Genetic diversity analysis of cacao (*Theobroma cacao* L.) germplasm in Samoa using microsatellite markers

A thesis submitted in partial fulfilment of the
requirements for the degree of

Master of Science

In

Plant Breeding

At Massey University, Palmerston North

New Zealand

Tianmiao Liu

2025

Abstract

Cacao (*Theobroma cacao* L.) is an important tropical crop with significant economic and ecological value. However, its genetic diversity is still insufficiently studied in Samoa. In this study, microsatellite markers were used to assess the genetic diversity of 183 cacao varieties collected from four plantations in Samoa. Six polymorphic SSR loci were used to identify 108 alleles revealing genetic variation in cacao among plantations (expected heterozygosity = 0.49). Both Structure and Neighbor-Net analyses showed that the SIU plantations were mainly Criollo and the SAT plantations were mainly Nanay, while the ALE and SAL plantations were mixtures of several varieties. Four different genetic groups were delineated by cluster analysis (UPGMA) and principal coordinate analysis (PCoA). It was finally determined that the cacao in the Samoa region consists of Amelonado, Trinitario, Nanay, Criollo and LAFI-7. The results of this study revealed the key role of microsatellite markers in analyzing cacao diversity, identified the major Samoa cacao types and provided feasible insights for Samoa molecular marker-assisted cacao breeding. It provides a direction for future improvement of cacao germplasm for biotic and abiotic resistance and the construction of modernized cacao plantations.

Acknowledgments

The successful completion of this study owes greatly to the dedicated guidance and support of my supervisor, Dr. Vaughan Symonds. I deeply appreciate your visionary direction in shaping the research framework and your pivotal advice during critical impasses. Your rigorous academic ethos, profound expertise, and relentless pursuit of scientific excellence have not only laid a robust foundation for this work but also profoundly shaped my understanding of scientific inquiry and its responsibilities.

I extend my heartfelt gratitude to Prashant Joshi, our laboratory manager, for his comprehensive support throughout this experimental research. From coordinating instrument access and ensuring the timely provision of critical reagents to optimizing complex experimental workflows, every phase of this study benefited from your expertise and dedication. Your efficient management of laboratory resources enabled the timely completion of data acquisition and validation. Special thanks for your patient mentorship in experimental techniques which significantly enhanced my independent research capabilities.

Lastly, I sincerely appreciate my family, colleagues, and friends for their unwavering support and encouragement throughout this research journey.

List of Abbreviations

%P	Percentage of polymorphic loci
AFLP	Amplified fragment length polymorphism
ALE	Aleipata Itupa I Luga area
EB	Early blight
EST	Expressed Sequence Tags
F_{ST}	Coefficient of population Differentiation
F_{IS}	Inbreeding coefficient of individuals within a population
F_{IT}	Total inbreeding coefficient
Ho	Heterozygosity Observed Heterozygosity
He	Expected Heterozygosity
HWE	Hardy-Weinberg Equilibrium
I	Information Index
N	Number of samples
Na	Number of Alleles
Ne	Effective number of alleles
Nm	Gene Flow
ORAC	Oxygen Radical Absorbance Capacity
Pa	Private alleles
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
PIC	Polymorphic information content
QTL	Quantitative Trait Loci
RFLP	Restriction fragment length polymorphism
RAPD	Random Amplified Polymorphic DNA
SAL	Salelologa
SAT	Satupaitea
SIU	Siumu
SNP	Single Nucleotide Polymorphism
SSR	Simple sequence repeat

Table of Contents

List of Figures

List of Tables

Chapter 1

Literature review

1. Introduction and Background

1.1 Origin and development of cacao

Theobroma cacao L., which belongs to the *Malvaceae* family, is a perennial tropical evergreen tree. The species typically thrives between the latitudes 20 degrees north to 20 degrees south of the equator (Almeida et al., 2007). The cultivation of this species for cocoa production is viable across a range of altitudes, extending from coastal areas reaching up to 1000 meters above sea level, depending on the latitude. Indigenous to South America, cacao has been widely dispersed and is now primarily cultivated in regions such as Ghana, Indonesia, Brazil, Mexico, and various Caribbean and Pacific islands, including Samoa (Motamayor et al., 2002). The origins of cacao use date back to around 400 AD with the Maya civilization believed to have been the first to cultivate it (Verna, 2013). Based on differences in morphology, genetics, and geographical origin, global cacao varieties are predominantly categorized into three main groups: Criollo, Forastero, and Trinitario (Barreto et al., 2018). In the contemporary era, the market for cacao has established a global presence across the continents of Europe, Asia, and the Americas, becoming one of the significant economic crops (Aikpokpodion., 2012).

1.1.1 Classification and morphological description of cacao trees

Cacao can be clearly distinguished in the ecosystem by morphological characteristics. The propagation of the cacao tree (*Theobroma cacao* L.) can be achieved through either seed propagation or cuttings. Under natural growth conditions, the cacao tree can reach heights of 20 to 25 meters; however, it is typically pruned to maintain a height of 3 to 5 meters for ease of management and harvesting in cultivated settings (Lachenaud et al., 1997; Almeida et al., 2007). The cacao tree possesses a dense canopy and broad leaves, which are ovate-lanceolate in shape with petioles and are glabrous on both

surfaces. The bark of the cacao tree is thick and dark gray-brown, while the young branches are brown. The flowers, which have five yellowish petals, are borne in cymose clusters on the trunk and branches and bloom throughout the year. In the cacao flower, the stamens are reduced to filamentous structures, and it has opposite stamens and petals. The ovary is obovate. The fruit is an ellipsoid drupe with a thick, fleshy rind, and individual fruit weights range from approximately 314 to 892 grams (Kuppers., 1953). The seeds are ovoid and enclosed in a white or red mucilage (Rodriguez-Campos et al., 2012). The cacao fruits mature twice a year, a characteristic that contributes to its production efficiency. Due to the presence of self-incompatibility in cacao, outcrossing strategies are predominantly employed (Zhang & Motilal., 2016). It thrives in warm, humid climates and rich, organic soils. The cacao tree's growth is stunted by poor drainage or harsh environmental conditions. Although the tree requires sunlight, young seedlings should be protected from prolonged direct exposure to prevent sunscald. Additionally, the cacao tree has a low tolerance for cold and frost, with an optimal annual temperature range of 20 to 30 degrees Celsius. Factors such as soil, temperature, and rainfall have a significant impact on the quality of the cacao fruit, which in turn influences the final flavor of the chocolate product (Ascrizzi et al., 2017; Apshara et al., 2013).

1.1.2 Cacao's beneficial effects, history and culture, and economic functions

Cacao (*Theobroma cacao* L.), as one of the world's major beverage crops, holds a significant position in the global economy of crops. Its seeds are rich in lipids, a variety of organic acids, trace elements, theobromine, vitamins, and numerous aromatic compounds. Through processes such as fermentation and roasting, cacao can be transformed into delicious beverages and chocolate (Aprotosoaie et al., 2016). These products are not only characterized by their rich taste and enticing aroma but are also nutritionally rich (Verna, 2013). The global market for cacao and its products is estimated to be worth approximately 200 billion US dollars per year on average (Zhang & Motilal, 2016).

The health benefits of cacao should not be overlooked, as it is believed to have potential benefits such as disease prevention, promotion of cardiovascular health, improvement of blood circulation, alleviation of headaches, and stress reduction (Verna, 2013). Notably, theobromine which extracted from cacao seeds, is a methylxanthine that positively modulates human mood and alertness. Compared to caffeine, theobromine is considered safer for human consumption (Martínez-Pinilla et al., 2015).

In terms of industrial applications, bio-based plastic made from a mixture of cacao bean husk powder and polyethylene has shown superior biodegradability than traditional plastic, offering new possibilities for sustainable development in the plastics industry (De Andrade Almeida et al., 2021). Additionally, cacao beans have agricultural uses, not only as a raw material for fertilizers but also as an alternative to corn in animal feed. The cacao fruit pulp, rich in sucrose and acids, was used to brew alcoholic beverages and produce vinegar as early as 3000 years ago, a practice that predates the making of chocolate (Henderson et al., 2007).

The cultural connotations of cacao are equally rich and diverse. Its rich taste and sweet aroma have made it a symbol of fairy tales, love, romance, and sensuality. This has given rise to a plethora of chocolate-themed films and literary works. For example, the Hollywood film "Charlie and the Chocolate Factory" critiques the capitalist framework of the factory through a fantastical narrative. Similarly, "The Chocolate War," adapted from the work of Robert Cormier, uses chocolate as a symbol of competition and change (Razdan, 2021).

According to the World Cocoa Foundation's report, in West and Southern Africa, over 90% of global cacao production comes from farmers living in Côte d'Ivoire and Ghana (World Cocoa Foundation, 2024). Cacao serves as a vital source of income, yet these farmers face challenges such as inadequate infrastructure, scarce environmental resources, low productivity, and even the exploitation of child labor (World Cocoa Foundation, 2024). Government-supported policies encouraging farmer participation in cacao cultivation have stimulated the development of related industry chains, including factory production, processing, and logistics trade. Through agricultural transformation

and innovative policies, ultimately, the local economies are diversified, and the socio-market economy achieves sustainable development. Moreover, as a primary export product, cacao has become a pillar of foreign trade for African countries, providing a bridge for some African nations to break economic barriers and integrate with the global financial system (Amuda et al., 2024).

1.2 Microsatellite overview

Within the domain of contemporary genetic research and plant breeding, the utilization of molecular marker technologies has become increasingly prevalent, establishing itself as an essential instrument for the investigation of crop genetic diversity and the execution of precision breeding. To date, a variety of molecular marker types have been successfully developed. These include Random Amplified Polymorphic DNA (RAPD), Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP), Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeats (SSR), also known as microsatellite markers (Varshney et al., 2005; Sehgal et al., 2008; Grover et al., 2016; Ji et al., 2013; Bohn & Melchinger., 1999). These molecular marker technologies have assumed significant roles across diverse areas of the biological sciences. Their applications extend, though are not limited to, the identification of individuals (Jeffreys et al., 1985), the localization of functional genes (Thelander et al., 1989; Pedersen et al., 2002), the construction of genetic linkage maps (Botstein et al., 1980; Faville et al., 2004), Molecular Marker-Assisted Selection (MAS) (Zhang et al., 2010), backcross breeding (Butruille et al., 1999; Hasan et al., 2015), the dissection of population genetic structures (Ouborg et al., 1999; Seelam et al., 2025), and the determination of phylogenetic relationships (Das et al., 2014; Kalia et al., 2011).

Among the numerous molecular markers, microsatellite markers have occupied an important position in biogenetic research due to many desirable attributes, such as high polymorphism, codominance, hypervariability, extensive coverage, and amenability to subsequent high-throughput analysis (Kalia et al., 2011). Consequently, they have been utilized more extensively than other types of markers. As early as 1982, a $(TG)_{25}$ alternating sequence of approximately 50kb was discovered within the Z-DNA

structure of eukaryotic organisms (Hamada et al., 1982). Subsequently, the first genetic linkage maps of the human (Dib et al., 1996) and mouse (Dietrich et al., 1996) genomes were constructed using microsatellite sequences. Jeffreys identified high variability in mini microsatellite DNA in identical twins by examining a pair of twin sisters, demonstrating that this specificity could be employed to ascertain kinship relationships (Jeffreys et al., 1985). Litt et al. (1989) amplified simple sequences repeats of the human genome and designated them as "Microsatellites". Morgante et al. (1993) systematically identified SSRs across multiple plant gene sequences up until 1993.

The efficacy of microsatellite markers is predominantly attributed to two pivotal factors: firstly, their high information content, which refers to the ability to detect a substantial count of discernible alleles and a high frequency of genetic variants; secondly, the procedural simplicity of genotyping (Powell et al., 1996). Their formation mechanism is intricately associated with the phenomenon of slipped mispairing and repair during DNA synthesis by DNA polymerase, as well as the retro-transcription process of certain transposons (Levinson et al., 1987; Vieira et al., 2016). Consequently, any alterations in microsatellite loci lead to changes in the length of the DNA strand. In conclusion, the evolution of microsatellites is a complex and dynamic process (Ellegren., 2004). Hence, it has become an important tool in evolutionary biology and genetic research.

1.2.1 Classification

Microsatellites, also known as short tandem repeats, are genomic DNA fragments that consist of a variable number of short, repetitive nucleotide sequences. The constitution of these sequences spans from mono- to hexanucleotide motifs, exemplified by $(A)_n$, $(AC)_n$, $(CTG)_n$, $(GATA)_n$, $(ATCGC)_n$, and $(ATTGCC)_n$, among others (Tautz, 1989; Lawson et al., 2008; Kumar et al., 2014). In these notations, 'n' denotes the number of repeats of the motif. Weber. (1990) categorized microsatellites in an innovative taxonomic approach to the configuration of the repeat motifs. It is described as the perfect microsatellites (only one repeating motif is connected in series to form a continuous array), imperfect microsatellites (interrupted by the insertion of non-repetitive sequences at certain positions), and compound microsatellites (a combination

of two or more distinct repeat motifs), which can be present in the genome in diverse combinations (Weber, 1990; Oliveira et al., 2006). Building upon this framework, it further introduced a more nuanced classification, distinguishing among simple perfect, simple imperfect, compound perfect, and compound imperfect microsatellites (Wang et al., 2009; Kalia et al., 2011).

1.2.2 Distribution and Positioning

In the plant genomic context, the distribution of simple sequence repeats (SSRs) is predominantly nuclear SSRs (nu-SSRs). However, a subset of SSRs is also found within the chloroplast (cp-SSRs) and mitochondrial (mt-SSRs) genomes. Soranzo et al. (1999) revealed the presence of mitochondrial SSRs in *Pinus* species, signifying the inaugural identification of microsatellites within plant mitochondria. In the mitochondrial NAD region of *Abies miller*, four distinct repeat sequences were identified, and they aggregated to an approximate length of 800 base pairs. (Jaramillo-Correa et al., 2013). More than 80 trinucleotide motifs were detected in the mitochondrial genome of *Selaginella moellendorffii* (Hecht et al., 2011). However, it is noteworthy that the diversity and variability of mitochondrial repeats are less than those of the chloroplast, amounting to approximately two-thirds (Ishii et al., 2006). Weising et al. (1999) identified chloroplast SSRs in *nicotiana* species. The origin of *Phaseolus vulgaris* from Mesoamerican was determined by comparing chloroplast microsatellites (Desiderio et al., 2013). It is generally observed that flowering plants have a higher SSR abundance, greater diversity, and higher density than algae, where the single nucleotide repeat sequence is the most prevalent; in contrast to higher plants, lower plants display relatively more genomic diversity (George et al., 2015). In summary, considering the typically uniparental inheritance characteristic of cp-SSRs and mt-SSRs, they are endowed with significant research value in exploring the variability of plant genomes, as well as phylogenetic relationships, gene flow between populations, and evolutionary history (Powell et al., 1996; Provan et al., 2001).

1.2.3 Microsatellite characteristics

1.2.3.1 Widely distributed and non-random

Microsatellites have now been identified across the genomes of most eukaryotic and prokaryotic organisms. The variation in repeat sequences forms the basis of microsatellite locus polymorphism (Zane et al., 2002; Morgante et al., 1993). Research indicates that microsatellites are distributed across introns, exons, intergenic regions, and regulatory areas. Typically, a microsatellite is found every 10-50 kilobases, and there is a high diversity of types (Tautz, 1986).

The distribution of SSR in the genome is not completely random but follows a discernible pattern. In the transcriptome of *Sargassum thunbergii* (*Phaeophyta*), the untranslated regions (UTRs) were found to contain a higher number of microsatellites compared to the coding DNA sequences (CDSs), with trinucleotide repeats being the generally abundant in the UTR region (Liu et al., 2016). Similar findings have been observed in the genomes of *Arabidopsis thaliana* and rice, where, based on comparative analysis, the SSRs density in the 5 'UTR region of *Arabidopsis thaliana* and rice was significantly higher than in other regions (Lawson et al., 2006). In addition, studies have shown that, aside from hexanucleotide types which are more frequent in coding regions, other SSRs are significantly denser in non-coding areas than in other regions (Zhang et al., 2004; Morgante et al., 2002; Xu et al., 2013). Tetranucleotide repeats are predominantly located in the non-coding intergenic regions of rice; in the maize genome, the density of SSRs decreases sequentially in non-coding regions, promoters, introns, intergenic spaces, and coding sequence regions (Qu et al., 2013).

The cause of this phenomenon may be attributed to the fact that the CDSs (Coding DNA Sequence) regions are directly involved in protein coding or RNA production, and even minor changes can alter the amino acid sequence. Consequently, microsatellites located within the CDSs are more likely to disrupt gene function and protein structure compared to those in the UTRs (Un-Translated Regions). Therefore, microsatellites in the CDSs are subject to more exacting evolutionary constraints, primarily limiting their variation and distribution within the genome. In contrast, UTRs, which are under less evolutionary pressure, exhibit a higher tolerance for mutations. This may result in a

higher frequency and polymorphism of microsatellite sequences in these areas (Kalia et al., 2011; Liu et al., 2016).

Moreover, repeat sequences exhibit conserved patterns among species. *Arabidopsis thaliana* has a higher frequency of mononucleotide repeats, while rice shows a stronger affinity for trinucleotide repeats. However, both plants share the same mononucleotide and dinucleotide repeat sequences, $(A)_n$, $(T)_n$, $(AG)_n$, and $(AT)_n$ (Lawson et al., 2006). Monocotyledons had an average of one short tandem repeat (STR) sequence per 64.6 kb of DNA, while dicotyledons had one STR per 21.2 kb.

1.2.3.2 Regulation of gene expression

As research into microsatellites progresses, it has been observed that these sequences are prone to slipped-strand mispairing during DNA replication, leading to mutations that can affect gene expression regulation. SSRs have been observed in both CDSs and UTRs of oil palm plants. Among them, the sequences encoding bZIP, NAC-like and MADS-box transcription factors have microsatellite polymorphism and most of the mutation sites are located within the open reading frame (Tranbarger et al., 2012). In addition, Nalavade. (2013) found that SSRs amplification in the 3'-UTR can disrupt mRNA transcription, affecting the formation of mature mRNA and its cytoplasmic transport, thereby disrupting cellular function. When trinucleotide repeat sequences are in the UTRs or intron regions, they have the potential to activate heterochromatin formation, thereby inducing gene silencing. Such gene silencing events can have profound effects on gene expression patterns, leading to phenotypic changes (Nalavade et al., 2013). Enrichment analysis has identified four pathways in *Sargassum thunbergii* (Ubiquitin-mediated proteolysis, Spliceosome, Terpenoid backbone biosynthesis, RNA degradation) that can regulate the function and evolution of microsatellites in the transcriptome (Liu et al., 2016). In the NCBI database, tandem repeat sequences are widely present in biological proteomes, potentially exerting significant impacts on the structure and function of proteins, thus promoting rapid protein evolution (Huntley et al., 2000). Additionally, studies have indicated that SSRs may influence the differential expression profiles between housekeeping genes and tissue-specific genes (Lawson et

al., 2008).

It also has been discovered that microsatellites co-evolve with transposable elements and are subject to natural selective pressures. Transposable elements (TEs), which include retrotransposons and DNA transposons, were once classified as "self-replicating junk DNA" in the genome (Kejnovsky et al., 2009). However, research has unveiled their active role in genome structure, function, and evolution through mechanisms such as gene rearrangement, amplification, and insertional mutagenesis. They are now considered major players in genomic evolution, exerting a profound influence on the diversity and evolution of organisms (Gao et al., 2013). In the rice genome, 100-200 microsatellite sequences composed mainly of $(TA)_n$ in subterminal inverted repeats within the Micron transposon family have been identified (Akagi et al., 2001). However, a negative correlation exists between the amplification mechanism of microsatellites and retrotransposons, which differs from the DNA replication process (Hua-Van et al., 2005). In Drosophila, the density of LTR retrotransposons (based on RNA elements) is higher in regions with low microsatellite recombination rates (Rizzon et al., 2002). In brief, microsatellite loci significantly impact gene expression patterns, potentially by altering the structure of gene regulatory regions or interacting with transposable elements, thereby affecting gene transcriptional activity and biological functions. These regulatory changes in gene expression may ultimately have profound biological impacts on the phenotype of organisms.

1.2.3.3 Gene mapping and linkage

In fish species (*Leporinus spp*.) with sex determination governed by the ZW system, the heterochromatic region of the W chromosome has been reported to contain a substantial number of conserved repetitive sequences. This suggests that certain microsatellite sequences are linked to the sex chromosomes controlling the W genes, thereby playing a significant role in maintaining the stability of genes on the heterochromosome (Poltronieri et al., 2013). Similarly, in the non-recombining regions of the Y chromosome in some plants, specific DNA motifs, such as $(CAA)_n$ and $(TAA)_n$, have also been documented to exhibit a high degree of amplification. The accumulation

of microsatellites on these heterochromosomes is considered to be fundamental to the evolution of sex chromosomes (Kejnovsky et al., 2009). These findings hold substantial scientific significance for understanding the unique structure and function of sex chromosomes, as well as providing important insights into their evolution within species.

1.2.4 Application of microsatellite loci

Microsatellite markers, as a superior class of codominant molecular markers, enable the tracking of changes in the genetic structure of populations by monitoring fluctuations in gene frequencies. They have become an indispensable tool in population ecology research and have been widely applied in biological and genetic studies.

In the field of botany, microsatellite markers have been extensively utilized to investigate the genetic characteristics of a variety of plants. Within the context of barley, genetic relationships among 18 varieties were analyzed using amplified fragment length polymorphism (AFLP) primers, simple sequence repeat (SSR) primer pairs, restriction fragment length polymorphism (RFLP) probes, and Random Amplified Polymorphic DNA (RAPD) methods, ultimately revealing the highest similarity between varieties “Triumph” and “Grit” (Russell et al., 1997). Microsatellite markers identified $(CA)_8$ repeat-rich sequences in the HNCA-1 and HNCA-2 loci of sunflower (Brunel et al., 1994). It is currently known that rice contains approximately 5700-10000 microsatellite loci, providing technical support for molecular breeding (McCouch et al., 1997). In cotton, the practicality of SSRs for the localization of transcribed genes was validated, and significant differences were determined in amplification using cDNA and DNA templates (Saha et al., 2003). Therefore, when designing microsatellite primers, it is feasible to identify microsatellite loci based on Expressed sequence tags (EST) and develop specific primers for amplifying these loci (Kalia et al., 2011). Moreover, microsatellite markers have often been employed to assess biodiversity and genetic structure. In grugru palm (*Acrocomia aculeata*), 16.8% of genetic variation was found in inter-population genetic variation and intra-population genetic variation accounted for 83.2% (Coelho et al., 2018). Sargent et al. (2004) used 78 molecular markers,

including 68 microsatellite loci in diploid hybrid strawberries, identifying seven discrete linkage groups and constructing a 448 cM diploid strawberry genetic map, laying the foundation for polyploid strawberry breeding. Additionally, microsatellite markers have demonstrated significant research value and application potential in the localization and linkage analysis of plant disease resistance genes. *Alternaria alternata* and *A. solani*, which affect potato leaves, show a certain degree of correlation with leaf cuticle wax thickness, suggesting that wax thickness could serve as a phenotypic marker for the identification of potato varieties resistant to early blight (EB) (Alizadeh-Moghaddam et al., 2024). Microsatellite markers have been used in backcross breeding to locate the Nax1 and Nax2 gene loci associated with salt stress resistance in the progeny of wheat varieties “WH1105” and “Kharchia 65” (Sharma et al., 2024).

In species conservation, microsatellite markers also play a crucial role. The wild celery (*K. odoratissima*) has experienced a drastic decline due to environmental changes and excessive harvesting. A study by Mahdavikia et al. (2024) indicates that wild celery (*K. odoratissima*) is rich in dinucleotide repeats (AC/GT; CA/TG) at 68.7%, trinucleotide repeats (TTC/GAA; GAA/TTC) at 30.2% and other repeats 1.1% with an average polymorphic information content (PIC) of 0.81. The high polymorphism of these dinucleotide and trinucleotide repeats is effective in analyzing the genetic diversity of wild celery and suggests guidelines for species conservation. All in all, these applications are not only of significant importance in agricultural science but also have a direct impact on sustainable biological development and the enhancement of breeding efficiency.

In the fields of forensic science and paternity testing, microsatellite markers also play a crucial role. Due to the considerable allelic diversity at microsatellite loci within the same species, this technique provides highly reliable biological evidence for confirming individual identity, kinship, and family lineage analysis (Veselinović et al., 2006).

Furthermore, microsatellite markers have significant value in disease linkage and association studies, enabling researchers to precisely locate genes related to diseases and gain a deeper understanding of the relationship between genetic factors and

diseases. This contributes to revealing the differences in organisms' adaptability to the environment, disease susceptibility, and drug tolerance, thereby advancing the development of disease prevention, diagnosis, and treatment (Holmes, 1994).

All things considered, these research findings not only aid in deciphering the processes of biological evolution but also provide powerful tools for mapping genes and conducting family lineage analysis, ultimately propelling the advancement of genetic and evolutionary research and biomedical sciences.

1.3 Overview of genetic diversity

Biological diversity refers to the aggregate of all biological variations within a given area, containing all species of life. Genetic diversity represents the sum of all genetic variations within a single species. Broadly speaking, it involves the exploration of all genetic information within entire ecosystems. Narrowly defined, it refers to the total genetic variation both within and between species. Genetic diversity is primarily manifested through biological morphological traits, cellular metabolism, isoenzyme markers, and the DNA molecular level. The first three are indirect reflections of genetic variation as a result of gene expression, while the latter is a direct reflection (Aikpokpodion, 2012). There is a central tool in the study of genetic diversity of species i.e. hpothesized that when one species has a sufficiently large population and there is largely random mating, its genotype and allele frequencies will exhibit stability, maintaining a balance in population numbers and preserving a certain level of diversity (Chen., 2010). In plant science research, the Hardy-Weinberg Principle is important for dealing with problems related to the genetic structure of populations, providing guidance for crop breeding, facilitating the calculation of quantitative relationships between heterozygotes and homozygotes, and so on (Ayiecho & Nyabundi., 2025). To sum up, genetic diversity is an important indicator for studying the adaptation of species to changes in the environment, as well as for measuring the survival of species. The genetic diversity of cacao will be discussed further in chapter II.

1.4 Application of molecular markers in cacao

Currently, molecular marker methods make an important contribution to the understanding of the genetic diversity of cacao trees. Firstly, it has been applied to the identification of cacao genotypes. As early as 1994, Figueira et al. (1994) used molecular markers to classify cacao trees. Fang et al. (2014) used SNP fingerprints generated from cacao seed coat DNA to isolate the correct cacao seeds from the suspected adult variety. Subsequently, molecular marker technology has been widely used in cacao evolution and diversity studies. SSR polymorphisms in the cacao plastid genome were used to study the differences between seven different cacao genotypes from Peru, and it was proposed that high differentiation of cacao populations occurs around the Pliocene or Miocene epochs (Tineo et al., 2025). Of the 93 cacao genotypes collected from Colombia, random amplification of microsatellite markers revealed that these cacao varieties were geographically clustered (Morillo et al., 2014). Everaert et al. (2017) analyzed 14 microsatellite markers and found that Vietnamese cacao varieties “Can Tho” and “Thu Duc” belonged to Trinitario and Forastero, respectively. Among them, mTcCIR 15, mTcCIR 33, and mTcCIR 37 had the highest polymorphisms among the microsatellite markers (Everaert et al., 2017).

In addition, molecular marker technology has been focused on the screening of cacao disease-resistance genes. Witches' Broom Disease caused by *Moniliophthora perniciosa* severely affects cacao growth and development. The search for genes associated with *M. perniciosa* - cacao protein interactions provide theoretical support for disease mitigation and precision treatment (Santos et al., 2023). Black pod disease is caused by *phytophthora* species (Barreto et al., 2015). It was found that primers designed from SNP loci generated from the CAT1 gene sequence were effective in distinguishing between susceptible and resistant cacao (Tarigan et al., 2021). This further improved the accuracy of screening for disease-resistant genotypes of cacao. The best solution to vascular streak dieback is to breed resistant varieties. Due to the fact that mTcCIR42 is linked to disease-resistance genes, cacao seeds with resistance genes in each generation can be quickly screened by molecular marker method (Chandrakant et al., 2020). However, there are still difficulties in breeding high-yielding

and disease-resistant cacao due to the narrow genetic bases and self-incompatibility of cacao (Maji et al., 2025). Future research is needed.

2. Conclusion

First, this chapter has introduced the origin, physiological characteristics, and classification of cacao, and revealed its rich historical and cultural background and genetic diversity. Secondly, it has elaborated on the development of microsatellite markers and their application in cacao research, emphasizing the significance of SSR markers in genetic research. Followed by an introduction to the relevant aspects of genetic diversity. Finally, this chapter has summarized the results of the application of molecular markers in cacao research in recent years. It provides theoretical support for further research on the genetic structure and development potential of cacao and sets the foundation for the next step of research. Despite these advances, there is still a gap in the literature on Samoa cacao. Future research will focus on exploring the diversity of Samoa cacao, aiming to provide a stronger scientific basis for collecting germplasm information, conservation, and improvement of cacao.

3. Thesis objectives

Currently, research on the genetic diversity of cocoa in the Samoa region is very limited. Given its extensive planting area and economic significance, the main objective of this study is as follows:

1. Summarize the state of the cocoa industry and research, with a focus on Samoa.
2. Examine and describe the genetic diversity of cocoa in Samoa from four geographical locations.
3. Provide insights into the future directions for cocoa research.

4. Reference List

Ayiecho, P. O., & Nyabundi, J. O. (2025). Population Genetic Variation, Genetic Structure of Crop Plant Populations, Inbreeding Depression and Heterosis. *In Conventional and Contemporary Practices of Plant Breeding* (pp. 79-102). Cham: Springer Nature Switzerland.

Barreto, M. A., Santos, J. C. S., Corrêa, R. X., Luz, E. D. M. N., Marelli, J., & Souza, A. P. (2015). Detection of genetic resistance to cocoa black pod disease caused by three *Phytophthora* species. *Euphytica*, 206, 677-687.

Barreto, M. A., Rosa, J. R. B. F., Holanda, I. S. A., Santos, J. C. S., Corrêa, R. X., & Souza, A. P. (2018). QTL mapping and identification of corresponding genomic regions for black pod disease resistance to three *Phytophthora* species in *Theobroma cacao* L. *Euphytica*, *214*(10), 188. https://doi.org/10.1007/s10681-018-2273-5

Bohn, M., Utz, H. F., & Melchinger, A. E. (1999). Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop Science*, *39*(1), 228–237. https://doi.org/10.2135/cropsci1999.0011183X003900010035x

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.

Brunel, D. (1994). A microsatellite marker in *Helianthus annuus* L. *Plant Molecular Biology*, *24*(2), 397–400. https://doi.org/10.1007/BF00020172

Butruille, D. V., Guries, R. P., & Osborn, T. C. (1999). Linkage analysis of molecular markers and quantitative trait loci in populations of inbred backcross lines of *Brassica napus* L. *Genetics*, *153*(2), 949–964. https://doi.org/10.1093/genetics/153.2.949

Chandrakant, E. S., Tulshiram, W. S., Mathew, D., Minimol, J. S., Jose, S., Kurian, S. P., & Shylaja, M. R. (2020). Microsatellite and inter-microsatellite markers linked with resistance to vascular streak dieback in cocoa (*Theobroma cacao* L.).

The Journal of Horticultural Science and Biotechnology, *95*(3), 294–302. https://doi.org/10.1080/14620316.2019.1689186

Chen, J. J. (2010). The Hardy-Weinberg principle and its applications in modern population genetics. *Frontiers in Biology*, *5*(4), 348–353. https://doi.org/10.1007/s11515-010-0644-y

Coelho, N. H. P., Tambarussi, E. V., Aguiar, B. I., Roque, R. H., Portela, R. M., Braga, R. C., Sanson, D., Silva, R. A. R., Ferraz, E. M., Moreno, M. A., Kageyama, P. Y., & Gandara, F. B. (2018). Understanding genetic diversity, spatial genetic structure, and mating system through microsatellite markers for the conservation and sustainable use of *Acrocomia aculeata* (Jacq.) Lodd. Ex Mart. *Conservation Genetics*, *19*(4), 879–891. https://doi.org/10.1007/s10592-018-1061-z

Das, S., Dash, H. R., Mangwani, N., Chakraborty, J., & Kumari, S. (2014). Understanding molecular identification and polyphasic taxonomic approaches for genetic relatedness and phylogenetic relationships of microorganisms. *Journal of Microbiological Methods*, *103*, 80–100. https://doi.org/10.1016/j.mimet.2014.05.013

De Andrade Almeida, A. C., De Moraes Pontes, J. G., Alvarenga, G. R., Finocchio, H., & Fill, T. P. (2021). The sustainable cycle of a new cacao-based bioplastic: From manufacturing to exploitable biodegradation products. *RSC Advances*, *11*(48), 29976–29985. https://doi.org/10.1039/D1RA04432J

Desiderio, F., Bitocchi, E., Bellucci, E., Rau, D., Rodriguez, M., Attene, G., Papa, R., & Nanni, L. (2013). Chloroplast microsatellite diversity in *Phaseolus vulgaris*. *Frontiers in Plant Science*, *4*, 312. https://doi.org/10.3389/fpls.2013.00312

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Kazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., & Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, *380*(6570), 152–154. https://doi.org/10.1038/380152a0

Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M. J., Ingalls, K. A., O'Connor, T. J., Evans, C. A., DeAngelis, M. M., Levinson, D. M., Kruglyak, L., Goodman, N., Copeland, N. G., Jenkins, N. A., Hawkins, T. L., Stein, L., & Lander, E. S. (1996). A comprehensive genetic map of the mouse genome. *Nature*, *380*(6570), 149–152. https://doi.org/10.1038/380149a0

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, *5*(6), 435–445. https://doi.org/10.1038/nrg1348

Everaert, H., Rottiers, H., Pham, P. H. D., Ha, L. T. V., Nguyen, T. P. D., Tran, P. D., De Wever, J., Maebe, K., Smagghe, G., Dewettinck, K., & Messens, K. (2017). Molecular characterization of Vietnamese cocoa genotypes (*Theobroma cacao* L.) using microsatellite markers. *Tree Genetics & Genomes*, *13*(5), 99. https://doi.org/10.1007/s11295-017-1180-6

Fang, W., Meinhardt, L. W., Mischke, S., Bellato, C. M., Motilal, L., & Zhang, D. (2014). Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *Journal of Agricultural and Food Chemistry*, *62*(2), 481–487. https://doi.org/10.1021/jf4044024

Faville, M. J., Vecchies, A. C., Schreiber, M., Drayton, M. C., Hughes, L. J., Jones, E. S., Guthridge, K. M., Smith, K. F., Sawbridge, T. I., Spangenberg, G. C., Bryan, G. T., & Forster, J. W. (2004). Functionally associated molecular genetic marker map construction in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics*, *110*(1), 12–32. https://doi.org/10.1007/s00122-004-1785-7

Figueira, A., Janick, J., Levy, M., & Goldsbrough, P. (1994). Reexamining the classification of *Theobroma cacao* L. using molecular markers. *Journal of the American Society for Horticultural Science*, *119*(5), 1073–1082. https://doi.org/10.21273/JASHS.119.5.1073

Gao, C., Ren, X., Mason, A. S., Li, J., Wang, W., Xiao, M., & Fu, D. (2013). Revisiting an important component of plant genomes: Microsatellites. *Functional Plant Biology*, *40*(7), 645–661. https://doi.org/10.1071/FP12343

George, B., Bhatt, B. S., Awasthi, M., George, B., & Singh, A. K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Current Genetics*, *61*(4), 665–677. https://doi.org/10.1007/s00294-015-0493-y

Grover, A., & Sharma, P. C. (2016). Development and use of molecular markers: Past and present. *Critical Reviews in Biotechnology*, *36*(2), 290–302. https://doi.org/10.3109/07388551.2014.959891

Hamada, H., Petrino, M. G., & Kakunaga, T. (1982). A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proceedings of the National Academy of Sciences*, *79*(21), 6465–6469.

Hasan, M. M., Rafii, M. Y., Ismail, M. R., Mahmood, M., Rahim, H. A., Alam, M. A., Ashkani, S., Malek, M. A., & Latif, M. A. (2015). Marker-assisted backcrossing: A useful method for rice improvement. *Biotechnology & Biotechnological Equipment*, *29*(2), 237–254. https://doi.org/10.1080/13102818.2014.995920

Hecht, J., Grewe, F., & Knoop, V. (2011). Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: The root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biology and Evolution*, *3*, 344–358.

Henderson, J. S., Joyce, R. A., Hall, G. R., Hurst, W. J., & McGovern, P. E. (2007). Chemical and archaeological evidence for the earliest cacao beverages. *Proceedings of the National Academy of Sciences USA*, *104*, 18937–18940.

Holmes, N. G. (1994). Microsatellite markers and the analysis of genetic disease. *British Veterinary Journal*, *150*(5), 411–421.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C., & Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: Similarities and differences. *Cytogenetic and Genome Research*, *110*(1–4), 426–440.

Huntley, M., & Golding, G. B. (2000). Evolution of simple sequence in proteins. *Journal of Molecular Evolution*, *51*, 131–140.

Improving farmer income through sustainable practices. (2024). *World Cocoa Foundation*. Retrieved August 10, 2024, from https://worldcocoafoundation.org/focus-areas/farmer-income

Ishii, T., Takahashi, C., Ikeda, N., Kamijima, O., & Mori, N. (2006). Mitochondrial microsatellite variability in common wheat and its ancestral species. *Genes & Genetic Systems*, *81*(3), 211–214.

Jaramillo-Correa, J. P., Aguirre-Planter, E., Eguiarte, L. E., Khasa, D. P., & Bousquet, J. (2013). Evolution of an ancient microsatellite hotspot in the conifer mitochondrial genome and comparison with other plants. *Journal of Molecular Evolution*, *76*, 146–157.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature*, *316*(6023), 76–79.

Ji, K., Zhang, D., Motilal, L. A., Boccara, M., Lachenaud, P., & Meinhardt, L. W. (2013). Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. *Genetic Resources and Crop Evolution*, *60*, 441–453.

Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, *177*(3), 309–334.

Kejnovsky, E., Hobza, R., Cermak, T., Kubat, Z., & Vyskot, B. (2009). The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity*, *102*(6), 533–541.

Kumar, M., Kapil, A., & Shanker, A. (2014). MitoSatPlant: Mitochondrial microsatellites database of Viridiplantae. *Mitochondrion*, *19*, 334–337.

Kuppers, J. R. (1953). Some biometric observations on cacao fruit. *Science*, *117*(3040), 354–355.

Lachenaud, P., Mooleedhar, V., & Couturier, C. (1997). Les cacaoyers spontanés de Guyane: Nouvelles prospections. *Plantations, Recherche, Développement*, *4*, 25–30.

Lawson, M. J., & Zhang, L. (2006). Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology*, *7*, 1–11.

Lawson, M. J., & Zhang, L. (2008). Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′-UTR region. *Gene*, *407*(1–2), 54–62.

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, *4*(3), 203–221.

Litt, M., & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, *44*(3), 397.

Liu, F., Hu, Z., Liu, W., Li, J., Wang, W., Liang, Z., Wang, F., & Sun, X. (2016). Distribution, function and evolution characterization of microsatellite in *Sargassum thunbergii* (Fucales, Phaeophyta) transcriptome and their application in marker development. *Scientific Reports*, *6*(1), 18947.

Mahdavikia, F., Ebadi, M. T., Shojaeiyan, A., Ayyari, M., & Falahati-Anbaran, M. (2024). Genetic variation and structure of endemic and endangered wild celery (*Kelussia odoratissima* Mozaff.) quantified using novel microsatellite markers developed by next-generation sequencing. *Frontiers in Plant Science*, *15*, 1301936.

Maji, A., Samanta, S., Mandal, A., Mandal, A., Imtiyaz, M., Gorai, S., Parimala, G., Hassan, W., & Hazra, S. (2025). Marker-assisted selection and genetic improvement of industrial crops. In *Industrial crops improvement: Biotechnological approaches for sustainable agricultural development* (pp. 97–117). Cham: Springer Nature Switzerland.

Martínez-Pinilla, E., Oñatibia-Astibia, A., & Franco, R. (2015). The relevance of theobromine for the beneficial effects of cocoa consumption. *Frontiers in Pharmacology*, *6*, Article 1663. https://doi.org/10.3389/fphar.2015.00030

McCouch, S. R., Cho, Y. G., Yano, M., Paul, E., Blinstrub, M., Morishima, H., & Kinoshita, T. (1997). Microsatellite marker development, mapping and applications in rice genetics and breeding. In T. Sasaki & G. Moore (Eds.),

Oryza: From molecule to plant (pp. 157–173). Dordrecht: Springer. https://doi.org/10.1007/978-94-011-5794-0_9

Morgante, M., & Olivieri, A. M. (1993). PCR-amplified microsatellites as markers in plant genetics. *The Plant Journal*, *3*(1), 175–182.

Morgante, M., Hanafey, M., & Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics*, *30*(2), 194–200.

Morillo, C. Y., Morillo, C. A. C., Muñoz, J. E. B., Ballesteros, W. P., & González, A. (2014). Molecular characterization of 93 genotypes of cocoa (*Theobroma cacao* L.) with random amplified microsatellites (RAMs). *Agronomía Colombiana*, *32*(3), 315–325.

Motamayor, J. C., & Lanaud, C. (2002). Molecular analysis of the origin and domestication of *Theobroma cacao* L. In *Managing plant genetic diversity: Proceedings of an international conference, Kuala Lumpur, Malaysia, 12–16 June 2000* (pp. 77–87). Wallingford: CABI Publishing.

Nalavade, R., Griesche, N., Ryan, D. P., Hildebrand, S., & Krauss, S. (2013). Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death & Disease*, *4*(8), e752.

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., & Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, *29*, 294–307.

Ouborg, N. J., Piquot, Y., & Van Groenendael, J. M. (1999). Population genetics, molecular markers and the study of dispersal in plants. *Journal of Ecology*, *87*(4), 551–568.

Pedersen, C., Rasmussen, S. W., & Giese, H. (2002). A genetic map of *Blumeria graminis* based on functional genes, avirulence genes, and molecular markers. *Fungal Genetics and Biology*, *35*(3), 235–246.

Poltronieri, J., Marquioni, V., Bertollo, L. A. C., Kejnovsky, E., Molina, W. F., Liehr, T., & Cioffi, M. B. (2013). Comparative chromosomal mapping of microsatellites in *Leporinus* species (Characiformes, Anostomidae): Unequal

accumulation on the W chromosomes. *Cytogenetic and Genome Research*, *142*(1), 40–45.

Powell, W., Machray, G. C., & Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science*, *1*(7), 215–222.

Provan, J., Powell, W., & Hollingsworth, P. M. (2001). Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution*, *16*(3), 142–147.

Qu, J., & Liu, J. (2013). A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. *BMC Research Notes*, *6*, 1–10.

Razdan, D., & Arora, J. (2021). Chocolate and the Holly Factory: Analysing the "Role" of Chocolate in Select Films from Hollywood. In S. Malhotra, K. Sharma, & S. Dogra (Eds.), *Food culture studies in India* (pp. 34–36). Springer, Singapore. https://doi.org/10.1007/978-981-15-5254-09

Rizzon, C., Marais, G., Gouy, M., & Biémont, C. (2002). Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research*, *12*(3), 400–407.

Rodriguez-Campos, J., Escalona-Buendía, H. B., Contreras-Ramos, S. M., Orozco-Avila, I., Jaramillo-Flores, E., & Lugo-Cervantes, E. (2012). Effect of fermentation time and drying temperature on volatile compounds in cocoa. *Food Chemistry*, *132*(1), 277–288. https://doi.org/10.1016/j.foodchem.2011.10.078

Russell, J. R., Fuller, J. D., Macaulay, M., Hatz, B. G., Jahoor, A., Powell, W., & Waugh, R. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *Theoretical and Applied Genetics*, *95*, 714–722.

Saha, S., Karaca, M., Jenkins, J. N., Zipf, A. E., Reddy, O. U. K., & Kantety, R. V. (2003). Simple sequence repeats as useful resources to study transcribed genes of cotton. *Euphytica*, *130*, 355–364.

Santos, A. S., Mora-Ocampo, I. Y., de Novais, D. P. S., Aguiar, E. R. G. R., & Pirovani, C. P. (2023). State of the art of the molecular biology of the interaction

between cocoa and witches broom disease: A systematic review. *International Journal of Molecular Sciences*, *24*(6), 5684.

Sargent, D. J., Davis, T. M., Tobutt, K. R., Wilkinson, M. J., Battey, N. H., & Simpson, D. W. (2004). A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. *Theoretical and Applied Genetics*, *109*, 1385–1391. https://doi.org/10.1007/s00122-004-1767-9

Seelam, R. T., Puram, V. R. R., Nallamothu, V., Injeti, S. K., & Madhunapantula, V. P. (2025). Unlocking the potential of rice for thermotolerance using simple sequence repeat (SSR) marker-based assessment of genetic variability and population structure. *Genetic Resources and Crop Evolution*, *72*(1), 947–963.

Sehgal, D., & Raina, S. N. (2008). DNA markers and germplasm resource diagnostics: New perspectives in crop improvement and conservation strategies. In *Utilization of biotechnology in plant sciences* (pp. 39–54). Rastogi Publications.

Sharma, K., Yashveer, S., Singh, V., Sharma, S., Redhu, M., & Rahimi, M. (2024). Microsatellite marker-assisted backcross breeding for improvement of wheat salt tolerance using Kharchia 65. *BMC Genomics*, *25*(1), 550.

Soranzo, N., Provan, J., & Powell, W. (1999). An example of microsatellite length variation in the mitochondrial genome of conifers. *Genome*, *42*(1), 158–161.

Tarigan, R., Maharijaya, A., & Izzah, N. K. (2021). Snap markers derived from catalase-1 gene sequence used for black pod disease resistance in cacao (*Theobroma cacao* L.). *SABRAO Journal of Breeding & Genetics*, *53*(3), 100-104.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic *DNA* markers. *Nucleic Acids Research*, *17*(16), 6463–6471.

Tautz, D., Trick, M., & Dover, G. (1986). Cryptic simplicity in *DNA* is a major source of genetic variation. *Nature*, *322*, 652–656. https://doi.org/10.1038/322652a0

Thelander, M., & Thelander, L. (1989). Molecular cloning and expression of the functional gene encoding the M2 subunit of mouse ribonucleotide reductase: A new dominant marker gene. *The EMBO Journal*, *8*(9), 2475–2479.

Tineo, D., Bustamante, D. E., Calderon, M. S., & Oliva, M. (2025). Comparative analyses of chloroplast genomes of *Theobroma cacao* from northern Peru. *PLOS ONE*, *20*(3), e0316148.

Tran, P. D., Van de Walle, D., De Clercq, N., De Winne, A., Kadow, D., Lieberei, R., Messens, K., Tran, D. N., Dewettinck, K., & Van Durme, J. (2015). Assessing cocoa aroma quality by multiple analytical approaches. *Food Research International*, *77*, 657–669. https://doi.org/10.1016/j.foodres.2015.09.019

Tranbarger, T. J., Kluabmongkol, W., Sangsrakru, D., Morcillo, F., Tregear, W. J., Tragoonrung, S., & Billotte, N. (2012). SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. *BMC Plant Biology*, *12*, 1–12.

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: Features and applications. *Trends in Biotechnology*, *23*(1), 48–55.

Verna, R. (2013). The history and science of chocolate. *The Malaysian Journal of Pathology*, *35*(2), 111.

Verner, V., Novy, P., Tauchen, J., Huml, L., Soon, J. W., Kudera, T., Laupua, F., & Kokoska, L. (2020). Diversity, economic value and regional distribution of plant food products at local tropical markets: A Samoa case study. *Sustainability*, *12*(23), 10014.

Veselinović, I. (2006). Microsatellite *DNA* analysis as a tool for forensic paternity testing (DNA paternity testing). *Medicinski Pregled*, *59*(5–6), 241–243.

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. D. F. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*, 312–328.

Wang, M. L., Barkley, N. A., & Jenkins, T. M. (2009). Microsatellite markers in plants and insects. Part I: Applications of biotechnology. *Genes, Genomes and Genomics*, *3*(1), 54–67.

Wattnem, T., Wiegel, J., González, C., & Reyes, B. (2022). Who defines fine chocolate? The construction of global cocoa quality standards from Latin

America. *The International Journal of Sociology of Agriculture and Food*, *28*(1), 73–87.

Weber, J. L. (1990). Informativeness of human (dC-dA) n·(dG-dT) n polymorphisms. *Genomics*, *7*(4), 524–530.

Weising, K., & Gardner, R. C. (1999). A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome*, *42*(1), 9–19.

Xu, J., Liu, L., Xu, Y., Chen, C., Rong, T., Ali, F., Zhou, S., Wu, F., Liu, T., Wang, J., Cao, M., & Lu, Y. (2013). Development and characterization of simple sequence repeat markers providing genome-wide coverage and high resolution in maize. *DNA Research*, *20*(5), 497–509.

Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, *11*(1), 1–16.

Zhang, D., & Motilal, L. (2016). Origin, dispersal, and current global distribution of cacao genetic diversity. In B. Bailey & L. Meinhardt (Eds.), *Cacao diseases* (pp. 1–25). Springer, Cham. https://doi.org/10.1007/978-3-319-24789-2_1

Zhang, L., Yuan, D., Yu, S., Li, Z., Cao, Y., Miao, Z., Qian, H., & Tang, K. (2004). Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. *Bioinformatics*, *20*(7), 1081–1086.

Zhang, W., He, H., Guan, Y., Du, H., Yuan, L., Li, Z., Yao, D., Pan, J., & Cai, R. (2010). Identification and mapping of molecular markers linked to the tuberculate fruit gene in the cucumber (*Cucumis sativus* L.). *Theoretical and Applied Genetics*, *120*, 645–654.

Chapter 2

Population genetics of *Theobroma cacao* L.

1. Introduction and Background

1.1 Samoa overview

Samoa is situated in Oceania --- in the central South Pacific --- at the heart of the Polynesian Islands. It comprises two main islands and several smaller, positioned between latitudes 13°15’ and 14°05’ South and longitudes 168° and 173° West. The topography of Samoa is characterized by a volcanic landscape that generally slopes from the northwest to the southeast, creating rugged terrain and lush tropical ecosystems (Whitkus et al., 1998). The economy of Samoa is predominantly based on agricultural exports and tourism. Agriculture focuses on the cultivation of tropical crops such as coconuts, taro, cacao, and papayas (Verner et al., 2020). Cacao cultivation in Samoa has a long history, with the initial introduction of the crop from Sri Lanka in 1883 (Dillon et al., 2024). Samoa is widely acclaimed for its superior cacao varieties, which are renowned for their exceptional quality and distinctive flavor profiles. This has earned the region the accolades of "Chocolate Island" and "Ecuador of the Pacific," reflecting its significant standing and influence in the global cacao industry (Verner et al., 2020). Reports indicate that the global consumption of chocolate is steadily increasing at an annual growth rate of 2% to 3% (Raju et al., 2024). This upward trend has had a notable economic impact on cacao-producing countries worldwide. As a vital participant in global cacao production, Samoa's economy is closely tied to the development of its cacao industry. Cacao trees, as a major cash crop in Samoa, play a pivotal role in boosting local economic development. Currently, it is estimated that Samoa's cacao cultivation covers approximately 2,000 hectares, accounting for over two-thirds of the arable land, highlighting the prominent position of the cacao industry within the country's agricultural structure (Raju et al., 2024).

Samoa commenced cacao cultivation in the year 1883, and by the year 1892, the

industry had scaled up significantly (Martínez., 2025). In Samoa, the initial cacao varieties introduced were Criollo and Forastero. Criollo, characterized by its yellow-green pods, is more susceptible to pests and diseases compared to other varieties, and has a lower yield, accounting for approximately 3% of the global cacao production. However, it is celebrated for its exceptional flavor profile, often associated with floral and fruity and low bitterness, which has made it highly favored in the premium chocolate market. Conversely, the Forastero variety, the most widely cultivated cacao type, is characterized by its intense and robust flavor, and high acidity, and is commonly described as having fruity, earthy, and bitter tastes. Its strong resistance to pests and diseases has resulted in it constituting 80% of the world's cacao output. This variety exhibits substantial genomic diversity, encompassing at least nine distinct genotypes (Ascrizzi et al., 2017).

The volatile compounds such as acetic acid and isobutyric acid, produced during the fermentation and drying processes of cacao, significantly influence the flavor and aroma, making them key indicators for assessing the quality of cacao germplasm (Rodriguez-Campos et al., 2011). These compounds serve not only as markers for the quality of cacao germplasm but also have a pronounced impact on the final product's quality, consumer acceptance, and preference (Wattnem et al., 2022; Tran et al., 2015). In the pursuit of superior flavor and enhanced genetic resistance, the Trinitario cacao variety, a hybrid of Criollo and Forastero genotypes, demonstrates significant genetic advantages. This variety combines the unique, superior flavor characteristics of Criollo with the disease resistance of Forastero, significantly enhancing its productivity and adaptability (Ascrizzi et al., 2017). Through this hybridization strategy, Trinitario cacao not only meets high-quality standards but also demonstrates greater sustainability and stability in agricultural production. These superior varieties hold significant value in the high-end chocolate market and are instrumental in the development of new products for the future.

1.2 Current understanding of genetic variation in cacao

SSR markers have revealed that the local cacao varieties on the island of Sulawesi in

Indonesia are primarily the result of hybridization between Trinitario and Forastero (Zhang & Motilal., 2016). Among them, Trinidad is the most common variety in this region, accounting for 51% of the total variation (Zhang & Motilal., 2016). Subsequently, an analysis of cacao germplasm from six farms in Côte d'Ivoire, using 12 microsatellite markers, indicated that the majority of cacao originated from Amazonian hybrid varieties: Amelonado and Upper Amazon (Pokou et al., 2009). Similarly, in Cameroon, 25.5% of cacao was closely related to Amelonado (Efombagn et al., 2006). Moreover, Amelonado has shown strong resistance to black rot disease (Pokou et al., 2009). However, Amelonado lacks resistance to cacao swollen shoot disease, which has facilitated the widespread dissemination of the disease (Toxopeus., 1968).

In the 1960s, the Nigerian hybrid is a third-generation Amazon hybrid from wild Amazonian cacao and has been found to be resistant to cocoa swollen shoot virus disease (Aikpokpodion., 2012). In addition, this third-generation Amazon cacao possesses traits such as high yield, short juvenile period, resistance to the black pod, and the ability to produce year-round. Therefore, it has become the cacao variety with the most potential for improvement in West Africa (Pokou et al., 2009; Aikpokpodion., 2012). Additionally, studies have shown that the Amazon and Amelonado hybrid varieties have increased cacao yield by 42% (Edwin & Masters., 2005).

A survey has revealed that in the West African region (Nigeria, Cameroon, and Ghana), approximately 50 parental genotypes were identified among nearly six million plants across 47 cacao plantations (Zhang & Motilal., 2016). This ratio indicates a relatively low level of cacao biodiversity in the area. In fact, the global cultivation of cacao germplasm is predominantly concentrated on a few varieties such as Amelonado, Criollo, Trinitario, and Forastero so on. The overall genetic diversity of these varieties is generally low, especially when compared with wild cacao germplasm, where the germplasm resources that have been identified and utilized represent only a small fraction of what could be available. A comparative study between wild cacao from varities of Guiana and Forastero demonstrated that the Oxygen Radical Absorbance

Capacity (ORAC) of cocoa “Guiana” (1097 ± 111.8 μM ET/g DM) is 1.3 times higher than that of Forastero (838.5 ± 67.8 μM ET/g DM). However, there is no significant difference in the impact on bioactivity between the two in terms of cacao fermentation. Therefore, it is conceivable to consider promoting “Guiana” wild cacao as an alternative to Forastero or to cross it with other germplasms to cultivate superior varieties (Jean-Marie et al., 2021). In cluster analysis, Criollo was discovered in the pristine forests of Central America. The remaining types are accessible in Central American farms. Individuals of the Nacional cluster group from the Amazonian Andes. The Amelonado cluster includes both wild and cultivated genotypes from various regions, with their origins yet to be precisely determined; however, historical data suggest that they may have been domesticated in the Para River area. And the distribution of these varieties is somewhat linked to human activities. The varieties of Iquitos, Purús, and Marañon are located in the upper Amazon River's tributaries, namely the Solimões River and the Içа River. Although there is an overlap in the geographical distribution of the cocoa varieties Nanay and Iquitos, significant differences in allele frequencies are observed between them (Motamayor et al., 2008).

All in all, the genetic diversity utilized in commercial cocoa production to date seems to represent only a small portion of the cacao gene pool (Zhang & Motilal., 2016). As Stander., (1993) has suggested, breeders need to expand the range of exotic germplasm they employ as a genetic base. Future cacao breeding should fully harness the diversity of wild cacao populations, introducing foreign alleles to address the pressures on cacao due to climate change and pest and disease interference.

The region of Samoa as a consequential cacao cultivation area in the Pacific has a century-long history of cacao farming. However, the collection of cacao germplasm resources and the assessment of genetic diversity have not been adequately completed. It is leading to an obvious deficiency in understanding the population structure and distribution characteristics, thus creating a research gap. To characterize their genetic diversity, this study aims to fill this gap by employing microsatellite marker technology to analyze 183 cacao DNA samples from four different areas in Samoa and comparing

them with 12 different cocoa varieties. Microsatellite markers serving as effective molecular markers are widely used for assessing the genetic diversity of populations and genetic distances among them. The findings of this study are expected to provide a scientific basis for the conservation of cacao germplasm in Samoa and contribute vital information to support the future development of cacao breeding.

2. Methods and materials

2.1 Sampling

Dried cacao leaves were collected from four different cocoa plantations in Samoa (~50 plants per plantation) by colleagues at the Scientific Research Organisation of Samoa (SROS) and local farmers. Two areas were selected on each of Savaii and Upolu Island: Satupaitea (SAT) located on the south coast of Savaii Island (SAT01 - SAT49); Salelologa (SAL) located on Savaii Island (SAL01 - SAL49); Siumu (SIU) located south of Upolu Island (SIU01 - SIU49); Aleipata Itupa i Luga area (ALE) on Upolu Island (ALE01 - ALE46). Visualisation was performed in R studio 4.2.0. (Fig. 1). Samples were desiccated using silica gel. In addition, freeze-dried samples from 12 different cacao varieties were imported from the C.A.T.I.E. germplasm center in Cost Rica (Table Supplement 1) following regulatory guidelines.

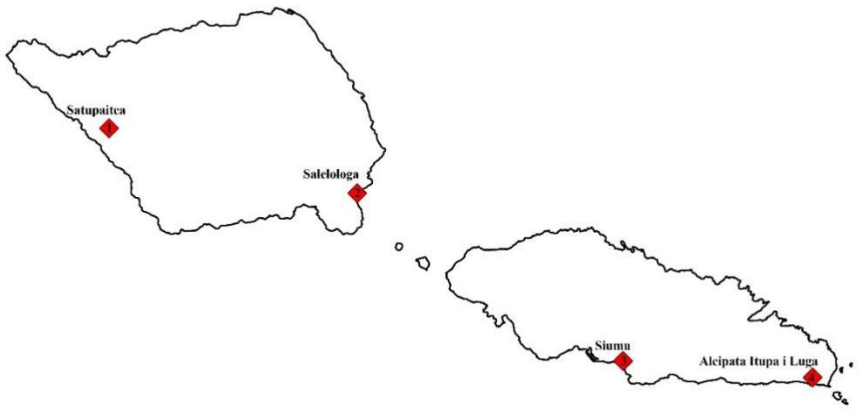

Fig 1: Map of Samoa. The red stars indicate the locations of the study plantations: 1 is Satupaitea (SAT); 2 is Salelologa (SAL); 3 is Siumu (SIU); 4 is Aleipata Itupa i Luga (ALE)

2.2 Selection of microsatellite markers

Based on previous findings on cocoa genetics (Pugh et al., 2004; Dadzie et al., 2013),

six microsatellite markers linked to cocoa black pod rot resistance/susceptibility were selected for assessing the genetic diversity of the Samoa cocoa samples (Table supplement 2).

2.3 Experiments:

2.3.1 DNA extraction

From Samoan samples, DNA was extracted by colleagues at SROS and imported to New Zealand following regulatory guidelines. For the C.A.T.I.E. germplasm samples, DNAs were extracted by the student. In both cases, DNA was extracted from dried cacao leaves of each sample using a modified Qiagen kit protocol (DNeasy® Plant). Eventually, 100 µl of DNA were eluted from each sample. DNA samples (3 µL each) were electrophoresed on a 1.35% agarose gel in TAE buffer at 75 V for 45 minutes. The samples were mixed with 3x loading dye prior to loading onto the gel. Visualization was achieved using ethidium bromide staining and subsequent imaging under ultraviolet light.

2.3.2 PCR assays and gel electrophoresis

The six microsatellite loci were screened on 190 cacao samples from the four different regions of Samoa and 12 different cacao varieties. Polymerase chain reaction (PCR) was used to amplify the microsatellite loci of DNA samples at a volume of 10µL. The specific reagent system included: 3.7µl ddH2O, 1.0µl of 10x Reaction Buffer BD (Solis Biodyne), 1.0µl of 2.5 µM $MgCL_2$, 1.0µl of 0.2µM Forward primer, 1.0µl of 4.5µM Reverse primer, 1.0µl of 4.5 µM M13 primer (FAM, VIC or NED), 0.2µl of 10µM dNTP, 0.1µl of 5U/µl FIREPol DNA polymerase (Solis Biodyne) and 1 µL of 1:10 diluted genomic DNA template was added to each sample. The PCR amplification procedure was as follows: The thermal lid temperature of the PCR instrument was set at 95°C; denaturation was performed at 94°C for 30 seconds; annealing at 46 °C or 51 °C for 1 minute, details are in table 2; extension at 72 °C for 1minute, cycle 35 times, and preservation at 4°C.

2.3.3 Electrophoresis detection

To evaluate the outcome of PCR amplification, PCR products were randomly sampled by electrophoresis on 1.35% agarose gels stained with ethidium bromide. 2 µl of PCR products were sampled and run on agarose gel at 75V for 45 minutes. The GeneRuler 1kb Plus DNA ladder (Thermo Scientific) was used for fragment analysis.

2.3.4 Genotyping Process

CASS size standard was added to 9µl Hi-Di formamide (Applied Biosystems by Thermo Fisher Scientific, UK) and 1.25µl PCR products (Symonds & Lloyd., 2004). Samples underwent sizing and separation via capillary electrophoresis using ABI 3730 and the ABI 3730XL DNA Analyzer at the Massey Genome Service sequencing and genotyping facility at Massey University in Palmerston North, New Zealand.

2.4 Data Analysis

Genotype data were manually scored using the Thermofisher Scientific application (https://apps.thermofisher.com/apps/spa/#/dashboard) to determine genotypes. Individuals with three or more loci of missing data (seven individuals) were removed from the analysis, and 183 individuals were selected for further analysis. GenAlEx v6.51 was used to calculate the average number of Alleles (Na), the effective number of alleles (Ne) and the Observed Heterozygosity (Ho), Expected Heterozygosity (He), Gene Flow (Nm), Information Index (I), F-Statistics: coefficient of population Differentiation (F_{ST}), inbreeding coefficient within a population (F_{IS}), measures the degree of inbreeding within the population (F_{IT}) (Peakall & Smouse., 2006).

The provesti method was used to examine individual genetic variation by using poppr and phangorn packages in the R studio 4.2.0. Principal coordinate analysis (PCoA) from the Pairwise Distance Matrix relationship between cultivars and germplasm was run and formatted using the R package ggplot2.Structure 2.3.4 was used to identify genetic groups (Pritchard et al., 2000). The number of burn-in iterations was set to 10,000 and the number of data collection iterations to 100,000. Structure Selector was used to calculate $\triangle K$ and *LnP (D)* to identify informative cluster numbers (*K* value). A model-

based Bayesian clustering algorithm was used to identify allele frequencies. An individual was placed in *K* clusters and the membership coefficients of all *K* clusters were summed to 1. The number of genetic groups (i.e. *K* values) were set from 1-5, with 15 replicates run for each *K* value and visualized using the CLUMPAK server. A Neighbor-net split graph was constructed by using the individual-level pairwise genetic differences through SplitsTree v.4.19.2.0.

Finally, Nei's I Similarity Matrix method was used for cluster analysis of only 12 cocoa varieties by building UPGAM trees using the phangorn package of R studio 4.2.0 for graphing.

3. Results

3.1 Features of SSRs and genetic diversity

A total of 190 cacao samples from four different locations were genotyped using six microsatellite loci. After excluding individuals with three or more missing microsatellite loci, a total of 183 samples were confirmed, for which there was just 3.68% missing data (Table supplement 3). The polymorphism percentage across the six pairs of SSR primers in the four regions was 100%, identifying a total of 108 alleles. The number of alleles per marker ranged from three to seven (Table 1). Notably, the ALE region displayed the highest number of alleles at marker mTcCIR280. The SAL region exhibited the highest number of alleles for markers mTcCIR200. Both SAT and SIU regions showed the highest number of alleles at mTcCIR61 and mTcCIR200, respectively. Expected heterozygosity (He) ranged from 0.10 to 0.66, with the highest and lowest values observed in the SIU region at markers mTcCIR168 and mTcCIR61, respectively. Observed heterozygosity (Ho) ranged from 0.11 to 0.81, with the lowest and highest values found in the SIU region at marker mTcCIR168 and SAL region at marker mTcCIR200, respectively. Overall data indicated that expected heterozygosity (He) was higher than observed heterozygosity (Ho) in the ALE region. However, the opposite was observed in the SAT and SIU regions, with the Ho value being slightly higher. Further statistical analysis of the four regions revealed an average number of

alleles (Na) for each group was as follows: 5 for ALT, 4.83 for SAL, 3.5 for SAT, and 4.67 for SIU (Table 1). Four private alleles were detected, with two in each of the ALE and SAL regions (Table 2).

Table 1. Population genetic metrics for the six microsatellite loci in *Theobroma cacao* L. samples from four regions.

Locus	ALE				SAL				SAT				SIU			
	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He
N61	42.00	5.00	0.31	0.46	45.00	5.00	0.60	0.64	42.00	5.00	0.50	0.54	45.00	6.00	0.51	0.66
F168	43.00	4.00	0.42	0.49	47.00	4.00	0.47	0.56	44.00	2.00	0.14	0.20	47.00	4.00	0.11	0.10
F200	43.00	5.00	0.56	0.55	43.00	6.00	0.81	0.62	44.00	4.00	0.70	0.56	39.00	6.00	0.49	0.47
V225	42.00	6.00	0.43	0.49	45.00	5.00	0.60	0.53	42.00	3.00	0.43	0.41	46.00	6.00	0.61	0.60
N236	42.00	3.00	0.40	0.46	47.00	4.00	0.60	0.52	45.00	3.00	0.44	0.35	46.00	3.00	0.54	0.51
V280	43.00	7.00	0.49	0.57	44.00	5.00	0.50	0.55	44.00	4.00	0.39	0.32	47.00	3.00	0.57	0.51
Mean	42.50	5.00	0.44	0.50	45.17	4.83	0.60	0.57	43.50	3.50	0.43	0.40	45.00	4.67	0.47	0.47

The columns are: sample number (N), allele number (NA), observed heterozygosity (Ho), expected heterozygosity (He). * Significant deviation from HWE (P< 0.01)

Table 2. The size of alleles at each microsatellite locus detected in each of the four collection regions.

Locus	Common alleles	N	Four regions			
			ALE	SAL	SAT	SIU
N61	202	7	202	202	202	202
	204		204		204	204
	206		206	206		
	214		214	214	214	214
	216		216	216	216	216
	218		218	218		218
	224				224	224
F168	192	4	192	192	192	192
	194		194	194	194	194
	196		196	196		196
	198		198	198		198
F200	250	7	250			250
	255			255		255
	257		257	257	257	257
	327			327		
	329		329	329	329	329
	331		331	331	331	331
	334		334	334	334	334
V225	320	6	320	320	320	320
	322		322	322	322	322
	332		332	332	332	332
	338		338			338
	340		340	340	340	340
	346		346	346	346	346
N236	195	5	195		195	
	197			197		
	199		199	199	199	199
	209		209	209	209	209
	215			215		215
V280	104	7	104	104		
	106		106	106	106	
	108		108	108	108	108
	112		112	112	112	112
	114		114			
	131		131			
	133		133	133	133	133

According to Table 3, The SAT region had the lowest average effective number of alleles (Ne) at 1.74, while SAL had the highest at 2.35. the ALE region had an average observed heterozygosity (Ho) of 0.43 (SE=0.03) and an average expected

heterozygosity of 0.50 (SE=0.02), while at SAL, both Ho and He were the highest at 0.6 (SE=0.05) and 0.57 (SE=0.02), respectively. The SAT region had the lowest values at 0.43 (SE=0.07) for Ho and 0.4 (SE=0.06) for He. The mean values of He and Ho in the SIU area are both 0.47 (SE=0.08). The Shannon's Information Index (I) average in the SAL region exceeded 1, indicating a higher level of genetic diversity in this area. The fixation index ranged from -0.06 to 0.14. Three Fis values less than zero indicate a slight excess of heterozygotes in the cacao locations of SAL, SAT, and SIU regions, but there was a slight excess of homozygotes in the ALE region (Table 3).

Table 3. Population genetic statistics of by location.

ID	N	%P	Pa		Na	Ne	I	Ho	He	Fis
ALE	44	100%	2	**Mean**	5.00	2.02	0.94	0.43	0.50	0.14
				SE	0.58	0.08	0.06	0.03	0.02	0.05
SAL	47	100%	2	**Mean**	4.83	2.35	1.07	0.60	0.57	-0.05
				SE	0.31	0.12	0.06	0.05	0.02	0.08
SAT	45	100%	0	**Mean**	3.50	1.74	0.69	0.43	0.40	-0.06
				SE	0.43	0.17	0.10	0.07	0.06	0.09
SIU	47	100%	0	**Mean**	4.67	2.08	0.88	0.47	0.47	-0.01
				SE	0.61	0.25	0.14	0.08	0.08	0.05

Note: number of samples (N), number of alleles (Na), number of effective alleles (Ne), number of private alleles (Pa), percentage of polymorphic loci (%P), Shannon's Information Index (I), observed heterozygosity (Ho), expected heterozygosity (He), and fixation index (Fis).

Pairwise F_{ST} values between ALE and SAL populations were 0.01, and between SAT and ALE and SAL populations were 0.02 and 0.03, respectively. The largest F_{ST} difference was observed between SIU and SAL at 0.04, and the paired F_{ST} values for SIU and ALE and SIU and SAT were consistent at 0.03 (Table 4).

Table 4. Pairwise population F_{ST} values of four sites.

Site	ALE	SAL	SAT	SIU
ALE	0.00			
SAL	0.01	0.00		
SAT	0.02	0.03	0.00	
SIU	0.03	0.04	0.03	0.00

Along with the Samoa cacao samples, the six SSR markers also were used for genotyping 12 cacao varieties. A total of 94 alleles were identified across these varieties. Each locus revealed just one to two alleles. Varieties such as Contamana, Criollo, Curaray, Guiana, Iquiotos, Maranon, Purus, and Trinitario were successfully amplified

for all six markers. Amelonado was missing data at the mTcCIR61 and mTcCIR200 loci. LAFI-7 lacked genotype information at mTcCIR168, mTcCIR225, and mTcCIR236, while Nacional and Nanay did not produce alleles at mTcCIR61 and mTcCIR236 (Table 5).

Table 5. Population genetic metrics for the six microsatellite loci in 12 samples of *Theobroma cacao* L.

Locus	Amelonado				Contamana				Criollo				Curaray			
	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He
N61	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
F168	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
F200	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
V225	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
N236	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
V280	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50
Mean	0.67	0.67	0.00	0.00	1.00	1.50	0.50	0.25	1.00	1.17	0.17	0.08	1.00	1.67	0.67	0.33

Locus	Guiana				Iquiotos				Maranon				Nacional			
	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He
N61	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	0.00	0.00	0.00	0.00
F168	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
F200	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
V225	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00
N236	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00
V280	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00
Mean	1.00	1.33	0.33	0.17	1.00	1.83	0.83	0.42	1.00	1.67	0.67	0.33	0.67	0.83	0.17	0.08

Locus	Nanay				LAFI-7				Purus				Trinitario			
	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He	N	Na	Ho	He
N61	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
F168	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50
F200	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50
V225	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	2.00	1.00	0.50
N236	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00
V280	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	2.00	1.00	0.50	1.00	1.00	0.00	0.00
Mean	0.83	1.00	0.17	0.08	0.50	0.67	0.17	0.08	1.00	1.67	0.67	0.33	1.00	1.67	0.67	0.33

The columns are: sample number (N), allele number (NA), observed heterozygosity (Ho), expected heterozygosity (He). * Significant deviation from HWE (P< 0.01)

Iquiotos had the highest average number of alleles at 1.83, and Amelonado and LAFI-7 had the lowest average at 0.67 across the six markers. The effective number of alleles (Ne) was consistent with the Na results. The mean values of Shannon's Information Index (I) were less than 1 for all varieties. Amelonado did not show any heterozygosity across the six markers. Iquiotos had the highest observed (Ho) and expected (He) heterozygosity values, which were 0.83 (SE=0.17) and 0.42 (SE=0.08), respectively. Criollo, Nacional, Nanay, and LAFI-7 had the lowest values at 0.17 (SE=0.17) and 0.08 (SE=0.08) (Table 6). Table supplement 4 presents the observation and genotypes of specific alleles amplified at each locus across the 12 varieties and marking private allele in red. A total of 21 private alleles and 21 common alleles were amplified by six pairs of primers. Criollo and Curare each have three private alleles. No private alleles were found in Amelonado and LAFI-7 (Table supplement 4).

Table 6. Population genetic statistics of by location.

ID	N	%P	Pa		Na	Ne	I	Ho	He	Fis
Amelonado	1	0.00%	0	**Mean**	0.67	0.67	0.00	0.00	0.00	—
				SE	0.21	0.21	0.00	0.00	0.00	—
Contamana	1	50.00%	1	**Mean**	1.50	1.50	0.35	0.50	0.25	—
				SE	0.22	0.22	0.16	0.22	0.11	—
Criollo	1	16.67%	3	**Mean**	1.17	1.17	0.12	0.17	0.08	—
				SE	0.17	0.17	0.12	0.17	0.08	—
Curaray	1	66.67%	3	**Mean**	1.67	1.67	0.46	0.67	0.33	—
				SE	0.21	0.21	0.15	0.21	0.11	—
Guiana	1	33.33%	2	**Mean**	1.33	1.33	0.23	0.33	0.17	—
				SE	0.21	0.21	0.15	0.21	0.11	—
Iquiotos	1	83.33%	2	**Mean**	1.83	1.83	0.58	0.83	0.42	—
				SE	0.17	0.17	0.12	0.17	0.08	—
Maranon	1	66.67%	2	**Mean**	1.67	1.67	0.46	0.67	0.33	—
				SE	0.21	0.21	0.15	0.21	0.11	—
Nacional	1	16.67%	1	**Mean**	0.83	0.83	0.12	0.17	0.08	—
				SE	0.31	0.31	0.12	0.17	0.08	—
Nanay	1	16.67%	2	**Mean**	1.00	1.00	0.12	0.17	0.08	—
				SE	0.26	0.26	0.12	0.17	0.08	—
LAFI-7	1	16.67%	0	**Mean**	0.67	0.67	0.12	0.17	0.08	—
				SE	0.33	0.33	0.12	0.17	0.08	—
Purus	1	66.67%	3	**Mean**	1.67	1.67	0.46	0.67	0.33	—
				SE	0.21	0.21	0.15	0.21	0.11	—
Trinitario	1	66.67%	2	**Mean**	1.67	1.67	0.46	0.67	0.33	—
				SE	0.21	0.21	0.15	0.21	0.11	—

Note: number of samples (N), number of alleles (Na), number of effective alleles (Ne), number of private alleles (Pa), percentage of polymorphic loci (%P), Shannon's Information Index (I), observed heterozygosity (Ho), expected heterozygosity (He), and fixation index (Fis).

3.2 Genetic structure of *Theobroma cacao* L.

The genetic population structure of *T. cacao* was analyzed by cluster analysis using the program STRUCTURE based on a Bayesian inference approach on the microsatellite marker data. In addition, ΔK statistics were used to determine the optimal K value to more accurately assess the degree of differentiation of the genetic population based on Evanno et al. (2005). Structure Selector had the greatest support for K=2 with a ΔK value of 55.95. ΔK value was 39.87 for K=3. In contrast, K=4 (5.09) had a very low ΔK value. At K=5, the low ΔK value was not of practical significance (Fig. 2, Table 7). The Evanno method was not able to determine the ΔK value at K=1.

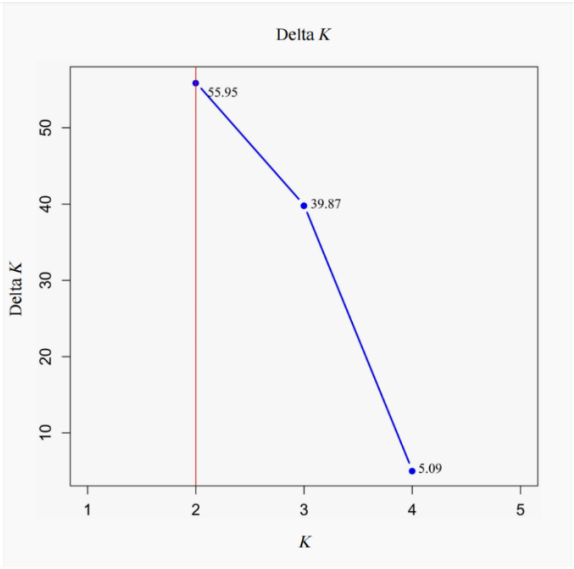

Fig 2: Δ*K* output from the Evanno method (Evanno et al., 2005) vs *K* STRUCTURE outputs based on 15 replicates at *K*=2, *K*=3 and *K*=4. This method suggests that there is structure detected at *K*=2 and *K*=3.

Table 7. Evanno method output from *K*=1-5 STRUCTURE runs.

K	Reps	Mean *LnP(K)*	Stdev *LnP(K)*	*Ln'(K)*	|*Ln''(K)*|	Delta-*K*
1	15	-2380.72	0.07	—	—	—
2	15	-2230.18	1.75	150.54	97.91	55.95
3	15	-2177.55	1.79	52.63	71.39	39.87
4	15	-2196.30	5.27	-18.75	26.83	5.09
5	15	-2241.89	33.69	-45.59	—	—

Note: — means NA

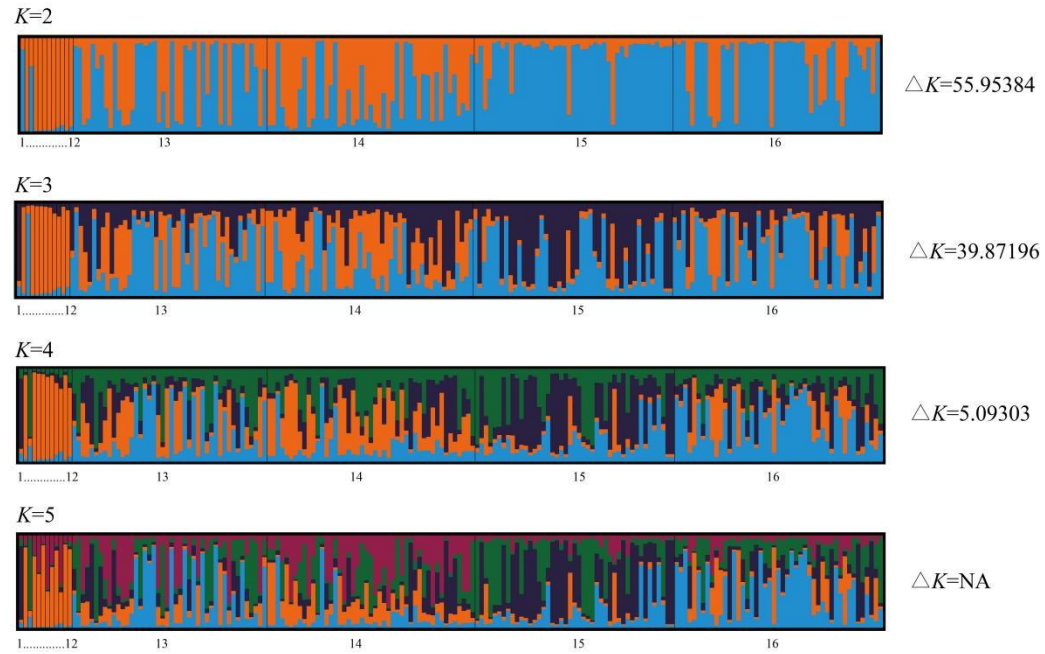

Fig 3: STRUCTURE output of *Theobroma cacao* L. dataset Individuals assigned to genetic clusters at *K*=2, *K*=3, *K*=4 and *K*=5. 1-12 are different cacao varieties. Individuals are segmented with black vertical bars. 1: Amelonado; 2: Contamana; 3: Criollo; 4: Curaray; 5: Guiana; 6: Iquiotos; 7: Maranon; 8: Nacional; 9: Nanay; 10: LAFI-7; 11: Purus; 12. Trinitario; 13: ALE; 14: SAL; 15: SAT; 16: SIU.

When divided into two clusters (*K*=2), one group (in blue) contains Amelonado and Criollo. The other group (in orange) included the remaining 10 cacao varieties (Contamana, Curaray, Guiana, Iquiotos, Maranon, Nacional, Nanay, LAFI-7, Purus). In the bar chart with *K*=2, the cacao varieties in ALE and SAL plantations are more similar to each other, and the SAT and SIU plantation areas exhibit greater similarity. Especially, the ALE and SAL populations showed a high degree of mixing with about 50 percent of each cluster. The Amelonado and Criollo continued to show a high degree of divergence from the other varieties at *K*=3. The SAT region showed greater affinity with Amelonado. Plants from the SIU region showed greater genetic similarity with Criollo. The results of the structure analyses showed that at *K*=4, the populations in the ALE and SAL regions showed significant genetic mixing. In the ALE and SIU regions, the cacao populations that used to be genetically more similar to Criollo varieties have diverged significantly. Approximately two-thirds of the individuals in the SAT region resemble the Amelonado variety, while a smaller number show similarity to the Criollo variety. In addition, three individuals in the SAT region (SAT38, SAT40, and SAT42) shared similar genetic characteristics with about two-thirds of the individuals in the SIU region, and the proportion of blue color in each of these individuals exceeded 50%. In the *K*=5 structure analysis, the color composition of samples Maranon and LAFI-7 in the bar plot significantly differed from the other samples, which had more than 50% of their genetic components in dark red. Thus, Maranon and LAFI-7 were separated from the previous genetic cluster classification. Meanwhile, at least 50% of the individuals in the SAL region were classified into the Maranon and LAFI-7 genetic clusters. In addition, individuals with a large proportion of blue color in the four areas missed matching with any of the 12 cacao varieties, showing unique genetic characteristics (Fig. 3).

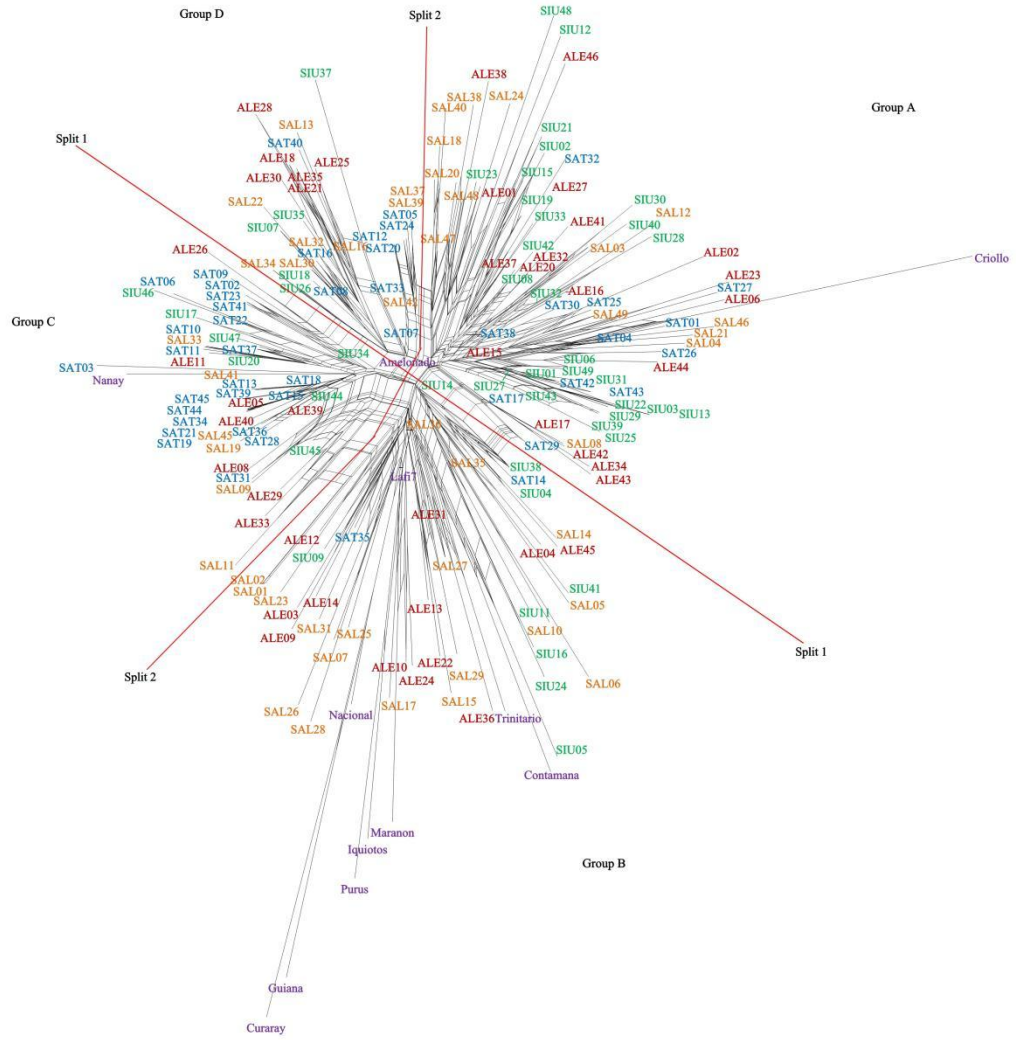

Fig 4: Individual-based Neighbor-Net analyses of *Theobroma cacao* L. populations. Different plantations are labelled with different colors. Red represents ALE cacao germplasm, yellow represents samples collected in SAL, blue is SAT, green represents SIU. Purple marks the different varieties of cacao. The red Split 1 and Split 2 divide the diagram into four groups A, B, C, D.

Using Neighbor-net analysis, 12 cacao varieties and cacao germplasm from four different regions were analyzed for comparative genetic distances (Fig. 4). The nodes are marked with different colors in the figure, clearly revealing the genetic relationships and differences among them. These nodes showed an overall radial distribution pattern. According to split 1, the figure is significantly divided into two parts, extending from the upper left to the lower right of the figure. This split divides the cacao varieties roughly into two major groups: one represented by Criollo with 100 samples and the

other extending outwards with Amelonado as the center point with 83 samples. Further observation reveals that the 12 cacao varieties are divided into four main sections, based on the two splits. Amelonado and LAFI-7 are located at the center of the neighbor-net diagram, which is probably due to missing data, while Criollo is located on the right side of the neighbor-net diagram, whereas Nanay is on the left side. The remaining cacao varieties exhibit relatively small genetic distances and are primarily distributed in the lower section of the neighbor-net diagram.

It is further divided into four parts according to split 2. As shown in the figure, group A belongs to Criollo with a total of 71 samples, containing 18 samples from the ALE plantation area, 12 from the SAT plantation area, 14 from the SAL and 27 from the SIU. group B has a total of 40 samples, i.e. ALE (12), SAT (2), SAL (18), SIU (8). And Group B was associated with nine varieties of cacao. Group C was genetically similar to Nanay with 43 cacao samples, specifically eight cacao samples from ALE plantation, 22 samples from SAT plantation, and six and seven samples from SAL and SIU, respectively. The 29 cacao samples from Group D (i.e. ALE (6), SAL (9), SAT (9), SIU (5) were genetically distant from the 12 cacao varieties and this occurrence is highly likely to be due to missing data for the time being. Therefore, further research is needed.

In terms of the overall distribution, cacao varieties in the SAT region are mainly clustered around the Nanay branch, but a few individuals are scattered around the Criollo; conversely, cacao in the SIU region is mainly situated in the same direction of the branch as the Criollo but closer to the center point. In contrast, the distribution of cacao species in the ALE and SAL regions is more dispersed. It is noteworthy that cocoa in the SAT region is mainly distributed in the left half of the figure, while cocoa samples in the SIU area are mainly distributed in the right half of the figure. Of particular interest is the fact that the cacao varieties in the ALE and SAL regions show a high degree of genetic similarity, further underscoring their similarity in genetic structure (Fig. 4).

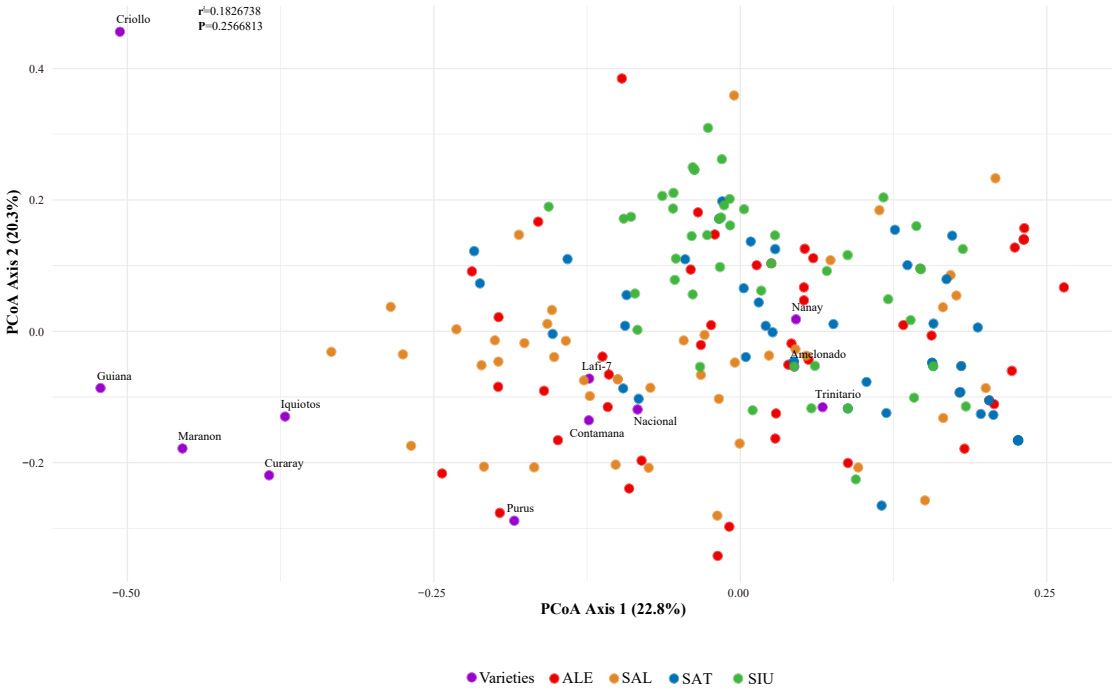

Fig 5: PCoA analysis plot. The coloration represents the varieties and the geographic locations of the populations.

Using Principal Coordinates Analysis (PCoA), the genetic distances of cacao germplasm from the SIU, SAT, ALE, and SAL regions were analyzed. The results indicated that cacao germplasm from the SIU region had the most concentrated distribution. It indicates high genetic similarity among the samples. This was followed by the SAT region, while the genetic distances of cacao germplasm in the ALE and SAL regions were relatively dispersed. Further analyses revealed that Amelonado, Trinitario, and Nanay are more closely related genetically to the SAT and SIU plantations. LAFI-7, Nacional, Contamana are exhibited greater genetic proximity to and ALE and SAL plantations. Notably, the Purus variety appears at the bottom of the graph. The Iquitos, Curaray, Maranon, and Guiana varieties are clustered in the lower left area of the graph. Criollo shows the greatest genetic difference among all varieties. Overall, the 12 cacao varieties can be broadly classified into five genetic groups. The two principal axes of PCoA explained 22.8% and 20.3% of the genetic variation, respectively.

Collectively, the outcomes from STRUCTURE analysis, Neighbor-net construction, and Principal Coordinates Analysis (PCoA) have delineated several consistent genetic patterns among the cacao varieties under study. Most notably, the Criollo variety has

significant genetic divergence from the remaining 11 varieties. Secondly, the genetic materials from the ALE and SAL regions are characterized by a more dispersed distribution, whereas the SIU region exhibits the most concentrated distribution of cacao varieties, followed closely by the SAT region. Furthermore, the Neighbor-net and PCoA plots visualizations suggest a closer genetic proximity between the SAT region and the Nanay variety. In summary, despite these four regions being geographically relatively close, they harbor different genetic variation.

3.3 Cluster analysis of *Theobroma cacao* L.

Based on the genetic similarity (Nei's Genetic Identity) among the 12 cacao populations, they were analyzed by clustering using the UPGMA method (Fig 6). The clusters were clustered in the range of 0.00 to 0.63. It can be seen that two main groups, Criollo and the remaining varieties. It can further be divided varieties into four main clusters: Amelonado, Trinitario, Nanay; Iquiotos, Contamana, Purus, LAFI-7, Maranon; Guiana, Curaray, Nacional; Criollo.

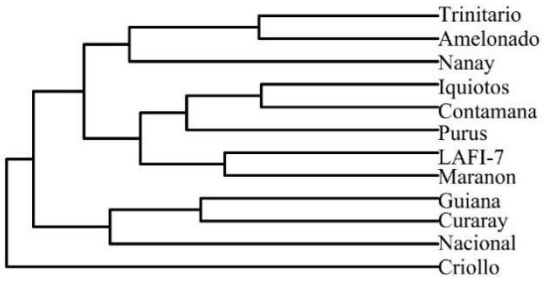

Fig 6: UPGMA Tree Based on Nei's I Similarity Matrix.

4. Discussion

4.1 Cacao genetic diversity analyses

The analyses showed that the overall level of genetic diversity in the cacao samples from the four regions was high. In the heterozygosity analysis, the average observed heterozygosity (Ho) (0.472) in the SIU region was close to the expected heterozygosity (He) (0.474). It indicated that this region was near the Hardy-Weinberg Equilibrium, (HWE). In the ALE region, Ho was lower than He indicating that either inbreeding has occurred in the region or that it has recently experienced events such as selection

pressure, genetic drift, and bottleneck effect (Spieth., 1979). Both SAL and SAT regions showed a different observation; i.e. observed heterozygosity was significantly higher than expected heterozygosity. Suggesting that levels of heterozygosity are above the mutation-drift equilibrium, which is conducive to increasing genetic diversity (Delph & Kelly, 2014).

Additionally, sample quantity can influence the determination of heterozygous and homozygous numbers, potentially causing genetic drift (Lynch et al., 2016). Moreover, in conjunction with other data indicates that the genetic diversity of cacao varieties in the ALE plantation is relatively high. It is hypothesized that if a population receives gene flow from another population with a higher frequency of homozygous resistance genes, the number of homozygous individuals could also increase (Balloux et al., 2004). At the same time, this does not exclude the possibility that the cacao from the plantation may be influenced by other plants in the vicinity of the plantation. For example, the presence of wild cacao around the plantation. Furthermore, Based on number of homozygotes, it is speculated that this plantation recently experienced a large-scale outbreak of black pod rot. Under the selection pressure of Black pod disease, cacao with disease resistant genotypes in plantations may be favored. Such directed selection eventually leads to an increase in the proportion of homozygous resistant individuals (Grant et al., 1995). Further exploration of other factors is needed.

The observed heterozygosity in the SAL and SAT regions is higher than the expected heterozygosity, indicating heterozygote advantage. It suggests that under natural selection, heterozygous individuals may have a higher degree of fitness than homozygous individuals. However, another perspective posits that heterosis is more often caused by pseudo-overdominance rather than true heterozygote advantage (Birchler et al., 2006). This has been validated in maize genetics (Moll & Robinson, 1964). In addition, gene flow and other factors can all lead to an increase in the amount of heterozygosity (Star & Spencer., 2013). Therefore, it is difficult to accurately determine whether there is a true heterozygote advantage. Consequently, it is not appropriate to solely associate heterozygosity with diversity.

Meanwhile, the F_{IS} coefficient for the ALE region was greater than 0.1 (0.139), indicating an excess of homozygotes and a high level of homozygotes in this plantation. The data also illustrates that the ALE region has the highest number of alleles (30) among the six markers. In contrast, the F_{IS} values for the remaining three districts tend toward 0, indicating that these regions are undergoing random mating and sharing genetic material to a greater extent, hence allowing for a broader spread of genetic diversity. This is consistent with the data from cacao in central and western Uganda. In central Uganda, through 96 single nucleotide polymorphism markers, the expected heterozygosity of cacao was obtained as 0.334 and the observed heterozygosity was 0.295 i.e. F_{IS}=0.117, whereas cocoa in western Uganda, the Ho was 0.317 and the He was 0.322 i.e. F_{IS}=0.016. The results from Uganda indicate that inbreeding occurs in the central region and that the genetic material is more abundant in the western region (Gopaulchan et al., 2019). The Shannon's Information Index (I) in the SAL region is greater than 1, indicating moderate to high species diversity in this area. Due to the fact that an increase in homozygosity does not necessarily lead to inbreeding depression (Pilon et al., 2021), combined with the value of 0.939 in the ALE region approaches 1. Therefore, it can be inferred that inbreeding depression has not occurred in this area.

F_{ST} is a commonly used statistical method in diversity analysis (Meirmans et al., 2011). The statistical analysis of F_{ST} values in this study showed that between all four regions are all less than 0.05. This indicate that the genetic variation of cacao between the four regions is low.

4.2 Genotype analysis

Although the genotypes of cacao in the four regions overall appear highly homogeneous, further analysis reveals that even within the same region, genetic differentiation among populations exists. Studies on the genotypes of cacao within the four regions indicate that these populations cannot be clearly categorized as a single variety. According to the results, the ALE region exhibits a broad range of genetic variation, with alleles associated with varieties of Guiana, Nacional, LAFI-7 (114 bp), and private alleles associated with Criollo (131 bp) were identified. Additionally, genotypes of

Curaray and Purus also overlap in the ALE region. In the SAL region, a private allele of Guiana variety (327 bp) was discovered, along with an allele (197 bp) that could not be matched with any variety. The cacao varieties in the SAT region are relatively concentrated, and alleles related to Nanay are also present. In addition, the SIU region is aligned with Criollo in the neighbor-net diagram, but in the PCoA plot, it is shown to be concentrated near Nanay. The alleles show associations with Nanay, Purus, and Marañon. Overall, alleles from all four regions intersect with Contamana, Marañon, Criollo, and Guiana varieties. This finding provides new insights into the diversity of cacao germplasm and its genetic history in the region.

Furthermore, Wright argued that genes that are self-incompatible are more susceptible to Negative Frequency Dependence and noted that strong natural selection can increase the frequency of private alleles (Wright, 1939). Rare superior private alleles help to maintain polymorphism in plants and even do well in resisting diseases (Haldane, 1992). In this study, a total of 21 private alleles were found. These private alleles probably be a mutation that favors biological evolution. However, the private allele (197bp) found in SAL plantations was not matched in the 12 cacao varieties. Both points need to be further verified in the future.

4.3 Phylogenetic Analysis

In this research, 183 cacao samples from Samoa and 12 cacao varieties were genetically analyzed and grouped. By integrating bar plot, Neighbor-net, and PCoA diagrams, it was preliminarily determined that the cacao varieties in the Samoa region consists of Amelonado, Trinitario, Nanay, Nacional, Contamana, Purus and LAFI-7. In the Honduras and Nicaragua regions, cacao varieties is mainly divided into five categories: Amelonado, Trinitario, Upper Amazon Forastero, and Criollo (Ji et al., 2013). The cacao variety in Ecuador is commonly referred to as Nacional (Loor et al., 2009). In central and western Uganda, most of the cacao varieties is Marañon, with a few Amelonado and Iquitos (Gopaulchan et al., 2019).

A UPGMA rooted tree was constructed using Nei's genetic distance matrix to delineate

genetic relationships. The results were divided varieties into three groups: Amelonado includes Trinitario, Nanay, Iquitos, Contamana, Purus, LAFI-7, and Maranon; Guiana, Curaray, and Nacional; and Criollo. This is similar to the cacao diversity study on the Puerto Rico Island, which also divided Criollo, Amelonado, and Nacional into three clusters and confirmed that the Criollo variety has the greatest genetic distance (Cosme et al., 2016).

In origin analyses, Nacional, Contamana, and Purus are most closely related to wild populations and are commonly found in the southern part of the Amazon region in Peru (Thomas et al., 2012). As populations migrated northward across the Andes, only a small ancestral population arrived. This early domestication process experienced a bottleneck event, which reduced genotypic diversity and ultimately resulted in smaller differences between these populations and human-selected varieties (Clement et al., 2010). Additionally, due to their adaptation to the environment of the rainforest, they are considered semi-domesticated crops (Clement, 1999). The origin centers of Iquitos and Purus are both located in the northwest of Brazil in the upper Amazon River region. In some studies, Iquitos have been considered to be grouped with Amelonado or to form a separate cluster (Thomas et al., 2012). The earliest discovery of Purus can be precisely traced to the State of Acre, Brazil (Sereno et al., 2006). Nanay originates from Peru and is a smaller cluster of cacaos. Studies have found that Nanay is closely related to Iquitos and Purus (Thomas et al., 2012), and it is hypothesized that they may have originated from the same gene pool, but with significant differences in gene frequencies (Motamayor et al., 2008). The Curaray population appears in the Amazon basin of Ecuador and can be traced back to 20,000 years ago or even earlier. Despite its long history, surprisingly, the distribution of this population seems to be limited to specific areas and has not spread widely throughout the Amazon region (Thomas et al., 2012). In this study, Neighbor-net and PCoA plots showed that cacao germplasm from the four plantations of Samoa is genetically distant from Curaray variety. Although F_{ST} values suggest some genetic association between them, analyses based on genetic distance do not support this connection. Marañon comes from the State of Rondônia, Brazil, and is

a population with a bias towards homozygosity. Its low heterozygosity and morphological consistency show significant signs of artificial selection. A similar situation also occurs in the Guiana variety (Thomas et al., 2012). Criollo is considered one of the earliest cacao varieties domesticated by humans and is mainly distributed in Central America. Although its dispersal path is still controversial, it is certain that Criollo cacao is found in large quantities in both Venezuela and southern Mexico (Motamayor & Lanaud, 2002; Clement et al., 2010; Whitkus et al., 1998).

The LAFI-7 variety, a cacao variety with excellent disease resistance selected by the Samoan government, especially for Western Samoa, has demonstrated significant resistance to the black pod, pink disease, root rot, and canker (Kelly, 1969). In the present study, the molecular markers of the four regional samples had alleles corresponding to the primer markers of the LAFI-7 variety, particularly the allele (114) at the mTcCIR280 locus in the ALE region, which showed uniqueness. Although the samples from the four regions showed some degree of similarity in genetic characteristics to the LAFI-7 variety, they do not exhibit significant genetic differences compared to other varieties, which may be related to the missing data for LAFI-7 in three markers, which could affect the accuracy of the results and introduce errors. Therefore, it is recommended to expand the sample size in future studies to ensure the reliability and accuracy of the results. Additionally, the genetic structure of LAFI-7 and Marañon was found to be relatively similar in the bar plot. This paper provides a reference for the genetic relationship between them, but further identification is still needed in the future. In general, the phylogenetic analysis results are consistent with previous studies, but they also reveal some new genetic relationships, which are of great significance for the protection, improvement, and dispersal of future cacao germplasm resources.

5. Conclusion

To sum up, this study has analyzed 195 cacao DNA samples from four different regions in Samoa by using microsatellite markers. The analysis has focused on the polymorphism of loci, determined the germplasm diversity, the causes affecting the

diversity and the formation of hypotheses. The results have identified the Samoa varieties as Amelonado, Trinitario, Nanay, Nacional, Contamana, Purus and LAFI-7. There still have gaps in the research on the association between microsatellite markers and cacao black rot, and further investigation is needed to determine the specific factors influencing cacao diversity in Samoa. In the future, future studies could explore the cultivation and disease status of cacao in the four regions of Samoa, as well as the impact of climate change and soil types on cacao quality and yield. Expanding the sample size and improving molecular marker methods are also recommended.

6. Reference List

Aikpokpodion, P. O. (2012). Defining genetic diversity in the chocolate tree (*Theobroma cacao* L.) grown in West and Central Africa. In M. Caliskan (Ed.), *Genetic diversity in plants* (pp. 185–212). Rijeka, Croatia: InTech.

Balloux, F., Amos, W., & Coulson, T. (2004). Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, *13*(10), 3021–3031.

Birchler, J. A., Yao, H., & Chudalayandi, S. (2006). Unraveling the genetic basis of hybrid vigor. *Proceedings of the National Academy of Sciences*, *103*(35), 12957–12958.

Clement, C. R. (1999). 1942 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. *Economic Botany*, *53*, 188–202.

Clement, C. R., de Cristo-Araújo, M., Coppens D'Eeckenbrugge, G., Alves Pereira, A., & Picanço-Rodrigues, D. (2010). Origin and domestication of native Amazonian crops. *Diversity*, *2*(1), 72–106.

Cosme, S., Cuevas, H. E., Zhang, D., Oleksyk, T. K., & Irish, B. M. (2016). Genetic diversity of naturalized cacao (*Theobroma cacao* L.) in Puerto Rico. *Tree Genetics & Genomes*, *12*, 1–13.

Delph, L. F., & Kelly, J. K. (2014). On the importance of balancing selection in plants. *New Phytologist*, *201*(1), 45–56.

Dadzie, A. M., Livingstone, D. S., Opoku, S. Y., Takrama, J., Padi, F., Offei, S. K., Danquah, E. Y., Motamayor, J. C., Schnell, R. J., & Kuhn, D. N. (2013). Conversion of microsatellite markers to single nucleotide polymorphism (SNP) markers for genetic fingerprinting of *Theobroma cacao* L. *Journal of Crop Improvement*, *27*(2), 215–241.

Dillon, N. L., Zhang, D., Nauheimer, L., Toramo, E., Nagalevu, P., Melteras, M.-V., Wallez, S., Finau, K., Nakidakida, S., Lepou, P., & Diczbalis, Y. (2024). Understanding the cocoa genetic resources in the Pacific to assist producers to supply the growing craft market. *New Zealand Journal of Crop and*

Horticultural Science, *52*(4), 306–320. https://doi.org/10.1080/01140671.2023.2278788

Edwin, J., & Masters, W. A. (2005). Genetic improvement and cocoa yields in Ghana. *Experimental Agriculture*, *41*(4), 491–503.

Efombagn, M. I. B., Sounigo, O., Nyassé, S., Dauleux, M. M., Cilas, C., Eskes, M. A. B., & Allen, K. (2006). Genetic diversity in cocoa germplasm of southern Cameroon revealed by simple sequence repeat (SSR) markers. *African Journal of Biotechnology*, *5*(16).

Gopaulchan, D., Motilal, L. A., Bekele, F. L., Clause, S., Ariko, J. O., Ejang, H. P., & Umaharan, P. (2019). Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. *Physiology and Molecular Biology of Plants, 25,* 361–375.

Grant, P. R., & Grant, B. R. (1995). Predicting microevolutionary responses to directional selection on heritable variation. *Evolution*, *49*(2), 241–251.

Haldane, J. B. S. (1992). Disease and evolution. *Current Science*, *63*(9-10), 599–604.

Jean-Marie, E., Bereau, D., Poucheret, P., Guzman, C., Boudard, F., & Robinson, J. C. (2021). Antioxidative and immunomodulatory potential of the endemic French Guiana wild cocoa “Guiana”. *Foods*, *10*(3), 522.

Ji, K., Zhang, D., Motilal, L. A., Boccara, M., Lachenaud, P., & Meinhardt, L. W. (2013). Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. *Genetic Resources and Crop Evolution*, *60*, 441–453.

Kelly, W. H. (1969). Selection of the Lafi 7 cocoa for Western Samoa Trust Estates Corporation. *South Pacific Bulletin*, *18*(3), 29–31.

Loor, R. G., Risterucci, A. M., Courtois, B., Fouet, O., Jeanneau, M., Rosenquist, E., & Lanaud, C. (2009). Tracing the native ancestors of the modern *Theobroma cacao* L. population in Ecuador. *Tree Genetics & Genomes*, *5*(3), 421–433.

Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714.

Martínez, J. T. (2025). Cultivating Pacific cocoa: Chinese plantation labor in colonial era Samoa and Vanuatu. *Journal of Social History*, *58*(4), 622–645.

Meirmans, P. G., & Hedrick, P. W. (2011). Assessing population structure: FST and related measures. *Molecular Ecology Resources*, *11*(1), 5–18.

Moll, R., Lindsey, M. F., & Robinson, H. F. (1964). Estimates of genetic variances and level of dominance in maize. *Genetics*, *49*(3), 411–423.

Motamayor, J. C., & Lanaud, C. (2002). Molecular analysis of the origin and domestication of *Theobroma cacao* L. In J. M. M. Engels, V. Ramanatha Rao, A. H. D. Brown, & M. T. Jackson (Eds.), *Managing plant genetic diversity: Proceedings of an international conference, Kuala Lumpur, Malaysia, 12–16 June 2000* (pp. 77–87). Wallingford, UK: CABI Publishing. https://doi.org/10.1079/9780851995229.0077

Motamayor, J. C., Lachenaud, P., Da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., & Schnell, R. J. (2008). Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLOS ONE*, *3*(10), e3311.

Peakall, R. O. D., & Smouse, P. E. (2006). GenAlex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, *6*(1), 288–295.

Pokou, N. D., N'Goran, J. A., Lachenaud, P. H., Eskes, A. B., Motamayor, J. C., Schnell, R., Kolesnikova-Allen, M., Clément, D., & Sangaré, A. (2009). Recurrent selection of cocoa populations in Côte d'Ivoire: comparative genetic diversity between the first and second cycles. *Plant Breeding*, *128*(5), 514–520.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Pugh, T., Fouet, O., Risterucci, A. M., Brottier, P., Abouladze, M., Deletrez, C., Courtois, B., Clement, D., Larmande, P., N'Goran, J. A. K., & Lanaud, C. (2004). A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theoretical and Applied Genetics*, *108*, 1151–1161.

Raju, R. N., & United Nations Economic and Social Commission for Asia and the Pacific. (2024). A review of the impact of climate change on cocoa production in the Pacific Islands. *Asia-Pacific Sustainable Development Journal*, *31*(1), 77–93.

Rodriguez-Campos, J., Escalona-Buendía, H. B., Orozco-Avila, I., Lugo-Cervantes, E., & Jaramillo-Flores, M. E. (2011). Dynamics of volatile and non-volatile compounds in cocoa (*Theobroma cacao* L.) during fermentation and drying processes using principal components analysis. *Food Research International*, *44*(1), 250–258. https://doi.org/10.1016/j.foodres.2010.10.028

Sereno, M. L., Albuquerque, P. S., Vencovsky, R., & Figueira, A. (2006). Genetic diversity and natural population structure of cacao (*Theobroma cacao* L.) from the Brazilian Amazon evaluated by microsatellite markers. *Conservation Genetics*, *7*, 13–24.

Spieth, P. T. (1979). Environmental heterogeneity: a problem of contradictory selection pressures, gene flow, and local polymorphism. *The American Naturalist*, *113*(2), 247–260.

Stander, J. R. (1993). Pre-breeding from the perspective of the private plant breeder. *Journal of Sugar Beet Research*, *30*(4), 197–207.

Symonds, V. V., & Lloyd, A. M. (2004). A simple and inexpensive method for producing fluorescently labelled size standard. *Molecular Ecology Notes*, *4*(4), 768–771.

Pilon, B., Hinterneder, K., Hay, E. H. A., & Fragomeni, B. (2021). Inbreeding calculated with runs of homozygosity suggests chromosome-specific inbreeding depression regions in line 1 Hereford. *Animals*, *11*(11), 3105.

Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., & van Etten, J. (2012). Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. *PLOS ONE*, *7*(10), e47676.

Toxopeus, H. (1968). Establishment of cacao clones in Nigeria. *Euphytica*, *17*(1), 38–45.

Whitkus, R., De la Cruz, M., Mota-Bravo, L., & Gómez-Pompa, A. (1998). Genetic

diversity and relationships of cacao (*Theobroma cacao* L.) in southern Mexico. *Theoretical and Applied Genetics*, *96*, 621–627.

Wright, S. (1939). The distribution of self-sterility alleles in populations. *Genetics*, *24*(4), 538–552.

Zhang, D., & Motilal, L. (2016). Origin, dispersal, and current global distribution of cacao genetic diversity. In B. Bailey & L. Meinhardt (Eds.), *Cacao diseases* (pp. 12–25). Springer, Cham. https://doi.org/10.1007/978-3-319-24789-2_1

Chapter 3

Future Directions

1. Introduction

Global cacao cultivation is largely concentrated in African, particularly in West Africa, where it consumes approximately six million hectares and accounts for 70 percent of the world's total production. Côte d'Ivoire and Ghana are the top two cacao-producing countries in the world, followed by Nigeria and Cameroon. At the beginning of the 21st century, the global production of cacao was about two million tons, whereas, after 2010, the production increased significantly to more than three million tons (Wessel & Quist-Wessel., 2015).

Côte d'Ivoire, the world's largest cacao producer, has an economy that is highly dependent on cacao exports. During periods of cacao price fluctuations, the government of Côte d'Ivoire has effectively stabilized farmers' earnings through a series of support and regulatory measures (Crook., 1990). Ghana is the second largest cacao producer in the world, and the cacao industry has been an important economic pillar of the country's export industry, as well as a centerpiece of its poverty alleviation, development, and reforms (Green., 2017). The industry provides a stable livelihood for over 700,000 farmers in the southern tropical region of Ghana (Adams., 2017). Indonesia's cacao industry occupies an important place in the economy and is the third largest cacao manufacturer in the world. Indonesia celebrates Indonesian Cacao Day every September to highlight the importance of cacao to its economy. In recent years, Indonesia's processed cacao bean market has shown great potential, especially in the cosmetic and pharmaceutical industries, where the demand for its processed cacao products has been growing. In 2022, Indonesia exported 385,421 tonnes of cacao, with a total value of approximately US$1.26 billion, and this figure shows a year-on-year growth trend (Ramdhani & Dewi., 2025).

The Pacific region is estimated to account for 2% of global cocoa production (Dillon et al., 2024). Samoa, the Solomon Islands, Fiji, Papua New Guinea, and so on, are the main production areas for cocoa in the Pacific (Hebbar et al., 2011). Cocoa cultivation in Samoa dates to the 19th century, and cocoa is one of the main cash crops in Samoa and an important part of the family income of local farmers (Kongor & Oduro-Yeboah., 2024; Dillon et al., 2024). Koko Samoa, a cocoa product, is one of the local population's favorite beverages (Janni, 2002).

2. Problems of cacao cultivation

To meet the demand, many countries have adopted the approach of expanding the area of cocoa cultivation to increase production. However, this has often been done at the expense of the original forest land and at the expense of the ecological environment. Therefore, the best approach should still be to increase the yield per hectare rather than depending solely on expanding the growing area However, the current yield per hectare of cacao cultivation is very low (Wessel & Quist-Wessel., 2015).

Over-aged cacao trees are one of the main reasons for low yields. There are two ways to address this problem: one is to rehabilitate existing farms. Since the economic life span of cacao trees is between 30 and 40 years, only farms with trees less than 30 years old are suitable for rehabilitation. Field trials conducted on ten cacao farms in Côte d'Ivoire showed that rehabilitation of cacao trees between 25 and 30 years old is feasible and most effective (Wessel & Quist-Wessel., 2015). Data showed that good maintenance and integrated pest management during rehabilitation increased average cacao yields by 40 percent over a four-year period, from about 500 kg to 700 kg per hectare (Alexiset al., 2012). Another approach is to replant parts or all the cacao from old farms. Farmers plant newer trees below the old trees and then gradually remove the old trees (Green., 2017). Experiments in Côte d'Ivoire have shown that planting young trees between old trees can successfully replace them step by step (Wessel & Quist-Wessel., 2015). However, this method needs an excessive amount of time. Due to the time it takes to mature the new trees, this process does not provide immediate profits. Besides, there is a risk that it can lead to depletion of soil nutrients (Wessel & Quist-

Wessel., 2015). Therefore, to save time and money, many farmers prefer the option of opening new land for cacao cultivation. The newly tilled soil is more fertile (Ruf, 2011). Overall, this approach exacerbates problems such as deforestation, occupation of arable land, environmental damage, and soil loss.

3. Methods

From a global perspective, Ghana is considered one of the most successful countries in cacao production and marketing (Wessel & Quist-Wessel., 2015). Madagascar is also actively searching for cultivation methods to combat natural disasters. This paper will look at Ghana's cacao policy and Madagascar's new "cacao landscapes" as examples and then will propose feasible options for Samoa.

One approach being taken by farmers in Ghana is to recycle nutrients left over from the cacao production process to achieve a Zero Waste Circular Economy. According to Kamp and Ostergard (2016), cacao shells can be used as an organic fertilizer and recycled into biogas for further production. Nutrient recycling eliminates the loss of carbon from the soil and reduces the demand for chemical fertilizers, contributing to environmentally friendly eco-agriculture (Kamp, A., & Østergård, H., 2016). Cacao factories in Bali use cacao pomace as a base material for the production of chocolate sauce (Sihwinarti et al., 2024). It has been estimated that the recycled agricultural residues in Ghana alone produce approximately the energy equivalent of 13 million barrels of oil (Kamp, A., & Østergård, H., 2016).

A major factor in the increase in cacao production in Ghana has been the provision of a range of support by the Ghana Cacao Board (COCOBOD). These include an increase in trade prices; the implementation of a free pest control program; the selection of hybrid germplasm; the provision of subsidies for fertilizers and insecticides; and the improvement of facilities. Meanwhile, the area under cacao cultivation was expanded, especially in the Western Region of Ghana (Wessel & Quist-Wessel., 2015). While expanding the area under cultivation, Ghana did not abandon the original cacao plantations. To address the aging problem, the Ghana Cacao Board announced a

national cacao rehabilitation program: 20 million cacao seedlings were provided free of charge to farmers in 2012, and a cacao rehabilitation and replanting plan was developed. 20% of the current size of cacao to be replanted by 2014 (Wessel & Quist-Wessel., 2015). It is calculated that over the past decade, the Ghana Cacao Board has replanted about 20 million cacao seedlings through the implementation of a cacao rehabilitation program to supplement and replace old and low-yielding trees on cacao farms (Hess., 2022).

Ghana is actively implementing projects in energy, hydraulic engineering, sanitation, and rural roads to attract more private and overseas investments. These cacao factories and enterprises have created more employment opportunities for the local population, reduced poverty, and improved the standard of living in Ghana (Hess, 2022). Ghana's cacao development has followed global trends, and with the intervention of the World Bank and the International Monetary Fund, Ghana has become one of the fastest-growing economies in sub-Saharan Africa (Adams et al., 2017).

In terms of pest and disease management, tree replacement and maintenance, as well as government support, Ghana's cacao policies can serve as a reference for the future development of cacao in Samoa. In addition, currently, a new model has been proposed and is being experimented with: the cacao landscape.

In Madagascar, although cacao exports account for only a small part of the global market. But there is a saying that a thing is valued if it is rare. Therefore, the cacao produced from Madagascar is highly sought after by chocolate marketers and connoisseurs. The cacao grown by smallholder farmers is cultivated with less fertilizer. However, the climate on Madagascar’s west coast is affected by hurricanes, leading to issues such as flooding and seawater intrusion. These challenges cause shifts between crops, forests, and other land uses. It makes it difficult to maintain stable and long-term agriculture. Land degradation is a serious problem, highlighting the need for sustainable agricultural landscapes to safeguard land resources. As a response, the concept of a sustainable cacao landscape has emerged. This approach integrates on-farm and off-farm interventions to protect the ecosystem functions within cacao landscapes,

prioritizing deforestation prevention and soil erosion control. Farmers' incomes and land utilization have been improved by reducing labor and optimizing cocoa germplasm (Tham-Agyekum et al., 2023). Samoa is an island nation surrounded by the ocean and is highly vulnerable to climate variations at sea. Therefore, it can draw inspiration from Madagascar’s cacao landscape strategy for its cacao sector development.

In addition to improving plantations and creating ecological landscapes to prevent natural disasters. The problem of diseases also seriously affects cocoa yields. Research has shown that there are five main diseases of cocoa. Black pod, vascular streak dieback, witch's broom, and frosty pod rot are caused by fungal infections. Cocoa swollen shoot virus is caused by viral infections (Oyenpemi et al., 2023). Scientists and farmers are actively using different methods to deal with these diseases. In field management, the pathogen of frosty pod rot is known to be spread by spores. Therefore, the prevalence of frosty pod rot can be reduced by controlling the amount of irrigation, tree height, and shade during cocoa cultivation (Phillips-Mora & Wilkinson 2007). It has been shown that a 20 % reduction in the shaded area of the cocoa canopy can reduce the prevalence of frosty pod rot by 90 % (Schroth et al., 2000). In addition, molecular markers are a fast and effective way to develop superior cocoa. Chandrakant et al., (2020) Rapid screening of cocoa germplasm for resistance to Vascular streak dieback by gene linkage. Queiroz et al., (2003) found a QTL locus flanking marker AV14.940 in cacao offspring obtained from the cross between the susceptible variety “ICS-1” and the resistant variety “Scavina-6” showing resistance to witches' broom. QTL loci linked to frosty pod rot and black pod rot were found on the linkage map constructed from SNPs in the progeny of the cross between “Pound 7” and “UF 273” (Gutiérrez et al., 2021). al., 2021). Cocoa in the Samoa region is also plagued by these diseases, especially witch’s broom and black pod rot (Kelly., 1969). The use of molecular markers will facilitate the development of superior disease-resistant cocoa plants in the future.

4. Suggestions for the future development of the Samoa cacao factory

Production and food need to be effectively integrated, and the transit point for achieving this is the factory operation. Cacao undergoes a series of fermentation and drying processes before entering the factory. These processes significantly affect the color, taste, texture, and aroma of the cacao and directly determine the future price of the product (Macedo., 2024). This process requires an optimum balance of time and temperature, and the harvested, fermented, and dried cacao is quickly transported to factories in various countries, where it is processed into a variety of chocolate products. Machinery and technology play a crucial role in cacao factories. In 20th-century Europe, the chocolate industry was controlled by a few large corporations (Poelmans, 2016). This indicates that while farmers worked hard to cultivate cacao, the low prices at which they sold allowed European factories to maximize profits—funneling wealth into the hands of capitalists rather than improving farmers' livelihoods. This could lead to an unfortunate situation where cacao farmers grow cacao but cannot afford to consume it (Sa'adah & Tuakia, 2024). In 2021, Ghana introduced a new policy announcing that it would halt the export of cacao to Europe's largest cacao producers. This policy aims to transform Ghana from a raw cacao exporter into a cacao product manufacturer in the future (Wätzold et al., 2025). This decision marks a significant milestone as well as a turning point in the country's cacao development history, impacting employment, foreign exchange earnings, international status, and economic recovery—especially following the downturn caused by COVID-19 (Boysen et al., 2023). Thus, the domestic demand and export of cacao products should be actively promoted and international cooperation should be broadened. A diversified export strategy could be implemented: exporting eastward to the Americas, westward to Asia, and increasing cacao exports to Pacific countries and islands, particularly Australia and New Zealand. This strategy could serve as a sustainable development strategy for Samoa, both now and in the future.

Samoa should also actively establish its own cacao factories. Not only limited to cacao cultivation and export, but also a comprehensive cacao chain of planting, harvesting, fermentation, factory production, product commercialization and company promotion. Samoa's unique geographic conditions for cacao production allow for direct

transportation from plantations to factories, reducing the quality degradation that occurs during long-distance shipping to other countries. This enables a plantation-to-factory-to-table supply chain, significantly benefiting regional development. By taking inspiration from Europe's well-established cacao industrial zones, such as Bourneville (Macedo, 2024), Samoa can develop a structured labor system and worker support facilities, including the establishment of employee housing, schools for workers' children, public facilities such as parks. The development of cacao handicrafts workshops would also encourage broader community participation in cacao-based creative industries (Seçuk & Seçim, 2024). These developments will directly help reduce unemployment pressures and boost Samoa's Gross Domestic Product (GDP). From a geographical perspective, Samoa's economic center is primarily concentrated in the northeastern coastal region, while inland areas remain underdeveloped. Therefore, cacao plantations and factories could be strategically located inland, creating a radiating, multi-layered urban development model. Under government coordination and management. In later stages, the country could further build cacao-themed parks, chocolate amusement centers, and museums, integrating local historical and cultural heritage and landscapes. A well-planned tourism ecosystem could be designed, consisting of the outer zone of maritime experiences, coastal landscapes, economic hubs, and culinary tourism. The middle zone of industrial parks, Community-oriented housing, and outdoor recreational activities. Inner zone of pristine rainforests eco-tourism routes.

5. Conclusion

In conclusion, the derivative industries of cacao are vast and diverse. Cacao is not only used in consumables such as beverages, pastries, and chocolate but also holds deep historical significance related to national development, including historical migration and the formation of labor forces. Additionally, cacao carries cultural value, symbolizing love and friendship in film and media, while also playing a role in the growth of handcrafted products. The economic and cultural impact of cacao production on a country is immense. However, its success depends on collective efforts from both

the government and the people. This chapter has examined several well-developed cacao-producing nations, ultimately providing recommendations and directions for Samoa's future cacao industry and factory establishment, and cacao community development. It is hoped that in the future, Samoa's cacao can become the "Pacific Icon" and gain global recognition.

6. Reference List

Adams, K., Nyuur, R. B., Ellis, F. Y., & Debrah, Y. A. (2017). South African MNCs' HRM systems and practices at the subsidiary level: Insights from subsidiaries in Ghana. *Journal of International Management*, *23*(2), 180–193.

Alexis, A. A., Abel, K. E., Arnaud, A. F., Kacou, E. K. R. A., Salomon, D. J. I., Yves, C. J., & Assamoi, Y. A. P. O. (2012). Rentabilité économique des techniques de réhabilitation et de replantation des vieux vergers de cacaoyers (*Theobroma cacao* L.) en Côte d'Ivoire. *Journal of Animal & Plant Sciences*, *14*(2), 1939–1951.

Boysen, O., Ferrari, E., Nechifor, V., & Tillie, P. (2023). Earn a living? What the Côte d'Ivoire–Ghana cocoa living income differential might deliver on its promise. *Food Policy*, *114*, 102389.

Chandrakant, E. S., Tulshiram, W. S., Mathew, D., Minimol, J. S., Jose, S., Kurian, S. P., & Shylaja, M. R. (2020). Microsatellite and inter-microsatellite markers linked with resistance to vascular streak dieback in cocoa (*Theobroma cacao* L.). *The Journal of Horticultural Science and Biotechnology*, *95*(3), 294–302.

Crook, R. C. (1990). Politics, the cocoa crisis, and administration in Côte d'Ivoire. *The Journal of Modern African Studies*, *28*(4), 649–669.

Dillon, N. L., Zhang, D., Nauheimer, L., Toramo, E., Nagalevu, P., Melteras, M.-V., Wallez, S., Finau, K., Nakidakida, S., Lepou, P., & Diczbalis, Y. (2024). Understanding the cocoa genetic resources in the Pacific to assist producers to supply the growing craft market. *New Zealand Journal of Crop and Horticultural Science*, *52*(4), 306–320. https://doi.org/10.1080/01140671.2023.2278788

Green, E. (2017). From extensive to involutionary growth: A dialectic interpretation of the boom and busts of cocoa production in the Gold Coast. *Journal of Agrarian Change*, *17*(3), 518–534.

Gutiérrez, O. A., Puig, A. S., Phillips-Mora, W., Bailey, B. A., Ali, S. S., Mockaitis, K., Schnell, R. J., Livingstone, D., Mustiga, G., Royaert, S., & Motamayor, J. C. (2021). SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of *Theobroma cacao* L. *Tree Genetics & Genomes*,

17(3), 28.

Hebbar, P., Bittenbender, H. C., & O'Doherty, D. (2011). Farm and forestry production and marketing profile for cacao (*Theobroma cacao*). *Specialty Crops for Pacific Island Agroforestry*. Retrieved August 10, 2011, from http://agroforestry.net/scps

Hess, J. P. (2022). A multi-level analysis of sustainability practices in Ghana: Examining the timber, cocoa, and gold mining industries. *International Journal of Organizational Analysis*, *30*(3), 760–777.

Janni, K. (2002). Plants in Samoan culture: The ethnobotany of Samoa. *Economic Botany*, *56*(1), 100–100.

Kamp, A., & Østergård, H. (2016). Environmental sustainability assessment of fruit cultivation and processing using fruit and cocoa residues for bioenergy and compost: Case study from Ghana. *Journal of Cleaner Production*, *129*, 329–340.

Kelly, W. H. (1969). Selection of the Lafi 7 cocoa for Western Samoa Trust Estates Corporation. *South Pacific Bulletin*, *18*(3), 29–31.

Kongor, J. E., Owusu, M., & Oduro-Yeboah, C. (2024). Cocoa production in the 2020s: Challenges and solutions. *CABI Agriculture and Bioscience*, *5*(1), 102.

Macedo, M. (2024). Cocoa at work: Materials and labor in the making of global chocolate. *Labor*, *21*(1), 42–59.

Oyenpemi, L. O., Tijani, A. A., & Kehinde, A. D. (2023). What determines a sustained use of approved pesticides for cleaner production and its impact on yield? Evidence from the cocoa industry in Osun State, Nigeria. *Cleaner and Responsible Consumption*, *9*, 100113.

Poelmans, E., & Swinnen, J. (2016). A brief economic history of chocolate. In M. P. Squicciarini & J. Swinnen (Eds.), *The economics of chocolate*(pp. 11–42). Oxford University Press.

Phillips-Mora, W., & Wilkinson, M. J. (2007). Frosty pod of cacao: a disease with a limited geographic range but unlimited potential for damage. *Phytopathology*, *97*(12), 1644–1647.

Queiroz, V. T., Guimarães, C. T., Anhert, D., Schuster, I., Daher, R. T., Pereira, M. G., & Wricke, G. (2003). Identification of a major QTL in cocoa (*Theobroma cacao*

L.) associated with resistance to witches' broom disease. *Plant Breeding*, *122*(3), 268–272.

Ramdhani, B. N. R., & Dewi, N. P. M. (2025). Analysis of factors influencing the volume of cocoa bean exports in Indonesia. *International Journal of Management Research and Economics*, *3*(1), 269–290.

Sa'adah, M., Santoso, I., & Tuakia, M. Z. (2024). Sustainable cocoa supply chain in East Java, Indonesia: Potential and risk identification. In *BIO Web of Conferences* (Vol. 90, p. 03006). EDP Sciences.

Schroth, G., Krauss, U., Gasparotto, L. J. A. D., Duarte Aguilar, J. A., & Vohland, K. (2000). Pests and diseases in agroforestry systems of the humid tropics. *Agroforestry Systems*, *50*, 199–241.

Seçuk, B., & Seçim, Y. (2024). Concept of artisan chocolate from the perspective of chocolatiers. *International Journal of Food Design*, *9*(1), 27–52.

Sihwinarti, D., Darmawan, A. J., Heptariza, A., Ramayu, I. M. S., & Erawati, K. N. (2024). Zero waste circular economy: Transforming chocolate production into chocolate jam in Bali. *Jurnal Ekonomi*, *13*(2), 1–14.

Tham-Agyekum, E. K., Ntem, S., Sarbah, E., Anno-Baah, K., Asiedu, P., Bakang, J. E. A., & Jones, E. O. (2023). Resilience against climate variability: The application of nature based solutions by cocoa farmers in Ghana. *Environmental and Sustainability Indicators*, *20*, 100310.

Wätzold, M. Y. L., Abdulai, I., Cooke, A., Krumbiegel, K., Ocampo-Ariza, C., Wenzel, A., & Wollni, M. (2025). Do voluntary sustainability standards improve socioeconomic and ecological outcomes? Evidence from Ghana's cocoa sector. *Ecological Economics*, *229*, 108474.

Wessel, M., & Quist-Wessel, P. F. (2015). Cocoa production in West Africa, a review and analysis of recent developments. *NJAS: Wageningen Journal of Life Sciences*, *74*(1), 1–7.

Supplement information

Table supplement 1. Cacao varieties used in this study.

IDs	Serial number	Varieties
1	Matina-1/6	Amelonado
2	SCA-9	Contamana
3	Criollo-13	Criollo
4	LCTEEN-37	Curaray
5	GU-156-B	Guiana
6	IMC-9	Iquiotos
7	PA-188	Maranon
8	UF-20	Nacional
9	NA-232	Nanay
10	LAFI-7	LAFI-7
11	RB-41	Purus
12	PMCT-58	Trinitario

Table supplement 2. Details for the six microsatellite loci used in this study.

SSR name	5' - 3' Forward primer	5' - 3' Reverse primer	Tm (°C)	M13 dye
mTcCIR61	CGC AGT CTG AAA CAA GAT AA	TGA CAT TAA TAT AAG GCG AA	46°C	NED
mTcCIR168	CGG TAG TAT TGA GGT GCG TAT	GTG AAT GAA TGG ATG TGA AA	51°C	FAM
mTcCIR200	CGC CAA TTT CTG ACC CA	CTT AAA ATA AGC CCA AAT AC	46°C	FAM
mTcCIR225	CAA GAC AAA GGG AAG AAG A	AGG GGA AGA GCA AAT C	51°C	VIC
mTcCIR236	CGA AGT CAA AGG AAA GTC AA	TCA GAA AAC GCA AAT AAA	51°C	NED
mTcCIR280	CAT TTG TCA TTT GTT GTT GT	GCC TTG GTA TTG ACT GT	51°C	VIC

Table supplement 3. Cacao gene locus information.

Sample	N61	N61	F168	F168	F200	F200	V225	V225	N236	N236	V280	V280
Amelonado	0	0	192	192	0	0	320	320	209	209	108	108
Contamana	214	214	192	196	257	257	322	322	209	213	108	112
Criollo	202	202	194	194	329	329	340	340	199	199	131	133
Curaray	206	214	198	198	255	257	332	332	195	217	106	113
Guiana	204	214	198	198	327	327	332	332	199	199	108	114
Iquiotos	214	218	192	194	255	257	322	322	199	211	108	118
Maranon	214	216	194	194	257	257	332	346	199	215	108	114
Nacional	0	0	196	204	0	0	332	332	209	209	114	114
Nanay	224	224	192	192	257	257	320	320	0	0	102	112
Lafi-7	216	216	0	0	257	257	0	0	0	0	106	114
Purus	216	216	194	202	250	257	322	322	195	211	108	117
Trinitario	196	212	192	196	257	329	322	340	209	209	108	108
ALE01	202	216	192	192	257	257	320	320	199	199	108	108
ALE02	216	216	192	194	257	329	320	340	199	199	133	133
ALE03	214	216	194	196	257	257	320	320	209	209	108	108
ALE04	214	214	192	192	257	331	320	320	209	209	108	108
ALE05	216	216	192	192	257	331	320	340	209	209	108	108
ALE06	202	216	194	194	329	329	320	320	199	209	108	133
ALE08	216	216	192	194	257	331	320	340	209	209	108	108
ALE09	202	202	194	196	257	257	320	320	199	209	108	108
ALE10	214	216	194	198	257	257	320	346	199	209	108	108
ALE11	202	202	192	192	257	257	320	320	209	209	108	108
ALE12	214	216	192	194	257	331	320	320	199	209	108	133
ALE13	216	216	194	198	257	331	320	320	0	0	104	108
ALE14	214	218	192	194	257	257	320	320	199	209	108	108
ALE15	216	216	192	198	257	257	0	0	199	209	0	0
ALE16	0	0	0	0	257	257	340	340	209	209	133	133
ALE17	216	216	192	192	257	329	320	320	199	209	108	133
ALE18	216	216	192	192	257	334	320	340	209	209	131	133
ALE20	216	216	192	192	257	257	320	340	199	209	108	133
ALE21	216	216	192	192	257	334	320	340	209	209	133	133
ALE22	216	216	194	194	257	331	320	320	195	209	108	108
ALE23	202	202	192	194	329	329	320	346	199	209	108	108
ALE24	216	218	194	198	257	331	320	322	199	209	108	108
ALE25	216	216	192	192	257	334	320	320	209	209	133	133
ALE26	216	218	192	194	257	331	320	340	209	209	106	108
ALE27	216	218	192	192	257	257	320	322	199	209	106	108
ALE28	216	216	192	192	257	334	340	340	209	209	133	131
ALE29	202	216	192	194	257	331	320	320	209	209	108	133
ALE30	216	216	192	192	257	334	340	340	209	209	108	108
ALE31	0	0	194	196	257	257	0	0	209	209	108	112
ALE32	216	216	192	192	257	257	320	340	0	0	133	133
ALE33	206	216	192	194	257	329	320	320	209	209	108	133
ALE34	216	216	192	192	257	329	320	320	199	199	108	133
ALE35	216	216	192	192	257	334	320	340	209	209	133	133
ALE36	216	216	192	194	250	329	322	340	209	209	106	108
ALE37	216	216	192	192	257	257	320	320	199	209	108	133
ALE38	216	216	192	196	257	257	320	332	199	199	106	108
ALE39	216	216	192	192	0	0	320	320	209	209	108	133
ALE40	216	216	192	192	257	329	320	320	209	209	108	108
ALE41	216	216	192	192	257	257	320	340	199	199	108	133
ALE42	216	216	192	192	257	329	320	320	199	209	108	114
ALE43	216	216	192	192	329	329	320	320	199	209	108	133
ALE44	206	216	192	192	257	329	340	340	199	209	108	133
ALE45	214	216	192	196	257	331	320	346	199	209	108	108
ALE46	202	202	192	192	257	257	320	338	199	199	133	133
SAL01	214	216	192	194	257	329	320	340	197	209	108	112
SAL02	214	216	192	194	257	329	320	340	197	209	108	112
SAL03	216	216	192	192	257	257	320	320	199	209	133	133
SAL04	202	216	198	198	257	329	320	340	199	209	108	133
SAL05	214	214	192	192	257	331	320	340	197	209	106	108
SAL06	206	218	192	198	257	331	320	332	209	215	106	108
SAL07	0	0	194	194	257	327	320	332	199	209	112	133
SAL08	216	216	192	192	257	329	320	320	199	209	106	108
SAL09	216	216	192	194	257	329	320	320	209	209	108	108
SAL10	206	218	192	194	257	331	320	320	209	215	108	108
SAL11	214	216	192	194	257	329	320	320	209	209	112	133
SAL12	202	216	192	192	329	329	320	340	199	209	133	133
SAL13	202	202	192	192	257	334	320	320	209	209	133	133
SAL14	216	216	192	196	257	331	320	346	209	209	108	108
SAL15	216	216	194	194	257	334	320	320	199	199	108	108
SAL16	216	216	192	192	0	0	320	320	209	209	133	133
SAL17	214	216	194	196	257	257	322	340	199	209	108	108
SAL18	202	216	198	192	257	331	320	332	199	209	106	108
SAL19	216	216	192	192	257	257	320	320	209	209	108	112
SAL20	202	214	192	198	257	331	320	332	199	209	0	0

SAL21	202	216	194	196	257	329	320	340	199	209	106	108
SAL22	202	202	192	192	257	329	320	346	209	209	108	133
SAL23	214	216	192	194	255	329	320	320	199	209	108	133
SAL24	202	206	192	194	257	329	320	320	199	199	108	108
SAL25	206	216	194	198	257	257	320	320	199	209	108	133
SAL26	202	206	194	194	257	257	320	322	199	209	104	108
SAL27	216	216	192	194	0	0	332	340	209	209	108	108
SAL28	206	216	194	194	257	331	320	320	199	209	106	112
SAL29	216	216	194	194	257	334	320	332	209	209	108	133
SAL30	202	216	192	192	257	329	320	320	209	209	108	133
SAL31	202	218	192	194	257	334	320	320	199	209	108	112
SAL32	216	216	192	192	257	329	320	340	209	209	108	133
SAL33	202	202	192	192	257	331	320	320	209	209	108	108
SAL34	216	218	192	192	257	331	320	340	209	209	108	133
SAL35	202	214	192	192	0	0	0	0	197	209	108	108
SAL36	216	216	192	194	0	0	0	0	199	209	108	108
SAL37	202	216	192	192	257	331	320	332	199	209	108	108
SAL38	216	216	192	198	257	331	320	332	199	199	108	108
SAL39	202	216	192	192	257	257	320	332	199	209	108	108
SAL40	202	216	192	198	257	331	332	332	199	209	108	133
SAL41	202	216	192	192	257	331	320	320	209	209	108	108
SAL42	0	0	192	192	257	331	320	332	199	209	108	108
SAL45	216	216	192	192	257	257	320	332	209	209	108	108
SAL46	202	214	194	194	257	329	320	340	199	209	108	108
SAL47	202	216	192	194	257	331	320	332	199	209	0	0
SAL48	202	216	192	194	257	331	320	346	199	199	0	0
SAL49	216	216	192	194	257	329	320	340	199	209	108	108
SAT01	202	216	194	194	257	329	320	340	199	209	108	133
SAT02	204	216	192	192	257	331	320	340	209	209	108	108
SAT03	202	224	192	192	329	329	320	320	209	209	108	112
SAT04	0	0	194	194	257	329	320	340	199	209	108	133
SAT05	202	216	192	192	257	329	320	340	199	209	108	108
SAT06	202	202	192	192	257	329	340	340	209	209	108	108
SAT07	202	216	192	192	257	331	0	0	209	209	0	0
SAT08	202	216	192	192	257	329	320	320	199	209	108	133
SAT09	204	216	192	192	257	331	320	320	209	209	106	108
SAT10	202	202	192	192	257	331	320	320	209	209	108	108
SAT11	202	202	192	192	257	329	320	320	209	209	108	108
SAT12	202	216	192	192	257	331	320	320	199	209	108	108
SAT13	202	216	192	192	257	331	320	320	209	209	108	108
SAT14	216	216	192	192	257	331	320	320	199	209	108	108
SAT15	216	216	192	192	257	329	0	0	209	209	108	108
SAT16	202	216	192	192	257	329	320	340	209	209	108	133
SAT17	216	216	192	192	0	0	320	320	199	209	108	133
SAT18	216	216	192	192	257	331	0	0	209	209	108	108
SAT19	216	216	192	192	257	331	320	320	209	209	108	108
SAT20	202	216	192	192	257	331	320	320	199	209	108	108
SAT21	216	216	192	192	257	331	320	320	209	209	108	108
SAT22	0	0	192	192	257	257	332	340	195	209	108	108
SAT23	204	216	192	192	257	257	320	340	209	209	108	108
SAT24	202	216	192	192	257	331	320	340	199	209	108	108
SAT25	216	216	192	194	257	329	320	340	199	209	108	133
SAT26	202	216	192	194	257	329	340	340	199	209	108	108
SAT27	202	202	192	194	329	329	320	340	209	209	108	133
SAT28	216	216	192	192	257	257	320	320	209	209	108	133
SAT29	216	216	192	192	257	331	320	320	199	209	108	133
SAT30	216	216	192	192	257	329	320	340	199	209	108	133
SAT31	216	216	192	194	257	331	320	320	209	209	108	108
SAT32	204	214	192	192	257	257	320	332	199	209	108	112
SAT33	202	216	0	0	257	331	320	320	199	209	108	108
SAT34	216	216	192	192	257	331	320	320	209	209	108	108
SAT35	202	216	192	194	257	257	320	320	199	209	108	108
SAT36	216	216	192	192	257	257	320	320	209	209	108	108
SAT37	202	216	192	192	257	257	320	320	209	209	108	108
SAT38	0	0	192	192	257	257	320	340	199	209	108	133
SAT39	216	216	192	192	257	331	320	340	209	209	108	108
SAT40	204	204	192	192	257	334	320	340	209	209	108	133
SAT41	204	216	192	192	257	257	320	340	209	209	108	108
SAT42	202	216	192	192	257	257	320	340	199	209	108	133
SAT43	202	216	192	194	257	257	320	340	199	209	108	133
SAT44	216	216	192	192	257	331	320	320	209	209	108	108
SAT45	216	216	192	192	257	331	320	320	209	209	108	108
SIU01	202	216	192	192	257	257	320	340	199	209	108	133
SIU02	202	204	192	192	257	257	320	320	199	199	108	133
SIU03	202	202	192	192	257	334	320	340	199	209	108	133
SIU04	216	216	192	192	257	331	320	322	199	209	108	108
SIU05	202	214	192	196	257	334	322	346	209	209	112	133
SIU06	202	216	192	192	0	0	340	340	199	209	108	133
SIU07	202	202	192	192	257	329	320	320	209	209	108	133
SIU08	216	216	192	192	257	257	320	340	199	209	108	133
SIU09	214	216	192	196	257	257	320	320	199	209	108	133

SIU11	216	224	192	198	250	257	320	322	209	209	108	108
SIU12	214	214	192	192	257	331	320	322	199	199	133	133
SIU13	202	202	192	194	257	334	320	340	199	209	108	133
SIU14	0	0	192	192	0	0	320	320	209	209	108	108
SIU15	202	216	192	192	257	334	320	320	199	199	108	133
SIU16	202	214	192	192	250	257	322	338	209	209	108	133
SIU17	202	216	192	192	257	257	340	340	209	209	108	108
SIU18	202	216	192	192	257	329	320	320	209	209	108	133
SIU19	202	204	192	192	257	257	320	320	199	209	108	133
SIU20	202	216	192	192	257	257	320	320	209	209	108	108
SIU21	202	202	192	192	257	334	320	320	199	199	108	133
SIU22	202	202	192	192	257	331	320	340	199	209	108	133
SIU23	202	216	192	192	257	257	320	340	199	199	108	108
SIU24	214	214	192	192	250	257	322	338	209	209	108	133
SIU25	202	202	192	192	257	257	320	340	209	209	108	133
SIU26	202	216	192	192	257	331	320	340	209	209	108	133
SIU27	202	216	192	192	0	0	320	340	199	209	108	133
SIU28	202	216	192	192	257	257	340	340	199	209	133	133
SIU29	202	202	192	192	257	257	320	340	199	209	108	133
SIU30	216	216	192	192	257	257	320	320	199	199	133	133
SIU31	202	216	192	192	257	334	320	340	199	209	108	133
SIU32	0	0	192	192	0	0	332	340	199	209	133	133
SIU33	202	216	192	192	257	257	320	322	199	209	108	112
SIU34	204	216	192	192	257	257	0	0	0	0	108	108
SIU35	216	216	192	192	0	0	338	340	209	215	108	133
SIU37	216	216	192	192	255	329	320	320	209	215	133	133
SIU38	216	216	192	192	257	331	320	320	199	209	108	108
SIU39	202	202	192	192	257	257	320	340	199	209	108	133
SIU40	216	218	192	192	257	257	320	332	199	209	133	133
SIU41	214	214	192	192	255	257	320	340	209	215	108	108
SIU42	216	216	192	192	257	257	320	322	199	209	108	133
SIU43	202	202	192	192	0	0	320	340	199	209	108	133
SIU44	216	216	192	192	0	0	320	320	209	209	108	108
SIU45	202	216	192	194	0	0	320	320	209	209	108	108
SIU46	202	202	192	192	257	257	322	340	199	209	108	108
SIU47	204	216	192	192	257	257	320	320	209	209	108	108
SIU48	204	204	192	192	329	329	320	322	199	199	108	108
SIU49	202	216	192	192	257	329	320	340	199	209	108	133

Table supplement 4. The size of alleles for the six microsatellite loci in 12 varieties cacao.

Locus	Common alleles	Alleles in 12 species										
		Amelonado	Contamana	Criollo	Curaray	Guiana	Iquiotos	Maranon	Nacional	Nanay	LAFI-7	Purus	Trinitario
N61	196, 202, 204, 206, 212, 214, 216, 218, 224	-	214	202	206, 214	204, 214	214, 218	214, 216	-	224	216	216	196, 212
F168	192, 194, 196, 198, 202, 204	192	192, 196	194	198	198	192, 194	194	196, 204	192	-	194, 202	192, 196
F200	250, 255, 257, 327, 329	-	257	329	255	327	255	257	0	257	257	250, 257	257, 329
V225	320, 322, 332, 340, 346	320	322	340	332	332	322	332, 346	332	320	-	322	322, 340
N236	195, 199, 209, 211, 213, 215, 217	209	209, 213	199	195, 217	199	199, 211	199, 215	209	-	-	195, 211	209
V280	102, 106, 108, 112, 113, 114, 117, 118, 131, 133	108	108, 112	131, 133	106, 113	108, 114	108, 118	108, 114	114	102, 112	106, 114	108, 117	108

"-": missing data. The red markers represent private alleles.