvement strategies that take into account regional adaptation, local genetic resources, and environmental heterogeneity.

Biological characteristics and traits of interest in shea

The economic uniqueness of shea resides primarily in its butter, extracted from the kernels of the fruit. Shea butter is distinguished by a fatty acid composition that makes it highly valued by the chocolate industry (Jahurul et al., 2013; Rodriguez-Negrette et al., 2019), where it can serve as an alternative to cocoa butter, as well as by the cosmetic and dermocosmetic industries, which exploit its rich content of bioactive compounds. Shea butter contains significant levels of stearic

and oleic acids, conferring desirable physicochemical properties for food applications. In addition, its unsaponifiable fraction is rich in tocopherols, triterpenes, sterols, flavonoids, and phenolic compounds, which are responsible for its anti-inflammatory, antioxidant, antibacterial, healing, and skin-regenerating properties (Akihisa et al., 2010b; Fiume et al., 2018; Fiume, 2002; Maranz et al., 2003a). These characteristics explain the growing international demand for high-quality shea butter and the increasing interest in differentiating products based on their biochemical composition. Consequently, genetic improvement efforts should focus not only on yield-related traits but also on oil quality, fatty acid composition, bioactive compound content, and the shortening of the juvenile phase, in order to enable faster dissemination of superior genotypes.

Rationale for omics-based improvement approaches

Conventional approaches to shea improvement, largely based on phenotypic selection and natural regeneration, have proven insufficient to achieve significant and rapid genetic gains (Merinosy et al., 2022). These limitations are primarily due to the long generation time of the species, its high level of heterozygosity, and the complex genetic control of key quality traits.

Recent advances in omics technologies, particularly genomics, metabolomics, and lipidomics, offer new opportunities to overcome these constraints (Teh et al., 2017). Genomic tools enable the characterization of genetic diversity, population structure, and the identification of genomic regions associated with traits of interest. Metabolomic and lipidomic analyses provide detailed insights into the variability of bioactive and fatty acid compounds respectively, facilitating the identification of distinct chemotypes with high industrial value (Razzaq et al., 2019). The integration of these approaches allows for a more comprehensive understanding of genotype–phenotype relationships and supports the development of targeted, efficient, and evidence-based improvement strategies for perennial tree species such as shea.

Research problem and knowledge gaps

Despite the recognized importance of shea, major gaps persist in the understanding of its genetic and biochemical diversity, particularly in Benin. Information on population genetic structure remains limited, and systematic characterization of lipid profiles and bioactive compound variability is still scarce. Moreover, the genomic regions underlying variation in these complex traits are poorly documented. The lack of integrative studies combining agromorphological, molecular, biochemical, and genomic data constitutes a significant barrier to the rational improvement and sustainable management of the species. Addressing these gaps is essential to

unlock the full potential of shea as a strategic resource for economic development and environmental sustainability.

Research questions and objectives of the study

Research Questions

In light of the ecological, socio-economic, and scientific challenges associated with the sustainable valorization of the shea tree in Benin, this study seeks to address the following research questions:

1. What is the extent and spatial organization of agromorphological and genetic diversity within shea populations across Benin's parklands?
2. How does genetic variability relate to differences in fat content, fatty acid composition, and bioactive compound content among shea trees?
3. Can shea trees be classified into distinct metabolomic chemotypes based on their fatty acid and unsaponifiable matter composition?
4. Which genomic regions are associated with the observed variability in bioactive compounds, and what do these associations reveal about the genetic control of these traits?
5. How can integrative omics approaches contribute to the identification of elite shea genotypes adapted to specific ecological zones and industrial requirements?

Objectives of the Study

The general objective of this study is to contribute to the genetic improvement of the shea tree in Benin by integrating agromorphological, molecular, biochemical, and omics approaches to identify elite genotypes with superior fatty acid and bioactive compound profiles.

Specifically, the study aims to:

- explore the application of omics approaches for the improvement of key traits in the shea tree;
- assess the phenotypic, genotypic, and biochemical diversity of shea genetic resources in Benin's parklands;
- characterize the metabolomic profiles of shea trees, including fatty acids and unsaponifiable matter, and to identify distinct chemotypes;

- identify genomic regions associated with variability in bioactive compound content using genome-wide association studies (GWAS).

Challenges and expected contributions of the study

This study is expected to generate original scientific knowledge on the genetic, biochemical, and metabolic diversity of shea. By adopting an integrative omics-based framework, it provides insights into the mechanisms underlying variation in economically important traits and supports the identification of high-value genotypes for breeding, conservation, and valorization. The results will contribute to the development of improved planting material, enhance the competitiveness of the shea value chain, and inform public policies related to agricultural development, biodiversity conservation, and climate resilience. More broadly, the methodological approach developed in this work may serve as a reference for the improvement of other underutilized perennial tree species in Africa.

Research Methodology Overview

To address the stated research questions and objectives, this study adopts an integrative and multidisciplinary methodological framework combining field surveys, laboratory analyses, and bioinformatic approaches.

Fieldwork was conducted across representative shea parklands in Benin, selected to capture major agroecological zones and spatial variability. Agromorphological data were collected on mature shea trees to document phenotypic diversity related to growth, fruiting, and yield attributes. Molecular analyses were performed using appropriate DNA-based markers to assess genetic diversity, population structure, and relationships among individuals and populations. Biochemical analyses focused on the characterization of fat content, fatty acid composition, and unsaponifiable fractions containing bioactive compounds using standardized analytical techniques. Metabolomic data were analyzed using multivariate statistical approaches to classify trees into distinct chemotypes based on their bioactive compound composition. Finally, genomic data were integrated with phenotypic and metabolomic information through genome-wide association studies (GWAS) to identify genomic regions associated with key quality traits. This integrated methodology provides a robust framework for linking genetic variation to metabolite expression and supports evidence-based strategies for shea improvement.

Structure of the Thesis

This thesis is structured into six chapters. Chapter 1 presents the general introduction and outlines the context, research questions, objectives, methodology, and significance of the study. Chapter 2 provides a literature review on the use of omics approaches in the genetic improvement of shea, drawing on oil palm and coconut as reference models, and examines the socio-economic context of the species. Chapter 3 analyzes the agromorphological, molecular, and biochemical diversity of shea populations in Benin. Chapter 4 focuses on the metabolic classification of shea trees and the identification of chemotypes based on bioactive compounds. Chapter 5 identifies genomic regions associated with variability in bioactive compound content and discusses the genetic mechanisms underlying these differences. Finally, Chapter 6 presents a general discussion along with the perspective of the study

CHAPTER 2

Toward omics-based breeding strategies for the Shea tree (*Vitellaria paradoxa* C.F. Gaertn.)

Guillaume J. Bodjrenou[1,2], Matice Gbetoho[1], Félicien Amakpé[3], Fernande G. Honfo[4], Alain Hehn[1], Alexandre Olry[1*], Enoch G. Achigan-Dako[2]

[1]Université de Lorraine, INRAE, LAE, 54000, Nancy, France.

[2]Genetics, Biotechnology and Seed Science Unit (GBioS), Laboratory of Crop Production, Physiology and Plant Breeding, Faculty of Agricultural Sciences, University of Abomey-Calavi, 01 BP 526 Abomey-Calavi, Republic of Benin.

[3]Direction Générale des Eaux, Forêts et Chasse, Ministère du Cadre de Vie et des Transport en charge du Développement Durable, BP 388 Abomey, République du Bénin

[4]Food Science Laboratory, University of Abomey-Calavi (UAC), 01 BP 526 Abomey-Calavi, Republic of Benin

Abstract

Shea tree (*Vitellaria paradoxa* C.F. Gaertn.), a semi-domesticated tree in Africa, is traded for its butter, which is a multibillion-dollar product with growing demand on the international market. However, limited scientific investigations were undertaken about its domestication and improvement. In this review, we examine the current state of shea tree improvement in comparison to that of oil palm and coconut trees. We analyze its socioeconomic importance and the genetic diversity underpinning selection and cultivar development. The transferability of breeding methods from oil palm and coconut was examined, with emphasis on fatty acid (FA) biosynthesis pathways and associated genes. Insights from these crops highlighted the importance of integrating omics tools, targeting product quality traits, and structuring breeding programs around well-defined selection criteria for the shea tree. Factors influencing FA biosynthesis in the shea tree were also discussed. Beyond fatty acid composition, traits such as specialized metabolite profiles were found to influence butter quality; integrating these as key traits in the literature supports a breeding strategy combining genomic selection, metabolite profiling, and environmental data as a concrete pathway forward. Omics tools have demonstrated effectiveness in enhancing oil palm and coconut breeding and could be applied in shea tree breeding. This integrated approach highlights the potential of the shea as a strategic resource for rural communities, as well as for the agri-food and cosmetics industries. Applying omics tools previously used in major oil crops to the shea tree opens new perspectives for selecting high-yielding genotypes.

Keywords: Fatty acids; Genomic selection; Oil palm; Coconut tree; Breeding pathway; Genetic diversity; Climate resilience.

Introduction

Perennial oilseed crops make a significant contribution to diversifying income sources and sustaining the livelihoods of local communities in the regions where they are cultivated (Eyog Matig et al., 2002; Hall et al., 1996; Jalani et al., 2003). They are among the primary sources of vegetable oils, which are essential not only for human and animal consumption, but also for various industrial applications, including the cosmetic and chocolate industries (Perera et al., 2010; Soh et al., 2009). In West Africa, the most commonly cultivated perennial oil crops are oil palm (*Elaeis guineensis* Jacq.) and coconut palm (*Cocos nucifera* L.) (Boateng et al., 2016). However, other perennial oleaginous plants are receiving increasing attention from local communities due to their growing economic value in local, regional, and international markets. Among them is the shea tree (*Vitellaria paradoxa* C.F. Gaertn.), widely valued for its butter extracted from kernels, an important export commodity worth multi-billion dollars, particularly for the chocolate, cosmetics, and pharmaceutical industries (Transparency Research Market, 2019).

A report by Transparency Market Research projected that the global shea butter market will reach USD 3.5 billion by 2028, with an annual growth rate of 5.2% during the 2018–2028 period (Transparency Research Market, 2019). More recently, Anil and Roshan (Anil and Roshan, 2021), in a report published by Research and Markets, provided an updated forecast estimating that the market will reach USD 4.7 billion by 2030. Although these projections differ slightly, both sources indicate substantial growth potential for the global shea butter market. One of the primary drivers of this growth is the increasing use of shea butter as a substitute for cocoa butter (Custom Market Insights, 2023). Jahurul et al. (2013) reported that the high cost of cocoa butter on the international market has led to increased interest in cocoa butter equivalents (CBEs) as cost-effective alternatives. Indeed, CBEs are particularly attractive because they can blend seamlessly with cocoa butter without altering its physicochemical properties (Alemawor et al., 2014). These CBEs include shea butter, which is very rich in stearic acid (30-50%), oleic acid (37-47%), linoleic acid (3-10%), and palmitic acid (2-7%) (Allal et al., 2011; Di Vincenzo et al., 2005b; Honfo et al., 2014; Nikiema and Umali, 2007; Okullo et al., 2010). This property makes shea butter a highly prized ingredient in the chocolate industry. In addition, shea butter contains high proportions of unsaponifiable matter, including triterpene alcohols and allantoin (Nikiema and Umali, 2007), tocopherols (Brigelius-Flohé, 2009; Kamal-Eldin and Appelqvist, 1996; Maranz and Wiesman, 2004a), phytosterols (Maranz et al., 2003b; Salau, 2019), and hydrocarbons (Maanikuu and Peker, 2017; Maranz et al., 2003b; Nahm et al., 2013). One of the distinctive features of shea butter is its suitability for use in cosmetics,

attributed to its antioxidant, anti-inflammatory, and skin-healing properties of some of the unsaponifiable matter. Furthermore, it has been revealed that the sterol fraction of the unsaponifiable matter, including α-spinasterol and δ-stigmasterol, campesterol, and β-sitosterol, confers antimicrobial properties, making shea butter highly valuable in the pharmaceutical industry (Moharram et al., 2006). This property makes shea butter a highly prized ingredient in the cosmetic industry. West Africa alone provides almost half of the global demand for shea nuts. The main producing countries include Nigeria, Mali, Burkina Faso, Ghana, Côte d'Ivoire, Benin, and Togo (Bup et al., 2014). The collection of shea nuts and the processing of kernels into butter are activities predominantly undertaken by women. Their extensive involvement in these stages of production has led to shea butter often being referred to as "women's gold" (Beczner et al., 2006; Moudio, 2013).

Despite its economic and socio-cultural significance, the shea tree faces multiple problems that limit the development of its economic potential in producing countries (Seghieri, 2019). These challenges include aging tree plantations, fluctuations in annual yield, very slow growth rates, and limited knowledge of the genetic determinants underlying its fatty acids (FAs) composition and bioactive metabolite potential, which are key factors for its use in the chocolate and cosmetics industries, respectively. Many of the challenges faced by the shea tree were common to other oleaginous perennial species in Africa, notably oil palm and coconut palm. However, significant efforts have been invested in improving these crops through the identification of key traits using omics tools such as genomics, transcriptomics, and metabolomics. For example, Ting et al. (2020) employed transcriptomic analysis to identify genes involved in FAs and triacylglycerol (TAG) biosynthesis in oil palm. In contrast, several important research questions remain unanswered or underexplored for the shea tree; these include: i) What is the socioeconomic importance of shea production in West Africa, and what are the main drivers and constraints affecting its development? ii) How can breeding strategies developed for more advanced perennial oilseed crops, such as coconut and oil palm, inform and guide shea tree improvement? iii) What is the extent of genetic diversity among shea tree populations across Africa, and how do biotic and abiotic factors influence their yield performance? iv) Which biochemical pathways and genetic mechanisms regulate FA biosynthesis in shea, and how do these influence the composition and quality of shea butter? v) How do environmental factors affect the expression of FA biosynthesis genes in oilseed plants, and what are the implications for shea tree physiology and productivity? vi) How can insights from socioeconomic, genetic, biochemical, and environmental analyses be integrated to develop a targeted breeding strategy for superior shea genotypes? vii) How do specialized metabolites contribute to the quality and

economic value of shea butter, and how can their variation be considered in breeding strategies for superior genotypes?

This review aims to explore the socioeconomic, genetic, biochemical, and environmental factors that influence shea production, with the ultimate goal of proposing targeted breeding strategies for its sustainable improvement in West Africa. Specifically, it seeks to assess the socioeconomic importance of shea, compile existing knowledge on the genetic diversity and yield performance of its populations, elucidate the biochemical and genetic regulation of FA biosynthesis, and evaluate the impact of environmental conditions on gene expression. To inform and guide breeding strategies, coconut and oil palm, two more advanced perennial oilseed crops, are used as comparative models to extract relevant insights applicable to shea tree improvement.

1. Economic potential and production distribution of shea in West Africa

Shea tree is increasingly gaining prominence in the international oil market through the commercialization of shea butter, which contains compounds that give it a wide range of roles and applications in cosmetics and nutraceuticals (Maranz et al., 2003b). These compounds are specialized metabolites and FAs that contribute to the quality, functionality, and value of shea butter (Bodjrenou et al., 2026a, 2026b). Among FAs, C18:0 (stearic acid), C16:0 (palmitic acid), C18:1n-9 (oleic acid) contents in particular make shea butter an alternative to cocoa butter in chocolate formulation (Jahurul et al., 2013). West Africa is currently the only region in Africa attempting to meet global demand for shea butter, with a total production of 854,887 tons of shea nuts in 2023 (FAOSTAT, 2024). Studies by Naangmenyele et al. (2023) reported that one kilogram of shea kernels produces around 0.36 kilogram of butter. Thus, an estimate of the total shea nut production reported by FAOSTAT (2024) would have generated a global production of about 307,759 tons of shea butter, for a total amount of 1.8 billion US dollars (USD), considering the export price of one ton of butter at USD 6,000 according to the American Shea Butter Institute (2011). However, it should be noted that the price of shea butter varies significantly depending on several factors that can influence market estimates (American Shea Butter Institute, 2011). The estimate above considers only the export of raw shea nuts. Most producing countries in West Africa export kernels without local processing, capturing only a small fraction of potential value. For example, In 2022, exports of shea nuts and derivatives were valued at USD 75.1 million, up 1.4% from 2021 (Global Shea Alliance, 2022), which represents a very small portion of the global shea market, estimated at around USD 2 billion in the same year. This is much lower than palm oil, valued at USD 67.2 billion, and coconut oil,

valued at USD 5 billion, and expected to reach USD 8.5 billion by 2032 (Grand View Research, 2021; Singh, 2019). Despite its limited cultivation, shea butter, an orphan crop with strong economic potential, is projected to reach a global market value of USD 4 billion by 2032, indicating its growing significance internationally.

In fact, all shea nut production originates from Africa, with West Africa being the primary source. Four major producing countries dominate the global shea nut supply: Nigeria leads with 50.2% of total production between 2017-2021, followed by Mali (21.7%), Burkina-Faso (17.9%), and Benin (10.2%). These countries represent the main contributors to global shea nut production (Figure 1).

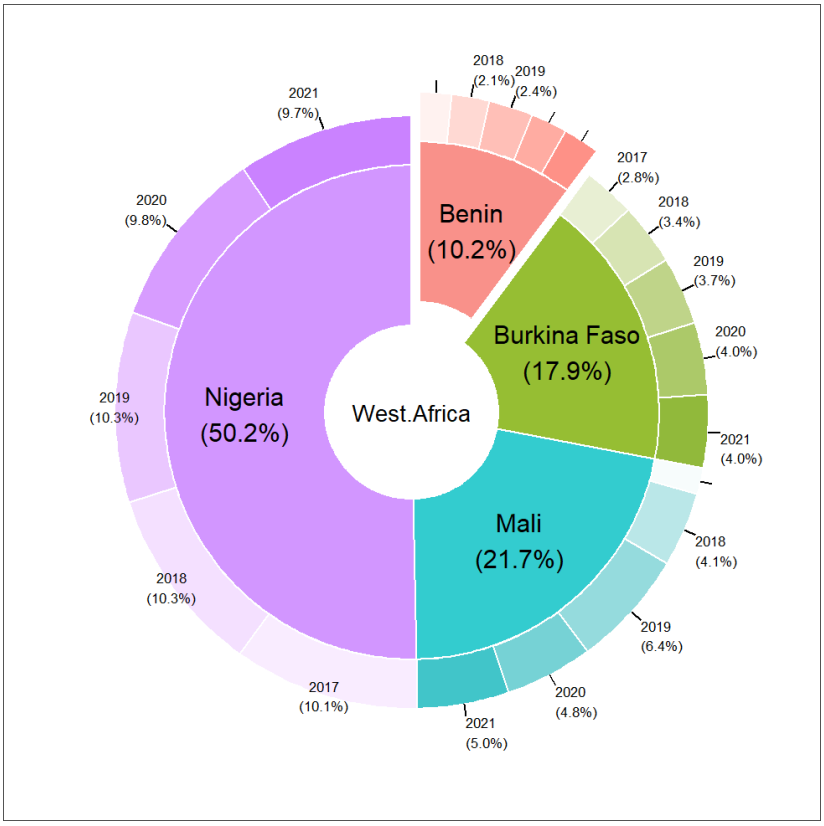


Figure 1: Main shea nut-producing countries worldwide over the last five years, showing the percentage contribution of each country to total annual shea nut production (2017-2021) (FAOSTAT, 2024)

During the last three years (2019–2021), the share of shea nut production in West Africa declined for Nigeria and Mali, whereas Burkina Faso's share increased (Figure 1), although this does not necessarily imply a decrease in absolute production (FAOSTAT, 2024). Although in Burkina Faso this growth coincided with an expansion of the cultivated area beginning in 2019 (Figure 2), the shea tree is characterized by a prolonged juvenile phase, typically lasting 10 to

25 years before reaching substantial fruit-bearing capacity (Choungo Nguekeng et al., 2021b). This biological constraint indicates that recent gains in production may not solely be the result of current interventions or recent planting efforts, but may also reflect the gradual entry into production of trees established many years earlier. In contrast, the area allocated to shea nut cultivation in Nigeria has declined since 2017, stabilizing at 99,389 ha in 2021, compared with 403,911 ha in 2017. Similarly, Mali saw a sharp reduction in cultivated area since 2020 (FAOSTAT, 2024). In Benin, the cultivated area remained relatively stable at ≥ 41,306 ha. However, production trends are influenced not only by cultivated area but also by factors such as national policies promoting the species, multidisciplinary research programs aimed at improving agronomic performance, and investment in the development of the shea value chain (Opoku-Mensah et al., 2024).



Figure 2: Percentage of area allocated to shea production by producer countries during 2017-2021 in West Africa (FAOSTAT, 2024); values represent the percentage contribution of each country to the total harvested area within each year (summing to 100%).

The market potential of shea products differs substantially depending on the level of processing within the value chain, particularly between the export of raw shea kernels and processed shea butter (Opoku – Mensah, 2023b). Existing literature on the global shea value chain consistently shows that most producing countries in West Africa primarily export kernels, thereby capturing only a limited share of the total value generated. In contrast, processing kernels into butter, whether crude or refined, significantly increases the unit value of exports and enables greater domestic value retention (Naangmenyele et al., 2023; Turreira-García, 2025). For instance, recent analyses of the shea sector in Ghana (Opoku – Mensah, 2023b) highlight that while countries such as Nigeria dominate kernel production, Ghana has developed relatively stronger processing and export capacities for butter and its derivatives. This structural difference illustrates how upgrading along the value chain can enhance export revenues and economic benefits.

However, despite the higher market value of shea butter, several constraints limit local processing, including technological barriers, quality standards required by international buyers, and capital-intensive refining processes (Lovett, 2004; Opoku – Mensah, 2023b). As a result, a large proportion of kernels continue to be exported unprocessed, particularly from countries with limited industrial infrastructure. This creates a structural imbalance in the distribution of value, whereby importing countries and multinational firms capture a significant share of the downstream benefits associated with refining and product differentiation (OECD, 2010; Opoku – Mensah, 2023b). Therefore, the comparative market potential of shea kernels versus shea butter is not only a matter of price differentials but also reflects broader issues of industrial capacity, market access, and value chain governance (Opoku – Mensah, 2023b). Promoting local processing into butter remains a key pathway for increasing value addition, improving producer incomes, and strengthening the participation of producing countries in higher-value segments of the global shea economy (OECD, 2010).

2. Specialized bioactive metabolites in shea: composition, variation, and breeding potential

Beyond its fatty acid composition, shea (*Vitellaria paradoxa*) kernels and butter are exceptionally rich in specialized bioactive metabolites, including tocopherols, phytosterols, triterpenes, phenolics, and tannins, that underpin both industrial and nutritional value (Toma et al., 2025). Recent metabolomic analyses by Bodjrenou et al. (2026b) revealed three chemotypic groups among shea trees, with certain accessions displaying particularly high levels of compounds relevant to cosmetics and food applications. Key metabolites such as α-tocopherol,

β-sitosterol, stigmasterol, gallic acid, and butyrospermol were identified as major constituents, while epigallocatechin, gallocatechin, and parkeol emerged as robust biomarkers for distinguishing tree clusters (Bodjrenou et al., 2026b).

This compositional richness distinguishes shea from other oil-bearing crops. Although coconut (*Cocos nucifera*) and palm oil (*Elaeis guineensis*) contain bioactive compounds, including tocopherols, phytosterols, and phenolics, studies have largely focused on general nutritional properties or functional applications, rather than exploiting intraspecific variation for targeted breeding (Ajibola, 2025; Olivares-Tenorio et al., 2024). In contrast, shea's natural variation in bioactive metabolites presents a unique opportunity: selecting and cultivating trees with superior profiles could enhance product quality, increase oxidative stability during harvest, processing, and storage, and provide functional benefits such as pest resistance through phenolics and tannins (Bodjrenou et al., 2026b).

Optimizing these traits would not only strengthen the cosmetic and food value of shea products but also translate into tangible economic gains for rural communities, particularly for women who dominate shea collection and processing (FAO, 2017). By integrating metabolite profiling into breeding programs, shea could be elevated from a traditional commodity to a strategically optimized crop, maximizing both its industrial potential and the livelihoods of those who depend on it.

3. Oil palm and coconut tree as examples for shea tree breeding

The review by Chiriacò et al. (2022) highlights the major role of oil palm in contributing to the Sustainable Development Goals, underscoring its substantial involvement in the global economy as well as its environmental implications. Extensively used in the food, energy, and cosmetics industries, palm oil is particularly rich in unsaturated fatty acids. Owing to its high economic value, palm oil production has expanded markedly across the regions where it is cultivated, especially in Asia, the Americas, and Africa.

The coconut palm supports the nutrition and livelihoods of millions of people through the commercialization of its derived products, particularly coconut water, coconut milk, and coconut oil, which are widely used in the food industry. Among these, coconut oil is the primary product driving the intensive exploitation of the coconut palm (Jayasekhar and Chandran, 2021). Extensively utilized in the food, cosmetics, and pharmaceutical sectors, coconut oil makes a substantial contribution to the foreign exchange earnings of producing countries.

The shea tree is an opportunistic woody species that has received limited attention from research institutions. Consequently, little data is available on developed varieties, despite extensive research efforts focused on improving oil palm through the development of high-performance cultivars (Choungo Nguekeng et al., 2021b). These varieties are being developed based on preliminary studies aimed at identifying traits of particular interest to the food and cosmetics industries. Recent research has also examined the existing genetic diversity of shea tree populations across several West African countries (Choungo Nguekeng et al., 2021b). In most of these countries, local communities classify ethno-varieties based on preference criteria, among which fruit or nut size is often cited as an important trait (Merinosy et al., 2022). However, nut size does not necessarily correlate with butter yield. In practice, trees producing smaller nuts are often associated with higher butter yield. This observation highlights that commonly used morphological traits, such as nut size, may not reliably indicate overall productivity, and partly explains the slow progress in developing high-performance shea varieties. By contrast, oil palm breeding in West Africa began in the 1920s, focusing on classical hybridization between *dura*, *pisifera*, and *tenera* types to select superior morphotypes. This breeding work was conducted in countries including Benin, Côte d'Ivoire, Nigeria, and Ghana (Corley and Tinker, 2008). However, given the evolving production requirements and the constraints and challenges encountered in the production process, the urgent need to develop more adapted varieties has become evident. In particular, the incidence of pathogen attacks, plant size, plantation lifespan, and oil yield have driven several research centers to develop high-performance varieties through more efficient selection methods. For instance, the French agricultural research and international cooperation organization (CIRAD), through its PalmElit company, developed many hybrid varieties grouped into three main categories (Table 1). The breeding methods used to develop oil palm varieties mainly involved molecular approaches such as marker-assisted selection, functional genomics, mutation breeding, and genome editing (GE) (Lei et al., 2014; Shearman et al., 2013; Wang et al., 2019; Yue et al., 2021; Zuhar et al., 2021). Marker-assisted selection and functional genomics have shown very good results, enabling the obtaining of elite varieties, unlike mutation breeding and GE, which still appear to be at the experimental stage (Cros et al., 2015). In addition, all these selection methods involved the use of omics approaches such as genomics, transcriptomics, and metabolomics. These approaches have proven their effectiveness in the case of oil palm breeding and could also be applied to other perennial oilseed plants.

Table 1: Major groups of varieties developed by CIRAD from experiments carried out in Benin, Colombia, Ecuador, Indonesia and Thailand (PalmElit, 2017).

Variety groups	Performance of the varieties	Variety Codes
Protection against disease	Moderate resistance to *Ganoderma*	#G
	High resistance to *Fusariosis*	#F
	Moderate resistance to heart rot complex	#PC
Improved oil quality	High oleic with high olein content	#HO
	Low Lipase with low oil acidity	#L
Optimized plantation life cycle	Short with long planting time	#S
	Compact, high planting density	#C

Another example to illustrate the existing gaps between the level of improvement of shea tree and that of other perennial oilseeds can be illustrated by the coconut palm. The differences in the level of improvement indirectly affect the differences observed in global markets. In fact, in the case of the coconut tree, the best-known classification of coconut varieties is based primarily on tree size and agronomic performance (Perera et al., 2010). The two existing groups of varieties are: tall, known as *typica*, and dwarf, known as *nana*. The tall coconut varieties include the following: Nigerian Tall, West African Tall, Zanzibar Tall, Mozambique Tall, Kenya Tall and Seychelles Tall. In the dwarf groups, there are green (Equatorial green dwarf, Guinea green dwarf), yellow (Ghana yellow dwarf), and red/orange (Cameroon red) varieties (Niral et al., 2019; Ramesh et al., 2022). The breeding techniques used are mainly based on mass selection with free hybridization, as in the case of the varieties developed by the International Coconut Genetic Resources Network (COGENT) (Batugal and Bourdeix, 2005). However, Arhin et al. (2023) have more recently identified many coconut breeding projects for which marker-assisted selection and omics approaches were used. These new advances in coconut breeding techniques range from classical selection approaches through the development of SSR-type microsatellite markers, the study of linkage mapping, to the study of the transcriptome and proteome (Azevedo et al., 2018; Fan et al., 2013; Herrán et al., 2000; Huang et al., 2016; Karp, 1999). The application of these techniques to coconut palms has been made possible by the development of reference genomes, enabling high precision in the identification of genes of interest and a reduction in selection time.

The varieties developed and improvement methods described above show a significant gap compared with the state of improvement in shea. Although the optimum level of improvement has not been reached for all three species, there has been strong progress in the improvement of oil palm for the various traits of interest (Ramli et al., 2020; Subhi et al., 2017). This is also the case for coconut, where the use of omics tools has increased genetic gain in dwarf cultivars (Arhin et al., 2023). Despite the growing economic importance of the shea tree, great progress still needs to be made to reach a state of significant improvement like that of the oil palm and coconut palm. The successful application of omics tools in oil palm and coconut offers relevant insights for shea improvement. However, their transfer to shea should not be considered automatic, but rather adapted to the biological and breeding constraints of the species. In particular, the long juvenile phase, limited domestication, and high variability observed in shea populations highlight the need for a stepwise improvement strategy.

As a first priority, efforts should focus on establishing well-funded breeding and domestication programs based on conventional approaches, including germplasm collection, *in situ* characterization, elite tree selection, and the development of improved planting material through appropriate propagation methods (Ma et al., 2025). These foundational steps are essential for building the phenotypic and genetic resources required for efficient improvement. In a second stage, omics approaches could be progressively integrated to complement conventional breeding by identifying genes, markers, and molecular mechanisms associated with important agronomic and quality traits, such as fat content and quality, fatty acid profile, high content of specialized metabolites with antioxidative and/or health benefits, fruit and nut yield, pest resistance, and production cycle length. Thus, the most promising pathway for shea improvement lies in combining traditional breeding efforts with modern omics-assisted approaches (Ma et al., 2025). Further studies should focus on the application of omics tools in the case of oilseed plant improvement, and thus, the shea tree (Teh et al., 2017).

4. Genetic diversity among shea tree populations in Africa

4.1. Agromorphological traits variation in shea tree populations

Initial classifications of shea trees were based primarily on the indigenous knowledge of local communities. Although lacking scientific or empirical foundations, this approach allowed people to identify trees of interest based on phenotypic traits (Merinosy et al., 2022). Agromorphological characterization was later applied to analyze variations within shea trees using morphological descriptors (Sandwidi et al., 2018), forming the first step in assessing genetic diversity at the species level.

In *Vitellaria paradoxa*, approximately twenty studies have explored genetic diversity using agromorphological traits, revealing significant variability within local germplasm. In Côte d'Ivoire, Diarrassouba et al. (2009) assessed 300 trees using 13 descriptors and found that differences in fruit and leaf shape and size primarily distinguished morphological varieties. In Ghana, Anyomi et al. (2023) initially grouped 282 accessions into two clusters based on leaf traits, later refined to three groups using molecular markers. Despite this, morphological variables remain essential for studying variability in shea trees.

In Nigeria, Ugese et al. (2010) reported that fruit and nut size and weight vary across collection zones, reflecting strong environmental influences. In Benin, producers prefer small nuts, which produce more fruits and higher shea butter content (Merinosy et al., 2022). Similarly, Odoi et al. (2023) in Uganda found an inverse relationship between nut size and oil content, supporting Merinosy et al.'s observations (Merinosy et al., 2022). Other studies have also reported strong correlations among fruit and nut traits (Djekota et al., 2014; Gwali et al., 2012; Yao et al., 2020).

4.2. Performance and key factors affecting fruit yields of shea trees in Africa

Within shea populations, fruit production shows considerable variability, which can be attributed to multiple factors. Beyond climatic, edaphic, geographic, and land-management conditions (Akpona et al., 2016), intrinsic tree characteristics also play an important role in determining yield. Shea fruit production has been extensively studied across West Africa. In their analysis of yield variability and estimation, Bondé et al. (2019) reported yields ranging from 12.77 to 26.31 kg per tree, with the highest values recorded in the Sudanian zone and marked interannual fluctuations. Similar patterns were observed by Aleza et al. (2018) in Benin, who reported an average yield of 17.4 kg per tree at the Farmland site. Furthermore, Kelly et al. (2004) demonstrated the significant effect of location and land management practices on shea productivity. Their study showed that trees located in the southern part of the Sudanian zone had higher yields than those located in the northern part of this zone or in the northern part of the Guinean zone in Mali. In addition, older stands had higher average yields than those observed in fallow land, highlighting the importance of soil management in production. The improved performance of trees in the South West Sudanian zone was attributed to particularly favorable climatic and soil conditions. These include annual rainfall ranging from 600 to 1,400 mm, a prolonged dry season lasting 5 to 7 months, and deep colluvial soils that exceed 30 cm in depth (Boffa, 1999; Hall et al., 1996).

Relationships among phenotypic traits are highly valuable for selection. For instance, studies by Bondé et al. (2019) and Yao et al. (2020) have demonstrated that diameter at breast height

(DBH) and tree circumference measured at 130 cm above ground level are key indicators on which breeding programs can rely to identify productive trees. Bondé et al. (2019) further reported that shea fruit production increases proportionally with tree size and showed a positive association between yield and crown diameter. Their study focused on naturally established shea trees in parklands, where populations are spatially separated to account for genetic diversity, indicating that the observed variation in yield likely reflects high genetic variation among individual trees. Trees with larger crowns capture more light and, through more developed root systems, have greater access to underground water resources. Similarly, Boffa (1996) found that tree size positively influences nut production and fruit weight. The work of Diarrassouba et al. (2021) also supports these findings, showing strong relationships between dendrometric parameters and fruit or nut traits, particularly between tree age and yield. However, this correlation is not always positive: older trees tend to produce lower yields.

The phytosanitary conditions of shea trees are also of vital importance to their performance. Shea is threatened by three hemiparasitic species of the Loranthaceae family (*Agelanthus dodoneifolius*, *Tapinanthus globiferus,* and *Tapinanthus ophiodes*), diseases, and even insect pests. Infestations of parasitic plants have been recorded at different rates in shea parks in Burkina-Faso (Boussim et al., 1993), Nigeria (Odebiyi et al., 2004), Ghana (Asare et al., 2019) and the northern Ivory Coast (Yao et al., 2020). However, few studies have been carried out on the effect of parasite infestations on shea fruit yield.

4.3. Shea tree genetic diversity analysis

Molecular diversity studies on shea tree have been conducted using various molecular markers, including Random Amplified Polymorphic DNA (RAPD), microsatellites (SSRs), Single Nucleotide Polymorphism (SNP), and SilicoDart (Allal et al., 2008; Anyomi et al., 2023; Attikora et al., 2024; Fontaine et al., 2004; Odoi et al., 2023). The first-ever molecular genetic characterization of shea tree was carried out by Fontaine et al. (2004) using RAPD markers on 179 shea individuals and SSR markers on 116 shea individuals in eight countries within the shea's natural distribution zone: Benin, Burkina-Faso, Côte d'Ivoire, Mali, Senegal, Central African Republic, Cameroon, and Uganda. This study enabled the distinction between West African and East African shea populations. The diversity and genetic structure of shea populations were also studied in Mali using various microsatellite loci. The use of eight microsatellite loci in Mali's shea populations revealed a level of heterozygosity ranging from 0.24 to 0.42 (Sanou et al., 2005) and from 0.035 to 0.507 (Cardi et al., 2005), demonstrating moderate genetic diversity between populations.

Furthermore, the impact of human activities on the spatial and temporal genetic structure of shea populations has shown little differentiation between shea populations at the microsatellite locus level, due to exchanges of gene flow (Kelly et al., 2004). However, there is high intra-population genetic diversity, often marked by a lack of tree selection by the local population. Abdulai et al. (2017) reported a high heterozygosity rate (He = 0.67) in their molecular characterization studies (SSR marker) on 200 individuals sampled in four agroecological zones. These results were confirmed by Gwali et al. (2015) in Uganda, who reported heterozygosity levels of 0.63 (Ho) and 0.72 (He), enabling the identification of three populations. This is a case of Allal et al. (2008) who also reported a high level of heterozygosity, reaching 0.87, using 14 microsatellite loci. The research conducted by Hale et al. (2021) provided the reference genome of shea, thereby enhancing the reliability of genomic-based studies on the species' genetic diversity. Recently, Odoi et al. (2023) studied the genetic diversity and population structure of 623 shea half-sibs in Uganda using both SNP and SilicoDart markers. Their results revealed a moderate level of heterozygosity both within the population (He = 0.26) and between populations (He = 0.21). However, the Polymorphism Information Content (PIC) values showed that SNP markers have a higher discriminatory power than SilicoDart. Studies of shea tree populations consistently report high levels of genetic diversity. Anyomi et al. (2023) documented substantial variation both within and among populations. More recently, Attikora et al. (2024) confirmed this trend in Côte d'Ivoire, reporting a heterozygosity rate of He = 0.26 based on SNP markers. These results support the use of SNPs as reliable tools for assessing genetic diversity in shea. Overall, the accumulated evidence points to the presence of a broad and heterogeneous genetic reservoir, which represents a valuable resource for improving shea tree performance across sub-Saharan Africa.

4.4. FAs variation in shea butter

Shea butter is dominated by Stearic–Oleic–Stearic triacylglycerides in West Africa (~31%) and Stearic–Oleic–Oleic triacylglycerides in East Africa (~28–30%) (Akihisa et al., 2011; Honfo et al., 2014), reflecting the distinction between *V. paradoxa* subsp. *paradoxa* and subsp. *nilotica*. Accordingly, stearic-type saturated FA are more abundant in subsp. *paradoxa* (45–57%), whereas subsp. *nilotica* is richer in oleic-type monounsaturated FA (39–46%) (Chibor Bariwere et al., 2017; Davrieux et al., 2010; Garti et al., 2019). Extraction methods only have a limited influence on the FA profile; both chemical processes and traditional methods result in comparable proportions of stearic and oleic acids (Ofoegbu-Chibuzo et al., 2022). Conversely, environmental and climatic conditions, such as latitude, altitude, and microclimate, appear to correlate with compositional differences, with some studies highlighting statistically significant

variations (Animasaun et al., 2019; Davrieux et al., 2010; Di Vincenzo et al., 2005b). Furthermore, reported refractive (1.46–1.5), saponification (172 – 210.37), and iodine (36.63 – 45.6g I2/100g) indices vary depending on the study, reflecting the diversity of analytical techniques and samples (Animasaun et al., 2019; Chibor Bariwere et al., 2017; Garti et al., 2019; Ofoegbu-Chibuzo et al., 2022).

5. Biochemistry and genetic regulation of FA biosynthesis in Shea tree

FA biosynthesis in oleaginous plants generally follows the same principles, and shea (*Vitellaria paradoxa*) is no exception. The initiation of FA synthesis occurs in the plastids (chloroplasts in photosynthetic tissues or non-green plastids in seeds). The key precursor is acetyl-CoA, which is converted into malonyl-CoA by the enzyme acetyl-CoA carboxylase (ACC). Two forms of ACC exist: a heteromeric form in plastids, composed of four subunits [biotin carboxylase (BC, gene *CAC2*), biotin carboxyl carrier protein (BCCP, genes *BCCP1* and *BCCP2*), α-carboxyltransferase (αCT, gene *CAC3*), and β-carboxyltransferase (βCT, gene *accD*)], and a homomeric form (HmACC) localized in the cytosol (Salie and Thelen, 2016). The heteromeric ACC catalyzes the plastidial conversion of acetyl-CoA into malonyl-CoA, which is subsequently transferred to an acyl carrier protein (ACP) by malonyl-CoA:ACP transferase (MAT), forming malonyl-ACP (Misra et al., 2009).

5.1. FAs biosynthesis in plastids and genes associated

FA chain elongation within the plastid proceeds through iterative cycles of condensation catalyzed by enzymes such as ketoacyl-ACP synthases (KAS I–III). Specifically, KAS III initiates chain formation, KAS I catalyzes chain elongation, and KAS II carries out the final extension of the carbon chain (He et al., 2020; Pidkowich et al., 2007). These reactions drive the elongation of the acyl chain through iterative cycles of condensation, consisting of reduction (KAR: ketoacyl-ACP reductase), dehydration (HAD: hydroxyacyl-ACP dehydratase), and a final reduction (ENR: enoyl-ACP reductase) (Baud and Lepiniec, 2010; Tehlivets et al., 2007). This process yields primarily C16:0 (palmitic acid) and C18:0 (stearic acid). The latter can be desaturated by the plastidial stearoyl-ACP desaturase (FAB2, also known as SAD) to form C18:1 (oleic acid) (Zhang et al., 2008). These FAs are released from ACP by acyl-ACP thioesterases (FATA and FATB) and exported from plastids as free FAs. Upon activation by long-chain acyl-CoA synthetases (LACS), they form acyl-CoA esters, which are transported to the endoplasmic reticulum (ER) for further modification and assembly into storage lipids (Li et al., 2016; Zhao et al., 2021). Thus, the plastidial pathway provides the basic set of C16 and C18 FAs that dominate the shea lipid profile.

The expression of genes involved in FA synthesis in shea butter has been studied mainly by Wei et al. (2019) and Hale et al. (2021) using transcriptomic approaches (Table 2). Their work used *Arabidopsis thaliana* L. and *Theobroma cacao* L. as reference species. In the plastidial pathway (known as the "prokaryotic pathway"), the BC, αCT, and βCT enzymes are monogenic and encoded by the CAC2, CAC3, and accD genes, respectively, while BCCP is encoded by two distinct genes: CAC1A/BCCP1 and CAC1A/BCCP2 (Hale et al., 2021).

Wei et al. (2019) showed that the expression of the KAS I, KAS II, and KAS III, which are involved in the condensation reactions between acetyl-CoA and malonyl-ACP, as well as the expression of the KAR, HAD, and ENR genes, which are involved in the reduction and dehydration reactions during each condensation cycle, is higher in shea than in cocoa and *Arabidopsis* (Table 2). One exception is KAR and HAD in *Arabidopsis*, where a similar number of gene copies was observed. Hale et al. (2021) confirm these observations, with values comparable to those reported by Wei et al. (2019). These differences in gene expression in the FA biosynthesis pathway could explain the variations observed in lipid accumulation and FA composition in oilseed plants. However, the biosynthesis pathway does not stop at the plastids. In the cytosol, they are stabilized and transported by acyl-CoA binding proteins (ACBP), forming the so-called FA-CoA pools. These pools constitute the substrate reservoir for glycerolipid assembly in the ER (Li et al., 2016).

5.2. FA modification and gene regulation

Within the ER, FAs undergo further desaturation and elongation. C18:1 esterified to phosphatidylcholine is the substrate of ER oleate desaturase (FAD2), producing C18:2, which in turn can be desaturated by ER linoleate desaturase (FAD3) to yield C18:3 (Lou et al., 2014). Importantly, PC serves as a central "platform" for desaturation, with enzymes such as lysophosphatidylcholine acyltransferase (LPCAT), diacylglycerolcholine phosphotransferase (PDCT), and diacylglycerolcholine phosphotransferase (CPT) allowing dynamic exchange of acyl groups between the acyl-CoA pool, PC, and diacylglycerol (Bates et al., 2009; Lou et al., 2014). In parallel, the FA elongase (FAE) complex, composed of β-ketoacyl-CoA synthase (KCS), ketoacyl-CoA reductase (KCR), and enoyl-CoA reductase (ECR), elongates C18 acyl-CoAs to very-long-chain FAs (VLCFAs, ≥ C20) using cytosolic malonyl-CoA as a carbon donor (Todd et al., 2002).

Overexpression of the FAD2, FAD3, and LACS genes has been observed in shea compared to cocoa and *Arabidopsis* (Hale et al., 2021; Wei et al., 2019) (Table 2). This differential regulation could explain the high proportion of unsaturated FAs in shea, particularly in *V. paradoxa* subsp.

nilotica, compared to cocoa. It is important to note that the number of gene copies reported by Hale et al. (2021) was higher than that described by Wei et al. (2019) (Table 2). This discrepancy could be explained by the use of more diverse germplasm in the study by Hale et al. (2021), including provenances from West and East Africa, unlike the samples from Wei et al. (2019), which came only from West Africa. Nevertheless, the conclusion remains unchanged: these genes are significantly overexpressed in shea trees, corroborating their role in the particular lipid composition of this species.

5.3. Triacylglycerol assembly and gene expression

The assembly of TAGs in shea follows the Kennedy pathway. Initially, glycerol-3-phosphate (G3P) is sequentially acylated by glycerol-3-phosphate acyltransferase (GPAT) and lysophosphatidic acid acyltransferase (LPAT) to form lysophosphatidic acid (LPA) and phosphatidic acid (PA), respectively (Board et al., 2022). PA is dephosphorylated by phosphatidate phosphatase (PAP) to generate DAG, which is then acylated by diacylglycerol acyltransferase (DGAT) or, alternatively, by phospholipid:diacylglycerol acyltransferase (PDAT) to form TAG (Bourgis et al., 1999; Miège and Maréchal, 1999; Ståhl et al., 2004). Finally, the TAGs synthesized in the ER are stored in specialized oil bodies (oleosomes), stabilized by structural proteins such as oleosins, which prevent coalescence and ensure proper mobilization during seed germination (Board et al., 2022).

Through transcriptomics and functional heterologous expression of genes, Wei et al. (2019) has identified a total of 14 functional genes involved in the biosynthesis pathway, particularly in the Kennedy pathway: three GPAT (*VpG1, VpG2, VpG3*), cinq LPAT (*VpL1, VpL3, VpL4, VpL6, VpL7*) et six DGAT genes (*VpD1, VpD4, VpD6, VpD7, VpD8 et VpD11*). Gene identification revealed that overexpression of genes such as VpL7, VpG3, or VpD7 could increase FA and TAG production in shea. The study also revealed that VpD1 or VpD7 might be the functional forms of DGAT in the shea tree. A co-expression study of these different genes would be desirable to assess their impact on lipid content and FA profile.

The specificity of shea oil lies in its high content of C18:0 and C18:1, and relatively low levels of polyunsaturated FAs. This composition reflects a limited activity of FAD2 and FAD3 in shea seeds, as demonstrated by transcriptomic analyses (Wei et al., 2019), and explains the semi-solid texture of shea butter, which is rich in stearic-oleic TAGs.

Table 2: Comparative profile of genes and enzymes involved in FA biosynthesis in *V. paradoxa*, *T. cacao*, and *A. thaliana* (Hale et al., 2021; Wei et al., 2019).

Genes	Enzymes	Gene copy number			
		Wei et al. (2019)			Hale et al. (2021)
		A. thaliana	*T. cacao*	*V. paradaxa*	*V. paradaxa*
ACC2	Homomeric Acetyl-CoA Carboxylase	1	1	1	-
CAC2	Homomeric Acetyl-CoA Carboxylase BC subunit	1	1	1	1
BCCP	Homomeric Acetyl-CoA Carboxylase BCCP subunit	2	3	1	3
CAC3	Homomeric Acetyl-CoA Carboxylase alpha-CT subunit	1	2	2	2
ACCD	Homomeric Acetyl-CoA Carboxylase beta-CTsubunit	1	1	1	-
MAT	Plastidial malonyl-CoA: ACP Malonyltransferase	1	1	1	-
KAS	I Ketoacyl-ACP synthase I	1	2	4	4
KAS	II Ketoacyl-ACP synthase II	1	3	4	3
KAS	III Ketoacyl-ACP synthase III	1	1	2	2
KAR	Plastidial ketoacyl-ACP Reductase	5	3	5	-
HAD	Plastidial Hydroxyacyl-ACP Dehydrase	2	1	2	-
ENR1	Plastidial Enoyl-ACP Reductase	1	2	2	-
FAB2	Stearoyl-ACP Desaturase	7	8	2	-
ACP	Plastidial Acyl Carrier Protein	5	3	3	-
FATA	Acyl-ACP Thioesterase Fat A	2	1	2	2
FATB	Acyl-ACP Thioesterase Fat B	1	5	3	2
FAD2	ER Oleate Desaturase	1	2	2	4
FAD3	ER Linoleate Desaturase	1	1	2	5
KCS	beta-Ketoacyl-CoA synthase	21	20	5	-
KCR	Ketoacyl-CoA Reductase	2	2	2	-
ECR	Enoyl-CoA Reductase	1	1	1	-
LACS	Long Chain Acyl-CoA Synthetase	2	7	5	8
GPAT	glycerol-3-phosphate acyltransferase	10	13	4	-
LPAT	lysophosphatidic acid acyltransferase	9	10	8	-
PAP	Phosphatidic acid phosphatase	2	2	2	
DGAT	Acyl-CoA: Diacylglycerol acyltransferase	3	2	3	-

6. Environmental conditions affecting fatty acid gene expression

FA composition is the key determinant of oil quality in oilseeds. The expression of FA-related genes in plants is complex and regulated at several levels. This gene expression is influenced by various environmental factors (Zhang et al., 2023). However, in the case of shea butter, the impact of environmental conditions on these gene expression pathways remains understudied

and poorly documented. Given the conserved nature of FA biosynthesis pathways across oil-producing species and the well-characterized roles of the associated enzymes, insights from studies on other plant species can be used as a reference. This comparative approach can offer a preliminary of the environmental factors that may similarly regulate gene expression in shea trees.

Studies in *A. thaliana* and *Brassica napus* have identified temperature, drought, salinity, light, and nitrogen as environmental factors affecting FA gene expression (Huang et al., 2008; Soitamo et al., 2008; Zhu and Provart, 2003). In this context, transcriptional or post-translational regulatory mechanisms may be activated, enabling the plant to adapt to changing environmental conditions. The environmental factors listed below could potentially influence the oil content and FA profile of the shea tree.

6.1. Temperature and light

It has been demonstrated that a change in the expression of the gene responsible for the enzyme FAD2 in cotton was observed under the influence of low temperature, allowing the plant membrane to adapt to cold (Kargiotidou et al., 2008). The effect of temperature on the expression level of FA-related genes has also been studied in olive trees by Contreras et al. (2022). They have shown that the thermal factor exerts a regulatory effect on the transcription of the FAB2 and FAD2 genes, affecting FA content, particularly oleic and linoleic acid. This differential expression is sometimes also due to the cultivars involved in the experiment and the plant's stage of development (Wong et al., 2017). Some cultivars may be more susceptible to temperature than others. The FAD2 gene plays an important role in regulating monounsaturated lipid levels in most oilseeds. As for the FAB2 gene, it is considered, according to Parvini et al. (2016), to be the main precursor in oleic acid synthesis in the olive mesocarp. In short-cycle oilseed crops such as rapeseed, sunflower and soybean, high temperatures induce a drop in linolenic acid content but an increase in stearic acid content (Bellaloui et al., 2013; Fernández-Moya et al., 2002).

It has also been shown that the presence or absence of light has a strong influence on the expression of genes involved in FA synthesis (Collados et al., 2006). Likewise, the duration of light exposure and light intensity are also contributing factors (Collados et al., 2006; Soitamo et al., 2008). Light variation during seed filling also affects oil content in oilseed species (Seiler, 1986). Studies by Seiler (1986) show a high proportion of oleic acid and a low proportion of linoleic acid under high radiation conditions in sunflower. However, under low-light conditions, the FAD2 and FAD3 desaturase genes are down-regulated at transcription level.

6.2. Water stress

Thus, temperature and light, as well as water stress, also affect the oil content and FA composition of oilseeds (Allal et al., 2013; Maranz et al., 2004). For example, shea trees, individuals growing in arid regions, have been shown to produce higher levels of tocopherols compared to those in more humid environments (Maranz and Wiesman, 2004a). However, these effects of water stress vary depending on the species. For instance, *Moringa oleifera* has exhibited opposite responses under similar conditions (Anwar et al., 2006). Drought influences seed composition through several mechanisms: it shortens the seed filling period (Desclaux and Roumet, 1996; Meckel et al., 1984; Westgate et al., 1989), negatively impacts the enzymatic activities involved in lipid biosynthesis enzymes (Baldini et al., 2002), and reduces the plant´s ability to assimilate carbon dioxide (Yang et al., 2004).

6.3. Epigenetics behind FA gene expression in perennial oilseed plants

The expression of FA-related genes is also regulated by epigenetic mechanisms, including histone modification. Histone methylation, acetylation, phosphorylation, and ubiquitination are processes that generally take place at the plant level (Gallusci et al., 2017; Gupta and Salgotra, 2022; Kumar et al., 2017). The influence of epigenetic mechanisms on FA synthesis has been very poorly documented, notably in shea tree. However, in species such as *A. thaliana*, Wang et al. Wang et al. (2016) showed modifications at the enzymatic levels (GCN5 acetyltransferase) in histones affecting the FA composition of seeds (transformation of alpha linoleic acid to linoleic acid). Moreover, Moreno-Pérez et al. (2021) characterized the epigenetic modifications that influence FA biosynthesis in sunflowers. Their results identified transcriptional regulators that arise from histone modification and influence FA biosynthesis by regulating expression of the WRINKLED 1 gene. These results show that by acting on a few transcriptional regulatory factors of genes involved in FA synthesis, it is possible to control FA synthesis in oilseed plants. The study of epigenetic phenomena in shea tree remains largely unexplored and constitutes one of the many limitations to its improvement.

7. Leveraging omics for climate resilience in shea: opportunities and contextual challenges

Recent advances in omics technologies offer promising opportunities for improving the resilience of shea tree populations to climate-related stresses, particularly drought and heat, which are major constraints in Sahelian parklands (Elankumaran et al., 2026; Zhou et al., 2022). In several plant species, genomics and transcriptomics have enabled the identification of molecular markers and candidate genes associated with stress tolerance, including those

involved in osmotic regulation, heat-shock responses, and oxidative stress pathways (Huang et al., 2008; Soitamo et al., 2008; Zhang et al., 2023; Zhu and Provart, 2003). For instance, environmental factors such as temperature, light, and water availability have been shown to significantly influence the expression of genes involved in fatty acid biosynthesis and broader metabolic pathways in oilseed species (Huang et al., 2008; Soitamo et al., 2008; Zhu and Provart, 2003). Applying similar approaches in shea could facilitate the identification of SNP markers and gene expression profiles linked to drought and heat resilience, thereby supporting the selection of genotypes better adapted to increasingly variable climatic conditions. Integrating these traits into genomic selection models would be particularly valuable for ensuring both productivity and stability under climate change scenarios (Amin et al., 2025; Palermo et al., 2024).

However, the implementation of such advanced breeding strategies in West African contexts presents significant practical challenges (Tibiri et al., 2025). The high cost of omics technologies, limited access to laboratory infrastructure, and the need for specialized technical expertise may constrain their large-scale application (Tibiri et al., 2025). In addition, the shea value chain is largely managed by women-led cooperatives, whose activities are primarily focused on kernel collection and butter processing, forming a critical source of income for rural households (Beczner et al., 2006; FAO, 2017; Moudio, 2013). Therefore, successful implementation will require the development of context-adapted strategies, including capacity building, strengthening local research infrastructures, and fostering collaborations between international research institutions and local stakeholders (Seghieri, 2019). Participatory approaches that actively involve women producers in the selection and dissemination of improved planting material will be essential to ensure both adoption and sustainability of breeding innovations (Sandwidi et al., 2018).

8. Breeding pathway for the targeted selection of shea trees

The improvement of shea trees based on high FA content, bioactive compounds or high fruit/nut/oil yields requires a study of the genetic diversity of shea tree stands, involving the use of molecular SNP markers. This genetic diversity study should be coupled with a study of the genetic structure of shea tree populations on the basis of morphological traits in relation to existing nucleotide differences. In view of the growing importance of shea butter in the chocolate, cosmetics, food and pharmaceutical industries, a targeted selection of high-triglyceride content trees with a high level of bioactive/and antioxidative compounds profile should be carried out.

In addition to selecting for biochemical and yield traits, reproductive traits such as self-pollination should be investigated (Ortega, 2021). While the shea tree is primarily allogamous, self-compatible genotypes may offer advantages by ensuring fruit set under conditions of low pollinator availability, potentially increasing yield stability in cultivated settings (Shojaiefar et al., 2021). Similar mechanisms in other perennial species ensure fruit set when pollinator availability or compatible mates are limited. For example, *Magnolia grandiflora* L., a predominantly outcrossing tree, exhibits pollinator-mediated self-pollination that guarantees reproductive assurance in isolated individuals, and the perennial *Kosteletzkya virginica* (L.) Presl, displays delayed self-pollination that provides reproductive assurance under conditions of low pollinator visitation (Ruan et al., 2011; Sukumaran et al., 2020). Harvesting practices are another critical factor influencing productivity (Frank, 2024; Selvi et al., 2025). Studies on optimal harvesting times, frequency, and methods are essential to maximize yield while maintaining tree health (Frank, 2024; Selvi et al., 2025). Integration of improved harvesting strategies with breeding programs can substantially enhance overall production.

For the chocolate industry, since shea butter is equivalent to cocoa butter, tree selection should focus firstly on fat content and secondly on the proportions of stearic, oleic, and palmitic acids (Jahurul et al., 2013). In cosmetics, the genetic groups of shea trees identified should be investigated based on the composition of the various phenolic, triterpene, tocopherol and sterols compounds in shea kernels (Akihisa et al., 2011; Maranz et al., 2003b).

Molecular tools have demonstrated their performance and usefulness in the case of improving perennial species. This is the case for oil palm, where improvement has focused on various traits, including yield and oleic acid content (Edem, 2002; Montoya et al., 2014). Molecular methods range from tissue cutting, haplodiploidization, and mutation breeding to more recent methods such as marker-assisted selection, genomic selection, and GE (Figure 3) (Yue et al., 2021).

8.1. Germplasm assembly and conservation efforts

Traditional knowledge systems, particularly those preserved and transmitted by women, play a vital role in the management and conservation of shea trees. The establishment of a breeding program should begin with the conservation of existing germplasm, which represents the only genetic reservoir available for improving the species (Quazi et al., 2021) (Figure 3). At present, conservation efforts rely primarily on in situ collections. Therefore, conservation strategies should prioritize the integration of local communities into improvement programs, both to ensure the long-term maintenance of shea tree populations and to generate essential baseline

knowledge derived from local and indigenous expertise (Ashby, 2009; Sandwidi et al., 2018). Attikora et al. (Attikora et al., 2023) reported that *in situ* conservation of shea trees is not considered a viable option for this species. The main threats to shea conservation relate to land tenure systems in rural areas, damage from pests, insufficient improved plant material, and poverty, which hinders the cultivation of the species (Seghieri, 2019). National governments and local and international conservation organizations should therefore introduce appropriate incentives to promote the preservation of this highly valuable species within agricultural systems, taking into account the constraints mentioned above (Seghieri, 2019). In addition, *ex-situ* conservation will need to be developed in parallel to ensure more sustainable long-term conservation of the species (see section 8.4) (Attikora et al., 2023).

8.2. Genetic diversity, genome-wide studies, and core collection establishment

Assessing genetic diversity with molecular tools, notably Genotyping by Sequencing, will highlight the major genetic groups of shea trees through a genetic structuring of shea tree populations, and this in relation to their agromorphological performance using Genome Wide Association Study (GWAS) (Attikora et al., 2025; Deschamps et al., 2012; Montoya et al., 2014). Agromorphological characterization must take into account a temporal and spatial dimension for better reliability of tree performance results (Jarvis et al., 2016). Although the work carried out to date suggests the existence of genetic diversity in shea trees (Ho ≈ 0.24), most of the genetic characterization work on shea trees has been carried out on local collections of relatively moderate size, which does not provide comprehensive information on subregional or regional genetic diversity in Africa. For this reason, in addition to a characterization study at the local level, it is recommended that a large collection of samples from several countries (≥300 shea trees per country) be integrated, covering all or almost all of the species' spatial distribution. These data should be pooled to generate a robust and informative database that captures the full subregional diversity of the shea tree (Jarvis et al., 2016). One option for implementing such a project could be to create a framework for exchange, sharing, and collaboration among all researchers or institutions working on shea improvement. The collection of agromorphological data should focus primarily on traits of interest in shea for a core collection establishment. Although the work of Attikora et al. (Attikora et al., 2025) has identified potential Quantitative Trait Nucleotides (QTNs) associated with fat content and FA profile in shea, a more comprehensive GWAS enriched with a sufficiently large sample size, as mentioned above, would gain in reliability, precision, and robustness.

8.3. Using Omics tools for trait profiling and cultivar development

As lipid content, FA profile, and unsaponifiable compounds are major traits of interest in shea butter, transcriptomic and metabolomic profiling of shea trees should be established at the same time. The availability of the reference genome through the work of Hale et al. (2021) and the identification of genes involved in the FA biosynthesis pathway by Wei et al. (2019) now offer great potential for exploring the genome and transcriptome of the shea tree. Gas chromatography and ultra-high-performance liquid chromatography, when coupled with high-resolution mass spectrometry using an Orbitrap analyzer, have proven highly effective for the precise identification of compounds in complex samples. These techniques enhance sensitivity and mass accuracy, enabling the detection of low-abundance metabolites with reduced background noise. In the case of shea tree, they can establish a comprehensive metabolomic profile by quantifying FAs across morphotypes, as well as identifying phenolic compounds, tocopherols, triterpenic alcohols, and phytosterols (Bodjrenou et al., 2026a, 2026b; García-Cañas et al., 2014; Ovbude et al., 2024; Zhou et al., 2021). Recent studies confirm that these approaches provide both qualitative and quantitative insights into shea metabolites, which are highly sensitive to post-harvest handling and processing practices. For example, analyses of neutral and polar lipids have shown that traditional fermentation methods influence the lipid composition, reducing free fatty acids and polar lipids, thereby improving both quality and extractability (Turreira-García et al., 2025). Similarly, GC–MS-based quantification of triterpenes, tocopherols, and sterols has highlighted the predominance of triterpene esters in unsaponifiable matter and their potential contribution to bioactivity such as antioxidant and anti-inflammatory effects (Granby et al., 2025; Toma et al., 2025).

Importantly, recent metabolomic profiling of phenolic compounds demonstrates that their diversity and abundance are significantly modulated by local kernel treatments such as storage duration, sun-drying, boiling, and fermentation (Forte Neran et al., 2025; Nielsen et al., 2025). This indicates that phenolic profiles are not only genetically determined but also strongly influenced by environmental and processing factors, reinforcing the need to integrate metabolomics with agronomic and post-harvest data when assessing shea quality. Such integrative profiling provides actionable information for breeding programs and cultivar development, enabling the selection of morphotypes that combine superior metabolite profiles with desirable processing characteristics. Overall, these combined chromatographic and high-resolution mass spectrometric approaches allow a comprehensive mapping of shea's metabolome, supporting trait selection and the development of improved cultivars with enhanced nutritional, functional, and industrial value.

The transcriptomic strategy based on differential gene expression, as applied by Hale et al. (2021) and Wei et al. (2019) to identify genes involved in FA biosynthesis, represents a valuable approach that should be leveraged to investigate biochemical variability among individual shea trees. Applying this approach to the shea core collection could reveal the differences in gene expression responsible for the observed biochemical performance among shea trees. All this information should help set up genomic prediction models, as Shea's long production cycle makes Marker-Assisted selection less preferable (Degen and Müller, 2023).

The influence of environmental effects on elite genotypes must be considered when evaluating genotype stability and adaptability. The effect of the environment on variations in the agronomic performance of shea had already been demonstrated by Aleza et al. (2018), Kelly et al. (2019), and Bondé et al. (2019) (see section 4.2). Thus, taking environmental parameters into account is essential in genomic selection (Figure 3) (Papin et al., 2024).

8.4. Propagation and shortening of the Shea production cycle

Furthermore, the development of tissue culture protocols specific to the shea tree will help provide healthy, vigorous shea seedlings for the rapid regeneration of shea parks. Existing vegetative propagation techniques, such as grafting, need to be optimized and the length of the production cycle reduced (Tchokponhoué et al., 2017). Yeboah et al. (2020) reported that propagating shea trees from cuttings yielded satisfactory results in terms of rapid root development when the cuttings were pretreated with a growth regulator such as indole-3-butyric acid (IBA, an auxin) with a concentration of 10,000 mg/L. Furthermore, Tchokponhoué et al. (2017) demonstrated that, in *Synsepalum dulcificum* (Schum. et Thonn.) Daniell, the production cycle could be shortened through the use of appropriate fertilizers, soil amendments, and effective water management.

Yao et al. (2020) reported that terminal slit grafting allowed for rapid plant recovery within 21 days, contrary to the findings of Sanou et al. (2004), who identified side veneer grafting as the most effective method. *In vitro* culture was also explored by Lovett and Haq (2013), who showed that the use of diluted Murashige and Skoog media combined with specific cytokinin (BA: Benzyladenine and NAA: 1-Naphthaleneacetic acid for shoots) and auxin (IBA: Indole-3-butyric acid for roots) growth regulators can effectively induce vegetative regeneration of shea trees, paving the way for the selection and propagation of superior genotypes. Further research is needed to determine the most effective vegetative technique for promoting rapid growth in shea trees. An integrated approach should also be considered, combining different propagation methods, the application of growth hormones, and optimized soil fertility and water

management, in order to overcome the constraint associated with the long production cycle of this species. Sexual reproduction remains to be investigated with the application of various treatments to lift the dormancy of shea nuts (Jatto et al., 2012). Given the allogamous nature of the shea tree, haplodiploidization remains an alternative for elite genotype selection (Ramawat et al., 2014).

Modern cultivation strategies should aim not only at reducing the production cycle through vegetative propagation and tissue culture, but also at developing cultivars with early fruiting, high fruit set, uniform maturation, pest resistance, and adaptability to varied climatic conditions. In addition, selection should target key bioactive compounds, including phenolic compounds, tocopherols, flavonoids, triterpenic alcohols, phytosterols, and allantoin, which are important for cosmetic, dermocosmetic, and pharmaceutical applications (Honfo et al., 2014; Masyita et al., 2022) . Given the difficulty of finding individual trees that combine all desirable agronomic and biochemical traits, a targeted selection approach focusing on specific traits according to the intended end-use is recommended. This strategy would allow the development of cultivars optimized either for industrial chocolate production or for cosmetic and pharmaceutical applications, while ensuring high productivity and adaptability in smallholder and industrial systems.

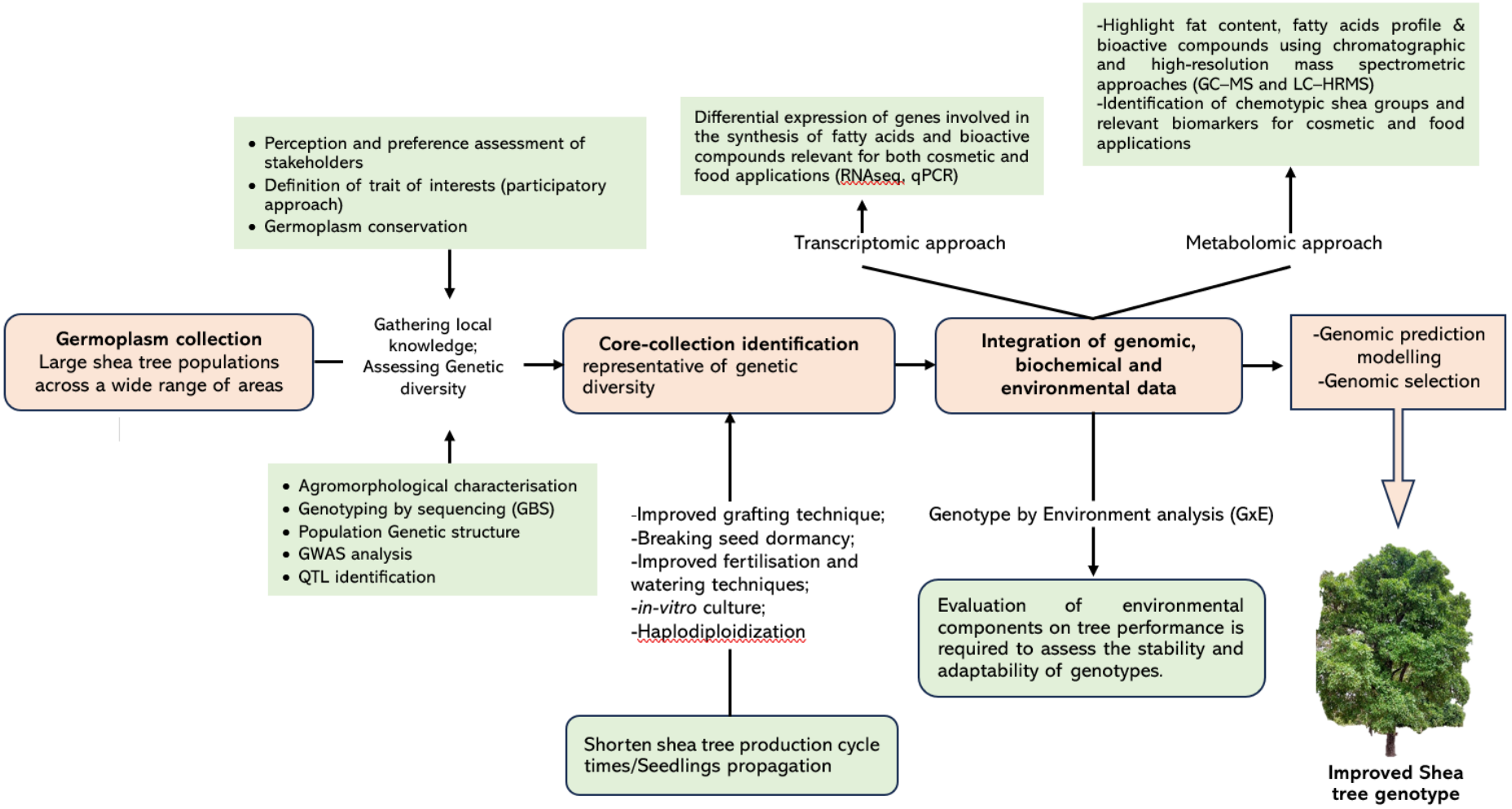


Figure 3: A roadmap for shea tree improvement based on the trait of interest using omics approaches

Conclusion

The present review highlights that, despite its major socio-economic importance, the shea tree remains an under-improved perennial oilseed species compared to oil palm and coconut. However, its high genetic diversity and recent advances in omics technologies provide a strong basis for targeted improvement.

Key traits identified for breeding include reduced juvenile phase (shorter production cycle), yield and yield stability, and fatty acid composition, which determines suitability for cocoa butter equivalents. Additional important targets include total lipid content (TGA) and reduced lipase activity, which influence oil quality. Beyond lipids, specialized metabolites such as phenolics, flavonoids, tocopherols, sterols, and triterpene esters are critical, as they contribute to oxidative stability, pest resistance, and added nutritional and cosmetic value (e.g., antioxidant, antimicrobial, and skin-protective properties). Experiences from oil palm and coconut demonstrate that combining structured breeding programs with omics-assisted approaches can accelerate genetic gains. In these species, marker-assisted selection, genomics, and transcriptomics have enabled the development of elite varieties with improved oil quality, disease resistance, and optimized agronomic traits. For shea, a stepwise strategy is recommended. Initial efforts should focus on germplasm characterization and elite tree selection, followed by the gradual integration of omics tools to identify markers linked to key traits and support genomic selection. These approaches should be coupled with participatory breeding involving local communities, particularly women, to ensure adoption and sustainability.

In conclusion, integrating conventional breeding, omics tools, and local knowledge offers a realistic pathway to develop climate-resilient, high-value shea varieties adapted to West African conditions.

CRediT authorship contribution statement

Guillaume Jesugnon Bodjrenou: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; validation; visualization; writing—original draft; writing— review and editing. Matice Gbètoho: investigation; methodology; writing—review. Félicien Amakpé.: writing—review and editing. Fernande Gbenato Honfo: supervision; investigation; writing—review and editing. Alexandre Olry: Supervision; methodology; validation; visualization; writing—review and editing. Alain Hehn: Supervision; methodology; validation; visualization; writing—review and editing. Enoch Gbènato Achigan-Dako: Conceptualization; funding acquisition; methodology; project administration; supervision; validation; visualization; writing—review and editing.

Declaration of generative AI use

The authors used ChatGPT and Grammarly to assist with the identification of grammatical errors and to improve the academic writing in parts of the manuscript. The authors carefully reviewed and edited the text where necessary and take full responsibility for the scientific content of the publication.

Funding

This work was fully supported by the PEA-BIOVALOR project: *Renforcement de la démarche compétences et de la culture entrepreneuriale dans la formation supérieure agronomique : leviers pour l'insertion professionnelle et le développement de la bioéconomie au Bénin*, financed by the Agence Française de Développement (AFD) through the Agence Nationale de Recherche (ANR). Grant Number: ANR-21-PEA2-0006.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The manuscript is approved by all authors for publication.

Acknowledgment

We express our gratitude to the Genetics, Biotechnology and Seed Sciences (GBioS) Unit of the Faculty of Agricultural Sciences, University of Abomey-Calavi (Benin), for organizing and supporting the supervision of this work. We also extend our sincere thanks to the entire team of the Agronomy and Environment Laboratory (LAE) of the University of Lorraine for their invaluable technical assistance.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request

CHAPTER 3

Genetic structure of the shea tree (*Vitellaria paradoxa* C.F. Gaertn.) population and fatty acid-based selection of elite genotypes in Benin

Guillaume J. Bodjrenou[1,2], Matice Gbètoho[2], Fernande Honfo[3], Cyril J.F. Kahn[5], M. Eric Schranz[6], Paul Césaire Gnanglè[4], Alain Hehn[1], Alexandre Olry[1], Enoch G. Achigan-Dako[2*]

[1]Laboratoire Agronomie et Environnement de l'Université de Lorraine (Université de Lorraine, INRAE, LAE, 54000, Nancy, France)

[2]Genetics, Biotechnology and Seed sciences Unit, Faculty of Agronomic Sciences, University of Abomey-Calavi (GBioS/UAC), Abomey-Calavi, Republic of Benin

[3] Laboratoire des Sciences et Technologie des Aliments, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi (LaSTA/UAC), République du Bénin

[4]Institut National des Recherches Agricoles du Bénin (INRAB), Abomey-Calavi, Republic of Benin

[5]Laboratoire d'Ingénierie des Biomolécules de l'Université de Lorraine (LIBio, Université de Lorraine, F-54000 Nancy, France)

[6]Biosystematics group, Wageningen University & Research (WUR), Netherlands

DOI : https://doi.org/10.1007/s12298-026-01734-2

Published in Physiology and Molecular Biology of Plants, 1–16.

Abstract

The shea tree (*Vitellaria paradoxa* C.F. Gaertn.) is an important source of oil, primarily used in the chocolate, cosmetics, and food industries. However, the absence of elite domesticated varieties limits breeding and optimal utilization. This study explores the genetic diversity, population structure and biochemical variation to identify genotypes with high oilseed potential in Benin. A total of 167 trees were genotyped using DArTseq technology, generating 3,720 high-quality single nucleotide polymorphism (SNP) markers after strict filtering. Data for two years (2023-2024) were collected, including agromorphological and biochemical data (fat content and fatty acid profile). Genetic diversity analysis revealed a moderate genetic diversity, with expected heterozygosity values of He = 0.21. Analysis of Molecular Variance (AMOVA) revealed significant differentiation ($p < 0.01$), with 7.9% of variance attributed to differences among populations. Population structure analysis distinguished three genetic groups (Pop1, Pop2, Pop3), with strong differentiation between Pop1 and Pop3 (FST > 0.85). Correlation analysis revealed positive relationships between nut traits and fatty acid composition. Five major fatty acids were found in the shea kernels: stearic (C18:0; 141.5 mg/g ≈ 45.6%), oleic (C18:1n9, 125.49 mg/g ≈ 43.7%), linoleic (C18:2n6, 17.43 mg/g ≈ 5.98%), palmitic (C16:0, 9.9 mg/g ≈ 3.6%), and arachidic (C20:0, 4.63 mg/g ≈ 1%). Genotypes BG104, BG109, BG114, BG214 and BG258 were elite candidates for fat and fatty acid content. This study offers an overview of the genetic potential of shea trees in Benin, identifying the best genotypes with high concentrations of fat and fatty acid composition for future breeding programs.

Key words: Shea tree; Fat content; Fatty acids; genetic diversity; Population structure

Background

The shea tree (*Vitellaria paradoxa*), which is endemic to West Africa, is an oil-bearing plant species that is becoming increasingly important from economic, ecological and biological perspectives (Moore, 2008; Pouliot, 2012). Shea is mainly used for its butter, derived from its kernels, a product with high economic value on the international oil market. Shea butter is equivalent to cocoa butter and is therefore used as an essential ingredient in chocolate formulation and confectionery (Norazlina et al., 2021). Shea butter is rich in five main types of fatty acid: palmitic (C16:0), stearic (C18:0), arachidic (C20:0), oleic (C18:1n9), and linoleic (C18:2n6). Four of these fatty acids make shea butter a popular alternative to cocoa butter in chocolate production. These fatty acids are: stearic (C18:0), palmitic (C16:0), oleic (C18:1n9), and linoleic acid (C18:2n6) (Davrieux et al., 2010).

Unfortunately, the vast majority of shea nuts are sourced from wild trees, some of which are still in their natural habitat, while others have been semi-domesticated, creating a shortfall in the shea value chain (Lovett and Haq, 2000; Seghieri, 2019). Boffa (1996) and Lamien et al. (2004) revealed that about 15% of the shea trees in a parkland actually contribute to the overall production recorded. Intensive investigation is needed to select high-performing shea trees effectively (Maranz et al., 2004). Developing varieties with high fruit and fat yields is one of the key objectives of improving the shea tree. Similarly, the secondary metabolites contained in the shea kernel are of growing importance to the cosmetics industry, thanks to the antioxidant, anti-inflammatory and healing properties conferred by phenolic compounds, tocopherols, phytosterols and triterpene alcohols (Maranz et al., 2003a).

A project to select or improve the shea tree involves first defining the traits of interest and studying genetic diversity, which should provide information on the species' genetic reservoir (Chakrabarty and Külheim, 2025). In addition to fruit and nut yield, fat content and metabolomic profile are traits of interest. It is worth highlighting that the long production cycle and mixed breeding system or clonal production of the species are an obstacle to the improvement process (Norazlina et al., 2021; Seghieri, 2019). Research has focused on optimizing vegetative propagation techniques to shorten the production cycle of trees. For instance, studied grafting methods that included several splitting techniques. However, they observed a mortality rate of 20.7% after 16 weeks, in the most successful treatment, which utilized side veneer grafting (Sanou et al., 2004).

When it comes to cuttings, successful rooting necessitates the selection of apical cuttings, which should be pruned or cut at the meristem level to ensure an acceptable rooting rate. Additionally,

applying a hormonal treatment such as auxin can lead to satisfactory results (Yeboah et al., 2020, 2009). Recently, *in vitro* culture techniques have been utilized on shea tree root explants to facilitate rapid and healthy multiplication of the tree. The results demonstrated that applying hormonal treatment with auxin (1-naphthaleneacetic acid) and cytokinins (6-benzylaminopurine and kinetin) at specific concentrations in a Murashige and Skoog (MS) culture medium significantly enhanced root and shoot development (Afful et al., 2022; Lovett and Haq, 2013). Although this research can be further refined, it offers valuable insights and opportunities to tackle the challenges posed by the prolonged reproductive phase in shea. Additionally, the agronomic and biochemical performance of the clones produced during this propagation process is inherently linked to the genetic quality of the original plant material. Therefore, it is essential to identify the best-performing genotypes that can serve as reference material. This emphasizes the need for high-quality base materials to achieve optimal agronomic outcomes in clonal propagation (Mayer et al., 2021; Werner et al., 2023). The selection of improved shea tree genotypes involves a pre-breeding process that requires knowledge of the species' genetic reservoir, making the study of its genetic diversity essential (Sukumaran et al., 2022). Sequencing the reference shea genome has provided essential information on its structure and composition, chromosome mapping, and gene function. The genome is approximately 658 Mbp in size, has 2n = 2x = 24 chromosomes, and contains around 38,500 identified functional genes (Hale et al., 2021). This makes it possible to explore genetic diversity more accurately and reliably.

The genetic diversity of shea trees was studied by several researchers who observed moderate genetic variability within the species using agromorphological descriptors and molecular genetic markers, which present potential selection opportunities. Investigations into genetic diversity using isoenzymes and SSR markers reported a range of genetic diversity, from moderate (He = 0.2142; FST = 0.0124) to high (He = 0.667; FST = 0.113; FIS = 0.136). Conversely, diversity assessed through SNP (Single Nucleotide Polymorphism) markers revealed relatively low genetic diversity (Ho = 0.034; Hs = 0.0406; FIS = 0.002; FST = 0.0001) with very slight differentiation among the identified genetic groups (Anyomi et al., 2023). Attikora et al. (2024) also explored genetic diversity in shea using markers, and found, unlike previous studies by Anyomi et al. (2023), moderate genetic diversity (Ho = 0.17; He = 0.26) with a low level of differentiation between populations (FST = 0.004, Nm = 59.02). All these studies indicate moderate genetic diversity within shea, which may be attributed to its semi-domesticated status. This situation creates a relatively homogeneous subpopulation within the species, thereby reducing the observed genetic diversity.

In Benin, the genetic diversity of shea trees has been insufficiently explored. Existing studies primarily rely on morphological descriptors and local knowledge from producers. For instance, Merinosy et al. (2022) noted that local varieties of shea trees in northern Benin are identified based on the size of their fruits and nuts. These findings highlight a significant gap in our understanding of the genetic diversity of shea in Benin. To address this gap, the genetic structure of shea tree populations in Benin was investigated using SNP markers. Furthermore, selecting shea trees, based on their fat content and fatty acid composition, will help identify elite genotypes from Benin's shea germplasm which can be integrated into future breeding programs. The research questions addressed in this article are: i) What is the extent of SNP polymorphism in the Shea tree populations, and how are the individuals structured across agroecological regions and shea parks in Benin? ii) What are the elite genotypes of the most expanded population of shea trees based on the fatty acid profiles? To address all these questions, the main objective of this study is to establish the genetic structuring of shea tree populations in Benin. Specifically, this study will (i) highlight the genetic diversity of Benin's shea trees; (ii) identify the number of gene pools existing within the trees studied; and (iii) determine the fat content of some genotypes, highlighting their fatty acid composition. This study presents three key hypotheses: (i) there is genetic diversity among the 167 shea trees, allowing selection process; (ii) the shea trees show significant genetic differentiation, enabling their classification into distinct genetic groups; and (iii) the fat content and fatty acid composition vary among the genotypes in the bigger population, which makes it possible to identify the most promising genotypes adapted to several mega environments.

Methodology

Plant Materials

The plant material comprised a total of 167 shea trees collected from four agro-ecological zones in the northern region of Benin during 2023 and 2024 years. These trees were located across nine communes: Parakou (19), Tchaourou (25), Ndali (10), Boukoumbé (21), Bembèrèkè (15), Pehunco (26), Kandi (24), Gogounou (30), and Segbana (26) (see Figure 1). The agro-ecological zones include Zone 2 (North Benin Cotton Zone), Zone 3 (South Borgou Food Zone), Zone 4 (West-Atacora Zone), and Zone 5 (Central Benin Cotton Zone) (see Figure 1). The pedoclimatic characteristics of the agroecological zones related to the municipalities in our sampling plan were presented in Table 1. A minimum distance of 100 meters was maintained between sampled individuals to prevent genetic pollution. The GPS coordinates of each tree

were recorded, enabling accurate mapping of the sampled trees. Young leaves were collected from each tree and immediately dried using silica gel to preserve DNA integrity for subsequent genomic analyses.

Table 1: Pedoclimatic characteristics of the agroecological zones (ZAE) with the municipalities involved in the sampling plan for shea tree accessions, table adapted from studies of Aholoukpè et al. (2020)

N°	ZAE	Communes	Area (km²)	Climate	Average annual rainfall (mm)	Soil types	Vegetation
2	North Benin cotton zone	Kandi, Gogounou, Ségbana	20,930	Sudanian zone with an influence of the Sudano-Sahelian climate near Kandi.	Very irregular, spanning from 800 to 1200.	Ferruginous (Lixisols), Vertisols and hydromorphic (Gleysols)	The vegetation primarily consists of wooded savannas featuring species such as *Adansonia digitata, Bombax costatum, Lannea microcarpa, Parkia biglobosa, Sclerocarya birrea, Hyphaene thebaica, Ceiba pentandra, Blighia sapida,* and *Vitellaria paradoxa*. (Ganglo and Maître, 2003)
3	South Borgou food zone	Péhunco, Bembèrèkè, N'dali	23,442	Sudanian	1100 - 1200	Leached Ferruginous (Lixisols)	
4	West-Atacora zone	Boukoumbé	16,936	Sudanian	Strong disparity in rainfall 800 - 1500	Ferruginous soils (Leptosols), less evolved raw minerals (Aerosols and Fluvisols), and hydromorphic soils (Fluvisols & Gleysols) relatively clayey and fertile.	
5	Central Benin cotton zone	Parakou, Tchaourou,	31,722	Sudano-Guinean with 1 rainy season	600 - 1400	Tropical ferruginous soils, leached and relatively concretionary (Leptosols and Luvisols); ferralitic soils with a sandy to clayey-soil texture (Acrisols)	Vegetation in the Guinean region is characterized by wooded savannas, dominated by a mosaic of different savanna types. The main species include *Daniellia oliveri, Isoberlinia doka, Blighia sapida, Vitellaria paradoxa,* and *Parkia biglobosa*.

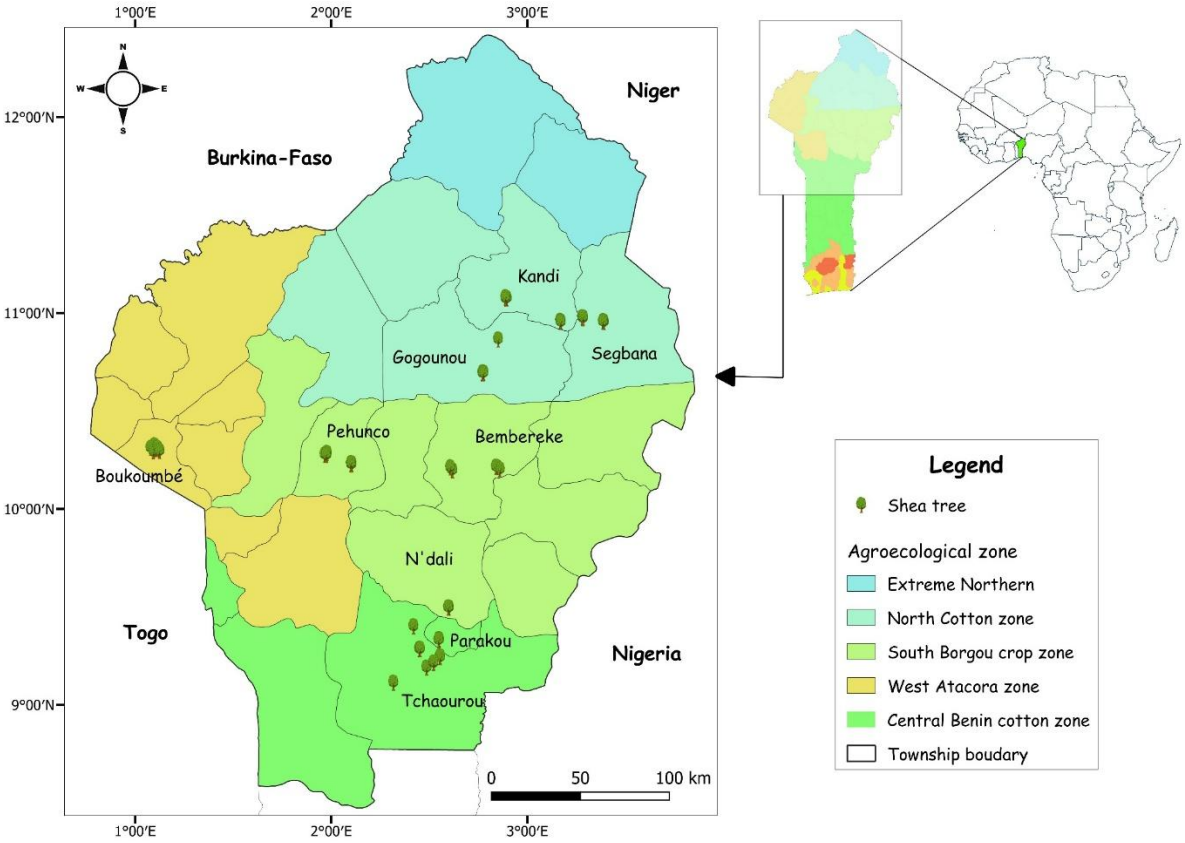


Figure 1: Distribution of Shea tree sampled across the agroecological zones of Benin.

DNA extraction and genotyping-by-sequencing data acquisition

Shea tree leaf samples were sent to SEQART AFRICA at the International Livestock Research Institute in Nairobi for Genotyping. DNA extraction was done from leaf tissue using the Nucleomag® Plant DNA extraction kit. The genomic DNA extracted was in the range of 50 - 100ng/ul. The genomic DNA quality and quantity were checked on 0.8% agarose gel. Libraries were constructed according to Kilian et al. (2012) DArTSeq complexity reduction method through digestion of genomic DNA using a combination of *Pst*I and *Mse*I enzymes and ligation of barcoded adapters and a common adapter, followed by PCR amplification of adapter-ligated fragments. Libraries were sequenced using single-read sequencing runs for 138 cycles. Next-generation sequencing was carried out using NovaseqX.

SEQART AFRICA uses genotyping by Sequencing (GBS) DArTseqTM technology, which provides rapid, high-quality and affordable genome profiling. DArTseq markers scoring was achieved using DArTsoft14, which is an in-house marker scoring pipeline based on algorithms. Two types of DArTseq markers were scored, SilicoDArT markers and SNP markers, which were both scored as binary for presence /absence (1 and 0, respectively) of the restriction

fragment with the marker sequence in genomic representation of the sample. Both SilicoDArT markers and SNP markers were aligned to the Vitpa_HiCP0_Assembly to identify chromosome positions.

Filtering of SNP markers

To ensure high data quality for future analyses, SNP markers underwent several treatments. Markers with a call rate below 0.8 (or less than 80%) were removed. Similarly, individuals with a missing data rate greater than 20% were also excluded. A reproducibility rate of 99% was maintained to ensure marker consistency across sample replicates, thereby confirming that the retained SNP markers were not artifacts or the result of sequencing errors. Additionally, a threshold of 0.05 for minor allele frequency (MAF) was applied to eliminate rare alleles that may not be informative for population genetics studies. All monomorphic alleles were removed, as they do not contribute to population differentiation, genetic diversity estimates, or association studies.

Linkage disequilibrium

Linkage disequilibrium (LD) refers to a non-random association between alleles at two different loci. High LD between alleles can influence the analysis of genetic structure. To address this, a pairwise LD correlation analysis was conducted to identify and remove highly correlated alleles. Consequently, only independent alleles exhibiting an LD of less than 0.2 were retained for the analysis (Kaeuffer et al., 2007).

Genetic diversity parameters

To describe genetic diversity within the shea tree population, four genetic parameters were used: (i) the expected average heterozygosity under the Hardy-Weinberg equilibrium hypothesis (HE), which provides information on the allelic richness of a population; (ii) the observed average heterozygosity (Ho), which quantifies the rate of heterozygotes in the population by comparison with HE; (iii) the inbreeding index (FIS), used to evaluate the rate of reproduction between related individuals leading to an excess of homozygotes; (iv) the percentage of polymorphic markers (PIC), which provides information on allele polymorphism. Analyses were carried out in R v4.4.2 using the dartR package.

Genetic differentiation and structure in shea tree populations

Genetic differentiation among populations and within groups was analyzed using Analysis of Molecular Variance (AMOVA), which is implemented in the poppr package in R (Kamvar et al., 2014). AMOVA quantifies the proportion of genetic variation that can be attributed to

differences within groups versus between groups (Excoffier et al., 1992). The genetic groups defined by STRUCTURE and DAPC were used as population units for AMOVA analysis. Population structure analysis was conducted to establish the genetic structure using the Bayesian method implemented in STRUCTURE v2.3.4 software. The Markov Chain Monte Carlo (MCMC) approach, utilizing a mixture model without LOCPRIOR information, was applied to estimate individual assignments to genetic groups. Each K value, ranging from 1 to 5, was tested 10 times, with a burn-in period of 10,000 iterations, followed by 100,000 MCMC iterations.

To determine the number of clusters (K) that best represents the genetic structure of the population, we used the ΔK method. This method evaluates the rate of change in the likelihood of the data for different K values. The results from STRUCTURE were then utilized to define genetic groups, which were subsequently analyzed using AMOVA (Porras-Hurtado et al., 2013; Pritchard et al., 2003).

Discriminant Analysis of Principal Components (DAPC) was conducted using the adegenet package in R to further explore population clustering. Unlike STRUCTURE, DAPC is a multivariate method that does not rely on the assumptions of Hardy-Weinberg equilibrium. It aims to maximize genetic differentiation between predefined groups while minimizing variation within each group, offering an alternative visualization of population structure (Jombart et al., 2010).

Fat content and Fatty acids profile determination

To conduct studies on the quantification of fat and fatty acid composition, a subsample was selected from the most representative genetic group based on population structure analysis. Over two years (2023 and 2024), agromorphological variables (see supplementary Table 1) were collected from these 25 individuals to investigate the correlation between biochemical and agromorphological variables. Total fat was extracted using petroleum ether as a solvent, following the Soxhlet® extraction technique (Nielsen and Carpenter, 2017).

The shea nuts were collected from the northern part of each tree and first shelled to extract the kernels. The kernels were then subjected to an initial drying process in an oven at 70°C for 72 h. After drying, the dry matter (DM) content was calculated using the following formula (AOAC, 1995) :

$$DM(\%) = \left(\frac{Mass\ of\ kernel\ after\ drying}{Mass\ of\ kernel\ before\ drying}\right) \times 100$$

The dried kernels were ground using a ball mill (RETSCH MM400, Retsch GmbH, Rheinische Straße 36, 42781 Haan, Germany) with the P4 program. The resulting crude powders were freeze-dried before being placed into cellulose cartridges measuring 33 mm x 94 mm for total fat extraction. A triplicate of 3 g of powder from each sample was put into cellulose extraction cartridges lined with glass wool, which were then inserted into the extractor. For the extraction process, a pre-weighed and dried 500 ml flask containing 350 ml of petroleum ether was used to recover the fat during each 4-hour extraction cycle. Once the extraction was complete, the solvent was evaporated at a pressure of 800 mbar using a Heidolph Laborota 4000 rotavapor (Heidolph Instruments GmbH & Co. KG, Walpersdorfer Straße 12, D-91126 Schwabach Bavière, Allemagne). Finally, the flask containing the extracted lipids was dried in an oven at 103°C for approximately 5 min.

The fat content was calculated by finding the mass difference between the empty flask and the flask containing the lipid after extraction, divided by the mass of the sample. The formula used is as follows (AOAC, 1995):

The mass of fat is given by:

$$Fat\ wet\ (\%) = [(w2 - w1)/w0] * 100 \text{ with;}$$

Fat wet = percentage of fat on the wet basis;
w2 = weight of flask with lipids after extraction (g);
w1 = weight of empty flask before extraction (g);
w0= sample weight (g)

To calculate fat content on a dry basis, the following formula was used:

$$Fat\ dry\ (\%) = (Fat\ wet/DM) * 100 \text{ with;}$$

Fat wet = percentage of fat on wet basis;
DM = dry matter content of the sample.

After gravimetric quantification, 100 mg of fats were esterified using the Ackman procedure to produce fatty acid methyl esters (FAMEs) needed for gas chromatography (GC) (Ackman, 1998; Ackman and Sipos, 1964; Morrison and Smith, 1964). 1 mL (1mg/mL) of nonadecanoic standard (C19:0) acid internal was added during FAMEs preparation and evaporated with N_2

flow to determine the concentration of each fatty acid. After preparing the fatty acid methyl esters (FAMEs), the fatty acid (FA) composition was analyzed using gas chromatography (GC) with flame ionization detection (FID) to measure and quantify the fatty acids.

The FAMEs were injected directly into the GC, with hydrogen as the carrier gas at a flow rate of 40 mL/min. The injector temperature was set to 120°C, while the FID temperature was maintained at 250°C. The temperature program was as follows: it was ramped from 120°C to 180°C over two minutes, and then held at 220°C for 25 min. The instrument used for this analysis was a SHIMADZU GC-2010, equipped with an AOC-20i autosampler. The column specifications included a length of 60 m, an internal diameter of 0.25 mm, a film thickness of 0.20 µm, and a maximum column temperature of 275°C. The total program execution time was 43.83 min. Fatty acids were identified using retention times compared to commercially available standards (C19:0). The following formula was used to calculate the concentration (AOAC, 1999) :

$$FA\ (mg/g) = [(A_X * W_{IS} * CF_X)/(A_{IS} * W_S * 1.04)] * 1000 \quad \text{with;}$$

A_X= area counts of the FAMEs; A_{IS} = area counts of internal standard (IS); CF_X = theoretical detector correction factor for FAMEs; W_{IS} = weight of IS added to the sample (mg); W_S = sample weight (mg); and 1.04 is factor necessary to express result as mg fatty acid/g oil

The lipid content and composition of the 25 subsamples were analyzed using descriptive statistics. To examine the relationship between agromorphological parameters and biochemical traits, specifically fat content and fatty acids, Pearson's correlation test was applied. All analyses were conducted using R software version 4.4.2.

Results

SNP selection and genetic diversity analysis in shea tree populations

SNP filtering retained 3,720 SNP markers out of a total of 61,885 SNPs, distributed unevenly across the 12 chromosomes (see Supplementary Figure S1a). The highest number of SNP markers was found on chromosome 2, which had a total of 495 SNP markers (Table 2). The marker density ranged from 210.37 to 145.81 Kbp, with an average of 177.07 Kbp, indicating that an SNP marker is encountered approximately every ≈ 177.1 Kbp in the genome. The highest marker densities were observed on chromosomes 6 and 2 (see Supplementary Figure S1b). Chromosome 1 had the longest total length (Table 2), which may be due to its role in the conservation and protection of DNA (heterochromatin) or through evolutionary processes

causing chromosomal rearrangements (Alberts et al., 2002). Transition mutations are the most common type, with a total of 2,115 single-nucleotide polymorphisms (SNPs), which represents 56.85% of all SNPs. These transitions are nearly evenly distributed among their variants: 28.49% involve the reciprocal substitution between bases C and T, while 28.36% involve the A and G variant (see Supplementary Figure S1c). In contrast, transversions comprise 1,605 SNPs, accounting for 43.1% of the total. The distribution of transversion variants is as follows: A to T accounts for 12.3%, C to G for 11.4%, A to C for 9.8%, and G to T for 9.6%.

Table 2: Distribution of the 3720 SNP markers retained after filtering and their respective densities across the 12 *Vitellaria paradoxa* chromosomes mapped.

Var1	SNP count	%SNPs	Density (Kpb)	Length (bp)
Chr01	392	10.54	205.15	80,417,609
Chr02	495	13.31	149.63	74,066,213
Chr03	336	9.03	170.6	57,320,494
Chr04	282	7.58	210.37	59,323,248
Chr05	291	7.82	202.42	58,903,358
Chr06	335	9.01	145.81	48,846,170
Chr07	296	7.96	186.17	55,105,912
Chr08	328	8.82	158.56	52,006,606
Chr09	248	6.67	188.6	46,772,806
Chr10	251	6.75	185.12	46,465,377
Chr11	234	6.29	162.82	38,100,608
Chr12	232	6.24	159.56	37,017,589
Mean	310	8.33	177.07	
Total	3720	100		

The observed heterozygosity rate (Ho) was 0.19, while the expected heterozygosity rate (He) was slightly higher at 0.21, resulting in an inbreeding index (Fis) of 0.09. Markers that exhibited low polymorphism and provided less information had an average Polymorphism Information Content (PIC) of 0.20, comprising 1,540, or 41.4%, of the total SNPs. In contrast, polymorphic and informative markers had a PIC that ranged from 0.25 to 0.5, with an average of 0.38, accounting for 2,180, or 58.6%, of the total SNP markers. Overall, the mean PIC for all markers was 0.30.

Genetic differentiation and population structure in shea tree

Molecular Analysis of Variance (AMOVA) revealed that 88.4% of genetic variation occurs within individuals, while 7.97% is attributable to differences among populations. Even less variation, at 3.67%, is due to differences between individuals within populations (Table 3). To assess genetic differentiation at different levels, we calculated Phi-statistics. The results indicate a moderate genetic differentiation among all individuals, estimated at around 11.6%, with approximately 8% of this differentiation explained by population structure. This finding suggests that there is limited but notable genetic structuring among the populations (see Table 3). The significance of the observed AMOVA results was evaluated using the "randtest" function from the "poppr" package (Table 4).

Table 3: Analysis of Molecular Variance (AMOVA) among the 167-shea tree, grouped from STRUCTURE analysis results

Source of Variation	Df	Sum Sq	Mean Sq	Variance	%	Phi
Among pops	2	4517.433	2258.716	59.071	7.97	0.797
Within populations	164	116344.6	709.418	27.192	3.67	0.0399
Within Individuals	167	109390.8	655.034	655.034	88.36	0.1164
Total	333	230252.8	691.45	741.298	100	

The standard deviation of genetic variation observed within individual shea trees showed a low negative value (Std.Obs = -4.30466) with a p-value of 0.01 ($p < 0.05$). This result suggests that the genetic differentiation within individuals was significantly lower than expected (see Table 4). Conversely, when examining individuals within the same population, we observed a positive but low standard deviation (Std.Obs = 2.213975) with a p-value of 0.02 ($p < 0.05$). This indicates that genetic differentiation between individuals within populations was significantly greater than expected, highlighting moderate genetic differences among individuals within the same population (see Table 4).

Furthermore, genetic differentiation among populations was highly significant and much greater than expected, with a standard deviation of 49.31827 ($p < 0.05$). This finding showed clear genetic structuring and differentiation between populations (see Table 4).

Table 4: Significance of AMOVA Permutation Test for Genetic Differentiation of shea tree genetic groups

Variation	Observed	Std.Obs	Alternative	P-value
Within individuals	655.0344	-4.30466	less	0.01
Within populations	27.19189	2.213975	greater	0.02
Among pops	59.07141	49.31827	greater	0.01

The genetic structure of shea tree populations was analyzed using STRUCTURE v2.3.4, revealing three distinct genetic groups, labeled Pop1, Pop2, and Pop3, represented by the colors blue, red, and green, respectively (Figure 2). In this analysis, a less negative log-likelihood value (LnProb) indicates a better model fit, while a similarity score closer to 1 reflects more stable assignments of individuals (Evanno et al., 2005; Pritchard et al., 2000).

For K = 3, the model demonstrated a strong fit with a significantly improved log-likelihood of LnProb = - 431,259.5 and a high similarity score of 0.994, suggesting stable and reliable groupings of individuals. In comparison, K = 2 resulted in a poorer fit (LnProb = - 435,678.3) and less stable assignments of individuals (similarity ≈ 0.93). Although K = 4 produced a slightly better log-likelihood (LnProb = -429,536.8), the similarity score decreased to 0.987, indicating reduced reliability in assignments compared to K = 3. Therefore, K = 3 provided the best balance between model fit and stability of clustering (see Figure 2).

In Supplementary Figure S2 a, the log likelihood for K (LnP(K)) showed a peak at K = 4, followed by a downward trend after this point. According to Evanno et al. (2005), LnP(K) indicates stabilization or a decrease in the mean value of LnP(K) when the optimal number of clusters (K) is reached, which justified the selection of K = 3. Furthermore, the use of deltaK helped prevent the selection of an inappropriate number of K (Evanno et al., 2005), whether too high or too low, by considering the increase in log likelihood relative to the rise in K (Supplementary Figure S2 b). The determination of the optimal number of clusters was further supported by the deltaK value ($\Delta K = 65.513$), which represents the difference in log likelihood between K and K+1 (Table 5).

Cluster 1 (Pop1) consisted of 153 individuals, with 149 of them sharing genes with Cluster 2 (Pop2), indicating that 97.4% of the individuals were admixed. Cluster 2 included eight (08) individuals, each exhibiting a diverse allelic composition; all these individuals shared genes from at least two different ancestors. A similar trend was observed in Cluster 3 (Pop3), which consisted of six (06) individuals, all of which were mixed. The divergence among alleles between populations, measured as net nucleotide distance, was 0.1228 between Pop1 and Pop2,

0.2953 between Pop1 and Pop3 (the highest), and 0.1894 between Pop2 and Pop3. In contrast to Cluster 2, which has an average heterozygosity distance of 0.4084, the average distances between individuals in Cluster 1 were 0.0567 and 0.0572 in Cluster 3, making the individuals within these populations more genetically similar. The fixation index (Fst) quantifies the proportion of genetic variation attributed to differences between populations relative to the total genetic variation of the species. The highest Fst values were observed in Cluster 3 (Fst = 0.9809) and Cluster 1 (Fst = 0.8557), while Cluster 2 recorded a very low value (Fst = 0.0029).

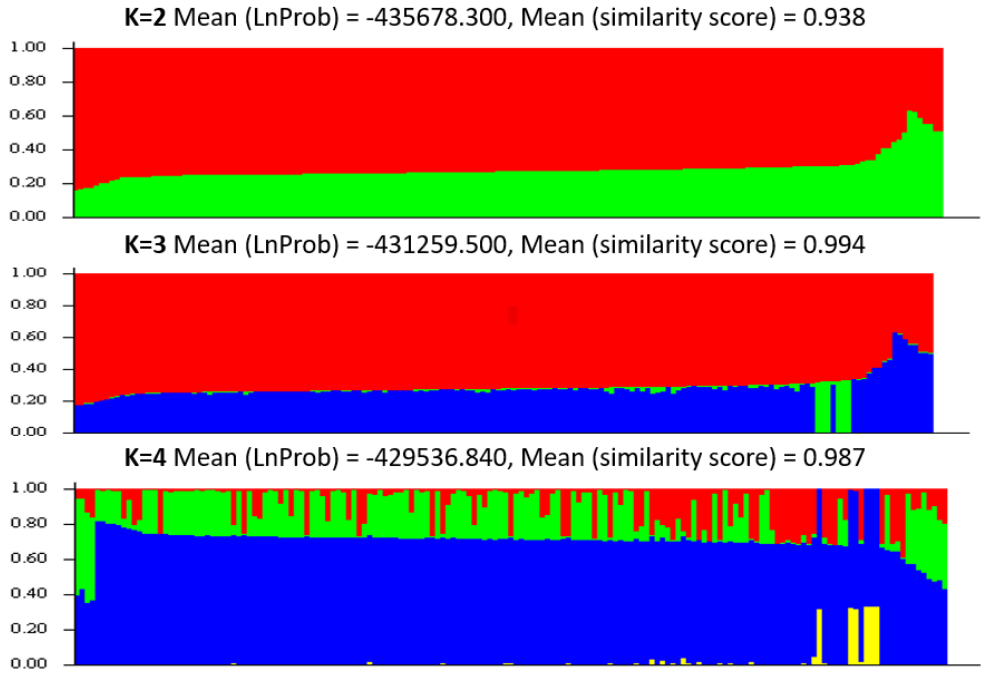


Figure 2: Barplot resulting from the genetic structuring of 167 *Vitellaria paradoxa* individuals sorted by Q matrix and showing each possible K clustering with similarity score and LnProb values

Table 5: Value of LnP(K) and deltaK for each probable value of K derived from the results of genetic structuring of the 167 individuals of *Vitellaria paradoxa*.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln"(K)|	Delta K
1	10	-438757	5.635	NA	NA	NA
2	10	-435678	348.663	3078.88	1339.92	3.843
3	10	-431260	41.713	4418.8	2732.72	65.513
4	10	-429573	1738.938	1686.08	2634.54	1.515
5	10	-430522	3615.009	-948.46	NA	NA

Population 1 (Pop 1) was predominantly made up of shea trees from the Gogounou and Segbana regions, with proportions of 16.34% and 15.0%, respectively. Following these were the

moderately represented individuals from the Pehunco and Tchaourou regions, contributing 13.73% and 13.1% respectively (Figure 3). Additionally, there were individuals with lower representation from the Parakou (11.1%), Kandi (9.1%), Boukoumbé (8.5%), Bembèrèkè (7.8%), and N'dali (5.2%) regions. Population 2 (Pop 2) consisted equally of individuals from the Gogounou and Segbana regions, each accounting for 25%. In contrast, there was a minority representation (12.5% each) from the regions of Kandi, N'dali, Parakou, and Tchaourou (Figure 3). Population 3 (Pop 3) was exclusively characterized by shea trees sourced from Kandi (Figure 3).

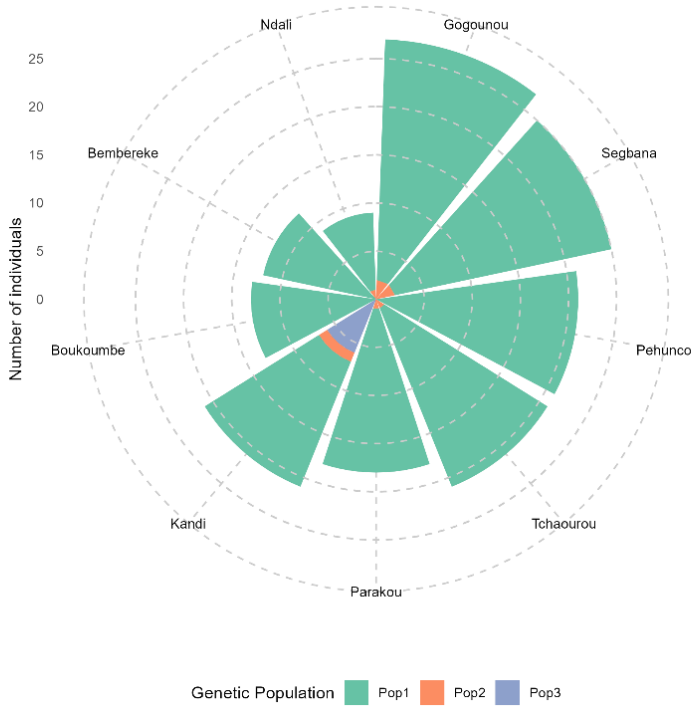


Figure 3: Distribution of shea trees based on their origin in the three clusters identified from the genetic analysis of 167 genotyped individuals.

Discriminant principal component analysis (DAPC) also revealed three (03) genetic clusters, confirming the STRUCTURE results. The optimal number of clusters (K = 3) was obtained using the Bayesian Information Criterion (BIC = 544.3486), with the BIC value first slowing down and then stabilizing slightly (Supplementary Figure S3a). Four (04) principal components (PC1, PC2, PC3 and PC4) were selected for discriminant analysis (Supplementary Figure S3b).

The number of individuals per cluster obtained from the DAPC (Pop1 = 150, Pop2 = 11, and Pop3 = 6) was very close to that obtained from the STRUCTURE v.2.3.4 (Pop1 = 153, Pop2 = 8, and Pop3 = 6) analysis, showing a clear separation between the three (03) populations (Figure

4). These results confirmed those obtained by STRUCTURE, which suggested three genetic groups with admixed individuals. Membership probability patterns was presented in the Supplementary Figure S4, showing the admixture individuals among populations. Supplementary Figure S5 displays the assignment of each individual to their respective genetic group.

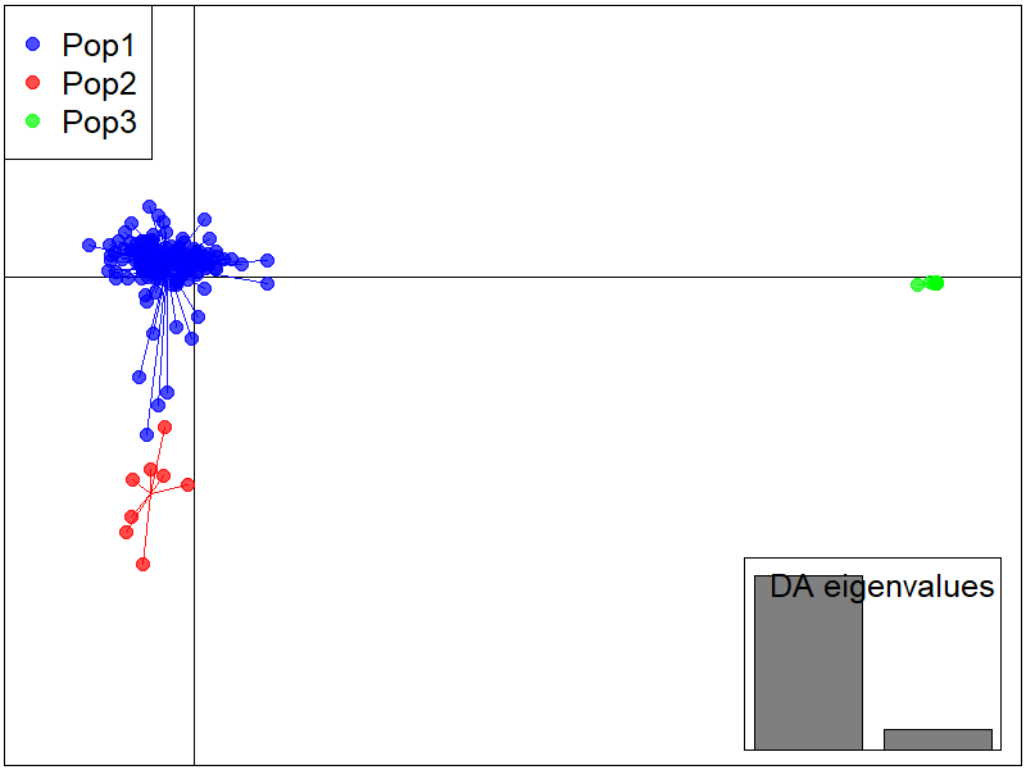


Figure 4: Principal component discriminant analysis of 167 Shea tree individuals grouped into three distinct genetic clusters.

Elite shea tree selection based on Fat content, Fatty acid composition and Yield

Descriptive statistics on agromorphological parameters (Table 6) and total fat content measured on a dry basis were performed from the 25 subsamples of shea tree genotypes of Pop1.

Table 6: Descriptive statistics on the 25 shea tree individuals from population 1 (Pop1)

Traits	Mean	Stdev	min	max	CV (%)
Fat content (%)	64.29	11.33	34.74	82.24	18
Height (m)	11.49	2.88	6.8	17.45	25
Cirf: Circumference (m)	1.36	0.72	0.4	2.9	53
HFW: Hundred Fruit Weight (g)	297.79	94.26	122.48	508.75	32
FL: Fruit Length (cm)	4.07	0.58	2.96	5.088	14
FW: Fruit Width (cm)	3.26	0.45	2.32	4.031	14
HNW: Hundred Nut Weight (g)	136.93	41.99	65.81	262.02	31
NL: Nut Length (cm)	3.2	0.41	2.32	4.163	13
NW: Nut Width (cm)	2.31	0.29	1.79	2.947	13
Yield (kg/year)	24.91	16.68	2.56	58.30	67

The analysis revealed an average fat content of 64.3 ± 11%. The genotype identified as exhibiting the lowest fat content (34.7%) is BG143, while the highest fat content was recorded in BG196 (82.2%). Additionally, an analysis of the FA composition identified five main FAs in various proportions (Table 7). The majority of fatty acids identified was saturated (SFA = 50.29%) and are divided into three types: palmitic acid (C16:0), stearic acid (C18:0) and arachidic acid (C20:0). A single monounsaturated fatty (MUFA) acid, oleic acid (C18:1n9), was observed, as well as a single polyunsaturated fatty acid (PUFA), linoleic acid (C18:2n6).

Table 7: Concentration and proportion of the five fatty acids among the 25 shea tree samples

Fatty acids	%	StDev	min	max
Palmitic acid (C16:0)	3.60	0.6	3	4.67
Stearic acid (C18:0)	45.65	4.9	39.33	59.67
Arachidic acid (C20:0)	1.04	0.11	1	1.33
SFA	50.29	-	-	-
Oleic acid (C18:1n9)	43.73	3.51	38	50
MUFA	43.73	-	-	-
Linoleic acid (C18:2n6)	5.98	1.28	4.5	9
PUFA	5.98	-	-	-

SFA= Saturated Fatty acid, MUFA = Monounsaturated Fatty acid, PUFA= Polyunsaturated Fatty acid

Five genotypes displayed an interesting profile regarding FA composition: BG109, BG104, BG214, BG258, and BG114. These genotypes exhibited high concentrations of the five fatty acids, although in varying concentrations and proportions (Table 8).

Table 8: Fatty acid composition of the top five (05) genotypes with high fatty acid potential from the 25 shea tree samples analyzed.

Genotypes	Unit	C16:0	C18:0	C20:0	C18:1n9	C18:2n6
BG109	mg.g	19.14±2.08	407.64±22.27	9.71±0.75	212.06±19.77	34.72±3.75
	%	3	59.67±0.58	1	31.33±0.58	5
BG104	mg.g	20.26±0.08	354.46±2.71	9.29±0.32	277.4±2.4	38.18±1.44
	%	3	50.67±0.58	1	40	5.33±0.58
BG214	mg.g	18.88±0.72	338.12±1.38	8.35±0.08	268.77±7.78	37.18±0.84
	%	3	50.67±0.58	1	40.33±0.58	6
BG258	mg.g	24.64±1.41	333.5±2.12	11.26±1.41	319.05±4.95	47.6±1.41
	%	3	45	1	44	6
BG114	mg.g	11.7±0.34	176.41±4.81	4.92±0.1	139.97±4.77	28.45±4.61
	%	3	48.67±1.15	1	39	7.67±0.58

In terms of fruit yield, the genotypes BG117 (58.30 kg), BG146 (53.88 kg), and BG078 (53.84 kg) performed the best. However, they exhibited low fatty acid (FA) concentration and total fat content, except BG117, which had a high FA content of 81.9% (Figure 5).

Regarding total fat content, genotypes BG196, BG117, BG214, and BG104 showed high fat content values of 82.2%, 81.9%, 80.3%, and 80.1%, respectively, making them the top performers in total lipid content (Figure 5). Among these previous genotypes, only BG117 and BG104 stood out with notably high yield values. The genotype with the best overall profile regarding fat content, FA composition, and yield was BG104, which had a fat content of 80.11%, a high FA composition (as shown in Table 8), and a moderate yield of 41.49 kg, which is above the average of 24.9 kg ± 16.7 (Figures 5 and 6).

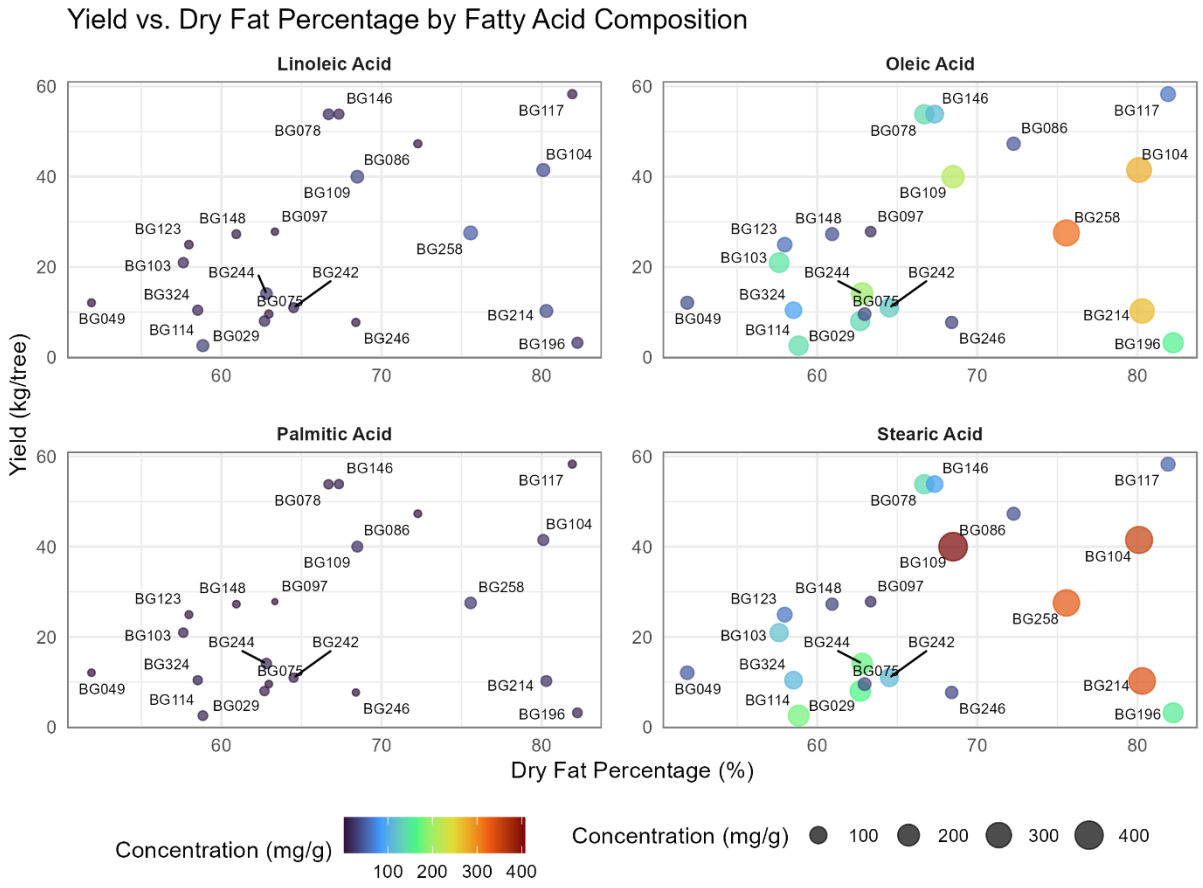


Figure 5: Fat content as a function of yield and concentration of shea tree genotypes by fatty acid type.

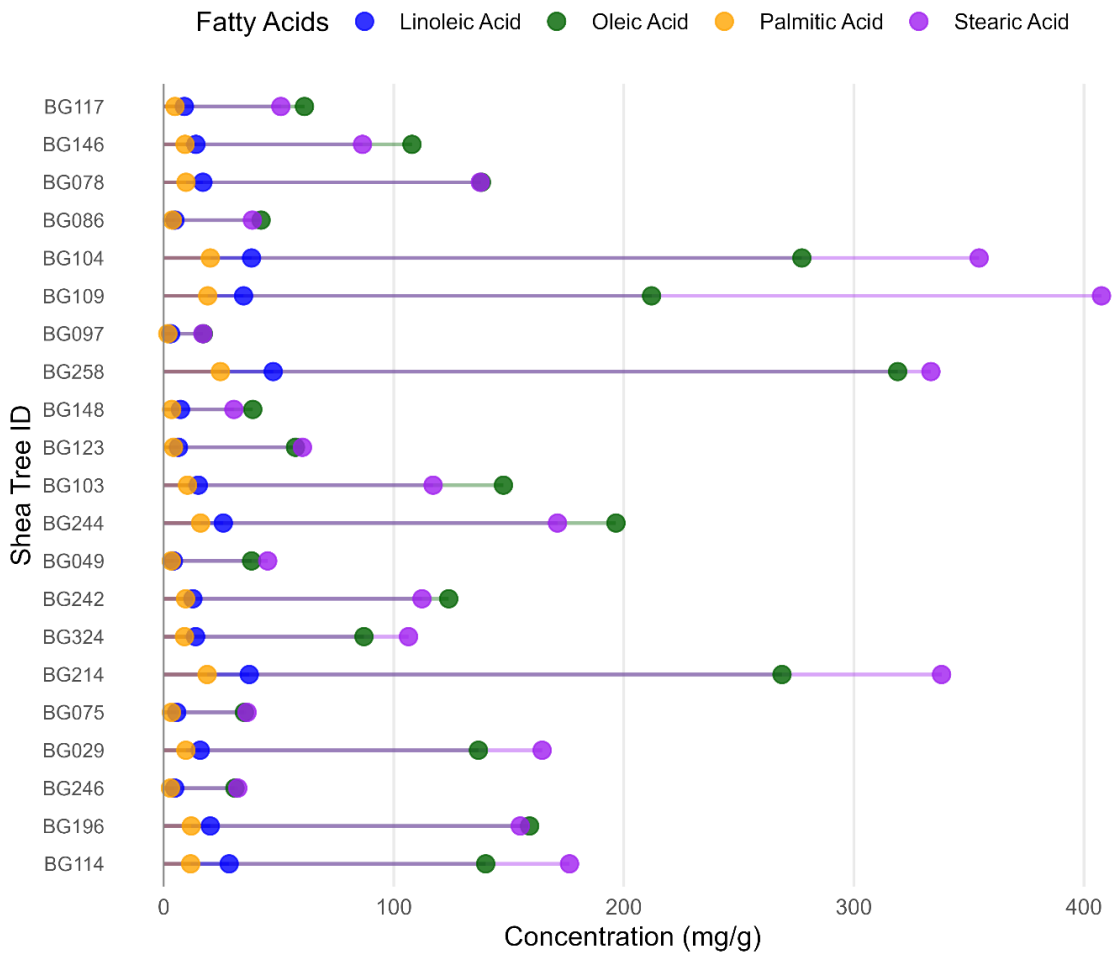


Figure 6**:** Four major fatty acid compositions of 25 genotypes of shea tree ordered by fruit Yield (kg/ha)

Phenotypic correlation among agromorphological and biochemical traits in shea tree

The correlation matrix between agromorphological and biochemical traits, created using the Pearson method, highlighted the linear relationships among various variables. Figure 7 illustrates the only significant correlation coefficients ($p<0.05$) between dendrometric, agronomic, and biochemical parameters. All observed significant correlations were positive, indicating a unidirectional linear relationship among the variables. The analysis revealed that fatty acids (FAs) were very strongly correlated with one another, with coefficients ranging from 0.83*** to 0.99*** (Figure 7). The fat content) exhibited a strong correlation with Hundred Nut Weight (HNW) (r = 0.71***), Nut Weight (NW) (r = 0.64**), and Nut Length (NL) (r = 0.62**). Furthermore, there was a moderate correlation between fat content and arachidic acid (r = 0.43*) and linoleic acid (r = 0.39*). Notably, no significant correlation was found between yield and fat content. The only variable with a moderate correlation to yield was fruit length (FL) (r = 0.47**).

In terms of dendrometric parameters, the only variable that had a strong correlation with tree height was circumference (r = 0.74***), which was also moderately correlated with Hundred Fruit Weight (HFW) (r = 0.39*). Very strong, significant correlations were observed between yield components, except NW and FL. Thus, NL was very strongly correlated with NW (r = 0.8***), HNW (r = 0.72***); strongly correlated with FW (r = 0.62**), FL (r = 0.61**), HFW (r = 0. 55**); and showed moderate but significant correlations with fatty acids C20:0 (r = 0.5**), C18:0 (r = 0.43*), C18:2n6 (r = 0.43*), C18:1n9 (r = 0.43*), C16:0 (r = 0.39*). HNW was very strongly correlated with NW (r = 0.81 ***), strongly correlated with HFW (r = 0.55**), and moderately correlated with FL (r = 0.39*), FW (r = 0.45*). It also showed moderate correlations with some FAs, including: C18:0 (r = 0.39*), C18:2n6 (r = 0.45*), and C18:1n9 (r = 0.41*). NW also showed moderate to strong correlations for fatty acids C20:0 (r = 0.48*), C18:0 (r = 0.48*), C18:2n6 (r = 0.52**), C18:1n9 (r = 0.5**), C16:0 (r = 0.45*); also, with HFW (r = 0.47*) and FW (r = 0.48*). Between the fruit parameters, very strong correlations were recorded, with HFW correlating very strongly with FL (r = 0.73***) and FW (r = 0.78***), which then correlated with each other with a coefficient of r = 0.86***. FW showed a moderate correlation with C20:0 (r = 0.41*) (Figure 7).



Figure 7: Phenotypic correlation matrix between two years (2023-2024) of agromorphological data collection and biochemical variables of shea tree genotypes.
Cirf = Circumference, Fruit Width (cm) = FW, Fruit Length (cm) = FL, Nut Width (cm) = NW, Nut Length (cm) = NL, Hundred Fruit Weight (g) = HFW, Hundred Nut Weight (g) = HNW, dry_Fat (%) = Fat content, Aa (%) = Arachidic acid (C20:0), Sa (%) = Stearic acid (C18:0), La (%) = Linoleic acid (C18:0), Oa (%) = Oleic acid (C18:0) and Pa (%) = Palmitic acid (C16:0).

Discussion

SNP selection and genetic diversity analysis in Benin's shea trees

The number of SNP markers retained in this study (3,720 SNPs out of 61,885) was lower than that used by Attikora et al. (2024), who reported 7,559 SNPs out of a total of 42,705 after filtering, in their study of genetic diversity and population structure of shea tree accessions in Côte d'Ivoire. Unlike the filtering parameters of Attikora et al. (2024), which include CallRate (>0.80), MAF (>5%) and PIC (≥0.1), the present study incorporated, in addition to these filtering parameters, a CallRate of 80% on individuals (≥80), a marker reproducibility rate of 99% and a linkage disequilibrium of 0.2 (LD<0.2). The filtering parameters were therefore quite rigorous and optimized in the present study to reduce artefacts introduced during library preparation and bioinformatic processing (Andrews et al., 2016; O'Leary et al., 2018; Shafer et al., 2017). The highest number of SNP markers recorded on chromosome 2 is equivalent to the same results reported by Odoi et al. (2023) and Attikora et al. (2024). Unlike Attikora et al. (2024), who reported a transition/tranversion ratio of 1.55, a transition/tranversion ratio of 1.34 was reported by Odoi et al. (2023), in line with the 1.31 reported in the present study. This slight discrepancy observed could be related among other things to: sequencing depth and quality, since, low sequencing depth can increase the number of false-positive SNPs, particularly transversions; the sample size and genetic diversity (a higher genetic diversity in the sampled population can introduce more transversions due to random mutations), and data filtering criteria (stringency in filtering SNPs for quality metrics) (Boussaha et al., 2012; Nielsen et al., 2011; Pirooznia et al., 2014).

The heterozygosity rate, being the proportion of individuals in a population that are heterozygous for a given locus, the observed heterozygosity rate value Ho = 0.19 indicates that on average 19% of loci are heterozygous among the 167 shea tree individuals sampled, which was smaller than the 21% expected (He = 0.21). This findings suggests that there is a possible deficit in heterozygotes compared with what is expected at Hardy Weinberg equilibrium, which could be explained by inbreeding or selection effects in our population samples (Campbell,

2001; Tsetsos et al., 2019). Although the shea tree primarily reproduces through cross-pollination, it still exhibits a notable rate of self-pollination, approximately 23% (Nasare et al., 2019). This self-pollination serves as an adaptive strategy, providing a viable alternative when faced with biotic or abiotic challenges, such as a lack of pollinators, high temperatures that impact pistil viability, or strong winds (Lassen et al., 2018; Nasare et al., 2019; Okullo, 2004). Odoi et al. (2023) reported an identical heterozygosity rate (He = 0.21), confirming the findings of the current study (He = 0.21). Moreover, Lovett and Haq (2000) in Ghana, reported also identical results (He = 0.2142), even though their study was based on the use of isoenzyme to illustrate genetic diversity within 294 shea tree accessions. Attikora et al. (2024) found in their study a He = 0.26, which provides similar information on the genetic diversity, supporting our results. The various perspectives presented are closely tied to our findings and emphasize the moderate genetic diversity among shea trees in Benin. This offers promising opportunities for future improvement of the species. The positive FIS value of 0.09, along with the observed heterozygosity (Ho) of 0.17 being lower than the expected heterozygosity (He) of 0.21, indicates a certain degree of inbreeding within these trees. This may be attributed to human activities (Koffi et al., 2017; Pienaah et al., 2024) that impact the survival and distribution of the species. Factors such as selection pressures, bushfires, monoculture, deforestation, and pesticide use may reduce the diversity among shea trees by producing a sub-population. Reproduction among individuals within these sub-populations can lead to increased genetic homogeneity, contributing to a higher F_{IS} value. Additionally, environmental conditions may affect the trees to self-pollinate, which raises the rate of homozygosity (Stout et al., 2018). The collection areas predominantly exhibit a Sudanian climate, characterized by high temperatures, irregular rainfall, and strong winds—factors that can significantly influence the reproductive mode of the trees (Nasare et al., 2019).

Genetic differentiation and population structure

The shea population showed three distinct genetic groups. The highest ΔK (65.513) was observed at K = 3) with evidence of admixture, meaning that there is gene flow between populations. Clustering was confirmed by DAPC analysis with some slight changes in the composition of individuals by genetic groups. Similar results were reported by Larson et al. (2021) in their study to examine the genetic structure of *Eschweilera* species in the Lecythidaceae family, which were important cross-pollinating trees in the Amazon rainforest. Three distinct genetic groups were also identified among *Eschweilera* populations, with high levels of admixture, and individuals sharing genetic material between groups. Similar results

were reported by Attikora et al. (2024) and Anyomi et al. (2023), who identified a high proportion of individuals sharing gene flow between the three genetic populations they obtained. These studies underline the importance of considering admixture in understanding the evolutionary history and conservation of tropical tree species (Larson et al., 2021; Peñaloza-Ramírez et al., 2010). The differentiation index (FST) obtained for Populations 1 and 3 (FST values of 0.8557 and 0.9809, respectively) suggests that these two populations are highly differentiated. By contrast, Pop 2 (FST = 0.0029) showed minimal differentiation, indicating significant gene flow with other groups, especially with Pop1. Pop2 acts as a genetic bridge, facilitating gene flow between Pop1 and Pop3, which is confirmed by the value of its net nucleotide distance obtained (Pop2 vs Pop3 = 0.1894, intermediate divergence; Pop1 vs Pop2: 0.1228 = low divergence), in contrast to that of Pop1 vs Pop2, which was equal to 0.2953.

These results support our hypothesis that significant genetic differentiation exists among individual shea trees, enabling their classification into distinct genetic groups. Additionally, describing the agromorphological traits of individuals from these populations is crucial for assessing the performance of trees within these groups.

AMOVA analysis, supported by a permutation test (randtest), revealed that the observed genetic structuring was significant at all levels (among populations, between individuals in the same population, and within individuals). The variance between populations (7.97%; p = 0.01) suggests moderate and significant genetic differentiation, probably linked to geographical or ecological barriers limiting gene flow (all the individuals of Pop3 are from Kandi only), thus validating the groups identified by the STRUCTURE analysis. Within population diversity (3.67%; p = 0.02) is low but remains significant, indicating a certain degree of local polymorphism, which can be explained by the wide variations observed in the yield of the 25 individuals sub-sampled from population 1, for example (24.91 ± 17 with a coefficient of variation of 67%). Finally, the intra-individual variance, although a majority (88.36%), is significantly lower than expected (p = 0.01), suggesting a potential linkage disequilibrium or allelic homogeneity, which could be due to the tree's semi-domesticated state (Sanou et al., 2005). The AMOVA results support the clustering results, emphasizing the molecular differences among the three identified shea groups.

Elite shea tree based on fat content, fatty acid composition and yield

The 25 shea tree genotypes showed an average fat content on a dry basis of 64.3% ± 11, which was around 10% higher than any recorded to date (Attikora et al., 2025; Honfo et al., 2017; Odoi et al., 2022; Ofoegbu-Chibuzo et al., 2022; Tano-Debrah and Ohta, 1994). Indeed, the dry

fat content depends on sample size, spatial distance between samples, extraction and drying method, and subspecies (Honfo et al., 2017; Odoi et al., 2022; Ofoegbu-Chibuzo et al., 2022). The fatty acid composition revealed (05) five principle types of fatty acid, including C16:0, C18:0, C20:0, C18:1n9 and C18:2n6, with stearic acid (C18:0) and oleic acid (C18:1n9) as the most represented. These results are in line with those reported by Di Vincenzo et al. (2005) and Oussou et al. (2022), who showed in their studies that samples from West Africa were richer in stearic and oleic acids, in contrast to those from East Africa. The fatty acid proportions per sample obtained in the present study (C16:0 ≈ 3.6%, C18:0 ≈ 45.7%, C20:0 ≈ 1%, C18:1n9 ≈ 43.7% and C18:2n6 ≈ 6 %) are similar to those reported in previous studies, particularly for C18:0 (Attikora et al., 2025). The proportions of C16:0, C20:0, and C18:2n9 are almost identical to those reported by Garti et al. (2019). A high proportion of oleic acid was noted, C18:1n9 ≈ 43.7%, in contrast to the values ranging from 29 to 36% often found in *Vitellaria paradoxa* subsp. *paradoxa* (Alale et al., 2024; Garti et al., 2019). The high values obtained for all the fatty acids in the present study are higher than those found in cocoa butter (Figueira and Luccas, 2022), and, therefore, constitute an alternative ingredient in chocolate formulation (Chakrabarty and Külheim, 2025). The observed dispersion in the average genotypes for fat content and fatty acid composition, despite being from the same gene pool (Pop1), suggests an environmental influence on the expression of genes responsible for fatty acid biosynthesis in shea (Baud and Lepiniec, 2010; Hale et al., 2021; Wei et al., 2019). A genotype x environment interaction study is recommended to assess the underlying bioclimatic parameters. The genotypes with the best yields for fatty acid composition were: BG109, BG104, BG214, BG258 and BG114. Their C18:0 concentrations ranged from 176.41 ± 4.81 mg/g (i.e. 48.67 ± 1.15%) to 407.64 ± 22.27 mg/g (i.e. 59.67 ± 0.58%) and for oleic acid, from 139.97 ± 4.77 mg/g (i.e. 39%) to 212.06 ± 19.77mg/g (i.e. 31.33 ± 0.58 %). Of the five (05) genotypes identified, BG104 from Bembèrèkè presented an interesting profile for fat content, fruit yield and fatty acid composition, and should be considered for further research programs. Parameters related to shea nuts (NL, NW and HNW) and fruit length (FL) are significantly positively correlated with fatty acid composition, highlighting the existence of a linear relationship between these parameters.

Conclusion and perspectives on future research

In this study, we highlighted the genetic diversity and structuring of Benin shea tree populations using SNP data and agromorphological variables. The results showed a moderate genetic diversity (He = 0.21). Three (03) genetic groups were obtained from the population structure analysis, showing significant genetic differentiation. Pop2 acts as a gene transfer bridge between Pop1 and Pop3. Fat content and fatty acid composition revealed five best genotypes: BG109, BG104, BG214, BG258 and BG114, all belonging to population 1 and coming respectively from four regions : Bembèrèkè, Pehunco, Gogounou, Gogounou and Kandi; divided into two agro-ecological zones (North cotton zone for genotypes BG214, BG258 and BG114; and South Borgou crop zone for genotypes BG109 and BG104). Genotype BG104 alone combines the best performance in terms of yield, fat content and fatty acid composition. The fatty acid profile of shea trees is strongly correlated with nut length, nut width and hundred nut weight. The larger the shea nuts, the higher the fatty acid concentrations. The results of the present study constitute a fundamental basis for the genetic improvement of the shea tree in Benin, based on the traits of interest, namely fat content, yield and fatty acid composition. A total of 25 genotypes were investigated for fat content and fatty acid composition, all from individuals in population 1 (Pop1). As the individuals of Pop1 and Pop3 are genetically distant, future research should focus on collecting agromorphological and biochemical data from Pop3 and Pop2. It should also extend sampling to all agroecological zones, and include data on secondary metabolites (phenols, tocopherols, triterpene, etc.) to capture most of the existing genetic diversity within Benin's shea tree populations. These studies are all the more important for establishing genomic prediction models for genomic selection. Future research should also focus on reducing the tree's production cycle by optimizing propagation methods, developing optimal fertilization plans and using plant growth promoting bacteria (PGPB) to stimulate rapid tree growth (essential for full domestication). In addition, the search for metabolites such as phenolics, tocopherols, phytosterols, triterpenes, etc. within the genetic groups identified in the present study could help identify elite genotypes recommended for the cosmetics industry. The study of the whole genome association (GWAS) highlighting genes linked to biochemical components would optimize targeted selection and a proven genetic gain.

Acknowledgements

Our gratitude to the GBioS Unit of the Faculty of Agricultural Sciences of the University of Abomey-Calavi in Benin for organizing and facilitating the supervision, data collection activities with us. We would like to thank the entire team of the Biomolecule Engineering Laboratory of the University of Lorraine (LIBio) for allowing us to carry out the biochemical experiments. Our sincere thanks to the entire team of the Agronomy and Environment Laboratory (LAE) of the University of Lorraine for the various technical and financial facilitations.

Author contributions

Guillaume Bodjrenou: Conceptualization, Investigation, Methodology, Data curation, Software, Formal analysis, Writing – original draft, Writing – review & editing. Matice Gbètoho: Investigation, Methodology, Data curation, Visualization. Fernande Honfo: Methodology, Supervision, Validation, Writing – review & editing. Cyril J.F. Kahn: Methodology, Supervision, Validation, Writing – review & editing. M. Eric Schranz: Methodology, Supervision, Validation, Writing – review & editing. Paul Césaire Gnanglè: Methodology, Supervision, Validation, Writing – review & editing. Alain Hehn: Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. Alexandre Olry: Investigation, Methodology, Supervision, Validation, Writing – review & editing. Enoch G. Achigan-Dako: Funding acquisition, Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Funding: This work was fully supported by the PEA-BIOVALOR project: *Renforcement de la démarche compétences et de la culture entrepreneuriale dans la formation supérieure agronomique : leviers pour l'insertion professionnelle et le développement de la bioéconomie au Bénin*, financed by the Agence Française de Développement (AFD) through the Agence Nationale de Recherche (ANR)

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request. All methods were performed in accordance with relevant guidelines and
regulations.

Declarations

Competing interests: The authors declare that they have no competing interests.

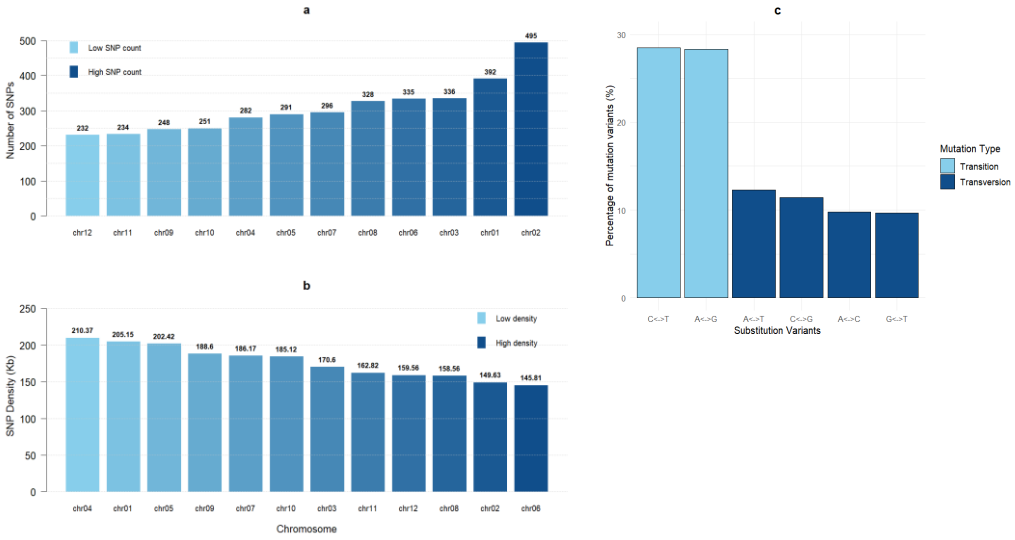
Supplementary Information

Supplementary Table S1: Morphological descriptors and measurement methods used to collect agromorphological data.

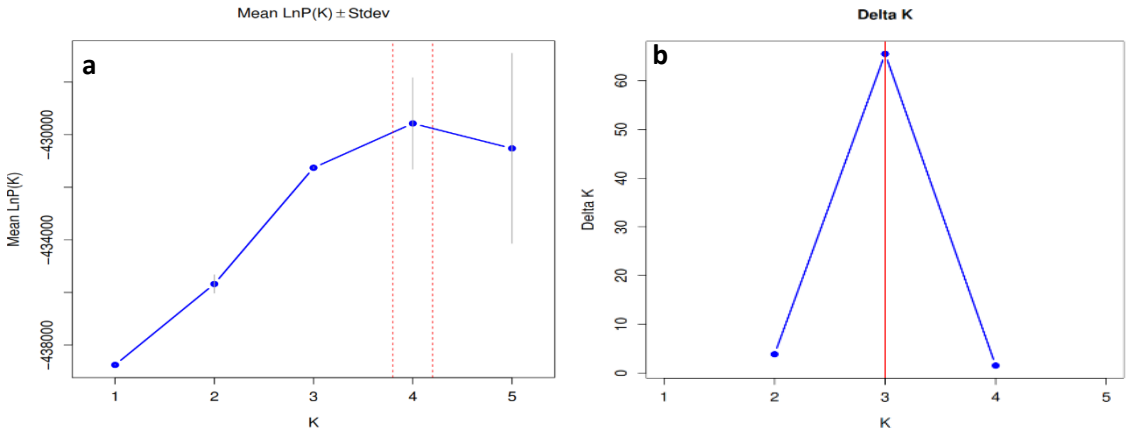
Category	Descriptors	Unit	Measurements
Dendrometric data	Circumference (Cirf)	m	Measure at a height of approximately. 1.40 m using a forestry tape.
	Height	m	Determine with a clinometer (Akplo et al., 2019)
Fruiting data	Hundred Fruits Weights (HFW)	g	Estimated using the mean value of ten fruit samples per tree. The estimated value is validated when the coefficient of variation is less than 5%.
	Hundred nut weight (HNW)	g	Estimated using the mean value of ten fruit samples per tree. The estimated value is validated when the coefficient of variation is less than 5%.
	Fruit length (FL)	cm	Mean of the measure recorded on five shea fruits
	Fruit width (FW)	cm	Mean of the measure recorded on five shea fruits
	Nut length (NL)	cm	Mean of the measure recorded on five shea nuts
	Nut width (NW)	cm	Mean of the measure recorded on five shea nuts
Productivity	Yield	Kg/year	Total fruit weight per tree by year of production (Kouyaté et al., 2006)

Supplementary Table S2: Value of LnP(K) and deltaK for each probable value of K derived from the results of genetic structuring of the 167 individuals of *Vitellaria paradoxa*.

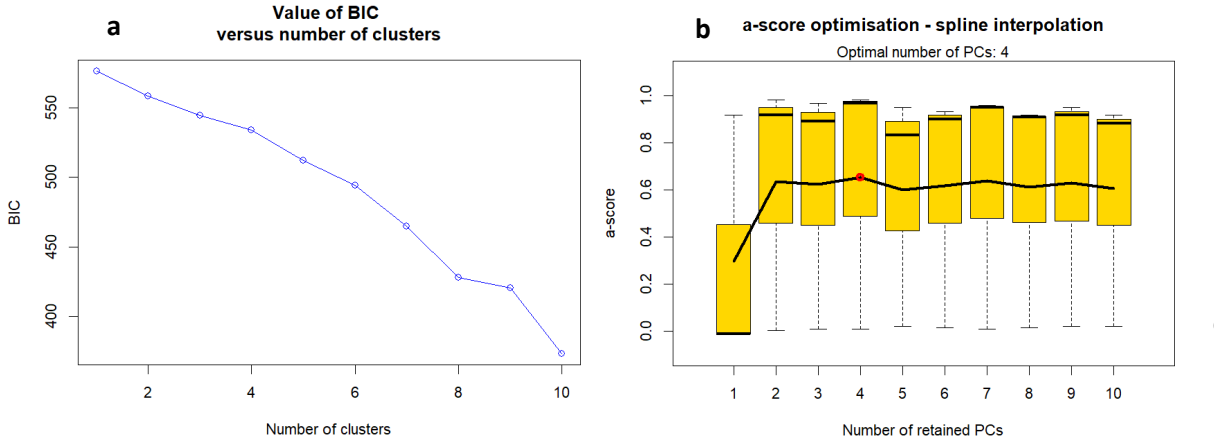
K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	10	-438757	5.635	NA	NA	NA
2	10	-435678	348.663	3078.88	1339.92	3.843
3	10	-431260	41.713	4418.8	2732.72	65.513
4	10	-429573	1738.938	1686.08	2634.54	1.515
5	10	-430522	3615.009	-948.46	NA	NA



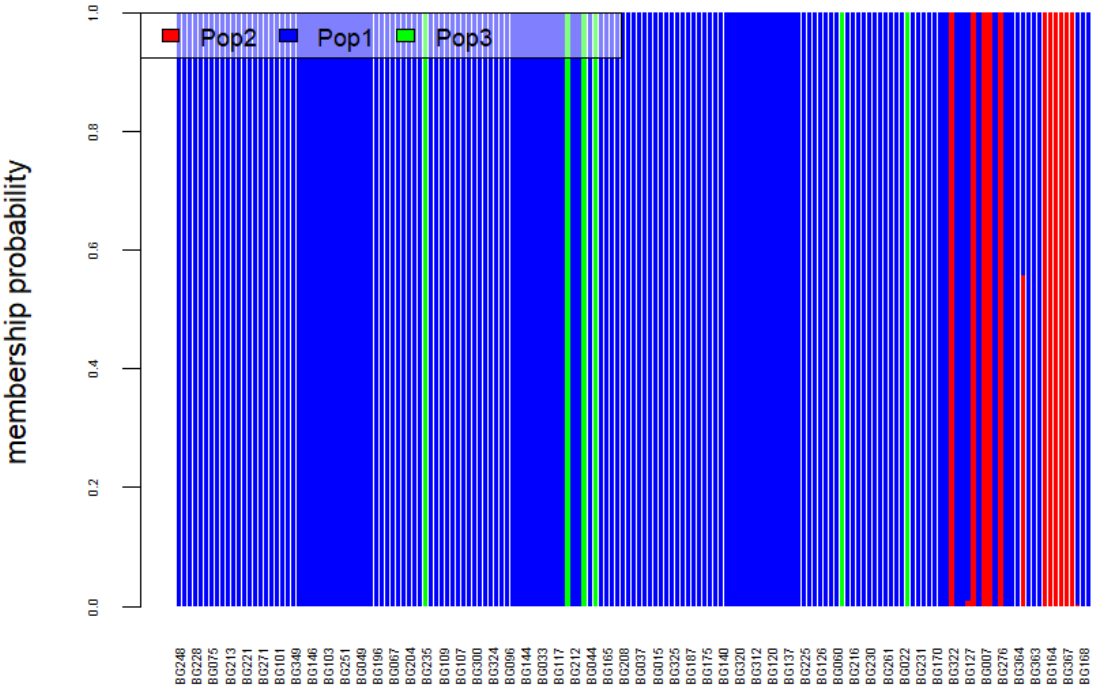
Supplementary Figure S1: **a** Number of SNP markers observed on the 12 V. paradoxa chromosomes after filtering; **b** Marker density per chromosome; **c** Proportion of different mutation types (transition and transversion) recorded in the genome.



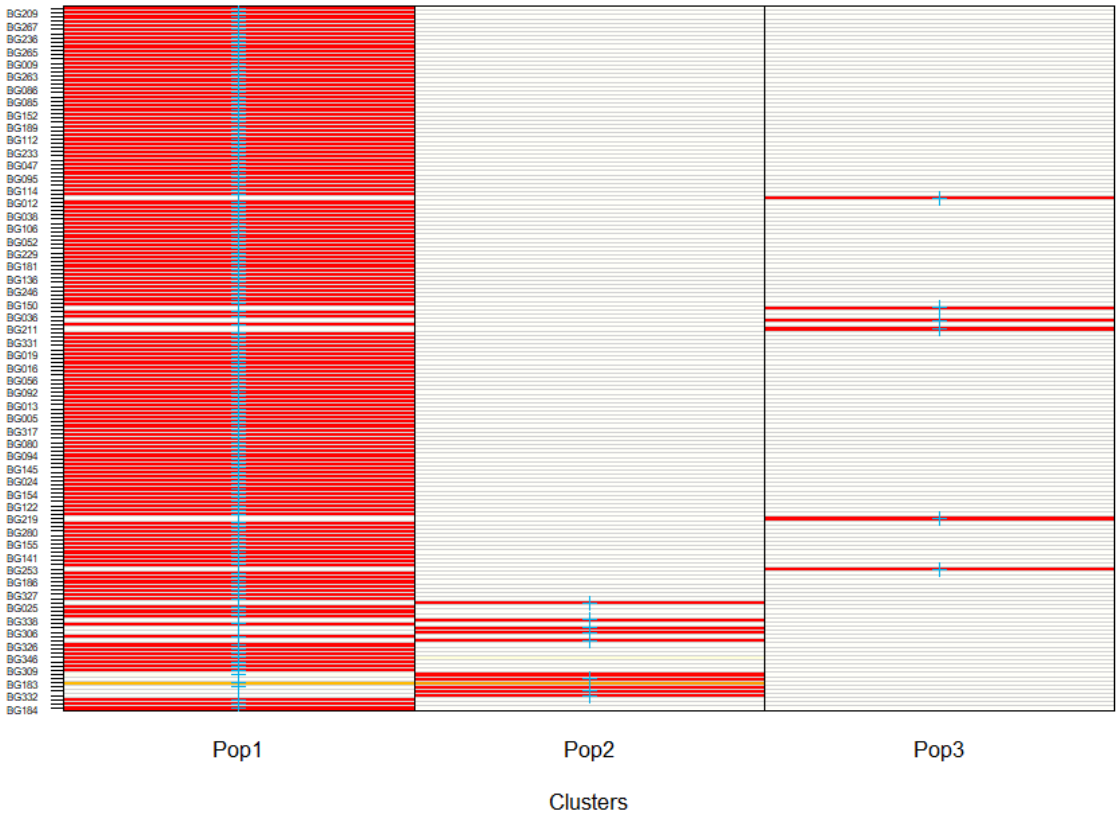
Supplementary Figure S2: **a** Log likelihood for k (LnP(K)) obtained from the population structure of 167 shea tree individuals; **b** Value of delta K for each possible number of clusters (K) best grouping the 167 shea tree individuals studied.



Supplementary Figure S3: **a** Bayesian Information Criterion against the number of clusters showing a decrease of BIC value; **b** Number of Principal components retained based on the a-score.



Supplementary Figure S4: Probability of shea tree individuals belonging to each genetic cluster identified from the DAPC analysis.



Supplementary Figure S5: Assignment of shea tree individuals by genetic groups identified by DAPC analysis

CHAPTER 4

Metabolomic-Guided Diversity Analysis Reveals Chemotypic Groups and Cosmetic-Relevant Biomarkers in Shea Trees (*Vitellaria paradoxa* C.F. Gaertn.) for Industrial Applications

Guillaume Bodjrenou[1,2], Jérémy Grosjean[1], Fernande G. Honfo[3], Alain Hehn[1], Enoch G. Achigan-Dako[2*], Alexandre Olry[1]

[1]Laboratoire Agronomie et Environnement de l'Université de Lorraine (Université de Lorraine, INRAE, LAE, 54000, Nancy, France).

[2]Genetics, Biotechnology and Seed Science Unit (GBioS), Laboratory of Crop Production, Physiology and Plant Breeding, Faculty of Agricultural Sciences, University of Abomey-Calavi, 01 BP 526 Abomey-Calavi, Republic of Benin.

[3]Laboratoire des Sciences et Technologie des Aliments, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi (LaSTA/UAC).

DOI : https://doi.org/10.1016/j.indcrop.2025.122354

Published in Industrial Crops and Products, 239 (2026) 122354

Abstract

Shea tree (*Vitellaria paradoxa* C.F. Gaertn) is valued mainly for its butter, which is rich in bioactive compounds, making it important for the cosmetics, dermocosmetics, and pharmaceutical industries. While some studies have quantified these compounds, the metabolic diversity of shea trees in Benin remains largely underexplored. This study aims to characterize the metabolic diversity of shea trees in Benin, using both targeted and untargeted metabolomics approaches to identify bioactive compounds with potential cosmetic and pharmaceutical applications. Freeze-dried shea kernel powders from 69 trees in nine localities of Benin were analyzed by ultra-high-performance liquid chromatography. Multivariate statistical analyses, including Principal Component Analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA), were used to identify metabolic diversity among the trees. A total of 4,634 signals were detected, with 70% annotated *in silico*, revealing a diverse chemical composition dominated by fatty acids (42%), amino acids/peptides (22%), alkaloids (15%), and phenolics, polyketides and terpene compounds (14%). PCA and K-means revealed three distinct metabolic groupings, and a targeted analysis identified six key metabolites contributing strongly to the variance, including gallic acid, gallocatechin, α-amyrin, epigallocatechin, α-tocopherol, and campesterol. PLS-DA confirmed these groupings, with eight metabolites showing high discriminatory power and potential as biomarkers. Notably, shea tree BG037 and BG143 exhibited particularly high concentrations of compounds of interest for cosmetic applications. This study highlights the substantial metabolic variability in Benin's shea trees, which offers opportunities for tailored applications in the dermocosmetic and nutraceutical industries.

Keywords: *Vitellaria paradoxa*, Metabolic diversity, Targeted and untargeted metabolomics, Bioactive compounds, Cosmetic applications.

1. Introduction

The shea tree, a valued plant resource of West African origin, is of significant economic importance on a global scale (Adedeji et al., 2024). The tree is primarily renowned for the butter extracted from its kernels, a product that is highly prized on the international market. Recent studies have indicated a consistent and gradual increase in global demand for shea butter (Grand View Research, 2024). Shea butter has significant applications in the cosmetics and pharmaceutical sectors. In 2024, over 80% of global shea butter volume was destined for specific uses in cosmetics and personal care products (Grand View Research, 2024). Consequently, it is legitimate to speculate on the specific properties of shea butter that render it a preferred ingredient in the cosmetics and dermocosmetics industries. As demonstrated in the relevant literature, the unsaponifiable fraction of the substance under investigation contains a variety of bioactive compounds for which there is proven scientific evidence of efficacy (Karaca and Güzel Kara, 2024; Segman et al., 2012). These compounds include tocopherols (vitamin E), which are potent antioxidants that prevent skin ageing (Fiume et al., 2018; Fiume, 2002); triterpenes and triterpene alcohols (cinnamic acid esters, α/β-amyrin, parkeol, butyrospermol, and lupeol), which are known for their anti-inflammatory, healing, and cell regeneration effects, particularly useful for irritated or damaged skin (Akihisa et al., 2010b); phytosterols (e.g. campesterol, stigmasterol, β-sitosterol and α-spinasterol), which have been demonstrated to strengthen the skin barrier and soothe redness and eczema (Trautwein and Demonty, 2007); phenol derivative Gallic acid and flavonoid polyphenols (gallocatechin, epigallocatechin, epigallocatechin gallate, gallocatechin gallate, catechin, epicatechin and epicatechin gallate), which are powerful antioxidants with anti-inflammatory, anticancer, antimicrobial, antiviral, bone-protective and neuroprotective properties (Capasso et al., 2025; Maranz et al., 2003a; Segman et al., 2012).

Despite growing international interest in shea butter due to its many biological properties, research into precisely quantifying its bioactive compounds is limited. However, optimising the scientific and industrial utilisation of this resource requires a better understanding of the potential of each tree. Among the few studies that have addressed the quantification of bioactive compounds in shea butter, the majority have focused on triterpenes and phenolic compounds. For example, Akihisa et al. (2010a) found an average proportion of 48.1% triterpene alcohols in the unsaponifiable fraction of shea butter. The most prevalent compounds were α-amyrin (31.3–41.4%), lupeol (17.5–25.1%), butyrospermol (14.9–26.3%), and β-amyrin (8.2–13.2%), which were obtained from samples collected in seven regions of sub-Saharan Africa. However,

no significant differences were observed between the different origins. Additionally, Maranz and Wiesman (2004) reported an average concentration of 220 µg/g of total tocopherols, 64% of which was α-tocopherol, in trees from various African regions. They also found that trees growing in more arid areas had higher tocopherol levels than those in less arid zones. Maranz et al. (2003) quantified several phenolic compounds, including gallic acid, gallocatechin, epigallocatechin, epigallocatechin gallate, epicatechin and epicatechin gallate, in 40 trees from various African regions. Their results showed significant variability according to tree provenance.

It should be noted that the aforementioned studies did not include samples from Benin, creating a geographical gap in shea research. In this context, the following scientific questions deserve consideration in order to better characterize the biochemical composition of the local germplasm: what is the bioactive metabolite composition of the shea trees present in Benin? Is there significant variation in the content of these metabolites according to the geographical origin of Benin's shea trees? Which metabolites, or metabolite families, selected from the untargeted and targeted approach, are key to discriminating between different groups of shea trees?

This study aimed to explore most of the metabolites present in Benin shea tree germplasm, and to quantify those particularly relevant to the cosmetics and dermo-cosmetics industries using the Ultra High Performance Liquid Chromatography (UHPLC) combined with an Orbitrap mass spectrometer with high sensitivity of detection and resolution. To achieve this objective and answer the research questions, the following hypotheses were formulated: Benin shea trees exhibit a large bioactive metabolites of interest to the cosmetics and dermo-cosmetics industries. The content of bioactive compounds varies significantly according to the geographical origin of Benin shea trees. There are specific metabolites that distinguish between different groups of shea trees.

2. Methodology

2.1. Plant materials

The study material consisted of 69 shea trees from nine localities in Benin: Bembèrèkè (n = 9), Boukoumbé (n = 7), Gogounou (n = 10), Kandi (n = 9), N'Dali (n = 3), Parakou (n = 4), Pehunco (n = 9), Segbana (n = 9) and Tchaourou (n = 9). The pedoclimatic characteristics of the collection areas are presented in Supplementary Table S1. Shea fruits were collected in 2022, and trees were randomly selected in each collection zone at 100 m intervals at least. The

geographic coordinates of each shea tree involved in this study are provided in Supplementary Table S2.

2.2. Sample preparation and extraction of bioactive metabolites for untargeted and targeted metabolomic analysis

Five shea fruits were collected and manually shelled for each tree to extract the kernels. These were then dried at 70 °C for 72 hours and ground at 30 Hz for 20 sec using a ball mill (RETSCH MM400, Retsch GmbH, Haan, Germany). The resulting samples were freeze-dried to stabilise the material for metabolomic analysis. Approximately 10 mg of the lyophilised, ground sample was weighed into 2 ml Eppendorf tubes. In a fume hood, 25 µL of taxifolin (2 mg/mL) and 1.4 mL of methanol/H20 60/40 (v/v) were added to the tubes. The mixture was vortexed for 10 sec and then left to macerate overnight at room temperature in the dark. After centrifugation for 10 min at 13.2 krpm, 1 mL of the supernatant was recovered. The residue was re-extracted with 800 µL of methanol/H20 60/40, shaken and macerated for four hours, then centrifuged for 10min at 13.2 krpm. The two extracts were then combined and evaporated using a speed-vac overnight. The residues were solubilised in 200 µL of methanol/H20 70/30 (v/v), adjusted for initial mass and then incubated for 2 hours at room temperature. After vortexing and centrifugation (10 minutes at 13.2 krpm), the supernatant was filtered through a 0.22 µm filter and transferred into vials for UHPLC-Orbitrap analysis. The extracts were stored at -20°C prior to analysis.

2.3. Preparation of analytical standards for targeted metabolomic approach

Pure standards of the bioactive compounds of interest were purchased from three suppliers: Sigma-Aldrich (Saint-Quentin-Fallavier, France), ChemFaces (Wuhan, China) and Carl Roth (France), for the identification and quantification of metabolites present in shea extracts (Supplementary Table S3). The selection of the fifteen metabolites was guided by evidence reported in previous studies on the unsaponifiable fraction of shea kernels, which is known to contain a wide range of bioactive compounds with demonstrated biological activities (Karaca and Güzel Kara, 2024; Segman et al., 2012). These standards included allantoin, gallic acid, gallocatechin, epigallocatechin, trans-cinnamic acid, α-amyrin, β-amyrin, lupeol, butyrospermol, parkeol, campesterol, stigmasterol, β-sitosterol, α-tocopherol and γ-tocopherol. The analytical methodology used for the identification and quantification of these metabolites was adapted from validated UHPLC–Orbitrap protocols previously described in metabolomic studies (Sheng et al., 2021; Zhao et al., 2024), with minor modifications to accommodate the

physicochemical properties of the selected compounds and the specific characteristics of shea extracts.

Each standard was solubilised in dimethyl sulfoxide to obtain a stock solution at a concentration of 100 µmol/L. Successive dilutions were then made from these stock solutions to generate standard solutions covering a range of concentrations from X to Y for establishing calibration curves. The standards were divided into several mixtures according to their molecular weight to avoid overlap or confusion during mass spectrometric detection. Standard mixtures were filtered at 0.22 µm, then stored at -20°C until injection into the UHPLC-Orbitrap system.

2.4. Analysis of extracts by UHPLC–HRMS (Orbitrap ID-X)

Chromatographic analyses were conducted using an Orbitrap ID-X (Thermo Fisher Scientific) coupled with a Vanquish UHPLC system (Thermo Fisher Scientific), which enables the accurate identification (< 3 ppm RMS with external calibration and < 1 ppm RMS with internal calibration using EASY-IC) of the structure of bioactive compounds through dual detection in MS and MS/MS modes (Kotadiya, 2025). The compounds were separated on a Kinetex C18 column (Phenomenex) measuring 150 mm × 2.1 mm with a particle size of 2.6 µm, over a run time of 45 minutes. Two solvents were used: A (0.1% formic acid in acetonitrile) and B (0.1% formic acid in ultrapure water). The gradient conditions were as follows: starting at 10% B for 2 min, then 20% B at 4 min, 25% B at 5 min, 31% B between 6.9 and 9 min, and finally 50% A and B at 15 min. The B concentration was 62% at 17 min, increasing to 77.5% at 25 min and reaching 80% at 32 min. It then rises to 83% at 35 minutes, finally achieving a 100% B concentration between 35.5 and 39 minutes. The system returns to the initial conditions within one minute. The column temperature was maintained at 40 °C throughout the analysis. Samples were kept at 15 °C prior to the injection of 10 µL of each sample.

Analyses were performed in MS/MS mode in both positive and negative ionisation modes, with a resolution of 60,000 for MS1 and 15,000 for MS2 (utilising both HCD and CID fragmentation modes). The mass range measured was between 120 and 1200 m/z. In ESI mode, the vaporizer temperature was set to 320 °C, and the transfer tube temperature was set to 275 °C. The gas settings for the ionisation block were as follows: sheath gas at 40 (AU), auxiliary gas at 8, and sweep gas at 1, with voltages set to 35kV in positive mode and 2.5kV in negative mode.

2.5. Data processing and analysis

Raw data from the Orbitrap ID-X were processed using the MZmine 3 software (Schmid et al., 2023) with a non-targeted compound detection approach. Feature extraction was performed on the chromatograms and mass spectra in several steps.

2.5.1. Chromatographic parameters (UHPLC)

The chromatographic signals were smoothed, and the *"Stable ionisation across samples"* option was activated to improve consistency between samples. The retention time range was defined as 0.50–45.00 min, with a maximum of 15 peaks per chromatogram. A minimum of four consecutive scans was required for peak validation. The width at half maximum (FWHM) was set to 0.10 min and retention tolerances were defined as 0.08 min (intra-sample) and 0.40 min (inter-sample).

2.5.2. Spectrometric parameters (Orbitrap)

The analysis was performed only in positive ionisation mode, with detection based on the lowest signal factor. The noise threshold was set at 5.0 for MS^1 and 2.5 for MS^2. The minimum intensity threshold for features was set at 1.0×10^6. The m/z tolerances were set as follows: 0.0020 m/z or 10 ppm for scan-to-scan variations, 0.0015 m/z or 3 ppm for intra-sample variations, and 0.0015 m/z or 5 ppm for inter-sample variations (Heuckeroth et al., 2024). Post-alignment filtering was applied to retain only features present in at least one sample. No filter specific to the ^{13}C isotope was applied to maximise metabolomic coverage and avoid the exclusion of relevant compounds.

2.5.3. Feature annotation

The detected compounds were annotated using SIRIUS software, version 6.1.0 (Dührkop et al., 2019). Features extracted via MZmine were exported in the compatible mgf (Mass Spectrum File) format, along with the associated MS/MS spectra for each detected peak. This approach is based on the reconstruction of molecular formulas and the prediction of chemical structures from isotopic patterns and MS^2 fragments. The CSI:FingerID module, which is integrated into SIRIUS, was used to compare the experimental spectra with the public molecular databases integrated into SIRIUS (Dührkop et al., 2015), such as PubChem, GNPS, Coconut and Lotus. This enabled structural similarity scores to be obtained for each annotation (Hoffmann et al., 2022). Annotations were considered reliable when the probability score was above a fixed

threshold (score ≥ 0.7) and when the proposed molecular formula matched expectations based on exact weight and isotopy.

2.6. Statistical analysis

Metabolomic data were analysed using two complementary statistical approaches: a non-targeted approach and a targeted approach. This was done to explore the overall variability of metabolic profiles and the dynamics of specific compounds of interest.

2.6.1. Non-targeted metabolomic approach

Log transformation was applied to the peak area data, which was then normalized, centered, and reduced prior to analysis. Unsupervised PCA was used to explore global trends. Classification was performed based on the Euclidean distance between individuals using the k-means function in the stats package. Permutational multivariate analysis of variance (PERMANOVA) was performed using the 'adonis2' package in R to test whether the obtained groups were significantly different. A qualitative approach was then employed to categorize the groups according to their origin.

2.6.2. Targeted metabolomic approach

A targeted analysis was performed using injected reference standards (allantoin, gallic acid, gallocatechin, epigallocatechin, trans-cinnamic acid, α-amyrin, β-amyrin, lupeol, butyrospermol, parkeol, campesterol, stigmasterol, β-sitosterol, α-tocopherol and γ-tocopherol). The compounds were quantified based on extracted peak intensities corresponding to the expected m/z and retention times. The identity of each compound was confirmed by comparing the MS/MS spectra with those of the standards. Descriptive statistics were combined with analysis of variance (ANOVA) to observe the dispersion of the variables and determine whether there was a significant difference in the metabolite content of shea trees depending on their provenance. Corrections for multiple comparisons (FDR type) were applied where necessary. The multivariate analyses were identical to those carried out in the non-targeted approach, except that PLS-DA, coupled with a non-linear Random Forest approach, was used to identify the variables that best distinguish the groups obtained from clustering. A multiclass ROC (Receiver Operating Characteristic) curve was used to validate the Random Forest model and confirm the discriminating variables identified by both methods. A heatmap was generated to visualize the concentrations of the discriminating variables in each cluster.

3. Results

3.1. Untargeted analysis of metabolic profiles

Using MZmine software for feature extraction, 4,634 signals were identified, of which 3,255 (around 70%) could be annotated based on chemical family name in silico using SIRIUS version 6.1.0. Of these, 67% were retained on the basis of a predictive fingerprint score of at least 0.7, which guarantees the reliability of the structural annotation. Analysis of the chemical classification of these annotated features revealed seven major compound families to be predominant: fatty acids (42%), amino acids and peptides (22%), alkaloids (15%), terpenoids (9%), carbohydrates (7%), shikimates and phenylpropanoids (3%), and polyketides (2%).

In our case, particular attention was paid to families of biological and cosmetic interest, such as terpenoids, shikimates, phenylpropanoids and polyketides (Figure 1). A more in-depth analysis of these revealed the constituent metabolites of each class.

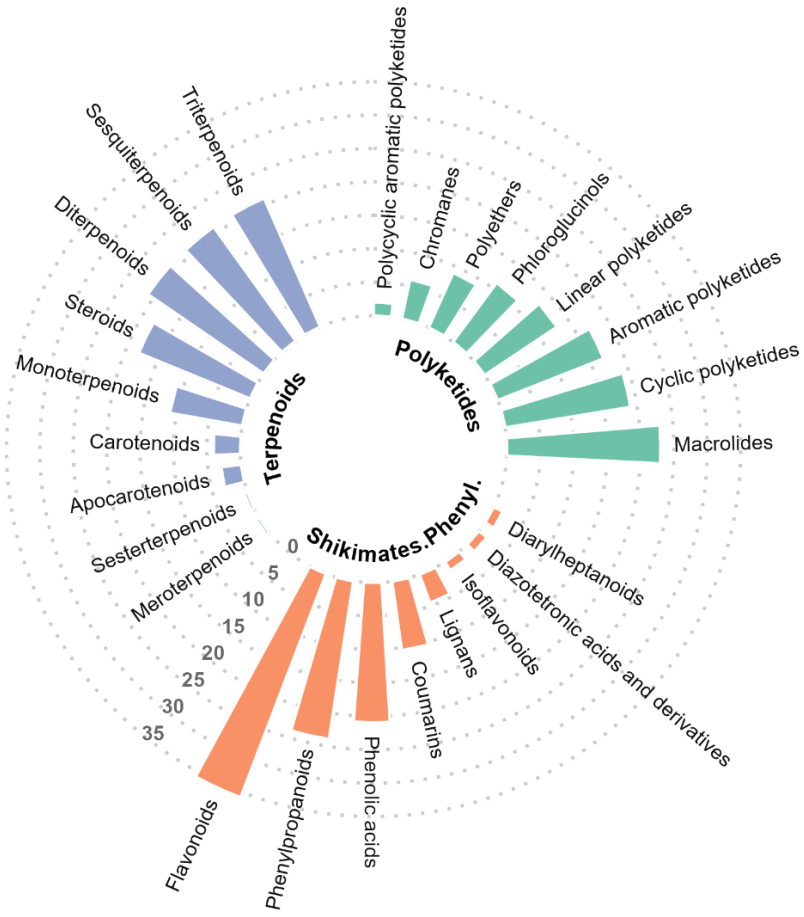


Figure 1: Circular percentage distribution of metabolites by class of interest via in silico annotation in SIRIUS *version 6.1.0*. The terpenoid family is represented in blue; shikimates and phenylpropanoids are in orange; polyketide compounds are shown in green.

Within the terpenoid group, nine (09) compounds were annotated using chemical family named, with five predominant. Among the five predominant compounds are di-, tri-, and sesquiterpenoids, each accounting for approximately 21%. Steroids and monoterpenoids represent approximately 18% and 11%, respectively. Eight (08) compounds were annotated within the shikimates and phenylpropanoids, the most abundant of which were flavonoids (35.8%), phenylpropanoids (23.9%), phenolic acids (20.9%), and coumarins (10.4%). The last class, consisting of polyketides, includes eight (08) compounds, with three major ones being macrolides (23%), cyclic polyketides (19%), and aromatic polyketides (17%) (Figure 1).

3.1.1. Principal component analysis (PCA) for detecting relevant features

When applied to all variables, PCA explained around 19.4% of the total variance on the first two axes (Supplementary Table S4). For each axis, the 50 features that contributed the most to the variance were retained (loadings $\geq$ 0.04) (Supplementary Table S5). Feature-based annotation via the GNPS platform identified around 73% of the 100 analysed compounds. Lipids, which were dominated by fatty acids and derivatives (57.5%), constituted the largest class, followed by flavonoids (15%). Amino acid derivatives were poorly represented (2.7%).

3.1.2. Clustering and PERMANOVA analysis

A k-means classification based on the PCA results was used to group the shea tree individuals according to similarity. The elbow plot, based on within-cluster variance (Supplementary Figure 1), indicated that the optimal number of groups was k = 3. Three clusters of shea trees were distinguished: G1, G2 and G3, comprising 10, 53 and 4 shea trees, respectively (Figure 2). The differences between the three clusters were highlighted using PERMANOVA analysis performed in the R package 'adonis2' and based on the Euclidean distance matrix between individuals.

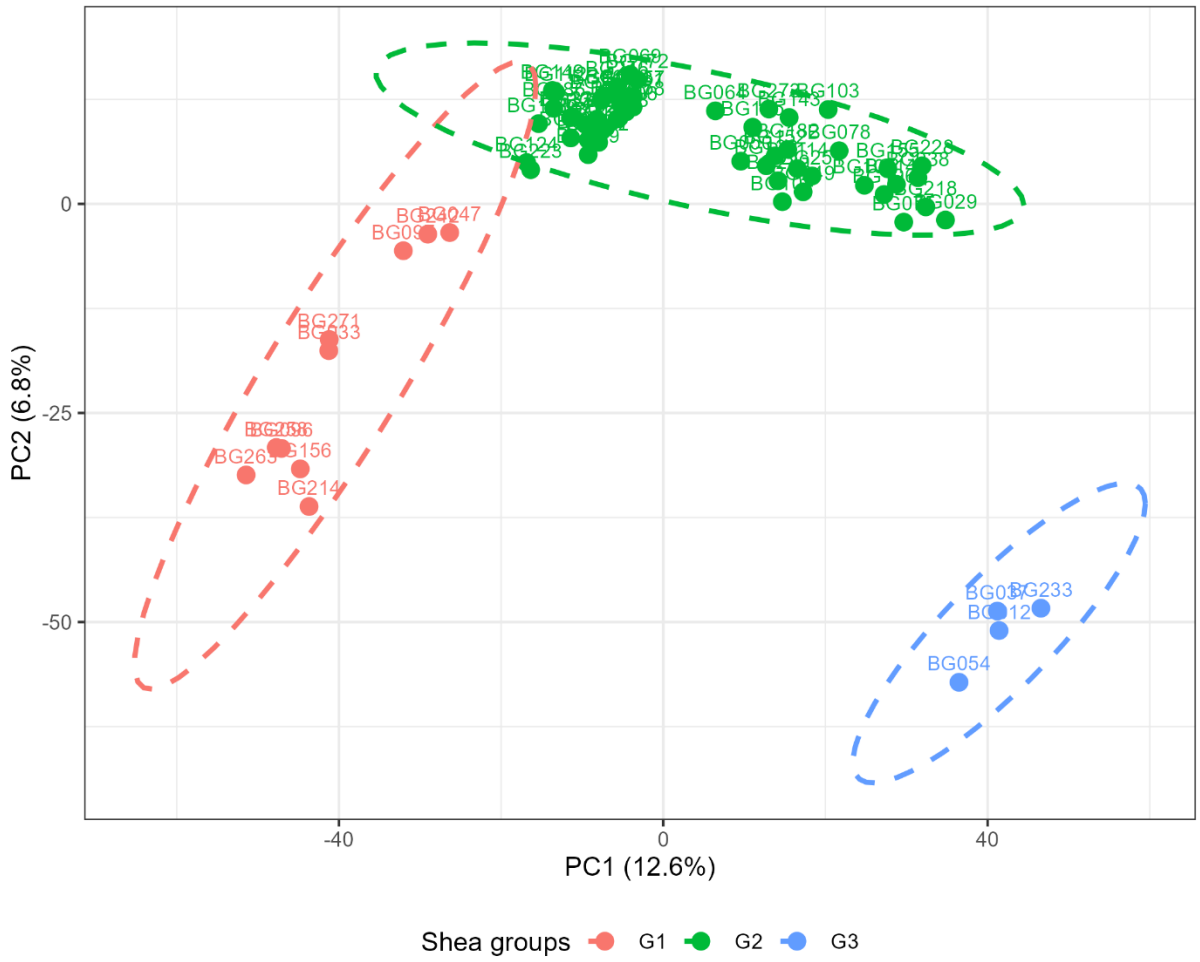


Figure 2: Metabolomic clustering of shea tree based on the set of 3,255 detected metabolic signals projected onto the first two (PCA). The k-means function was used with Euclidean distance between individuals.

The results revealed a statistically significant difference between the three groups (F = 5.93, p = 0.001) based on all features (Table 1). The 'cluster' factor explained 15.6% of the total variance, suggesting robust structuring of the samples according to their metabolic profiles.

Table 1: Permutational multivariate analysis of variance (PERMANOVA) of shea tree clusters according to the 3,255-feature set.

Source of variation	Degree of freedom	Sum of squares	R^2	F	P (>F)
Shea groups	2	47416	0.15638	5.9319	0.001
Residual	64	255788	0.84362		
Total	66	303204	1		

3.1.3. Description of clusters according to the geographic origin of the shea tree individuals

Clustering analysis revealed that 79% of individuals were grouped within cluster G2 (Supplementary Figure 2). This cluster mainly comprises individuals from Tchaourou and Ségbana (approximately 15% each), followed by Kandi and Gogounou (approximately 13% each). Boukoumbé and Bembèrèkè were moderately represented at around 11% each. Group G1, representing around 14% of the total sample, did not show any marked geographical dominance. Nevertheless, it includes individuals from Bembèrèkè, Gogounou, Kandi and Pehunco (20% each), as well as Ségbana and Tchaourou (10% each). Finally, the marginal G3 group, accounting for 6% of individuals, comprises 50% from Pehunco, with Boukoumbé and Gogounou contributing 10% each (Supplementary Figure 2).

3.2. Targeted analysis of metabolic profiles

A descriptive statistical analysis of the fifteen target metabolites was carried out to identify the compounds predominantly present in shea tree samples. Six metabolites stood out due to their high concentrations, expressed in milligrams per gram (mg/g): α-tocopherol, campesterol, stigmasterol, β-sitosterol, gallic acid and butyrospermol (Table 2). Conversely, nine compounds showed lower levels expressed in ppm (parts per million), including trans-cinnamic acid, lupeol, α-amyrin, γ-tocopherol, β-amyrin, parkeol, epigallocatechin, gallocatechin and allantoin. No significant differences were observed between individuals depending on their provenance for any variable (Supplementary Figure 3).

Table 2: Descriptive statistics for the concentrations of the fifteen targeted metabolites in the shea samples expressed in µmol/L, mg/g and ppm

Compounds	Minimum	Maximum	Mean	Standard error	Mean (mg/g or ppm)
			µmol/L		mg/g
α-tocopherol	9707.48	23952.49	17527.14	385.2	150.98
Campesterol	9878.77	28634.16	15535.25	428.53	124.5
Stigmasterol	991.08	2424.16	1849.55	33.09	15.27
β-sitosterol	51.75	7744.35	1595.05	218.16	13.23
Gallic acid	113.26	1059.48	581.35	31.21	1.98
Butyrospermol	6.17	776.51	139.39	25.29	1.19
			µmol/L		ppm
Transcinnamic acid	29.29	2445.48	232.48	67.89	688.9
Lupeol	1.82	273.85	60.6	7.49	517.17
α-amyrin	10.29	214.77	44.52	7.62	379.91

Compounds	Minimum	Maximum	Mean	Standard error	Mean (mg/g or ppm)
γ-tocopherol	1.29	100.22	31.34	2.35	261.14
β-amyrin	1.29	123.46	23.97	5.31	204.56
Parkeol	3.17	58.8	21.69	3.15	185.1
Epigallocatechin	0.49	26.43	6.66	0.56	40.8
Gallocatechin	0.56	51.69	6.24	1.21	38.24
Allantoin	0.04	15.36	2.37	0.74	7.5

The highest concentrations (expressed in mg/g) of different metabolites were observed in specific accessions. Accessions BG233, BG037, BG037, BG229, BG267 and BG249 showed high levels of α-tocopherol, campesterol, stigmasterol, β-sitosterol, gallic acid and butyrospermol, respectively.

The highest levels of compounds present in lower concentrations were recorded in accessions BG050 (7.25 mg/g), BG143 (2.34 mg/g), BG047 (874.78 ppm) and BG112 (835 ppm), BG271 (742 ppm), BG143 (501.8 ppm), BG258 (161.89 ppm), BG037 (316.62 ppm) and BG265 (48.57 ppm) for trans-cinnamic acid, lupeol, α-amyrin, γ-tocopherol, β-amyrin, parkeol, epigallocatechin, gallocatechin and allantoin, respectively. Two accessions stood out in particular for their metabolic profiles: BG037 (from Pehunco), which showed higher concentrations of campesterol, stigmasterol and gallocatechin than all other accessions, and BG143 (from Boukoumbé), which stood out with remarkably high levels of lupeol and parkeol.

3.2.1 PCA and clustering of shea trees according to their metabolic profile

The first two axes of the PCA, which were carried out on the 15 studied metabolites, explained 40.5% of the total variance in the dataset (Supplementary Figure 4). The metabolites that contribute most strongly to this variance ($r \geq 0.60$) were: gallic acid (-0.84), gallocatechin (0.81), α-amyrin (0.78), epigallocatechin (-0.75), α-tocopherol (-0.67) and campesterol (0.62).

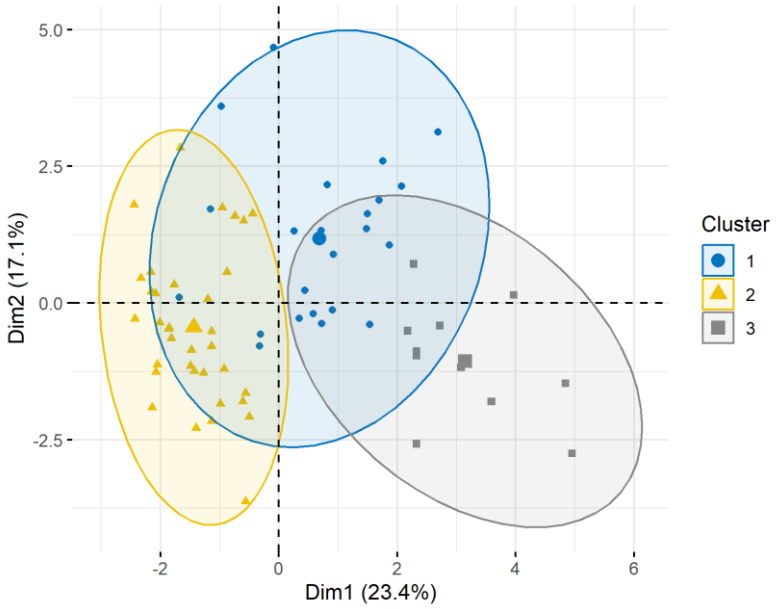


Figure 3: Hierarchical ascending clustering of shea tree individuals based on 15 metabolites of interest, using "ward.D2" distance among individuals.

The PCA results were used to group the shea tree individuals into three main categories (Figure 3b). The optimal number of classes (k = 3) was determined by combining the Dunn index and elbow methods (Supplementary Figure 5). Cluster 1 comprises 23 individuals (approximately 33%), cluster 2 comprises 35 individuals (approximately 51%), and cluster 3 comprises 11 individuals (approximately 16%).

A comparison of the classifications obtained using the targeted and non-targeted approaches revealed partial correspondence between the clusters. Indeed, 47% of individuals classified as G2 using the non-targeted approach were found in Cluster 2 using the targeted approach. Conversely, all individuals in cluster G3 from the non-targeted approach belonged to cluster 3 from the targeted approach. However, only 20% of the accessions in group G1 from the non-targeted approach were found in cluster 1 of the targeted approach.

To test the significance of the multivariate differences observed between clusters, a similarity analysis (ANOSIM) was performed. The results revealed a significant difference between clusters, with an R statistic of 0.5077 and a p-value of 0.001, indicating notable separation between clusters (Table 3). The distribution of dissimilarity ranks shows that dissimilarities between clusters are generally higher than within clusters. PERMANOVA analysis confirmed this separation, indicating that the clusters explained 44% of the variability in the 15 metabolic variables (Table 4).

Table 3: Analysis of similarities (ANOSIM) and PERMANOVA pairwise comparison between clusters of shea tree samples

Global test: R = 0.5077***			
Clusters comparisons	R^2	F-value	p-value
Cluster1 vs Cluster2	0.3419***	29.09	0.001
Cluster1 vs Cluster3	0.357***	17.77	0.001
Cluster2 vs Cluster3	0.4224***	32.18	0.001

After Bonferroni p-value correction, comparisons of pairwise differences between clusters were statistically significant (Table 3), suggesting that each cluster differs significantly from the other two. Additionally, the homogeneity of dispersions test revealed significant variance heterogeneity (p = 0.006) between clusters, suggesting that the observed differences may partly result from varying internal variability within clusters (Supplementary Table S6).

Table 4: PERMANOVA analysis results of the clusters identified by targeted metabolomics clustering

Source variation	Df	Sum of squares	R^2	F	P(>F)
Clusters	2	7.57E+08	0.44012	25.942	0.001***
Residual	66	9.62E+08	0.55988		
Total	68	1.72E+09	1		

Analysis of variance (ANOVA) combined with a multiple comparison of means revealed nine metabolites (α-tocopherol, campesterol, β-sitosterol, butyrospermol, gallic acid, lupeol, parkeol, epigallocatechin and gallocatechin) that showed highly significant differences in responses between the three clusters (Table 5). The dispersion of variables by cluster is shown in Supplementary Figure 6.

Table 5: Analysis of variance between clusters, with a comparison of the mean values for each metabolite of interest per cluster expressed in mg/g and ppm.

Metabolites	Units	Cluster 1	Cluster 2	Cluster 3	p-value
		n = 23	n = 35	n = 11	
α-tocopherol	mg/g	128.27^{b}	167.25^{a}	146.44^{b}	< 0.0001
Campesterol	mg/g	114.64^{b}	117.17^{b}	168.28^{a}	< 0.0001
β-sitosterol	mg/g	24.31^{a}	5.26^{b}	13.3^{b}	< 0.0001
Stigmasterol	mg/g	15.43^{a}	14.91^{a}	16.13^{a}	0.268
Butyrospermol	mg/g	2.45^{a}	0.54^{b}	0.61^{b}	< 0.0001
Gallic acid	mg/g	1.74b	2.49^{a}	0.83^{c}	< 0.0001
Transcinnamic acid	ppm	891.35^{a}	583.86^{a}	550.14^{a}	0.387

Metabolites	Units	Cluster 1	Cluster 2	Cluster 3	p-value
		n = 23	n = 35	n = 11	
Lupeol	ppm	704.17[a]	434.49b	336[b]	0.005
α-amyrin	ppm	448.38[a]	408.52[a]	301.68[a]	0.365
γ-tocopherol	ppm	243.76[a]	274.51[a]	277.01[a]	0.722
Parkeol	ppm	199.35[b]	123.32[c]	295.02[a]	< 0.0001
β-amyrin	ppm	184.93[ab]	266.52[a]	106.16[b]	0.026
Epigallocatechin	ppm	34.73[b]	52.86[a]	14.7[b]	< 0.0001
Gallocatechin	ppm	26.71[b]	13.66[b]	134.45[a]	< 0.0001
Allantoin	ppm	7.4[a]	8.38[a]	9.17[a]	0.747

Qualitative analysis of the clusters revealed a relatively homogeneous distribution of origins within the groups. The distribution of individuals' metabolic profiles within clusters does not show any obvious traditional geographical structure, which seems to contradict the expected spatial organization of "Shea Parklands" as described by Gnanglè (2005) (Figure 4). All individuals from the three clusters were represented within each "Shea Parklands," showing no singularity by Parklands (Figure 4).

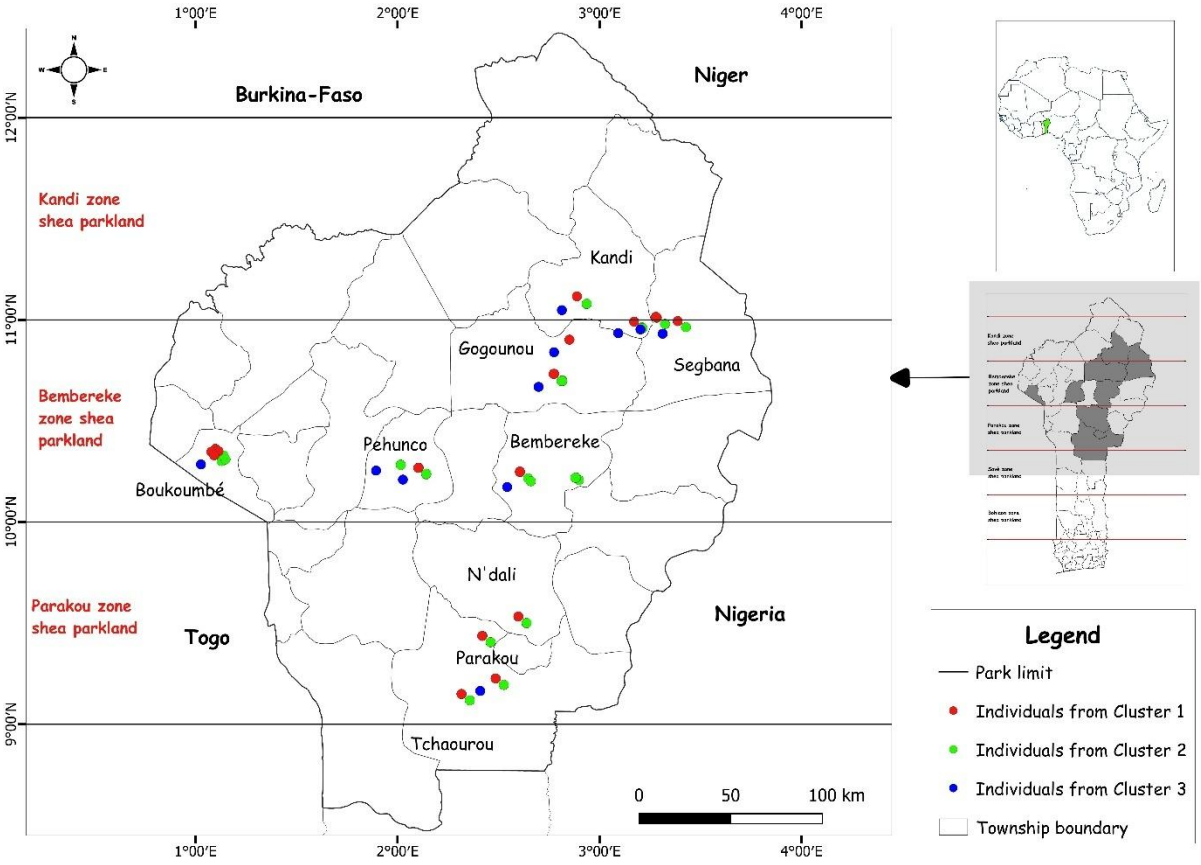


Figure 4: Geographic distribution of *Vitellaria paradoxa* individuals grouped by metabolite-based clusters overlaid on Shea Parkland zones in Benin

3.2.2. PLS-DA and Random Forest analyses for discriminant metabolites identification

A combined analysis integrating a linear PLS-DA and non-linear (Random Forest) model was carried out, considering the clusters obtained from the principal component analysis as shea tree groups, to determine the most discriminating metabolites. The two axes of the PLS-DA analysis captured around 39% of the total variation contained within the given sets. Using the VIP (Variable Importance in Projection) of the two components to select highly discriminating variables, eight compounds with a VIP ≥ 1 were identified (Supplementary Figure 7). Lupeol has a relatively moderate VIP value (VIP = 0.80) and contributes moderately to the observed discrimination between shea tree groups (see Supplementary Table S7).

A random forest analysis of discriminant variables enabled us to identify discriminant variables using two parameters: MeanDecreaseAccuracy and MeanDecreaseGini. The former measures the extent to which model accuracy decreases when the values of a variable are randomly permuted, while the latter measures the average reduction in Gini impurity caused by a metabolite across all trees in the model. A high value for either parameter reflects the variable's contribution to improving the efficiency and therefore the discriminatory power of the model. Thus, variables whose importance (as measured by MeanDecreaseAccuracy and MeanDecreaseGini scores) was greater than or equal to the median for all variables were considered highly discriminative (Supplementary Table S8). Consequently, eight (08) variables were selected from the Random Forest analysis: gallic acid, gallocatechin, epigallocatechin, campesterol, α-tocopherol, β-sitosterol, butyrospermol and parkeol (Figure 5).

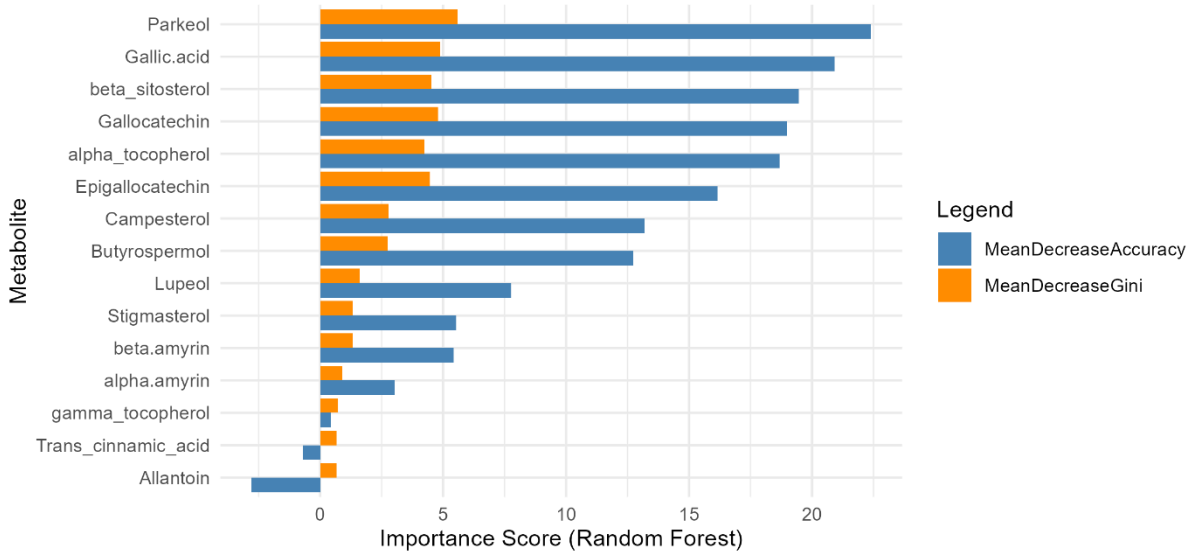


Figure 5: Projection of the Mean Decrease Accuracy and Mean Decrease Gini values for the 15 metabolites obtained from the Random Forest analysis

The PLS-DA and Random Forest analyses produced identical results for the identification of discriminating metabolites, demonstrating the robustness and reliability of the selected variables. Furthermore, a multiclass analysis of the ROC (Receiver Operating Characteristic) curves confirmed the relevance of selecting the eight metabolites common to both models, with an AUC (area under the curve) value of at least 0.75, indicating moderate to good discriminating power (Schober et al., 2021) (Figure 6).

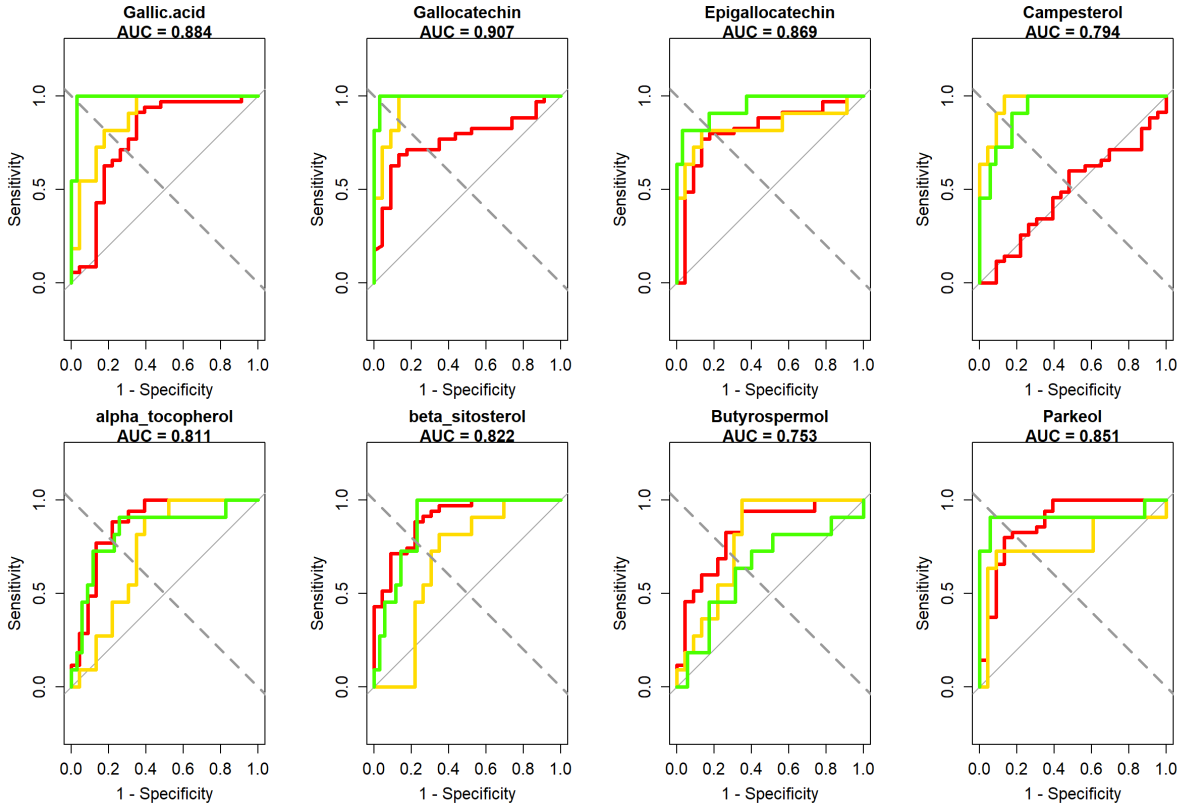


Figure 6: Receiver operating characteristic (ROC) curves for cluster-discriminating compounds: comparison of predictive performance (AUC) of metabolites

The curves for all compounds were close to the upper left-hand corner, indicating a high sensitivity value and good metabolite discrimination power.

3.2.3. Annotated heatmap with hierarchical clustering based on the discriminant metabolites

The heatmap of discriminating metabolites clearly shows how metabolic profiles are structured according to the groups defined by clustering. There are marked variations in the majority of

compounds between the different clusters. This distribution indicates the significant role of these metabolites in the chemical differentiation of shea trees. Cluster 1 is characterized by high levels of butyrospermol and β-sitosterol, cluster 2 by notable accumulations of epigallocatechin, gallic acid and α-tocopherol, and cluster 3 by higher concentrations of gallocatechin, campesterol and parkeol (Figure 7). Aggregating accessions according to the similarity of their metabolic profiles thus confirms the coherence of the groups identified by multivariate analysis.

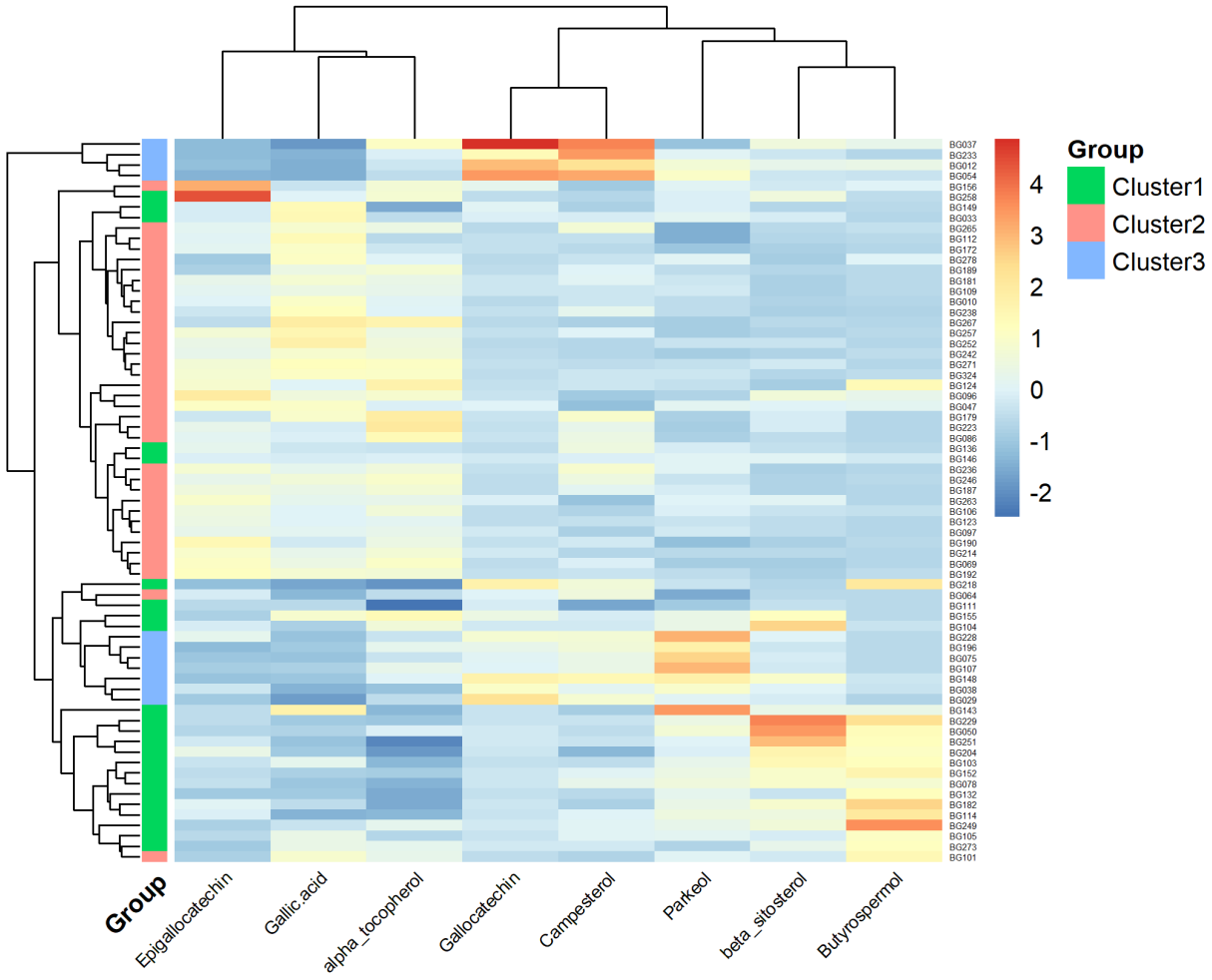


Figure 7: Annotated heatmap of discriminating metabolites with hierarchical clustering of shea tree individuals

4. Discussion

4.1. Bioactive compounds with cosmetic and dermocosmetic value in *Vitellaria paradoxa*

A non-targeted metabolomic analysis of Benin shea trees using UHPLC-Orbitrap systems identified various chemical families, around 14% of which corresponded to polyketides, terpenoids, shikimates and phenylpropanoids. These families are well known for containing

metabolites with significant potential in cosmetics and dermocosmetics. Due to their anti-inflammatory, antioxidant, antimicrobial and anti-allergic properties, as well as their ability to protect and regenerate the skin, these compounds are particularly promising candidates for the development of innovative skincare products (Câmara et al., 2024; Mrudulakumari Vasudevan and Lee, 2020; Wang et al., 2023). The presence of these chemical families highlights the great phytochemical richness of Benin's shea trees and the diversity of bioactive metabolites they contain. Several of these metabolites are of major interest for cosmetic and dermocosmetic applications. The presence of polyketides among these families emphasises the key role of shea in preventing and treating skin disorders of bacterial or fungal origin (Wang et al., 2023; Albisoru et al., 2025). The work of Maranz et al. (2003) and Akihisa et al. (2010a) reported the presence of significant concentrations of polyphenols (notably flavonoids) and triterpene alcohols in West African shea trees. These compounds are used in cosmetics and pharmaceutical products (Masyita et al., 2022; Jesus et al., 2025). These findings corroborate the results of the present study, which confirm the presence of major bioactive metabolites in Benin shea tree germoplasm.

4.2. Analysis of chemical families contributing to the metabolic discrimination of shea trees

Principal component analysis (PCA) was carried out on all metabolic features to identify the 100 most instrumental compounds in structuring the data. This led to the identification of three distinct groups of shea trees. Chemical family name-based annotation of 73% of these discriminating compounds revealed a predominance of lipid compounds (such as fatty acids and their derivatives, as well as glycerolipids), accompanied by around 15% flavonoids. These results emphasise the significant role of polyphenols, particularly flavonoids, in metabolic differentiation between the identified groups. This inter-group discrimination was also statistically confirmed by PERMANOVA analysis. In their study of the quantification of six phenolic compounds, notably flavonoids, in 40 shea trees from ten African countries, Maranz et al. (2003) revealed significant variations in flavonoid content depending on geographical origin. This heterogeneity could explain the significant presence of flavonoids among the discriminating compounds identified in Benin shea trees in the present study.

4.3. Targeted analytical approach to bioactive metabolites

The fifteen metabolites involved in the present study were found to have higher concentrations in Benin shea trees than those reported in the literature. Among the tocopherols, α-tocopherol

predominates, in line with the findings of Adejumo et al. (2022) and Maranz and Wiesman (2004). However, the α-tocopherol concentrations obtained in the present study (α-tocopherol = 150.98 mg/g) are much higher than those reported by these two authors: α-tocopherol = 2.60 mg/100 g and α-tocopherol = 220 µg/g, respectively. The concentrations of phytosterols recorded are also higher than those reported for species studied by Vecka et al. (2019), as is the case for triterpenes (Akihisa et al., 2010a; Di Vincenzo et al., 2005a).

Indeed, these authors quantify metabolites in whole shea butter (in micrograms per gram of butter), whereas in the present study, we standardise on crude dry matter (freeze-dried shea kernel powder). In the absence of lipid/intermediate dilution, this naturally concentrates the extracted compounds (Górnaś et al., 2022; Liu et al., 2025). Furthermore, our targeted and iterative extraction protocol maximises the recovery of lipophilic compounds, producing a more concentrated fraction than overall fat extraction, which is often used in analyses of raw butter. This matrix contains lipids that have not yet been altered by extraction processes and may have a higher relative content of lipophilic compounds (Górnaś et al., 2022). Moreover, UHPLC coupled with Orbitrap MS detection has a higher resolution, sensitivity and accuracy than conventional HPLC MS methods. This enables the separation and detection of close and/or low-proportion analytes with increased efficiency, reducing overlap and analytical losses (Kotadiya, 2025). Consequently, our higher value is consistent with the literature, reflecting the quantification of a highly concentrated dry matter fraction made possible by optimized extraction and ultra-efficient detection. This suggests that differences in matrix, method and normalization basis must be carefully considered when comparing compound levels in shea butter across different studies.

4.4. Variation in metabolic profiles and spatial distribution of Benin shea trees

This study identifies two accessions with exceptional metabolic profiles that merit particular attention due to their high concentrations of several metabolites of interest. Accession BG037, from Pehunco, has a notable profile characterized by high concentrations of campesterol (229.47 mg/g of shea kernel powder), stigmasterol (20 mg/g) and gallocatechin (316.62 ppm). Similarly, accession BG143, from Boukoumbé, has particularly high levels of lupeol (2.34 mg/g) and parkeol (501.8 ppm), highlighting its significant bioactive potential. However, analysis of variance (ANOVA) showed no significant difference in shea tree concentration according to geographical origin. Moreover, the spatial projection of metabolite-based clusters over the established Shea parklands reveals a mismatch between biochemical diversity and geographical parkland boundaries. Individuals assigned to the same metabolic cluster are

scattered across distinct parklands, indicating that the metabolic profiles do not align with current agro-ecological zoning. This suggests that factors beyond geography, likely genetic variation and local environmental microconditions play a more dominant role in shaping the metabolomic landscape of *Vitellaria paradoxa*.

This contradicts the findings of Maranz and Wiesman (2004), who noted a regional influence on tree tocopherol content, and Di Vincenzo et al. (2005), who observed a significant difference in triterpene compound content according to tree origin. However, our results are consistent with those of Akihisa et al. (2010a), who similarly found no significant difference in terpene content based on accession origin. The contradictions observed in Maranz and Wiesman (2004) and Di Vincenzo et al. (2005) works, could be explained by the fact that all the shea trees involved in the present study originate from the northern region of Benin and are subject to minimal environmental influence. However, a study of genotype-environment interaction would be better suited to assessing the influence of environmental components on metabolite content. Such a study would require metabolic data collected over several years (at least three years) from a large number of shea trees and covering a wide geographical area (Crossa et al., 2022).

4.5. Identification of discriminant metabolites and differentiation of chemical profiles in Benin shea trees

Principal component analysis (PCA) of the 15 selected metabolites explained approximately 40% of the total variance. Six of these metabolites: gallic acid, gallocatechin, α-amyrin, epigallocatechin, α-tocopherol and campesterol, were found to be particularly important in structuring the data. Three distinct groups of shea trees were obtained, which were statistically differentiated according to the ANOSIM and PERMANOVA tests. The absence of any apparent correlation with geographical origin suggests that this structuring is not random, but rather results from an organisation based on metabolic distance between individuals. Shakuri et al. (2025) conducted a phytochemical analysis of *Quercus sp*. fruits (oak), which are renowned for their high content of phenolic acids, triterpenoids, and polyphenols of plant origin. Their study identified compounds such as gallic acid and certain flavonoids as biomarkers that discriminate between trees based on their origin. In the present study, the combined PLS-DA and Random Forest approach identified, with 100% concordance, the most discriminating metabolites among Benin shea trees. These include: gallic acid, gallocatechin, epigallocatechin, campesterol, α-tocopherol, β-sitosterol, butyrospermol and parkeol. These compounds exhibit moderate to good discriminatory power ($0.75 \leq AUC \leq 0.90$), positioning them as potential

biomarkers for the analysis of metabolic diversity within shea accessions. This metabolic structuring highlighted distinct clusters of trees, characterized by specific bioactive metabolite profiles and differentiated functional potential. Cluster 1 is distinguished by high levels of butyrospermol and β-sitosterol; cluster 2 is enriched in epigallocatechin, gallic acid and α-tocopherol; while cluster 3 is characterized by high concentrations of gallocatechin, campesterol and parkeol. These findings highlight the crucial role of metabolomic approaches in selecting trees with a high concentration of compounds of interest.

4.6. Industrial applications of Shea Trees Cosmetic-Relevant Biomarkers

Each of these compounds, identified as discriminating biomarkers in shea butter, has complementary and synergistic properties in cosmetics and dermocosmetics. Their numerous applications have been demonstrated by several scientific studies. For example, Yang et al. (2016) reported on the powerful antioxidant and wound-healing properties of gallic acid. They demonstrated that gallic acid increases the expression of antioxidant genes and promotes skin repair, even under conditions of oxidative stress. Monteiro e Silva et al. (2017) corroborated these results by directly demonstrating the antioxidant effect on the outermost layer of the skin (the stratum corneum) using a non-invasive method. They strongly recommended its use in dermocosmetic products such as body lotions, ointments and gels. Furthermore, the US Food and Drug Administration has recognised it as generally recognised as safe (GRAS) and it is used as an alternative to corticosteroids in dermatoses (Saini et al., 2025). Gallocatechin and epigallocatechin, which are flavonoid compounds generally recognised for their cell regeneration, antioxidant and anti-inflammatory properties, have anti-ageing and revitalising properties and increase epidermal hydration. While it has not been quantified in the present study, the presence of epigallocatechin gallate in this family offers protection against UV, apoptosis and inhibits melanogenesis (useful for the production of lightening cosmetic products) (Braegelmann et al., 2022; Capasso et al., 2025; Di et al., 2022; Kim et al., 2018; Wisuitiprot et al., 2011; Zhu et al., 2000). The identified triterpenes (butyrospermol and parkeol) offer anti-inflammatory, healing and regenerative properties, as well as anticancer and skin-soothing effects (Burnett et al., 2024; Gill et al., 2016; Maanikuu and Peker, 2017; Metelmann et al., 2012). Alpha-tocopherol is widely used in cosmetics due to its moisturising, repairing, anti-ageing and skin barrier effects. It protects cell membranes from free radicals, neutralises UV-induced damage, supports healing and prevents pigmentation imbalances (Keen and Hassan, 2016). Furthermore, studies have shown that the effects of alpha-tocopherol are enhanced in combination with vitamin C and ferulic acid (Burns et al., 2013; Lin et al., 2003;

Wu et al., 2013). The phytosterols identified as potential biomarkers in this study are related to the aforementioned cosmetic and dermocosmetic properties. Campesterol and β-sitosterol, for example, are known to strengthen the skin barrier, improve epidermal lipid cohesion and reduce transepidermal water loss (TEWL) (Rani et al., 2025; Takeda et al., 2021). Gupta (2020) recently highlighted one of the most significant applications of these compounds : β-sitosterol's ability to reduce bad cholesterol, thereby lowering the risk of coronary heart disease, heart attacks, and atherosclerosis, while also playing a role in preventing many cancers. Additionally, their steroid structures can stabilise cosmetic formulations and improve skin tolerance (Chang et al., 2023).

The unique combination of compounds found in certain shea tree varieties offers the chance to create innovative and distinctive formulations. Their rich potential justifies their targeted use in high-value formulations with a range of properties, including antioxidant protection and deep skin regeneration, as well as combating bacterial, fungal, carcinogenic, viral and allergic diseases.

5. Conclusion and perspective

This study employed a non-targeted metabolomics approach, coupled with targeted analysis of bioactive metabolites, to characterize the phytochemical diversity of shea trees (*Vitellaria paradoxa*) in Benin. The analysis focused particularly on compounds with cosmetic and dermocosmetic potential. The analysis revealed a significant presence of chemical families such as polyketides, terpenoids, flavonoids and phenylpropanoids, which are renowned for their beneficial biological properties for the skin, including antioxidant, anti-inflammatory and antimicrobial properties. Several of the 15 targeted metabolites stood out for their high concentrations, particularly α-tocopherol, phytosterols and triterpenes, which far exceeded the levels reported in previous studies. Principal component analysis (PCA), combined with PLS-DA and Random Forest methods, enabled three groups of shea trees with distinct chemical profiles to be discriminated, regardless of their geographical origin. Eight metabolites (gallic acid, gallocatechin, epigallocatechin, campesterol, α-tocopherol, β-sitosterol, butyrospermol and parkeol) were identified as discriminating biomarkers with strong classification power (area under the curve (AUC) $\geq$ 0.75). Each group is characterised by a distinct composition of bioactive metabolites, indicating distinct functional potential. Finally, while no significant effect of geographical origin was observed, two accessions (BG037 and BG143) showed great potential for developing high-value cosmetic products.

These results emphasize the importance of utilizing the intraspecific chemical diversity of shea butter for selection, conservation and innovation in natural cosmetics. However, further studies incorporating multi-year data and genotype-by-environment analyses are needed to confirm the stability of these metabolic profiles. Despite the wealth of data generated, this study has certain limitations that should be highlighted. While the metabolomic approach is effective in revealing differences in chemical profiles, it does not allow for the precise identification of the genetic mechanisms underlying the biosynthesis of the observed bioactive metabolites. To identify, annotate and validate the genes responsible for regulating these compounds in shea, a multi-omics approach integrating genomics, transcriptomics and biochemical data is essential. Furthermore, the absence of a clear geographical signature in this study suggests that environmental factors may play a role, and these should be incorporated as covariates in a future genotype x environment (GxE) interaction study. A more robust study would include a larger number of trees spread across West Africa with multi-year metabolic data monitoring, making it possible to better isolate genetic effects from environmental noise and increase genetic gain in selection. Finally, the metabolic biomarkers identified in this study will require experimental validation, particularly through functional tests and genetic correlation studies, to confirm their role in chemical discrimination and relevance to shea improvement programs.

Credit authorship contribution statement

Guillaume Bodjrenou: Conceptualization, Investigation, Methodology, Data curation, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Jérémy Grosjean**: Methodology, Data curation, Writing – review & editing. **Alexandre Olry**: Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Fernande Honfo**: Methodology, Supervision, Validation, Writing – review & editing. **Alain Hehn**: Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. **Enoch G. Achigan-Dako**: Funding acquisition, Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Funding

This work was fully supported by the PEA-BIOVALOR project: *Renforcement de la démarche compétences et de la culture entrepreneuriale dans la formation supérieure agronomique : leviers pour l'insertion professionnelle et le développement de la bioéconomie au Bénin*, financed by the Agence Française de Développement (AFD) through the Agence Nationale de Recherche (ANR).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The manuscript is approved by all authors for publication.

Acknowledgments

We express our deep gratitude to the Laboratory of Agronomy and Environment (LAE) at the University of Lorraine (UL/Nancy) for its valuable support in carrying out the experimental activities, as well as for providing access to its PASM platform for UHPLC-Orbitrap analysis. We would also like to warmly thank the Genetics, Biotechnology, and Seed Sciences Unit (GBioS) at the University of Abomey-Calavi for its crucial support in collecting shea nut samples in the parklands of northern Benin.

Appendix A. Supplementary material

Supplementary data associated with this article can be online version at https://doi:10.1016/j.indcrop.2025.122354

Data Availability

Data will be made available on request.

CHAPTER 5

Multi-models genome-wide association study revealed candidate genomic regions linked with bioactive metabolites in shea tree (*Vitellaria paradoxa* C.F. Gaertn.)

Guillaume J. Bodjrenou[1,2], Alexandre Olry[1], Dedeou A. Tchokponhoue[2], Fernande G. Honfo[3], Alain Hehn[1], Enoch G. Achigan-Dako[2*]

[1]Université de Lorraine, INRAE, LAE, 54000 Nancy, France

[2]Genetics, Biotechnology and Seed sciences Unit, Faculty of Agronomic Sciences, University of Abomey-Calavi (GBioS/UAC), Abomey-Calavi, Republic of Benin

[3] Laboratoire des Sciences et Technologie des Aliments, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi (LaSTA/UAC), République du Bénin

Abstract

Bioactive compounds such as tocopherols, phytosterols, triterpenes and triterpene alcohols, phenol derivative gallic acid, and flavonoid polyphenols are naturally found in Shea tree (*Vitellaria paradoxa*) and valued in cosmetics, dermocosmetics, and pharmaceutical industries. To expand our knowledge of the genetic determinants underlying variation in these metabolites, 52 shea trees selected across five phytogeographical regions in Benin were analyzed using UHPLC–Orbitrap high-resolution mass spectrometry. They were further genotyped using genotyping-by-sequencing (DArTseq™) that produced SNPs, filtered for quality, and aligned to the *V. paradoxa* reference genome. Genome-wide association studies were conducted using multilocus models (MLMM, FarmCPU, and BLINK) and compared with conventional single-locus models (GLM and MLM). Population structure revealed shallow differentiation with no distinct cluster, while overall genetic diversity was high (Ho = 0.27; He = 0.28; PIC = 0.36), providing robust power for association analyses. Marker-trait associations (MTAs) were consistently identified across the GWAS models. For instance, Chr7_2247894 and Chr12_9908474 were associated with allantoin, Chr8_19176240 and Chr12_32881566 associated with β-amyrin, Chr5_42200948 and Chr6_20018672 associated with epigallocatechin. Additional MTAs, including Chr5_57215940 for gallocatechin and Chr12_2602746 for lupeol, were detected. Trans-cinnamic acid was linked with Chr3_53275580, Chr5_56427098, and Chr10_14462034. Polygenic traits, including allantoin and flavan-3-ols (epigallocatechin, gallocatechin), were influenced by multiple small- to moderate-effect loci across several chromosomes, whereas major-effect loci controlling lupeol and β-amyrin were identified on chromosomes 8 and 12. Candidate gene annotation highlighted regulatory, transport, and compartmentalization processes rather than structural biosynthetic enzymes as central drivers of metabolite variation. These results provide the first genome-wide insights into the genetic architecture of cosmetic-related metabolites in shea.

Keywords: *Vitellaria paradoxa*, GWAS, bioactive metabolites, polygenic traits, triterpenes, regulatory genes

1. Introduction

Vitellaria paradoxa C.F. Gaertn. (shea tree) is a major plant genetic resource that thrives in sub-Saharan Africa. Although it is frequently classified as an orphan crop, its socio-economic importance is substantial, largely due to the butter extracted from its kernels. Shea butter provides income for rural communities and producing countries. Globally, the shea butter market was valued at USD 2.63 billion in 2025 and is projected to reach USD 5.28 billion by 2034, corresponding to a compound annual growth rate (CAGR) of 8.05% (Fortune Business Insights, 2025). This growth is primarily driven by increasing demand for raw and minimally processed shea butter, which is valued for its functional properties, including moisturising, regenerative, wound-healing, antioxidant, and skin-protective effects.

The biological and cosmetic properties of shea butter are largely attributed to the diversity and concentration of its bioactive compounds. Previous studies have reported the presence and variability of phenolic compounds, triterpenes and triterpene alcohols, tocopherols, phytosterols, as well as polyphenols and flavonoids (Akihisa et al., 2010a; Burnett et al., 2024; Capasso et al., 2025; Fiume et al., 2018; Maranz et al., 2003a). More recently, Bodjrenou et al. (2026) quantified fifteen metabolites of cosmetic interest—including allantoin, gallic acid, gallocatechin, epigallocatechin, trans-cinnamic acid, several pentacyclic triterpenes (α-amyrin, β-amyrin, lupeol, butyrospermol, and parkeol), phytosterols (campesterol, stigmasterol, and β-sitosterol), and tocopherols (α and γ), and revealed substantial inter-individual variation within shea germplasm of Benin. This pronounced metabolic diversity suggests the existence of underlying genetic determinants that remain largely unexplored, despite their relevance for genetic improvement and valorization of the species.

Despite growing interest in shea bioactive metabolites and their importance for butter quality, no study has yet investigated the genetic basis of variation in the concentration of these compounds or identified the genomic regions and genes involved in their biosynthesis. To date, genome – wide association studies (GWAS) in shea have mainly focused on genomic regions associated with variation in fat and fatty acid composition (Attikora et al., 2025). This knowledge gap represents a major constraint for the development of genomics-assisted breeding and conservation strategies. Without a clear understanding of the genetic architecture underlying cosmetic-related metabolites, it remains difficult to identify functional genes, develop molecular markers, or implement targeted improvement programs (Fang and Luo, 2019; Luo, 2015). This challenge is compounded by the limited genomic resources available for shea and the relatively small sizes of phenotyped populations typically used in this species (Hale et al., 2021).

Genome-wide association studies are inherently sensitive to several confounding factors that can lead to false positives or reduced statistical power, including population structure, kinship among individuals, sample size, SNP marker density and quality, linkage disequilibrium (LD), and environmentally driven phenotypic variation (Korte and Farlow, 2013; Tang et al., 2016). To address these issues, modern GWAS approaches incorporate appropriate corrections, such as population structure covariates (Q-matrix or principal components), kinship matrices, and environmental covariates, within robust statistical models (Wang and Zhang, 2021).

The Genomic Association and Prediction Integrated Tool (GAPIT) is one of the most widely used software packages for GWAS and genomic selection analyses in plants (Lipka et al., 2012). Recent versions of GAPIT implement advanced multilocus GWAS methods, including Multiple-Locus Mixed linear Model (MLMM), the Fixed and Random Model Circulating Probability Unification (FarmCPU), and the Bayesian Information and Linkage Disequilibrium Iteratively Nested Keyway (BLINK). These models allow the simultaneous inclusion of multiple SNPs in the association framework, improving control of false positives, statistical power, and computational efficiency, particularly for complex traits controlled by multiple loci (Wang and Zhang, 2021). Each method has demonstrated strong performance in detecting small-effect quantitative trait loci (QTLs) across a range of agronomic and biochemical traits in crop species (Hong et al., 2025; Huang et al., 2019; Zeng et al., 2025).

In this context, the present study addresses the following research questions: (i) which genomic variants are associated with variation in the concentration of cosmetic-related bioactive compounds in shea; (ii) the extent to which GWAS models such as GLM, MLM, MLMM, FarmCPU, and BLINK are suitable to identify genomic regions linked with any of the fifteen metabolites analyzed. The overall objective of this study is to elucidate the genetic basis of variation in cosmetic-related bioactive compounds in shea germplasm from Benin. Specifically, we aim to conduct a GWAS on multiple metabolites of cosmetic interest, identify significant QTLs, annotate genes located within the corresponding QTL regions, and propose candidate genes potentially involved in the biosynthetic pathways of these compounds. We hypothesize that (i) variation in the concentration of bioactive compounds is partly under genetic control and detectable through GWAS, (ii) multilocus models such as MLMM, FarmCPU, and BLINK enable robust detection of associated QTLs and genes involved in metabolic pathways relevant to cosmetic bioactive compounds.

2. Materials and Methods

2.1. Plant Material and Plant Extracts

The plant material consisted of 52 shea tree individuals sampled in the northern region of Benin (Supplementary Table 1) during 2022, with a minimum distance of 100 meters between trees. These trees were previously included in a metabolomic evaluation study focusing on compounds of cosmetic interest in Benin (Bodjrenou et al., 2026b).

2.2. Quantification of Bioactive Compounds

Bioactive compound extraction, UHPLC–HRMS (Orbitrap ID-X) analysis, metabolite identification, and quantification were performed following the protocol previously described by Bodjrenou et al. (2026). Briefly, metabolites were extracted from kernels collected from phenotyped shea trees using a methanol–water-based extraction procedure. The resulting extracts were analyzed by ultra-high-performance liquid chromatography coupled with high-resolution Orbitrap mass spectrometry (UHPLC–Orbitrap HRMS). Compound identification was based on accurate mass measurements, retention times, and MS/MS fragmentation patterns by comparison with authentic standards.

fifteen metabolites of cosmetic and dermocosmetic interest—allantoin, β-amyrin, campesterol, epigallocatechin, gallocatechin, lupeol, and trans-cinnamic acid, gallic acid, butyrospermol, parkeol, stigmasterol, β-sitosterol, tocopherols (α and γ) —were included from the reference study and quantified using external calibration curves, as detailed in Bodjrenou et al. (2026).

2.2. Genomic data acquisition through Genotyping by Sequencing (GBS)

Leaf samples from shea trees were sent to SEQART AFRICA at the International Livestock Research Institute (Nairobi) for genotyping. DNA was extracted using the Nucleomag® Plant DNA extraction kit, yielding 50–100 ng/µL, and quality was assessed on 0.8% agarose gels. Libraries were prepared following the DArTseq complexity reduction protocol (Kilian et al., 2012) by digesting genomic DNA with PstI and MseI, ligating barcoded and common adapters, and PCR-amplifying adapter-ligated fragments. Single-read sequencing (138 cycles) was performed on a NovaSeq X platform.

Genotyping was conducted using DArTseq™ technology, providing rapid and cost-effective genome profiling. Marker scoring was performed with DArTsoft14, generating two types of markers: SilicoDArT and SNPs, both scored as binary presence/absence (1/0). Marker positions were aligned to the *V. paradoxa* HiCP0 genome assembly to determine chromosomal locations.

2.3. SNP Data processing

SNP markers were filtered to ensure high-quality genotypic data with R software using the *dartR* package. Markers with a call rate below 95% and individuals with more than 10% missing data were excluded. SNP markers with a minor allele frequency (MAF) <0.05 and monomorphic markers were removed, as they provide limited information for population genetics or association analyses.

2.5. Population structure and relatedness

Population structure was assessed using principal coordinate analysis (PCoA) implemented via the *gl.pcoa* function of the *dartR* package, capturing major axes of genetic variation among individuals. The resulting coordinates were included as covariates in association models to correct for population stratification. Pairwise genetic relatedness was estimated using a kinship matrix (after incorporating Linkage disequilibrium information, LD = 0.2) computed with the *gl.grm* function of *dartR*, which was incorporated in the Mixed Linear Model (MLM) model to account for familial relationships and reduce confounding due to shared ancestry.

3. Genome-wide association analysis (GWAS)

Recent multilocus GWAS methods were employed in this study, enabling the simultaneous inclusion of multiple SNPs as fixed and/or random effects in the model. Specifically, the MLMM, FarmCPU and BLINK approaches were applied. These methods have been widely adopted because of their ability to detect quantitative trait loci (QTLs) with small effects, reduce false-positive associations arising from population structure or linkage among SNPs, and account for SNP interactions and correlations through linkage disequilibrium (LD) and iterative model updating (Huang et al., 2019; Liu et al., 2016; Wang and Zhang, 2021).

In the MLMM framework, the phenotype is modeled as:

$$y = X_{SNP}\beta_{SNP} + X_{cof}\beta_{cof} + X_Q\beta_Q + ZKZ^T + e$$

where $X_{SNP}\beta_{SNP}$ represents the fixed effect of the tested SNP, $X_{cof}\beta_{cof}$ the fixed effects of iteratively selected cofactor SNPs, $X_Q\beta_Q$ the fixed effects of population structure, ZKZ^T the random genetic effect captured by the kinship matrix, and e the residual error.

In FarmCPU, the phenotype y is modelled as:

$$y = X_{SNP}\beta_{SNP} + X_{PQTNs}\beta_{PQTNs} + e$$

$$y = ZKZ^T + e$$

where $X_{SNP}\beta_{SNP}$ represents the fixed effect of candidate SNP markers, $X_{PQTNs}\beta_{PQTNs}$ represents the fixed effects of pseudo-QTLs identified during iterations, ZKZ^T accounts for the random effect of the kinship matrix, capturing residual genetic variance, and e represents residual errors (Liu et al., 2016; Wang and Zhang, 2021). FarmCPU operates iteratively: significant markers identified in the fixed-effect model are added as pseudo-QTLs in subsequent rounds, while the random-effect model estimates the variance of remaining loci, taking into account population structure and relatedness. This alternation continues until convergence, providing an optimal balance between statistical power and false positive control. By explicitly separating major and minor genetic effects and incorporating kinship, FarmCPU reduces spurious associations due to population structure and linkage disequilibrium among markers, making it highly effective for detecting QTLs of small effect in structured population (Liu et al., 2016; Wang and Zhang, 2021).

BLINK is an extension of FarmCPU designed to optimize the selection of candidate markers (Huang et al., 2019; Wang and Zhang, 2021). It uses the Bayesian Information Criterion (BIC) to identify the most informative SNPs and explicitly accounts for linkage disequilibrium to avoid redundancy among highly correlated markers. Similar to FarmCPU, BLINK relies on an iterative framework based on fixed-effect models, but with improved computational efficiency. In BLINK, the phenotype y is modelled as

$$y = X_{SNP}\beta_{SNP} + X_S\beta_s + e$$

Where is the vector of phenotypic values, $X_{SNP}\beta_{SNP}$ the fixed effect of the tested candidate SNP, $X_S\beta_S$ the fixed effects of the iteratively selected pseudo-QTNs (set S), and e the residual errors.

To evaluate the relative performance of multilocus GWAS approaches, results obtained using the MLMM, FarmCPU, and BLINK models were compared with those generated by conventional single-locus models, namely the General Linear Model (GLM) and MLM. Model performance was assessed using quantile–quantile (QQ) plots, which examine deviations from the expected diagonal under the null hypothesis, as well as the genomic inflation factor (λ), which was used to quantify the extent of test statistic inflation and to evaluate model adequacy in controlling for population structure and relatedness.

3.3. Identification and annotation of candidate genes surrounding significant SNPs

3.3.1. Extraction of significant SNPs

Significant SNPs were identified from the GWAS results obtained using each of the three models (MLMM, FarmCPU, and BLINK) implemented in the GAPIT package. P-values were adjusted for multiple testing using the Benjamini–Hochberg procedure, and SNPs with a false discovery rate (FDR) < 0.05 were retained for downstream analyses. The SNP showing the strongest association signal was designated as the "lead SNP" and was used to illustrate the candidate gene annotation pipeline.

3.3.2. Definition of the QTL region

For each significant SNP, a genomic window of ±100 kb was defined to capture potentially locally regulated genes, in accordance with previous studies indicating that the majority of cis-regulatory effects occur within approximately 100 kb of the associated variant (Korte and Farlow, 2013). Consequently, the quantitative trait locus (QTL) region for the lead SNP extended from 100,000 bp upstream to 100,000 bp downstream of the SNP position.

3.3.3. Candidate gene annotation

Genes located within the ±100 kb window surrounding the lead SNP were extracted from the *Vitellaria paradoxa* genome annotation (GFF3 format; ORCAE database: https://bioinformatics.psb.ugent.be/gdb/vitellaria_paradoxa/) using the *rtracklayer* and *GenomicRanges* R packages. Only features annotated as "gene" were retained. Functional annotations, including gene descriptions, Gene Ontology (GO) terms, and predicted protein domains, were integrated for the identified genes. GO terms provide a standardized framework for describing gene-associated molecular functions, biological processes, and cellular components. Candidate genes were subsequently ranked according to their physical distance from the lead SNP to facilitate the identification of the most biologically relevant candidates.

3.4. Allelic effect analysis

For each trait, GWAS results from multiple models (MLMM, FarmCPU, BLINK) were considered. The top SNP for each major signal was selected based on the lowest p-value across all models. Genotypes at each SNP were converted into additive dosages (0, 1, 2) corresponding to homozygous reference, heterozygous, and homozygous alternative individuals.

To quantify genetic effects, linear models including both additive (Dosage) and dominance (heterozygosity) terms were fitted for each SNP-trait combination. Additive effects (A) and dominance effects (D) were estimated, and allelic substitution effects (α) were calculated as:

$$\alpha = A + D \cdot (q - p)$$

where p is the frequency of the alternative allele and $q = 1 - p$ (Beavis et al., 2023). Phenotypic values with missing genotypes or missing trait data were excluded. Descriptive statistics (mean, standard deviation, sample size) were computed for each genotype class. For traits in which the same top SNP was detected across multiple models, results were combined into a single plot to compare allelic effects across methods. Violin plots with overlaid medians and jittered individual points were used for visualization.

All statistical analyses were performed using R software version 4.5.2.

4. Results

4.1. Metabolite concentration variation

Metabolite concentrations (μmol·L⁻¹) showed substantial variability across the population. Low-abundance compounds such as allantoin, gallocatechin, and epigallocatechin exhibited narrow concentration ranges and low mean values, whereas sterols and triterpenes displayed markedly higher concentrations and broader distributions (Table 1).

High-abundance metabolites, including campesterol (mean = 15,473.02 ± 479.9), α-tocopherol (17,521.34 ± 422), β-sitosterol (1,622.30 ± 242.95), and Stigmasterol (1,854.70 ± 39.06) showed large standard deviations, indicating substantial phenotypic heterogeneity (Table 1).

Several metabolites, such as trans-cinnamic acid, butyrospermol, and lupeol, exhibited wide min–max ranges and high variability, suggesting strong dispersion and potential genetic control. In contrast, metabolites like Allantoin and Epigallocatechin showed relatively limited dispersion (Table 1). Overall, the observed variability supports the suitability of these traits for genetic association and allelic effect analyses.

Table 1: Descriptive statistics of metabolite concentrations (μmol·L⁻¹) across the shea tree population

Metabolites	min	max	mean	SE.mean	std.dev
allantoin	0.04	15.36	2.49	0.32	2.29
gallic acid	113.26	1059.48	594.4	34.83	251.16
gallocatechin	0.56	51.69	5.7	1.23	8.88
stigmasterol	991.08	2424.16	1854.7	39.06	281.67
epigallocatechin	0.96	26.43	6.8	0.65	4.71
campesterol	10801.74	28634.16	15473.02	479.9	3460.61
trans cinnamic acid	29.29	2445.48	240.15	45.31	326.74
α-amyrin	10.29	214.77	50.08	4.96	35.76
β-amyrin	1.29	86.14	24.43	2.77	19.96
α-tocopherol	10652.48	23426.27	17521.34	422	3043.07
γ-tocopherol	2.75	100.22	33.79	2.61	18.81
β-sitosterol	82.85	7744.35	1622.3	242.95	1751.94
lupeol	4.93	239.06	54.73	4.87	35.15
butyrospermol	9.74	776.51	150.06	26.2	188.96
parkeol	3.77	54.53	20.94	1.43	10.31

4.2. Distribution of SNPs and Genetic Diversity across Shea Tree Populations

After quality filtering, a total of 8,112 SNP markers were retained from an initial set of 61,885. SNP density was evaluated across the 12 chromosomes of *V. paradoxa* using non-overlapping 1-Mb windows. The resulting genome-wide map revealed a highly heterogeneous distribution of SNPs, with substantial variation both among chromosomes and along individual chromosomes (Figure 1). Most genomic windows exhibited low to moderate SNP density (7–18 SNPs per Mb), corresponding to the Low and Moderate density classes. In contrast, regions with high (19–23 SNPs per Mb) and very high (>23 SNPs per Mb) SNP density were sporadically distributed across several chromosomes, particularly chr02, chr03, chr06, chr07, and chr08 (Figure 1). Conversely, multiple chromosomal segments showed very low SNP density (1–6 SNPs per Mb), suggesting regions with limited polymorphism or reduced marker coverage.

Observed (Ho = 0.27) and expected (He = 0.28) heterozygosity values were high and comparable, indicating substantial genetic diversity within *V. paradoxa* populations. The average Polymorphism Information Content (PIC) was 0.36.

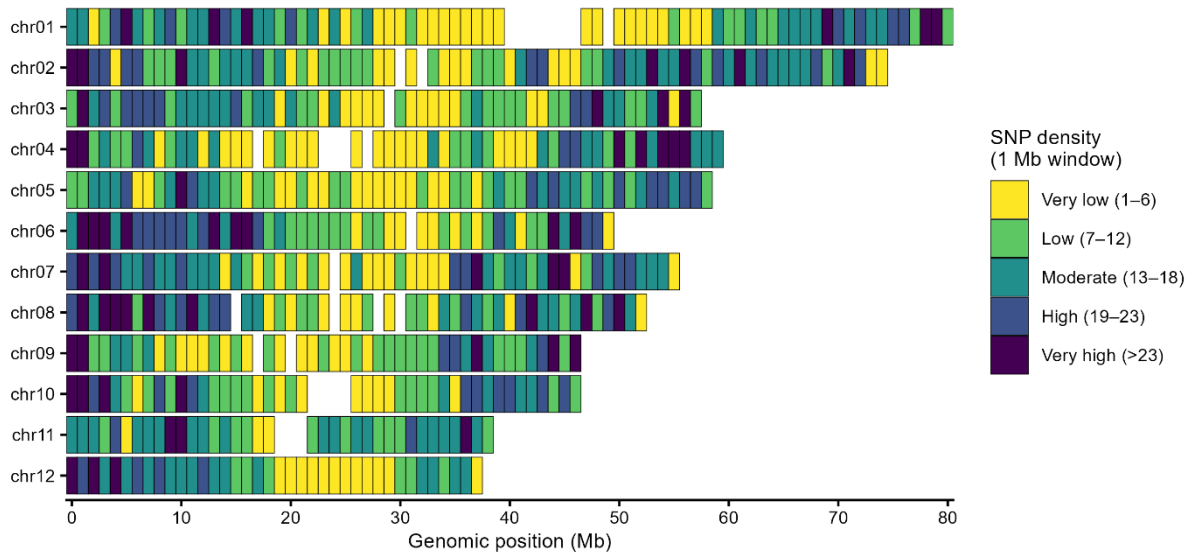


Figure 1: Genome-wide distribution of SNP density across *V. paradoxa* chromosomes using 1-Mb windows

4.3. Population Structure and Relatedness

Principal coordinates analysis (PCoA) showed that the first axis explained 3.09% of the total genetic variation, while the second axis accounted for 2.68%. The cumulative variance explained by the first two axes was 5.77% (Supplementary Figure 1). The low proportion of variance explained by the first PCoA axes indicates a weak genetic structure, with genetic variation distributed across multiple dimensions. The genomic kinship matrix was estimated using LD-pruned SNP data ($r^2 = 0.2$) to minimize the effect of marker redundancy. The resulting matrix revealed predominantly low to moderate pairwise relatedness among individuals, with most kinship coefficients centered around zero (Supplementary Table 2). Only a limited number of individual pairs exhibited elevated kinship values, indicating the absence of strong familial structure within the population.

4.4 GWAS association

A total of 8,112 genetic variants were tested across all GWAS models. The results presented here focus on the eight metabolites for which at least one significant association was detected across the GWAS models (MLMM, FarmCPU, BLINK). Although GWAS was conducted on 15 metabolites, no significant associations were observed for the remaining seven, and thus they are not included in this summary. The analyses performed using the MLMM, FarmCPU and BLINK models identified several significant associations with multiple metabolic traits across the fifteen tested bioactive compounds (allantoin, β-amyrin, campesterol,

epigallocatechin, gallocatechin, lupeol, trans-cinnamic acid, and stigmasterol). To control for multiple testing, false discovery rate (FDR) correction was applied using the method of Benjamini and Hochberg (1995) and only markers that passed this correction were retained. Manhattan plots display SNPs exceeding the genome-wide Bonferroni threshold, whereas association tables report all SNPs significant after FDR correction. Consequently, some FDR-significant associations are not visible in the plots

For allantoin content, the genome-wide association analysis conducted with the MLMM revealed one significant SNP association on chromosome 7. Using the FarmCPU model, seven (07) loci exceeding the genome-wide significance threshold (FDR-adjusted p-values < 0.05) were identified (Table 2). The strongest association was detected on chromosome 8, where the lead SNP reached a $-\log_{10}(P) \approx 8.0$ (Figure 2). Significant loci were distributed across several chromosomes (chr1, chr7, chr8, chr10, chr11, and chr12), suggesting a polygenic genetic architecture underlying variation in allantoin content. Using the BLINK model, significant associations for the same trait were detected on chromosomes 2, 7, and 12, with two chromosomes (chr7 and chr12) overlapping with those identified by the FarmCPU model (Figure 3), further supporting the robustness of this association and indicating that it is unlikely to result from multiple-testing artifacts (Table 2).

In total, the three models identified 12 significant SNP markers associated with allantoin content, with two markers shared between the models and the lead SNP located on chromosomes 8 and 7, respectively, for the FarmCPU and BLINK models (Table 2).

Table 2: Significant genome-wide association results for seven metabolic traits in 52 *Vitellaria paradoxa* samples identified using MLMM, FarmCPU and BLINK models, including adjusted p-values and overlapping SNPs

Traits	Models	SNP	Chr	Pos	P.value	MAF	effect	H&B.P.Value	Overlapping (SNP)
allantoin	MLMM	100077840-6-G/A	7	2247894	2.78E-07	0.07692308	-3.541231	0.00225445	100077840-6-G/A 100218985-33-T/C
	FarmCPU	100189376-56-C/A	8	46418346	9.55E-09	0.03846154	-2.2824394	7.75E-05	
		100077840-6-G/A	7	2247894	4.09E-07	0.07692308	-1.23292587	0.00121402	
		100214430-9-A/G	10	10867116	4.49E-07	0.03846154	-2.14827347	0.00121402	
		100137606-34-A/C	1	5204894	1.64E-06	0.03846154	-1.75589981	0.00265751	
		100218985-33-T/C	12	9908474	1.53E-06	0.05769231	-1.21191405	0.00265751	
		100111389-53-T/C	1	73282408	1.41E-05	0.30769231	-0.52683715	0.01909592	
		100111996-67-C/T	11	3142167	3.39E-05	0.06730769	-1.16636849	0.03924497	
	BLINK	100077840-6-G/A	7	2247894	2.54E-14	0.07692308	-3.47512779	2.06E-10	
		100145659-66-A/T	12	33725044	8.59E-10	0.13461538	-1.65811417	3.48E-06	
		100074646-15-A/C	2	59635751	3.48E-07	0.05769231	1.746706233	0.00094029	
		100218985-33-T/C	12	9908474	1.17E-06	0.05769231	-1.57313469	0.00237602	
β-.amyrin	MLMM	100155037-34-T/G	8	19176240	6.52E-07	0.05769231	-32.21889	0.005292787	100155037-34-T/G 100110543-55-G/A
		100213725-54-A/T	1	14394784	3.10E-06	0.03846154	-35.37475	0.012555649	
		100195557-21-A/G	11	6355179	1.14E-05	0.07692308	-20.83696	0.030854346	
	FarmCPU & BLINK	100155037-34-T/G	8	19176240	6.08E-06	0.05769231	-35.3014803	0.04116732	
		100110543-55-G/A	12	32881566	1.01E-05	0.08653846	-25.6205716	0.04116732	
campesterol	MLMM	100185734-25-G/A	5	57215940	1.10E-06	0.06730769	-5669.714	0.008960843	100185734-25-G/A
	FarmCPU	100223580-56-T/G	12	34518123	1.11E-10	0.06730769	-3001.04124	8.99E-07	
		100146120-66-G/A	11	26263224	1.62E-09	0.13461538	-2274.91375	6.57E-06	
		100151083-35-C/G	8	21538674	1.98E-08	0.05769231	3028.93193	5.35E-05	
		100110913-52-G/C	3	20465903	3.18E-08	0.35576923	-1239.83776	6.46E-05	
		100192581-36-A/T	10	2830907	8.40E-07	0.45192308	-1421.87518	0.00136224	
	BLINK	100218528-20-C/T	8	4197561	1.20E-07	0.18269231	-2127.4538	0.00097007	

Traits	Models	SNP	Chr	Pos	P.value	MAF	effect	H&B.P.Value	Overlapping (SNP)
		100185734-25-G/A	5	57215940	1.27E-05	0.06730769	-3369.30276	0.03488681	
		100231782-27-T/C	7	1697012	1.29E-05	0.06730769	-2917.23978	0.03488681	
epigallocatechin	MLMM	100191722-57-C/T	5	42200948	2.86E-06	0.1153846	-6.099868	0.02320212	100191722-57-C/T 100234631-23-T/C
	FarmCPU	100191722-57-C/T	5	42200948	2.14E-06	0.11538462	-6.09986287	0.01737272	
		100234631-23-T/C	6	20018672	4.67E-06	0.14423077	-5.70939809	0.0189418	
		100157075-25-T/C	6	24270532	1.04E-05	0.06730769	-6.22318518	0.02823053	
	BLINK	100191722-57-C/T	5	42200948	2.03E-07	0.11538462	-4.78217817	0.00164882	
		100234631-23-T/C	6	20018672	4.67E-07	0.14423077	-4.37722293	0.00189366	
gallocatechin	MLMM	100223580-56-T/G	12	34518123	1.28E-08	0.06730769	-11.89546	0.00010369	100185734-25-G/A 100223580-56-T/G
	FarmCPU	100185734-25-G/A	5	57215940	1.58E-14	0.06730769	-9.53159701	1.28E-10	
		100072105-40-A/G	9	1848228	3.28E-14	0.06730769	-9.80804994	1.33E-10	
		100197809-10-A/G	12	33179659	1.40E-09	0.07692308	8.258122518	3.78E-06	
		100071091-14-C/T	7	49771253	5.03E-08	0.19230769	-4.07892994	0.00010193	
		100225211-52-T/C	11	37878289	4.66E-07	0.25961538	-2.40970862	0.00075533	
		100194631-43-A/G	12	10484551	3.60E-06	0.03846154	-6.02039587	0.00486265	
	BLINK	100223580-56-T/G	12	34518123	4.95E-19	0.06730769	-12.4229371	4.01E-15	
		100185734-25-G/A	5	57215940	5.92E-15	0.06730769	-9.24935458	2.40E-11	
		100101704-11-T/A	1	67759722	3.25E-10	0.16346154	4.491703652	8.78E-07	
		100144019-21-T/C	1	59920896	1.59E-08	0.20192308	-3.2846263	3.22E-05	
		100124600-39-G/A	2	31704731	5.90E-08	0.28846154	2.631016356	9.57E-05	
		100077099-33-G/A	1	79308204	9.80E-08	0.08653846	-4.04292164	0.00013247	
		100152417-16-T/C	8	10193225	1.45E-07	0.31730769	-2.40405667	0.00016838	
lupeol	MLMM	100228303-14-T/C	12	2602746	1.09E-07	0.07692308	-49.13113	0.00088737	100228303-14-T/C
	FarmCPU	100228303-14-T/C	12	2602746	4.97E-12	0.07692308	-33.6284621	4.03E-08	
		100195630-27-G/A	8	725349	2.31E-06	0.125	-18.9718332	0.00936911	
		100169167-47-A/G	7	50937855	8.28E-06	0.05769231	-22.4829072	0.02239988	
	BLINK	100228303-14-T/C	12	2602746	1.32E-07	0.07692308	-35.1187133	0.0010694	

Traits	Models	SNP	Chr	Pos	P.value	MAF	effect	H&B.P.Value	Overlapping (SNP)
trans cinnamic acid	MLMM	100186831-32-T/C	3	53275580	7.75E-11	0.03846154	-775.5663	6.29E-07	100200597-22-G/C 100186831-32-T/C 100208314-27-G/C
		100204898-18-G/A	4	46748073	7.20E-07	0.05769231	384.5794	2.04E-03	
		100200597-22-G/C	5	56427098	7.55E-07	0.05769231	-428.5146	2.04E-03	
		100116567-44-A/G	3	17168286	1.50E-05	0.07692308	311.7627	3.05E-02	
		100113488-8-A/G	1	74618177	2.19E-05	0.05769231	342.6327	3.55E-02	
	FarmCPU	100200597-22-G/C	5	56427098	3.15E-15	0.05769231	-486.321609	2.56E-11	
		100186831-32-T/C	3	53275580	3.12E-08	0.03846154	-280.544616	0.00012648	
		100217570-30-T/C	3	13331858	2.61E-06	0.05769231	-173.121719	0.00423784	
		100108333-24-G/T	8	3045789	2.27E-06	0.44230769	-80.4463823	0.00423784	
		100208314-27-G/C	10	14462034	2.52E-06	0.29807692	-80.3548436	0.00423784	
		100082387-46-T/C	3	36115382	4.42E-06	0.05769231	172.9336668	0.00597607	
	BLINK	100200597-22-G/C	5	56427098	2.35E-10	0.05769231	-366.045823	1.91E-06	
		100186831-32-T/C	3	53275580	2.30E-07	0.03846154	-306.607477	0.00062108	
		100208314-27-G/C	10	14462034	2.16E-07	0.29807692	-124.459124	0.00062108	
Stigmasterol	MLMM	100208022-11-G/T	7	18705710	3.44E-07	0.5	260.7306	0.002793442	-
		100080492-58-A/G	1	13789908	1.24E-05	0.3461538	-200.4930	0.049990600	

Both models, FarmCPU (Figure 2) and BLINK (Figure 3), consistently identified two SNPs associated with β-amyrin content, located on chromosomes 8 and 12, with the lead SNP on chromosome 8 (Table 2). The MLMM model identified three SNPs on chromosomes 8, 1, and 11, with the same lead SNP as FarmCPU and BLINK (Supplementary Figure 2). Nine SNPs were significantly associated with campesterol (FDR < 0.05): one identified by MLMM (Supplementary Figure 3), five by FarmCPU (Figure 2), and three by BLINK (Figure 3).

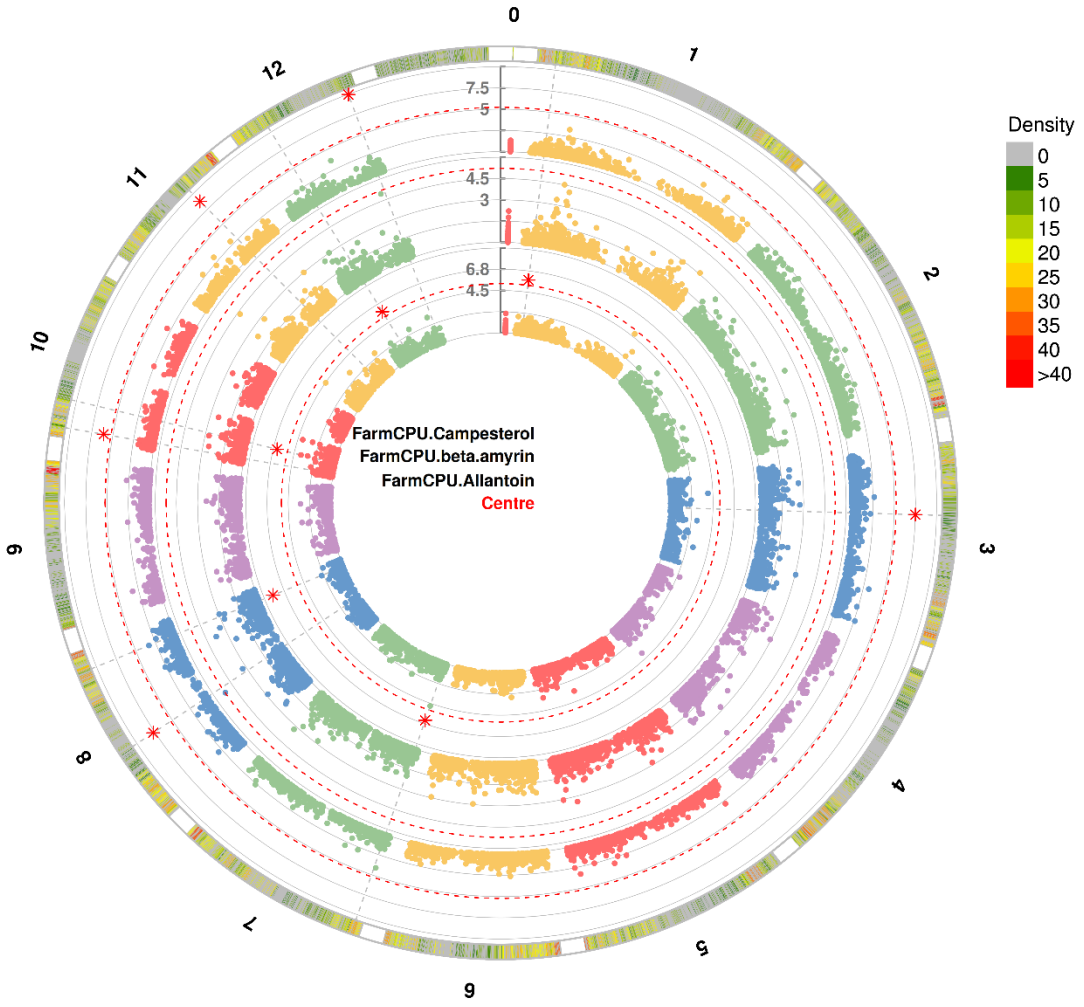


Figure 2. Circular Manhattan plot illustrating multi-trait genome-wide association signals for allantoin, β-amyrin, and campesterol contents in a shea tree (*Vitellaria paradoxa*) population, identified using the FarmCPU model across the 12 chromosomes. Asterisks indicate SNPs exceeding the genome-wide Bonferroni-corrected significance threshold. From the center outward, the concentric circles correspond to allantoin, β-amyrin, and campesterol, respectively.

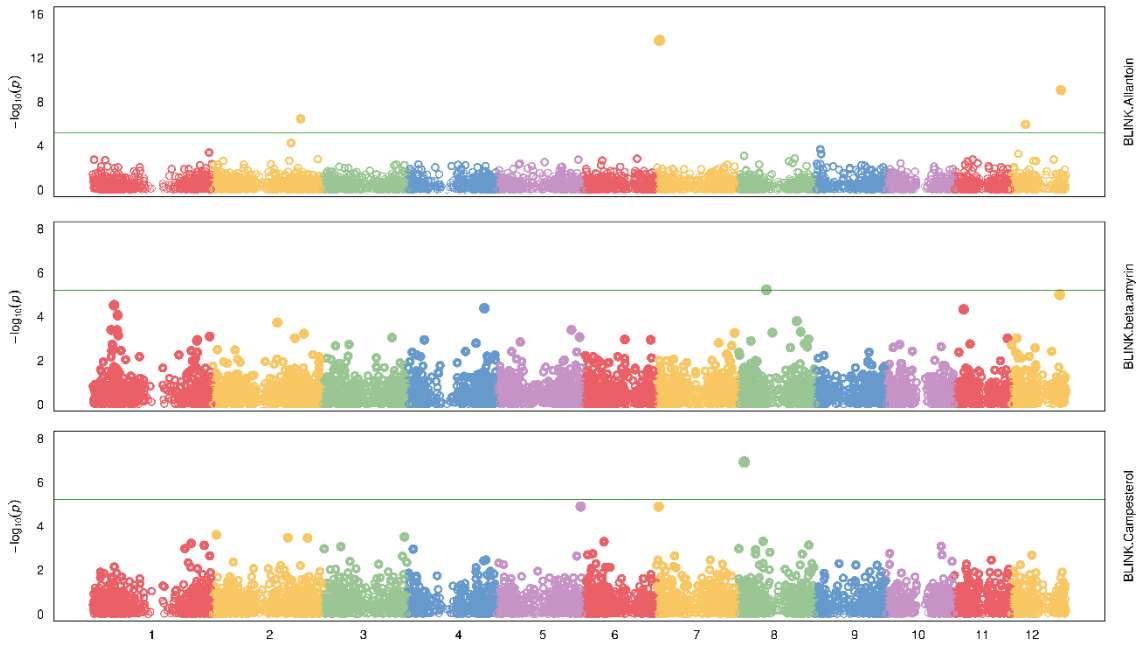


Figure 3. Manhattan plots depicting genome-wide association signals for allantoin, β-amyrin, and campesterol contents in a shea tree (*Vitellaria paradoxa*) population, identified using the BLINK model across the 12 chromosomes. Panels from top to bottom correspond to allantoin, β-amyrin, and campesterol, respectively. The horizontal green line indicates the genome-wide significance threshold, and the y-axis represents $-\log_{10}(p)$ values for SNP associations.

Genome-wide association analysis revealed consistent and robust loci associated with epigallocatechin content, detected by all three MLMM, FarmCPU, and BLINK models on chromosomes 5 and 6. The lead SNP for epigallocatechin was located on chromosome 5 at 42.2 Mb (SNP 100191722-57-C/T), which showed the strongest association in all the models (Table 2). The consistency of this lead SNP across models highlights it as a high-confidence locus underlying epigallocatechin variation. A second lead SNP on chromosome 6 (100234631-23-T/C, 20.0 Mb) was also consistently detected by FarmCPU and BLINK models, while FarmCPU identified an additional nearby SNP (100157075-25-T/C, 24.3 Mb). Across all detected lead SNPs, allelic effects were predominantly negative, suggesting that these loci are associated with reduced epigallocatechin content (Table 2). Overall, the identification of shared lead SNPs across independent GWAS models reinforces the biological relevance of chromosomes 5 and 6 as key genomic regions controlling epigallocatechin biosynthesis in *Vitellaria paradoxa*.

For gallocatechin, 14 significant SNP associations (FDR < 0.05) were identified across the three models. MLMM identified one SNP on chromosome 12 (Table 2), FarmCPU detected six SNPs

on chromosomes 5, 7, 9, 11, and 12 (lead SNP: 100185734-25-G/A, Chr 5) (Figure 4), while BLINK detected seven SNPs on chromosomes 1, 2, 5, 8, and 12 (lead SNP: 100223580-56-T/G/G, Chr 12) (Figure 5). One SNP on chromosome 5 (100185734-25-G/A) was common to FarmCPU and BLINK models (Table 2).

Using the FarmCPU model, three SNPs were significantly associated with lupeol content (Figure 4). These SNPs were located on chromosomes 7 (SNP 100169167-47-A/G), 8 (SNP 100195630-27-G/A), and 12 (SNP 100228303-14-T/C) (Table 2). Among these, SNP 100228303-14-T/C on chromosome 12 was identified as the lead SNP, as it showed the strongest statistical association with lupeol content (p-value = 4.97×10^{-12}; FDR = 4.03×10^{-8}) (Table 2). This SNP also exhibited a large negative allelic effect, indicating that the minor allele was probably associated with reduced lupeol content. Consistently, the MLMM and BLINK models detected only this same SNP (100228303-14-T/C) as significant at the 5% FDR threshold (FDR = 0.00107; Figure 5), further confirming it as the most robust and reliable genetic marker associated with lupeol content in the shea tree population.

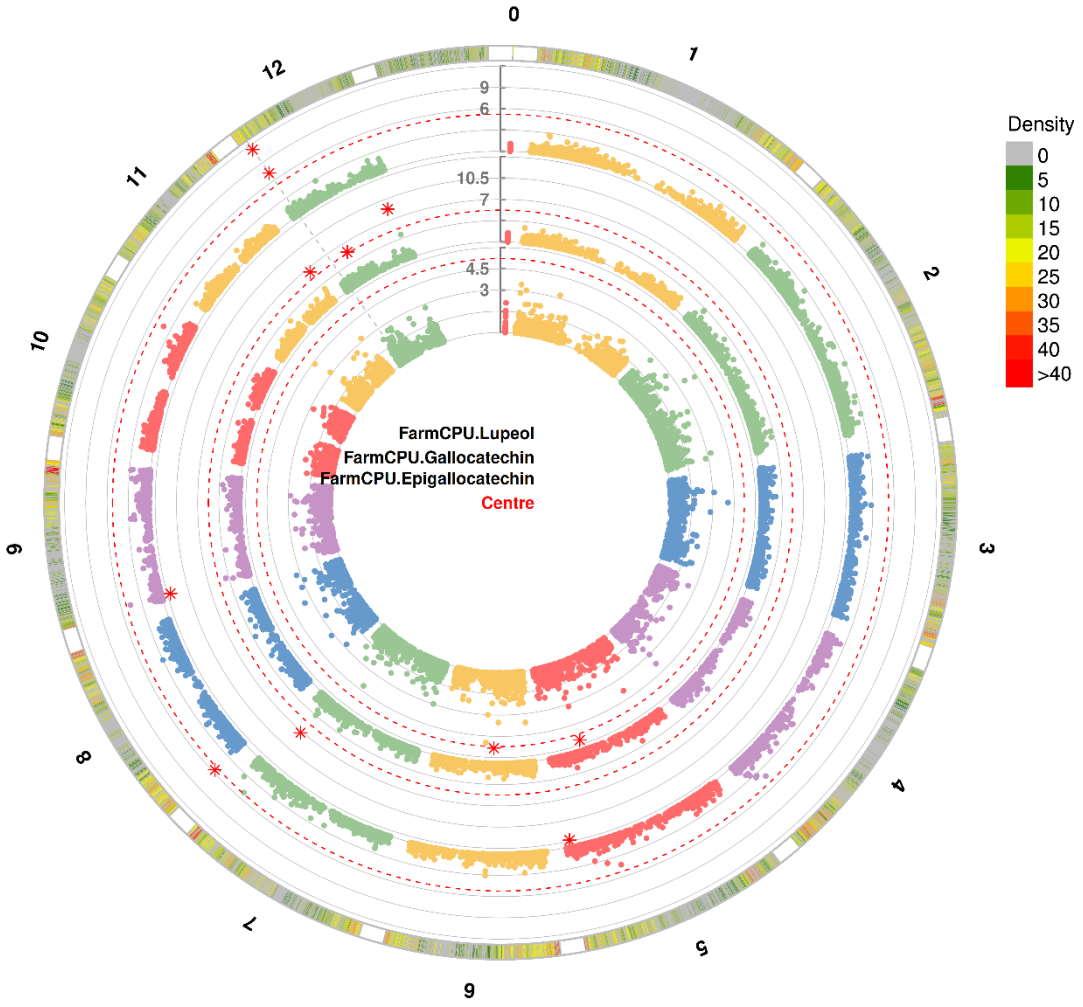


Figure 4. Circular Manhattan plot illustrating multi-trait genome-wide association signals for epigallocatechin, gallocatechin, and lupeol contents in a shea tree (*Vitellaria paradoxa*) population, identified using the FarmCPU model across the 12 chromosomes. Asterisks indicate SNPs exceeding the genome-wide Bonferroni-corrected significance threshold. From the center outward, the concentric circles correspond to epigallocatechin, gallocatechin, and lupeol, respectively.

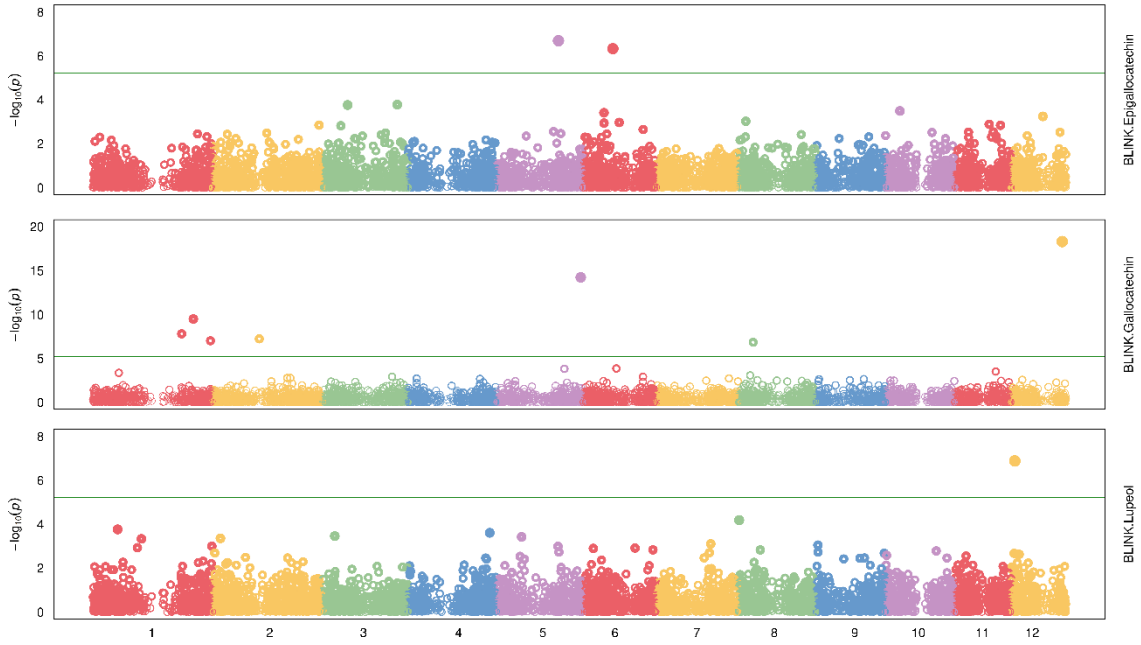


Figure 5. Manhattan plots depicting genome-wide association signals for epigallocatechin, gallocatechin, and lupeol contents in a shea tree (*Vitellaria paradoxa*) population, identified using the BLINK model across the 12 chromosomes. Panels from top to bottom correspond to epigallocatechin, gallocatechin, and lupeol, respectively. The horizontal green line indicates the genome-wide significance threshold, and the y-axis represents $-\log_{10}(p)$ values for SNP associations.

The MLMM model identified five SNPs significantly associated with trans-cinnamic acid content (Table 2), located on chromosomes 1,3,4 and 5 (Table 2). Using the FarmCPU model, six SNPs were found (Figure 6a). These SNPs were located on chromosomes 3, 5, 8, and 10, with three loci detected on chromosome 3 (SNPs 100217570-30-T/C, 100082387-46-T/C, and 100186831-32-T/C) (Table 2). Single significant SNPs were also identified on chromosome 5 (SNP 100200597-22-G/C), chromosome 8 (SNP 100108333-24-G/T), and chromosome 10 (SNP 100208314-27-G/C). Among these loci, SNP 100200597-22-G/C on chromosome 5 showed the strongest association with trans-cinnamic acid content ($P = 3.15 \times 10^{-15}$; FDR = 2.56×10^{-11}) and was therefore considered the lead SNP. Using the BLINK model, three SNPs

were significantly associated with trans-cinnamic acid content (Figure 6b), including SNP 100186831-32-T/C on chromosome 3, SNP 100200597-22-G/C on chromosome 5, and SNP 100208314-27-G/C on chromosome 10 (Table 2). Notably, the lead SNP identified by FarmCPU (100200597-22-G/C) was also detected by BLINK (FDR = 1.91×10^{-6}), providing independent support for the involvement of this genomic region in the genetic control of trans-cinnamic acid content in the shea tree population (Table 2).

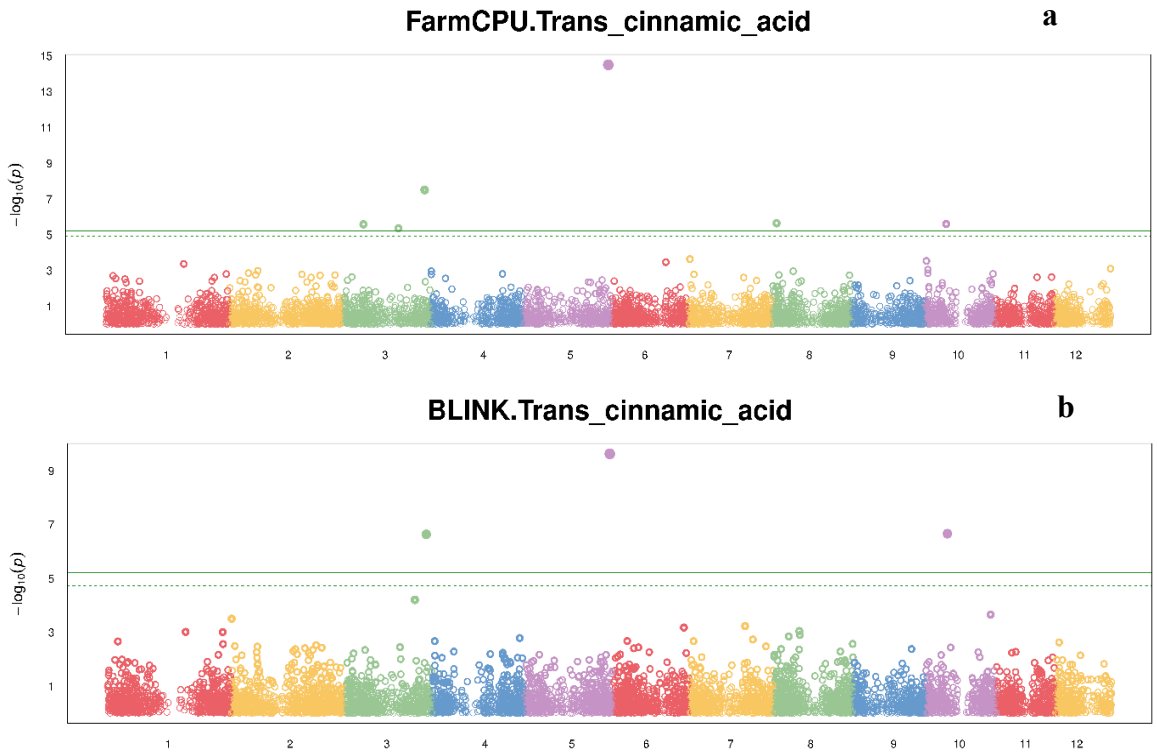


Figure 6: Manhattan plots summarizing genome-wide association results for trans-cinnamic acid content across 12 chromosomes using the FarmCPU and BLINK models. Panels (a) and (b) correspond to FarmCPU and BLINK-based analyses, respectively. The dashed green line indicates the genome-wide significance threshold, and association strength is expressed as $-\log_{10}(p)$.

4.5. Comparison of GWAS Model Performance

Figure 7 presents QQ plots of observed versus expected p-values for three metabolic traits (allantoin, β-amyrin, and campesterol) analyzed using the BLINK and FarmCPU models. The red diagonal line represents the expectation under the null hypothesis of no association, while the shaded area corresponds to the confidence interval. Deviations above this region indicate potential genetic associations.

For all three traits under the BLINK model, the points closely follow the diagonal at low $-\log_{10}(p)$ values, indicating effective control of the false positive rate and limited statistical inflation ($\lambda = 0.9610$). A pronounced deviation in the right tail is observed, particularly for allantoin, with several SNPs reaching very high $-\log_{10}(p)$ values (>10). Campesterol also exhibits strong but less extreme signals, whereas β-amyrin shows a more moderate deviation (Figure 7). This pattern is characteristic of a well-calibrated model capable of detecting a limited number of robustly associated loci. Overall, BLINK appears to provide good detection power while maintaining adequate control of inflation, thereby enhancing the credibility of the identified associations (Figure 7). Similar trends were observed in MLMM, with genomic inflation close to 1 ($\lambda = 0.97$), indicating good control of false-positive associations.

Similarly, under the FarmCPU model, the close alignment of points with the diagonal at low p-values indicates satisfactory overall calibration (Figure 7). Right-tail deviations are observed for all three traits; however, the maximum $-\log_{10}(p)$ values are generally lower than those obtained with BLINK for allantoin. Campesterol still displays several substantial signals, while β-amyrin remains the most moderately associated trait. The dispersion of points suggests a more conservative detection of strong associations (Figure 7). Overall, FarmCPU appears slightly more conservative, attenuating extreme signals ($\lambda = 1.0058$) while retaining the ability to identify biologically relevant associations.

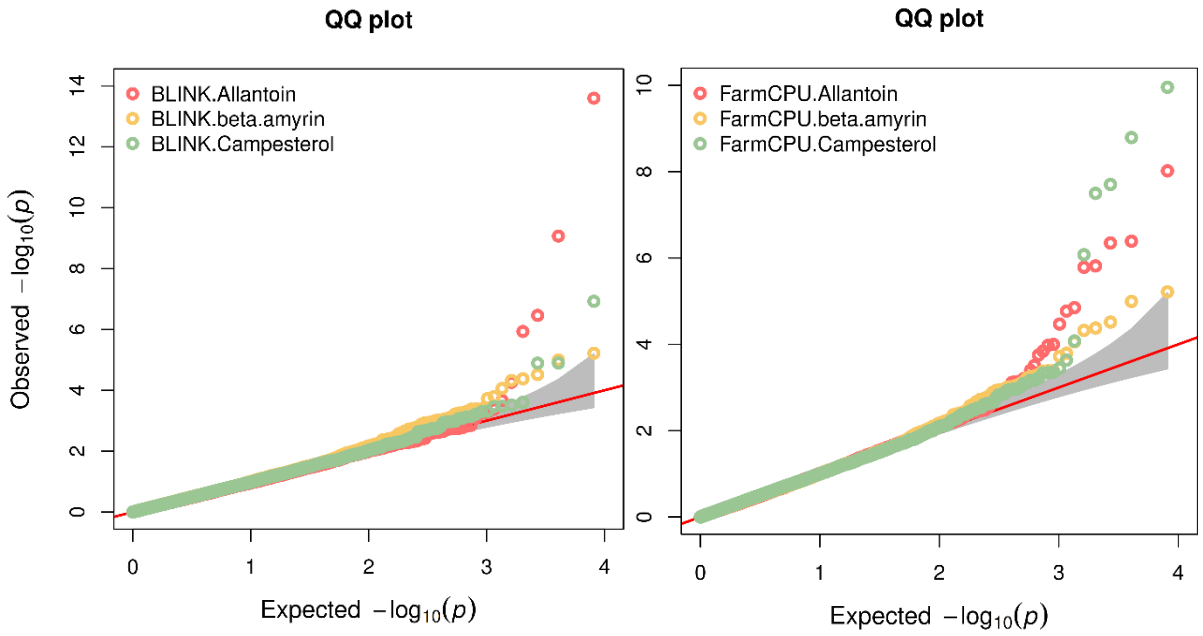


Figure 7: QQ plots of GWAS associations for allantoin, β-amyrin, and campesterol traits: BLINK model on the left and FarmCPU model on the right

Similarly, for the epigallocatechin, gallocatechin, and lupeol traits, QQ plots generated using the two GWAS models, FarmCPU and BLINK, show an overall excellent agreement between observed and expected p-values under the null hypothesis, indicating effective control of genomic inflation and adequate correction for population structure across all three traits (Figure 8). Under the FarmCPU model, deviations from the diagonal are observed only in the right tail of the distribution, suggesting the presence of true genetic associations, with the strongest signal detected for gallocatechin, followed by lupeol, while epigallocatechin exhibits more modest effects (Figure 8).

The BLINK model displays a similar pattern but with a more pronounced right-tail deviation, reflecting higher statistical power and the detection of a larger number of strongly associated SNPs, particularly for gallocatechin (Figure 8). Thus, although both models are well calibrated and yield consistent results, BLINK appears more sensitive in detecting strong associations (λ = 1.0139), whereas FarmCPU is more conservative and robust (λ = 0.9093), making it particularly suitable for validating major genetic signals (Figure 8).

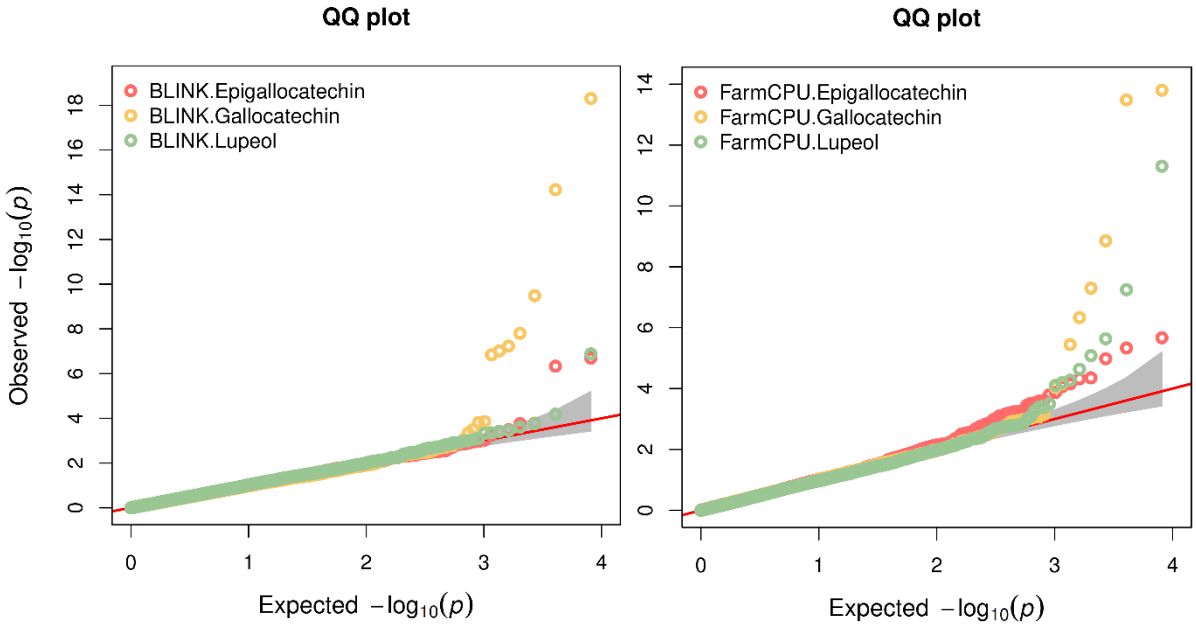


Figure 8: QQ plots of GWAS associations for epigallocatechin, gallocatechin, and lupeol traits: BLINK model on the left and FarmCPU model on the right

For trans-cinnamic acid content, both models are excellently calibrated (λ_{BLINK} = 1.0080 and $\lambda_{FarmCPU}$ = 1.0104; Supplementary Figure 2). The consistency of the signals observed in the Manhattan plots further supports the robustness of the detected associations.

The MLM and GLM models failed to identify significant genomic associations for most of the traits analyzed. An exception was trans-cinnamic acid, for which significant associations were detected on chromosomes 3 and 5 by both models (Supplementary Figures 3 and 4). However, variance component analysis from the MLM and GLM models indicates that nearly all phenotypic variance is attributable to the residual component, with a negligible contribution from SNP-explained genetic variance (Supplementary Figures 5 and 6). In this context, the associations observed for trans-cinnamic acid are likely to represent false positives. Furthermore, both models exhibit marked and comparable genomic deflation ($\lambda = 0.8771$), suggesting an absence of detectable genetic signal and limited statistical power.

4.6. Annotation of genes near the lead SNP associated with allantoin

Based on the SNP lead identified by FarmCPU model, 27 genes with putative regulatory functions were identified within the QTL interval on chromosome 08 (Supplementary Table 3). The closest gene to the lead SNP, Vitpa08g27650 (4.6 kb), encodes a pre-mRNA splicing factor (Prp8-like), suggesting a possible role in transcriptional regulation. Additional nearby candidates include protein kinases (Vitpa08g27660) and receptor-like proteins (Vitpa08g27630 / 27670 / 27680), supporting the hypothesis that this locus is involved in signal transduction and gene expression regulation. In MLMM and BLINK models, on chromosome 07, the strongest candidate gene associated with allantoin content is Vitpa07g01800, located only 112 bp from the lead SNP and encoding an ABC transporter (Supplementary Table 4). This gene, together with nearby regulators of vesicle trafficking and vacuolar protein sorting, suggests that natural variation in allantoin content in shea kernels is driven by differential transport and vacuolar accumulation rather than by biosynthetic capacity.

4.7. Annotation of genes near the lead SNP associated with β-amyrin

To investigate the genetic basis underlying the association with β-amyrin content, genes located in the genomic region surrounding the lead SNP on chromosome 08 were identified and annotated. A total of ten genes were detected within the region flanking the lead SNP, spanning approximately 19.09 Mb to 19.26 Mb on chromosome 08 (Supplementary Table 5). The distances between these genes and the lead SNP ranged from 312 bp to approximately 58 kb. The genes closest to the lead SNP were Vitpa08g15240, located at 312 bp, Vitpa08g15230, located at 2,088 bp, and Vitpa08g15220, located at 20,238 bp. The closest gene to the lead SNP, Vitpa08g15240 (312 bp from the SNP), displayed a rich functional annotation related to intracellular trafficking, including the Gene Ontology (GO) terms GO:0006886 (intracellular

transport), GO:0015031 (protein transport), GO:0016192 (vesicle-mediated transport), and GO:0030117 / GO:0030123 (AP-3 adaptor complex). This gene is predicted to encode protein domains associated with the delta subunit of the adaptor protein complex AP-3, armadillo-like domains, and adaptor proteins involved in vesicle sorting and vacuolar trafficking (Supplementary Table 5).

Several neighboring candidate genes (Vitpa08g15210, Vitpa08g15200, Vitpa08g15160, and Vitpa08g15170) encode proteins containing transmembrane domains, signal peptides, and glycosyltransferase domains (family 14). These genes are predominantly associated with the endomembrane system, including the plasma membrane, endoplasmic reticulum, and Golgi apparatus (Supplementary Table). A more distal gene, Vitpa08g15250, encodes a protein containing RING/CHY-type zinc finger domains, suggesting a potential role in transcriptional or post-translational regulation (Supplementary Table 5).

4.8. Annotation of genes near the lead SNP associated with campesterol

Within the ±100 kb interval surrounding the FarmCPU lead SNP on chromosome 12, ten protein-coding genes were identified, spanning approximately 34.43 Mb to 34.59 Mb. Gene Ontology (GO) annotations indicate that this region is predominantly enriched in genes involved in RNA binding and post-transcriptional regulation (Supplementary Table 6). The distances between these genes and the lead SNP ranged from 867 bp to ~73 kb. The closest gene to the lead SNP was Vitpa12g20510, located 867 bp away. Other nearby genes included Vitpa12g20520 (15.5 kb), Vitpa12g20500 (30.3 kb), and Vitpa12g20530 (47.3 kb). Functional annotation revealed that several genes in this region encode proteins associated with RNA binding and protein–protein interactions, as indicated by the recurrent Gene Ontology terms GO:0003723 (RNA binding) and GO:0005515 (protein binding). These GO terms were detected in three genes: Vitpa12g20510, Vitpa12g20520, and Vitpa12g20540.

Protein domain analysis showed a strong enrichment for domains related to pre-mRNA splicing and RNA metabolism, including MIF4G-like domains, armadillo-type folds, and CWC22 (pre-mRNA splicing factor) homologs. These domains were particularly prevalent in the closest genes to the lead SNP, notably Vitpa12g20510 and Vitpa12g20520. One more distal gene, Vitpa12g20500 (~30 kb from the SNP), encodes a protein containing a beta-L-arabinofuranosidase (GH127) domain and signal peptide motifs, suggesting a potential role in carbohydrate metabolism or cell wall modification. Another distant gene, Vitpa12g20480 (~73 kb), harbors an IPP transferase / dimethylallyltransferase domain, involved in isoprenoid-

related biochemical processes. Overall, no genes directly annotated with GO terms related to sterol biosynthesis, lipid metabolic processes, or isoprenoid pathways were detected within the FarmCPU-defined locus.

Analysis of the genomic region flanking the BLINK lead SNP on chromosome 8 revealed 18 annotated genes exhibiting a wide range of biological functions, as reflected by their GO annotations (Supplementary Table 7). Several genes in close proximity to the lead SNP are associated with protein binding and regulatory functions, including Vitpa08g04010, annotated with GO:0005515 (protein binding) and encoding a RING-type E3 ubiquitin ligase, suggesting a role in protein turnover and regulatory signaling. Transcriptional regulation is represented by Vitpa08g04020, annotated with GO:0003677 (DNA binding) and GO:0008289 (lipid binding), encoding a homeobox–leucine zipper transcription factor. Additionally, Vitpa08g03990 is annotated with GO:0003700 (DNA-binding transcription factor activity) and GO:0006355 (regulation of transcription). Signal transduction-related functions are suggested by Vitpa08g04000, annotated with GO:0004672 (protein kinase activity), GO:0005524 (ATP binding), and GO:0006468 (protein phosphorylation), encoding a leucine-rich repeat receptor-like kinase. Metabolic functions are represented by Vitpa08g03970, annotated with GO:0009058 (biosynthetic process), GO:0016746 (transferase activity), and GO:0016747 (acyltransferase activity), encoding a chalcone/stilbene synthase-like enzyme. Notably, Vitpa08g03940, located within ~49 kb of the BLINK lead SNP, is annotated with GO:0005525 (GTP binding) and encodes a small ARF-family GTPase, a protein class known to regulate vesicle-mediated transport and membrane dynamics (Supplementary Table 7).

4.9. Annotation of genes near the lead SNP associated with epigallocatechin

The FarmCPU and MLMM models identified a significant association for epigallocatechin on chromosome 5, with the lead SNP located at 42,200,948 bp. A ±100 kb window around this SNP (42.10–42.30 Mb) contained 12 annotated protein-coding genes (Supplementary Table 8). Functional annotation revealed that several genes in this region are involved in post-transcriptional regulation, protein turnover, transcriptional control, and detoxification processes, as supported by Gene Ontology (GO) annotations (Supplementary Table 8). The closest gene to the lead SNP is Vitpa05g19650, located approximately 1.6 kb away. This gene is annotated with GO:0008270 (zinc ion binding) and encodes a RING-type E3 ubiquitin ligase of the MARCH family, containing a RING-CH zinc finger domain and a transmembrane region. Such proteins are known to regulate protein stability through ubiquitination (Supplementary Table 8).

Other notable genes in the region include Vitpa05g19640, annotated with GO:0016787 (hydrolase activity) and encoding an apyrase-like protein, potentially involved in nucleotide metabolism and signaling. Vitpa05g19680 carries multiple GO terms related to transcription initiation (GO:0001164, GO:0001188, GO:0006360, GO:0070860) and encodes a TAF1B/Rrn7-like factor, a component of RNA polymerase I transcription machinery. Additionally, Vitpa05g19700 is associated with ribosomal structure and translation (GO:0003735, GO:0005840, GO:0006412), while Vitpa05g19710 encodes a glutathione S-transferase (GST) annotated with GO:0005515 (protein binding) and GO:0006749 (glutathione metabolic process). No gene encoding a core enzyme of the flavonoid biosynthetic pathway was identified within the FarmCPU locus (Supplementary Table 8).

At the same chromosomal position (42,200,948 bp on chromosome 5), overlapping the FarmCPU and MLMM signals, the ±100 kb interval identified by BLINK contained the same set of 12 genes, confirming the robustness of this locus across GWAS models (Supplementary Table 9). Importantly, BLINK refined the signal strength and reinforced the proximity ranking of candidate genes, again highlighting Vitpa05g19650 as the closest gene to the lead SNP. The GO annotations and protein domain composition of genes in this region were consistent with those identified under FarmCPU, emphasizing enrichment in protein ubiquitination, transcriptional regulation, and detoxification-related processes (Supplementary Table 9).

4.10. Annotation of genes near the lead SNP associated with the gallocatechin

A ±100 kb window around the FarmCPU lead SNP on chromosome 5 revealed 21 annotated genes. Among them, several genes presented functional annotations, protein domains, or Gene Ontology (GO) terms relevant to metabolic regulation (Supplementary Table 10). The closest gene to the lead SNP was Vitpa05g31190 (distance: 928 bp), encoding a ubiquitin-conjugating enzyme E2, supported by protein domains related to ubiquitin-mediated protein modification. Nearby genes included Vitpa05g31210, annotated as a DNA-directed RNA polymerase subunit (GO:0003677, GO:0003899, GO:0006351), and Vitpa05g31180, encoding a ribosomal protein L21-like (GO:0003723, GO:0003735, GO:0005840, GO:0006412) (Supplementary Table 10). Of particular interest, Vitpa05g31220, located ~31 kb from the lead SNP, encodes an allantoinase-like protein with GO terms associated with metal-dependent hydrolase activity and nitrogen compound metabolism (GO:0004038, GO:0016787, GO:0016810). In addition, Vitpa05g31170 and Vitpa05g31150 encode lipoxygenase-related proteins (GO:0016491, GO:0016702, GO:0055114), suggesting a potential involvement in redox-related metabolic processes (Supplementary Table 10).

The ±100 kb region surrounding the MLMM and BLINK lead SNP on chromosome 12 contained 10 annotated genes, several of which showed coherent functional annotations (Supplementary Table 11). The closest gene to the lead SNP was Vitpa12g20510 (distance: 867 bp), encoding a protein homologous to the pre-mRNA splicing factor CWC22, characterized by RNA-binding activity (GO:0003723) and protein binding (GO:0005515), and containing MIF4G and armadillo-type fold domains. Closely related paralogs (Vitpa12g20520 and Vitpa12g20540) with similar GO terms and domains were also present in the region, suggesting a local cluster of RNA metabolism–related genes. Another notable gene, Vitpa12g20500, encodes a glycoside hydrolase (GH127) with GO:0005975 (carbohydrate metabolic process), while Vitpa12g20480 is annotated as a tRNA dimethylallyltransferase (GO:0008033), a key enzyme in tRNA modification and translational regulation (Supplementary Table 11).

4.11. Annotation of genes near the lead SNP associated with lupeol

A ±100 kb window surrounding the shared lead SNP on chromosome 12 (2.50–2.70 Mb) was explored to identify candidate genes. This interval contained 19 annotated genes (Vitpa12g02570 to Vitpa12g02750), several of which displayed informative GO annotations and conserved protein domains (Supplementary Table 12). The functional landscape of this region was dominated by the following categories: i) Transcriptional regulation and DNA binding**:** Multiple genes were annotated with GO terms related to transcriptional control, including GO:0006355 (regulation of transcription, DNA-templated), GO:0003690 (double-stranded DNA binding), and GO:0001228 (DNA-binding transcription factor activity). These genes encode CCCH-, C2H2-, and CW-type zinc finger proteins, as well as an NF-Y transcription factor subunit (Vitpa12g02590), suggesting a regulatory role at the transcriptional level; ii) Membrane-associated transport and ion homeostasis: Vitpa12g02640 was annotated with GO:0006812 (cation transport), GO:0008324 (cation transmembrane transporter activity), GO:0016021 (membrane), and GO:0055085 (transmembrane transport), and harbors conserved cation efflux transporter domains (Supplementary Table 12). This gene represents the most explicit link to membrane transport within the QTL; iii) Carbohydrate metabolic processes: Vitpa12g02700 was associated with GO:0005975 (carbohydrate metabolic process) and glycosidase-related domains, indicating potential involvement in primary metabolism rather than direct triterpenoid biosynthesis; and iv) Translation-related functions : Vitpa12g02690 encodes a peptide chain release factor, with GO:0003747 (translation release factor activity) and GO:0006415 (translational termination), suggesting housekeeping functions.

No gene within this interval was annotated as a known triterpene biosynthetic enzyme (e.g. oxidosqualene cyclases or cytochrome P450s) (Supplementary Table 12).

4.12. Annotation of genes near the lead SNP associated with trans-cinnamic acid

Based on the shared lead SNP identified by both models, a ±100 kb window around position 56.43 Mb on chromosome 5 (56.33–56.53 Mb) was defined as the primary quantitative trait locus (QTL). This region contained 22 annotated genes, spanning from Vitpa05g30420 to Vitpa05g30630 (Supplementary Table 13).

Distances between individual genes and the lead SNP ranged from 23 bp to approximately 98 kb, allowing fine-scale prioritization of candidate genes based on physical proximity and functional annotation. Functional annotation of genes located within the QTL region revealed a diverse repertoire of biological processes and molecular functions. Several genes were associated with oxidation–reduction processes and stress responses, including Vitpa05g30510, which encodes a class III plant peroxidase (Supplementary Table 13). This gene is annotated with GO terms such as GO:0004601 (peroxidase activity), GO:0006979 (response to oxidative stress), and GO:0055114 (oxidation–reduction process), and contains conserved peroxidase domains.

Two closely located genes (Vitpa05g30530 and Vitpa05g30540) encode GDSL lipase/esterase family proteins, characterized by SGNH hydrolase domains and annotated with GO:0016788 (hydrolase activity), suggesting roles in lipid or ester metabolism (Supplementary Table 13). The QTL also harbors multiple transcription factor genes, including members of the AP2/ERF, TCP, LOB, and homeobox (WOX) families. These genes are associated with GO terms such as GO:0003700 (DNA-binding transcription factor activity) and GO:0006355 (regulation of transcription), indicating potential regulatory roles.

Additional genes encode proteins involved in signal transduction and membrane-associated processes, including Rho/Rac small GTP-binding proteins (annotated with GO:0007264, small GTPase-mediated signal transduction) and EF-hand calcium-binding proteins linked to GO:0055085 (transmembrane transport). Importantly, no gene encoding a canonical enzyme of the core phenylpropanoid pathway (e.g. phenylalanine ammonia-lyase, cinnamate 4-hydroxylase, or 4-coumarate-CoA ligase) was directly identified within this QTL interval (Supplementary Table 13).

4.13. Annotation of genes near the lead SNP associated with stigmasterol

The lead SNP for Stigmasterol was identified at position 18,705,710 on chromosome 7 (chr07) appliying the MLMM model. Using a ±250 kb window around this SNP, the closest gene captured is Vitpa07g11490, which lies 243 kb away from the SNP, indicating that the variant is located in a potentially regulatory or intergenic region. Functional annotation shows an unnamed protein containing signal peptide and extracellular domains, suggesting it may be secreted or membrane-associated. Although no Gene Ontology (GO) labels were available, this gene represents the best candidate underlying the observed allelic effect on Stigmasterol. The absence of genes within the narrower ±100 kb window highlights the importance of selecting a window size that reflects the species' linkage disequilibrium. Further functional validation is required to confirm the gene's role.

4.14. Allelic effect analysis

The effects of GWAS lead SNPs on metabolite concentrations were assessed by comparing allelic dosage classes (0, 1, and 2 copies of the alternative allele) using MLMM, FarmCPU, and BLINK models. For allantoin, metabolite levels increased markedly with allelic dosage, indicating a predominantly additive genetic effect, particularly well captured by the MLMM and BLINK models (Supplementary Table 14), consistent with the progressive shift of distributions observed in the violin plots (Supplementary Figure 7). Similarly, β-amyrin displayed a strong differentiation between genotypic classes, with a clear separation of distributions and a high additive effect, with no evidence of dominance (Supplementary Table 14). Campesterol showed a gradual increase in concentration across allelic dosages, associated with robust additive effects that were consistent across GWAS models, despite increased variability for rare homozygous genotypes (Supplementary Table 14, Supplementary Figure 7). In contrast, gallocatechin and epigallocatechin exhibited non-linear responses across genotypes, with heterozygotes deviating from intermediate values, suggesting partial dominance effects (Supplementary Table 14). Trans-cinnamic acid was characterized by extremely high concentrations in certain homozygous individuals, resulting in highly asymmetric distributions and very large additive effects. Finally, stigmasterol showed a consistent decrease in concentration with increasing allelic dosage, indicative of a stable negative additive effect across genotypes (Supplementary Table 14). Linear regression analyses confirmed the additive genetic effects suggested by allelic dosage plots across multiple metabolites (Supplementary Table 15). Consistent with the progressive shifts observed in the violin plot distributions

(Supplementary Figure 7), strong positive additive effects were detected for gallocatechin (13.85 ± 1.82 µmol·L^{-1} per allele), epigallocatechin (7.66 ± 1.92 µmol·L^{-1}), lupeol (63.05 ± 9.14 µmol·L^{-1}), β-amyrin (35.72 ± 7.14 µmol·L^{-1}), allantoin (4.72–6.64 µmol·L^{-1} per allele), and campesterol (2986.93–6417.77 µmol·L^{-1} per allele), whereas a significant negative additive effect was observed for stigmasterol (–256.16 ± 50.90 µmol·L^{-1}) (Supplementary Table 15). Trans-cinnamic acid showed a particularly large additive effect (1123.71 ± 50.22 µmol·L^{-1} per allele), reflected by a marked separation between allelic classes in the violin plots (Supplementary Table 15). Significant dominance effects for gallocatechin, trans-cinnamic acid, lupeol, and allantoin explained deviations from strictly intermediate heterozygous distributions, while epigallocatechin, β-amyrin, campesterol, and stigmasterol displayed violin plot patterns consistent with predominantly additive genetic control (Supplementary Figure 7).

5. Discussion

5.1. Genome-wide SNP distribution and genetic diversity in shea tree populations

The genome-wide SNP distribution in *V. paradoxa* revealed a highly heterogeneous pattern, with substantial variation in SNP density among chromosomes and along individual chromosomes. While most 1-Mb windows exhibited low to moderate SNP densities, several regions particularly on chromosomes 2, 3, 6, 7, and 8, showed clusters of high polymorphism, likely reflecting elevated recombination and increased marker coverage. Conversely, regions with very low SNP density may represent conserved genomic segments, reduced variability, or technical limitations in sequencing and alignment. The genome-wide distribution of SNP density obtained in this study is highly consistent with the pattern previously reported by Attikora et al. (2025) for *V. paradoxa*. Both analyses reveal a pronounced heterogeneity in SNP density along and among chromosomes, indicating a non-random organization of genetic variation across the genome. Despite this heterogeneity, overall genetic diversity was high, with observed heterozygosity (Ho = 0.27) closely matching expected values (He = 0.28) and a mean polymorphism information content (PIC = 0.36), as reported by Attikora et al. (2024). These observations indicate that the SNP dataset provides sufficient coverage and allelic diversity to support robust GWAS analyses, while highlighting the need for cautious interpretation of signals in marker-sparse regions.

5.2. Weak population structure and implications for GWAS

Principal coordinates analysis revealed that the first two axes explained only 5.77% of the total genetic variation, indicating weak population structure. Similar results were reported by

Attikora et al. (2024), who observed low genetic differentiation (1.96% of the total variation) within a core collection of 100 *Vitellaria* individuals, suggesting that this species exhibits limited genetic structure when sample size is relatively small. The genomic kinship matrix showed predominantly low to moderate relatedness, consistent with a largely panmictic population. This limited structure reduces the risk of confounding due to population stratification in GWAS and supports the use of multi-locus models such as FarmCPU and BLINK (Enyew et al., 2022; Segura et al., 2012). Consequently, significant associations detected are unlikely to be artifacts of structure or relatedness, increasing confidence in the identified candidate loci.

5.3. Complementarity and comparative performance of MLMM, FarmCPU and BLINK

The use of MLMM, FarmCPU, and BLINK models allowed us to leverage their complementary strengths for mapping metabolic traits. Both models were well-calibrated, with genomic inflation factors near 1 and QQ plots indicating effective control of population structure (Sul et al., 2018).

BLINK consistently detected stronger signals and more associated SNPs, particularly for polygenic traits such as allantoin and gallocatechin, reflecting its higher sensitivity to small-effect loci (Huang et al., 2019). FarmCPU, in contrast, provided a more conservative profile, attenuating extreme $-\log_{10}(p)$ values, which is advantageous for confirming major-effect loci like lupeol and β-amyrin (Liu et al., 2016). For instance, in the study by Sallam et al. (2026), FarmCPU demonstrated enhanced efficiency in reliably detecting associations between peduncle-related traits and spike productivity in wheat. The convergence of significant SNPs across both models, such as the lead SNP on chromosome 5 for trans-cinnamic acid, underscores the robustness of these associations. This complementarity ensures a balance between sensitivity (BLINK) and specificity (FarmCPU), providing a reliable framework for interpreting complex metabolite GWAS. Prateep-Na-Thalang et al. (2024) in a genome-wide association study in tomato demonstrated that the FarmCPU and BLINK models outperformed other GWAS approaches, effectively identifying SNPs associated with citric acid content as well as a range of other fruit quality traits.

5.4. Genetic architecture of metabolic traits

Metabolic traits in *V. paradoxa* display diverse genetic architectures. Polygenic traits, including allantoin and flavan-3-ols (epigallocatechin and gallocatechin), are influenced by multiple loci with moderate to small effects distributed across several chromosomes, reflecting the

complexity of flavonoid biosynthesis and nitrogen metabolism. These observations are consistent with previous reports in other plant species. For instance, in rice (*Oryza sativa* L.), GWAS highlighted that approximately one-third of metabolites are mainly regulated by mQTLs (metabolic Quantitative Trait Loci) with large effects, whereas many metabolites are controlled by multiple mQTLs weakly associated with metabolic phenotypes (Matsuda et al., 2015). Similarly, in *Arabidopsis thaliana* L., Naake et al. (2024) demonstrated that metabolic GWAS (mGWAS) in seeds identified numerous loci associated with specialized metabolites, including flavonoids, illustrating that these compounds are regulated by multiple genetic factors. In contrast, major-effect loci were detected for triterpenes such as lupeol and β-amyrin. For instance, strong SNPs on chromosomes 8 and 12 for β-amyrin, and on chromosome 12 for lupeol, likely correspond to key biosynthetic or regulatory elements controlling metabolite accumulation. Major-effect loci have been identified for certain specialized metabolites in other plants, where individual loci explain a large fraction of phenotypic variance. For example, GWAS in rapeseed (*Brassica napus* L.) uncovered multiple strong association intervals corresponding to known biosynthetic genes for glucosinolate metabolites (Tang et al., 2023), and rice mGWAS demonstrated that metabolites are mainly regulated by mQTLs with large effects (Matsuda et al., 2015). These differences have practical implications: polygenic traits may require genomic selection to capture cumulative effects, while major-locus traits are suitable targets for marker-assisted selection.

5.5. Functional insights from candidate gene annotation

Integration of candidate gene annotation across multiple metabolites reveals consistent biological patterns. This pattern of genetic control is consistent with findings from Chan et al. (2010), who showed in *A. thaliana* that metabolite variation is under polygenic control with numerous genotype–metabolite associations across the genome, indicating a complex genetic architecture for the metabolome. In general, natural variation in shea kernel metabolites appears to be largely governed by regulatory, transport, and compartmentalization processes rather than by structural variation in core biosynthetic enzymes. Soltis and Kliebenstein (2015) showed that natural variation in plant metabolites often involves regulatory mechanisms in addition to structural factors, suggesting an important role for genetic regulation and potentially the organization of genetic networks in shaping metabolic diversity.

For allantoin, two major QTLs were identified. On chromosome 7, the lead SNP is positioned just 112 bp from Vitpa07g01800, an ABC transporter involved in vacuolar transport. This locus is supported by neighboring genes associated with vesicle trafficking, such as Rho-like

GTPases, and vacuolar protein sorting (VPS35), suggesting that allantoin accumulation is strongly influenced by intracellular transport and storage. On chromosome 8, a QTL encompassing a Prp8-like splicing factor likely regulates purine catabolism, indicating that both transport and transcriptional regulation contribute to trait variation. Collectively, these findings highlight a dual-level control of allantoin content via transcriptional and post-biosynthetic processes. Indeed, Saurabh Singh et al. (2022) reported that specialized metabolites in plants are frequently modulated by complex regulatory networks involving transcription, co-expression, and multi-omics integration, indicating that metabolite variation depends largely on regulation and pathway integration rather than solely on biosynthetic enzymes. Moreover, Lunn (2007) demonstrated that plant metabolism is strongly influenced by intracellular compartmentalization and the transport of metabolites between organelles, which can alter metabolic profiles independently of variation in the biosynthetic enzymes themselves (Lunn, 2007).

Triterpenes, including β-amyrin and lupeol, are predominantly influenced by major-effect loci linked to genes controlling intracellular trafficking and membrane transport. The β-amyrin locus is tightly associated with Vitpa08g15240, encoding a component of the AP-3 adaptor complex, critical for vesicle trafficking and vacuolar targeting. Lupeol variation is similarly linked to a locus enriched in transcriptional regulators and a cation transporter (Vitpa12g02640), with minimal representation of canonical triterpene biosynthetic enzymes. These observations indicate that intracellular compartmentalization and transcriptional control, rather than enzymatic activity, are key drivers of triterpene accumulation, as demonstrated by Yao et al. (2023). Similar observations have been reported by Méteignier et al. (2023), who show that natural variation in specialized plant metabolites is largely governed by complex regulatory processes, including transcriptional regulation, chromatin dynamics, and other levels of gene regulation, rather than by variation in the biosynthetic enzymes themselves.

This complex regulatory architecture underlies the diversity of metabolic profiles observed in specialized pathways.

Variation in campesterol appears to be mediated by post-transcriptional and transport-related processes, as candidate genes include RNA-binding proteins and ARF-family GTPases, implicating vesicle-mediated transport as a determinant of sterol content. Similarly, flavan-3-ols, such as epigallocatechin and gallocatechin, are controlled by polygenic loci enriched in ubiquitin-mediated protein turnover, RNA-binding proteins, and transcriptional regulators, pointing to regulation at both the post-transcriptional and post-translational levels.

For trans-cinnamic acid, lead SNPs are located near genes encoding peroxidases, transcription factors, and lipid-modifying enzymes, with no core phenylpropanoid biosynthetic genes identified. This suggests that variation in trans-cinnamic acid content is driven primarily by redox processes, transcriptional regulation, and intracellular compartmentalization rather than direct enzymatic control (Lunn, 2007; Saurabh Singh et al., 2022; Yao et al., 2023).

Across all traits, the convergence of FarmCPU and BLINK signals highlights robust candidate loci, particularly those involved in transport, compartmentalization, and regulatory functions. While major-effect traits such as β-amyrin and lupeol are primarily shaped by intracellular trafficking, polygenic traits—including flavan-3-ols, allantoin, and trans-cinnamic acid—are controlled by distributed regulatory networks. These findings collectively underscore the central role of cellular regulation and compartmentalization in shaping the metabolite landscape of shea kernels.

5.6. Allelic dosage effects of GWAS lead SNPs on metabolite accumulation

The strong agreement between quantitative estimates derived from the linear regression table and the distributional patterns observed in the violin plots highlights the robustness of the detected GWAS signals. Metabolites such as allantoin, β-amyrin, and campesterol predominantly exhibited additive genetic effects, suggesting relatively simple and cumulative genetic control, potentially reflecting direct modulation of key metabolic pathways. A similar pattern was reported by Elsafy et al. (2026) in their study on candidate genomic regions underlying capsule shattering in sesame, where allelic effect analyses validated the functional relevance of key allelic variants. In contrast, the non-linear allelic responses observed for gallocatechin, epigallocatechin, and trans-cinnamic acid point to more complex genetic architectures, possibly involving partial dominance, threshold effects, or regulatory constraints within biosynthetic pathways. Genetic studies have shown that non-additive effects, such as partial dominance and epistasis (gene–gene interactions), substantially contribute to phenotypic variation and can generate non-linear responses to allelic variation. This is consistent with studies in *Arabidopsis* by Rodriguez Cubillos et al. (2018), who demonstrated that compounds derived from secondary metabolism exhibited higher variability and more pronounced non-additive inheritance patterns, which were largely attributable to epistatic interactions. The presence of extreme values and highly skewed distributions, particularly for rare homozygous genotypes, underscores the influence of sample size imbalance on effect estimation and supports the combined use of descriptive statistics and violin plot visualizations for robust interpretation. Overall, the consistency of effects across multiple GWAS models strengthens

the biological relevance of these lead SNPs and supports their prioritization as candidates for functional validation of the genetic mechanisms underlying metabolite variation.

5.7. Biological relevance and implications for shea tree improvement

The insights gained from our GWAS and functional annotation analyses have important implications for both breeding and conservation of *V. paradoxa*. The identification of major-effect loci, particularly those controlling triterpenes such as lupeol and β-amyrin, provides clear targets for marker-assisted selection, allowing breeders to directly manipulate metabolite content with predictable outcomes. In contrast, polygenic traits, including flavan-3-ols and allantoin, are influenced by multiple small-effect loci, suggesting that genomic selection approaches capturing the cumulative effects of these loci may be more effective for improving these traits (Bhat et al., 2016; Kumar et al., 2024).

Beyond direct selection, our results emphasize the central role of regulatory, transport, and compartmentalization mechanisms in shaping metabolite accumulation. Genes involved in vesicle trafficking, vacuolar transport, transcriptional regulation, and post-translational modification are prime targets for manipulating metabolite levels without disrupting core biosynthetic pathways, thereby minimizing potential pleiotropic effects on growth, development, or other metabolic processes (Friso and van Wijk, 2015; Li et al., 2025; Nogia and Pati, 2021).

Furthermore, understanding the genetic basis of key secondary metabolites provides guidance for the conservation of genetic resources (Liveri et al., 2024; Salgotra et al., 2023). Maintaining allelic diversity at loci controlling essential metabolites ensures the long-term sustainability and adaptive potential of shea tree populations, supporting both ecological resilience and the continued availability of high-value compounds (Katz et al., 2021; Salgotra et al., 2023). Collectively, these findings bridge the gap between genomics and practical applications, offering actionable insights to improve the nutritional, medicinal, and industrial value of shea kernels while preserving the genetic integrity of natural populations.

Beyond association-based inferences, the functional validation of candidate genes identified in this study will be a critical step toward translating genomic insights into practical applications. The development and mastery of in vitro culture systems combined with targeted genome editing approaches, such as CRISPR/Cas-based mutagenesis, would enable direct testing of gene function and causal relationships between loci and metabolite accumulation. Such systems provide a controlled framework to assess the impact of gene knock-out or modulation on specific secondary metabolites, while minimizing environmental and developmental variability.

In long-lived tree species like *V. paradoxa*, these approaches offer a powerful strategy to accelerate functional genomics, reduce the time required for phenotypic evaluation, and refine breeding targets before field deployment, thereby strengthening the link between GWAS discoveries and applied improvement programs.

5.8. Limitations related to sample size and study design

While the results provide robust insights, the limited sample size (n = 52) imposes constraints on GWAS power. Small-effect loci may remain undetected, particularly for polygenic traits, potentially underestimating the true genetic complexity. Korte and Farlow (2013) review the advantages and limitations of GWAS for plant trait analysis, highlighting that sample size, effect sizes, and genetic heterogeneity are critical determinants of the power to detect associations, and that small sample sizes can restrict the detection of loci with small effects, potentially underestimating the true genetic complexity of polygenic traits.

However, the detection of strong, convergent signals across both FarmCPU and BLINK suggests that the identified loci represent the most robust and biologically meaningful contributors to metabolite variation. Future studies with larger panels could uncover additional minor-effect loci, refine effect size estimates, and improve the resolution of polygenic architectures. Despite this limitation, the present study offers a valuable foundation for functional validation and breeding applications.

Conclusion and perspective

This study provides the first comprehensive genome-wide association analysis of cosmetic-related bioactive metabolites in *V. paradoxa*. Using high-density SNP data and multilocus GWAS models (MLMM, FarmCPU, and BLINK), we identified both major-effect and polygenic loci controlling the concentration of key metabolites, including triterpenes, flavan-3-ols, allantoin, and trans-cinnamic acid. Major-effect loci, particularly those controlling lupeol and β-amyrin, are promising candidates for marker-assisted selection, while polygenic traits such as flavan-3-ols and allantoin are better suited for genomic selection approaches that capture cumulative small-effect loci.

Functional annotation of candidate genes revealed that natural variation in shea kernel metabolites is largely governed by regulatory processes, intracellular transport, and compartmentalization, rather than by structural variation in core biosynthetic enzymes. These findings highlight the central role of transcriptional, post-transcriptional, and post-translational regulation in shaping metabolite accumulation, consistent with reports from other plant species. Genes involved in vesicle trafficking, vacuolar transport, and transcriptional regulation emerge as key targets for metabolite manipulation with minimal pleiotropic effects.

From a practical perspective, this study bridges fundamental genomics with applied breeding. The identification of robust candidate loci and underlying genes provides actionable insights for shea improvement programs, supporting both the enhancement of high-value cosmetic compounds and the preservation of allelic diversity within natural populations. Furthermore, the genetic information generated can guide conservation strategies by maintaining diversity at loci controlling key metabolites, thereby promoting long-term adaptive potential and ecological resilience of shea tree populations.

Finally, while the current study is limited by sample size, which may have restricted the detection of small-effect loci, the convergence of signals across MLMM, FarmCPU, and BLINK models ensures that the identified loci are robust and biologically meaningful. Future studies with expanded germplasm panels and integrated multi-omics approaches will refine our understanding of the genetic architecture of specialized metabolites in shea, paving the way for more precise breeding and conservation interventions.

In summary, this work provides a critical foundation for the genomics-assisted valorization of shea, elucidates the complex regulatory landscape governing its bioactive metabolites, and sets the stage for targeted improvement and sustainable management of this economically and culturally important tree species.

Acknowledgements

We express our sincere gratitude to the GBioS Unit of the Faculty of Agricultural Sciences at the University of Abomey-Calavi, Benin, for organizing and facilitating supervision and data collection activities. Furthermore, we are deeply grateful to the team of the Agronomy and Environment Laboratory (LAE) at the University of Lorraine for their valuable technical and financial support.

Author Contributions

GJB: Conceptualization, Investigation, Methodology, Data curation, Software, Formal analysis, Writing – original draft, Writing – review & editing. AO: Investigation, Methodology, Supervision, Validation, Writing – review & editing. FGH: Methodology, Supervision, Validation, Writing – review & editing. AH: Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. EGAD: Funding acquisition, Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Funding

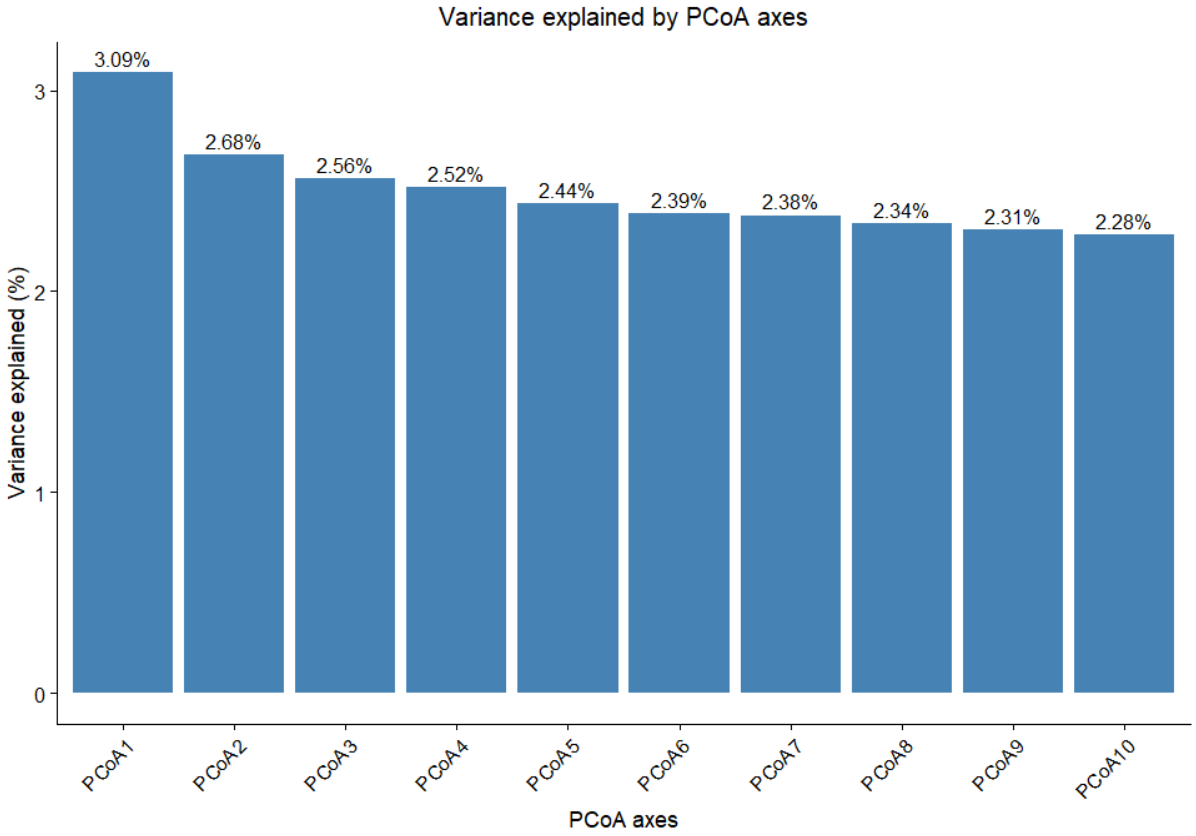
This work was fully supported by the PEA-BIOVALOR project: *Renforcement de la démarche compétences et de la culture entrepreneuriale dans la formation supérieure agronomique : leviers pour l'insertion professionnelle et le développement de la bioéconomie au Bénin*, financed by the Agence Française de Développement (AFD) through the Agence Nationale de Recherche (ANR)

Data Availability

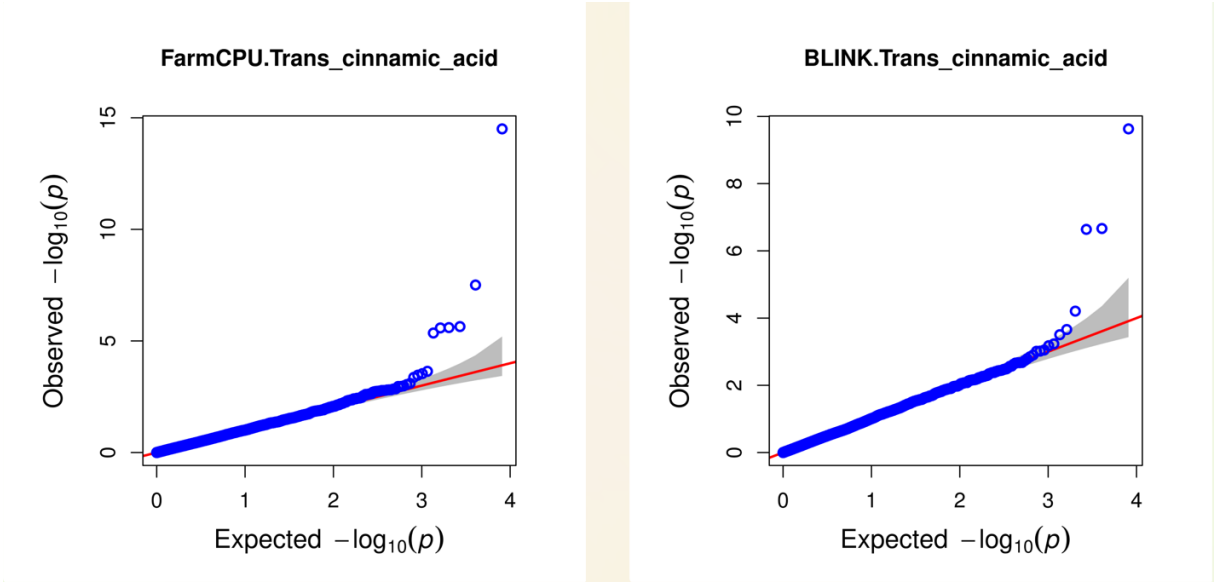
The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interests

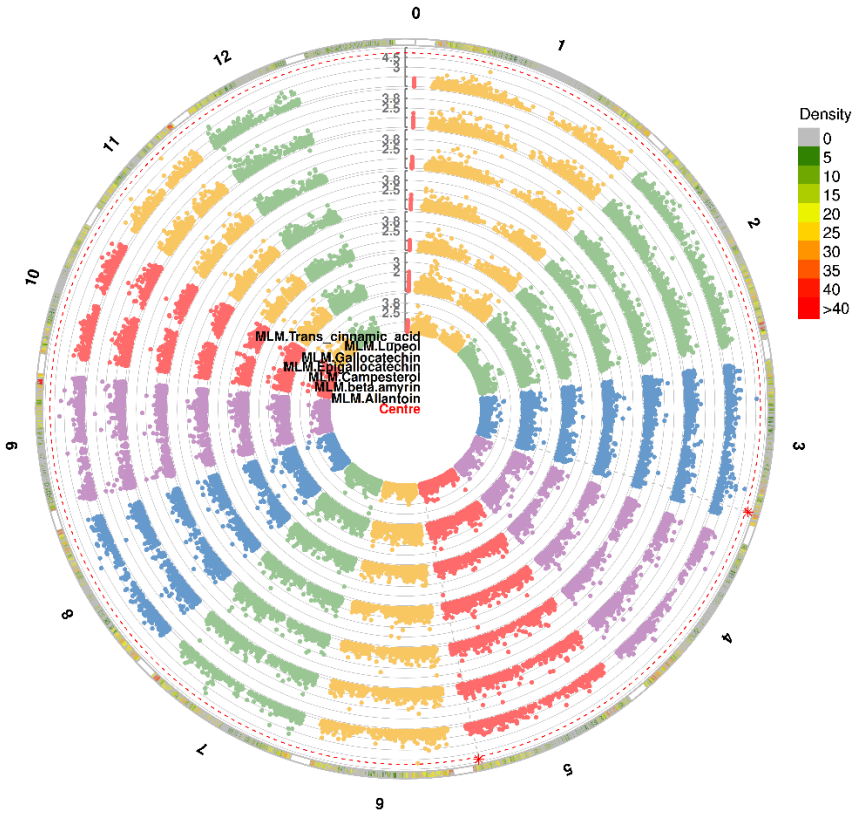
The authors have no relevant financial or non-financial interests to disclose



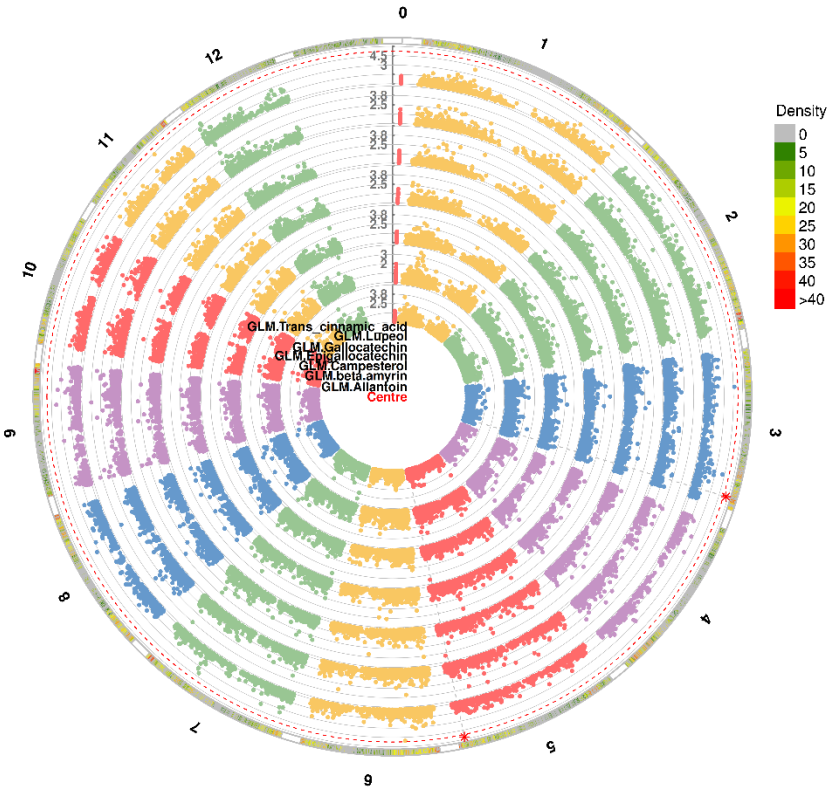
Suppl. Figure 1: Variance explained by PCoA controlling the population structure



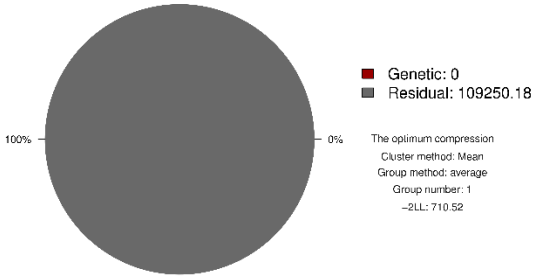
Suppl. Figure 2: Trans acid cinnamic QQ plots obtained from FarmCPU and BLINK models



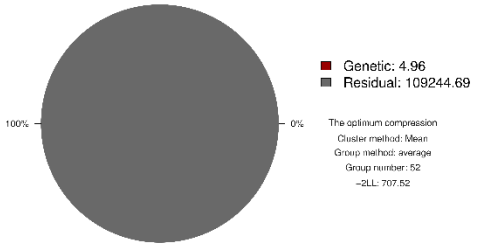
Suppl. Figure 3: Circular Manhattan plot illustrating multi-trait genome-wide association signals for all the traits involved in the study in shea tree (*Vitellaria paradoxa*) population, identified using the MLM model across the 12 chromosomes. Asterisks indicate SNPs exceeding the genome-wide Bonferroni-corrected significance threshold.



Suppl. Figure 4: Circular Manhattan plot illustrating multi-trait genome-wide association signals for all the traits involved in the study in shea tree (*Vitellaria paradoxa*) population, identified using the GLM model across the 12 chromosomes. Asterisks indicate SNPs exceeding the genome-wide Bonferroni-corrected significance threshold.

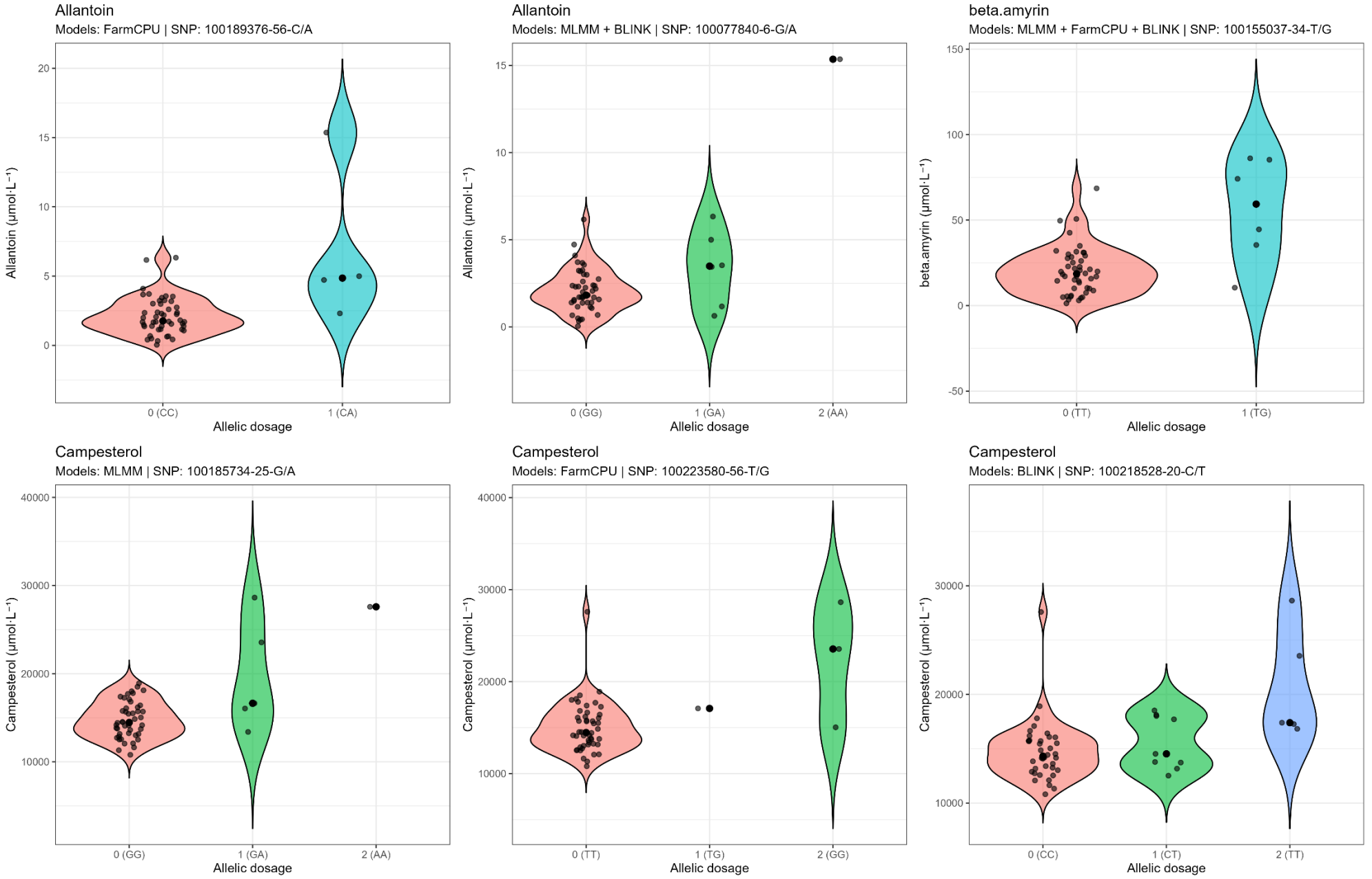


Suppl. Figure 5: Decomposition of the phenotypic variance in the GLM model

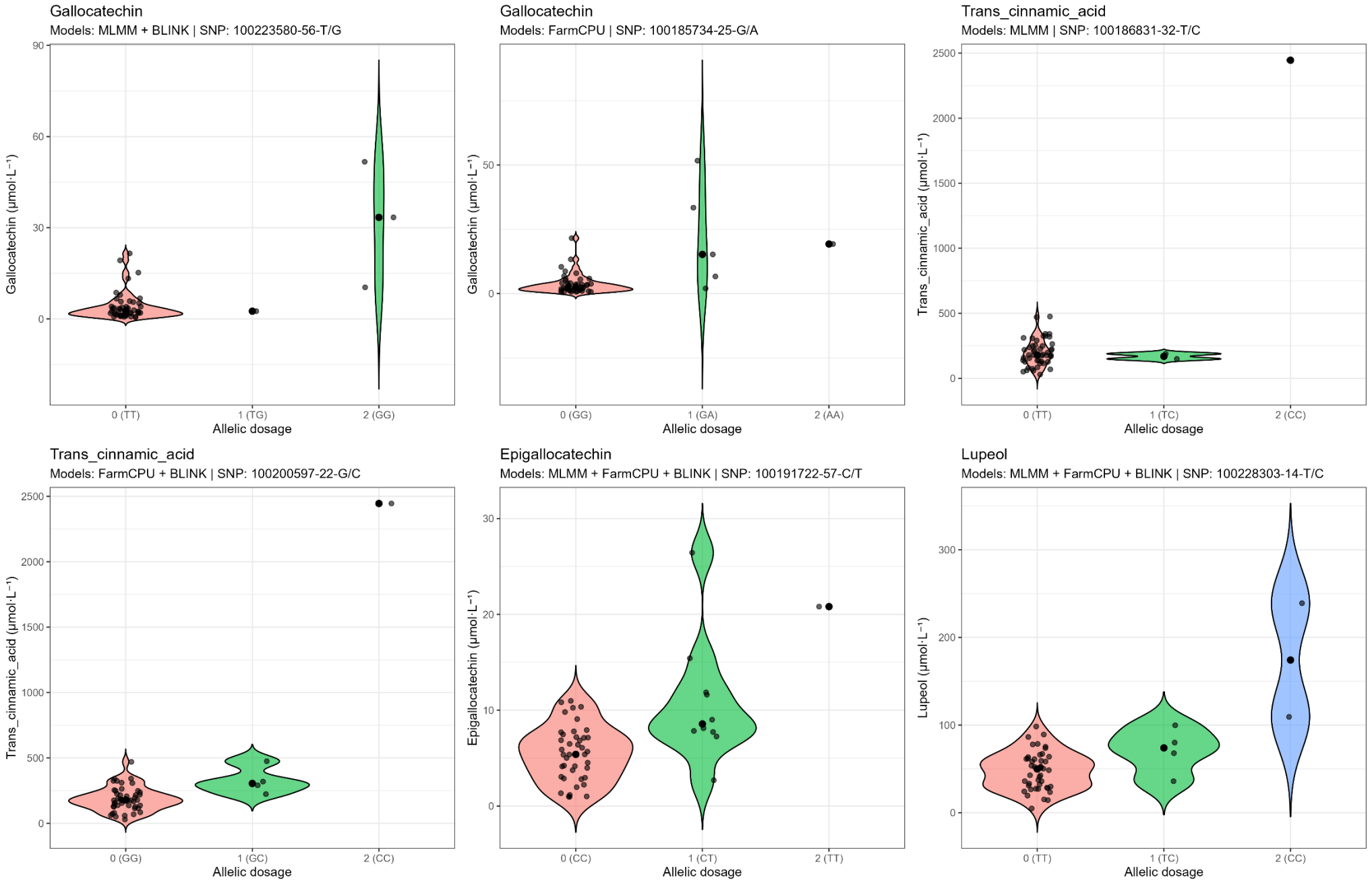


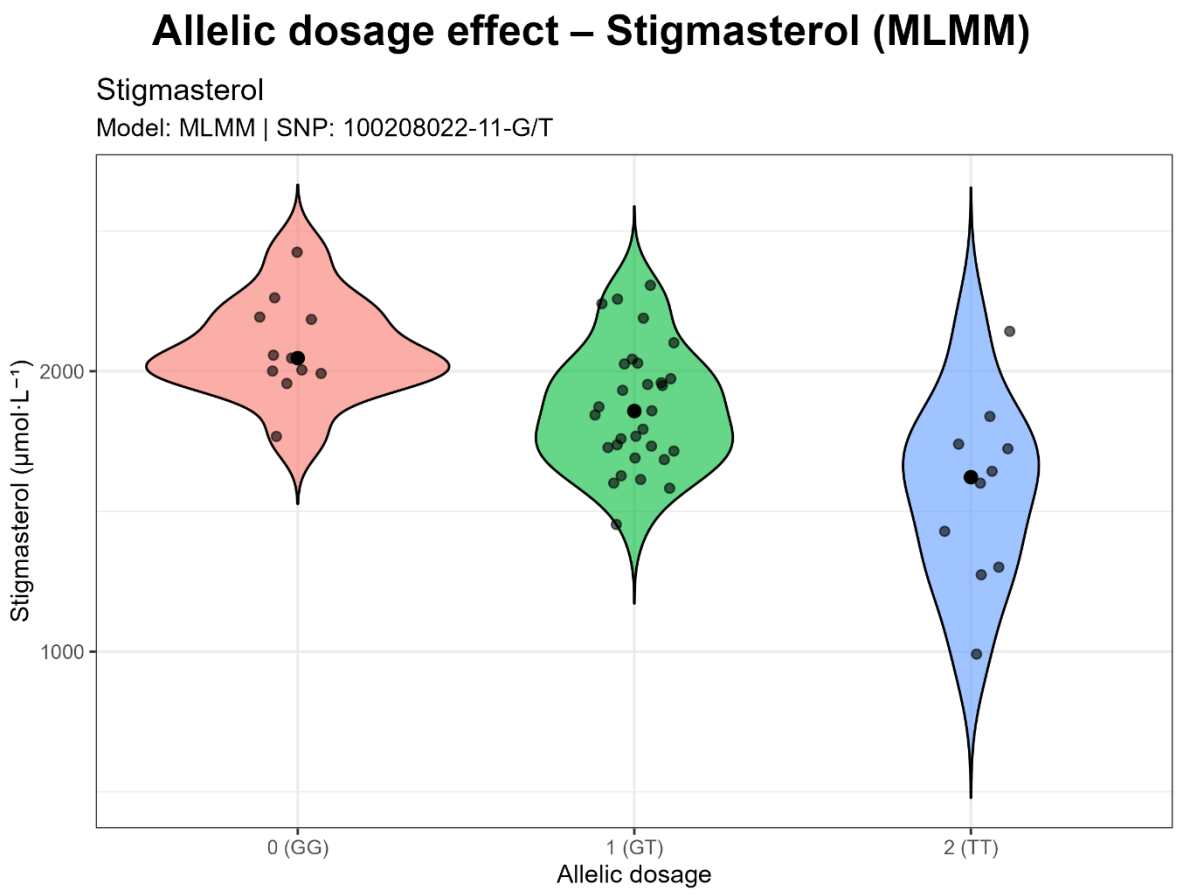
Suppl. Figure 6: Decomposition of the phenotypic variance in the GLM mode

Allelic dosage effects grouped by equivalent GWAS signals



Allelic dosage effects grouped by equivalent GWAS signals





Suppl. Figure 7: Allelic dosage effect grouped by equivalent GWAS signals for all eight significant traits.

Suppl. Table 1: Geographic origin and sampling coordinates of *Vitellaria paradoxa* individuals used in this study

id	origin	latitude	longitude	altitude
BG155	Boukoumbe	10.2948	1.092421	244.1903
BG136	Boukoumbe	10.32677	1.098174	243.7755
BG012	Boukoumbe	10.31088	1.102817	246.0516
BG112	Tchaourou	9.114695	2.317115	308.6611
BG149	Tchaourou	9.114685	2.316492	311.4174
BG242	Kandi	10.95565	3.171205	304.1
BG228	Kandi	10.95997	3.168618	291.8
BG192	Kandi	10.95974	3.169266	292.9
BG249	Kandi	10.9565	3.170928	303.2
BG265	Ndali	9.497592	2.59863	385.5336
BG267	Bembereke	10.21683	2.843076	405.9509
BG271	Bembereke	10.21517	2.840749	392.1788
BG278	Bembereke	10.21579	2.839675	371.8627
BG263	Bembereke	10.20431	2.857546	406.0131
BG152	Segbana	10.97536	3.283676	297.7
BG172	Segbana	10.97648	3.283331	296
BG182	Segbana	10.97978	3.279331	303.3
BG038	Segbana	10.97862	3.278676	298
BG010	Segbana	10.69417	2.775456	338.9
BG033	Segbana	10.95878	3.386852	322.1
BG324	Segbana	10.96032	3.38702	317.7
BG075	Segbana	10.95722	3.388683	332.8
BG246	Gogounou	9.499083	2.597901	378.1566
BG214	Gogounou	10.69666	2.774368	338.1
BG229	Gogounou	10.86686	2.851279	331.3
BG233	Gogounou	10.86579	2.850968	332.1
BG096	Kandi	11.07532	2.895293	302.1
BG114	Kandi	11.08124	2.88826	331.9
BG196	Kandi	11.07414	2.889673	318.5
BG106	Kandi	11.07649	2.895978	328.9
BG086	Pehunco	10.23565	2.101521	356.6079
BG047	Pehunco	10.23232	2.101562	363.5438
BG037	Pehunco	10.23573	2.102597	356.8981
BG050	Pehunco	10.23398	2.104294	359.3112
BG078	Pehunco	10.23269	2.102478	365.0767

BG181	Bembereke	10.1984	2.619263	393.424
BG179	Bembereke	10.21291	2.6063	416.9127
BG156	Pehunco	10.27909	1.975935	419.7538
BG109	Pehunco	10.28035	1.973886	430.2668
BG251	Gogounou	10.69785	2.775186	313.8
BG236	Gogounou	10.69701	2.775367	338.3
BG238	Gogounou	10.69666	2.774368	338.1
BG187	Gogounou	10.6939	2.771176	350.7
BG258	Gogounou	10.69763	2.772474	344.6
BG107	Tchaourou	9.190614	2.485486	354.1948
BG204	Parakou	9.401699	2.419825	336.3551
BG101	Tchaourou	9.191434	2.485849	354.528
BG105	Tchaourou	9.191214	2.485441	356.5342
BG097	Tchaourou	9.190871	2.485622	349.5751
BG103	Tchaourou	9.191302	2.486288	351.5874
BG189	Parakou	9.40243	2.419758	334.4642
BG218	Parakou	9.401714	2.419108	339.8552

Suppl. Table 2: Kinship matrix of the 52 shea tree samples (available on request)

Suppl. Table 3: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU lead SNP associated with allantoin content on chromosome 8 in *Vitellaria paradoxa*.

Name	Distance_SNP	Description	GO-labels	protein domains
Vitpa08g27650	4593	unnamed product	GO:0000398;GO:0003723;GO:0005515;GO:0005681;GO:0008237;GO:0017070;GO:0030623;GO:0070122	G3DSA:3.40.140.10: Cytidine Deaminase; IPR042516: Pre-mRNA-processing-splicing factor 8, U5-snRNA-binding domain superfamily; IPR012337: Ribonuclease H-like superfamily; IPR019581: Pre-mRNA-processing-splicing factor 8, U5-snRNA-binding; IPR012592: PROCN domain; IPR043172: Prp8 RNase domain IV, palm region; IPR037518: MPN domain; cd08056: MPN_PRP8; IPR012984: PROCT domain; mobidb-lite: consensus disorder prediction; IPR043173: Prp8 RNase domain IV, fingers region; IPR000555: JAB1/MPN/MOV34 metalloenzyme domain; G3DSA:3.90.1570.40: -; IPR019582: RNA recognition motif, spliceosomal PrP8; IPR012591: PRO8NT domain; IPR021983: PRP8 domain IV core; PTHR11140:SF2: PRE-MRNA-PROCESSING-SPLICING FACTOR 8A-RELATED; IPR027652: Pre-mRNA-processing-splicing factor 8; Coil: Coil; IPR019580: Pre-mRNA-processing-splicing factor 8, U6-snRNA-binding
Vitpa08g27660	9416	unnamed product	GO:0004672;GO:0005524;GO:0006468	IPR008271: Serine/threonine-protein kinase, active site; IPR011009: Protein kinase-like domain superfamily; G3DSA:3.30.200.20: Phosphorylase Kinase; domain 1; G3DSA:1.10.510.10: Transferase(Phosphotransferase) domain 1; PTHR47989:SF1: BNAA05G14800D PROTEIN; cd14066: STKc_IRAK; PTHR47989: OS01G0750732 PROTEIN; IPR017441: Protein kinase, ATP binding site; IPR000719: Protein kinase domain
Vitpa08g27640	12953	unnamed product	GO:0000398;GO:0005689	cd17058: Ubl_SNRNP25; IPR039690: U11/U12 small nuclear ribonucleoprotein 25kDa protein; IPR029071: Ubiquitin-like domain superfamily; G3DSA:3.10.20.90: -; mobidb-lite: consensus disorder prediction; IPR040610: SNRNP25, ubiquitin-like domain; PTHR14942:SF9: OS02G0188500 PROTEIN
Vitpa08g27630	21343	unnamed product	GO:0004672;GO:0005524;GO:0006468	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR47209: OS06G0639500 PROTEIN; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR000719: Protein kinase domain; mobidb-lite: consensus disorder prediction; IPR001245: Serine-threonine/tyrosine-protein kinase, catalytic domain; IPR011009: Protein kinase-like domain superfamily; G3DSA:3.30.200.20: Phosphorylase Kinase; domain 1; G3DSA:1.10.510.10: Transferase(Phosphotransferase) domain 1; cd14014: STKc_PknB_like; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR47209:SF6: KINASE, PUTATIVE-RELATED
Vitpa08g27670	25577	unnamed product	GO:0005515	CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR001611: Leucine-rich repeat; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR032675: Leucine-rich repeat domain superfamily; PTHR48004:SF42: PROTEIN TOO MANY MOUTHS; SSF52058: L domain-like; PTHR48004: OS01G0149700 PROTEIN
Vitpa08g27680	26263	unnamed product	GO:0005515	PTHR27000:SF714: PROTEIN TOO MANY MOUTHS; IPR001611: Leucine-rich repeat; SSF52058: L domain-like; PTHR27000: LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE FAMILY PROTEIN-RELATED; PR00019: Leucine-rich repeat signature
Vitpa08g27620	27574	unnamed product	-	-
Vitpa08g27690	29162	unnamed product	-	mobidb-lite: consensus disorder prediction; PTHR13780: AMP-ACTIVATED PROTEIN KINASE, GAMMA REGULATORY SUBUNIT; PTHR13780:SF101: SNF1-RELATED PROTEIN KINASE REGULATORY SUBUNIT GAMMA-LIKE PV42A
Vitpa08g27700	30792	unnamed product	-	PTHR13780:SF101: SNF1-RELATED PROTEIN KINASE REGULATORY SUBUNIT GAMMA-LIKE PV42A; IPR000644: CBS domain; G3DSA:3.10.580.10: -; SSF54631: CBS-domain pair; cd04641: CBS_euAMPK_gamma-like_repeat2; PTHR13780: AMP-ACTIVATED PROTEIN KINASE, GAMMA REGULATORY SUBUNIT
Vitpa08g27610	39018	unnamed product	GO:0003999;GO:0005737;GO:0006168;GO:0009116	IPR005764: Adenine phosphoribosyl transferase; PTHR11776:SF27: ADENINE PHOSPHORIBOSYLTRANSFERASE 5-LIKE ISOFORM X1; PTHR11776: ADENINE PHOSPHORIBOSYLTRANSFERASE; IPR029057: Phosphoribosyltransferase-like; IPR000836: Phosphoribosyltransferase domain

Name	Distance_SNP	Description	GO-labels	protein domains
Vitpa08g27710	43301	unnamed product	-	TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR34674: PHOSPHATIDYLCHOLINE:DIACYLGLYCEROL CHOLINEPHOSPHOTRANSFERASE 1-RELATED; mobidb-lite: consensus disorder prediction; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR34674:SF2: PHOSPHATIDYLCHOLINE:DIACYLGLYCEROL CHOLINEPHOSPHOTRANSFERASE 1-LIKE; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.
Vitpa08g27600	44666	unnamed product	-	-
Vitpa08g27590	46337	unnamed product	GO:0005515;GO:0043130	IPR003892: Ubiquitin system component CUE; PTHR31245: UBIQUITIN SYSTEM COMPONENT CUE PROTEIN; G3DSA:1.10.8.10: DNA helicase RuvA subunit; cd14279: CUE; IPR009060: UBA-like superfamily; PTHR31245:SF22: -; Coil: Coil
Vitpa08g27720	49908	unnamed product	GO:0016717;GO:0055114	TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR11351:SF94: BNAC05G37460D PROTEIN; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR015876: Acyl-CoA desaturase; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa08g27580	52620	unnamed product	GO:0005524;GO:0006457;GO:0009408;GO:0031072;GO:0051082	cd10747: DnaJ_C; IPR018253: DnaJ domain, conserved site; IPR012724: Chaperone DnaJ; IPR036869: Chaperone J-domain superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; G3DSA:2.10.230.10: -; G3DSA:2.60.260.20: Urease metallochaperone UreE; PTHR43096:SF22: MOLECULAR CHAPERONE HSP40/DNAJ FAMILY PROTEIN; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR036410: Heat shock protein DnaJ, cysteine-rich domain superfamily; IPR001305: Heat shock protein DnaJ, cysteine-rich domain; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; PTHR43096: DNAJ HOMOLOG 1, MITOCHONDRIAL-RELATED; IPR002939: Chaperone DnaJ, C-terminal; IPR008971: HSP40/DnaJ peptide-binding; IPR001623: DnaJ domain; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; mobidb-lite: consensus disorder prediction
Vitpa08g27730	59107	unnamed product	-	PTHR31087:SF60: PROTEIN LURP-ONE-RELATED 5; IPR025659: Tubby-like, C-terminal; IPR007612: LURP-one-related; IPR038595: LURP-one-related superfamily
Vitpa08g27740	63314	unnamed product	GO:0009055	SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR039391: Phytocyanin; IPR008972: Cupredoxin; IPR003245: Phytocyanin domain; cd13920: Stellacyanin; SignalP-noTM: SignalP-noTM; IPR028871: Blue (type 1) copper protein, binding site; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; mobidb-lite: consensus disorder prediction; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; PTHR33021:SF189: STELLACYANIN; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.
Vitpa08g27750	66889	unnamed product	GO:0004553;GO:0005975	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; G3DSA:3.20.20.80: Glycosidases; IPR000490: Glycoside hydrolase family 17; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; PTHR32227: GLUCAN ENDO-1,3-BETA-GLUCOSIDASE BG1-RELATED-RELATED; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR017853: Glycoside hydrolase superfamily; PTHR32227:SF357: GLYCOSYL HYDROLASE SUPERFAMILY PROTEIN
Vitpa08g27760	70862	unnamed product	GO:0003677	mobidb-lite: consensus disorder prediction; PTHR33729:SF12: METHYL-CPG-BINDING DOMAIN-CONTAINING PROTEIN 11; IPR016177: DNA-binding domain superfamily; IPR039622: Methyl-CpG-binding domain-containing protein 10/11; IPR001739: Methyl-CpG DNA binding; G3DSA:3.30.890.10: -

Name	Distance_SNP	Description	GO-labels	protein domains
Vitpa08g27570	71055	unnamed product	-	PTHR37224: OS02G0804400 PROTEIN; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; mobidb-lite: consensus disorder prediction
Vitpa08g27560	78618	unnamed product	GO:0003924;GO:0005525;GO:0006614;GO:0008312;GO:0048500	IPR022941: Signal recognition particle, SRP54 subunit; IPR036891: Signal recognition particle, SRP54 subunit, M-domain superfamily; IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR036225: SRP/SRP receptor, N-terminal; IPR013822: Signal recognition particle SRP54, helical bundle; IPR042101: Signal recognition particle SRP54, N-terminal domain superfamily; PTHR11564:SF33: SIGNAL RECOGNITION PARTICLE 54 KDA PROTEIN; cd17875: SRP54_G; G3DSA:3.40.50.300: -; IPR000897: Signal recognition particle, SRP54 subunit, GTPase domain; IPR004125: Signal recognition particle, SRP54 subunit, M-domain
Vitpa08g27770	82462	unnamed product	GO:0004672;GO:0004674;GO:0005524;GO:0006468	G3DSA:1.10.510.10: Transferase(Phosphotransferase) domain 1; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR47976: G-TYPE LECTIN S-RECEPTOR-LIKE SERINE/THREONINE-PROTEIN KINASE SD2-5; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; IPR011009: Protein kinase-like domain superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; G3DSA:3.30.200.20: Phosphorylase Kinase; domain 1; PTHR47976:SF26: G-TYPE LECTIN S-RECEPTOR-LIKE SERINE/THREONINE-KINASE SD2-5; IPR001480: Bulb-type lectin domain; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR017441: Protein kinase, ATP binding site; IPR000719: Protein kinase domain; cd01098: PAN_AP_plant; cd14066: STKc_IRAK; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR036426: Bulb-type lectin domain superfamily; SignalP-noTM: SignalP-noTM; IPR008271: Serine/threonine-protein kinase, active site; IPR024171: S-receptor-like serine/threonine-protein kinase; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa08g27550	87427	unnamed product	-	PTHR46856: PX DOMAIN-CONTAINING PROTEIN EREL1-RELATED; Coil: Coil; PTHR46856:SF3: PX DOMAIN-CONTAINING PROTEIN EREX
Vitpa08g27780	88271	unnamed product	GO:0004672;GO:0004674;GO:0005524;GO:0006468	IPR011009: Protein kinase-like domain superfamily; G3DSA:3.30.200.20: Phosphorylase Kinase; domain 1; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; PTHR47976: G-TYPE LECTIN S-RECEPTOR-LIKE SERINE/THREONINE-PROTEIN KINASE SD2-5; G3DSA:1.10.510.10: Transferase(Phosphotransferase) domain 1; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; cd14066: STKc_IRAK; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; cd01098: PAN_AP_plant; IPR001480: Bulb-type lectin domain; PTHR47976:SF26: G-TYPE LECTIN S-RECEPTOR-LIKE SERINE/THREONINE-KINASE SD2-5; IPR000719: Protein kinase domain; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR017441: Protein kinase, ATP binding site; IPR024171: S-receptor-like serine/threonine-protein kinase; SignalP-noTM: SignalP-noTM; IPR008271: Serine/threonine-protein kinase, active site; IPR036426: Bulb-type lectin domain superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa08g27540	89446	unnamed product	-	PTHR46856: PX DOMAIN-CONTAINING PROTEIN EREL1-RELATED; Coil: Coil; mobidb-lite: consensus disorder prediction
Vitpa08g27530	91163	unnamed product	GO:0035091	PTHR46856:SF3: PX DOMAIN-CONTAINING PROTEIN EREX; mobidb-lite: consensus disorder prediction; IPR001683: Phox homology; PTHR46856: PX DOMAIN-CONTAINING PROTEIN EREL1-RELATED; IPR036871: PX domain superfamily
Vitpa08g27790	93846	unnamed product	GO:0005515	IPR013026: Tetratricopeptide repeat-containing domain; IPR011990: Tetratricopeptide-like helical domain superfamily; PTHR26312: TETRATRICOPEPTIDE REPEAT PROTEIN 5; PTHR26312:SF73: TETRATRICOPEPTIDE REPEAT (TPR)-LIKE SUPERFAMILY PROTEIN; mobidb-lite: consensus disorder prediction

Suppl. Table 4: List and functional annotation of candidate genes identified within ±100 kb of the BLINK lead SNP associated with allantoin content on chromosome 7 in *Vitellaria paradoxa*.

Name	Distance SNP	description	GO-labels	protein domains
Vitpa07g01800	112	unnamed product	GO:0005524;GO:0016021;GO:0016887;GO:0042626;GO:0055085	IPR003593: AAA+ ATPase domain; IPR039421: Type I protein exporter; PTHR24221:SF415: ABC TRANSPORTER B FAMILY MEMBER 13-RELATED; IPR011527: ABC transporter type 1, transmembrane domain; IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR017871: ABC transporter, conserved site; IPR036640: ABC transporter type 1, transmembrane domain superfamily; IPR003439: ABC transporter-like; G3DSA:3.40.50.300: -; cd03249: ABC_MTABC3_MDL1_MDL2; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; mobidb-lite: consensus disorder prediction; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; cd18578: ABC_6TM_Pgp_ABCB1_D2_like; cd18577: ABC_6TM_Pgp_ABCB1_D1_like; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa07g01790	8101	unnamed product	GO:0003924;GO:0005525;GO:0007264	IPR005225: Small GTP-binding protein domain; cd04133: Rop_like; IPR001806: Small GTPase; IPR027417: P-loop containing nucleoside triphosphate hydrolase; SM00174: rho_sub_3; G3DSA:3.40.50.300: -; PTHR24072: RHO FAMILY GTPASE; PR00449: Transforming protein P21 ras signature; PTHR24072:SF336: RAC-LIKE GTP-BINDING PROTEIN 5; IPR003578: Small GTPase Rho
Vitpa07g01780	13950	unnamed product	GO:0015031;GO:0030906;GO:0042147	IPR016024: Armadillo-type fold; IPR042491: Vacuolar protein sorting-associated protein 35, C-terminal; PTHR11099:SF3: VACUOLAR PROTEIN SORTING-ASSOCIATED PROTEIN 35A; IPR005378: Vacuolar protein sorting-associated protein 35
Vitpa07g01770	29120	unnamed product	-	mobidb-lite: consensus disorder prediction; PTHR34660: MYB-LIKE PROTEIN X; PTHR34660:SF7: DNA LIGASE; Coil: Coil
Vitpa07g01760	35612	unnamed product	GO:0005515;GO:0006749;GO:0009055;GO:0015035;GO:0050220	PTHR12782:SF11: GLUTATHIONE S-TRANSFERASE FAMILY PROTEIN; G3DSA:1.20.1050.10: -; IPR034335: Prostaglandin E synthase 2, C-terminal; IPR036282: Glutathione S-transferase, C-terminal domain superfamily; IPR036249: Thioredoxin-like superfamily; PTHR12782: MICROSOMAL PROSTAGLANDIN E SYNTHASE-2; IPR011767: Glutaredoxin active site; IPR002109: Glutaredoxin; IPR004045: Glutathione S-transferase, N-terminal; IPR034334: Prostaglandin E synthase 2; SFLDG01182: Prostaglandin E synthase like; cd03040: GST_N_mPGES2
Vitpa07g01810	36683	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa07g01820	37122	unnamed product	GO:0008289	IPR039010: Synaptotagmin, SMP domain; IPR031468: Synaptotagmin-like mitochondrial-lipid-binding domain; PTHR10774:SF197: CALCIUM-DEPENDENT LIPID-BINDING (CALB DOMAIN) FAMILY PROTEIN-RELATED; IPR000008: C2 domain; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR035892: C2 domain superfamily; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR10774: EXTENDED SYNAPTOTAGMIN-RELATED; mobidb-lite: consensus disorder prediction; cd00030: C2; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; SSF49562: C2 domain (Calcium/lipid-binding domain, CaLB)
Vitpa07g01750	39763	unnamed product	-	-
Vitpa07g01740	41845	unnamed product	-	PTHR42647:SF22: BOI-RELATED E3 UBIQUITIN-PROTEIN LIGASE 3-RELATED; Coil: Coil; IPR013083: Zinc finger, RING/FYVE/PHD-type; IPR017066: E3 ubiquitin-protein ligase BOI-like
Vitpa07g01830	47343	unnamed product	-	mobidb-lite: consensus disorder prediction; PTHR47999: TRANSCRIPTION FACTOR MYB8-RELATED-RELATED; G3DSA:1.10.10.60: -; IPR001005: SANT/Myb domain; PTHR47999:SF23: TRANSCRIPTION REPRESSOR MYB4; IPR009057: Homeobox-like domain superfamily; IPR017930: Myb domain
Vitpa07g01730	49633	unnamed product	GO:0016788	PTHR45642:SF95: GDSL-LIKE LIPASE/ACYLHYDROLASE; IPR035669: GDSL lipase/esterase-like, plant; PTHR45642: GDSL ESTERASE/LIPASE EXL3; SSF52266: SGNH hydrolase; IPR001087: GDSL lipase/esterase; IPR036514: SGNH hydrolase superfamily
Vitpa07g01720	65651	unnamed product	GO:0016788	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region;

Name	Distance SNP	description	GO-labels	protein domains
				SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; Coil: Coil; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; PTHR45642:SF95: GDSL-LIKE LIPASE/ACYLHYDROLASE; IPR035669: GDSL lipase/esterase-like, plant; SignalP-noTM: SignalP-noTM; PTHR45642: GDSL ESTERASE/LIPASE EXL3; SSF52266: SGNH hydrolase; IPR001087: GDSL lipase/esterase; IPR036514: SGNH hydrolase superfamily
Vitpa07g01840	72799	unnamed product	GO:0003777;GO:0007018	Coil: Coil; PTHR24115:SF600: KINESIN-LIKE PROTEIN KIF23; IPR027640: Kinesin-like protein; mobidb-lite: consensus disorder prediction
Vitpa07g01850	87425	unnamed product	GO:0005634	PTHR47292:SF1: TRANSCRIPTION ELONGATION FACTOR (TFIIS) FAMILY PROTEIN; PTHR47292: TRANSCRIPTION ELONGATION FACTOR (TFIIS) FAMILY PROTEIN-RELATED; mobidb-lite: consensus disorder prediction; IPR035441: TFIIS/LEDGF domain superfamily; IPR017923: Transcription factor IIS, N-terminal
Vitpa07g01860	93581	unnamed product	-	CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR37714:SF1: PROTEIN, PUTATIVE-RELATED; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; PTHR37714: PROTEIN, PUTATIVE-RELATED; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa07g01870	99643	unnamed product	-	-

Suppl. Table 5: List and functional annotation of candidate genes identified within ±100 kb of the lead SNP associated with β-amyrin content on chromosome 8 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa08g15240	312	unnamed product	GO:0006886; GO:0015031; GO:0016192; GO:0030117; GO:0030123	mobidb-lite: consensus disorder prediction; IPR017105: Adaptor protein complex AP-3, delta subunit; IPR011989: Armadillo-like helical; IPR016024: Armadillo-type fold; IPR002553: Clathrin/coatomer adaptor, adaptin-like, N-terminal
Vitpa08g15230	2088	unnamed product	-	SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa08g15220	20238	unnamed product	-	SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa08g15210	30302	unnamed product	GO:0016020	SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; mobidb-lite: consensus disorder prediction; PTHR31042:SF20: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR003406: Glycosyl transferase, family 14; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR31042: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN-RELATED; SIGNAL_PEPTIDE_H_REGION:

Name	Distance_SNP	description	GO-labels	protein domains
				Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.
Vitpa08g15200	43258	unnamed product	GO:0016020	PTHR31042: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN-RELATED; IPR003406: Glycosyl transferase, family 14; PTHR31042:SF20: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN
Vitpa08g15190	58445	unnamed product	-	SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.
Vitpa08g15250	61291	unnamed product	GO:0008270	IPR013083: Zinc finger, RING/FYVE/PHD-type; IPR001841: Zinc finger, RING-type; IPR037275: Zinc finger, CTCHY-type superfamily; PTHR21319: RING FINGER AND CHY ZINC FINGER DOMAIN-CONTAINING PROTEIN 1; PTHR21319:SF50: RING FINGER AND CHY ZINC FINGER DOMAIN-CONTAINING PROTEIN 1-LIKE; IPR017921: Zinc finger, CTCHY-type; cd16464: RING-H2_Pirh2; IPR012312: Haemerythrin-like; SSF57850: RING/U-box; G3DSA:1.20.120.520: nmb1532 protein domain like; cd12108: Hr-like; IPR037274: Zinc finger, CHY-type superfamily; IPR008913: Zinc finger, CHY-type
Vitpa08g15180	72194	unnamed product	-	SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SignalP-TM: SignalP-TM; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; mobidb-lite: consensus disorder prediction; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa08g15170	82964	unnamed product	GO:0016020	PTHR31042:SF71: -; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR003406: Glycosyl transferase, family 14; PTHR31042: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN-RELATED; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa08g15160	89063	unnamed product	GO:0016020	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR003406: Glycosyl transferase, family 14; PTHR31042:SF113: CORE-2/I-BRANCHING ENZYME; PTHR31042: CORE-2/I-BRANCHING BETA-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE FAMILY PROTEIN-RELATED

Suppl. Table 6: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU lead SNP associated with campesterol content on chromosome 12 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g20510	867	unnamed product	GO:0003723;GO:0005515	IPR003891: Initiation factor eIF-4 gamma, MA3; IPR016021: MIF4G-like domain superfamily; PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; mobidb-lite: consensus disorder prediction; IPR016024: Armadillo-type fold; IPR003890: MIF4G-like, type 3
Vitpa12g20520	15457	unnamed product	GO:0003723;GO:0005515	PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; IPR016021: MIF4G-like domain superfamily; mobidb-lite: consensus disorder prediction; IPR003890: MIF4G-like, type 3; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR016024: Armadillo-type fold

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g20500	30309	unnamed product	GO:0005975	IPR012878: Beta-L-arabinofuranosidase, GH127; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; IPR008928: Six-hairpin glycosidase superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR31151: PROLINE-TRNA LIGASE (DUF1680); NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; PTHR31151:SF6: -; PS51257: Prokaryotic membrane lipoprotein lipid attachment site profile.; SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa12g20530	47275	unnamed product	-	PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; mobidb-lite: consensus disorder prediction; PTHR18034:SF5: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR003891: Initiation factor eIF-4 gamma, MA3
Vitpa12g20540	51522	unnamed product	GO:0003723;GO:0005515	IPR016024: Armadillo-type fold; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR003890: MIF4G-like, type 3; mobidb-lite: consensus disorder prediction; PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; IPR016021: MIF4G-like domain superfamily
Vitpa12g20490	54597	unnamed product	-	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa12g20550	56798	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g20560	69182	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g20570	70089	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g20480	73334	unnamed product	GO:0008033	IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR018022: IPP transferase; PTHR11088:SF60: TRNA DIMETHYLALLYLTRANSFERASE 9; IPR039657: Dimethylallyltransferase; PF01715: IPP transferase; G3DSA:3.40.50.300: -

Suppl. Table 7: List and functional annotation of candidate genes identified within ±100 kb of the BLINK lead SNP associated with campesterol content on chromosome 8 in *Vitellaria paradoxa*.

Name	Distance_SNP	Description	GO-labels	protein domains
Vitpa08g04010	3202	unnamed product	GO:0005515	IPR013083: Zinc finger, RING/FYVE/PHD-type; IPR001841: Zinc finger, RING-type; PF13637: Ankyrin repeats (many copies); PTHR24128:SF69: E3 UBIQUITIN-PROTEIN LIGASE XBOS32; IPR002110: Ankyrin repeat; mobidb-lite: consensus disorder prediction; SSF57850: RING/U-box; IPR020683: Ankyrin repeat-containing domain; IPR017907: Zinc finger, RING-type, conserved site; IPR036770: Ankyrin repeat-containing domain superfamily; PTHR24128: HOMEOBOX PROTEIN WARIAI
Vitpa08g04020	8138	unnamed product	GO:0003677;GO:0008289	IPR001356: Homeobox domain; SSF55961: Bet v1-like; IPR042160: Homeobox-leucine zipper protein GLABRA2/ANL2/PDF2/ATML1-like; IPR002913: START domain; mobidb-lite: consensus disorder prediction; PTHR45654:SF52: HOMEOBOX-LEUCINE ZIPPER PROTEIN HDG1; G3DSA:1.10.10.60: -; IPR009057: Homeobox-like domain superfamily; cd08875: START_ArGLABRA2_like
Vitpa08g04000	10813	unnamed product	GO:0004672;GO:0005515;GO:0005524;GO:0006468	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; IPR032675: Leucine-rich repeat domain superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; G3DSA:3.30.200.20: Phosphorylase Kinase; domain 1; PTHR27006: PROMASTIGOTE SURFACE ANTIGEN PROTEIN PSA; SSF52058: L domain-like; IPR001245: Serine-threonine/tyrosine-protein kinase, catalytic domain; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; G3DSA:1.10.510.10: Transferase(Phosphotransferase) domain 1; IPR008271: Serine/threonine-protein kinase, active site; cd14066: STKc_IRAK;

Name	Distance_SNP	Description	GO-labels	protein domains
				IPR000719: Protein kinase domain; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR017441: Protein kinase, ATP binding site; IPR001611: Leucine-rich repeat; IPR011009: Protein kinase-like domain superfamily; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; IPR013210: Leucine-rich repeat-containing N-terminal, plant-type; PTHR27006:SF250: LEUCINE-RICH REPEAT RECEPTOR PROTEIN KINASE EMS1; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; SignalP-noTM: SignalP-noTM
Vitpa08g03990	16102	unnamed product	GO:0003700;GO:0006355;GO:0043565	IPR036390: Winged helix DNA-binding domain superfamily; IPR027725: Heat shock transcription factor family; Coil: Coil; IPR036388: Winged helix-like DNA-binding domain superfamily; mobidb-lite: consensus disorder prediction; PTHR10015:SF333: HEAT STRESS TRANSCRIPTION FACTOR B-2A; IPR000232: Heat shock factor (HSF)-type, DNA-binding
Vitpa08g04030	19927	unnamed product	-	SSF57716: Glucocorticoid receptor-like (DNA-binding domain); IPR000988: Ribosomal protein L24e-related; G3DSA:3.30.160.440: -; IPR038630: Ribosomal protein L24e/L24 superfamily; Coil: Coil; PTHR10792:SF41: LOW QUALITY PROTEIN: 60S RIBOSOMAL PROTEIN L24-LIKE; IPR023442: Ribosomal protein L24e, conserved site; mobidb-lite: consensus disorder prediction
Vitpa08g03980	28980	unnamed product	-	PTHR33676:SF15: -; PTHR33676: COLD REGULATED PROTEIN 27; mobidb-lite: consensus disorder prediction
Vitpa08g03970	33640	unnamed product	GO:0009058;GO:0016746;GO:0016747	IPR001099: Chalcone/stilbene synthase, N-terminal; IPR012328: Chalcone/stilbene synthase, C-terminal; cd00831: CHS_like; IPR018088: Chalcone/stilbene synthase, active site; PTHR11877:SF82: CHALCONE SYNTHASE 2; IPR011141: Polyketide synthase, type III; IPR016039: Thiolase-like
Vitpa08g03960	36424	unnamed product	GO:0004864;GO:0009966;GO:0043666	mobidb-lite: consensus disorder prediction; IPR007062: Protein phosphatase inhibitor 2 (IPP-2); PTHR12398:SF30: PROTEIN GLC8-LIKE ISOFORM X1
Vitpa08g04040	36594	unnamed product	-	SSF53756: UDP-Glycosyltransferase/glycogen phosphorylase; PTHR11926:SF1099: GLYCOSYLTRANSFERASE; PTHR11926: GLUCOSYL/GLUCURONOSYL TRANSFERASES; G3DSA:3.40.50.2000: Glycogen Phosphorylase B;; cd03784: GT1_Gtf-like
Vitpa08g03950	45037	unnamed product	-	-
Vitpa08g03940	48604	unnamed product	GO:0005525	PS51417: small GTPase Arf family profile.; PTHR45909: ADP-RIBOSYLATION FACTOR-RELATED PROTEIN 1; G3DSA:3.40.50.300: -; PTHR45909:SF3: BNAA03G12710D PROTEIN; SM00177: arf_sub_2; cd04160: Arfrp1; IPR006689: Small GTPase superfamily, ARF/SAR type; IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR005225: Small GTP-binding protein domain
Vitpa08g03930	52972	unnamed product	GO:0007064;GO:0031390	PTHR47475: CHROMOSOME TRANSMISSION FIDELITY PROTEIN 8; IPR018607: Chromosome transmission fidelity protein 8
Vitpa08g03920	57063	unnamed product	GO:0003677	IPR001739: Methyl-CpG DNA binding; cd00122: MBD; PTHR34067:SF15: METHYL-CPG-BINDING DOMAIN-CONTAINING PROTEIN 13; mobidb-lite: consensus disorder prediction; IPR016177: DNA-binding domain superfamily; IPR038945: Methyl-CpG-binding domain-containing protein 13-like; G3DSA:3.30.890.10: -
Vitpa08g03910	65553	unnamed product	-	PTHR35472:SF8: CLAVATA3/ESR (CLE) GENE FAMILY MEMBER MTCLE10; mobidb-lite: consensus disorder prediction; PTHR35472: -
Vitpa08g03900	69783	unnamed product	-	PTHR10281:SF90: STEROID-BINDING PROTEIN 3-RELATED; Coil: Coil; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR001199: Cytochrome b5-like heme/steroid binding domain; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; mobidb-lite: consensus disorder prediction; PTHR10281: MEMBRANE-ASSOCIATED PROGESTERONE RECEPTOR COMPONENT-RELATED; IPR036400: Cytochrome b5-like heme/steroid binding domain superfamily; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.

Name	Distance_SNP	Description	GO-labels	protein domains
Vitpa08g03890	83271	unnamed product	GO:0005515	G3DSA:1.20.1280.50: -; PTHR38926: F-BOX DOMAIN CONTAINING PROTEIN, EXPRESSED; IPR001810: F-box domain; SSF52047: RNI-like; IPR036047: F-box-like domain superfamily; IPR032675: Leucine-rich repeat domain superfamily; PTHR38926:SF27: BNAANNG30020D PROTEIN
Vitpa08g04050	93452	unnamed product	GO:0004252;GO:0006508;GO:0016021	PTHR22936:SF86: RHOMBOID-LIKE PROTEIN 3; IPR022764: Peptidase S54, rhomboid domain; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR035952: Rhomboid-like superfamily; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR002610: Peptidase S54, rhomboid; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; SSF144091: Rhomboid-like
Vitpa08g03880	94387	unnamed product	-	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR34286:SF2: BNAA03G54340D PROTEIN; PTHR34286: TRANSMEMBRANE PROTEIN; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.

Suppl. Table 8: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU lead SNP associated with epigallocatechin content on chromosome 5 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa05g19650	1566	unnamed product	GO:0008270	SSF57850: RING/U-box; IPR022143: Protein of unknown function DUF3675; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR011016: Zinc finger, RING-CH-type; cd16495: RING_CH-C4HC3_MARCH; IPR033275: E3 ubiquitin-protein ligase MARCH-like; PTHR23012:SF181: E3 UBIQUITIN-PROTEIN LIGASE MARCH-RELATED; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR013083: Zinc finger, RING/FYVE/PHD-type; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; Coil: Coil; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa05g19660	10801	unnamed product	-	TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR038821: CLE-related protein 45-like; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa05g19640	16348	unnamed product	GO:0016787	cd00012: NBD_sugar-kinase_HSP70_actin; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR000407: Nucleoside phosphatase GDA1/CD39; PTHR11782:SF91: APYRASE 6-RELATED; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; G3DSA:3.30.420.40: -; G3DSA:3.30.420.150: Exopolyphosphatase. Domain 2; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.
Vitpa05g19670	32833	unnamed product	-	IPR009057: Homeobox-like domain superfamily; G3DSA:1.10.10.60: -; PTHR31948:SF102: HOMEOBOX DOMAIN, ZF-HD CLASS PROTEIN; PTHR31948: ZINC-FINGER HOMEODOMAIN PROTEIN 2; mobidb-lite: consensus disorder prediction; IPR006456: ZF-HD homeobox protein, Cys/His-rich dimerisation domain; IPR006455: Homeodomain, ZF-HD class
Vitpa05g19630	49790	unnamed product	-	PTHR48017:SF109: OS06G0228500 PROTEIN; PTHR48017: OS05G0424000 PROTEIN-RELATED; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; NON_CYTOPLASMIC_DOMAIN:

Name	Distance_SNP	description	GO-labels	protein domains
				Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR013057: Amino acid transporter, transmembrane domain
Vitpa05g19680	51812	unnamed product	GO:0001164;GO:0001188;GO:0006360;GO:0070860	IPR033599: RNA polymerase I transcription initiation factor TAF1B/Rrn7; mobidb-lite: consensus disorder prediction
Vitpa05g19690	65623	unnamed product	-	-
Vitpa05g19620	70199	unnamed product	GO:0003723;GO:0033897	IPR036430: Ribonuclease T2-like superfamily; IPR033697: Ribonuclease T2, eukaryotic; SignalP-noTM: SignalP-noTM; PTHR11240:SF67: BNAA02G26660D PROTEIN; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR033130: Ribonuclease T2, His active site 2; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR018188: Ribonuclease T2, His active site 1; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR001568: Ribonuclease T2-like
Vitpa05g19700	81682	unnamed product	GO:0003735;GO:0005840;GO:0006412	IPR008195: Ribosomal protein L34Ae; PTHR46741: OS09G0413600 PROTEIN; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR012870: Protein of unknown function DUF1666; mobidb-lite: consensus disorder prediction; PTHR46741:SF4: FINGER FYVE DOMAIN PROTEIN, PUTATIVE (DUF1666)-RELATED; IPR038562: Ribosomal protein L34Ae superfamily; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR018065: Ribosomal protein L34e, conserved site
Vitpa05g19610	84089	unnamed product	GO:0003723;GO:0033897	SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR018188: Ribonuclease T2, His active site 1; IPR001568: Ribonuclease T2-like; PTHR11240:SF55: F7A19.32 PROTEIN; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SignalP-noTM: SignalP-noTM; IPR033697: Ribonuclease T2, eukaryotic; IPR036430: Ribonuclease T2-like superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa05g19710	94739	unnamed product	GO:0005515;GO:0006749	cd03058: GST_N_Tau; IPR036282: Glutathione S-transferase, C-terminal domain superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; PTHR11260:SF622: GLUTATHIONE S-TRANSFERASE, AMINO-TERMINAL DOMAIN PROTEIN; IPR036249: Thioredoxin-like superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR004045: Glutathione S-transferase, N-terminal; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR040079: Glutathione Transferase family; SFLDG01152: Main.3: Omega- and Tau-like; SIGNAL_PEPTIDE: Signal peptide region; PTHR11260: GLUTATHIONE S-TRANSFERASE, GST, SUPERFAMILY, GST DOMAIN CONTAINING; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR010987: Glutathione S-transferase, C-terminal-like; cd03185: GST_C_Tau; G3DSA:1.20.1050.10: -
Vitpa05g19720	97592	unnamed product	GO:0003677;GO:0006355	IPR003441: NAC domain; IPR036093: NAC domain superfamily; PTHR31744:SF151: PROTEIN FEZ ISOFORM X1; PTHR31744: PROTEIN CUP-SHAPED COTYLEDON 2-RELATED

Suppl. Table 9: List and functional annotation of candidate genes identified within ±100 kb of the BLINK lead SNP associated with epigallocatechin content on chromosome 5 in *Vitellaria paradoxa*.

Name	Distance SNP	description	GO-labels	protein domains
Vitpa05g19650	1566	unnamed product	GO:0008270	SSF57850: RING/U-box; IPR022143: Protein of unknown function DUF3675; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR011016: Zinc finger, RING-CH-type; cd16495: RING_CH-C4HC3_MARCH; IPR033275: E3 ubiquitin-protein ligase MARCH-like; PTHR23012:SF181: E3 UBIQUITIN-PROTEIN LIGASE MARCH-RELATED; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR013083: Zinc finger, RING/FYVE/PHD-type; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; Coil: Coil; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa05g19660	10801	unnamed product	-	TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR038821: CLE-related protein 45-like; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa05g19640	16348	unnamed product	GO:0016787	cd00012: NBD_sugar-kinase_HSP70_actin; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR000407: Nucleoside phosphatase GDA1/CD39; PTHR11782:SF91: APYRASE 6-RELATED; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; G3DSA:3.30.420.40: -; G3DSA:3.30.420.150: Exopolyphosphatase. Domain 2; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.
Vitpa05g19670	32833	unnamed product	-	IPR009057: Homeobox-like domain superfamily; G3DSA:1.10.10.60: -; PTHR31948:SF102: HOMEOBOX DOMAIN, ZF-HD CLASS PROTEIN; PTHR31948: ZINC-FINGER HOMEODOMAIN PROTEIN 2; mobidb-lite: consensus disorder prediction; IPR006456: ZF-HD homeobox protein, Cys/His-rich dimerisation domain; IPR006455: Homeodomain, ZF-HD class
Vitpa05g19630	49790	unnamed product	-	PTHR48017:SF109: OS06G0228500 PROTEIN; PTHR48017: OS05G0424000 PROTEIN-RELATED; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR013057: Amino acid transporter, transmembrane domain
Vitpa05g19680	51812	unnamed product	GO:0001164 GO:0001188 GO:0006360 GO:0070860	IPR033599: RNA polymerase I transcription initiation factor TAF1B/Rrn7; mobidb-lite: consensus disorder prediction
Vitpa05g19690	65623	unnamed product	-	-
Vitpa05g19620	70199	unnamed product	GO:0003723 GO:0033897	IPR036430: Ribonuclease T2-like superfamily; IPR033697: Ribonuclease T2, eukaryotic; SignalP-noTM: SignalP-noTM; PTHR11240:SF67: BNAA02G26660D PROTEIN; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR033130: Ribonuclease T2, His active site 2; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR018188: Ribonuclease T2, His active site 1; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR001568: Ribonuclease T2-like
Vitpa05g19700	81682	unnamed product	GO:0003735 GO:0005840 GO:0006412	IPR008195: Ribosomal protein L34Ae; PTHR46741: OS09G0413600 PROTEIN; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR012870: Protein of unknown function DUF1666; mobidb-lite: consensus disorder prediction; PTHR46741:SF4: FINGER FYVE DOMAIN PROTEIN, PUTATIVE (DUF1666)-RELATED; IPR038562: Ribosomal protein L34Ae superfamily; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR018065: Ribosomal protein L34e, conserved site

Name	Distance SNP	description	GO-labels	protein domains
Vitpa05g19610	84089	unnamed product	GO:0003723; GO:0033897	SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR018188: Ribonuclease T2, His active site 1; IPR001568: Ribonuclease T2-like; PTHR11240:SF55: F7A19.32 PROTEIN; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SignalP-noTM: SignalP-noTM; IPR033697: Ribonuclease T2, eukaryotic; IPR036430: Ribonuclease T2-like superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa05g19710	94739	unnamed product	GO:0005515; GO:0006749	cd03058: GST_N_Tau; IPR036282: Glutathione S-transferase, C-terminal domain superfamily; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; PTHR11260:SF622: GLUTATHIONE S-TRANSFERASE, AMINO-TERMINAL DOMAIN PROTEIN; IPR036249: Thioredoxin-like superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; IPR004045: Glutathione S-transferase, N-terminal; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR040079: Glutathione Transferase family; SFLDG01152: Main.3: Omega- and Tau-like; SIGNAL_PEPTIDE: Signal peptide region; PTHR11260: GLUTATHIONE S-TRANSFERASE, GST, SUPERFAMILY, GST DOMAIN CONTAINING; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR010987: Glutathione S-transferase, C-terminal-like; cd03185: GST_C_Tau; G3DSA:1.20.1050.10: -
Vitpa05g19720	97592	unnamed product	GO:0003677; GO:0006355	IPR003441: NAC domain; IPR036093: NAC domain superfamily; PTHR31744:SF151: PROTEIN FEZ ISOFORM X1; PTHR31744: PROTEIN CUP-SHAPED COTYLEDON 2-RELATED

Suppl. Table 10: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU lead SNP associated with gallocatechin content on chromosome 5 in *Vitellaria paradoxa*.

Name	Distance SNP	description	GO-labels	protein domains
Vitpa05g31190	928	unnamed product	-	IPR016135: Ubiquitin-conjugating enzyme/RWD-like; SM00212: ubc_7; IPR000608: Ubiquitin-conjugating enzyme E2; PTHR46116:SF15: UBIQUITIN-CONJUGATING ENZYME E2 24-RELATED; PTHR46116: (E3-INDEPENDENT) E2 UBIQUITIN-CONJUGATING ENZYME
Vitpa05g31200	2046	unnamed product	-	PTHR11615: NITRATE, FORMATE, IRON DEHYDROGENASE; PTHR11615:SF263: CASP-LIKE PROTEIN 1C1; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR006702: Casparian strip membrane protein domain; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; SignalP-TM: SignalP-TM; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR006459: Casparian strip membrane protein
Vitpa05g31210	5404	unnamed product	GO:0003677;GO:0003899;GO:0006351;GO:0032549	IPR014724: RNA polymerase Rpb2, OB-fold; IPR007642: RNA polymerase Rpb2, domain 2; PTHR20856:SF28: DNA-DIRECTED RNA POLYMERASE SUBUNIT BETA; mobidb-lite: consensus disorder prediction; IPR015712: DNA-directed RNA polymerase, subunit 2; IPR007646: RNA polymerase Rpb2, domain 4; IPR007121: RNA polymerase, beta subunit, conserved site; IPR007644: RNA polymerase, beta subunit, protrusion; IPR007641: RNA polymerase Rpb2, domain 7; G3DSA:3.90.1100.10: -; IPR007120: DNA-directed RNA polymerase, subunit 2, hybrid-binding domain; IPR007645: RNA polymerase Rpb2, domain 3; SSF64484: beta and beta-prime subunits of DNA dependent RNA-polymerase; IPR037034: RNA polymerase Rpb2, domain 2 superfamily; IPR007647: RNA polymerase Rpb2, domain 5; G3DSA:3.90.1800.10: RNA polymerase alpha subunit dimerisation domain; IPR037033: DNA-directed RNA polymerase, subunit 2, hybrid-binding domain superfamily
Vitpa05g31180	18945	unnamed product	GO:0003723;GO:0003735;GO:0005840;GO:0006412	IPR036164: L21-like superfamily; IPR028909: Ribosomal protein L21-like; mobidb-lite: consensus disorder prediction; PTHR21349:SF5: BNAA08G12830D PROTEIN; IPR001787: Ribosomal protein L21; IPR018258: Ribosomal protein L21, conserved site
Vitpa05g31220	31441	unnamed product	GO:0000256;GO:0004038;GO:0008270;G	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR43668: ALLANTOINASE; G3DSA:3.20.20.140: -; IPR011059: Metal-dependent hydrolase, composite domain superfamily; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.;

Name	Distance SNP	description	GO-labels	protein domains
			O:0016787;GO:0016810;GO:0050897	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR032466: Metal-dependent hydrolase; PTHR43668:SF2: ZGC:103559; SignalP-noTM: SignalP-noTM; IPR017593: Allantoinase; IPR006680: Amidohydrolase-related
Vitpa05g31170	32564	unnamed product	GO:0016491;GO:0016702;GO:0046872;GO:0055114	IPR000907: Lipoxygenase; IPR013819: Lipoxygenase, C-terminal; IPR001246: Lipoxygenase, plant; mobidb-lite: consensus disorder prediction; IPR036226: Lipoxigenase, C-terminal domain superfamily; PTHR11771:SF133: LIPOXYGENASE; G3DSA:4.10.375.10: -
Vitpa05g31160	36764	unnamed product	-	-
Vitpa05g31150	39430	unnamed product	GO:0016702;GO:0046872;GO:0055114	IPR013819: Lipoxygenase, C-terminal; IPR000907: Lipoxygenase; PTHR11771:SF133: LIPOXYGENASE; IPR036226: Lipoxigenase, C-terminal domain superfamily; G3DSA:1.20.245.10: -
Vitpa05g31140	40492	unnamed product	-	SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; SignalP-TM: SignalP-TM; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR34274: TRANSMEMBRANE PROTEIN; PTHR34274:SF7: PROTEIN, PUTATIVE-RELATED; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.
Vitpa05g31130	44214	unnamed product	-	mobidb-lite: consensus disorder prediction; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR006852: Protein of unknown function DUF616; PTHR12956: ALKALINE CERAMIDASE-RELATED; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR12956:SF65: -
Vitpa05g31230	47831	unnamed product	-	mobidb-lite: consensus disorder prediction; PTHR36053: OSJNBB0017I01.18 PROTEIN; PTHR36053:SF1: OSJNBB0017I01.18 PROTEIN
Vitpa05g31240	58577	unnamed product	GO:0000439;GO:0001671;GO:0006289	IPR004598: Transcription factor TFIIH subunit p52/Tfb2; G3DSA:3.30.70.2610: -; IPR040662: Transcription factor Tfb2, C-terminal domain
Vitpa05g31120	63017	unnamed product	-	PTHR31676:SF121: -; IPR007493: Protein of unknown function DUF538; IPR036758: At5g01610-like superfamily
Vitpa05g31250	71773	unnamed product	-	TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR19282: TETRASPANIN; PTHR19282:SF432: TETRASPANIN-19-LIKE; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.
Vitpa05g31260	76307	unnamed product	GO:0006629	IPR005804: Fatty acid desaturase domain; PTHR12879: SPHINGOLIPID DELTA 4 DESATURASE/C-4 HYDROXYLASE PROTEIN DES2; PTHR12879:SF17: SPHINGOLIPID DELTA(4)-DESATURASE DES1-LIKE
Vitpa05g31270	76706	unnamed product	-	IPR017907: Zinc finger, RING-type, conserved site; SSF57850: RING/U-box; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR001841: Zinc finger, RING-type; PTHR15315:SF80: PEROXISOME BIOGENESIS FACTOR 10-LIKE; PTHR15315: RING FINGER PROTEIN 41, 151; IPR013083: Zinc finger, RING/FYVE/PHD-type; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa05g31280	83186	unnamed product	-	PTHR15315:SF80: PEROXISOME BIOGENESIS FACTOR 10-LIKE; PTHR15315: RING FINGER PROTEIN 41, 151
Vitpa05g31110	92385	unnamed product	-	IPR036758: At5g01610-like superfamily; IPR007493: Protein of unknown function DUF538; PTHR31676:SF121: -

Name	Distance SNP	description	GO-labels	protein domains
Vitpa05g31290	96893	unnamed product	-	-
Vitpa05g31300	98275	unnamed product	-	G3DSA:3.10.580.10: -; PTHR13780: AMP-ACTIVATED PROTEIN KINASE, GAMMA REGULATORY SUBUNIT; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR13780:SF39: CBS; SSF54631: CBS-domain pair
Vitpa05g31310	98670	unnamed product	-	PTHR13780: AMP-ACTIVATED PROTEIN KINASE, GAMMA REGULATORY SUBUNIT; cd02205: CBS_pair_SF; PTHR13780:SF108: CBS DOMAIN-CONTAINING PROTEIN CBSX5-LIKE; SSF54631: CBS-domain pair

Suppl. Table 11: List and functional annotation of candidate genes identified within ±100 kb of the BLINK lead SNP associated with gallocatechin content on chromosome 12 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g20510	867	unnamed product	GO:0003723;GO:0005515	IPR003891: Initiation factor eIF-4 gamma, MA3; IPR016021: MIF4G-like domain superfamily; PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; mobidb-lite: consensus disorder prediction; IPR016024: Armadillo-type fold; IPR003890: MIF4G-like, type 3
Vitpa12g20520	15457	unnamed product	GO:0003723;GO:0005515	PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; IPR016021: MIF4G-like domain superfamily; mobidb-lite: consensus disorder prediction; IPR003890: MIF4G-like, type 3; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR016024: Armadillo-type fold
Vitpa12g20500	30309	unnamed product	GO:0005975	IPR012878: Beta-L-arabinofuranosidase, GH127; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; IPR008928: Six-hairpin glycosidase superfamily; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR31151: PROLINE-TRNA LIGASE (DUF1680); NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; PTHR31151:SF6: -; PS51257: Prokaryotic membrane lipoprotein lipid attachment site profile.; SignalP-noTM: SignalP-noTM; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.
Vitpa12g20530	47275	unnamed product	-	PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; mobidb-lite: consensus disorder prediction; PTHR18034:SF5: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR003891: Initiation factor eIF-4 gamma, MA3
Vitpa12g20540	51522	unnamed product	GO:0003723;GO:0005515	IPR016024: Armadillo-type fold; PTHR18034:SF3: PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG; IPR003890: MIF4G-like, type 3; mobidb-lite: consensus disorder prediction; PTHR18034: CELL CYCLE CONTROL PROTEIN CWF22-RELATED; IPR016021: MIF4G-like domain superfamily
Vitpa12g20490	54597	unnamed product	-	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region
Vitpa12g20550	56798	unnamed product	-	mobidb-lite: consensus disorder prediction

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g20560	69182	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g20570	70089	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g20480	73334	unnamed product	GO:0008033	IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR018022: IPP transferase; PTHR11088:SF60: TRNA DIMETHYLALLYLTRANSFERASE 9; IPR039657: Dimethylallyltransferase; PF01715: IPP transferase; G3DSA:3.40.50.300: -

Suppl. Table 12: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU and BLINK lead SNP associated with lupeol content on chromosome 12 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g02660	218	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa12g02670	2766	unnamed product	GO:0046872	G3DSA:2.30.30.1190: -; PTHR12506:SF47: TRANSCRIPTION FACTOR C3H FAMILY-RELATED; PTHR12506: PROTEIN PHOSPHATASE RELATED; IPR036855: Zinc finger, CCCH-type superfamily; IPR000571: Zinc finger, CCCH-type; mobidb-lite: consensus disorder prediction
Vitpa12g02680	9340	unnamed product	GO:0046872	G3DSA:4.10.1000.10: CCCH zinc finger; PTHR12506: PROTEIN PHOSPHATASE RELATED; IPR036855: Zinc finger, CCCH-type superfamily; IPR000571: Zinc finger, CCCH-type; PF14608: RNA-binding, Nab2-type zinc finger; PTHR12506:SF43: ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 32
Vitpa12g02690	15612	unnamed product	GO:0003747;GO:0006415;GO:0016149	PTHR43804: LD18447P; SSF75620: Release factor; G3DSA:1.20.58.410: Release factor; G3DSA:3.30.70.1660: -; Coil: Coil; IPR000352: Peptide chain release factor class I; IPR005139: Peptide chain release factor; G3DSA:3.30.160.20: -; PTHR43804:SF4: PEPTIDE CHAIN RELEASE FACTOR APG3, CHLOROPLASTIC; IPR004373: Peptide chain release factor 1
Vitpa12g02650	30163	unnamed product	-	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; mobidb-lite: consensus disorder prediction; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.
Vitpa12g02700	31241	unnamed product	GO:0005975	IPR036249: Thioredoxin-like superfamily; IPR008928: Six-hairpin glycosidase superfamily; IPR024705: Spermatogenesis-associated protein 20; IPR012341: Six-hairpin glycosidase-like superfamily; IPR004879: Domain of unknown function DUF255
Vitpa12g02710	40853	unnamed product	GO:0003690;GO:0006355	PTHR13068: CGI-12 PROTEIN-RELATED; IPR038538: MTERF superfamily, mitochondrial/chloroplastic; IPR003690: Transcription termination factor, mitochondrial/chloroplastic; PTHR13068:SF133: MITOCHONDRIAL TRANSCRIPTION TERMINATION FACTOR FAMILY PROTEIN

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa12g02640	41174	unnamed product	GO:0006812;GO:0008324;GO:0016021;GO:0055085	IPR027470: Cation efflux protein, cytoplasmic domain; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; mobidb-lite: consensus disorder prediction; PTHR43840:SF15: MITOCHONDRIAL METAL TRANSPORTER 1-RELATED; IPR002524: Cation efflux protein; IPR036837: Cation efflux protein, cytoplasmic domain superfamily; IPR027469: Cation efflux transmembrane domain superfamily; PTHR43840: MITOCHONDRIAL METAL TRANSPORTER 1-RELATED
Vitpa12g02720	48808	unnamed product	GO:0003690;GO:0006355	IPR038538: MTERF superfamily, mitochondrial/chloroplastic; PTHR13068:SF133: MITOCHONDRIAL TRANSCRIPTION TERMINATION FACTOR FAMILY PROTEIN; IPR003690: Transcription termination factor, mitochondrial/chloroplastic; PTHR13068: CGI-12 PROTEIN-RELATED
Vitpa12g02630	54507	unnamed product	-	mobidb-lite: consensus disorder prediction; Coil: Coil
Vitpa12g02620	55664	unnamed product	-	PTHR10593: SERINE/THREONINE-PROTEIN KINASE RIO; G3DSA:3.30.160.60: Classic Zinc Finger; mobidb-lite: consensus disorder prediction; IPR013087: Zinc finger C2H2-type; IPR036236: Zinc finger C2H2 superfamily; IPR022755: Zinc finger, double-stranded RNA binding; PTHR10593:SF181: PROTEIN INDETERMINATE-DOMAIN 2-LIKE
Vitpa12g02730	59377	unnamed product	GO:0003690;GO:0006355	PTHR13068: CGI-12 PROTEIN-RELATED; PTHR13068:SF133: MITOCHONDRIAL TRANSCRIPTION TERMINATION FACTOR FAMILY PROTEIN; IPR003690: Transcription termination factor, mitochondrial/chloroplastic; IPR038538: MTERF superfamily, mitochondrial/chloroplastic
Vitpa12g02610	63897	unnamed product	-	G3DSA:1.10.10.60: -; PTHR13859:SF24: ARGININE-GLUTAMIC ACID DIPEPTIDE REPEAT PROTEIN; IPR009057: Homeobox-like domain superfamily; PTHR13859: ATROPHIN-RELATED; mobidb-lite: consensus disorder prediction
Vitpa12g02740	85718	unnamed product	-	PTHR46524:SF7: CW-TYPE ZINC FINGER; mobidb-lite: consensus disorder prediction; Coil: Coil; PTHR46524: CW-TYPE ZINC FINGER
Vitpa12g02600	88940	unnamed product	-	-
Vitpa12g02750	94420	unnamed product	-	PTHR46524: CW-TYPE ZINC FINGER; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR46524:SF7: CW-TYPE ZINC FINGER; mobidb-lite: consensus disorder prediction
Vitpa12g02590	94848	unnamed product	GO:0001228;GO:0005634;GO:0006355;GO:0016602;GO:0043565;GO:0046982	mobidb-lite: consensus disorder prediction; IPR003956: Transcription factor, NFYB/HAP3, conserved site; PR00615: CCAAT-binding transcription factor subunit A signature; IPR027113: Transcription factor NFYB/HAP3; IPR003958: Transcription factor CBF/NF-Y/archaeal histone domain; IPR009072: Histone-fold; PTHR11064:SF106: NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT B-5
Vitpa12g02580	97547	unnamed product	-	-
Vitpa12g02570	98287	unnamed product	GO:0015079;GO:0016020;GO:0071805	mobidb-lite: consensus disorder prediction; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR003855: Potassium transporter; PTHR30540:SF99: POTASSIUM TRANSPORTER; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside

Name	Distance_SNP	description	GO-labels	protein domains
				the membrane, in the extracellular region; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.

Suppl. Table 13: List and functional annotation of candidate genes identified within ±100 kb of the FarmCPU lead SNP associated with trans-cinnamic acid content on chromosome 5 in *Vitellaria paradoxa*.

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa05g30500	23	unnamed product	GO:0005739;GO:0065003	IPR010591: ATP11; PTHR13126:SF1: BNAA04G19940D PROTEIN
Vitpa05g30490	6095	unnamed product	-	Coil: Coil; IPR016024: Armadillo-type fold; cd17980: DEXHc_DHX35; PTHR12758:SF20: APOPTOSIS INHIBITOR 5-LIKE ISOFORM X1; mobidb-lite: consensus disorder prediction; IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR008383: Apoptosis inhibitory 5; IPR014001: Helicase superfamily 1/2, ATP-binding domain; G3DSA:3.40.50.300: -
Vitpa05g30510	12301	unnamed product	GO:0004601;GO:0006979; GO:0020037;GO:0042744; GO:0055114	SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR010255: Haem peroxidase superfamily; PTHR31235:SF56: PEROXIDASE 19; IPR019793: Peroxidases heam-ligand binding site; PTHR31235: PEROXIDASE 25-RELATED; G3DSA:1.10.420.10: Peroxidase; IPR019794: Peroxidase, active site; SignalP-noTM: SignalP-noTM; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; IPR002016: Haem peroxidase; IPR000823: Plant peroxidase; IPR033905: Secretory peroxidase; G3DSA:1.10.520.10: -; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region
Vitpa05g30520	14948	unnamed product	GO:0005634;GO:0006974; GO:0030915;GO:0031348; GO:0045892	IPR034561: Negative regulator of systemic acquired resistance SNI1
Vitpa05g30530	24699	unnamed product	GO:0016788	NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR45642:SF3: GDSL-LIKE LIPASE/ACYLHYDROLASE; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR035669: GDSL lipase/esterase-like, plant; PTHR45642: GDSL ESTERASE/LIPASE EXL3; SSF52266: SGNH hydrolase; IPR001087: GDSL lipase/esterase; IPR036514: SGNH hydrolase superfamily
Vitpa05g30540	28404	unnamed product	GO:0016788	PTHR45642:SF61: GDSL ESTERASE/LIPASE; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; PTHR45642: GDSL ESTERASE/LIPASE EXL3; SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR036514: SGNH hydrolase superfamily; IPR001087: GDSL lipase/esterase
Vitpa05g30480	31768	unnamed product	GO:0004386	IPR014001: Helicase superfamily 1/2, ATP-binding domain; G3DSA:3.40.50.300: -; G3DSA:1.20.120.1080: -; cd17980: DEXHc_DHX35; IPR001650: Helicase, C-terminal; Coil: Coil; IPR007502: Helicase-associated domain; IPR027417: P-loop containing nucleoside triphosphate hydrolase; IPR011709: Domain of unknown function DUF1605; cd18791: SF2_C_RHA; PTHR18934:SF208: OS05G0389800 PROTEIN; PTHR18934: ATP-DEPENDENT RNA HELICASE

Name	Distance_SNP	description	GO-labels	protein domains
Vitpa05g30550	41207	unnamed product	GO:0003677;GO:0003700; GO:0006355	PTHR31677: AP2 DOMAIN CLASS TRANSCRIPTION FACTOR; IPR036955: AP2/ERF domain superfamily; PTHR31677:SF98: ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR ERF091; mobidb-lite: consensus disorder prediction; IPR001471: AP2/ERF domain; IPR016177: DNA-binding domain superfamily
Vitpa05g30560	46480	unnamed product	GO:0003677;GO:0003700; GO:0006355	IPR036955: AP2/ERF domain superfamily; IPR001471: AP2/ERF domain; PTHR31190:SF173: PATHOGENESIS-RELATED GENES TRANSCRIPTIONAL ACTIVATOR PTI5; PTHR31190: DNA-BINDING DOMAIN; IPR016177: DNA-binding domain superfamily; mobidb-lite: consensus disorder prediction
Vitpa05g30570	55147	unnamed product	GO:0004553;GO:0005618; GO:0005975;GO:0006073; GO:0010411;GO:0016762; GO:0042546;GO:0048046	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; IPR016455: Xyloglucan endotransglucosylase/hydrolase; cd02176: GH16_XET; IPR013320: Concanavalin A-like lectin/glucanase domain superfamily; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; IPR000757: Glycoside hydrolase family 16; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; PTHR31062: XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE PROTEIN 8-RELATED; IPR008264: Beta-glucanase; IPR008263: Glycoside hydrolase, family 16, active site; PTHR31062:SF147: XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE; IPR010713: Xyloglucan endo-transglycosylase, C-terminal; G3DSA:2.60.120.200: -
Vitpa05g30470	56300	unnamed product	GO:0003824;GO:0006520; GO:0016831;GO:0019752; GO:0030170	PTHR11999: GROUP II PYRIDOXAL-5-PHOSPHATE DECARBOXYLASE; IPR015421: Pyridoxal phosphate-dependent transferase, major domain; IPR010977: Aromatic-L-amino-acid decarboxylase; PTHR11999:SF148: OS10G0400500 PROTEIN; IPR002129: Pyridoxal phosphate-dependent decarboxylase
Vitpa05g30580	57522	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa05g30590	58339	unnamed product	-	IPR004883: Lateral organ boundaries, LOB
Vitpa05g30460	61872	unnamed product	GO:0005509;GO:0016021; GO:0055085	TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR10891: EF-HAND CALCIUM-BINDING DOMAIN CONTAINING PROTEIN; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; G3DSA:1.10.238.10: -; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; SIGNAL_PEPTIDE: Signal peptide region; IPR011992: EF-hand domain pair; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR002048: EF-hand domain; PTHR10891:SF818: EXCHANGER FAMILY PROTEIN / CALCIUM-BINDING EF HAND FAMILY PROTEIN, PUTATIVE-RELATED; IPR004837: Sodium/calcium exchanger membrane region
Vitpa05g30600	65496	unnamed product	-	Coil: Coil; IPR004883: Lateral organ boundaries, LOB; PTHR31301: LOB DOMAIN-CONTAINING PROTEIN 4-RELATED; PTHR31301:SF135: LOB DOMAIN-CONTAINING PROTEIN 2
Vitpa05g30450	68824	unnamed product	GO:0005509	IPR018247: EF-Hand 1, calcium-binding site; TRANSMEMBRANE: Region of a membrane-bound protein predicted to be embedded in the membrane.; SignalP-TM: SignalP-TM; NON_CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region; CYTOPLASMIC_DOMAIN: Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.; PTHR10891:SF896: -; PTHR10891: EF-HAND CALCIUM-BINDING DOMAIN CONTAINING PROTEIN; TMhelix: Region of a membrane-bound protein predicted to be embedded in the membrane.; G3DSA:1.10.238.10: -; SIGNAL_PEPTIDE_N_REGION: N-terminal region of a signal peptide.;

Name	Distance_SNP	description	GO-labels	protein domains
				SIGNAL_PEPTIDE: Signal peptide region; SIGNAL_PEPTIDE_H_REGION: Hydrophobic region of a signal peptide.; IPR011992: EF-hand domain pair; SIGNAL_PEPTIDE_C_REGION: C-terminal region of a signal peptide.; IPR002048: EF-hand domain
Vitpa05g30440	74502	unnamed product	-	mobidb-lite: consensus disorder prediction
Vitpa05g30610	75646	unnamed product	-	PTHR31301: LOB DOMAIN-CONTAINING PROTEIN 4-RELATED; mobidb-lite: consensus disorder prediction; PTHR31301:SF135: LOB DOMAIN-CONTAINING PROTEIN 2; Coil: Coil; IPR004883: Lateral organ boundaries, LOB
Vitpa05g30620	85299	unnamed product	GO:0003924;GO:0005525; GO:0007264	G3DSA:3.40.50.300: -; PTHR24072: RHO FAMILY GTPASE; PR00449: Transforming protein P21 ras signature; IPR003578: Small GTPase Rho; IPR005225: Small GTP-binding protein domain; cd04133: Rop_like; IPR001806: Small GTPase; IPR027417: P-loop containing nucleoside triphosphate hydrolase; SM00174: rho_sub_3; PTHR24072:SF300: RAC-LIKE GTP-BINDING PROTEIN ARAC2
Vitpa05g30430	88514	unnamed product	GO:0005515	PTHR32295: IQ-DOMAIN 5-RELATED; mobidb-lite: consensus disorder prediction; PTHR32295:SF93: IQ-DOMAIN 9; IPR000048: IQ motif, EF-hand binding site; Coil: Coil
Vitpa05g30420	94869	unnamed product	GO:0003700	IPR005333: Transcription factor, TCP; Coil: Coil; mobidb-lite: consensus disorder prediction; PTHR31072:SF166: TRANSCRIPTION FACTOR TCP4; IPR017887: Transcription factor TCP subgroup
Vitpa05g30630	98292	unnamed product	GO:0003677	mobidb-lite: consensus disorder prediction; IPR009057: Homeobox-like domain superfamily; PTHR47288: WUSCHEL-RELATED HOMEOBOX 9; G3DSA:1.10.10.60: -; IPR001356: Homeobox domain

CHAPTER 6

General Discussion

Integrative understanding of diversity in *Vitellaria paradoxa*

This thesis provides an integrative assessment of agromorphological, genetic, biochemical, and metabolomic diversity in *Vitellaria paradoxa* populations from Benin. By combining field-based phenotyping, genome-wide SNP genotyping, targeted metabolomics, and genome-wide association analyses, this work offers a comprehensive view of the factors shaping variation in this semi-domesticated tree species. One of the central outcomes is the demonstration that diversity in shea is substantial at all levels of biological organization, yet structured in a non-congruent manner across traits and scales. Genetic analyses based on SNP markers revealed moderate levels of diversity and clear population structuring, consistent with previous studies in West African shea germplasm (Anyomi et al., 2023; Attikora et al., 2024; Hale et al., 2021) . In contrast, agromorphological traits and biochemical profiles exhibited high intra-population variability and weak geographic differentiation, as also observed in earlier phenotypic and biochemical surveys (Davrieux et al., 2010; Honfo et al., 2014). This partial decoupling between molecular structure and phenotypic expression is characteristic of long-lived, outcrossing tree species subject to both natural and anthropogenic selection pressures.

In shea, the absence of formal domestication and the persistence of farmer-mediated management practices appear to maintain broad variation within populations while limiting strong regional differentiation (Hall et al., 1996; Boffa, 1999). These findings highlight the limitations of single-layer analyses and demonstrate that integrative approaches are essential for capturing the biological complexity of perennial agroforestry species such as shea.

Genetic structure and its drivers in Benin shea populations

The population structure identified in this study confirms the presence of genetically differentiated groups within Benin's shea parklands, despite relatively low overall differentiation. The predominance of within-population variation is consistent with earlier reports based on both microsatellite and SNP markers (Anyomi et al., 2023; Attikora et al., 2024). However, the level of differentiation detected here suggests that genetic structuring is not negligible and should be accounted for in breeding and conservation strategies. The observed structure does not strictly follow agroecological boundaries, supporting the hypothesis that geography alone does not explain genetic differentiation in shea. Similar patterns have been reported in Côte d'Ivoire and Ghana, where gene flow and human-mediated selection were shown to blur geographic signals (Attikora et al., 2024; Odoi et al., 2023). In traditional parkland systems, selective retention of trees based on preferred traits such as fruit size, kernel

quality, or butter yield likely contributes to shaping genetic structure independently of climate or soil conditions (Merinosy et al., 2022).

Methodologically, the stringent SNP filtering applied in this study reduced potential artefacts associated with genotyping-by-sequencing approaches (Andrews et al., 2016; Shafer et al., 2017). Although this resulted in a lower number of retained markers compared to some studies, it enhanced the robustness of downstream analyses, particularly GWAS.

Agromorphological traits and their limited predictive value for quality

Agromorphological characterization revealed substantial phenotypic variability among shea trees for fruit, nut, and yield-related traits, confirming earlier observations in West African shea populations (Honfo et al., 2014; Merinosy et al., 2022). Strong correlations among certain morphological traits reflected coordinated growth patterns and developmental constraints, while relationships between morphology and biochemical quality were generally weak. In particular, fat content and fatty acid composition showed limited correlation with yield-related traits, in agreement with previous studies reporting weak or inconsistent associations between productivity and butter quality in West African shea population (Animasaun et al., 2019; Davrieux et al., 2010). This result indicates that high-yielding trees are not necessarily optimal in terms of industrial quality, challenging traditional selection strategies based primarily on visible or easily measurable traits.

These findings support the hypothesis that quality-related traits are governed by distinct genetic and regulatory mechanisms and cannot be reliably inferred from morphology alone. Consequently, selection strategies based solely on agromorphological descriptors risk overlooking genotypes with high economic value.

Biochemical and metabolomic diversity as a central asset of shea

A major contribution of this thesis is the detailed characterization of biochemical and metabolomic diversity in shea kernels, particularly for bioactive compounds of cosmetic and functional interest such as triterpenes, flavan-3-ols, phytosterols, and phenolic acids. The wide variation observed for these metabolites confirms earlier reports of chemical heterogeneity in shea butter (Akihisa et al., 2010b; Honfo et al., 2014), while providing a higher-resolution, metabolite-specific perspective.

The identification of distinct chemotypes demonstrates that biochemical diversity constitutes an independent dimension of variation, largely decoupled from morphology and geographic

origin. Similar conclusions have been drawn in metabolomic studies of other perennial crops, where specialized metabolites showed strong genotype dependence and limited spatial structuring (Matsuda et al., 2015; Naake et al., 2024).

However, environmental influences on metabolite accumulation cannot be excluded. Previous studies have shown that climatic factors such as latitude and rainfall may affect fatty acid composition and secondary metabolites in shea (Animasaun et al., 2019; Davrieux et al., 2010). The absence of multi-year and multi-site data in the present study therefore limits inference on trait stability, highlighting the need for genotype × environment analyses.

Genetic architecture of bioactive metabolites

Genome-wide association analyses revealed contrasting genetic architectures among metabolites. Polygenic control was evident for compounds such as allantoin and flavan-3-ols, whereas major-effect loci were identified for triterpenes such as lupeol and β-amyrin. This diversity of genetic architectures is consistent with metabolic GWAS findings in rice and *A. thaliana*, where both polygenic regulation and major-effect loci have been reported, depending on the metabolic pathway (Chan et al., 2010; Matsuda et al., 2015; Naake et al., 2024).

The convergence of association signals across MLMM, FarmCPU, and BLINK models strengthens confidence in the detected loci, as previously demonstrated in plant GWAS studies (Lipka et al., 2012; Liu et al., 2016). Functional annotation of candidate genes suggested that regulatory and transport-related mechanisms play a key role in metabolite accumulation, rather than structural variation in core biosynthetic enzymes, a pattern also reported in other plant metabolomic studies (Chan et al., 2010). To our knowledge, this study provides the first GWAS-based characterization of the genetic architecture of unsaponifiable metabolites, thereby filling a critical knowledge gap in plant metabolomic genetics. From a breeding perspective, these results support the use of marker-assisted selection for traits controlled by major loci and genomic selection for polygenic traits, as advocated for complex traits in perennial species (Fang and Luo, 2019; Luo, 2015).

Implications for breeding and domestication strategies

The combined analysis of genetic, phenotypic, and biochemical data provides a scientific basis for redefining shea improvement strategies. The identification of elite genotypes with favorable fatty acid profiles and high concentrations of bioactive compounds confirms the feasibility of targeted selection, even in a semi-domesticated species. Rather than pursuing uniform

ideotypes, breeding programs should embrace product-oriented selection pathways, distinguishing genotypes suited for chocolate production from those optimized for cosmetic or nutraceutical applications (Garti et al., 2019; Norazlina et al., 2021). The integration of genomic and metabolomic tools into participatory domestication programs could accelerate genetic gain while preserving diversity, as previously suggested for shea and other underutilized tree crops (Eyog Matig et al., 2002; Fang and Luo, 2019).

Conservation and sustainable valorization

The high intra-population diversity observed across all trait categories emphasizes the importance of in situ conservation. Shea parklands function as dynamic reservoirs of genetic and chemical diversity shaped by ecological processes and long-standing human management (Boffa, 1999; Hall et al., 1996).

By linking genetic diversity to biochemical quality and end-use potential, this thesis contributes to bridging conservation biology and socio-economic development. The identification of high-value genotypes supports differentiated value chains and aligns with current strategies promoting the sustainable valorization of non-timber forest products in West Africa (FAO et al., 2022; UNDP, 2019).

Limitations and Future research directions for shea improvement

Despite its integrative scope, this study is constrained by sample size, limited temporal replication, and the absence of functional validation. Similar limitations have been acknowledged in previous shea and metabolomic GWAS studies (Attikora et al., 2025; Hale et al., 2021). Future work should prioritize multi-year trials, expanded geographic sampling, and integration of transcriptomic and epigenomic data to validate candidate genes and refine breeding strategies.

Future research on *V. paradoxa* should aim to move beyond diversity characterization toward the development of improved planting material, while explicitly accounting for the species' slow reproductive cycle and semi-domesticated status. A major priority is to disentangle genetic control from environmental plasticity by conducting multi-site and multi-year evaluations of key quality traits, particularly fatty acid composition and bioactive metabolites, in order to assess their stability and heritability. Assessing genotype × environment interactions should rely on opportunistic multi-site and multi-year evaluations of already productive trees, combined with temporal replication and targeted clonal testing of elite genotypes, rather than on classical

breeding trials. Integrating transcriptomic and regulatory analyses with existing GWAS results will be essential to clarify the functional mechanisms underlying metabolite accumulation and to validate candidate genes. Given the long juvenile phase of shea, breeding strategies should focus on early selection approaches, combining marker-assisted selection for traits controlled by major loci with genomic selection for polygenic traits. In parallel, the development of clonal propagation systems based on elite genotypes represents a realistic pathway to accelerate genetic gain without relying on conventional breeding cycles. Future improvement efforts should also shift toward use-oriented ideotypes, distinguishing genotypes tailored for cosmetic, nutraceutical, or food applications, and co-defined with industrial stakeholders.

Finally, sustainable shea improvement will require an integrative and participatory framework that combines genomics, metabolomics, and traditional farmer knowledge, while ensuring the conservation of existing genetic resources. Such an approach will be critical to enhance the economic value of shea while preserving the biological diversity that underpins its long-term resilience.

Concluding synthesis

In conclusion, this thesis provides the first integrative dissection of genetic, agromorphological, biochemical, and metabolomic diversity in Benin shea germplasm. By demonstrating the complex, multi-layered nature of variation in *V. paradoxa*, it establishes a robust foundation for conservation, improvement, and sustainable valorization of this economically and culturally important species.

References

Abdulai, I., Krutovsky, K.V., Finkeldey, R., 2017. Morphological and genetic diversity of shea tree (Vitellaria paradoxa) in the savannah regions of Ghana. Genet. Resour. Crop Evol. 64, 1253–1268. https://doi.org/10.1007/s10722-016-0434-8

Ackman, R.G., 1998. Remarks on official methods employing boron trifluoride in the preparation of methyl esters of the fatty acids of fish oils. J. Am. Oil Chem. Soc. 75, 541–545. https://doi.org/10.1007/s11746-998-0263-9

Ackman, R.G., Sipos, J.C., 1964. Application of specific response factors in the gas chromatographic analysis of methyl esters of fatty acids with flame ionization detectors. J. Am. Oil Chem. Soc. 41, 377–378. https://doi.org/10.1007/BF02654818

Adedeji, A., Ogundari, I.O., Adenike Adedeji, A., Ogundolie, F.A., Sunday Ogunsina, B., Emmanuel, O.O., 2024. Shea Butter (Vitellaria paradoxa CF Gaertn) Production in Nigeria: Constraints and Prospects a Multifaceted Industry in the 21st Century. Presented at the International Conference on Science, Engineering and Business for Driving Sustainable Development Goals, SEB4SDG 2024. https://doi.org/10.1109/SEB4SDG60871.2024.10629804

Adejumo, O.E., Edun, T.A., Kolapo, A.L., Ayodele, O.A., 2022. Quality Characteristics and Alpha-Tocopherol Content of Shea Butter Samples from Selected Markets Within Three Southwestern States of Nigeria. Dhaka Univ. J. Pharm. Sci. 21, 105–115.

Afful, N.T., Abdulai, I., Azu, E., Elegba, W., Annor, C., Akama, C., Asare, K., Dentey, J., Amoatey, H.M., 2022. In vitro regeneration of Vitellaria paradoxa from shoot tip explants. BioTechnologia 103, 71.

Aholoukpè, H.S.N., Amadji, G.L., Koussihouèdé, H.K.I., 2020. Chapitre 5. Stocks de carbone dans les sols des zones agro-écologiques du Bénin, in: Chevallier, T., Razafimbelo, T.M., Chapuis-Lardy, L., Brossard, M. (Eds.), Carbone des sols en Afrique : Impacts des usages des sols et des pratiques agricoles, Synthèses. IRD Éditions, Marseille, pp. 101–122. https://doi.org/10.4000/books.irdeditions.34917

Ajibola, O.O., 2025. Nutritional composition and bioactive properties of tropical coconut palm fruit with functional food applications. Discov. Food 6, 21. https://doi.org/10.1007/s44187-025-00746-3

Akihisa, T., Kojima, N., Katoh, N., Ichimura, Y., Suzuki, H., Fukatsu, M., Maranz, S., Masters, E.T., 2010a. Triterpene alcohol and fatty acid composition of shea nuts from seven African countries. J. Oleo Sci. 59, 351–360.

Akihisa, T., Kojima, N., Katoh, N., Kikuchi, T., Fukatsu, M., Shimizu, N., Masters, E.T., 2011. Triacylglycerol and triterpene ester composition of shea nuts from seven African countries. J. Oleo Sci. 60, 385–391.

Akihisa, T., Kojima, N., Kikuchi, T., Yasukawa, K., Tokuda, H., T. Masters, E., Manosroi, A., Manosroi, J., 2010b. Anti-Inflammatory and Chemopreventive Effects of Triterpene Cinnamates and Acetates from Shea Fat. J. Oleo Sci. 59, 273–280. https://doi.org/10.5650/jos.59.273

Akplo, T.M., Dan, B.S.C., Houessou, L.G., Houinato, M.R.B., Sinsin, A.B., 2019. Typologie et structure des systèmes agroforestiers dans la commune de Djidja (Bénin). Rev. Int. Sci. 1, 29–39.

Akpona, J.D., Akpona, H., Djossa, B., Savi, M.K., Daïnou, K., Ayihouenou, B., Glele Kakaï, R.L., 2016. Impact of land use practices on traits and production of shea butter tree (Vitellaria paradoxa

C.F. Gaertn.) in Pendjari Biosphere Reserve in Benin. Agrofor. Syst. 90. https://doi.org/10.1007/s10457-015-9847-1

Alale, E.M., Tulashie, S.K., Miyittah, M., Baidoo, E.B., Adukpoh, K.E., Dadzie, E.O., Osei, C.A., Gah, B.K., Acquah, D., Quasi, P.A., 2024. Oil produced from Ghana Shea Nut crop for prospective industrial applications. Heliyon 10, e31171. https://doi.org/10.1016/j.heliyon.2024.e31171

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. The Global Structure of Chromosomes, in: Molecular Biology of the Cell. 4th Edition. Garland Science.

Alemawor, F., Agbenorhevi, J.K., Poku, A.K., 2014. Partial substitution of cocoa butter with processed shea butter in milk chocolate.

Aleza, K., Villamor, G.B., Nyarko, B.K., Wala, K., Akpagana, K., 2018. Shea (Vitellaria paradoxa Gaertn C. F.) fruit yield assessment and management by farm households in the Atacora district of Benin. PLOS ONE 13, e0190234. https://doi.org/10.1371/journal.pone.0190234

Allal, F., Piombo, G., Kelly, B.A., Okullo, J.B.L., Thiam, M., Diallo, O.B., Nyarko, G., Davrieux, F., Lovett, P.N., Bouvet, J.-M., 2013. Fatty acid and tocopherol patterns of variation within the natural range of the shea tree (Vitellaria paradoxa). Agrofor. Syst. 87, 1065–1082. https://doi.org/10.1007/s10457-013-9621-1

Allal, F., Sanou, H., Millet, L., Vaillant, A., Camus-Kulandaivelu, L., Logossa, Z.A., Lefèvre, F., Bouvet, J.-M., 2011. Past climate changes explain the phylogeography of Vitellaria paradoxa over Africa. Heredity 107, 174–186. https://doi.org/10.1038/hdy.2011.5

Allal, F., Vaillant, A., Sanou, H., Kelly, B., Bouvet, J.-M., 2008. Isolation and characterization of new microsatellite markers in shea tree (Vitellaria paradoxa C. F. Gaertn). Mol. Ecol. Resour. 8, 822–824. https://doi.org/10.1111/j.1755-0998.2007.02079.x

American Shea Butter Institute, 2011. The Factors That Determine the Price of Certified Premium Grade A Shea Butter.

Amin, A., Zaman, W., Park, S., 2025. Harnessing Multi-Omics and Predictive Modeling for Climate-Resilient Crop Breeding: From Genomes to Fields. Genes 16, 809. https://doi.org/10.3390/genes16070809

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92. https://doi.org/10.1038/nrg.2015.28

Anil, K., Roshan, D., 2021. Shea Butter Market by Type (Raw & Unrefined and Refined & Ultra-Refined) and Application (Food and Cosmetics): Global Opportunity Analysis and Industry Forecast, 2021-2030 (No. 5480870). Research and Markets.

Animasaun, D.A., Oyedeji, S., Olorunmaiye, K.S., Azeez, M.A., Tijani, I.A., Morakinyo, J.A., 2019. Morpho-chemical divergence and fatty acid profile of shea tree seeds (Vitellaria paradoxa) collected from different locations in Kwara State, Nigeria. Acta Bot. Croat. 78, 17–24. https://doi.org/10.2478/botcro-2019-0002

Anwar, F., Zafar, S.N., Rashid, U., 2006. Characterization of Moringa oleifera seed oil from drought and irrigated regions of Punjab, Pakistan. Grasas Aceites 57, 160–168. https://doi.org/10.3989/gya.2006.v57.i2.32

Anyomi, W.E., Barnor, M.T., Eleblu, J.S.Y., Danquah, A., Avicor, S.W., Ofori, K., Hale, I., Padi, F.K., Danquah, E.Y., 2023. Elucidation of the Genetic Diversity within Some in situ Shea Germplasm in Ghana. Agronomy 13, 2256. https://doi.org/10.3390/agronomy13092256

AOAC, 1999. Official methods of analysis of AOAC International. AOAC international.

AOAC, 1995. Official Methods of Analysis of AOAC International. AOAC International.

Arhin, L., Abdullah, S. nor A., Jaafar, J.N., Izan Ramlee, S., 2023. Conventional and modern breeding technologies for improving dwarf coconut cultivars: A review. J. Hortic. Sci. Biotechnol. 98, 551–562. https://doi.org/10.1080/14620316.2023.2195863

Asare, E.K., Avicor, S.W., Dogbatse, J.A., Anyomi, E.W., 2019. Occurrence of mistletoes on shea trees in Northern Ghana. Afr. Crop Sci. J. 27, 679–686. https://doi.org/10.4314/acsj.v27i4.9

Ashby, J.A., 2009. The impact of participatory plant breeding. Plant Breed. Farmer Particip. 649–671.

Attikora, A.J., Diarrassouba, N., Martial Yao, S.D., De Clerck, C., Souleymane, S., Taofic, A., Ludivine, L., 2023. Morphological traits and sustainability of plus shea trees (Vitellaria paradoxa C.F.Gaertn.) in Côte d'Ivoire. BASE. https://doi.org/10.25518/1780-4507.20462

Attikora, A.J.P., Kouassi, K.A., Yao, S.D.M., Dago, D.N., Silué, S., De Clerck, C., Diarrassouba, N., Alabi, T., Achigan-Dako, E.G., Fauconnier, M.-L., Danthine, S., Lassois, L., 2025. Genome-wide association study of fat content and fatty acid composition of shea tree (Vitellaria paradoxa C.F. Gaertn subsp. paradoxa). BMC Genomics 26, 164. https://doi.org/10.1186/s12864-025-11344-z

Attikora, A.J.P., Yao, S.D.M., Dago, D.N., Silué, S., De Clerck, C., Kwibuka, Y., Diarrassouba, N., Alabi, T., Achigan-Dako, E.G., Lassois, L., 2024. Genetic diversity and population structure of superior shea trees (Vitellaria paradoxa subsp. paradoxa) using SNP markers for the establishment of a core collection in Côte d'Ivoire. BMC Plant Biol. 24, 913. https://doi.org/10.1186/s12870-024-05617-0

Ayelazuno, R.A., Yaro, J.A., 2024. Shea butter livelihoods and challenges faced by women in the Kassena Nankana districts, Ghana. SN Soc. Sci. 4, 18. https://doi.org/10.1007/s43545-023-00810-8

Azevedo, A.O.N., Azevedo, C.D. de O., Santos, P.H.A.D., Ramos, H.C.C., Boechat, M.S.B., Arêdes, F.A.S., Ramos, S.R.R., Mirizola, L.Â., Perera, L., Aragão, W.M., Pereira, M.G., 2018. Selection of legitimate dwarf coconut hybrid seedlings using DNA fingerprinting. Crop Breed. Appl. Biotechnol. 18, 409–416. https://doi.org/10.1590/1984-70332018v18n4a60

Baldini, M., Giovanardi, R., Tahmasebi-Enferadi, S., Vannozzi, G.P., 2002. Effects of water regime on fatty acid accumulation and final fatty acid composition in the oil of standard and high oleic sunflower hybrids. Ital. J. Agron. 6, 119–126.

Bates, P.D., Durrett, T.P., Ohlrogge, J.B., Pollard, M., 2009. Analysis of Acyl Fluxes through Multiple Pathways of Triacylglycerol Synthesis in Developing Soybean Embryos. Plant Physiol. 150, 55–72. https://doi.org/10.1104/pp.109.137737

Batugal, P., Bourdeix, R., 2005. Chapter 5: Germoplasm Use: Conventional coconut breeding, in: Coconut Genetic Resources. International Plant Genetic Resources Institute – Regional Office for Asia, the Pacific and Oceania (IPGRI-APO), Serdang, Selangor DE, Malaysia. IPGRI-APO PO Box 236, UPM Post Office Serdang 43400 Selangor Darul Ehsan, Malaysia, p. 251pp.

Baud, S., Lepiniec, L., 2010. Physiological and developmental regulation of seed oil production. Prog. Lipid Res. 49, 235–249. https://doi.org/10.1016/j.plipres.2010.01.001

Beavis, W., Lamkey, K., Espinosa, K., Mahama, A.A., Suza, W., 2023. Chapter 5: Gene Effect, in: Quantitative Genetics for Plant Breeding. Iowa State University Digital Press.

Beczner, C., Gower, B., Vizzoni, P., 2006. Women's Gold: Finding a Market for Dagara Shea Butter, in: Corporate Citizenship in Africa. Routledge.

Bellaloui, N., Mengistu, A., Kassem, M.A., 2013. Effects of Genetics and Environment on Fatty Acid Stability in Soybean Seed. Food Nutr. Sci. 4, 165–175. https://doi.org/10.4236/fns.2013.49A1024

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.

Bhat, J.A., Ali, S., Salgotra, R.K., Mir, Z.A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P.K., Singh, G.P., Prabhu, K.V., 2016. Frontiers | Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. https://doi.org/10.3389/fgene.2016.00221

Board, A.J., Crowther, J.M., Acevedo-Fani, A., Meisrimler, C.-N., Jameson, G.B., Dobson, R.C.J., 2022. How plants solubilise seed fats: revisiting oleosin structure and function to inform commercial applications. Biophys. Rev. 14, 257–266. https://doi.org/10.1007/s12551-021-00923-5

Boateng, L., Ansong, R., Owusu, W.B., Steiner-Asiedu, M., 2016. Coconut oil and palm oil's role in nutrition, health and national development: A review. Ghana Med. J. 50, 189–196.

Bodjrenou, G.J., Gbètoho, M., Honfo, F., Kahn, C.J.F., Schranz, M.E., Gnanglè, P.C., Hehn, A., Olry, A., Achigan-Dako, E.G., 2026a. Genetic structure of the shea tree (Vitellaria paradoxa C.F. Gaertn.) population and fatty acid-based selection of elite genotypes in Benin. Physiol. Mol. Biol. Plants 1–16. https://doi.org/10.1007/s12298-026-01734-2

Bodjrenou, G.J., Grosjean, J., Honfo, F.G., Hehn, A., Achigan-Dako, E.G., Olry, A., 2026b. Metabolomic-guided diversity analysis reveals chemotypic groups and cosmetic-relevant biomarkers in shea trees (Vitellaria paradoxa C.F. Gaertn.) for industrial applications. Ind. Crops Prod. 239, 122354. https://doi.org/10.1016/j.indcrop.2025.122354

Boffa, J.M., 1999. Agroforestry parklands in sub-Saharan Africa. Food and Agriculture Organization of the United Nations, Rome.

Boffa, J.-M.J., 1996. Productivity and management of agroforestry parklands in the Sudan zone of Burkina Faso, West Africa. Purdue University.

Bondé, L., Ouédraogo, O., Ouédraogo, I., Thiombiano, A., Boussim, J.I., 2019. Variability and estimating in fruiting of shea tree (Vitellaria paradoxa C.F. Gaertn) associated to climatic conditions in West Africa: implications for sustainable management and development. Plant Prod. Sci. 22, 143–158. https://doi.org/10.1080/1343943X.2018.1541712

Bourgis, F., Kader, J.-C., Barret, P., Renard, M., Robinson, D., Robinson, C., Delseny, M., Roscoe, T.J., 1999. A Plastidial Lysophosphatidic Acid Acyltransferase from Oilseed Rape1. Plant Physiol. 120, 913–922. https://doi.org/10.1104/pp.120.3.913

Boussaha, M., Guyomard, R., Cabau, C., Esquerré, D., Quillet, E., 2012. Development and characterisation of an expressed sequence tags (EST)-derived single nucleotide polymorphisms

(SNPs) resource in rainbow trout. BMC Genomics 13, 238. https://doi.org/10.1186/1471-2164-13-238

Boussim, J.I., Sallé, G., Guinko, S., 1993. Tapinanthus, parasite du Karité au Burkina Faso. 1ère partie : identification et distribution. BOIS FORETS Trop. 238, 45–52. https://doi.org/10.19182/bft1993.238.a19772

Braegelmann, C., Niebel, D., Ferring-Schmitt, S., Fetter, T., Landsberg, J., Hölzel, M., Effern, M., Glodde, N., Steinbuch, S., Bieber, T., Wenzel, J., 2022. Epigallocatechin-3-gallate exhibits anti-inflammatory effects in a human interface dermatitis model—implications for therapy. J. Eur. Acad. Dermatol. Venereol. 36, 144–153. https://doi.org/10.1111/jdv.17710

Brigelius-Flohé, R., 2009. Vitamin E: the shrew waiting to be tamed. Free Radic. Biol. Med. 46, 543–554. https://doi.org/10.1016/j.freeradbiomed.2008.12.007

Bup, D.N., Mohagir, A.M., Kapseu, C., Mouloungui, Z., 2014. Production zones and systems, markets, benefits and constraints of shea (Vitellaria paradoxa CF Gaertn) butter processing. OCL 21, D206. https://doi.org/10.1051/ocl/2013045

Burnett, C.L., Bergfeld, W.F., Belsito, D.V., Hill, R.A., Klaassen, C.D., Liebler, D.C., MarksJr, J.G., Shank, R.C., Slaga, T.J., Snyder, P.W., Fiume, M., Heldreth, B., 2024. Safety Assessment of Butyrospermum parkii (Shea)-Derived Ingredients as Used in Cosmetics. Int. J. Toxicol. 43, 82S-95S. https://doi.org/10.1177/10915818231224230

Burns, E.M., Tober, K.L., Riggenbach, J.A., Kusewitt, D.F., Young, G.S., Oberyszyn, T.M., 2013. Differential Effects of Topical Vitamin E and C E Ferulic® Treatments on Ultraviolet Light B-Induced Cutaneous Tumor Development in Skh-1 Mice. PLOS ONE 8, e63809. https://doi.org/10.1371/journal.pone.0063809

Câmara, J.S., Perestrelo, R., Ferreira, R., Berenguer, C.V., Pereira, J.A.M., Castilho, P.C., 2024. Plant-Derived Terpenoids: A Plethora of Bioactive Compounds with Several Health Functions and Industrial Applications—A Comprehensive Overview. Molecules 29, 3861. https://doi.org/10.3390/molecules29163861

Campbell, A., 2001. Genetics, in: Brenner, S., Miller, J.H. (Eds.), Encyclopedia of Genetics. Academic Press, New York, pp. 856–857. https://doi.org/10.1006/rwgn.2001.0545

Capasso, L., De Masi, L., Sirignano, C., Maresca, V., Basile, A., Nebbioso, A., Rigano, D., Bontempo, P., 2025. Epigallocatechin Gallate (EGCG): Pharmacological Properties, Biological Activities and Therapeutic Potential. Molecules 30, 654. https://doi.org/10.3390/molecules30030654

Cardi, C., Vaillant, A., Sanou, H., Kelly, B.A., Bouvet, J.-M., 2005. Characterization of microsatellite markers in the shea tree (Vitellaria paradoxa C. F Gaertn) in Mali. Mol. Ecol. Notes 5, 524–526. https://doi.org/10.1111/j.1471-8286.2005.00980.x

Chakrabarty, S., Külheim, C., 2025. Trends in tree improvement methods: from classical breeding to genomic technologies. Tree Genet. Genomes 21, 1–21. https://doi.org/10.1007/s11295-025-01698-6

Chan, E.K.F., Rowe, H.C., Hansen, B.G., Kliebenstein, D.J., 2010. The Complex Genetic Architecture of the Metabolome. PLoS Genet. 6, e1001198. https://doi.org/10.1371/journal.pgen.1001198

Chang, Z.-Y., Chen, C.-W., Tsai, M.-J., Chen, C.-C., Alshetaili, A., Hsiao, Y.-T., Fang, J.-Y., 2023. The elucidation of structure-activity and structure-permeation relationships for the cutaneous delivery of phytosterols to attenuate psoriasiform inflammation. Int. Immunopharmacol. 119, 110202. https://doi.org/10.1016/j.intimp.2023.110202

Chibor Bariwere, S., Kiin-Kabari, D.B., Eke-Ejiofor, J., 2017. Physicochemical properties and fatty acid profile of shea butter and fluted pumpkin seed oil, a suitable blend in bakery fat production. Int. J. Nutr. Food Sci. 6, 122–128.

Chiriacò, M.V., Bellotta, M., Jusić, J., Perugini, L., 2022. Palm oil's contribution to the United Nations sustainable development goals: outcomes of a review of socio-economic aspects. Environ. Res. Lett. 17, 063007. https://doi.org/10.1088/1748-9326/ac6e77

Choungo Nguekeng, P.B., Hendre, P., Tchoundjeu, Z., Kalousová, M., Tchanou Tchapda, A.V., Kyereh, D., Masters, E., Lojka, B., 2021a. The Current State of Knowledge of Shea Butter Tree (Vitellaria paradoxa C.F.Gaertn.) for Nutritional Value and Tree Improvement in West and Central Africa. Forests 12, 1740. https://doi.org/10.3390/f12121740

Choungo Nguekeng, P.B., Hendre, P., Tchoundjeu, Z., Kalousová, M., Tchanou Tchapda, A.V., Kyereh, D., Masters, E., Lojka, B., 2021b. The current state of knowledge of shea butter tree (Vitellaria paradoxa cf gaertner.) for nutritional value and tree improvement in west and central africa. Forests 12, 1740.

Collados, R., Andreu, V., Picorel, R., Alfonso, M., 2006. A light-sensitive mechanism differently regulates transcription and transcript stability of omega3 fatty-acid desaturases (FAD3, FAD7 and FAD8) in soybean photosynthetic cell suspensions. FEBS Lett. 580, 4934–4940. https://doi.org/10.1016/j.febslet.2006.07.087

Contreras, C., Pierantozzi, P., Maestri, D., Tivani, M., Searles, P., Brizuela, M., Fernández, F., Toro, A., Puertas, C., Trentacoste, E.R., Kiessling, J., Mariotti, R., Baldoni, L., Mousavi, S., Fernandez, P., Moschen, S., Torres, M., 2022. How Temperatures May Affect the Synthesis of Fatty Acids during Olive Fruit Ripening: Genes at Work in the Field. Plants Basel Switz. 12, 54. https://doi.org/10.3390/plants12010054

Corley, R.H.V., Tinker, P.B.H., 2008. The Oil Palm. John Wiley & Sons.

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., Nouy, B., Omoré, A., Pomiès, V., Riou, V., Suryana, E., Bouvet, J.-M., 2015. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). TAG Theor. Appl. Genet. Theor. Angew. Genet. 128, 397–410. https://doi.org/10.1007/s00122-014-2439-z

Crossa, J., Montesinos-López, O.A., Pérez-Rodríguez, P., Costa-Neto, G., Fritsche-Neto, R., Ortiz, R., Martini, J.W.R., Lillemo, M., Montesinos-López, A., Jarquin, D., Breseghello, F., Cuevas, J., Rincent, R., 2022. Genome and Environment Based Prediction ModelsPrediction models and Methods of Complex TraitsComplex traits Incorporating Genotype × Environment Interaction, in: Ahmadi, N., Bartholomé, J. (Eds.), Genomic Prediction of Complex Traits: Methods and Protocols. Springer US, New York, NY, pp. 245–283. https://doi.org/10.1007/978-1-0716-2205-6_9

Custom Market Insights, 2023. Shea Butter Market Size, Trends and Insights By Grade (Pharmaceutical, Industrial), By Application (Pharmaceuticals, Food and Preservatives, Cosmetics, Others), and By Region - Global Industry Overview, Statistical Data, Competitive Analysis, Share, Outlook, and Forecast 2023–2032 (No. CMI22814). Custom Market Insights. (No. No. CMI22814).

Davrieux, F., Allal, F., Piombo, G., Kelly, B., Okulo, J.B., Thiam, M., Diallo, O.B., Bouvet, J.-M., 2010. Near Infrared Spectroscopy for High-Throughput Characterization of Shea Tree (Vitellaria paradoxa) Nut Fat Profiles. J. Agric. Food Chem. 58, 7811–7819. https://doi.org/10.1021/jf100409v

Degen, B., Müller, N.A., 2023. A simulation study comparing advanced marker-assisted selection with genomic selection in tree breeding programs. G3 GenesGenomesGenetics 13, jkad164. https://doi.org/10.1093/g3journal/jkad164

Deschamps, S., Llaca, V., May, G.D., 2012. Genotyping-by-Sequencing in Plants. Biology 1, 460–483. https://doi.org/10.3390/biology1030460

Desclaux, D., Roumet, P., 1996. Impact of drought stress on the phenology of two soybean (Glycine max L. Merr) cultivars. Field Crops Res. 46, 61–70. https://doi.org/10.1016/0378-4290(95)00086-0

Di, M., Zhang, Q., Wang, J., Xiao, X., Huang, J., Ma, Y., Yang, H., Li, M., 2022. Epigallocatechin-3-gallate (EGCG) attenuates inflammatory responses and oxidative stress in lipopolysaccharide (LPS)-induced endometritis via silent information regulator transcript-1 (SIRT1)/nucleotide oligomerization domain (NOD)-like receptor pyrin domain-containing 3 (NLRP3) pathway. J. Biochem. Mol. Toxicol. 36, e23203. https://doi.org/10.1002/jbt.23203

Di Vincenzo, D., Maranz, S., Serraiocco, A., Vito, R., Wiesman, Z., Bianchi, G., 2005a. Regional variation in shea butter lipid and triterpene composition in four African countries. J. Agric. Food Chem. 53, 7473–7479. https://doi.org/10.1021/jf0509759

Di Vincenzo, D., Maranz, S., Serraiocco, A., Vito, R., Wiesman, Z., Bianchi, G., 2005b. Regional variation in shea butter lipid and triterpene composition in four African countries. J. Agric. Food Chem. 53, 7473–7479.

Diarrassouba, N., Dougba, D.N., Yao Saraka, D.M., 2021. Estimation of shea trees (Vitellaria paradoxa C.F. Gaertn.) fruit production by assessing the correlation between yield parameters and dendrometric features in northern of Côte d'ivoire. J. Exp. Biol. Agric. Sci. 9, 745–758. https://doi.org/10.18006/2021.9(6).745.758

Diarrassouba, N., Fofana, I.J., Issali, A.E., Bup, N.D., Sangare, A., 2009. Typology of shea trees (Vitellaria paradoxa) using qualitative morphological traits in Côte d'Ivoire.

Djekota, C., Diouf, D., Sane, S., Noba, K., 2014. Morphological characterization of shea tree (Vitellaria paradoxa subsp. paradoxa) populations in the region of Mandoul in Chad. Int. J. Biodivers. Conserv. 6, 184–193. https://doi.org/10.5897/IJBC2013.0662

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat. Methods 16, 299–302. https://doi.org/10.1038/s41592-019-0344-8

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc. Natl. Acad. Sci. 112, 12580–12585. https://doi.org/10.1073/pnas.1509788112

Edem, D.O., 2002. Palm oil: Biochemical, physiological, nutritional, hematological and toxicological aspects: A review. Plant Foods Hum. Nutr. 57, 319–341. https://doi.org/10.1023/A:1021828132707

Egbule, P.O., Okonta, E.C., 2024. How climate change induced land conflicts and food insecurity in Africa: A case of herdsmen-farmers crisis in Nigeria: Climate Change, Land conflicts, Food insecurity, Herdsmen, Farmers, Land policies. Afr. J. Land Policy Geospatial Sci. 7, 251–265.

Elankumaran, M., Singh, C., Saravanan, K., Devi, L.A., 2026. Integrated Breeding and Biotechnological Strategies for Climate-Resilient Crop Development. Crop Breed. Genet. Genomics 8. https://doi.org/10.20900/cbgg20260006

Elsafy, M., Badawi, W., Ibrahim, A., Baillo, E.H., Assar, A.H.A., Brhane, H., Mahmood, U., Bajgain, P., Abdelhalim, T., Rahmatov, M., 2026. Candidate genomic regions underlying capsule shattering in sesame revealed by multi-model GWAS and field-based phenotyping. Theor. Appl. Genet. 139, 42. https://doi.org/10.1007/s00122-026-05151-7

Enyew, M., Feyissa, T., Carlsson, A.S., Tesfaye, K., Hammenhag, C., Seyoum, A., Geleta, M., 2022. Genome-wide analyses using multi-locus models revealed marker-trait associations for major agronomic traits in Sorghum bicolor. Front. Plant Sci. 13, 999692. https://doi.org/10.3389/fpls.2022.999692

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14, 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131, 479–491. https://doi.org/10.1093/genetics/131.2.479

Eyog Matig, O., Gaoué, O.G., Dossou, B., éditeurs, 2002. Programme de ressources génétiques forestières en Afrique au sud du Sahara: Réseau « Espèces Ligneuses Alimentaires ». Compte rendu de la première réunion du Réseau tenue du 11 au 13 décembre 2000 au CNSF Ouagadougou, Burkina Faso. Institut International des Ressources Phytogénétiques. ISBN 92-9043-552-6 IPGRI Regional Office for sub-Saharan Africa, IPGRI Regional Office for sub-Saharan Africa c/o ICRAF, P.O. Box 30677, Nairobi, Kenya. ed.

Fan, H., Xiao, Y., Yang, Y., Xia, W., Mason, A.S., Xia, Z., Qiao, F., Zhao, S., Tang, H., 2013. RNA-Seq Analysis of Cocos nucifera: Transcriptome Sequencing and De Novo Assembly for Subsequent Functional Genomics Approaches. PLOS ONE 8, e59997. https://doi.org/10.1371/journal.pone.0059997

Fang, C., Luo, J., 2019. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. Plant J. 97, 91–100. https://doi.org/10.1111/tpj.14097

FAO, 2018. The State of Agricultural Commodity Markets 2018. Agricultural trade, climate change and food security. Rome.

FAO, 2017. The impact of the shea nut industry on women's empowerment in Burkina Faso: A multi-dimensional study focusing on the Central, Central-West and Hauts-Bassins regions, Social Protection and Forestry Working Paper 3. Food and Agriculture Organization of the United Nations, Rome.

FAO, GSA, 2020. Shea value chain as a key pro-poor carbon-fixing engine in West Africa. Accra.

FAO, I., IFAD, UNICEF, WFP, WHO, 2022. The State of Food Security and Nutrition in the World 2022. Repurposing food and agricultural policies to make healthy diets more affordable. Rome, FAO. FAO ;, Rome.

FAOSTAT, 2024. FAOSTAT Database: Crops and livestock products. https://www.fao.org/faostat/en/#data/QCL [WWW Document]. URL https://www.fao.org/faostat/en/#data/QCL (accessed 9.7.24).

Fernández-Moya, V., Martínez-Force, E., Garcés, R., 2002. Temperature effect on a high stearic acid sunflower mutant. Phytochemistry 59, 33–37. https://doi.org/10.1016/S0031-9422(01)00406-X

Figueira, A.C., Luccas, V., 2022. Physicochemical characterization of national and commercial cocoa butter used in Brazil to make chocolate. Braz. J. Food Technol. 25, e2022033. https://doi.org/10.1590/1981-6723.03322

Fiume, M.M., Bergfeld, W.F., Belsito, D.V., Hill, R.A., Klaassen, C.D., Liebler, D.C., MarksJr, J.G., Shank, R.C., Slaga, T.J., Snyder, P.W., Andersen, F.A., Heldreth, B., 2018. Safety Assessment of Tocopherols and Tocotrienols as Used in Cosmetics. Int. J. Toxicol. 37, 61S-94S. https://doi.org/10.1177/1091581818794455

Fiume, Z.M., 2002. Final report on the safety assessment of Tocopherol, Tocopheryl Acetate, Tocopheryl Linoleate, Tocopheryl Linoleate/Oleate, Tocopheryl Nicotinate, Tocopheryl Succinate, Dioleyl Tocopheryl Methylsilanol, Potassium Ascorbyl Tocopheryl Phosphate, and Tocophersolan. Int. J. Toxicol. 21 Suppl 3, 51–116. https://doi.org/10.1080/10915810290169819

Fontaine, C., Lovett, P.N., Sanou, H., Maley, J., Bouvet, J.-M., 2004. Genetic diversity of the shea tree (Vitellaria paradoxa C.F. Gaertn), detected by RAPD and chloroplast microsatellite markers. Heredity 93, 639–648. https://doi.org/10.1038/sj.hdy.6800591

Forte Neran, R.A., Walbeogo, E., Dao, Y., Ravn, T.K., Kristensen, B., Christensen, J.H., Nielsen, N.J., 2025. Impact of West African treatment practices on phenolic profiles in shea kernels. Food Chem. 487, 144832. https://doi.org/10.1016/j.foodchem.2025.144832

Fortune Business Insights, 2025. Shea Butter Market Size, Share, and COVID-19 Impact Analysis, By Type (Raw and Unrefined, and Refined), BY Nature (Organic and Conventional), By Application (Food Processing, Pharmaceuticals, and Personal Care and Cosmetics), and Regional Forecast, 2026-2034 (No. FBI108282), Food & Beverages/Shea Butter Market. Food & Beverages.

Frank, D., 2024. Assessment of Harvesting Time and Its Influence on Yield Production in Agricultural Crops. Agrotechnology 13, 1–2. https://doi.org/10.35248/2168-9881.24.13.359

Friso, G., van Wijk, K.J., 2015. Posttranslational Protein Modifications in Plant Metabolism1. Plant Physiol. 169, 1469–1487. https://doi.org/10.1104/pp.15.01378

Gallusci, P., Dai, Z., Génard, M., Gauffretau, A., Leblanc-Fournier, N., Richard-Molard, C., Vile, D., Brunel-Muguet, S., 2017. Epigenetics for Plant Improvement: Current Knowledge and Modeling Avenues. Trends Plant Sci. 22, 610–623. https://doi.org/10.1016/j.tplants.2017.04.009

Ganglo, C.J., Maître, H.-F., 2003. Etat de la recherche forestière au Bénin : bilan et perspectives. In : Actes du XIIe Congrès forestier mondial : La forêt, source de vie, Québec, Canada. Bibliothèque nationale du Québec, p. 10p.

García-Cañas, V., Cifuentes, A., Simó, C., 2014. Applications of advanced omics technologies: from genes to metabolites, Wilson and Wilson's comprehensive analytical chemistry. Elsevier, Amsterdam.

Garti, H., Agbemafle, R., Mahunu, G.K., 2019. Physicochemical properties and fatty acid composition of Shea butter from Tamale, Northern Ghana. UDS Int. J. Dev. 6, 35–40. https://doi.org/10.47740/389.UDSIJD6i

Gbemavo, D.S.J.C., Dassou, A., Gbemavo, M., Ouinsavi, C., 2022. Occurrence and ecological habitat effect on Vitellaria paradoxa (C. F. Gaertn.) parasitism: implication for pest management and plant conservation. Heliyon 8, e10492. https://doi.org/10.1016/j.heliyon.2022.e10492

Gill, B.S., Kumar, S., Navgeet, 2016. Triterpenes in cancer: significance and their influence. Mol. Biol. Rep. 43, 881–896. https://doi.org/10.1007/s11033-016-4032-9

Global Shea Alliance, 2022. Annual Report: Growing the Future.

Gnanglè, P.C., 2005. Parcs à karité (Vitellaria paradoxa) (Gaertn. C. F.) (Sapotaceae) au Bénin: Importance socio-culturelle, caractérisations morphologique, structurale et régénération naturelle. Mémoire de DEA, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi.

Górnaś, P., Baškirovs, G., Siger, A., 2022. Free and Esterified Tocopherols, Tocotrienols and Other Extractable and Non-Extractable Tocochromanol-Related Molecules: Compendium of Knowledge, Future Perspectives and Recommendations for Chromatographic Techniques, Tools, and Approaches Used for Tocochromanol Determination. Molecules 27, 6560. https://doi.org/10.3390/molecules27196560

Granby, K., Toma, P., Christensen, P., Walbeogo, E., Dao, Y., Ravn, T.K., Christensen, J.H., Nielsen, N.J., 2025. Quantitative results of tocopherols, sterols, triterpene esters, and lipid degradation products in shea kernels using native or quantifier ion intensity corrected analogue standards. Talanta Open 12, 100570. https://doi.org/10.1016/j.talo.2025.100570

Grand View Research, 2024. Shea Butter Market Size, Share & Trend Analysis Report By Product (Raw/Unrefined, Refined), By Application (Cosmetics & Personal Care, Food, Others), By Region, And Segment Forecasts, 2025 - 2030 (Market Analysis Report No. GVR-3-68038-490-1). Grand View Research.

Grand View Research, 2021. Palm Oil Market Size, Share & Trends Analysis Report By Nature (Organic, Conventional), By Product (Crude Palm Oil, RBD Palm Oil), By End-use (Food & Beverage, Pharmaceuticals), By Region, And Segment Forecasts, 2023 - 2030 (No. 978-1-68038-462–8). (Market Analysis Report No. No. 978-1-68038-462–8).

Gupta, C., Salgotra, R.K., 2022. Epigenetics and its role in effecting agronomical traits. Front. Plant Sci. 13. https://doi.org/10.3389/fpls.2022.925688

Gupta, E., 2020. β-Sitosterol: Predominant Phytosterol of Therapeutic Potential, in: Mishra, P., Mishra, R.R., Adetunji, C.O. (Eds.), Innovations in Food Technology: Current Perspectives and Future Goals. Springer, Singapore, pp. 465–477. https://doi.org/10.1007/978-981-15-6121-4_32

Gwali, S., Nakabonge, G., Okullo, J.B.L., Eilu, G., Nyeko, P., Vuzi, P., 2012. Morphological variation among shea tree (Vitellaria paradoxa subsp. nilotica) 'ethnovarieties' in Uganda. Genet. Resour. Crop Evol. 59, 1883–1898. https://doi.org/10.1007/s10722-012-9905-8

Gwali, S., Vaillant, A., Nakabonge, G., Okullo, J.B.L., Eilu, G., Muchugi, A., Bouvet, J.-M., 2015. Genetic diversity in shea tree (Vitellaria paradoxa subspecies nilotica) ethno-varieties in Uganda assessed with microsatellite markers. For. Trees Livelihoods 24, 163–175. https://doi.org/10.1080/14728028.2014.956808

Hale, I., Ma, X., Melo, A.T.O., Padi, F.K., Hendre, P.S., Kingan, S.B., Sullivan, S.T., Chen, S., Boffa, J.-M., Muchugi, A., Danquah, A., Barnor, M.T., Jamnadass, R., Van de Peer, Y., Van Deynze, A., 2021. Genomic Resources to Guide Improvement of the Shea Tree. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.720670

Hall, J.B., Aebischer, D.P., Tomlinson, H.F., Osei-Amaning, E.O., Hindle, J.R., 1996. Vitellaria paradoxa: a monograph. School of Agricultural and Forest Sciences Publication No. 8. Bangor, UK, University of Wales. 105 pp. Sch. Agric. For. Sci. Publ. Bang. UK Univ. Wales No. 8., 105 pp.

He, M., Qin, C.-X., Wang, X., Ding, N.-Z., 2020. Plant Unsaturated Fatty Acids: Biosynthesis and Regulation. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.00390

Herrán, A., Estioko, L., Becker, D., Rodriguez, M.J.B., Rohde, W., Ritter, E., 2000. Linkage mapping and QTL analysis in coconut (Cocos nucifera L.). Theor. Appl. Genet. 101, 292–300. https://doi.org/10.1007/s001220051482

Heuckeroth, S., Damiani, T., Smirnov, A., Mokshyna, O., Brungs, C., Korf, A., Smith, J.D., Stincone, P., Dreolin, N., Nothias, L.-F., Hyötyläinen, T., Orešič, M., Karst, U., Dorrestein, P.C., Petras, D., Du, X., van der Hooft, J.J.J., Schmid, R., Pluskal, T., 2024. Reproducible mass spectrometry data processing and compound annotation in MZmine 3. Nat. Protoc. 19, 2597–2641. https://doi.org/10.1038/s41596-024-00996-y

Hien, B., Kaboré, P., 2025. République du Bénin Programme d'options stratégiques pour le pays.

Hoffmann, M.A., Nothias, L.-F., Ludwig, M., Fleischauer, M., Gentry, E.C., Witting, M., Dorrestein, P.C., Dührkop, K., Böcker, S., 2022. High-confidence structural annotation of metabolites absent from spectral libraries. Nat. Biotechnol. 40, 411–421. https://doi.org/10.1038/s41587-021-01045-9

Honfo, F.G., Akissoe, N., Linnemann, A.R., Soumanou, M., Van Boekel, M.A., 2014. Nutritional composition of shea products and chemical properties of shea butter: a review. Crit. Rev. Food Sci. Nutr. 54, 673–686.

Honfo, G.F., Linnemann, A.R., Guo, M., Akissoe, N., Soumanou, M.M., van BOEKEL, M.A., 2017. Influence of roasting of shea kernels on their fat content and some quality characteristics of shea butter. Int. J. Food Stud. 6, 66–80. https://doi.org/10.5296/jfs.v6i1.10881

Hong, M.J., Ko, C.S., Kim, D.Y., Hong, M.J., Ko, C.S., Kim, D.Y., 2025. Genome-Wide Association Study of Agricultural and Biochemical Traits in Radiation-Induced Colored Wheat. Agronomy 15. https://doi.org/10.3390/agronomy15081933

Huang, D., Wu, W., Abrams, S.R., Cutler, A.J., 2008. The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. J. Exp. Bot. 59, 2991–3007. https://doi.org/10.1093/jxb/ern155

Huang, J., Liu, X., Lan, Q., Lai, X., Luo, Z., Yang, G., 2016. Proteomic profile of coconuts. Eur. Food Res. Technol. 242, 449–455. https://doi.org/10.1007/s00217-015-2556-1

Huang, M., Liu, X., Zhou, Y., Summers, R.M., Zhang, Z., 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience 8, giy154. https://doi.org/10.1093/gigascience/giy154

Jahurul, M.H.A., Zaidul, I.S.M., Norulaini, N.A.N., Sahena, F., Jinap, S., Azmir, J., Sharif, K.M., Omar, A.K.M., 2013. Cocoa butter fats and possibilities of substitution in food products concerning cocoa varieties, alternative sources, extraction methods, composition, and characteristics. J. Food Eng., SI: Extraction and Encapsulation 117, 467–476. https://doi.org/10.1016/j.jfoodeng.2012.09.024

Jalani, B.S., Kushairi, A., Cheah, S.C., 2003. Production Systems and Agronomy | Oil Palm and Coconut. ISBN 978-0-12-227050-5, in: Thomas, B. (Ed.), Encyclopedia of Applied Plant Sciences. Elsevier, Oxford, pp. 960–969. https://doi.org/10.1016/B0-12-227050-9/00206-4

Jarvis, D.I., Hodgkin, T., Brown, A.H.D., Tuxill, J.D., Noriega, I.L., Smale, M., Sthapit, B., 2016. Crop Genetic Diversity in the Field and on the Farm: Principles and Applications in Research Practices. Yale University Press.

Jatto, M.I., Bello, L.L., Jakusko, B.B., 2012. Effect of Growth Hormones on the Germination of Shea-Nut (Vitellaria paradoxa C. F. Gaertn) after Pre-Treatment. Glob. J. Agric. Sci. 11, 99–103. https://doi.org/10.4314/gjass.v11i2.5

Jayasekhar, S., Chandran, K.P., 2021. World Economic Importance, in: Rajesh, M.K., Ramesh, S.V., Perera, L., Kole, C. (Eds.), The Coconut Genome. Springer International Publishing, Cham, pp. 1–12. https://doi.org/10.1007/978-3-030-76649-8_1

Jepsen, T., Stopponi, G., Jørgensen, N.O.G., 2024. Shea tree (Vitellaria paradoxa C.F. Gaertn.) agroforestry systems in Northern Ghana: population structure, management of trees and impact of below canopy microclimate. Agrofor. Syst. 98, 1493–1506. https://doi.org/10.1007/s10457-024-01019-1

Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94. https://doi.org/10.1186/1471-2156-11-94

Kaeuffer, R., Réale, D., Coltman, D.W., Pontier, D., 2007. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity 99, 374–380.

Kamal-Eldin, A., Appelqvist, L.-Å., 1996. The chemistry and antioxidant properties of tocopherols and tocotrienols. Lipids 31, 671–701. https://doi.org/10.1007/BF02522884

Kamvar, Z.N., Tabima, J.F., Grünwald, N.J., 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2, e281. https://doi.org/10.7717/peerj.281

Karaca, E., Güzel Kara, S., 2024. Nutritional Benefits of Shea Products and Medicinal Properties of Shea Butter. Hacet. Univ. J. Fac. Pharm. 44, 46–61. https://doi.org/10.52794/hujpharm.1328634

Kargiotidou, A., Deli, D., Galanopoulou, D., Tsaftaris, A., Farmaki, T., 2008. Low temperature and light regulate delta 12 fatty acid desaturases (FAD2) at a transcriptional level in cotton (Gossypium hirsutum). J. Exp. Bot. 59, 2043–2056. https://doi.org/10.1093/jxb/ern065

Karp, A., 1999. The use of polymorphic microsatellites for assessing genetic diversity in coconut, in: Oropeza, C., Verdeil, J.L., Ashburner, G.R., Cardeña, R., Santamaría, J.M. (Eds.), Current Advances in Coconut Biotechnology. Springer Netherlands, Dordrecht, pp. 121–129. https://doi.org/10.1007/978-94-015-9283-3_9

Katz, E., Li, J., Jaegle, B., Ashkenazy, H., Abrahams, S., Bagaza, C., Holden, S., Pires, C., Angelovici, R., Kliebenstein, D., 2021. Genetic variation, environment and demography intersect to shape Arabidopsis defense metabolite variation across Europe. PubMed.

Keen, M.A., Hassan, I., 2016. Vitamin E in dermatology. Indian Dermatol. Online J. 7, 311–315. https://doi.org/10.4103/2229-5178.185494

Kelly, A.B., Hardy, O., Bouvet, J., 2004. Temporal and spatial genetic structure in Vitellaria paradoxa (shea tree) in an agroforestry system in southern Mali. Mol. Ecol. 13, 1231–1240. https://doi.org/10.1111/j.1365-294X.2004.02144.x

Kelly, B.A., Poudyal, M., Bouvet, J.-M., 2019. Impact of land use and land use history on fruits production of Vitellaria paradoxa (Shea tree) according to agroclimatic zones in Mali (West Africa). Curr. Bot. 01–07. https://doi.org/10.25081/cb.2019.v10.3835

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K., Jaccoud, D., Hopper, C., Aschenbrenner-Kilian, M., Evers, M., Peng, K., Cayla, C., Hok, P., Uszynski, G., 2012. Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms, in: Pompanon, F., Bonin, A. (Eds.), Data Production and Analysis in Population Genomics: Methods and Protocols. Humana Press, Totowa, NJ, pp. 67–89. https://doi.org/10.1007/978-1-61779-870-2_5

Kim, E., Hwang, K., Lee, J., Han, S.Y., Kim, E.-M., Park, J., Cho, J.Y., 2018. Skin Protective Effect of Epigallocatechin Gallate. Int. J. Mol. Sci. 19, 173. https://doi.org/10.3390/ijms19010173

Koffi, C.K., Djoudi, H., Gautier, D., 2017. Landscape diversity and associated coping strategies during food shortage periods: evidence from the Sudano-Sahelian region of Burkina Faso. Reg. Environ. Change 17, 1369–1380. https://doi.org/10.1007/s10113-016-0945-z

Korte, A., Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9, 29. https://doi.org/10.1186/1746-4811-9-29

Kotadiya, R., 2025. Enhancing Pharmaceutical Analysis with Analytical Quality by Design in UHPLC: A Review of Methodological Innovations (2014–2025). Crit. Rev. Anal. Chem. 0, 1–21. https://doi.org/10.1080/10408347.2025.2516607

Kouyaté, A.M., Van Damme, P., Diawara, H., 2006. Évaluation de la production en fruits de Detarium microcarpum Guill. & Perr. au Mali. Fruits 61, 267–272. https://doi.org/10.1051/fruits:2006024

Kumar, R., Das, S.P., Choudhury, B.U., Kumar, A., Prakash, N.R., Verma, R., Chakraborti, M., Devi, A.G., Bhattacharjee, B., Das, R., Das, Bapi, Devi, H.L., Das, Biswajit, Rawat, S., Mishra, V.K., 2024. Advances in genomic tools for plant breeding: harnessing DNA molecular markers, genomic selection, and genome editing. Biol. Res. 57, 80. https://doi.org/10.1186/s40659-024-00562-6

Kumar, S., Singh, A.K., Mohapatra, T., 2017. Epigenetics : History, present status and future perspective. INDIAN J. Genet. PLANT Breed. 77, 445–463.

Kuusaana, E.D., 2022. Customary land governance dynamics and its implications for shea tenure and ecology in selected peri-urban communities in Ghana. Front. Sustain. Food Syst. 6. https://doi.org/10.3389/fsufs.2022.1033523

Lamien, N., Ouédraogo, S.J., Diallo, O.B., Guinko, S., 2004. Productivité fruitière du karité (Vitellaria paradoxa Gaertn. CF, Sapotaceae) dans les parcs agroforestiers traditionnels au Burkina Faso. Fruits 59, 423–429.

Larson, D.A., Vargas, O.M., Vicentini, A., Dick, C.W., 2021. Admixture may be extensive among hyperdominant Amazon rainforest tree species. New Phytol. 232, 2520–2534. https://doi.org/10.1111/nph.17675

Lassen, K.M., Nielsen, L.R., Lompo, D., Dupont, Y.L., Kjær, E.D., 2018. Honey bees are essential for pollination of Vitellaria paradoxa subsp. paradoxa (Sapotaceae) in Burkina Faso. Agrofor. Syst. 92, 23–34. https://doi.org/10.1007/s10457-016-0007-z

Lei, X., Xiao, Y., Xia, W., Mason, A.S., Yang, Y., Ma, Z., Peng, M., 2014. RNA-Seq Analysis of Oil Palm under Cold Stress Reveals a Different C-Repeat Binding Factor (CBF) Mediated Gene Expression Pattern in Elaeis guineensis Compared to Other Species. PLOS ONE 9, e114482. https://doi.org/10.1371/journal.pone.0114482

Li, N., Xu, C., Li-Beisson, Y., Philippar, K., 2016. Fatty Acid and Lipid Transport in Plant Cells. Trends Plant Sci. 21, 145–158. https://doi.org/10.1016/j.tplants.2015.10.011

Li, Y., Li, Z., Xu, C., Wang, Q., 2025. WRKYs as regulatory hubs of secondary metabolic networks: Diverse inducers and distinct responses. Plant Commun. 6, 101438. https://doi.org/10.1016/j.xplc.2025.101438

Lin, J.-Y., Selim, M.A., Shea, C.R., Grichnik, J.M., Omar, M.M., Monteiro-Riviere, N.A., Pinnell, S.R., 2003. UV photoprotection by combination topical antioxidants vitamin C and vitamin E. J. Am. Acad. Dermatol. 48, 866–874. https://doi.org/10.1067/mjd.2003.425

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genet. 12, e1005767. https://doi.org/10.1371/journal.pgen.1005767

Liu, Y., Lv, M., Wang, Y., Wei, J., Chen, D., 2025. Analytical Strategies for Tocopherols in Vegetable Oils: Advances in Extraction and Detection. Pharmaceuticals 18, 1137. https://doi.org/10.3390/ph18081137

Liveri, E., Passa, K., Papasotiropoulos, V., Liveri, E., Passa, K., Papasotiropoulos, V., 2024. The Contribution of Genetic and Genomic Tools in Diversity Conservation: The Case of Endemic Plants of Greece. J. Zool. Bot. Gard. 5, 276–293.

Lou, Y., Schwender, J., Shanklin, J., 2014. FAD2 and FAD3 desaturases form heterodimers that facilitate metabolic channeling in vivo. J. Biol. Chem. 289, 17996–18007. https://doi.org/10.1074/jbc.M114.572883

Lovett, P.N., 2004. The Shea Butter Value Chain: Production, Transformation, and Marketing in West Africa (No. Technical Report No. 2). USAID West Africa Trade Hub.

Lovett, P.N., Haq, N., 2013. Progress in developing in vitro systems for shea tree (Vitellaria paradoxa C.F. Gaertn.) propagation. For. Trees Livelihoods 22, 60–69. https://doi.org/10.1080/14728028.2013.765092

Lovett, P.N., Haq, N., 2000. Diversity of the Sheanut tree (Vitellaria paradoxa C.F. Gaertn.) in Ghana. Genet. Resour. Crop Evol. 47, 293–304. https://doi.org/10.1023/A:1008710331325

Lunn, J.E., 2007. Compartmentation in plant metabolism. J. Exp. Bot. 58, 35–47. https://doi.org/10.1093/jxb/erl134

Luo, J., 2015. Metabolite-based genome-wide association studies in plants. Curr. Opin. Plant Biol. 24, 31–38. https://doi.org/10.1016/j.pbi.2015.01.006

Ma, Z., Ren, J., Liu, Q., Li, J., Zhao, H., Tibesigwa, D.G., Matola, S.H., Gulfam, T., Yang, J., Wang, F., 2025. Integrating Traditional Breeding and Modern Biotechnology for Advanced Forest Tree Improvement. Int. J. Mol. Sci. 26, 8591. https://doi.org/10.3390/ijms26178591

Maanikuu, P.M.I., Peker, K., 2017. Medicinal and nutritional benefits from the shea tree (Vitellaria Paradoxa). J. Biol. Agric. Healthc. 7, 51–57.

Maranz, S., Wiesman, Z., 2004a. Influence of Climate on the Tocopherol Content of Shea Butter. J. Agric. Food Chem. 52, 2934–2937. https://doi.org/10.1021/jf035194r

Maranz, S., Wiesman, Z., 2004b. Influence of Climate on the Tocopherol Content of Shea Butter. J. Agric. Food Chem. 52, 2934–2937. https://doi.org/10.1021/jf035194r

Maranz, S., Wiesman, Z., Bisgaard, J., Bianchi, G., 2004. Germplasm resources of Vitellaria paradoxa based on variations in fat composition across the species distribution range. Agrofor. Syst. 60, 71–76. https://doi.org/10.1023/B:AGFO.0000009406.19593.90

Maranz, S., Wiesman, Z., Garti, N., 2003a. Phenolic Constituents of Shea (Vitellaria paradoxa) Kernels. J. Agric. Food Chem. 51, 6268–6273. https://doi.org/10.1021/jf034687t

Maranz, S., Wiesman, Z., Garti, N., 2003b. Phenolic Constituents of Shea (Vitellaria paradoxa) Kernels. J. Agric. Food Chem. 51, 6268–6273. https://doi.org/10.1021/jf034687t

Masyita, A., Mustika Sari, R., Dwi Astuti, A., Yasir, B., Rahma Rumata, N., Emran, T.B., Nainu, F., Simal-Gandara, J., 2022. Terpenes and terpenoids as main bioactive compounds of essential oils, their roles in human health and potential application as natural food preservatives. Food Chem. X 13, 100217. https://doi.org/10.1016/j.fochx.2022.100217

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J., Ebana, K., Yano, M., Saito, K., 2015. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J. 81, 13–23. https://doi.org/10.1111/tpj.12681

Mayer, N.A., Ueno, B., Antunes, L.E.C., Nava, G., Roth, F.M., 2021. Agronomic performance of 'BRS Kampai'peach on 15 clonal rootstocks and own-rooted trees in Pelotas-RS, Brazil. Rev. Bras. Frutic. 43, e-115. https://doi.org/10.1590/0100-29452021115

Meckel, L., Egli, D.B., Phillips, R.E., Radcliffe, D., Leggett, J.E., 1984. Effect of Moisture Stress on Seed Growth in Soybeans. Agron. J. 76, 647–650. https://doi.org/10.2134/agronj1984.00021962007600040033x

Merinosy, M.F., Achigan-Dako, E.G., Gnanglè, P.C., Kassa, E., Boffa, J.-M., 2022. Morphotype Classification Criteria and Influence of Sociocultural Factors on Perceived Shea Tree (Vitellaria paradoxa C.F. Gaertn) Natural Variation across Parklands in Benin. Plants Basel Switz. 11, 299. https://doi.org/10.3390/plants11030299

Méteignier, L.-V., Nützmann, H.-W., Papon, N., Osbourn, A., Courdavault, V., 2023. Emerging mechanistic insights into the regulation of specialized metabolism in plants. Nat. Plants 9, 22–30. https://doi.org/10.1038/s41477-022-01288-7

Metelmann, H.R., Brandner, J., Schumann, H., Bross, F., Hoffmann, M., Podmelle, F., 2012. Accelerating the aesthetic benefit of wound healing by triterpene. J. Cranio-Maxillofac. Surg. 40, e150–e154. https://doi.org/10.1016/j.jcms.2011.07.020

Miège, C., Maréchal, É., 1999. 1,2-sn-Diacylglycerol in plant cells: Product, substrate and regulator. Plant Physiol. Biochem. 37, 795–808. https://doi.org/10.1016/S0981-9428(99)00118-7

Misra, A., Surolia, N., Surolia, A., 2009. Catalysis and mechanism of malonyl transferase activity in type II fatty acid biosynthesis acyl carrier proteins. Mol. Biosyst. 5, 651–659. https://doi.org/10.1039/b820420a

Moharram, H., Ray, J., Ozbas, S., Juliani, H., Simon, J., 2006. Shea Butter: Chemistry, Quality, and New Market Potentials, in: Herbs: Challenges in Chemistry and Biology, ACS Symposium Series. American Chemical Society, pp. 326–340. https://doi.org/10.1021/bk-2006-0925.ch025

Monteiro e Silva, S.A., Calixto, G.M.F., Cajado, J., De Carvalho, P.C.A., Rodero, C.F., Chorilli, M., Leonardi, G.R., 2017. Gallic Acid-Loaded Gel Formulation Combats Skin Oxidative Stress: Development, Characterization and Ex Vivo Biological Assays. Polymers 9, 391. https://doi.org/10.3390/polym9090391

Montoya, C., Cochard, B., Flori, A., Cros, D., Lopes, R., Cuellar, T., Espeout, S., Syaputra, I., Villeneuve, P., Pina, M., Ritter, E., Leroy, T., Billotte, N., 2014. Genetic Architecture of Palm Oil Fatty Acid Composition in Cultivated Oil Palm (Elaeis guineensis Jacq.) Compared to Its Wild Relative E. oleifera (H.B.K) Cortés. PLoS ONE 9, e95412. https://doi.org/10.1371/journal.pone.0095412

Moore, S., 2008. The role of Vitellaria paradoxa in poverty reduction and food security in the Upper East region of Ghana. Earth Env. 3, 209–245.

Moreno-Pérez, A.J., Santos-Pereira, J.M., Martins-Noguerol, R., DeAndrés-Gil, C., Troncoso-Ponce, M.A., Venegas-Calerón, M., Sánchez, R., Garcés, R., Salas, J.J., Tena, J.J., Martínez-Force, E., 2021. Genome-Wide Mapping of Histone H3 Lysine 4 Trimethylation (H3K4me3) and Its Involvement in Fatty Acid Biosynthesis in Sunflower Developing Seeds. Plants Basel Switz. 10, 706. https://doi.org/10.3390/plants10040706

Morrison, W.R., Smith, L.M., 1964. Preparation of fatty acid methyl esters and dimethylacetals from lipids with boron fluoride–methanol. J. Lipid Res. 5, 600–608. https://doi.org/10.1016/S0022-2275(20)40190-7

Moudio, R., 2013. Shea butter nourishes opportunities for African women. Afr. Renew. Afr. Renew. Econ. Dev. URL https://africarenewal.un.org/en/magazine/shea-butter-nourishes-opportunities-african-women (accessed 9.10.25).

Mrudulakumari Vasudevan, U., Lee, E.Y., 2020. Flavonoids, terpenoids, and polyketide antibiotics: Role of glycosylation and biocatalytic tactics in engineering glycosylation. Biotechnol. Adv. 41, 107550. https://doi.org/10.1016/j.biotechadv.2020.107550

Naake, T., Zhu, F., Alseekh, S., Scossa, F., Perez de Souza, L., Borghi, M., Brotman, Y., Mori, T., Nakabayashi, R., Tohge, T., Fernie, A.R., 2024. Genome-wide association studies identify loci controlling specialized seed metabolites in Arabidopsis. Plant Physiol. 194, 1705–1721. https://doi.org/10.1093/plphys/kiad511

Naangmenyele, Z., Banye, E.Z., Bekoe, E.M.O., Zakaria, I., Amuah, E.E.Y., 2023. Fundamentals of shea butter production; input–output analyses and profit maximization in Northern Ghana. Res. Glob. 6, 100113. https://doi.org/10.1016/j.resglo.2023.100113

Nahm, H.S., Juliani, H.R., Simon, J.E., 2013. Quality characteristics of shea butter, Vitellaria paradoxa, in: African Natural Plant Products Volume II: Discoveries and Challenges in Chem, Health, and Nutrition. American Chemical Society, pp. 167–184. https://doi.org/10.1021/bk-2013-1127.ch012

Nasare, L.I., Kwapong, P.K., Doke, D.A., 2019. Insect pollinator dependence of shea (Vitellaria paradoxa C.F. Gaertn.) in the Guinea Savanna zone of Ghana. Ecol. Process. 8, 48. https://doi.org/10.1186/s13717-019-0202-8

Nielsen, N.J., Kronik, O.M., Forte Neran, R.A., Christensen, J.H., Ravn, T.K., de Villiers, A., 2025. Data-independent profiling of phenolic constituents in shea using comprehensive two-dimensional liquid chromatography (RPLC × HILIC) hyphenated to cyclic ion mobility-quadrupole-time-of-flight mass spectrometry. J. Chromatogr. A 1756, 466092. https://doi.org/10.1016/j.chroma.2025.466092

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–451. https://doi.org/10.1038/nrg2986

Nielsen, S.S., Carpenter, C., 2017. Fat Content Determination, in: Food Analysis Laboratory Manual, Food Science Text Series. Springer International Publishing, Cham, pp. 121–129. https://doi.org/10.1007/978-3-319-44127-6_12

Nikiema, A., Umali, B.E., 2007. Vitellaria paradoxa C.F.Gaertn. [Internet] Record from PROTA4U. van der Vossen, H.A.M. & Mkamilo, G.S. (Editors). PROTA (Plant Resources of Tropical Africa / Ressources végétales de l'Afrique tropicale)., Wageningen, Netherlands.

Niral, V., Samsudeen, K., Sudha, R., Ranjini, T.N., 2019. Genetic resource management and improved varieties of coconut. Indian Coconut Journal. ICAR-CPCRI, Kasargod.

Nogia, P., Pati, P.K., 2021. Frontiers | Plant Secondary Metabolite Transporters: Diversity, Functionality, and Their Modulation. https://doi.org/10.3389/fpls.2021.758202

Norazlina, M.R., Jahurul, M.H.A., Hasmadi, M., Mansoor, A.H., Norliza, J., Patricia, M., George, M.R.R., Noorakmar, A.W., Lee, J.S., Fan, H.Y., 2021. Trends in blending vegetable fats and oils for cocoa butter alternative application: A review. TRENDS FOOD Sci. Technol. 116, 102–114. https://doi.org/10.1016/j.tifs.2021.07.016

Odebiyi, J.A., Bada, S.O., Awodoyin, R., Oni, P.I., Omoloye, A.A., 2004. Population structure of Vitelaria paradoxa Gaertn. CF. and Parkia biglobosa (Jacq.) Benth.in the Agroforestry Parklands of Nigerian humid Savanna. West Afr. J. Appl. Ecol. 5. https://doi.org/10.4314/wajae.v5i1.45597

Odoi, J.B., Adjei, E.A., Hendre, P., Nantongo, J.S., Ozimati, A.A., Badji, A., Nakabonge, G., Edema, R., Gwali, S., Odong, T.L., 2023. Genetic diversity and population structure among Ugandan shea tree (Vitellaria paradoxa subsp. nilotica) accessions based on DarTSeq markers. Crop Sci. 63, 2297–2309. https://doi.org/10.1002/csc2.21013

Odoi, J.B., Okia, C.A., Gwali, S., Odong, T.L., Agaba, H., Okullo, J.B.L., 2022. Kernel morphometric characteristics and oil content among Shea tree genotypes in Uganda. Afr. Crop Sci. J. 30, 547–561. https://doi.org/10.4314/acsj.v30i4.12

OECD, 2010. The Economic Impact of Export Restrictions on Raw Materials, OECD Publishing. http://dx.doi.org/10.1787/9789264096448-en

Ofoegbu-Chibuzo, N.E., Chukwu, U.J., Okoye, I.P., 2022. Physicochemical Analysis and Fatty Acid Content of Chemical and Traditional Extracts of Shea Kernel (Vitellaria paradoxa) from Kwara State Nigeria. Open Access Libr. J. 9. https://doi.org/10.4236/oalib.1108295

Okullo, J.B.L., 2004. Vitellaria paradoxa in Uganda: population structures and reproductive characteristics.

Okullo, J.B.L., Omujal, F., Agea, J.G., Vuzi, P.C., Namutebi, A., Okello, J.B.A., Nyanzi, S.A., 2010. Physico-Chemical characteristics of shea butter (Vitellaria paradoxa C.F. Gaertn.) oil from the Shea district of Uganda. Afr. J. Food Agric. Nutr. Dev. 10. https://doi.org/10.4314/ajfand.v10i1.51484

O'Leary, S.J., Puritz, J.B., Willis, S.C., Hollenbeck, C.M., Portnoy, D.S., 2018. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Mol. Ecol. 27, 3193–3206. https://doi.org/10.1111/mec.14792

Olivares-Tenorio, M.-L., Cortes-Prieto, C.-M., Londoño-Univio, N.-C., Rojas-Díaz, D.-A., Quintanilla-Carvajal, M.-X., Tibaquira-Perez, L.-N., Ricaurte-Puentes, L.-Y., Rada-Bula, A.-I., Romero, H.-M., Garcia-Nuñez, J.-A., 2024. Bioactive compounds in palm oil: A comprehensive review of recent advances in physicochemical characteristics, health-promoting properties and technologies for extraction, concentration, fractionation, encapsulation and functional food applications. J. Food Compos. Anal. 132, 106306. https://doi.org/10.1016/j.jfca.2024.106306

Opoku – Mensah, S., 2023a. An assessment of the structure of shea global value chain in Ghana and implication for policy development. Heliyon 9, e20219. https://doi.org/10.1016/j.heliyon.2023.e20219

Opoku – Mensah, S., 2023b. An assessment of the structure of shea global value chain in Ghana and implication for policy development. Heliyon 9, e20219. https://doi.org/10.1016/j.heliyon.2023.e20219

Opoku-Mensah, S., Mahama Al-Hassan, R., Awo, M., 2024. Relationship, governance, and upgrading opportunities in the Shea value Chain: a descriptive analysis from Northern Ghana. For. Trees Livelihoods 33, 158–178. https://doi.org/10.1080/14728028.2024.2314126

Ortega, E., 2021. Pollination, fertilization and fruit production in fruit tree species, in: I International Symposium on Reproductive Biology of Fruit Tree Species 1342. pp. 173–180.

Oussou, K.F., Guclu, G., Sevindik, O., Starowicz, M., Kelebek, H., Selli, S., 2022. Comparative Elucidation of Aroma, Key Odorants, and Fatty Acid Profiles of Ivorian Shea Butter Prepared by Three Different Extraction Methods. Separations 9, 245. https://doi.org/10.3390/separations9090245

Ovbude, S.T., Sharmeen, S., Kyei, I., Olupathage, H., Jones, J., Bell, R.J., Powers, R., Hage, D.S., 2024. Applications of chromatographic methods in metabolomics: A review. J. Chromatogr. B 1239, 124124. https://doi.org/10.1016/j.jchromb.2024.124124

PAG, 2016. Programme d'Actions du Gouvernement (PAG 2016 – 2021).

Palermo, N., Buffagni, V., Vurro, F., Impollonia, G., Pignone, D., Janni, M., Nguyen, H.T., Dembech, E., Marmiroli, N., 2024. From landraces to haplotypes, exploiting a genomic and phenomic approach to identify heat tolerant genotypes within durum wheat landraces. Environ. Exp. Bot. 228, 105986. https://doi.org/10.1016/j.envexpbot.2024.105986

PalmElit, 2017. PalmElit, 2017. Catalogue : Semences de Palmier à huile.

Papin, V., Bosc, A., Sanchez, L., Bouffier, L., 2024. Integrating environmental gradients into breeding: application of genomic reactions norms in a perennial species. Heredity 133, 160–172. https://doi.org/10.1038/s41437-024-00702-4

Parvini, F., Sicardo, M.D., Hosseini-Mazinani, M., Martínez-Rivas, J.M., Hernández, M.L., 2016. Transcriptional Analysis of Stearoyl-Acyl Carrier Protein Desaturase Genes from Olive (Olea europaea) in Relation to the Oleic Acid Content of the Virgin Olive Oil. J. Agric. Food Chem. 64, 7770–7781. https://doi.org/10.1021/acs.jafc.6b02963

Peñaloza-Ramírez, J.M., González-Rodríguez, A., Mendoza-Cuenca, L., Caron, H., Kremer, A., Oyama, K., 2010. Interspecific gene flow in a multispecies oak hybrid zone in the Sierra Tarahumara of Mexico. Ann. Bot. 105, 389–399.

Perera, L., Perera, S.A.C.N., Bandaranayake, C.K., Harries, H.C., 2010. Coconut, in: Vollmann, J., Rajcan, I. (Eds.), Oil Crops, Handbook of Plant Breeding. Springer, New York, NY, pp. 369–396. https://doi.org/10.1007/978-0-387-77594-4_12

Pidkowich, M.S., Nguyen, H.T., Heilmann, I., Ischebeck, T., Shanklin, J., 2007. Modulating seed β-ketoacyl-acyl carrier protein synthase II level converts the composition of a temperate seed oil to that of a palm-like tropical oil. Proc. Natl. Acad. Sci. 104, 4742–4747. https://doi.org/10.1073/pnas.0611141104

Pienaah, C.K.A., Baruah, B., Luginaah, I., 2024. Women's perspectives on addressing the decline of Shea trees in Northern Ghana. GeoJournal 89, 9. https://doi.org/10.1007/s10708-024-11023-8

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., Zandi, P.P., 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum. Genomics 8, 14. https://doi.org/10.1186/1479-7364-8-14

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., Lareu, M.V., 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software. Front. Genet. 4, 98.

Pouliot, M., 2012. Contribution of "Women's Gold" to West African Livelihoods: The Case of Shea (Vitellaria paradoxa) in Burkina Faso. Econ. Bot. 66, 237–248. https://doi.org/10.1007/s12231-012-9203-6

Prateep-Na-Thalang, N., Tongyoo, P., Phumichai, C., Duangjit, J., 2024. Comparing different statistical models for association mapping and genomic prediction of fruit quality traits in tomato. Sci. Hortic. 327, 112838. https://doi.org/10.1016/j.scienta.2023.112838

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155, 945–959. https://doi.org/10.1093/genetics/155.2.945

Pritchard, J.K., Wen, W., Falush, D., 2003. Documentation for STRUCTURE software: Version 2.

Quazi, S., Golani, T., Capuzzo, A.M., Quazi, S., Golani, T., Capuzzo, A.M., 2021. Germplasm Conservation, in: Endangered Plants. IntechOpen. https://doi.org/10.5772/intechopen.96184

Ramawat, K.G., Mérillon, J.-M., Ahuja, M.R., 2014. Tree Biotechnology. CRC Press.

Ramesh, S.V., Sudha, R., Niral, V., Rajesh, M.K., 2022. Enhancing Genetic Gain in Coconut: Conventional, Molecular, and Genomics-Based Breeding Approaches, in: Gosal, S.S., Wani, S.H. (Eds.), Accelerated Plant Breeding, Volume 4: Oil Crops. Springer International Publishing, Cham, pp. 313–357. https://doi.org/10.1007/978-3-030-81107-5_10

Ramli, U.S., Othman, A., Tahir, N.I.M., Lau, B.Y.C., Shahwan, S., Hassan, H., Nurazah, Z., Dzulkafli, S.B., Rozali, N.L., Ishak, N.A., Sambanthamurthi, R., Parveez, G.K.A., Kushairi, A., 2020. Omics—A Potential Tool for Oil Palm Improvement and Productivity, in: Ithnin, M., Kushairi, A. (Eds.), The Oil Palm Genome. Springer International Publishing, Cham, pp. 141–157. https://doi.org/10.1007/978-3-030-22549-0_10

Rani, R., Kaushik, D., Kumar, A., Rasane, P., Oz, E., Kumar, V., Patra, K., Oz, F., Proestos, C., Kumar, M., 2025. Unveiling the approach of campesterol: Its application and circular bio-economy. CABI Rev. 20, 0049. https://doi.org/10.1079/cabireviews.2025.0049

Razzaq, A., Sadia, B., Raza, A., Hameed, M.K., Saleem, F., Razzaq, A., Sadia, B., Raza, A., Hameed, M.K., Saleem, F., 2019. Metabolomics: A Way Forward for Crop Improvement. Metabolites 9. https://doi.org/10.3390/metabo9120303

Rincón Barajas, J.A., 2024. Current Issues in Rural Development: Land Tenure, Land Use Change and Climate Change.

Rodriguez Cubillos, A.E., Tong, H., Alseekh, S., de Abreu e Lima, F., Yu, J., Fernie, A.R., Nikoloski, Z., Laitinen, R.A.E., 2018. Inheritance patterns in metabolism and growth in diallel crosses of Arabidopsis thaliana from a single growth habitat. Heredity 120, 463–473. https://doi.org/10.1038/s41437-017-0030-5

Rodriguez-Negrette, A.C., Huck-Iriart, C., Herrera, M.L., 2019. Physical Chemical Properties of Shea/Cocoa Butter Blends and their Potential for Chocolate Manufacture. J. Am. Oil Chem. Soc. 96, 239–248. https://doi.org/10.1002/aocs.12189

Ruan, C.-J., Qin, P., Teixeira da Silva, J.A., 2011. Relationship between reproductive assurance and mixed mating in perennial Kosteletzkya virginica. South Afr. J. Bot. 77, 280–291. https://doi.org/10.1016/j.sajb.2010.08.012

Saini, U., Sharma, A., Mittal, V., 2025. Gallic Acid: A Potent Antioxidant and Anti-inflammatory Agent in Modern Cosmeceuticals. Recent Adv. Drug Deliv. Formul. 19. https://doi.org/10.2174/0126673878341990250516063126

Salau, R.B., 2019. Comparative studies on the lipid profile of shea (Vitellaria paradoxa CF Gaertn.) fruit kernel and pulp. Journal of Applied Science, 19(5), 480 – 486.

Salgotra, R.K., Chauhan, B.S., Salgotra, R.K., Chauhan, B.S., 2023. Genetic Diversity, Conservation, and Utilization of Plant Genetic Resources. Genes 14.

Salie, M., Thelen, J., 2016. Regulation and structure of the heteromeric acetyl-CoA carboxylase. PubMed, Sep;1861(9 Pt B):1207-1213. Biochim Biophys Acta. 2016 Sep;1861(9 Pt B):1207-1213. https://doi.org/10.1016/j.bbalip.2016.04.004

Sallam, A., Hashem, M., Ahmed, A.A.M., Ismail, S.M., Baenziger, P.S., Börner, A., 2026. Genetic and phenotypic associations between peduncle characteristics and spike productivity in wheat under drought and normal conditions. Theor. Appl. Genet. 139, 34. https://doi.org/10.1007/s00122-025-05140-2

Sandwidi, A., Diallo, B.O., Lamien, N., Vinceti, B., Sanon, K., Coulibaly, P., Paré, S., Sawadogo, M., 2018. Participatory identification and characterisation of shea butter tree (Vitellaria paradoxa C.F. Gaertn.) ethnovarieties in Burkina Faso. Fruits 73, 141–152. https://doi.org/10.17660/th2018/73.3.1

Sanou, H., Kambou, S., Teklehaimanot, Z., Dembele, M., Yossi, H., Sina, S., Djingdia, L., Bouvet, J.-M., 2004. Vegetative propagation of Vitellaria paradoxa by grafting. Agrofor. Syst. 60, 93–99. https://doi.org/10.1023/B:AGFO.0000009408.03728.46

Sanou, H., Lovett, P.N., Bouvet, J. -M., 2005. Comparison of quantitative and molecular variation in agroforestry populations of the shea tree (Vitellaria paradoxa C.F. Gaertn) in Mali. Mol. Ecol. 14, 2601–2610. https://doi.org/10.1111/j.1365-294X.2005.02601.x

Saurabh Singh, K., Hooft, J.J.J. van der, Wees, S.C.M. van, H. Medema, M., 2022. Integrative omics approaches for biosynthetic pathway discovery in plants. https://doi.org/10.1039/D2NP00032F

Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T.S., Bushuiev, R., Murray, K.J., Hoffmann, N., Lu, M., Sarvepalli, A., Zhang, Z., Fleischauer, M., Dührkop, K., Wesner, M., Hoogstra, S.J., Rudt, E., Mokshyna, O., Brungs, C., Ponomarov, K., Mutabdžija, L., Damiani, T., Pudney, C.J., Earll, M., Helmer, P.O., Fallon, T.R., Schulze, T., Rivas-Ubach, A., Bilbao, A., Richter, H., Nothias, L.-F., Wang, M., Orešič, M., Weng, J.-K., Böcker, S.,

Jeibmann, A., Hayen, H., Karst, U., Dorrestein, P.C., Petras, D., Du, X., Pluskal, T., 2023. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat. Biotechnol. 41, 447–449. https://doi.org/10.1038/s41587-023-01690-2

Schober, P., Mascha, E.J., Vetter, T.R., 2021. Statistics From A (Agreement) to Z (z Score): A Guide to Interpreting Common Measures of Association, Agreement, Diagnostic Accuracy, Effect Size, Heterogeneity, and Reliability in Medical Research. Anesth. Analg. 133, 1633–1641. https://doi.org/10.1213/ANE.0000000000005773

Seghieri, J., 2019. Shea tree (Vitellaria paradoxa Gaertn.): from local constraints to multi-scale improvement of economic, agronomic and environmental performance in an endemic Sudanian multipurpose agroforestry species. Agrofor. Syst. 93, 2313–2330. https://doi.org/10.1007/s10457-019-00351-1

Segman, O., Wiesman, Z., Yarmolinsky, L., 2012. 17 - Methods and Technologies Related to Shea Butter Chemophysical Properties and to the Delivery of Bioactives in Chocolate and Related Products, in: Garti, N., Widlak, N.R. (Eds.), Cocoa Butter and Related Compounds. AOCS Press, pp. 417–441. https://doi.org/10.1016/B978-0-9830791-2-5.50020-7

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., Nordborg, M., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44, 825–830. https://doi.org/10.1038/ng.2314

Seiler, G.J., 1986. Analysis of the relationships of environmental factors with seed oil and fatty acid concentrations of wild annual sunflower. Field Crops Res. 15, 57–72. https://doi.org/10.1016/0378-4290(86)90101-2

Selvi, G.T., Suthakar, B., Surendrakumar, A., Kavitha, R., Masilamani, P., 2025. Tree fruit harvesting: Recent developments and future challenges for robotic harvesting. Plant Sci. Today 12. https://doi.org/10.14719/pst.8239

Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W., Wolf, J.B.W., 2017. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol. Evol. 8, 907–917. https://doi.org/10.1111/2041-210X.12700

Shakuri, F., Eghlima, G., Behboudi, H., Babashpour-Asl, M., 2025. Phytochemical variation, phenolic compounds and antioxidant activity of wild populations of Iranian oak. Sci. Rep. 15, 6534. https://doi.org/10.1038/s41598-025-90991-4

Shearman, J.R., Jantasuriyarat, C., Sangsrakru, D., Yoocha, T., Vannavichit, A., Tragoonrung, S., Tangphatsornruang, S., 2013. Transcriptome analysis of normal and mantled developing oil palm flower and fruit. Genomics 101, 306–312. https://doi.org/10.1016/j.ygeno.2013.02.012

Sheng, F., Hu, B., Jin, Q., Wang, J., Wu, C., Luo, Z., 2021. The Analysis of Phenolic Compounds in Walnut Husk and Pellicle by UPLC-Q-Orbitrap HRMS and HPLC. Molecules 26, 3013. https://doi.org/10.3390/molecules26103013

Shojaiefar, S., Sabzalian, M.R., Mirlohi, A., Tajdivand, A., 2021. Evidence for self-compatibility and variation for inbreeding depression within breeding populations of fennel (Foeniculum vulgare Mill.). J. Appl. Res. Med. Aromat. Plants 22, 100299. https://doi.org/10.1016/j.jarmap.2021.100299

Singh, S., 2019. Coconut Oil Market Research Report Information By Type (Virgin Coconut Oil, Refined Coconut Oil), By End-user (Industry, Food), And By Region (North America, Europe,

Asia-Pacific, And Rest Of The World) –Market Forecast Till 2032 (No. MRFR/F-B&N/5983-CR). Market Research Future.

Soh, A., Wong, C., Ho, Y., Choong, C., 2009. Oil palm. pp. 333–367. https://doi.org/10.1007/978-0-387-77594-4_11

Soitamo, A.J., Piippo, M., Allahverdiyeva, Y., Battchikova, N., Aro, E.-M., 2008. Light has a specific role in modulating Arabidopsis gene expression at low temperature. BMC Plant Biol. 8, 13. https://doi.org/10.1186/1471-2229-8-13

Soltis, N.E., Kliebenstein, D.J., 2015. Natural Variation of Plant Metabolism: Genetic Mechanisms, Interpretive Caveats, and Evolutionary and Mechanistic Insights1. Plant Physiol. 169, 1456–1468. https://doi.org/10.1104/pp.15.01108

Ståhl, U., Carlsson, A.S., Lenman, M., Dahlqvist, A., Huang, B., Banas, W., Banas, A., Stymne, S., 2004. Cloning and functional characterization of a phospholipid:diacylglycerol acyltransferase from Arabidopsis. Plant Physiol. 135, 1324–1335. https://doi.org/10.1104/pp.104.044354

Stout, J.C., Nombre, I., Bruijn, B. de, Delaney, A., Doke, D.A., Gyimah, T., Kamano, F., Kelly, R., Lovett, P., Marshall, E., Nasare, L.I., Nana, A., Roberts, J., Tankoano, P., Tayleur, C., Thomas, D., Vickery, J., Kwapong, P., 2018. Insect pollination improves yield of Shea (Vitellaria paradoxa subsp. paradoxa) in the agroforestry parklands of West Africa. J. Pollinat. Ecol. 22, 11–20. https://doi.org/10.26786/1920-7603(2018)two

Subhi, S., Tahir, N., Rasid, O., Ramli, U., Othman, A., Mat Yunus, A.M., Ghulam Kadir, A.P., Kushairi, A., 2017. Post-genomic technologies for the advancement of oil palm research. J. Oil Palm Res. 29, 469–486.

Sukumaran, A., Khanduri, V.P., Sharma, C.M., 2020. Pollinator-mediated self-pollination and reproductive assurance in an isolated tree of Magnolia grandiflora L. Ecol. Process. 9, 45. https://doi.org/10.1186/s13717-020-00254-5

Sukumaran, S., Rebetzke, G., Mackay, I., Bentley, A.R., Reynolds, M.P., 2022. Pre-breeding Strategies, in: Reynolds, M.P., Braun, H.-J. (Eds.), Wheat Improvement: Food Security in a Changing Climate. Springer International Publishing, Cham, pp. 451–469. https://doi.org/10.1007/978-3-030-90673-3_25

Sul, J.H., Martin, L.S., Eskin, E., 2018. Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet. 14, e1007309. https://doi.org/10.1371/journal.pgen.1007309

Takeda, S., Terazawa, S., Shimoda, H., Imokawa, G., 2021. β-Sitosterol 3-O-D-glucoside increases ceramide levels in the stratum corneum via the up-regulated expression of ceramide synthase-3 and glucosylceramide synthase in a reconstructed human epidermal keratinization model. PLoS ONE 16, e0248150. https://doi.org/10.1371/journal.pone.0248150

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A.E., Buckler, E.S., Zhang, Z., 2016. GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. Plant Genome 9, plantgenome2015.11.0120. https://doi.org/10.3835/plantgenome2015.11.0120

Tang, Yunshan, Zhang, G., Jiang, X., Shen, S., Guan, M., Tang, Yuhan, Sun, F., Hu, R., Chen, S., Zhao, H., Li, J., Lu, K., Yin, N., Qu, C., Tang, Yunshan, Zhang, G., Jiang, X., Shen, S., Guan, M., Tang, Yuhan, Sun, F., Hu, R., Chen, S., Zhao, H., Li, J., Lu, K., Yin, N., Qu, C., 2023. Genome-Wide Association Study of Glucosinolate Metabolites (mGWAS) in Brassica napus L. Plants 12.

Tano-Debrah, K., Ohta, Y., 1994. Enzyme-assisted aqueous extraction of fat from kernels of the shea tree, Butyrospermum parkii. J. Am. Oil Chem. Soc. 71, 979–983. https://doi.org/10.1007/BF02542265

Tchokponhoué, D.A., N'Danikou, S., Hale, I., Van Deynze, A., Achigan-Dako, E.G., 2017. Early fruiting in Synsepalum dulcificum (Schumach. & Thonn.) Daniell juveniles induced by water and inorganic nutrient management. F1000Research 6, 399.

Teh, H.F., Neoh, B.K., Ithnin, N., Daim, L.D.J., Ooi, T.E.K., Appleton, D.R., 2017. Review: Omics and Strategic Yield Improvement in Oil Crops. J. Am. Oil Chem. Soc. 94, 1225–1244. https://doi.org/10.1007/s11746-017-3033-8

Tehlivets, O., Scheuringer, K., Kohlwein, S.D., 2007. Fatty acid synthesis and elongation in yeast. Biochim. Biophys. Acta BBA - Mol. Cell Biol. Lipids, Regulation of Lipid Metabolism in Yeast 1771, 255–270. https://doi.org/10.1016/j.bbalip.2006.07.004

Tibiri, E.B., Boua, P.R., Soulama, I., Dubreuil-Tranchant, C., Tando, N., Tollenaere, C., Brugidou, C., Nanema, R.K., Tiendrebeogo, F., 2025. Challenges and opportunities of developing bioinformatics platforms in Africa: the case of BurkinaBioinfo at Joseph Ki-Zerbo University, Burkina Faso.

Ting, N.-C., Sherbina, K., Khoo, J.-S., Kamaruddin, K., Chan, P.-L., Chan, K.-L., Halim, M.A.A., Sritharan, K., Yaakub, Z., Mayes, S., Massawe, F., Chang, P.L., Nuzhdin, S.V., Sambanthamurthi, R., Singh, R., 2020. Expression of fatty acid and triacylglycerol synthesis genes in interspecific hybrids of oil palm. Sci. Rep. 10, 16296. https://doi.org/10.1038/s41598-020-73170-5

Todd, J., Post-Beittenmiller, D., Jaworski, J.G., 2002. KCS1 encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in Arabidopsis thaliana. https://doi.org/10.1046/j.1365-313X.1999.00352.x

Toma, P., Granby, K., Christensen, P., Walbeogo, E., Dao, Y., Ravn, T.K., Christensen, J.H., Nielsen, N.J., 2025. Identification of semi-volatile specialized metabolites from Vitellaria paradoxa (shea) kernels using GC-MS. J. Food Compos. Anal. 147, 108042. https://doi.org/10.1016/j.jfca.2025.108042

Transparency Research Market, 2019. Shea Butter Market (Grade-Grade A (Unrefined), Grade B (Refined), Grade C (Highly Refined); End Use - Cosmetics & Personal Care, Food Processing, Pharmaceuticals, Retail) - Global Industry Analysis, Size, Share,Growth, Trends and Forecast 2018 - 2028 (No. TMRGL23081).

Trautwein, E.A., Demonty, I., 2007. Phytosterols: natural compounds with established and emerging health benefits. Ol. Corps Gras Lipides 14, 259–266. https://doi.org/10.1051/ocl.2007.0145

Tsetsos, F., Drineas, P., Paschou, P., 2019. Genetics and Population Analysis, in: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 363–378. https://doi.org/10.1016/B978-0-12-809633-8.20114-3

Turreira-García, N., 2025. West African Shea Processors in a Changing Global Market: Effects of Market Integration on Traditional Knowledge Resilience. Econ. Bot. 79, 287–309. https://doi.org/10.1007/s12231-025-09646-0

Turreira-García, N., Nielsen, N.J., Pattison, D.I., Makowicz, E., Walbeogo, E., Dao, Y., Christensen, J.H., Pouliot, M., 2025. Fermenting shea nuts using the traditional pit method yields better

physicochemical properties with potential environmental benefits. Sci. Rep. 15, 9502. https://doi.org/10.1038/s41598-025-93921-6

Ugese, F.D., Baiyeri, P.K., Mbah, B.N., 2010. Agroecological variation in the fruits and nuts of shea butter tree (Vitellaria paradoxa C. F. Gaertn.) in Nigeria. Agrofor. Syst. 79, 201–211. https://doi.org/10.1007/s10457-009-9261-7

UNDP, 2022. Human Development Report 2021-22: Uncertain Times, Unsettled Lives: Shaping our Future in a Transforming World. New York., Human Development Reports. United Nations.

UNDP, 2019. Human Development Report 2019: Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. New York., Human Development Reports. United Nations, New York.

Vecka, M., Staňková, B., Kutová, S., Tomášová, P., Tvrzická, E., Žák, A., 2019. Comprehensive sterol and fatty acid analysis in nineteen nuts, seeds, and kernel. SN Appl. Sci. 1, 1531. https://doi.org/10.1007/s42452-019-1576-z

Wang, J., Zhang, Z., 2021. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics 19, 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, L., Lu, H., Jiang, Y., 2023. Natural Polyketides Act as Promising Antifungal Agents. Biomolecules 13, 1572. https://doi.org/10.3390/biom13111572

Wang, T., Xing, J., Liu, X., Liu, Z., Yao, Y., Hu, Z., Peng, H., Xin, M., Zhou, D.-X., Zhang, Y., Ni, Z., 2016. Histone acetyltransferase general control non-repressed protein 5 (GCN5) affects the fatty acid composition of Arabidopsis thaliana seeds by acetylating fatty acid desaturase3 (FAD3). Plant J. 88, 794–808. https://doi.org/10.1111/tpj.13300

Wang, Y., Htwe, Y.M., Li, J., Shi, P., Zhang, D., Zhao, Z., Ihase, L.O., 2019. Integrative omics analysis on phytohormones involved in oil palm seed germination. BMC Plant Biol. 19, 363. https://doi.org/10.1186/s12870-019-1970-0

Wei, Y., Ji, B., Siewers, V., Xu, D., Halkier, B.A., Nielsen, J., 2019. Identification of genes involved in shea butter biosynthesis from Vitellaria paradoxa fruits through transcriptomics and functional heterologous expression. Appl. Microbiol. Biotechnol. 103, 3727–3736. https://doi.org/10.1007/s00253-019-09720-3

Werner, C.R., Gaynor, R.C., Sargent, D.J., Lillo, A., Gorjanc, G., Hickey, J.M., 2023. Genomic selection strategies for clonally propagated crops. Theor. Appl. Genet. 136, 74. https://doi.org/10.1007/s00122-023-04300-6

Westgate, M.E., Schussler, J.R., Reicosky, D.C., Brenner, M.L., 1989. Effect of Water Deficits on Seed Development in Soybean : II. Conservation of Seed Growth Rate. Plant Physiol. 91, 980–985. https://doi.org/10.1104/pp.91.3.980

Wisuitiprot, W., Somsiri, A., Ingkaninan, K., Waranuch, N., 2011. In vitro human skin permeation and cutaneous metabolism of catechins from green tea extract and green tea extract–loaded chitosan microparticles. Int. J. Cosmet. Sci. 33, 572–579. https://doi.org/10.1111/j.1468-2494.2011.00673.x

Wong, Y.C., Teh, H.F., Mebus, K., Ooi, T.E.K., Kwong, Q.B., Koo, K.L., Ong, C.K., Mayes, S., Chew, F.T., Appleton, D.R., Kulaveerasingam, H., 2017. Differential gene expression at different stages of mesocarp development in high- and low-yielding oil palm. BMC Genomics 18, 470. https://doi.org/10.1186/s12864-017-3855-7

World Bank, 2020. The World Bank in Western and Central Africa.

Wu, Y., Zheng, X., Xu, X.-G., Li, Y.-H., Wang, B., Gao, X.-H., Chen, H.-D., Yatskayer, M., Oresajo, C., 2013. Protective effects of a topical antioxidant complex containing vitamins C and E and ferulic acid against ultraviolet irradiation-induced photodamage in Chinese women. J. Drugs Dermatol. JDD 12, 464–468.

Yang, D.J., Moh, S.H., Son, D.H., You, S., Kinyua, A.W., Ko, C.M., Song, M., Yeo, J., Choi, Y.-H., Kim, K.W., 2016. Gallic Acid Promotes Wound Healing in Normal and Hyperglucidic Conditions. Molecules 21, 899. https://doi.org/10.3390/molecules21070899

Yang, H.M., Zhang, X.Y., Wang, G.X., 2004. Relationships between stomatal character, photosynthetic character and seed chemical composition in grass pea at different water availabilities. J. Agric. Sci. 142, 675–681. https://doi.org/10.1017/S0021859605004831

Yao, L., Wu, X., Jiang, X., Shan, M., Zhang, Z., Li, Yiting, Yang, A., Li, Yu, Yang, C., 2023. Subcellular compartmentalization in the biosynthesis and engineering of plant natural products. Biotechnol. Adv. 69, 108258. https://doi.org/10.1016/j.biotechadv.2023.108258

Yao, S.D.M., Diarrassouba, N., Attikora, A.J.P., Fofana, I.J., Dago, D.N., Silue, S., 2020. Morphological diversity patterns among selected elite Shea trees (Vitellaria paradoxa CF Gaertn.) from Tchologo and Bagoué districts in Northern Côte d'Ivoire. Int. J. Genet. Mol. Biol. 12.

Yeboah, J., Lowor, S.T., Boateng, P.T., Amoah, F.M., 2009. Some Physiological and Environmental Factors Influencing the Vegetative Propagation of the Shea Tree (Vitellaria paradoxa Gaertn). Int. J. Trop. Agric. Food Syst. 3, 170–177.

Yeboah, J., Segbefia, M. a. D., Dadzie, A.M., Padi, F., Lowor, S.T., Agene, V.N., Owusu-Ansah, F., Owusu-Ansah, B., 2020. Cutting Propagation of Shea (Vitellaria paradoxa C.F. Gaertn) Tree Using Shoot Types and Application of Auxin. J. Agric. Sci. 12, p213. https://doi.org/10.5539/jas.v12n12p213

Yue, G.H., Ye, B.Q., Lee, M., 2021. Molecular approaches for improving oil palm for oil. Mol. Breed. New Strateg. Plant Improv. 41. https://doi.org/10.1007/s11032-021-01218-z

Zeng, Y., Xu, X., Jiang, J., Lin, S., Fan, Z., Meng, Y., Maimaiti, A., Wu, P., Ren, J., 2025. Genome-wide association analysis and genomic selection for leaf-related traits of maize. PLOS ONE 20, e0323140. https://doi.org/10.1371/journal.pone.0323140

Zhang, H., Yu, Y., Wang, M., Dang, P., Chen, C., 2023. Effect of Genotype-by-Environment Interaction on Oil and Oleic Fatty Acid Contents of Cultivated Peanuts. Horticulturae 9, 1272. https://doi.org/10.3390/horticulturae9121272

Zhang, P., Burton, J.W., Upchurch, R.G., Whittle, E., Shanklin, J., Dewey, R.E., 2008. Mutations in a Δ9–Stearoyl-ACP-Desaturase Gene Are Associated with Enhanced Stearic Acid Levels in Soybean Seeds. https://doi.org/10.2135/cropsci2008.02.0084

Zhao, H., Kosma, D.K., Lü, S., 2021. Functional Role of Long-Chain Acyl-CoA Synthetases in Plant Development and Stress Responses. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.640996

Zhao, X., Ren, D., Jin, R., Chen, W., Xu, L., Guo, D., Zhang, Q., Luo, Z., 2024. Development of UHPLC-Q-Exactive Orbitrap/MS Technique for Determination of Proanthocyanidins (PAs) Monomer Composition Content in Persimmon. Plants 13, 1440. https://doi.org/10.3390/plants13111440

Zhou, Q., Galindo-González, L., Hwang, S.-F., Strelkov, S.E., 2021. Application of genomics and transcriptomics to accelerate development of clubroot resistant canola. Can. J. Plant Pathol. 43, 189–208. https://doi.org/10.1080/07060661.2020.1794541

Zhou, R., Jiang, F., Niu, L., Song, X., Yu, L., Yang, Y., Wu, Z., 2022. Increase Crop Resilience to Heat Stress Using Omic Strategies. Front. Plant Sci. 13. https://doi.org/10.3389/fpls.2022.891861

Zhu, N., Sang, S., Huang, T.-C., Bai, N., Yang, C.S., Ho, C.-T., 2000. Antioxidant Chemistry of Green Tea Catechins: Oxidation Products of (-)-Epigallocatechin Gallate and (-)-Epigallocatechin with Peroxidase. J. Food Lipids 7, 275–282. https://doi.org/10.1111/j.1745-4522.2000.tb00178.x

Zhu, T., Provart, N.J., 2003. Transcriptional responses to low temperature and their regulation in Arabidopsis. Can. J. Bot. 81, 1168–1174. https://doi.org/10.1139/b03-115

Zuhar, L.M., Madihah, A.Z., Ahmad, S.A., Zainal, Z., Idris, A.S., Shaharuddin, N.A., 2021. Identification of Oil Palm's Consistently Upregulated Genes during Early Infections of Ganoderma boninense via RNA-Seq Technology and Real-Time Quantitative PCR. Plants Basel Switz. 10, 2026. https://doi.org/10.3390/plants10102026

Summary

Shea (*Vitellaria paradoxa* C.F. Gaertn.) is a key agroforestry tree species in sub-Saharan Africa, providing edible oil, cosmetic ingredients, and income to millions of rural households in West Africa. Despite its high socio-economic importance, shea remains a semi-domesticated species, and its genetic improvement has been limited by its long juvenile phase, complex reproductive biology, and the predominance of in situ management systems. Improving shea productivity and quality therefore requires a better understanding of the genetic, biochemical, and metabolomic diversity underlying traits of industrial relevance, as well as the development of breeding strategies adapted to perennial agroforestry species. The overall objective of this thesis was to characterize this diversity and to explore integrative approaches to support future shea improvement and sustainable valorization.

Chapter 1 provides a general introduction to the biological, ecological, and socio-economic context of shea. It reviews current knowledge on shea distribution, domestication status, and management in traditional parkland systems, and highlights the major constraints associated with breeding and genetic improvement in long-lived perennial species. The chapter emphasizes the need for integrative approaches combining socio-economic insights, genetic diversity analyses, and omics-based tools to guide shea improvement efforts.

In **Chapter 2**, an integrated conceptual framework for shea breeding is presented. This chapter synthesizes existing knowledge on socio-economic drivers, farmer-mediated management practices, genetic diversity, fatty acid biosynthesis, and recent advances in omics technologies. It outlines how these components can be combined to define breeding objectives, identify relevant traits, and design improvement strategies adapted to the biological and socio-cultural context of shea.

Chapter 3 focuses on the genetic structure of shea populations in Benin and on fatty acid-based selection of elite genotypes. Using single nucleotide polymorphism (SNP) markers, genetic diversity and population structure were analysed in a panel of 196 shea trees sampled across different agroecological zones. The results revealed moderate genetic diversity, with most variation occurring within populations and a low but significant differentiation among populations. Population structure analyses identified three genetic clusters. In parallel, fatty acid composition was characterized in a subset of accessions, allowing the identification of

genotypes with favourable quality profiles, although limited correspondence was observed between genetic structure and fatty acid variation.

In **Chapter 4**, metabolomic-guided diversity analysis was conducted to characterize chemical variation in shea kernels and to identify cosmetic-relevant biomarkers. Metabolomic profiling revealed pronounced variation in the accumulation of bioactive compounds, including triterpenes, flavan-3-ols, phytosterols, and phenolic acids. The analyses highlighted the existence of distinct chemotypic groups and identified metabolites of potential industrial interest, while showing weak geographic structuring of metabolomic variation.

Chapter 5 investigates the genetic basis of cosmetic bioactive metabolites through genome-wide association analyses. Using a subset of accessions for which both genomic and metabolomic data were available, contrasting genetic architectures were identified among metabolites, with some traits controlled by major-effect loci and others displaying polygenic regulation. Several candidate genes potentially involved in regulatory and metabolic pathways were identified, providing insights into the molecular mechanisms underlying variation in cosmetic-relevant compounds and their implications for breeding and conservation.

Finally, **Chapter 6** presents a general discussion that synthesizes the results obtained across all chapters. This chapter highlights the substantial genetic, biochemical, and metabolomic diversity observed in shea, as well as the limited congruence among these layers of variation. It discusses the influence of farmer-mediated management practices, environmental modulation, and complex genetic architectures on diversity patterns, and emphasizes the necessity of integrative approaches for understanding and exploiting shea diversity. The chapter concludes by identifying key research gaps and proposing future directions for shea improvement, including genomics-assisted selection, targeted clonal propagation of elite genotypes, and sustainable conservation strategies adapted to this long-lived agroforestry species.

Résumé

Le karité (*Vitellaria paradoxa* C.F. Gaertn.) est une espèce agroforestière clé en Afrique subsaharienne, fournissant une huile alimentaire, des ingrédients cosmétiques et des revenus à des millions de ménages ruraux. Malgré son importance socio-économique élevée, le karité demeure une espèce semi-domestiquée, et son amélioration génétique a été limitée par sa longue phase juvénile, sa biologie reproductive complexe et la prédominance de systèmes de gestion in situ. L'amélioration de la productivité et de la qualité du karité nécessite donc une meilleure compréhension de la diversité génétique, biochimique et métabolomique sous-jacente aux caractères d'intérêt industriel, ainsi que le développement de stratégies de sélection adaptées aux espèces agroforestières pérennes. L'objectif général de cette thèse était de caractériser cette diversité et d'explorer des approches intégratives pour soutenir l'amélioration future et la valorisation durable du karité.

Le chapitre 1 présente une introduction générale au contexte biologique, écologique et socio-économique du karité. Il passe en revue les connaissances actuelles sur sa distribution, son statut de domestication et sa gestion dans les systèmes traditionnels de parcs agroforestiers, et met en évidence les principales contraintes liées à la sélection et à l'amélioration génétique des espèces pérennes à longue durée de vie. Ce chapitre souligne la nécessité d'approches intégratives combinant des analyses socio-économiques, de la diversité génétique et des outils omiques pour orienter les efforts d'amélioration du karité.

Le chapitre 2 propose un cadre conceptuel intégré pour la sélection du karité. Ce chapitre synthétise les connaissances existantes sur les facteurs socio-économiques, les pratiques de gestion des producteurs, la diversité génétique, la biosynthèse des acides gras et les avancées récentes des technologies omiques. Il décrit comment ces composantes peuvent être combinées pour définir les objectifs de sélection, identifier les caractères pertinents et concevoir des stratégies d'amélioration adaptées au contexte biologique et socioculturel du karité.

Le chapitre 3 est consacré à l'étude de la structure génétique des populations de karité au Bénin et à la sélection de génotypes élites sur la base de la composition en acides gras. À l'aide de marqueurs de type polymorphisme nucléotidique simple (SNP), la diversité génétique et la structure des populations ont été analysées sur un ensemble de 196 arbres de karité échantillonnés dans différentes zones agroécologiques. Les résultats ont révélé une diversité génétique modérée, avec une majorité de la variation observée au sein des populations et une

différenciation faible mais significative entre populations. Les analyses de structure de population ont mis en évidence trois groupes génétiques. Parallèlement, la composition en acides gras a été caractérisée pour un sous-ensemble d'accessions, permettant d'identifier des génotypes présentant des profils de qualité favorables, bien qu'une correspondance limitée ait été observée entre la structure génétique et la variation des acides gras.

Le chapitre 4 présente une analyse de la diversité guidée par la métabolomique afin de caractériser la variation chimique des amandes de karité et d'identifier des biomarqueurs d'intérêt cosmétique. Le profilage métabolomique a révélé une forte variation dans l'accumulation de composés bioactifs, notamment les triterpènes, les flavan-3-ols, les phytostérols et les acides phénoliques. Les analyses ont mis en évidence l'existence de groupes chimiotypiques distincts et ont permis d'identifier des métabolites présentant un intérêt industriel potentiel, tout en montrant une faible structuration géographique de la variation métabolomique.

Le chapitre 5 examine les bases génétiques des métabolites bioactifs d'intérêt cosmétique à travers des analyses d'association à l'échelle du génome. En utilisant un sous-ensemble d'accessions pour lesquelles des données génomiques et métabolomiques étaient disponibles, des architectures génétiques contrastées ont été mises en évidence selon les métabolites, certains caractères étant contrôlés par des loci à effet majeur tandis que d'autres présentaient une régulation polygénique. Plusieurs gènes candidats potentiellement impliqués dans des voies de régulation et métaboliques ont été identifiés, apportant des éclairages sur les mécanismes moléculaires à l'origine de la variation des composés d'intérêt cosmétique et leurs implications pour la sélection et la conservation.

Enfin, le chapitre 6 présente une discussion générale qui synthétise l'ensemble des résultats obtenus. Ce chapitre met en évidence l'importante diversité génétique, biochimique et métabolomique du karité, ainsi que la faible concordance observée entre ces différents niveaux de variation. Il discute l'influence des pratiques de gestion paysannes, de la modulation environnementale et de la complexité des architectures génétiques sur les patrons de diversité, et souligne la nécessité d'approches intégratives pour comprendre et exploiter durablement la diversité du karité. Le chapitre se conclut par l'identification des principales lacunes de recherche et la proposition de perspectives pour l'amélioration future du karité, incluant la sélection assistée par la génomique, la propagation clonale ciblée de génotypes élites et des stratégies de conservation durable adaptées à cette espèce agroforestière pérenne à longue durée de vie.

Acknowledgements

As the African proverb reminds us, “One hand cannot clap”; this work is the result of the support and precious contributions of many people. It is therefore my opportunity to express my deepest gratitude to all those who have contributed, directly or indirectly, to the completion of this thesis.

I would like to extend my heartfelt thanks to my thesis supervisors, **Pr. Enoch Achigan-Dako, Pr. Alain Hehn**, and **Dr. Alexandre Olry**, for their guidance, scientific rigor, and unwavering support. A special note of gratitude to **Dr. Alexandre Olry** for his availability, technical guidance, and moral support throughout these four years of research. **Prof. Enoch and Prof. Alain**, thank you sincerely for your trust, your dedication, and your constant support. I can never thank you enough for believing in me, facilitating my stay at Wageningen University & Research through the DrEAM fellowship from the University of Lorraine, and for the constant care you showed me. I am also grateful for all administrative support that made my stay in France possible. Please accept my deepest appreciation, dear supervisors.

I sincerely thank **Prof. Guido Rychen,** director of ENSAIA, North-coordinator of PEA-BIOVALOR project and **Ms. Stecyna KIKI** for facilitating my enrollment at the University of Lorraine and supporting my stay in France as part of the BIOVALOR project.

Prof. Sévérine Piutti, thank you for welcoming me to the Laboratoire Agronomie et Environnement (LAE), for making my stay smooth, and for your invaluable support during the submission of my DrEAM fellowship application.

Prof. Elmira Arab-Tehrany, I cannot thank you enough for your guidance during my GC-MS analyses at your Laboratory of Biomolecule Engineering (LIBio), for your availability, and for the attention you gave to every committee meeting. Please accept my sincere gratitude.

Prof. Cyril Kahn and **Ms. Carol**, I warmly thank you for your technical assistance in acquiring metabolomics data at LIBio.

Prof. Eric Schranz, thank you deeply for welcoming me to the Biosystematic Group at WUR, allowing me to acquire bioinformatics skills in genomics and metabolomics. Your guidance was exceptional and unforgettable.

To **Wilma, Félicia, Cyril, Chloé, Eva, Qiong, and Pamela**, thank you for your kindness, warmth, and support during my stay at WUR. You were simply wonderful.

Mr. Jérémy Grosjean, thank you sincerely for your technical assistance with Orbitrap data acquisition and for your involvement in organizing social activities, especially the MPG.

A big thank you to **Ms. Aude Fauvet** for your technical support in metabolite extraction for UHPLC-HRMS analyses and for introducing me to laboratory biosafety principles.

Thanks to **Dr. Dèdéou** for mediating with SEQUART AFRICA in acquiring genomic data.

Dear **Matice, Albert, and Willbroad**, I can never thank you enough for your dedication and professionalism during field data collection in the north. Please accept my deepest gratitude.

My thanks also go to the members of my supervision team, **Dr Fernande Honfo** and **Dr Paul-Césaire Gnanglè**, for their constant support in protocol validation and data collection.

A very special tribute to the late **Mr. Claude Gallois**, a man with an angelic heart and an unwavering spirit of kindness. **Dear Claude**, wherever you are now, please receive my deepest gratitude and heartfelt appreciation for your countless forms of support and for your sincere, selfless friendship. Your generosity and humanity will forever remain in my heart. Rest in peace.

A very special thank you to **Dr. Florent Ducrocq**, my friend and office colleague, for his sincere and selfless friendship. Thank you for the joyful moments shared, our little jokes at the office, and your invaluable help during laboratory experiments. Thank you so much, “Flo.”

To my colleagues at LAE, **Omar, Marwa, Stéphane, Florent, Corentin, Valentin**, **Tania, Muqsita** and all those I could not mention, thank you for your warmth and support every day.

Sincere thanks to **Ms. Anne** and **Ms. Léocadie** for their assistance with financial procedures, acquiring materials and reagents, and supporting my participation in conferences.

To **Gérard, Wenceslas, and Donatien**, thank you deeply for your help with administrative and financial matters.

To my **parents, my brother, and my sisters**, you have been my foundation, giving me love, care, and stability in difficult times. Thank you for existing for me.

A special and deeply heartfelt thank you to **my wife** for her daily presence, love, and unwavering support. I will never forget it.

Finally, to all those who contributed, directly or indirectly, to the success of this work and whose names I have not mentioned, please accept my sincerest thanks.

List of publications

Peer-reviewed papers

Bodjrenou, G. J., Grosjean, J., Honfo, F. G., Hehn, A., Achigan-Dako, E. G., and Olry, A. (2026). Metabolomic-guided diversity analysis reveals chemotypic groups and cosmetic-relevant biomarkers in shea trees (*Vitellaria paradoxa* C.F. Gaertn.) for industrial applications. Ind. Crops Prod. 239, 122354. https://doi: 10.1016/j.indcrop.2025.122354

G.J. Bodjrenou, M. Gbètoho, F. Honfo, C.J.F. Kahn, M.E. Schranz, P.C. Gnanglè, A. Hehn, A. Olry, E.G. Achigan-Dako, Genetic structure of the shea tree (Vitellaria paradoxa C.F. Gaertn.) population and fatty acid-based selection of elite genotypes in Benin, Physiology and Molecular Biology of Plants (2026) 1–16. https://doi.org/10.1007/s12298-026-01734-2

Ongoing

Guillaume J. Bodjrenou, Matice Gbetoho, Félicien Amakpé, Fernande G. Honfo, Alain Hehn, Alexandre Olry, Enoch G. Achigan-Dako : Toward omics-based breeding strategies for the Shea tree (*Vitellaria paradoxa* C.F. Gaertn.) ***(****Accepted for publication in Current Plant Biology****)***

Guillaume J. Bodjrenou, Alexandre Olry, Dedeou A. Tchokpohoue, Fernande G. Honfo, Alain Hehn, Enoch G. Achigan-Dako: Multi-models genome-wide association study revealed candidate genomic regions linked with bioactive metabolites in shea tree (*Vitellaria paradoxa* C.F. Gaertn.) (***Pending submission****)*

Biography



Guillaume J. Bodjrenou is a Beninese researcher committed to the sustainable valorization of local plant resources through life sciences. Specialized in genetics, plant breeding, and agrobiological innovation, he has developed expertise at the interface of biotechnology, genomics, biochemistry, and multi-omics approaches.

Born on January 10, 1995, in Sèmè-Podji (Benin), he obtained his scientific baccalaureate in 2012 before enrolling at the University of Abomey-Calavi (UAC), where he earned a Bachelor's degree in Genetics, Biotechnology, and Biological Resources in 2015 from the Faculty of Science and Technology (FAST). In 2020, he completed a Master's degree in Biotechnology and Plant Breeding, further strengthening his orientation toward research applied to agriculture and the valorization of plant resources.

His scientific career has been enriched by several academic and professional experiences, notably as a Research Assistant at the GBioS Laboratory from late 2020, and later as Research Project Manager between 2021 and 2023. In this role, he coordinated initiatives funded by ARES and ENABEL, developing strong skills in project management, scientific coordination, and multi-stakeholder collaboration.

Awarded a PhD scholarship under the BIOVALOR project in 2022, he began a joint doctoral program between the University of Abomey-Calavi (Benin) and the University of Lorraine (France). His doctoral research focuses on the genetic improvement of shea tree (Vitellaria paradoxa C.F. Gaertn.), with the ambition of identifying high-value trees rich in specific fatty acids and bioactive compounds sought after by the chocolate, cosmetic, and dermocosmetic industries.

Throughout this PhD, he has developed advanced competencies in population genetics, bioinformatics, metabolomics, statistical analyses, multi-omics data integration, field experimentation, and scientific valorization. In 2024, a research stay at Wageningen University & Research (Netherlands), supported by a DREAM mobility grant, enabled him to further strengthen his expertise in complex biological data analysis and integrated omics approaches.

Author and co-author of several international scientific publications in high-impact journals, Guillaume J. Bodjrenou positions his research within the broader framework of African bioeconomy, based on local transformation, value creation, and the smart use of biodiversity. Through his work, he contributes to the development of innovative scientific solutions aimed at enhancing the competitiveness of the shea sector and promoting the sustainable use of natural resources for economic and social development.

Academic Statement

Details	Date/Duration
1) Specific Training	
➢ Molecular Biology Training and Open Labware Building workshop	1-5 April 2019
➢ Next Generation Genomics and Integrated Breeding for Crop Improvement on Genomics for food, health and Nutrition	14 May 2020
➢ Plant genetic resources and resilient seed systems for sustainable food security	29 Oct. – 21 Dec. 2021
➢ Mooc-Project management	25 sept. - 26 nov.2023
➢ Bioinformatic training (Metabolomic and genomic data analysis)	1st Nov 2024- 31 Jan. 2024
➢ Whole genome sequencing of bacterial genomes: Tools and Applications	09-13 June 2025
➢ Bioconductor training & Introduction to R for RNA-seq data analysis	17-21 nov. 2025

2) Scientific Exposure	
Oral presentation	
Talk: Genetic diversity and population structure of shea tree (*Vitellaria paradoxa*) in Benin| Colloquium, Biosystematic group, Wageningen University & Research (WUR), Netherlands	13 Jan. 2025
Talk: Targeted selection of shea trees (*Vitellaria paradoxa* CF Gaertn) rich in specific fatty acids and unsaponifiable matter for the chocolate and cosmetics industry in Benin, MétaSP, Bordeaux, France	Aug. 2024
Talk: Targeted selection of shea trees (*Vitellaria paradoxa* CF Gaertn) rich in specific fatty acids and unsaponifiable matter for the chocolate and cosmetics industry in Benin, Séminaire Scientifique Biovalor, University of Abomey-Calavi (UAC), Benin	Jan. 2024
Talk: Targeted selection of shea trees (*Vitellaria paradoxa* CF Gaertn) rich in specific fatty acids and unsaponifiable matter for the chocolate and cosmetics industry in Benin, les Doctoriales Multidisciplinaires, University of Abomey-Calavi (UAC), Benin	23 sept. 2023

3) Skills Development	
Scientific Writing	
Published: Metabolomic-guided diversity analysis reveals chemotypic groups and cosmetic-relevant biomarkers in shea trees (*Vitellaria paradoxa* C.F. Gaertn.) for industrial applications. Ind. Crops Prod. 239, 122354. doi: 10.1016/j.indcrop.2025.122354	Dec. 2025
Published: Genetic structure of the shea tree (*Vitellaria paradoxa* C.F. Gaertn.) population and fatty acid-based selection of elite genotypes in Benin (Accepted at Physiology and Molecular Biology of Plants)	Feb. 2026
Submitted: Toward omics-based breeding strategies for the Shea tree (*Vitellaria paradoxa* C.F. Gaertn.) (Under review)	Feb. 2026
Pending submission: Multi-models genome-wide association study revealed candidate genomic regions linked with bioactive metabolites in shea tree (*Vitellaria paradoxa* C.F. Gaertn.) (Pending submission)	ongoing
Annual Progress Presentations	
Réunion du comité de suivi (CSI), University of Lorraine (UL)	21 June 2023
CSI, UL & progress presentation, UAC Year 2	21 May 2024
CSI, UL & progress presentation, UAC Year 3	5 Oct. 2025

Titre : Sélection ciblée d'arbres à karité (*Vitellaria paradoxa* C.F. Gaertn.) riches en acides gras et en matières insaponifiables spécifiques pour l'industrie de la chocolaterie et de la cosmétique au Bénin

Résumé

Cette thèse s'inscrit dans le contexte de la valorisation durable du karité (*Vitellaria paradoxa* C.F. Gaertn.), une espèce oléagineuse pérenne d'Afrique de l'Ouest reconnue pour son importance économique, écologique et socioculturelle. Le beurre de karité constitue aujourd'hui une matière première stratégique pour les industries cosmétique, pharmaceutique et agroalimentaire, en raison de sa richesse en acides gras et en composés bioactifs. Malgré cette importance, l'espèce demeure faiblement domestiquée et peu améliorée génétiquement comparativement à d'autres cultures oléagineuses majeures telles que le palmier à huile et le cocotier. Cette situation limite fortement l'exploitation de son potentiel économique et industriel. Dans ce contexte, cette étude a été conduite afin de contribuer à l'amélioration génétique du karité au Bénin à travers une approche intégrative combinant analyses agromorphologiques, moléculaires, biochimiques et omiques.

Les principales questions de recherche abordées concernent : (i) l'ampleur et l'organisation spatiale de la diversité génétique et agromorphologique des populations de karité au Bénin ; (ii) les relations entre variabilité génétique et composition biochimique des amandes ; (iii) l'existence éventuelle de chimiotypes métaboliques distincts ; (iv) l'identification des régions génomiques associées aux composés bioactifs ; et (v) l'intérêt des approches omiques dans l'identification de génotypes élites adaptés aux exigences industrielles et écologiques.

Le travail adopte une approche multidisciplinaire combinant prospections de terrain, caractérisations phénotypiques, analyses moléculaires, métabolomiques et bioinformatiques. Les données collectées proviennent des principaux parcs à karité du Bénin couvrant différentes zones agroécologiques. Les analyses biochimiques ont porté sur la teneur en matière grasse, le profil en acides gras ainsi que les fractions insaponifiables riches en molécules bioactives. Les approches génomiques et statistiques multivariées ont permis d'étudier les relations entre diversité génétique et variabilité métabolique.

Le présent mémoire est structuré en six chapitres.

Le Chapitre 1 présente l'introduction générale de la thèse ainsi que le contexte scientifique et socio-économique de l'étude. Ce chapitre met en évidence l'importance croissante du karité dans l'économie mondiale des produits naturels et souligne les défis liés à sa valorisation durable. Une attention particulière est accordée à l'importance du beurre de karité comme substitut du beurre de cacao dans l'industrie chocolatière ainsi qu'à son utilisation dans les produits cosmétiques grâce à sa richesse en stérols, tocophérols, triterpènes et autres métabolites bioactifs. Le chapitre expose également les contraintes majeures auxquelles fait face l'espèce, notamment la longue phase juvénile, le vieillissement des peuplements, la forte variabilité de rendement et le manque de programmes structurés d'amélioration variétale. Les hypothèses de recherche, les objectifs généraux et spécifiques ainsi que la méthodologie globale y sont également détaillés. Ce chapitre sert ainsi de cadre conceptuel et méthodologique à l'ensemble du travail de recherche.

Le Chapitre 2 fournit une synthèse bibliographique approfondie sur les approches omiques appliquées à l'amélioration génétique du karité, en s'inspirant des progrès réalisés chez le palmier à huile et le cocotier. Ce chapitre met en évidence l'importance économique mondiale du karité dont le marché international connaît une croissance rapide portée par les secteurs cosmétique et agroalimentaire. Les analyses montrent que les pays ouest-africains, principaux producteurs de noix de karité, exportent encore majoritairement les amandes brutes avec une faible valorisation locale. Les limites structurelles liées à la transformation industrielle sont discutées, notamment les contraintes technologiques, les normes de qualité internationales et la faible capacité de raffinage. Le chapitre explore ensuite la composition biochimique du beurre de karité, particulièrement sa richesse en acides gras stéarique et oléique ainsi qu'en composés bioactifs tels que les tocophérols, phytostérols, composés phénoliques et triterpènes. Les variations observées entre populations de karité sont présentées comme une opportunité importante pour la sélection de génotypes à haute valeur industrielle. Les résultats de travaux antérieurs révèlent notamment l'existence de groupes métaboliques distincts caractérisés par des profils spécifiques de composés bioactifs.

Une large partie du chapitre est consacrée à l'analyse comparative des stratégies d'amélioration utilisées chez le palmier à huile et le cocotier. Les approches basées sur la sélection assistée par marqueurs, la génomique fonctionnelle, la transcriptomique et la métabolomique sont décrites comme des outils particulièrement efficaces pour accélérer le gain génétique. L'intérêt du développement de génomes de référence et des approches GWAS pour l'identification des

gènes d'intérêt est également discuté. Le chapitre souligne toutefois que le transfert direct de ces méthodes au karité nécessite des adaptations tenant compte de la biologie particulière de l'espèce, notamment sa longue durée juvénile et son faible niveau de domestication.

Le Chapitre 3 analyse la diversité agromorphologique, moléculaire et biochimique des populations de karité au Bénin. Les travaux réalisés montrent une importante variabilité phénotypique au sein des populations étudiées. Les caractères dendrométriques, les dimensions des fruits et des noix, ainsi que les paramètres liés au rendement présentent des différences significatives entre zones agroécologiques. Les résultats mettent en évidence des relations importantes entre certains traits morphologiques et la productivité des arbres. Par exemple, le diamètre du tronc et la taille de la couronne apparaissent positivement corrélés au rendement en fruits.

Les analyses moléculaires révèlent également un niveau élevé de diversité génétique au sein des populations de karité. Les marqueurs utilisés montrent une forte hétérozygotie et une structuration génétique modérée entre populations, indiquant l'existence d'un vaste réservoir génétique exploitable dans les programmes d'amélioration. Les échanges de flux géniques entre populations ainsi que l'absence historique de sélection intensive par les producteurs expliquent en partie cette diversité élevée. Les analyses biochimiques mettent en évidence des variations importantes dans la composition lipidique des amandes. Les teneurs en acides gras saturés et insaturés diffèrent selon les individus et les zones écologiques. Les proportions d'acides stéarique, oléique, palmitique et linoléique présentent une variabilité susceptible d'influencer directement la qualité industrielle du beurre. Les résultats suggèrent également l'influence des facteurs environnementaux tels que la pluviométrie, la température et les caractéristiques édaphiques sur la composition biochimique des amandes.

Le Chapitre 4 se focalise sur l'analyse métabolomique et l'identification de chimiotypes chez le karité. Une approche métabolomique guidée a permis de caractériser la diversité des composés bioactifs présents dans les amandes et le beurre de karité. Les analyses multivariées ont conduit à l'identification de plusieurs groupes métaboliques distincts caractérisés par des signatures chimiques spécifiques. Les résultats révèlent notamment des variations importantes dans les concentrations de tocophérols, phytostérols, composés phénoliques et triterpènes. Certains génotypes présentent des teneurs particulièrement élevées en α-tocophérol, β-sitostérol, stigmasterol, acide gallique et butyrospermol, molécules reconnues pour leurs propriétés antioxydantes, anti-inflammatoires et cosmétiques. D'autres métabolites, tels que

l’épigallocatéchine, la gallocatéchine et le parkeol, ont été identifiés comme biomarqueurs discriminants des différents chimiotypes.

Ce chapitre met également en évidence le potentiel des approches métabolomiques dans la sélection de génotypes élites à forte valeur ajoutée industrielle. Les chimiotypes identifiés pourraient servir de base à des programmes de sélection orientés vers des usages spécifiques, notamment dans les secteurs cosmétique, pharmaceutique et agroalimentaire. L’étude souligne aussi l’intérêt d’intégrer les profils métaboliques dans les stratégies modernes d’amélioration génétique du karité.

Le Chapitre 5 étudie les bases génétiques de la variabilité des composés bioactifs à travers des approches d’association pangénomique (GWAS). L’objectif principal de ce chapitre est d’identifier les régions génomiques impliquées dans la régulation de la biosynthèse des acides gras et des métabolites spécialisés. Les résultats obtenus permettent de mieux comprendre les mécanismes moléculaires responsables des différences observées entre individus. Le chapitre décrit en détail les principales voies biochimiques impliquées dans la synthèse des acides gras chez le karité. Les enzymes clés intervenant dans les étapes de condensation, de réduction, de désaturation et d’élongation des chaînes carbonées sont présentées. Plusieurs gènes impliqués dans la biosynthèse des acides gras, notamment ceux associés aux enzymes ACC, KAS, SAD, FAD2 et FAD3, sont discutés à partir des données transcriptomiques disponibles.

Les analyses GWAS ont permis d’identifier plusieurs régions génomiques potentiellement associées à la variation des composés bioactifs et des profils lipidiques. Les résultats suggèrent que certaines variations phénotypiques observées chez les populations de karité possèdent une base génétique importante susceptible d’être exploitée dans les programmes de sélection assistée par marqueurs. Ce chapitre démontre ainsi l’intérêt des approches génomiques modernes pour accélérer l’amélioration variétale du karité.

Enfin, le Chapitre 6 présente la discussion générale et les perspectives de recherche. Ce chapitre synthétise les principaux résultats obtenus et met en évidence les contributions scientifiques majeures de la thèse. L’étude confirme l’existence d’une forte diversité génétique, biochimique et métabolique au sein des populations de karité du Bénin. Les résultats démontrent également le potentiel considérable des approches intégratives combinant génomique, métabolomique et analyses phénotypiques pour l’identification de génotypes élites.

Les implications pratiques des résultats pour les programmes de domestication et d'amélioration génétique sont largement discutées. Le développement futur de variétés performantes nécessitera la mise en place de programmes structurés intégrant sélection conventionnelle et outils omiques avancés. Les perspectives concernent notamment le développement de banques de ressources génétiques, l'utilisation de la sélection génomique, l'identification de marqueurs moléculaires robustes et l'amélioration des techniques de multiplication végétative. Au-delà du karité, cette recherche propose un cadre méthodologique transférable à d'autres espèces ligneuses pérennes sous-utilisées en Afrique. Les résultats obtenus contribuent à une meilleure compréhension des mécanismes biologiques et génétiques associés à la qualité des produits dérivés du karité et ouvrent de nouvelles perspectives pour la valorisation durable de cette ressource stratégique dans un contexte de changement climatique et de développement des bioéconomies africaines.