# Samoa Cocoa Genotyping 2022-2023

Prepared by: Dr. Vaughan Symonds (Massey University) in collaboration with Seeseei Molimau-Samasoni (Scientific Research Organisation of Samoa)

## Introduction

*Theobroma cacao* L. (Malvaceae), or ‘cocoa’, has a long history of global dispersal and cultivation. The centre of origin for the species are the Orinoco and Amazon basins and the species is largely believed to be native to the region spanning from Mexico to Peru (Urquhart 1961). Upon discovery by European explorers, cocoa plants were distributed to all humid tropical regions of the world over several centuries (Figure 1). The development of the cocoa plantation industry was led by the Spaniards and was eventually followed by the Dutch, French, British and Germans (Urquhart 1961). Further movement of cocoa cultivation to Asia and the Pacific rim occurred in the sixteenth and seventeenth centuries, when cocoa was transported across the Philippines, Sulawesi and Java by the Dutch and the Spanish conquistadors (Urquhart 1961, Wood 1985, Young 1994). In the early 1800s, the cultivation of cocoa had expanded from Central America to South America and to several islands in the Caribbean and some areas in the East Indies (Young 1994). It is believed that cocoa was also introduced to Sri Lanka and India after it was brought to the East Indies (Wood 1985). It was in the nineteenth century and early twentieth century that Germans started to establish colonies in the Pacific Islands, which provided opportunities for Germany to pioneer the cocoa industry in Samoa (Stolberg 2013, Urquhart 1961).

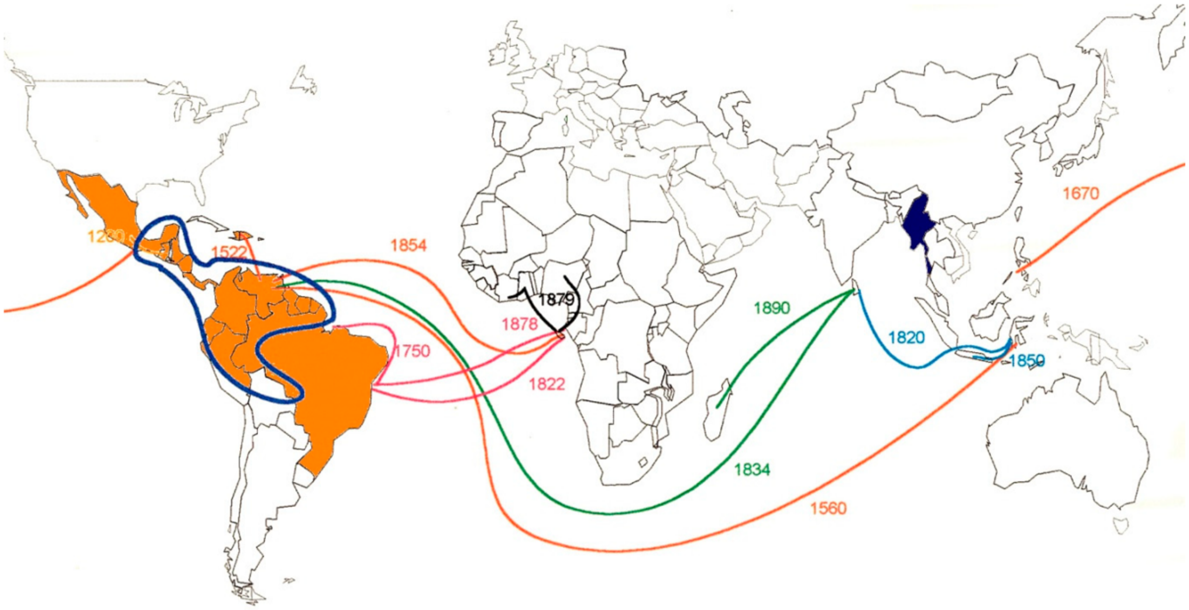


**Figure 1.** Map showing the timing of cocoa distribution around the globe. Figure lifted from Cilas and Bastide (2020). We acknowledge that both Samoa and New Zealand are missing from this map, but this was the best example from the literature for cocoa dispersal around the rest of the world.

Samoan cocoa is the product of a long history of development, cessation, re-introduction and transplantation of cocoa plant materials. The Criollo cocoa, which purportedly originated from Venezuela, was the first introduced cocoa variety in Samoa (Eden & Edwards 1952, Urquhart 1952). Later introductions of cocoa were sourced from Sri Lanka and Indonesia (Urquhart 1952). These Criollo varieties were introduced by a German company in 1883 (Kozicka et al 2018) as an effort to establish plantation holdings for tropical fruits (Droessler 2017, Slade 1984). However, when the cocoa industry in Samoa was still in its initial phase of development, the infestation of canker caused by the oomycete, *Phytophtora palmivora*, resulted in the abortion of cocoa cultivation and production (Eden & Edwards 1952). It was not until 1898 that the more resistant Forastero (Amelonado) variety from Sri Lanka was introduced to Samoa, as a replacement for the susceptible Criollo (Eden & Edwards 1952, Slade 1984, Urquhart 1952). It was also during the same period in the 1880s when the Trinitario cocoa was transplanted in Samoa, after it was re-planted in Sri Lanka in 1880 (Pohlan & Pérez 2010). In addition, due to mixed plantings of Amelonado and Criollo, natural hybridisation occurred, resulting in increased numbers of the hybrid-origin Trinitario (which was termed Samoa Trinitario) (Slade 1984). Currently, the predominant cultivated cocoa in Samoa are the open-pollinated Trinitario (sometimes called Koko Samoa, a term also used to refer to the roasted and pounded cocoa paste used for hot beverages), which are planted with Amelonado (Bourke 1992, Dillon et al 2014). The Amelonado cocoas, however, have been thought to comprise only about 12% of the total plantings (Bourke 1992). The history of introducing new planting materials triggered further mixture of the varieties across Samoa. While considerable research has gone into better resolving the genetic diversity and genetic groupings within *T. cacao*, resulting in several classification/naming systems (e.g., (Motamayor et al 2008)), the cocoa industry still largely utilises the traditional variety names: Forastero, Criollo, and Trinitario. An important note is that the term Trinitario is a blanket term used to refer to material derived from crosses between Forastero and Criollo, but the origins/nature of the Forastero and Criollo that have gone into a given cross, the nature of the cross (human-made or naturally occurring) and the generation of the hybrid (first generation, later generation, or back-crossed generations) are rarely understood or distinguished. As such, ‘Trinitario’ will refer to a huge range of genetic material, not one uniform type or variety and likely varies considerably by region based on introduction history and management.

As of 2018, ~10 million hectares of cocoa farms produced a total global export value of ~US$15 billion (Kozicka et al 2018) per year. Of course, as with the establishment of any new farming practice, the establishment of cocoa farms is often associated with habitat destruction (Maney et al 2022); however,  cocoa agroforests can actually maintain a high level of indigenous biodiversity (Schroth & Harvey 2007) if managed properly. Contemporary cocoa farming and production the world over face several challenges, many of which are likely to be exacerbated by a changing global climate (Delgado-Ospina et al 2021).

Given the potential (productivity and financial) gains to be made in cocoa, there is obvious pressure to enhance productivity (Kozicka et al 2018) wherever cocoa is farmed. In Samoa, cocoa bean exports remain an important cash crop, with significant room for growth of the industry. While there are several approaches that can be taken to improve various aspects of crops and their yields (Cilas & Bastide 2020, Dormatey et al 2020, Gao 2021, Pazhamala et al 2021, Yang et al 2021, Zhang & Batley 2020), all of them rely on understanding and best utilizing existing diversity. To provide a better understanding of the genetic diversity of Samoan holdings of cocoa, in this project, three approaches were taken:

(1) The first approach taken here explores the genetic diversity in cocoa plants from four locations around Samoa – two on Savai'i and two on Upolu. This follows up on previous work that examined genetic diversity from four other sites, also two from each of Savai'i and Upolu. This approach not only assesses genetic diversity in these collections but, importantly, puts this diversity in the context of recognized cocoa varieties by making comparisons with reference collections from The Tropical Agricultural Research and Higher Education Center (C.A.T.I.E.). Similar approaches have been taken in other cocoa regions of the world with similar aims (Adenet et al 2020, Gopaulchan et al 2019, Opoku et al 2007, Whitkus et al 1998, Zhang et al 2006). The results here reinforce the results of a previous project that examined cocoa genetic diversity in Samoa. Specifically, the analyses conducted indicate individual plants that have a high genetic proportion of Amelonado, plants that have a high proportion of Criollo, and a vast mixture of these two varieties. No significant genetic contributions by other cocoa genetic groups were identified.

(2) The second approach taken here aims to characterize the genetic change that takes place between generations of cocoa plants. Specifically, we assessed the degree of genetic similarity between eight maternal plants and their progeny by genotyping each at 12 microsatellite loci and directly comparing the compositions of maternal plants and their progeny. The main result of this part of the project is that each progeny pool is highly variable as is the genetic similarity between maternal plants and their progeny. Beyond the direct genetic contribution to their progeny, the extent to which progeny will be more or less similar to the maternal plant will depend on the paternal gene pool that contributes pollen to the progeny pool.

(3) The thirs part of the project seeks to identify genetic variation in the site collections that is associated with natural resistance to *Phytophthora palmivora*, the oomycete pathogen that causes block pod rot, which continues to be a problem in much of the cocoa-producing world, including Samoa. Where prevalent, this disease reduces productivity considerably. There is, however, variation among cocoa plants for resistance/susceptibility to this disease. Here, we sought to identify molecular markers associated with black pod rot resistance. Over the past two decades, there have been multiple studies that sought to identify parts of the cocoa genome that are associated with disease resistance (e.g., see (Barreto et al 2015, Brown et al 2007, Gutiérrez et al 2021)). Such association studies link particular genetic markers with resistance/susceptibility. As such, identifying markers from the literature is fairly straightforward; however, it is rare that these published works reveal which marker variants (alleles) are linked to resistance and which are linked to susceptibility, which can impede the utility of the mapping results. Regardless, several markers were selected and the plants from the four Samoa sites were genotyped at those loci. The genotype data reveal variation at all loci analysed, yet identifying which of those variants (if any) are linked with black pod resistance requires continued work. Details are provided later in the report on how this can be resolved.

# Materials and Methods

## Plant materials

### Site collections

Young leaf material from ~200 cocoa trees was sampled from four sites (50 trees per site) across Samoa: Aleisa (ALE), Saleimoa Uta (SAL), Siufaga (SIU), and Satupaitea (SAT); these sample collections were made by staff of the Scientific Research Organisation of Samoa (SROS). From each tree, leaf material was collected fresh for immediate processing and also dried (in silica gel) for long-term preservation. These samples are hereafter referred to as the 'site collections'.

### Maternal plant and seedling collections (progeny pools)

From each of eight maternal cocoa plants, five seed from each of five fruits were collected and germinated by the staff at SROS. Four maternal plants were sourced from Vaisala/Savaii (VAI-01, VAI-02, VAI-03, and VAI-04) and four from STEC/Upolu (STE-01, STE-02, STE-03, and STE-04). Young leaf tissue was collected (fresh and silica-dried) from each parent plant (eight) and all progeny (total of 200) for analysis.

## Molecular methods

### DNA extraction

For all samples, DNA was extracted from either fresh or silica dried leaf material using the Qiagen DNeasy Plant Mini Kit (Qiagen, USA). DNA quality was checked on 1% agarose gels through electrophoresis (see Supplementary Figure 1). The result was obtained by viewing the Ethidium Bromide-stained gel in a Bio-Rad Gel Documentation system (Bio-Rad, USA). The DNA concentration was quantified using a NanoDrop$^{TM}$ Spectrophotometer (Thermo Scientific, USA). High quality DNA samples were diluted at 1:5 with sterile water. The extractions were performed by the team at SROS in Samoa and posted to the Massey research group in New Zealand upon completion.

### Microsatellite genotyping

Microsatellite markers for assessing genetic diversity in the site collected plants and the maternal-progeny pools were selected to match previous work on Samoan cocoa material (SCIDI 2018 project; Supplementary Table 1). Those markers were initially selected based on their use in a large *T. cacao* genetic diversity study (Motamayor et al 2008). While that study used ~96 markers, our modelling work identified fifteen markers that captured nearly all of the genetic groups identified by Motamayor et al (2008), including the Criollo type. All microsatellite primers were ordered from Integrated DNA Technologies, Inc. Forward primers were 5'-tailed with an M13 tag sequence (CACGACGTTGTAAAACGAC) (Boutin-Ganache et al 2001). See Figure 2 and Supplementary Figure 1 for images of the molecular lab and examples of the equipment used in this work.

To screen the site collections for the presence of alleles that have been associated with black pod rot resistance, six microsatellite loci were selected from the literature (Supplementary Table 2). Plants were genotyped at these loci following the same methods used for the genetic diversity assessment.

The PCR amplifications were carried out using a Biometra Ti thermal-cycler. The reaction mix contained 3.7 µL ultrapure water, 1 µL 10x buffer BD, 1 µL $MgCl_2$ (4.50 µM), 0.2 µL dNTP (10 µM), 1 µL forward primer (0.20 µM), 1 µL reverse primer (4.50 µM), 1 µL M13 primer (labelled with FAM, VIC or NED) (4.50 µM), 0.1 µL FirePol Taq Polymerase (5U/ µL) and 1 µL DNA (diluted to 1:5). Two PCR profiles were used to check the optimum amplification condition for the markers, since the universal fluorescent primer M13 was also incorporated in the mix. The first PCR profile is a standard used often utilised for M13-tailed microsatellite PCR. The PCR conditions were: 95°C for 3 mins (initial denaturation), followed by 35 cycles of 95°C for 30 sec (denaturation), 52°C for 40 sec (primer annealing), and 72°C for 40 sec (extension), and 72°C for 20 mins for the final extension. The second thermal cycling profile (Everaert et al 2017) was: 94°C for 4 mins (initial denaturation), followed by 35 cycles of 94°C for 30 sec (denaturation), 46°C or 51°C for 1 min (primer annealing), and 72°C for 1 min (extension), and the final extension at 72°C for 15 mins. The protocol from Everaert et al (2017) was used for all subsequent runs as it produced more prominent bands during electrophoresis trials. The PCR products were checked by resolving 5 µL of amplicon and 3 µL of 3X loading dye in 1.5% agarose gel electrophoresis for 100 minutes at 75 volts.

Marker pooling was carried out for three microsatellite markers at a time, each labelled with a different dye in a ratio of 1.75 µL: 1.50 µL: 2.00 µL for markers labelled with FAM:VIC:NED, respectively. For genotyping, 1 µL of the pooled PCR products was then mixed with 9 µL of Hi-Di formamide mix (Applied Biosystems, California, USA). The Hi-Di mix was prepared by adding 14 µL of ROX-labelled CASS ladder (3.5 µL each of 100, 200, 300 and 400 bp ladder) (Symonds & Lloyd 2004) and 86 µL of water to 1000 µL of Hi-Di formamide. The prepared genotyping plate was submitted to the Massey Genome Service (Palmerston North New Zealand) for microsatellite genotyping through fragment separation (capillary electrophoresis) on an ABI 3730 Genetic Analyzer (Applied Biosystems).

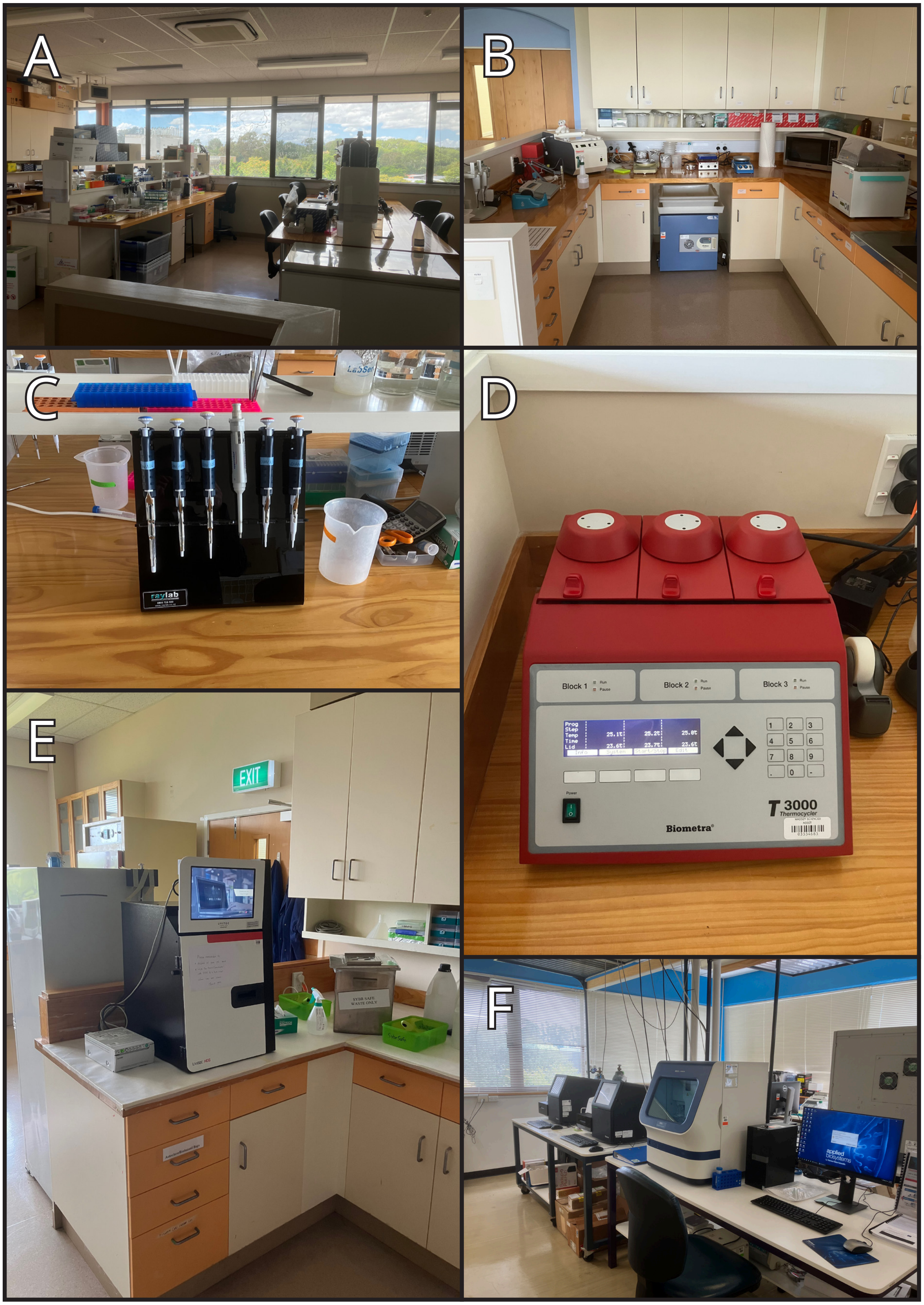

A
B
C
raylab
D
Block 1
Block 2
Block 3
Run
Pause
Power
T 3000
Thermocycler
Biometra
E
EXIT
STAR SAFE WASTE ONLY
F
applied biosystems

**Figure 2** (previous page). Photos of the spaces and equipment used in the molecular lab: (A) the main lab where PCR are set up and agarose gels are run; (B) an equipment bay in the lab for making solutions and incubating samples; (C) a standard set of micropipeters (each researcher has a set for themselves); (D) one of several thermalcylcers, which are used for temperature cycling of PCR; (E) a gel documentation system in the lab for observing and photographing DNA gels; (F) one of the ABI genotyping machines in the Massey Genome Service facility, which is in the same building as the molecular lab. This machine separates DNA fragments to a resolution of 1 base pair.

## Data analysis

### Allele calling

The raw genotype data were analysed using GeneMapper 5 Software (Applied Biosystems, USA) for initial allele calling. Alleles were called and recorded manually and final allele binning was accomplished by plotting and comparing all allele calls for a given marker to identify natural breaks in the distribution. The final data sets were assembled and initially analysed using GenAlEx v.6.51b2 software (Peakall & Smouse 2006). Samples with missing genotypes were re-genotyped and samples that still had more than three missing data points after re-genotyping were removed from analyses. For analyses that cannot handle missing data, missing genotype data were imputed.

### Genetic diversity

The GenAlEx software was used to: examine allele frequencies and observed and expected heterozygosities, generate summary statistics, produce genetic distances, including $F_{ST}$ (Wright 1949), and to run Principal Coordinate Analyses (PCoA) on the collected samples.

### Utilizing cocoa accession controls and previous Samoa cocoa genotyping results

The data from the four site sourced collections were analysed on their own and combined with the results of the 2018 SCIDI project and with the genotype data for 12 C.A.T.I.E. germplasm reference collections previously genotyped by our group (Table 1).

**Table 1.** *T. cacao* germplasm samples from C.A.T.I.E. used as controls.

| MASSEY Internal IDs | C.A.T.I.E. ID | Genotype group represented* |
|---|---|---|
| TC-1 | Matina-1/6 | Amelonado |
| TC-2 | SCA-9 | Contamana |
| TC-3 | Criollo-13 | Criollo |
| TC-4 | LCTEEN-37 | Curaray |
| TC-5 | GU-156-B | Guiana |
| TC-6 | IMC-9 | Iquiotos |
| TC-7 | PA-188 | Maranon |
| TC-8 | UF-20 | Nacional |
| TC-9 | NA-232 | Nanay |
| TC-10 | LAFI-7 | LAFI-7 |
| TC-11 | RB-41 | Purus |
| TC-12 | PMCT-58 | Trinitario |

*These samples represent the 10 genetic groups identified by Motamayor et al. (2008) plus a representative of Trinitario and the Samoan cocoa lineage, LAFI-7.

### Population structure analysis

The genetic structure of the ~200 site sourced cocoa samples was analysed with STRUCTURE v.2.3.4 software (Pritchard et al 2000). To obtain results with the broadest context, the ~200 samples from the current analysis were analysed along with the 2018 SCIDI and germplasm samples. The STRUCTURE parameters applied were: the admixture model and correlated allele frequencies. Each run consisted of a burn-in of 100,000 generations and MCMC (Markov chain Monte Carlo) length of 1,000,000 generations for each value of $K$ ($K$ = 1 to $K$ = 5) and 10 replicates were run for each $K$ value. $K$ represents the number of 'ancestral populations' or, effectively, the number of distinct genetic groups. These genetic groups consist of collections of alleles that make up a set of genotypes that either represent current observations (based on gene flow) or represent historical collections of genotypes, remnants of which are found in the current samples. To determine the best value of $K$, the log likelihood Ln P(D) = L($K$) was plotted against the $K$ values. As $K$ approaches its most likely value, L($K$) will start to increase in smaller increments until it plateaus. The best fitting $K$ value is typically selected based on the inflection point in the data.

STRUCTURE was also used to examine the structure of the progeny pools. For this analysis, all maternal plants and their progeny were analysed together following the same general parameters described above. Given the outcrossing nature of the cocoa plants, the expectation was that each maternal plant and its progeny would form a distinct genetic group. The STRUCTURE analyses are run blind and the software does not utilize the sample location information (i.e., it does not know which individuals are half-sibs or associated with a particular maternal plant). Instead, STRUCTURE simply seeks to identify collections of genotypes that form good Hardy-Weinberg groups and then assigns each individual's genotypes back to one or multiple genetic groups.

### Cluster analysis

From the site collections, a pairwise genetic distance matrix using the Provesti model (Prevosti et al 1975) was created using the *poppr* package in R (R Core Team 2013) and exported in the nexus file format using the *phangorn* package. This data set included individuals identified by the STRUCTURE analysis to be >80% from a single genetic group. Using this distance matrix, a cluster analysis was performed using the SplitsTree (v.4.15.1) software (Huson & Bryant 2006), employing the Neighbor-net distance-based method.

### Assessing genetic similarity between maternal plants and their progeny

The biological questions being addressed using the progeny data sets utilised some of the same methods employed above, but also required somewhat different analyses than those used on the site collections. In short, the genetic similarities between maternal plants and their progeny were to be assessed; several approaches were taken to achieve this. First, pairwise genetic distance matrices were generated using Genalex (Nei's genetic distance (Nei 1987)) and R (Prevosti's genetic distance (Prevosti et al 1975)) for each of the eight genotype datasets (a maternal parent plant and its progeny were treated as a data set in this analysis). Nei's distance matrix was used to run a Principal Coordinate Analysis for each of the eight data sets in Genalex. The Prevosti distance matrices were used to produce Neighbor-nets in Splitstree as described above.

Genetic distance-based metrics (as above) can display the genetic relationships among individuals (in this case progeny and their maternal parent and to one another), but not how they are similar or dissimilar. To better understand the nature of the differences between progeny and their maternal parent, we developed a novel analysis method that characterizes and summarises genotype similarities between progeny and the maternal parent. This method (Discreet Genotype Pattern Analysis) compares each individual's genotype (at each locus) to its maternal parent's genotype (at each locus) and then categorises each comparison into one of the patterns described below (Types 0, 1, 2, and 3). For a given locus (region of the genome), a progeny could be:

**Type 0** = an individual is identical to the homozygous maternal parent, having received the same allele from each parent:

Example: Parent genotype: 110/110
Progeny genotype: 110/110

**Type 1** = an individual has one copy of the homozygous maternal parent's allele and another allele contributed by the paternal parent:

Example: Parent genotype: 110/110
Progeny genotype: 110/120

**Type 2** = an individual has two copies of one of the maternal parent's alleles, but is missing the other:

Example: Parent genotype: 110/130
Progeny genotype: 110/110

**Type 3** = an individual has one copy of one of the maternal parent's alleles, but is missing the other, and has a different allele contributed by the paternal parent:

Example: Parent genotype: 110/130
Progeny genotype: 110/120

The proportion of each progeny's genotypes that fitted the four patterns above were then tabulated and summarised using stacked bar charts to give an indication of how similar or dissimilar progeny are from their maternal parent and precisely how they differ.

## Results and Discussion

### Genetic diversity in the Samoa site sourced collections

Descriptive genetic diversity metrics were generated for the 2022 site sourced samples and are presented alongside the 2018 SCIDI site sourced samples (Table 2). Among the most common genetic diversity metrics used in population genetic studies are the average number of alleles per locus and the expected heterozygosity. Considering these two metrics together, the SAT and VS sites have the lowest and the SAL and STC sites have the greatest genetic diversity. However, the range of values observed is not wildly variable. The $H_E$ values observed here (range: 0.39 – 0.52) are intermediate to those reported for cocoa in other regions; e.g., Java (0.67 (Susilo et al 2011)), Sulawesi (0.67 (Dinarti et al 2015)), and Aceh (0.37 (Lukman et al 2014)). While extreme variation in these values is likely to represent real differences in genetic diversity, the use of different microsatellite markers from one study to another should be expected to underlie some of this variability.

Having more or less genetic variation can be interpreted in a number of ways. More genetic diversity tends to mean that a population (site, farm) is potentially more resilient to environmental change (including new pests or diseases) over the long-term; however having less genetic variation that is well suited to the current conditions can mean better yields in the short-term. Moving forward, some balance between these two would need to be struck – particularly once genetic variation associated with particular beneficial traits (increased yield, cocoa quality, pest and disease resistance) has been identified.

The fixation index ($F_{IS}$) is often used to identify populations or groups of individuals that are undergoing either natural selection for or against particular genotypes or extreme inbreeding or outbreeding. This metric has a range of -1 to +1. Values close to zero indicate a population or group of individuals that result from random mating, while negative values indicate an excess of heterozygotes relative to the expectation under random mating and positive values indicate an excess of homozygotes relative to the expectation under random mating. While there is some degree of difference in $F_{IS}$ observed among the collection sites, none are significantly different from zero, indicating that reproduction from one generation to the next has largely been through random mating – likely through naturally occurring regeneration, but potentially through breeding efforts that involved a large number of individuals.

**Table 2.** Genetic diversity summary statistics for the current farm sourced samples (ALE, SAL, SAT, SIU) and from the 2018 SCIDI farm samples (MZ, STC, SV, VS) : N = average number of samples per locus, $N_a$ = average number of alleles per locus, $H_O$ = Observed heterozygosity, $H_E$, expected heterozygosity, $F_{IS}$ = Wright's fixation index.

| Location | | N | $N_a$ | $H_O$ | $H_E$ | $F_{IS}$ |
|---|---|---|---|---|---|---|
| ALE | Mean | 47.80 | 4.73 | 0.41 | 0.47 | 0.12 |
| | SE | 0.14 | 0.36 | 0.04 | 0.04 | 0.05 |
| SAL | Mean | 47.00 | 4.93 | 0.49 | 0.51 | 0.04 |
| | SE | 0.00 | 0.56 | 0.04 | 0.04 | 0.03 |
| SAT | Mean | 49.87 | 3.60 | 0.38 | 0.39 | 0.00 |
| | SE | 0.09 | 0.38 | 0.03 | 0.03 | 0.04 |
| SIU | Mean | 46.60 | 4.20 | 0.41 | 0.45 | 0.07 |
| | SE | 0.24 | 0.34 | 0.03 | 0.04 | 0.03 |
| MZ | Mean | 50 | 4.27 | 0.40 | 0.46 | 0.13 |
| | SE | 0 | 0.42 | 0.03 | 0.02 | 0.06 |
| STC | Mean | 50 | 4.80 | 0.48 | 0.52 | 0.08 |
| | SE | 0 | 0.55 | 0.04 | 0.03 | 0.06 |
| SV | Mean | 47.67 | 4.20 | 0.42 | 0.46 | 0.10 |
| | SE | 0.19 | 0.39 | 0.04 | 0.03 | 0.06 |
| VS | Mean | 49 | 3.07 | 0.41 | 0.45 | 0.07 |
| | SE | 0 | 0.30 | 0.04 | 0.02 | 0.07 |

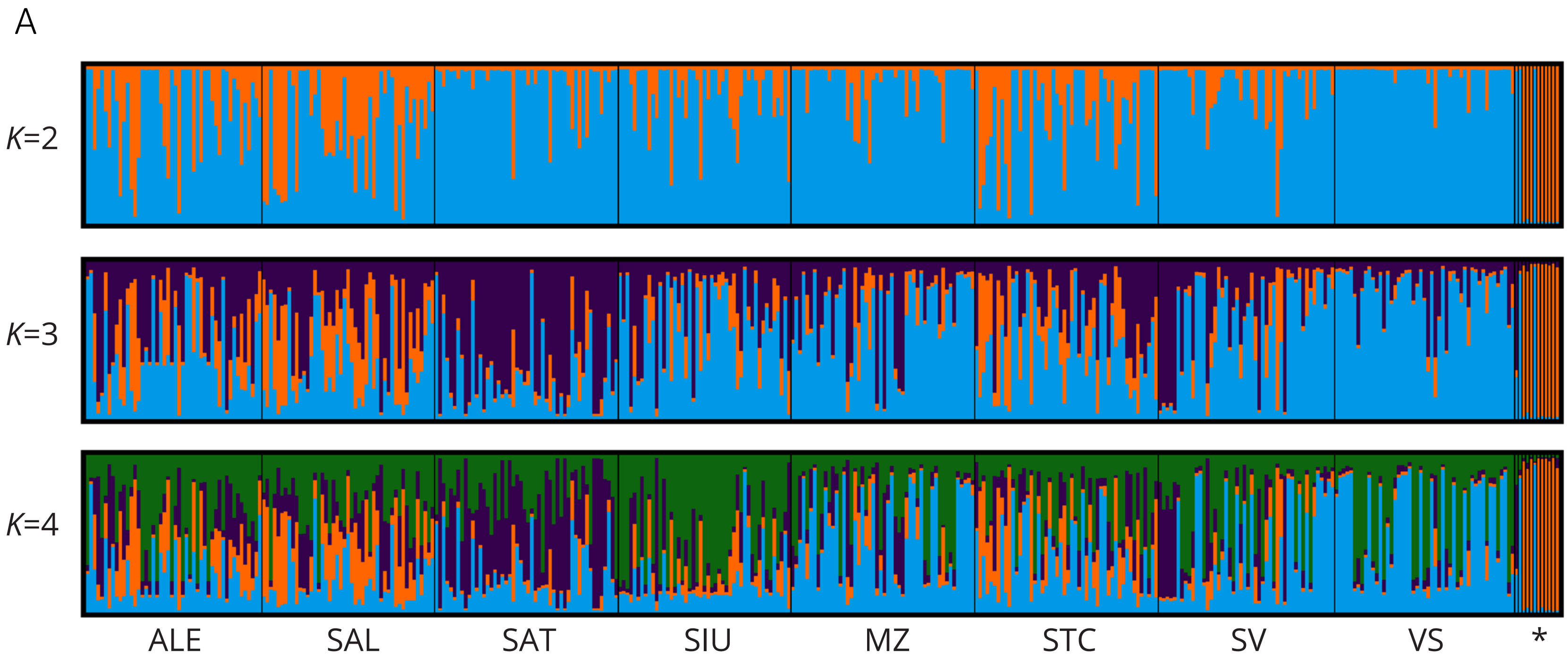
## Patterns of genetic structure in the Samoa site sourced samples

### *STRUCTURE results*

Among the most common analyses for identifying and visualising genetic groupings is the program STRUCTURE (Pritchard et al 2000). This program uses a maximum likelihood approach to construct genetic groups based on genotype data from multiple individuals and an analytical model based on the Hardy-Weinberg equilibrium criterion. In short, hypothetical genetic groups are sought that best fit a given data set and the allele frequency parameters of those hypothetical genetic groupings are then used to assign the genotypes of real samples back to one or more of the identified genetic groups. The genetic assignments for each individual are represented by a stacked bar graph where different colours represent the different genetic groups, each showing the proportion of an individual's genotypes that are assigned to each genetic grouping. The analysis is run multiple times, assuming a numerical range of genetic groups. For example, when assuming two genetic groups, one would be indicated by *K*=2, three genetic groups would be indicated as *K*=3, etc. Importantly, the collection of samples that are analysed can have a large impact on the results obtained as the hypothetical genetic groups that are identified are based on the samples in the data set.

Here, we analysed the plants from the four 2022 site sourced collections along with the four site sourced collections from 2018 and 12 germplasm samples from C.A.T.I.E. The results identify strong signals at *K*=2, *K*=3, and *K*=4 (Figure 3). Beyond *K*=4, there is little support for greater resolution and the genetic groupings begin to disintegrate. Recall that the hypothetical genetic groups are represented by different colours in the STRUCTURE outputs. At *K*=4 most of the reference germplasm samples, including the Trinitario representative, cluster together (orange), Criollo and LAFI-7 cluster together (blue), and Amelonado is distinguished from the others (mostly purple).

Collection sites MZ, VS, and to a lesser extent, SV have the greatest proportion of samples/genotypes that cluster with Criollo/LAFI-7 (Figure 3; panels A and B). This is consistent with the findings from the SCIDI 2018 project. The genetic group (represented by orange) that most of the germplasm collections belong to (including Trinitario) has the lowest proportion of all groupings across collection sites, but is at greatest frequency at the SAL site. The genetic group that Amelonado clustered with in the 2018 study is essentially split into two overlapping groups with the wider sampling included here. These are the groups represented by green and purple in all figures. The genetic group represented by purple is at greatest frequency at the SAT site, while the genetic group represented by green is at greatest frequency at the SIU site. It is most likely due to the inclusion of samples from these two sites that this split has occurred. The Amelonado sample included in this study mostly consists of the purple genetic group, but possesses some proportion of both the green and blue as well. This is the only germplasm sample to have a significant split in its genotype and the only one with a significant proportion of the green genetic group represented.


A
K=2
K=3
K=4
ALE
SAL
SAT
SIU
MZ
STC
SV
VS
*

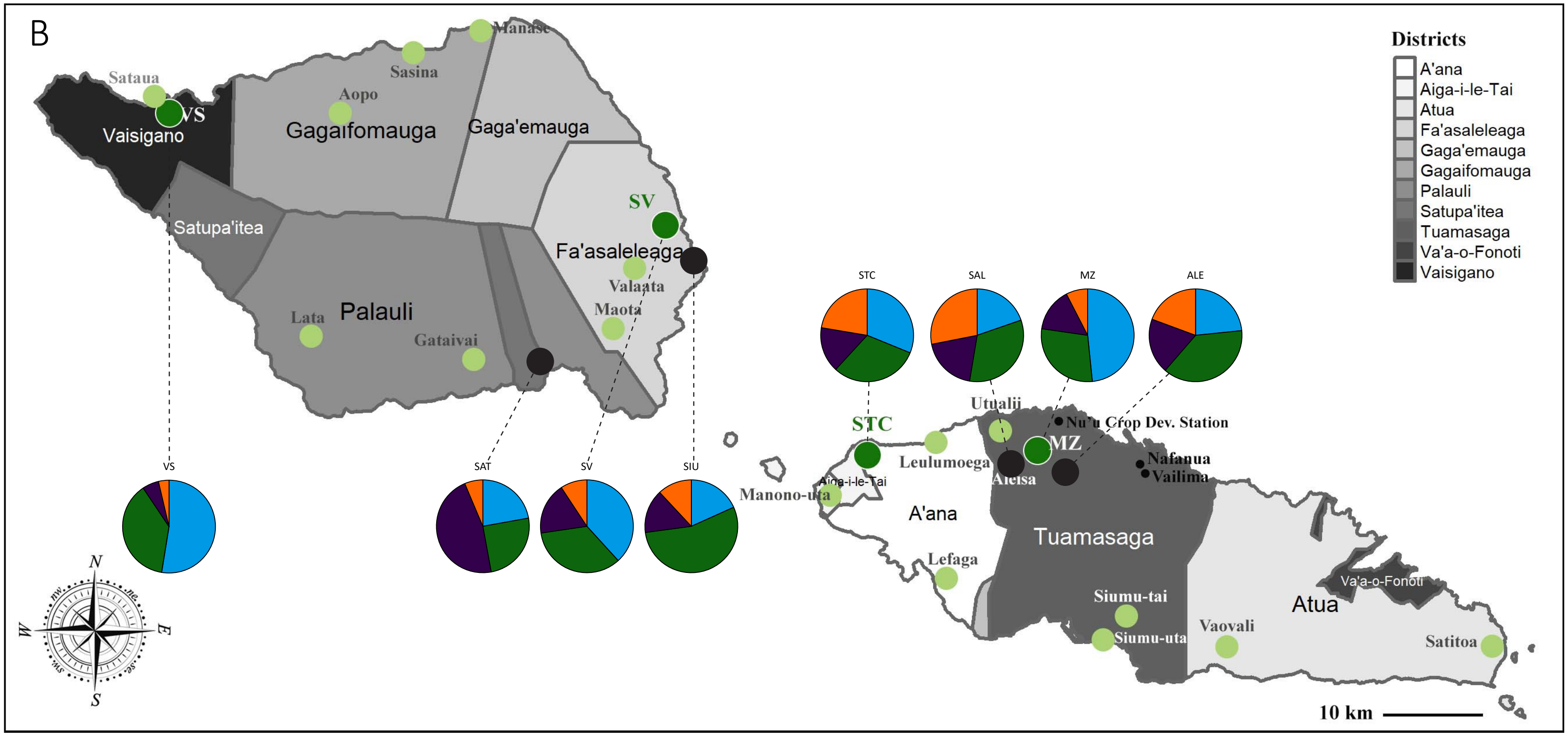


**Figure 3.** (A) STRUCTURE results for the data set including samples from all eight Samoa collection sites and the 12 C.A.T.IE. germplasm samples. Each block of samples is labelled by site. The block labelled with an * consists of the germplasm reference samples (from left to right): Amelonado, LAFI-7, Purus, Trinitario, Contamana, Criollo, Curaray, Guiana, Iquiotos, Maranon, Nacional, and Nanay (text colors indicate primary affiliation at *K*=4 in panel A). (B) Map of Samoa showing the locations of the eight farm collection sites and the proportions of genotypes at *K*=4 from the STRUCTURE results. Dark green circles represent the 2018 collection sites and black circles represent the 2022 collection sites. The pie charts indicate the relative proportion of genotypes from a given collection site that are from each of the four genetic groups identified by STRUCTURE; the colours match those in panel A.

### *Neighbor-net results*

The Neighbor-net presented here was generated by including only the samples with >80% affinity to one of the STRUCTURE-identified genetic groups (at *K*=4) plus all of the germplasm samples (Figure 4). The result shows essentially three major groups, similar to those identified by STRUCTURE. One of these groups is comprised of individuals that clustered together in the STRUCTURE blue group, one consists of individuals in the orange group, and the last one has individuals from a green group with individuals from the purple group nested within it. This outcome is consistent with the SCIDI report that identified three major genetic groups, but there is slightly more resolution in the larger data set. Based on the placement of the germplasm reference samples, we observe that the clusters seem to coincide with a 'Criollo' group (blue), an 'Amelonado' group (green and purple), and a group that initially appeared to be a 'Trinitario' group (orange); however nearly all other reference samples also fall into this group, along with very few Samoa samples. This is interesting as the germplasm samples are meant to best represent genetic diversity within natural populations of *T. cacao*. As STRUCTURE can only identify genetic groupings that are sufficiently sampled, one might conclude that, beyond Amelonado and Criollo, no or little other native genetic groupings are well represented in the Samoan material examined. Among the ~400 Samoa cocoa plants genotyped, the vast majority of the gene pool seems to stem from Amelonado and Criollo (ranging from 62-96% per site). This is different from similar studies on cocoa from Martinique (Adenet et al 2020) and Vietnam (Everaert et al 2020), where genotypes associated with other reference genetic groups was seemingly clear. Of course, each geographic region will have its own unique history of cocoa introduction and management. Potentially, deeper sampling of germplasm material could yield different results (as STRUCTURE relies on sampling depth to some extent); however, of those germplasm individuals sampled analysed here, one would expect them to find some affinity with Samoan genotypes if there was any shared history – as was the result for the Criollo, Amelonado, and LAFI-7 samples. Also, in a Neighbor-net analysis that included all Samoan samples and all germplasm samples (not shown), the nine germplasm samples (excluding Criollo, Amelonado, and LAFI-7) still clustered together. This is significant as the Neighbor-net analysis utilises only genetic distances and has no reliance on sampling depth, further supporting the supposition that the bulk of Samoan cocoa stems from the Criollo and Amelonado lineages.

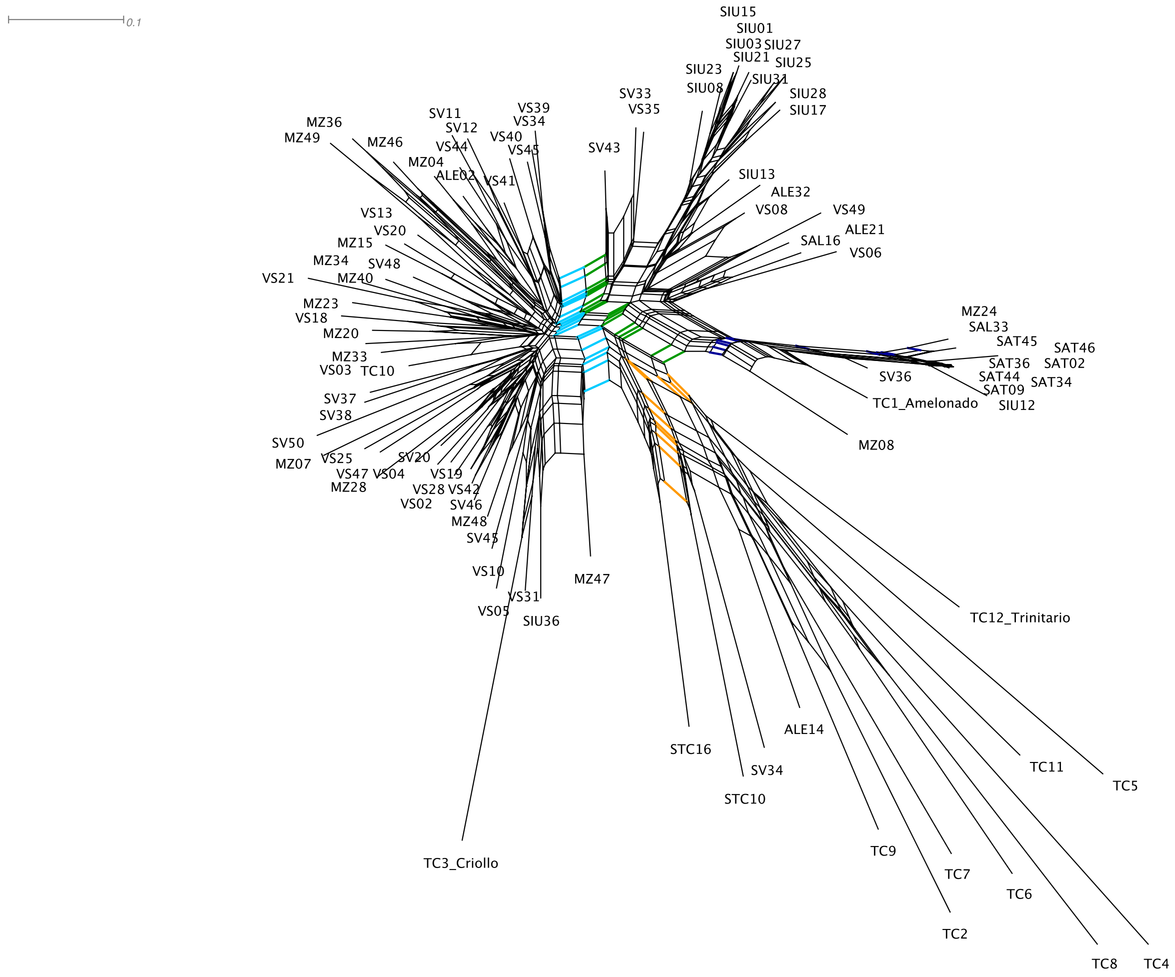


**Figure 4.** Neighbor-net of site collection individuals identified as having genotype compositions that are 80% or greater from a single genetic grouping identified by the STRUCTURE *K*=4 analysis and the reference germplasm samples. This approach allows for a more simple visualisation of the individuals that were identified as predominantly of one genetic type. Splits that differentiate the four groups are highlighted in blue, orange, green, and purple. Note that most of the germplasm samples (TC#) cluster together along with very few Samoa samples (the orange grouping).

## Assessing genetic similarity between maternal plants and their progeny

### *STRUCTURE results*

STRUCTURE analyses were run on the maternal/progeny pool data set assuming from 1- 9 genetic groupings. The number of groupings with the greatest likelihood score (best fit to the data) was seven (shown in Figure 5). For most of the samples, the progeny and maternal parent for a given pool clustered together, which was the expectation as these should constitute groupings that fit the Hardy-Weinberg criterion used by STRUCTURE. While there are a few individuals that show affiliations with groups other than their progeny group, some missing data might account for this; however, they tend to be samples that best match groupings from the same island, which suggests that there may be some gene flow taking place among sites (e.g., in pools 5 and 6). The surprise result is progeny pool 3 (VAI03), which shows a very broad mixture of genotypes from other progeny pools. Even at higher *K* values (not shown), this progeny pool never resolves as a distinct grouping. It is not entirely clear why this grouping does not form a good genetic group. This progeny pool does have the highest genetic diversity of the four VAI pools, but some of the STE pools have greater diversity yet still cluster as distinct groups.  This outcome may be due to two factors: potentially a broader set of paternal genotypes that are similar to those present at other locations have contributed to this progeny pool and (2) the maternal parent for VAI03 has a genotype that matches the VAI04 genetic group. In sum, this means that this progeny pool does not form a genetically unique grouping, which is an interesting result that indicates that this pool has a genetic history that is somehow different from the other pools.

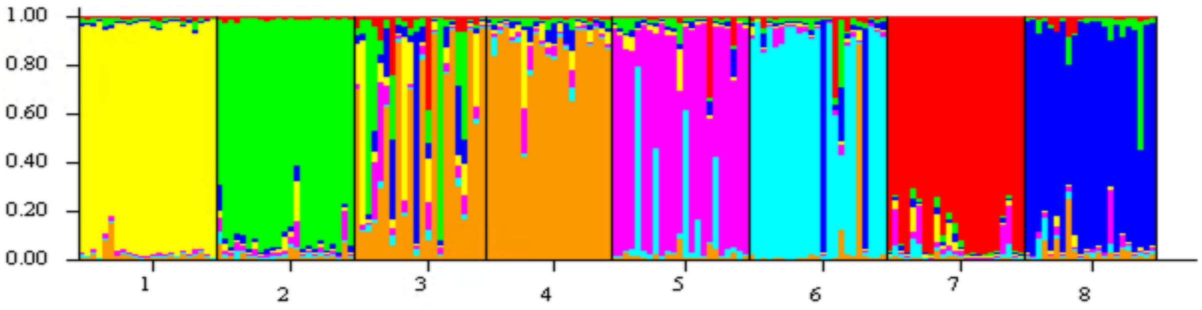


**Figure 5.** STRUCTURE result for *K*=7 showing the genetic composition of individual samples and progeny pools. Each block (1-8) consists of the progeny from a single maternal plant and the maternal plant is the last individual in each of the eight blocks. The blocks are: VAI01 (1), VAI02 (2), VAI03 (3), VAI04 (4), STE01 (5), STE02 (6), STE03 (7), STE04 (8). Note that a different colour scheme was used here relative to the STRUCTURE results in Figure 3 to make it clear that these results are not related to or comparable to those results.

### *PCoA results*

Among the most common ways of determining genetic similarity among individuals is to use multilocus genotype data to generate a pairwise genetic distance matrix. In essence, this sums up the genotypes that are in common among all pairs of individuals. Such a matrix is then analysed to produce a figure that best displays those relationships. While there are several methods available, here we have chosen a traditional approach (Principal Coordinate Analysis (PCoA)) and an alternative method that helps discern genetic relationships among individuals (Neighbor-nets).

Figure 6 shows the PCoA results for each maternal plant and its progeny, where the physical distance between plot points reflects genetic distance. Perhaps the first thing to point out is that progeny are not genetically identical to the maternal plant. If they were, their plot points would sit directly on top of the maternal individual's plot point (indicated in red in each graph). Of course, this is to be expected as cocoa plants do not clonally reproduce; instead, they undergo sexual reproduction. But even when cocoa plants self-pollinate, their progeny are not expected to have identical genotypes to the maternal plant as segregation and recombination will produce a wide array of genetic variability. The second note here is that the dispersion of points varies from one progeny pool to another. This also is to be expected as each set of progeny samples is a random collection of many possible genotypes. Finally, it is clear that some progeny are more genetically similar to their respective maternal plant than others (as reflected by their proximity to the red plot point in each graph). This is because some progeny will receive alleles from their paternal parent that are more similar (or not) to the maternal plant. This is discussed further in a later section that explores *how* progeny differ from their maternal parent plant.

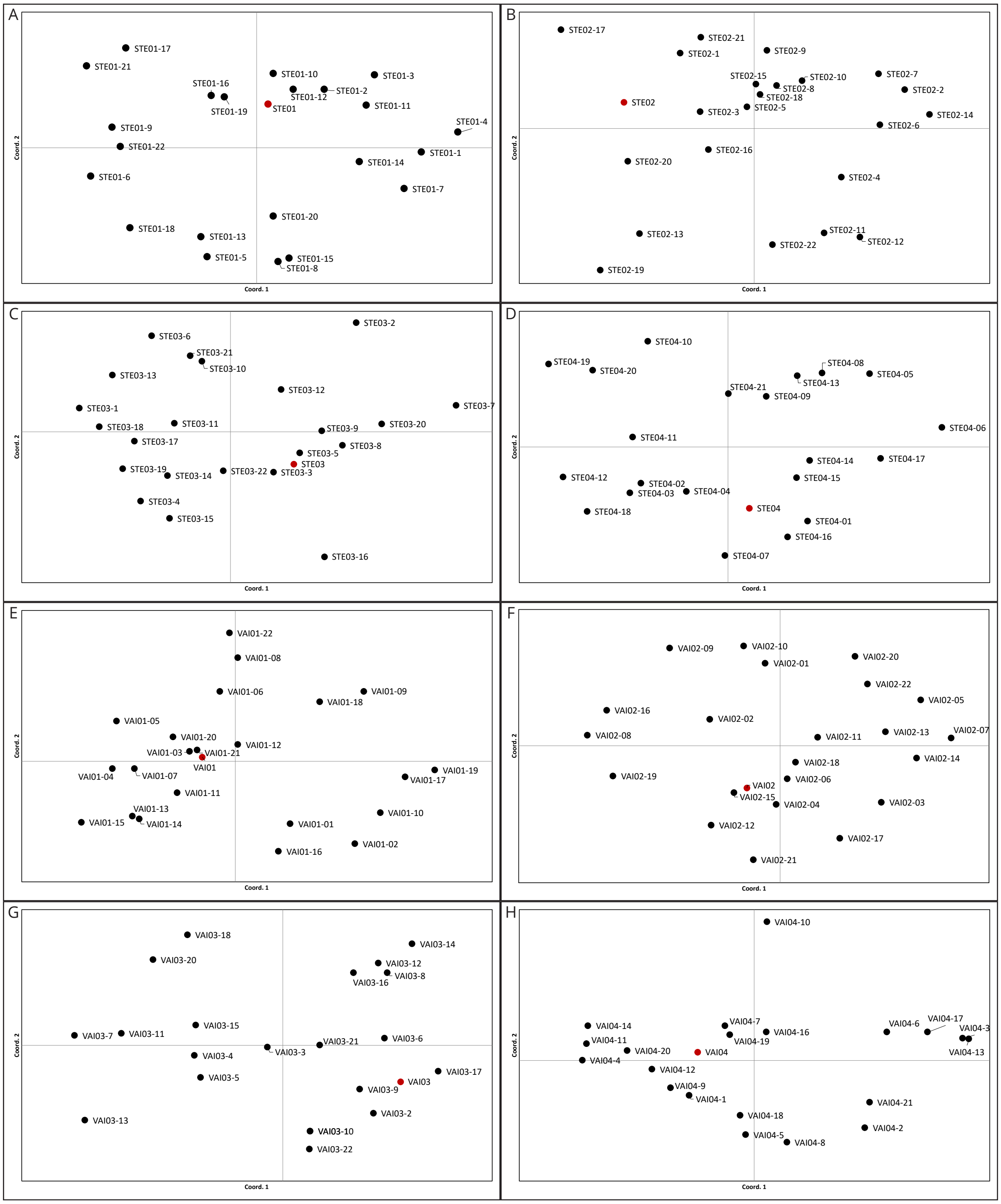


**Figure 6.** PCoA of each maternal parent plant (STE01 (A), STE02 (B), STE03 (C), STE04 (D), VAI01 (E), VAI02 (F), VAI03 (G), and VAI04 (H)) and its progeny based on pairwise genetic distances (Nei's distance). Each maternal parent individual is plotted in red. The position of each plot point indicates the genetic similarity among samples, where the greater the plot distance, the greater the genetic variation between samples. Note that some samples were removed from analyses due to missing data.

### *Neighbor-net results*

Another common method for visualizing the genetic relationships among samples is the use of a Neighbor-net analysis to generate network plots. For this data set, each progeny pool and their maternal parent plant were analysed together and the results are shown in Figure 7. Neighbor-nets attempt to reflect as many pairwise relationships as possible in two dimensions. In each graph, a series of parallel lines are referred to as a 'split'. In effect, there is information in the data that separates individuals on one side of a split from individuals on the other side. The length of a split reflects how much support there is in the data for that particular division.

Independent analyses such as PCoA and Neighbor-nets often (but don't always) reveal similar patterns. Discordant results between different approaches can arise due to the use of different genetic distance metrics prior to final analysis and the ways that each analysis handles data. For example, a PCoA seeks the greatest variation in a data set and assigns that to axis (coord.) 1 and then takes the same approach to determine axis 2; however further variation in the data set may be partitioned to subsequent axes (3, 4, etc.), which are not plotted in the two-dimensional PCoA. Such variation may be better identified and represented in other analyses though and result in different relationships being revealed.

It is often useful then to make comparisons across different analysis methods. For example, the Neighbor-net in Figure 7 shows a clear affiliation between individuals STE01-17 and STE01-21; they are differentiated from other samples by a long split in the graph. If we look at the PCoA plot in Figure 6, we see that these two individuals cluster together in one corner of the plot, thus corroborating the Neighbor-net results. Similarly, individuals STE01-08 and STE01-15 appear together as do STE01-05, STE01-13, and STE01-18 in both analyses. Such concordant examples can be found in each progeny data set.

One interesting, though potentially frustrating result is that the maternal plants often appear at the base of Neighbor-net plots and toward the center of PCoA plots. Given the obvious relationship between maternal plants and all of their progeny, the average genetic distance between them and progeny is lower than the average genetic distance among all pairs of progeny. As a result, the maternal plants tend to appear more central in the plots relative to the progeny, thus somewhat obscuring the relative similarities between maternal plants and their progeny.

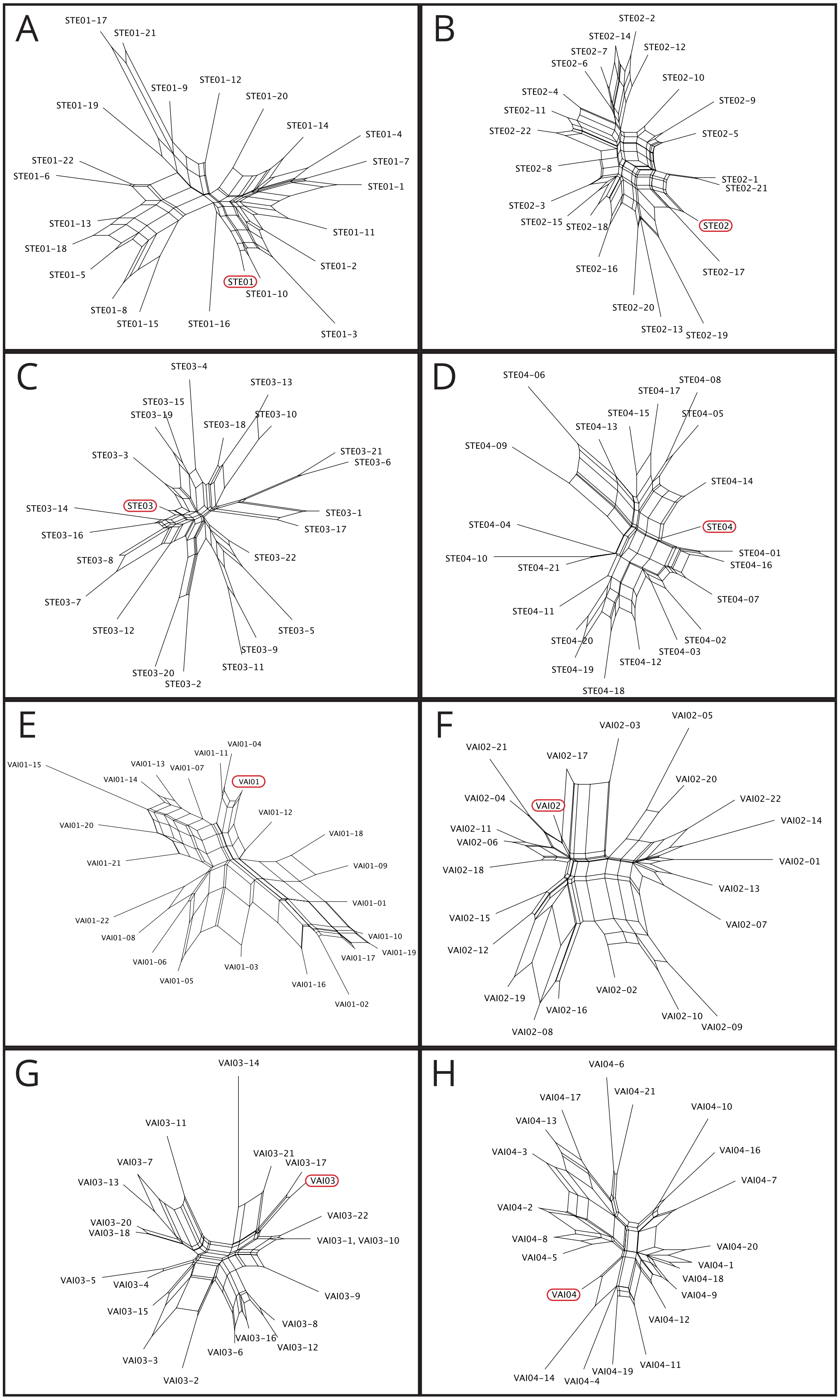

A
STE01-17
STE01-21
STE01-9
STE01-12
STE01-20
STE01-19
STE01-14
STE01-4
STE01-7
STE01-22
STE01-6
STE01-1
STE01-13
STE01-11
STE01-18
STE01-2
STE01-5
STE01
STE01-10
STE01-8
STE01-15
STE01-16
STE01-3
B
STE02-2
STE02-14
STE02-12
STE02-7
STE02-6
STE02-10
STE02-4
STE02-9
STE02-11
STE02-22
STE02-5
STE02-8
STE02-1
STE02-21
STE02-3
STE02-15
STE02-18
STE02
STE02-16
STE02-17
STE02-20
STE02-13
STE02-19
C
STE03-4
STE03-13
STE03-15
STE03-19
STE03-18
STE03-10
STE03-3
STE03-21
STE03-6
STE03-14
STE03
STE03-1
STE03-16
STE03-17
STE03-8
STE03-22
STE03-7
STE03-12
STE03-5
STE03-9
STE03-11
STE03-20
STE03-2
D
STE04-06
STE04-08
STE04-17
STE04-15
STE04-05
STE04-13
STE04-09
STE04-14
STE04-04
STE04
STE04-10
STE04-01
STE04-21
STE04-16
STE04-07
STE04-11
STE04-20
STE04-02
STE04-03
STE04-12
STE04-19
STE04-18
E
VAI01-04
VAI01-11
VAI01-15
VAI01-13
VAI01-07
VAI01-14
VAI01
VAI01-12
VAI01-20
VAI01-18
VAI01-21
VAI01-09
VAI01-01
VAI01-22
VAI01-08
VAI01-10
VAI01-19
VAI01-17
VAI01-06
VAI01-03
VAI01-05
VAI01-16
VAI01-02
F
VAI02-05
VAI02-03
VAI02-21
VAI02-17
VAI02-20
VAI02-04
VAI02
VAI02-22
VAI02-14
VAI02-11
VAI02-06
VAI02-01
VAI02-18
VAI02-13
VAI02-15
VAI02-07
VAI02-12
VAI02-02
VAI02-19
VAI02-16
VAI02-10
VAI02-08
VAI02-09
G
VAI03-14
VAI03-11
VAI03-21
VAI03-7
VAI03-17
VAI03-13
VAI03
VAI03-22
VAI03-20
VAI03-18
VAI03-1, VAI03-10
VAI03-5
VAI03-4
VAI03-9
VAI03-15
VAI03-8
VAI03-16
VAI03-12
VAI03-6
VAI03-3
VAI03-2
H
VAI04-6
VAI04-21
VAI04-17
VAI04-10
VAI04-13
VAI04-16
VAI04-3
VAI04-7
VAI04-2
VAI04-8
VAI04-20
VAI04-5
VAI04-1
VAI04-18
VAI04
VAI04-9
VAI04-12
VAI04-11
VAI04-19
VAI04-14
VAI04-4

**Figure 7** (previous page)**.** Neighbor-net graphs for each maternal parent (STE01 (A), STE02 (B), STE03 (C), STE04 (D), VAI01 (E), VAI02 (F), VAI03 (G), and VAI04 (H)) and its progeny based on pairwise genetic distances. Each maternal parent individual is highlighted with a red oval. Each panel shows a Neighbor-net graph depicting the genetic similarity among maternal plants and their progeny.

### *Discreet Genotype Pattern Analysis results*

Genetic distance-based metrics (as above) can display the genetic relationships among individuals (in this case progeny and their maternal parent and to one another), but not how they are similar or dissimilar. To better understand the nature of the differences between progeny and their maternal parent, we developed a novel analysis method that characterizes and summarises the nature of genotype similarities between progeny and the maternal parent.

Progeny pools represent a complicated mixture of (1) individuals that result from self-pollination, which does not generate progeny that are identical to the maternal parent, but an array of combinations of the parent's genotype and (2) individuals that result from outcrossing events (receiving pollen from another individual), which may deliver the same or different alleles to progeny relative to the maternal parent. As most reproduction in cocoa requires outcrossing, we would expect progeny to largely be the result of sexual reproduction with other individuals.

Ultimately, how similar or dissimilar progeny will be to the maternal parent depends on the paternal contribution. Progeny will always have at least 50% similarity to the maternal genotype (based on the genetic contribution by the egg cell), but if the paternal genotype(s) are genetically quite different from the maternal genotype, then the progeny will display more differentiation from the maternal parent's genotype (closer to 50% than 100% similarity). If, however, the paternal genotype(s) are similar to the maternal genotype (due to shared ancestry), then the progeny will receive more alleles from the paternal parent that are in common with the maternal parent. In effect, this is inbreeding. The more genetically similar the parents are, the more likely it is that the progeny will receive the same alleles (copies of a gene or locus) from both parents and, as such, would have a genotype that is more similar to the maternal parent's genotype. To assess this, we developed an analysis that compares each individual's genotype (at each locus) to its maternal parent's genotype (at each locus) and then categorises each comparison into one of four patterns (Types 0, 1, 2, and 3 described in the Methods section).

Types 1 and 3 patterns can only result from outcrossing as the progeny possess alleles not present in the maternal parent. Technically, these could represent new mutations, but, as mutation rates are relatively low per generation, it is far more likely that they are the result of outcrossing events, which is the most common form of reproduction in cocoa. As nearly all progeny examined revealed multiple loci with Type 1 and Type 3 patterns, it is clear that the vast majority of (if not all) progeny are produced through outcrossing, which is not surprising.

Figure 8 shows the summary results of this analysis for each progeny pool. In these graphs, the proportion of loci fitting the Type 0, 1, 2, or 3 pattern is plotted for each of the progeny and the maternal parent. The maternal parent was included as a control for the analysis method; because the characterizations are relative to the maternal parent, the genotype at all of each of the maternal parent's loci should be Type 0 ('identical' to the respective maternal parent).

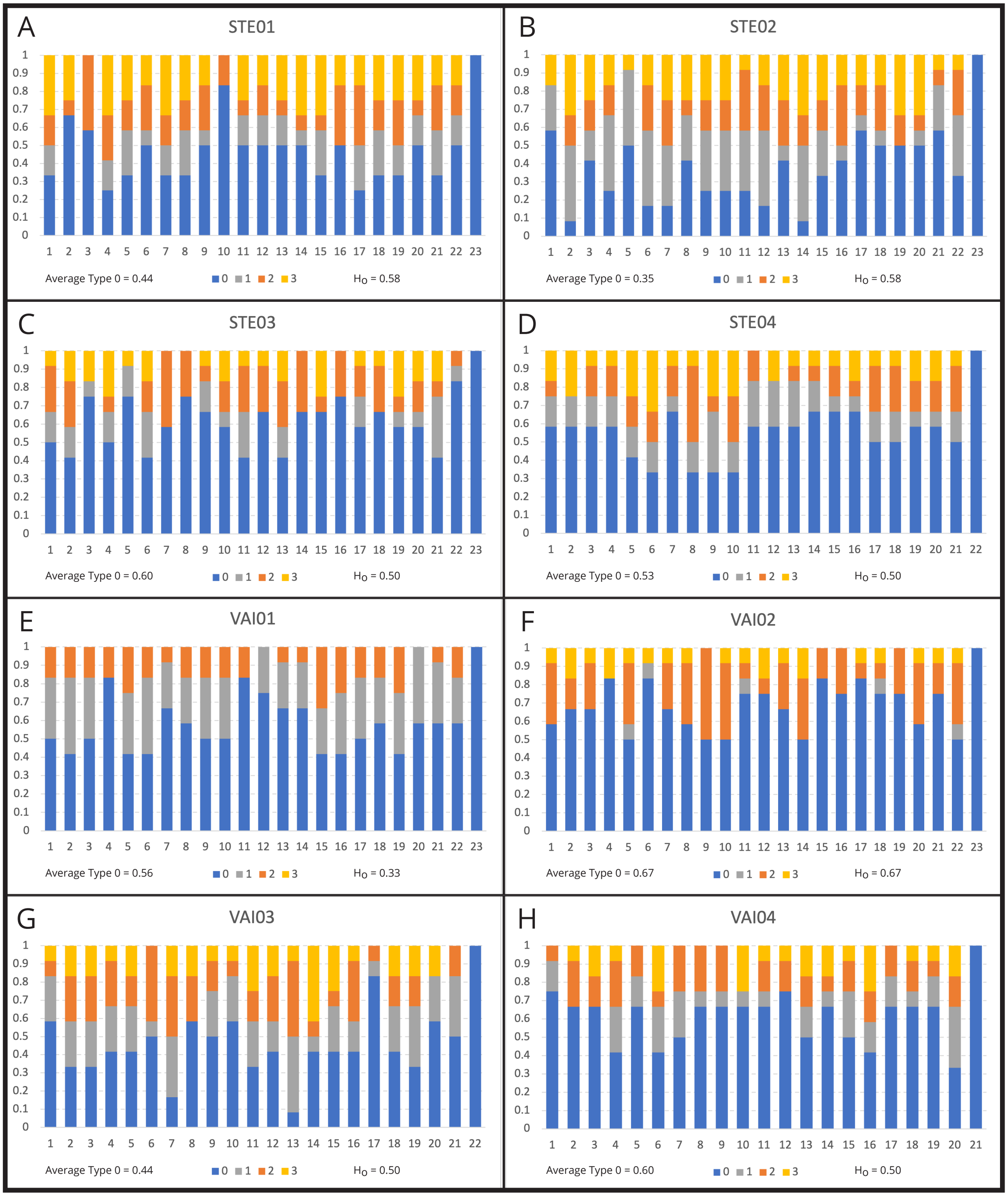


**Figure 8.** Histograms showing the proportion of loci that fit the Type 0-3 patterns described in the text for multiple progeny from each of eight maternal cocoa plants. Results for the maternal parent individual is plotted at the far right of each graph; all other individuals are the progeny. The average proportion of loci that are classified as Type 0 (exact matches of the maternal plant's genotype) across all progeny from a given maternal plant (e.g., from STE01) is indicated at the bottom left of each graph and the observed heterozygosity (H$_O$) of the maternal plant is shown at bottom right of each graph.

So how do we interpret these results? As a predominantly outcrossing species, the genetic similarity of individuals to the maternal plant will largely depend on the extent of similarity of the paternal parent(s) to the maternal parent. As it is likely that pollinators are delivering pollen from multiple plants to any one cocoa plant, there could be considerable variation among progeny (half-sibs) from the same maternal plant. A good example of this can be seen in the progeny of VAI03 where the proportion of genotypes that are identical to the maternal plant ranges from 0.08 to 0.83. This is a lot of variation from individual to individual among the progeny. The range of variation among progeny is even clearer when we sort by the proportion of loci that conform to the Type 0 pattern (loci that have an identical genotype to the maternal parent) as in Figure 9. Recall that the VAI03 progeny pool also stood out in the STRUCTURE results, showing a composition that appeared to be more broad than that of the other pools of individuals.

One of the somewhat surprising results was that the progeny of VAI01 displayed no loci conforming to the Type 3 pattern. Although at least some progeny in seven of the eight progeny pools had such a result, the VAI01 progeny pool was extreme. Upon closer examination, it was discovered that the maternal parent of this pool had the lowest observed heterozygosity ($H_O$ = 0.33) of all maternal plants examined. As the Type 3 pattern consists of genotypes that only have one of the two alleles of the maternal plant (plus a different allele contributed by the paternal parent), it makes sense that the maternal parent with the lowest heterozygosity would produce offspring with the lowest proportion of Type 3 loci. In this vein, it makes sense that this pool also has the lowest proportion of loci with the Type 2 pattern, which also depends on maternal heterozygosity.

In conclusion, there is considerable variation both among progeny pools and within progeny pools. Given the nature of open pollination typically observed in *T. cacao*, this should not be surprising. While basic genetic principles allow us to predict patterns of inheritance, the genetic changes observed between maternal plants and their progeny will depend heavily on (1) the nature of the genotype(s) of the maternal plant (homozygous vs heterozygous) and (2) the genotypes of the pollen donors (i.e. the gene pool of individuals within the range of pollen dispersal).

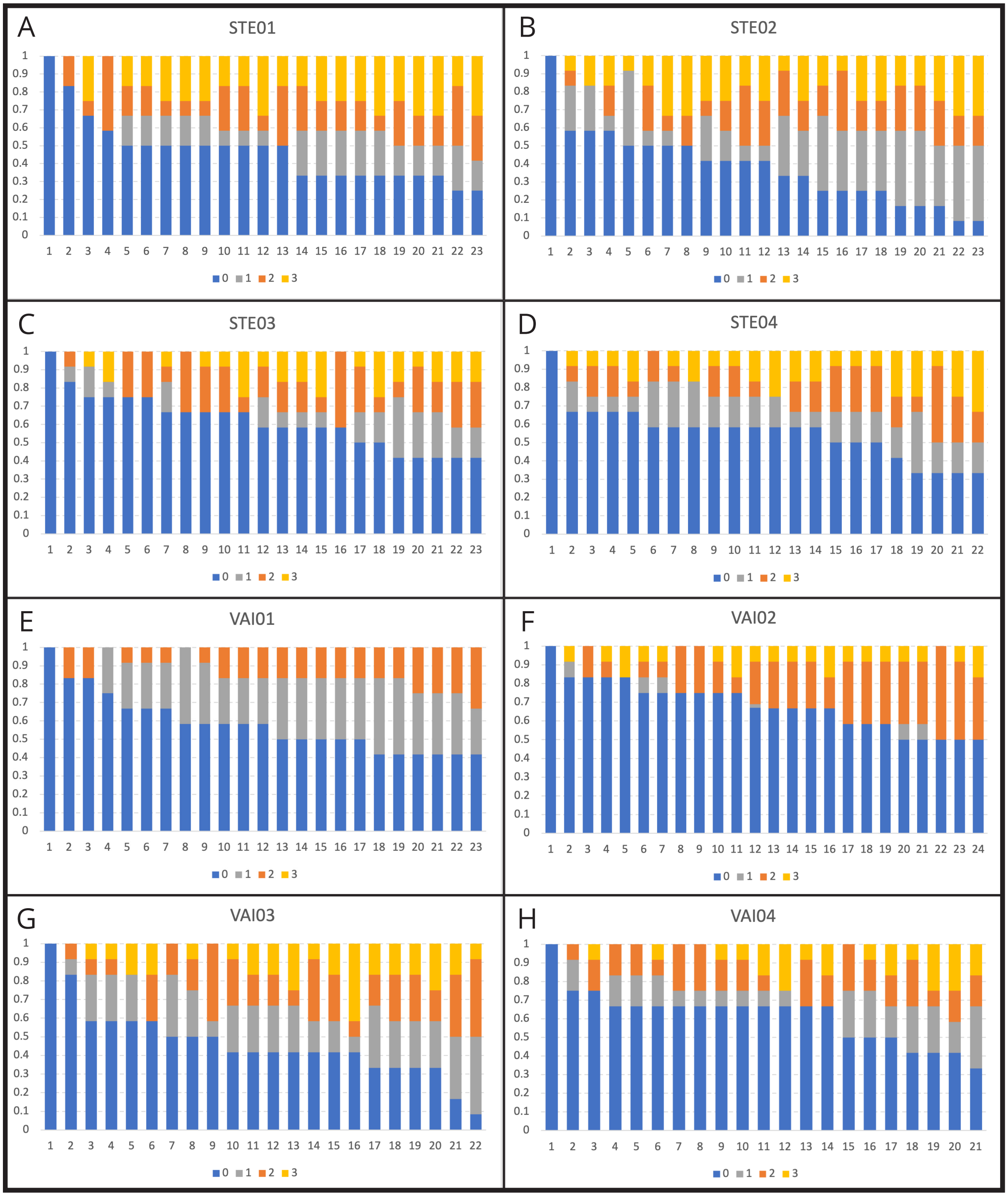


**Figure 9.** Histograms showing the proportion of loci that fit the Type 0-3 patterns described in the text for multiple progeny from each of eight maternal cocoa plants. These are the same plots as those presented in Figure 8, but are sorted from the highest to lowest frequency of Type 0 loci, which puts the maternal plant at the far left. Note that the numbers on the x-axis in this plot do not indicate sample ID.

### Genotyping the site collections for markers linked to black pod rot

Based on association mapping results from the literature, we identified six candidate microsatellite markers to trial that had been reported to be genetically linked to black pod rot resistance/susceptibility in cocoa. The ~200 plants from the four sites were genotyped for those markers; however, it has been difficult to obtain the raw genotype data from the authors of the mapping studies. While the original publications provide information on which loci (markers) are associated with disease resistance, they typically do not indicate which alleles at each locus are linked with resistance and which alleles are linked with susceptibility. As this point, we have only had feedback from one author for one marker. While transparency should be forthcoming, it has been exceedingly slow up to this point. We will continue to press those authors for transparency.

As many of the black pod resistance alleles have been discovered in the Pound-7 *T. cacao* accession, we are working to obtain tissue from that line from the C.A.T.I.E. germplasm center in Costa Rica; however, communication response times here too are remarkably slow and we cannot predict when that material will be made available to us. Once obtained, we would then be able to determine which of the Samoan plants examined carry the resistance alleles identified in Pound-7. There is the possibility though that the genetic diversity (alleles) linked to black pod resistance in the literature have not part of the history of cocoa introductions to Samoa. We point this out here as the analyses of genetic diversity from site collections indicate potentially narrower genetic diversity in Samoa relative to other locations. This is not to say that genetic diversity for resistance doesn't exist in Samoa, but it could be distinct from that identified elsewhere.

In summary, the molecular work for this part of the project is complete; however, the interpretation of the results is not possible until the acquisition of either the relevant genotype data or the Pound-7 plant material from the germplasm center (which we would genotype upon arrival. To this end, we will continue to pursue (1) the researchers responsible for the association mapping work to acquire genotype data and (2) the germplasm center for Pound-7 leaf material.

**Supplementary Table 1.** Genetic parameters and primer details for *Theobroma cacao* L. screened microsatellite loci.

| Marker Name | EMBL Acc. Number / Locus | Primer sequence (5'–3') | Chr | $T_a$(°C) | Size (bp) | Repeat structure | $N_a$ | $H_O$ | $H_E$ | Screened/ selected | | | | Source |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 1 | 2 | 3 | 4 | |
| mTcCIR1* | Y16883 | GCAGGGCAGGCTCAGTGAAGCA<br>TGGGCAACCAGAAAACGAT | 8 | 51 | 143 | $(CT)_{14}$ | 3 | 0.50 | 0.62 | * | * | | | 1, 2, 3 |
| mTcCIR2 | Y16978 | CAGGGAGCTGTGTTATTGGTCA<br>AGTTATTGTCGGCAAGGAGGAT | 5 | 51 | 254 | $(GA)_3$ $N_5$ $(AG)_2$ GG $(AG)_4$ | 3 | 0.55 | 0.51 | * | | | | 1 |
| mTcCIR6* | Y16980 | TTCCCTCTAAACTACCCTAAAT<br>TAAAGCAAAGCAATCTAACATA | 6 | 46 | 231 | $(TG)_7(GA)_{13}$ | 8, 7 | 0.54 | 0.57 | * | * | | * | 1, 2, 3, 4 |
| mTcCIR7* | Y16981 | ATGCGAATGACAACTGGT<br>GCTTTCAGTCCTTTGCTT | 7 | 51 | 160 | $(GA)_{11}$ | 6, 4 | 0.42 | 0.75 | * | | * | | 1, 2, 3 |
| mTcCIR8* | Y16982 | CTAGTTTCCCATTTACCA<br>TCCTCAGCATTTTCTTTC | 9 | 46 | 301 | $(TC)_5$ TT $(TC)_{17}$ TTT $(CT)_4$ | 5, 6 | 0.33 | 0.55 | * | * | * | | 1, 2, 3 |
| mTcCIR10 | Y16984 | ACAGATGGCCTACACACT<br>CAAGCAAGCCTCATACTC | 5 | 46 | 208 | $(TG)_{13}$ | 4 | 0.56 | 0.71 | * | | | | 1 |
| mTcCIR11* | Y16985 | TTTGGTGATTATTAGCAG<br>GATTCGATTTGATGTGAG | 2 | 46 | 298 | $(TC)_{13}$ | 9, 11 | 0.46 | 0.81 | * | * | | | 1, 2, 3 |
| mTcCIR12* | Y16986 | TCTGACCCCAAACCTGTA<br>ATTCCAGTTAAAGCACAT | 4 | 46 | 188 | $(CATA)_4$ $N_{18}$ $(TG)_6$ | 10, 11 | 0.62 | 0.87 | * | * | * | * | 1, 2, 3, 4 |
| mTcCIR15* | Y16988 | CAGCCGCCTCTTGTTAG<br>TATTTGGGATTCTTGATG | 1 | 46 | 254 | $(TC)_{19}$ | 10, 11, 13 | 0.62 | 0.84 | * | * | * | * | 1, 2, 3, 4 |
| mTcCIR18* | Y16991 | GATAGCTAAGGGGATTGAGGA<br>GGTAATTCAATCATTTGAGGATA | 4 | 51 | 345 | $(GA)_{12}$ | 8 | 0.46 | 0.72 | * | * | * | * | 1, 2, 3, 4 |
| mTcCIR22* | Y16995 | ATTCTCGCAAAAACTTAG<br>GATGGAAGGAGTGTAAATAG | 1 | 46 | 289 | $(TC)_{12}$ $N_{146}$ $(CT)_{10}$ | 4, 8 | 0.29 | 0.43 | * | | * | | 1, 2, 3 |
| mTcCIR24* | Y16996 | TTTGGGGTGATTTCTTCTGA<br>TCTGTCTCGTCTTTTGGTGA | 9 | 46 | 198 | $(AG)_{13}$ | 4, 5, 9 | 0.35 | 0.31 | * | | * | * | 1, 2, 3, 4 |
| mTcCIR26* | Y16998 | GCATTCATCAATACATTC<br>GCACTCAAAGTTCATACTAC | 8 | 46 | 298 | $(TC)_9$ C $(CT)_4$ TT $(CT)_{11}$ | 6, 8, 12 | 0.41 | 0.67 | * | * | * | * | 1, 2, 3, 4 |
| mTcCIR33* | AJ271826 | TGGGTTGAAGATTTGGT<br>CAACAATGAAAATAGGCA | 4 | 51 | 265–348 | $(TG)_{11}$ | 12 | 0.82 | 0.69 | | * | | | 2, 3 |
| mTcCIR37* | AJ271942 | CTGGGTGCTGATAGATAA<br>AATACCCTCCACACAAAT | 10 | 46 | 136–187 | $(GT)_{15}$ | 13 | 0.83 | 0.73 | | * | * | | 2, 3 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mTcCIR40* | AJ271943 | AATCCGACAGTCTTTAATC<br>CCTAGGCCAGAGAATTGA | 3 | 51 | 262–288 | $(AC)_{15}$ | 10 | 0.71 | 0.65 | * | | 2, 3 |
| mTcCIR60* | AJ271958 | CGCTACTAACAAACATCAAA<br>AGAGCAACCATCACTAATCA | 2 | 51 | 190–218 | $(CT)_{7}(CA)_{20}$ | 9 | 0.80 | 0.80 | * | * | 2, 3 |

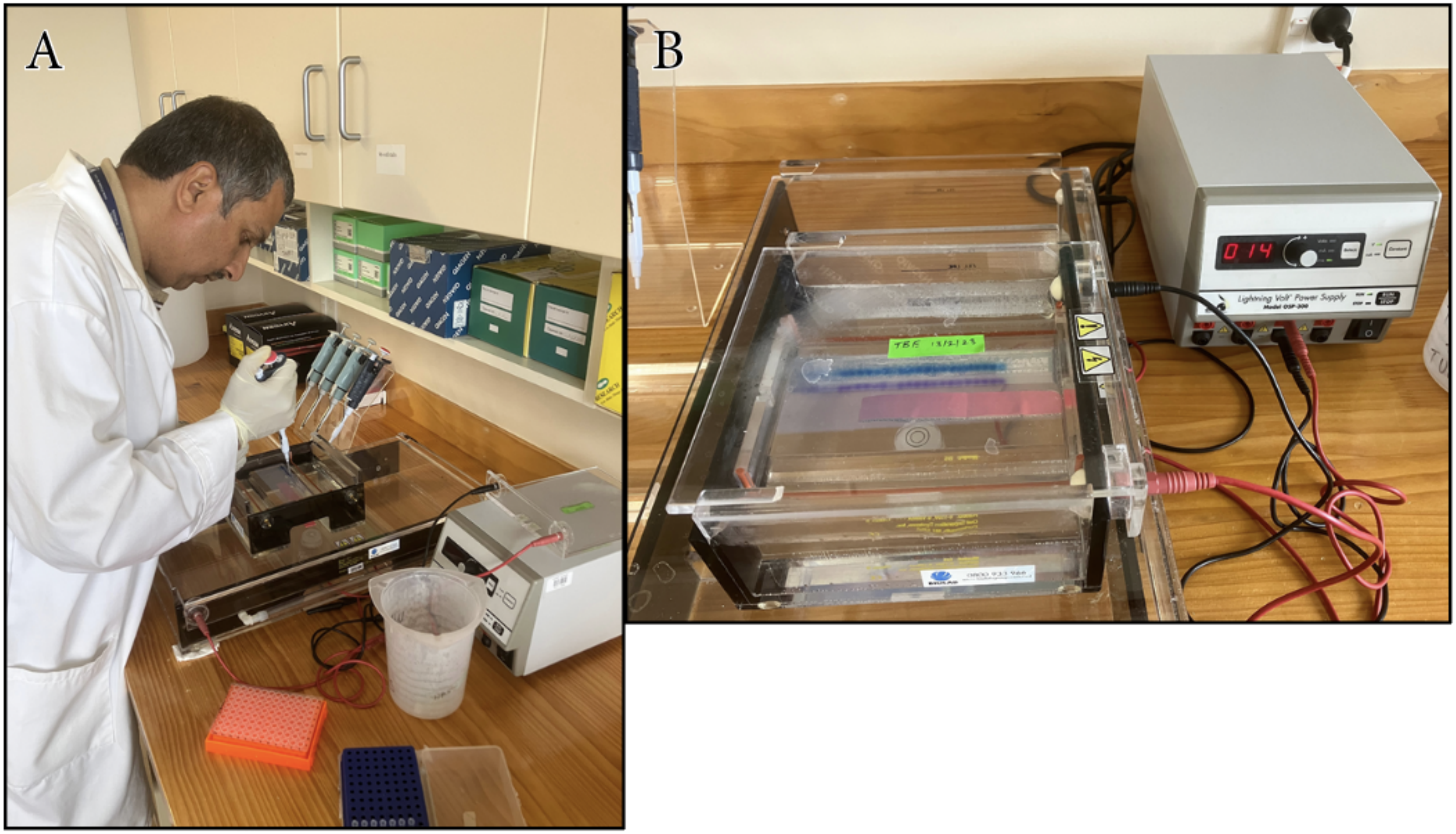
*Genetic Parameters*: $N_a$ number of alleles, $H_o$ observed heterozygosity, $H_e$ expected heterozygosity;

*Sources:* (1) Lanaud et al (1999); (2) Saunders et al (2004); (3) Everaert et al (2017); (4) Aikpokpodion et al (2009);

* Screened and selected primers (from the cited literatures) with high quality profiles and polymorphism.

**Supplementary Table 2.** Genetic parameters and primer details for *Theobroma cacao* L. screened microsatellite loci.

| Marker Name | Primer sequence (5'–3') | Reported effect | Size (bp) | Repeat structure | Source |
|---|---|---|---|---|---|
| mTcCIR61 | GCAGTCTGAAACAAGATAA<br>TGACTATAATATAAGGCGAA | 23% | | | Brown et al. 2007 |
| mtcCIR168 | GGTAGTATTGAGGTGCGTAT<br>GTGAATGAATGGATGTGAAA | 8.7% | | | Brown et al. 2007 |
| mTcCIR200 | GCCAATTTCTGACCCA<br>CTTAAAATAAGCCCAAATAC | 7.3% | | | Brown et al. 2007 |
| mTcCIR225 | AAGACAAAGGGAAGAAGA<br>AGGGGAAGAGCAAATC | 7.3% | | | |
| mTcCIR236 | GAAGTCAAAGGAAAGTCAA<br>TCAGAAAACGCAAATAAA | | | | |
| mTcCIR280 | ATTTGTCATTTGTTGTTGT<br>GCCTTGGTATTGACTGT | | | | Motilal et al. |



**Supplementary Figure 1.** Scenes from the Massey University molecular lab. (A) One of the lab technicians loads cocoa DNA samples in an agarose gel for visual analysis. (B) This image shows the agarose gel while running – the power pack at right applies an electrical charge to  the gel, which forces the DNA fragments to migrate through the gel and separate according to size. The gel is later stained with an intercalating dye that fluoresces under UV light allowing for visualization of the DNA.

Adenet S, Regina F, Rogers D, Bharath S, Argout X, et al. 2020. Study of the genetic diversity of cocoa populations (Theobroma cacao L.) of Martinique (FWI) and potential for processing and the cocoa industry. *Genetic Resources and Crop Evolution* 67: 1969-79

Aikpokpodion PO, Motamayor JC, Adetimirin VO, Adu-Ampomah Y, Ingelbrecht I, et al. 2009. Genetic diversity assessment of sub-samples of cacao, *Theobroma cacao* L. collections in West Africa using simple sequence repeats marker. *Tree Genetics & Genomes* 5: 699-711

Barreto M, Santos J, Corrêa R, Luz E, Marelli J, Souza A. 2015. Detection of genetic resistance to cocoa black pod disease caused by three Phytophthora species. *Euphytica* 206: 677-87

Bourke TV. 1992. Pests and diseases of cocoa in Western Samoa and the Philippines In *Cocoa pest and disease management in Southeast Asia and Australasia*, ed. PJ Keane, CAJ Putter, pp. 195-98. Rome: Food and Agriculture Organisation of the United Nations (FAO)

Boutin-Ganache I, Raposo M, Raymond M, Deschepper CF. 2001. M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. *Biotechniques* 31: 25-28

Brown JS, Phillips-Mora W, Power EJ, Krol C, Cervantes-Martinez C, et al. 2007. Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in Theobroma cacao L. *Crop science* 47: 1851-58

Cilas C, Bastide P. 2020. Challenges to cocoa production in the face of climate change and the spread of pests and diseases. *Agronomy* 10: 1232

Delgado-Ospina J, Molina-Hernández JB, Chaves-López C, Romanazzi G, Paparella A. 2021. The role of fungi in the cocoa production chain and the challenge of climate change. *Journal of Fungi* 7: 202

Dillon N, Hucks L, Diczbalis Y, Kete T. 2014. Evaluation of molecular marker technology for the identification of elite cocoa germplasm in the South Pacific, ACIAR, Australia

Dinarti D, Susilo AW, Meinhardt LW, Ji K, Motilal LA, et al. 2015. Genetic diversity and parentage in farmer selections of cacao from Southern Sulawesi, Indonesia revealed by microsatellite markers. *Breeding science* 65: 438-46

Dormatey R, Sun C, Ali K, Coulter JA, Bi Z, Bai J. 2020. Gene pyramiding for sustainable crop improvement against biotic and abiotic stresses. *Agronomy* 10: 1255

Droessler H. 2017. *Colonialism by deferral: Samoa under the tridominium, 1889–1899*. pp. 203-224. Emerald Publishing Limited

Eden DRA, Edwards WL. 1952. *Cocoa plantation management in Western Samoa*. Noumea: South Pacific Commission.

Everaert H, De Wever J, Tang TKH, Vu TLA, Maebe K, et al. 2020. Genetic classification of Vietnamese cacao cultivars assessed by SNP and SSR markers. *Tree Genetics & Genomes* 16: 43

Everaert H, Rottiers H, Pham PHD, Ha LTV, Nguyen TPD, et al. 2017. Molecular characterization of Vietnamese cocoa genotypes (*Theobroma cacao* L.) using microsatellite markers. *Tree Genetics & Genomes* 13: 99

Gao C. 2021. Genome engineering for crop improvement and future agriculture. *Cell* 184: 1621-35

Gopaulchan D, Motilal LA, Bekele FL, Clause S, Ariko JO, et al. 2019. Morphological and genetic diversity of cacao (Theobroma cacao L.) in Uganda. *Physiology and Molecular Biology of Plants* 25: 361-75

Gutiérrez OA, Puig AS, Phillips-Mora W, Bailey BA, Ali SS, et al. 2021. SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of Theobroma cacao L. *Tree Genetics & Genomes* 17: 28

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23: 254-67

Kozicka M, Tacconi F, Horna D, Gotor E. 2018. Forecasting cocoa yields for 2050.

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL. 1999. Isolation and characterization of microsatellites in *Theobroma cacao* L. *Molecular Ecology* 8: 2141-43

Lukman, Zhang D, Susilo AW, Dinarti D, Bailey B, et al. 2014. Genetic identity, ancestry and parentage in farmer selections of cacao from Aceh, Indonesia revealed by single nucleotide polymorphism (SNP) markers. *Tropical plant biology* 7: 133-43

Maney C, Sassen M, Hill SL. 2022. Modelling biodiversity responses to land use in areas of cocoa cultivation. *Agriculture, Ecosystems & Environment* 324: 107712

Motamayor JC, Lachenaud P, Da Silva e Mota JW, Loor R, Kuhn DN, et al. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PloS one* 3: e3311

Nei M. 1987. *Molecular evolutionary genetics*. Columbia university press.

Opoku S, Bhattacharjee R, Kolesnikova-Allen M, Motamayor J, Schnell R, et al. 2007. Genetic diversity in cocoa (Theobroma cacao L.) germplasm collection from Ghana. *Journal of Crop Improvement* 20: 73-87

Pazhamala LT, Kudapa H, Weckwerth W, Millar AH, Varshney RK. 2021. Systems biology for crop improvement. *The plant genome* 14: e20098

Peakall R, Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular ecology notes* 6: 288-95

Pohlan HAJ, Pérez VD. 2010. Growth and production of cacao  In *Soils, plant growth and crop production*

Prevosti A, Ocana J, Alonso G. 1975. Distances between populations of Drosophila subobscura, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45: 231-41

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-59

R Core Team. 2013. R: A Language and Environment for Statistical Computing.

Saunders JA, Mischke S, Leamy EA, Hemeida AA. 2004. Selection of international molecular standards for DNA fingerprinting of *Theobroma cacao*. *Theoretical and Applied Genetics* 110: 41-47

Schroth G, Harvey CA. 2007. Biodiversity conservation in cocoa production landscapes: an overview. *Biodiversity and Conservation* 16: 2237-44

Slade DA. 1984. A review of small scale cocoa production in Western Samoa. *Alafua Agricultural Bulletin* 9: 30-39

Stolberg D. 2013. German in Samoa: Historical traces of a colonial variety. *Poznan Studies in Contemporary Linguistics* 49: 321-53

Susilo AW, Zhang D, Motilal LA, MISCHKE S, MEINHARDT LW. 2011. Assessing genetic diversity in Java fine-flavor cocoa (Theobroma cacao L.) germplasm by using simple sequence repeat (SSR) markers. *Tropical Agriculture and Development* 55: 84-92

Symonds VV, Lloyd AM. 2004. A simple and inexpensive method for producing fluorescently labelled size standard. *Molecular Ecology Notes* 4: 768-71

Urquhart DH. 1952. *Cocoa growing in Western Samoa: A report on a survey made for the South Pacific Commission from 16th to 30th April, 1952*. Noumea: South Pacific Commission. 18 pp.

Urquhart DH. 1961. *Cocoa*. London: Longmans. 293 pp.

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: 621-27

Wood GAR. 1985. Cocoa  In *Cocoa*, ed. GAR Wood, RA Lass. New York: Longman Group Limited

Wright S. 1949. The genetical structure of populations. *Annals of eugenics* 15: 323-54

Yang Y, Saand MA, Huang L, Abdelaal WB, Zhang J, et al. 2021. Applications of multi-omics technologies for crop improvement. *Frontiers in Plant Science* 12: 563953

Young AM. 1994. *The chocolate tree*. Washington: Smithsonian Institution Press. 200 pp.

Zhang D, Arevalo-Gardini E, Mischke S, Zuniga-Cernades L, Barreto-Chavez A, Aguila JAD. 2006. Genetic diversity and structure of managed and semi-natural populations of cocoa (Theobroma cacao) in the Huallaga and Ucayali Valleys of Peru. *Annals of botany* 98: 647-55

Zhang F, Batley J. 2020. Exploring the application of wild species for crop improvement in a changing climate. *Current Opinion in Plant Biology* 56: 218-22