Retrospective Determination of the Parental Population of Superior Cacao (*Theobroma cacao* L.) Seedlings and Association of Microsatellite Alleles with Productivity

R.J. Schnell, C.T. Olano, J.S. Brown, A.W. Meerow, and C. Cervantes-Martinez
Subtropical Horticulture Research Station, U.S. Department of Agriculture, Agricultural Research Service, 13601 Old Cutler Road, Miami, FL 33158

C. Nagai
Hawaii Agriculture Research Center, 99-193 Aiea Heights Drive, Aiea, HI 96701

J.C. Motamayor
Mars Inc., U.S. Department of Agriculture, Agricultural Research Service, 13601 Old Cutler Road, Miami, FL 33158

ADDITIONAL INDEX WORDS. parentage analysis, molecular markers, productivity QTL, association tests

ABSTRACT. Commercial production of cacao in Hawaii is increasing, and this trend is expected to continue over the next several years. The increased acreages are being planted with seedlings from introduced and characterized cacao populations from at least three initial introductions of cacao into the islands. Productive seedlings have been selected from a planting at Waialua, Oahu. The parents of these selections were believed to be the population at the Hawaii Agriculture Research Center (HARC) at Kunia; however, potential parental populations also exist at Univ. of Hawaii research stations at Waimanalo and Malama Ki. Using microsatellite markers, we analyzed the potential parental populations to identify the parents and determine the genetic background for 99 productive and 50 unproductive seedlings from the Waialua site. Based on 19 polymorphic microsatellite loci the parental population was identified as trees from Waimanalo and not trees from Malama Ki or Kunia. The Kunia and Malama Ki populations were very similar with low allelic diversity (A = 1.92) and low unbiased gene diversity (Hnb) of 0.311 and 0.329, respectively, and were determined to be Trinitario in type. The Waimanalo, productive seedling, and unproductive seedling populations had much higher levels of genetic diversity with Hnb of 0.699, 0.686, and 0.686, respectively, and were determined to be Trinitario in type. An additional 46 microsatellite markers were amplified and analyzed in the Waimanalo parents, productive, and unproductive seedlings for a total of 65 loci. Seventeen loci contained alleles that were significantly associated with productive seedlings as determined by Armitage’s trend test. Of these, 13 loci (76.4%) co-located with previously reported quantitative trait loci for productivity traits. These markers may prove useful for marker assisted selection and demonstrate the potential of association genetic studies in perennial tree crops such as cacao.

The genus *Theobroma* L. contains 22 species. Of these, *Theobroma cacao* is the only species whose seeds are used as raw material for chocolate containing products (Cuatrecasas, 1964). Cacao is an out-crossing diploid (n = x = 10) with a genome size approximately 2.8 times larger than *Arabidopsis thaliana* (L.) Heynh. (Lanaud et al., 1992). Much genetic diversity exists within *T. cacao*; however, current commercial cultivars have a very narrow genetic base (Warren and Kennedy, 1991). The primary types of cacao grown have traditionally been referred to as Forastero, Trinitario, and Criollo, with most of the world’s production based on Forastero-derived types. The Forastero material is usually further divided into upper Amazon and lower Amazon types. The Trinitario are hybrids between Forastero and Criollo types.

Traditional cacao breeding has been only marginally successful due to the use of a narrow genetic base, mislabeling in breeding germplasm collections, and the disruption in the continuity of breeding programs (Lockwood, 2003; Motilal et al., 2002). The lack of a long-term recurrent selection program for cacao has resulted in the selection of productive seedlings as parents.

These seedlings are cloned and placed in seed gardens to mate with other selected seedlings. This “hybrid seed” is then used to establish cocoa farms. Unfortunately, as many as 80% of the genotypes from the “hybrid seeds” are unproductive in farmers’ fields (Hunter, 1990). Progressive breeding programs have selected and released clones for commercial planting; however, the establishment of cocoa farms with clones requires greater sophistication in farm management.

Microsatellite markers have been isolated and characterized for cacao, and currently over 300 microsatellite markers are available for use (Brown et al., 2003; Lanaud et al., 1999; Pugh et al., 2004). Using these and other molecular markers, genetic linkage maps have been produced and quantitative trait loci (QTL) for yield components, vigor, bean traits, ovule number, and resistance to *Phytophthora palmivora* Butler and *Crinipellis perniciosa* (Stahel) Singer, have been mapped (Brown et al., 2003; Clement et al., 2003a, 2003b). These markers have also been used for fingerprinting and to investigate the genetic diversity in current cacao breeding populations (Motamayor et al., 2003).

The U.S. Dept. of Agriculture, Agricultural Research Service (USDA-ARS) in collaboration with Mars Inc. (Hackettstown, N.J.) and other international research institutes have in progress...
a marker assisted selection (MAS) program to develop cacao cultivars with resistance to witches’ broom (C. perniciosi) and frosty pod diseases [Moniliophthora roerii] (Ciif. & Par.) Evans] (Saunders, 2002). The development and evaluation of mapping populations for QTL identification is very time consuming and current QTL identification for productivity traits and disease resistance is based on a few F1 families and a single F2 population (Brown et al., 2003; Clement et al., 2003a, 2003b). An attractive alternative approach for perennial tree crops like cacao is association mapping. Association mapping was developed for medical genetics but has been successfully applied to plants where a QTL for flowering time was associated with a DNA polymorphism in a candidate gene locus among inbred lines of maize (Thornsberry et al., 2001). Association analysis or linkage disequilibrium (LD) mapping in plants has recently been reviewed by Flint-Garcia et al. (2003) and provides an alternative method for evaluating the basis of quantitative variation in perennial fruit crops where family designs are difficult and recombinant inbred lines impossible to produce. As a major goal of the USDA/Mars breeding program is to identify QTL for disease resistance and productivity, the use of association genetics could enhance this effort.

Cacao was first introduced into Hawaii in 1850 and evaluated for commercial production between 1905 and 1918. The Hawaii Experimentation Station demonstrated that cacao produced well in sheltered localities near sea level; however, cacao remained a novelty without any commercial development (Crawford, 1937). A second introduction occurred in the early 1980s when Hawaii Vintage Chocolate established a nursery at Keaau on the island of Hawaii. Cacao plantings also exist at the Univ. of Hawaii (UH) research station at Waimanalo on the island of Oahu, the UH research station at Malama Ki on the island of Hawaii, and at Hawaiian Sugar Planters’ Association (HSPA) now Hawaii Agriculture Research Center (HARC) research station at Kunia, Oahu. Limited documentation is available as to source and type of material imported, and these populations are being used as seed sources for current commercial plantings. Commercial production of cacao in Hawaii is limited to 40 ha with ~70,000 trees planted. The industry is growing, with a significant increase in planting area predicted over the next few years. The lack of any serious diseases in Hawaii, and existence of long quarantine restrictions, increases the need to identify local material that is highly productive and has unique flavor characteristics.

While visiting the Oahu production area in 2001, it was noted that a large population at the site of the old Waialua Sugar Plantation had a strikingly high proportion of productive seedlings in comparison with previously observed seedling plantings in Central America (B.K. Matlick and R. Schnell, personal observation). These seedlings were originally planted by Dole Food Co. Hawaii in 1998. All the trees at Waialua were derived from one or more of the three populations at HARC, UH—Waimanalo, or UH—Malama Ki. Anecdotal information available suggested the HARC Kunia population as parents, but seed was also collected from the UH—Malama Ki and UH—Waimanalo research stations by workers for Dole Food Co. Hawaii.

Our objectives were to determine the most likely parental population for the Waialua planting, and then identify the parents of the productive and unproductive seedlings using microsatellite markers. Furthermore, we hypothesized that associations could be detected between microsatellite alleles and the productive vs. unproductive seedlings, and that microsatellite loci associated with productivity would co-locate near previously mapped QTLs for productivity traits.

### Materials and Methods

**Plant Materials.** Leaf material was collected from each of the three source populations and from two seedling populations: unproductive and productive selections. Candidate parent leaf material was collected from the UH research station at Malama Ki on the island of Hawaii, from HARC research station at Kunia, and the UH research station at Waimanalo, Oahu. Seedling leaf material was sampled from the Dole planting at the old Waialua Sugar Plantation, Oahu. Productive genotypes were also collected from the Hawaii Gold Cacao Tree Inc. planting at Ophihako, Hawaii; these plants were clones of selected seedlings from the original Waialua planting. The productive seedlings were selected over two years as trees that produced at least 40 pods per year. The unproductive trees were those who produced few or no pods over the 2-year period.

In total, individuals of the following backgrounds were studied: 99 productive seedlings and 50 unproductive seedlings from the Waialua planting; 76 potential parents from Kunia, 20 from Waimanalo, and 8 from Malama Ki (Table 1). Three DNA samples stored at the Subtropical Horticulture Research Station (SHRS) at Miami were also included in the study to be used as controls: ‘ICS1’, ‘SCA6’, and ‘UF273’.

DNA extraction was performed on 200 mg samples of leaf tissue using the Fast DNA kit (BIO 101, Carlsbad, Calif.) and a FastPrep FP 120 Cell Disrupter (Savant Instruments, Holbrook, N.Y.). The kit protocol for plant tissue was followed including the optional SPIN protocol. Tissue was homogenized using the Garnet Matrix and two 0.64-cm spheres as the Lysing Matrix combination, at speed 5 for 30 s, repeated three times. DNA was quantified on a DynaQuant 200 Spectrophotometer (Amersham Pharmacia, Piscataway, N.J.) All samples were diluted 1:100 to obtain a concentration of 0.1-2.5 µg·µL⁻¹.

**Microsatellite Markers and Polymerase Chain Reaction (PCR) Amplification.** The microsatellite markers used in this study were previously reported (Brown et al., 2003; Lanaud et al., 1999; Pugh et al., 2004). Initially, 19 primer pairs were used (primer set 1). An additional 46 primers pairs (primer set 2) were used after the most probable parents were identified. The additional microsatellite loci were not evenly spaced over the linkage map; instead, linkage groups containing identified productivity QTLs were more highly sampled. Microsatellite locus names and linkage groups are listed in Table 2. PCR amplification reactions were carried out in a total volume of 5 µL, containing 0.25 ng·µL⁻¹ genomic DNA. All PCR reactions contained 0.05 U/µL Amplitaq (Applied Biosystems, Foster City, Calif.), 0.2 mM dNTPs, 0.4 µM each forward and reverse primers, 2 mM·µL⁻¹ BSA, 1X GeneAmp PCR buffer (1.5 mM MgCl₂, 10 µM Tris-HCl pH 8.3, 50 mM KCl, 0.001% (w/v) gelatin). Thermal cycling profile consisted of the following: 4 min denaturation at 94 °C; followed by 33 cycles of denaturation at 94 °C for 30 s, 1 min at appropriate annealing temperature for each primer, 1 min extension at 72 °C; with a final 5 min 72 °C extension. For multiplex primer reactions the extension temperature was changed to 65 °C and final extension increased to 7 min. PCR was carried out on a DNA Engine tetrad thermal cycler (MJ Research, Watertown, Mass.). Multiplex primer reactions were performed for combinations of primers with matching annealing temperatures but differing size ranges and dye labels.

**Electrophoresis.** Capillary electrophoresis (CE) was performed on an ABI Prism 3730 Genetic Analyzer (Applied Biosystems) using Performance Optimized Polymer 7 (POP 7;
Table 1. Number of Individuals and location of cacao populations studied. HARC = Hawaii Agricultural Research Center, UH = University of Hawaii, CATIE = Centro Agronomico Tropical de Investigacion y Ensenanza, UAF = upper Amazon Forastero.

| Group               | Location               | No. | Source                        | Comments                        | Type   |
|---------------------|------------------------|-----|-------------------------------|---------------------------------|--------|
| Potential parents   | HARC research station at Kunia, Oahu | 76  | UH research station at Malama Kī, Hawaii | Selected from CATIE hybrids | Trinitario |
|                     |                        |     | Hershey Hummingbird Farms in Belize | Selected from CATIE hybrids | Trinitario |
|                     |                        | 8   | UH research station at Malama Kī, Hawaii | Seed to Keaau from Tagnanan Estate, Mindanao Philippines | UAF × Trinitario hybrid seeds |
|                     |                        | 20  | Hawaii Keaau Nursery |                                  | UAF × Trinitario hybrid seeds |
| Seedlings           | Waialua Sugar Plantation, Oahu | 20  | UH research station at Waimanalo, Oahu | Productive seedlings | UAF × Trinitario hybrid seeds |
|                     |                        | 79  | Waialua Sugar Plantation, Oahu | Clones of productive seedlings | UAF × Trinitario hybrid seeds |
|                     |                        | 50  | UH research station at Waimanalo, Oahu | Unproductive seedlings | UAF × Trinitario hybrid seeds |

9.0 (SAS Institute, Cary, N.C.) with modified Rogers’ distance (Wright, 1978).

**Population comparisons with known Central and South American populations.** Microsatellite primer set 1 is a standard set of markers that have been used in the USDA Miami lab to fingerprint 1098 accessions belonging to 75 diverse cacao populations from South and Central America. To determine the genetic origin of the Kunia and Waimanalo populations, an assignment test was performed as described in Paetkau et al. (1995). The assignment test involves calculating the expected frequency of each individual’s genotype in each of the 75 populations and subsequent assignment of each individual to the population where its expected genotype frequency is highest. Using the software Doh (Brzustowski, 2002) the product of the expected genotype frequency at 12 of the 19 loci that comprise primer set 1 was calculated to determine the expected frequency of each genotype in the 75 populations (Table 2). This test was performed using all 20 samples from Waimanalo and 20 random samples from the Kunia population.

**Parentage analysis.** Parentage analysis was performed using the program Cervus ver. 2.0 (Marshall et al., 1998; Slate et al., 2000). This software uses a simulation program to generate log-likelihood scores and provides a confidence statistic for assigning paternity. Since neither parent of the seedlings was known, the analysis was conducted in which the most-likely parent was identified. Every individual from the three locations:

Applied Biosystems). Samples were prepared immediately prior to electrophoresis by adding 1 µL of PCR product to 20 µL of deionized water and 0.05 µL of GeneScan 500 ROX size standard (Applied Biosystems), then denatured at 95 °C for 2 min, and chilled on ice. Samples were injected electrokinetically at 2 kV for 10 s and run at 15 kV and 60 °C for 16 min. Resulting data were analyzed with GeneMapper 3.0 (Applied Biosystems) for internal standard and fragment size determination, and for allelic designations.

**Diversity analysis.** Gene diversity values for each locus, and averages across all loci, were calculated using Nei’s (1987) unbiased estimate $\hat{H}$:

$$\hat{H} = n(1 – \sum \hat{\rho}^2)/n – 1$$

Where $n$ = number of individuals sampled, $\hat{\rho}$ is the frequency of the $i$th allele. The variance of this statistic was estimated as:

$$\hat{V}(\hat{H}) = 2/[(n – 2)[\sum \hat{\rho}^2 – (\sum \hat{\rho})^2] + \sum \hat{\rho}^2 – (\sum \hat{\rho})^2]$$

Unbiased gene diversity ($H_{nb}$) and observed heterozygosity ($H_{obs}$) were estimated from the allele frequencies of the five populations studied at each locus using GENETIX ver. 4.0 (Montpellier, France). The principal coordinate analyses [PCA (Gower, 1966)] were performed on allele frequencies of the data using SAS ver. 9.0 (SAS Institute, Cary, N.C.) with modified Rogers’ distance (Wright, 1978).
Table 2. Microsatellite loci, linkage group for each locus, gene diversity for loci in primer set 1, significant associations with productivity, and number of alleles for each locus for the Hawaiian cacao populations.

| Locus    | Linkage Group | Significant Association with Productivity | Number of alleles | Gene Diversity [H+/-V(H)] |
|----------|---------------|------------------------------------------|-------------------|-----------------|
| mTcCIR15 | 1             |                                          | 10                | 0.77346(0.037032) |
| mTcCIR22 | 1             |                                          | 7                 | 0.547748(0.025838) |
| mTcCIR24 | 1             |                                          | 7                 | 0.429948(0.024496) |
| mTcCIR19 | 2             |                                          | 10                | 0.568612(0.024554) |
| mTcCIR3  | 2             |                                          | 15                | 0.782966(0.040789) |
| mTcCIR60 | 2             | *                                        | 8                 | 0.69735(0.030847) |
| mTcCIR21 | 3             | *                                        | 9                 | 0.518214(0.00597) |
| mTcCIR12 | 3             |                                          | 10                | 0.754144(0.040169) |
| mTcCIR17 | 4             |                                          | 5                 | 0.60553(0.028452) |
| mTcCIR18 | 4             | *                                        | 8                 | 0.762929(0.040475) |
| mTcCIR33 | 4             |                                          | 12                | 0.683724(0.028731) |
| SHRSTc11 | 5             |                                          | 5                 | 0.627551(0.030948) |
| mTcCIR6  | 6             |                                          | 9                 | 0.738059(0.035851) |
| mTcCIR25 | 6             |                                          | 11                | 0.664851(0.028529) |
| mTcCIR56 | 7             |                                          | 14                | 0.748706(0.039331) |
| mTcCIR1  | 8             |                                          | 5                 | 0.436229(0.026261) |
| mTcCIR26 | 8             |                                          | 8                 | 0.529504(0.024982) |
| mTcCIR8  | 9             |                                          | 7                 | 0.583333(0.025749) |
| mTcCIR37 | 10            | *                                        | 13                | 0.852183(0.074516) |

| Locus    | Linkage Group | Significant Association with Productivity | Number of alleles | Locus    | Linkage Group | Significant Association with Productivity | Number of alleles |
|----------|---------------|------------------------------------------|-------------------|----------|---------------|------------------------------------------|-------------------|
| mTcCIR29 | 1             |                                          | 6                 | mTcCIR107 | 4             | *                                        | 4                 |
| mTcCIR34 | 1             |                                          | 4                 | mTcCIR115 | 4             | *                                        | 5                 |
| mTcCIR54 | 1             | *                                        | 8                 | mTcCIR199 | 4             |                                          | 11                |
| mTcCIR66 | 1             |                                          | 8                 | mTcCIR207 | 4             |                                          | 2                 |
| mTcCIR102 | 1            |                                          | 9                 | mTcCIR231 | 4             |                                          | 4                 |
| mTcCIR106 | 1            |                                          | 3                 | mTcCIR233 | 4             |                                          | 6                 |
| mTcCIR184 | 1            |                                          | 5                 | mTcCIR242 | 4             | *                                        | 4                 |
| mTcCIR249 | 1            |                                          | 14                | SHRSTc4  | 5             |                                          | 9                 |
| mTcCIR11 | 2             |                                          | 10                | mTcCIR148 | 5             |                                          | 6                 |
| mTcCIR44 | 2             |                                          | 4                 | mTcCIR232 | 5             |                                          | 6                 |
| mTcCIR073 | 2            |                                          | 4                 | mTcCIR239 | 5             |                                          | 4                 |
| mTcCIR100 | 2            | *                                        | 5                 | mTcCIR257 | 5             |                                          | 6                 |
| mTcCIR129 | 2            |                                          | 8                 | mTcCIR274 | 5             |                                          | 10                |
| mTcCIR152 | 2            | *                                        | 10                | mTcCIR279 | 5             |                                          | 3                 |
| mTcCIR162 | 2            | *                                        | 6                 | mTcCIR53  | 6             |                                          | 16                |
| mTcCIR224 | 2            |                                          | 7                 | mTcCIR136 | 6             |                                          | 10                |
| mTcCIR228 | 2            |                                          | 3                 | mTcCIR46  | 7             |                                          | 8                 |
| mTcCIR240 | 2            | *                                        | 10                | mTcCIR35  | 9             | *                                        | 6                 |
| mTcCIR261 | 2            |                                          | 7                 | mTcCIR126 | 9             |                                          | 6                 |
| mTcCIR247 | 3            | *                                        | 4                 | mTcCIR160 | 9             |                                          | 2                 |
| mTcCIR32  | 4             |                                          | 3                 | mTcCIR205 | 9             | *                                        | 5                 |
| mTcCIR76  | 4             |                                          | 13                | mTcCIR212 | 9             | *                                        | 3                 |
| mTcCIR95  | 4             |                                          | 10                | mTcCIR223 | 10            |                                          | 6                 |

\(^1\)Used in population comparison

\(^2\)Significant association with productivity found at this locus using Armitage's trend test.
Kunia, Malama Ki, and Waimanalo; as well as the control clones ‘UF273’, ‘ICS1’, and ‘SCA6’ were listed as candidate parents for the analysis.

Armitage’s Trend Test. Associations between productivity and markers were determined by Armitage’s (1955) trend test using the CASECONTROL procedure of SAS ver. 9.0 (SAS, 2002). Both primer sets were used on the Waimanalo, productive, and unproductive populations for a total of 65 loci (Table 2). After identifying loci with a significant linear (additive) trend associated with productivity vs. nonproductivity, a contingency table was created for each locus, testing the frequencies of productive vs. unproductive trees for heterogeneity among alleles at each locus. The difference(s) in frequencies were observed to identify the allele(s) most likely associated with productivity. The FREQ procedure of SAS ver. 9.0 was used to create the tables and associated chi-squared statistics.

Map Comparisons. The location of QTLs for productivity were assigned from linkage maps reported by Clement (2001). The map positions for all microsatellite markers are based on the reference linkage map as described by Pugh et al. (2004). Using the consensus map developed by Pugh et al. (2004), we were able to give approximate locations to these QTLs using microsatellite markers and restriction fragment-length polymorphism (RFLP) markers common to the respective maps. The productivity traits evaluated by Clement et al. (2003a, 2003b) included dry bean weight, bean length, shape index, number of ovules per ovary, wet bean weight, pod number, pod weight, trunk circumference, and canopy width.

Population Structure. A model-based clustering analysis was performed with the software Structure ver. 2.0 (Falush et al., 2003; Pritchard et al., 2000) to test for possible population subdivision within the 149 seedlings and 20 parental plants. Structure is a program that implements a Bayesian approach for deducing population structure from multilocus data. The analysis was performed with the admixture model on 65 microsatellite markers and allowing for correlated allele frequencies. The initial value of the Dirichlet parameter (α) to determine the degree of admixture was set to 1.0, considering the same value for all subpopulations. The starting values of the allele frequencies (λ) and rate of drift (FST) parameters of the Dirichlet prior distribution were set to 1.0 and 0.01, respectively. The λ parameter was assumed to be constant, but the FST was allowed to vary among different subpopulations. Three independent runs of the Gibbs sampler were run for each K (number of subpopulations) from 1 to 4, based on runs of 10^6 MCMC replications following a burn-in period of 20000 iterations. The symmetry of the proportions of samples assigned to each population was tested for every K using a lack of fit chi-square test. Finally, an overall chi-square test was performed with the ALLELE procedure of SAS ver. 9.0, to test for Hardy–Weinberg (HW) equilibrium at every locus.

Results

Level of Polymorphism. The original 19 microsatellite loci (primer set 1) were not highly polymorphic in these populations. The number of alleles ranged from five at locus mTcCIR1, mTcCIR17, and SHRSTc11 to 15 for locus mTcCIR3 with an average of 9.1 alleles per locus. The average gene diversity was 0.5131 (Table 2). An example of the allelic diversity is given in Fig. 1 for seven genotypes containing seven alleles for locus mTcCIR19.

The initial PCA using the five different populations and the 19 microsatellite markers clearly grouped the Malami Ki with the Kunia population and the Waimanalo with both the productive and unproductive seedling populations (Fig. 2). In the Kunia and Malama Ki populations, the average number of alleles per locus was 2.00 and 1.84 respectively, and the Hₑ was 0.32. In contrast, in the Waimanalo, productive, and unproductive seedling populations the average number of alleles per locus and Hₑ were both much higher (Table 3).

Population Origin Identification. All 20 individuals sampled from the Kunia population and three from the Waimanalo population were assigned to the Trinitario type as well as the clones ‘ICS1’ and ‘UF273’ using the assignment test. The 17 remaining individuals from Waimanalo were assigned to several upper Amazon populations as shown in Table 4. These upper Amazon populations correspond to germplasm collected by Pound (1943). Two individuals from Waimanalo could not be assigned since their expected genotype frequency was equal in two populations; however, both of these populations correspond to germplasm collected by Chalmers (1968) in the upper Amazon.

Parentage Analysis. All 19 loci comprising primer set 1 were used in the parentage analysis, and the most likely parent for each individual was identified. The three clones used as controls for the diversity analysis were identified as the most likely parent for seven of the seedlings: two seedlings had ‘ICS1’, four ‘UF273’, and one ‘SCA6’ identified as the most likely parent. The remaining 142 seedlings all had most-likely parents identified from the Waimanalo population. No most-likely parent was identified from the Kunia or Malama Ki populations (Table 5). Each of these 149 estimates was performed with a confidence level of 80%. Forty-three (29%) of the seedlings had no mismatched loci with the parent candidate while another 47 (32%) had from one to three mismatched loci; the remaining 59 (39%) had from four to seven mismatched loci with the candidate parent.

Association of Markers with Productivity. Seventeen of 65 loci were identified as having a significant linear (additive) trend with an α level of 0.05 or lower (Table 2). Of these, 13 (76.4%) were located in areas of the chromosomes previously associated with QTLs for productivity (Fig. 3). No significant association of microsatellite alleles with productive seedlings were detected for markers on linkage groups 5, 6, 7, or 8. Linkage group 1 contained one significant association with mTcCIR54, and linkage group 3 contained two significant associations with mTcCIR21 and mTcCIR247; however, these did not co-locate with any known QTLs. Linkage group 2 contained five significant associations; of these, three (mTcCIR162, mTcCIR60, and mTcCIR152) co-located with QTLs identified for bean length and pod weight, while mTcCIR240 and mTcCIR100, co-located with wet bean weight. Linkage group 4 contained five significant associations with mTcCIR111, mTcCIR18, and mTcCIR107, co-locating with a QTL for pod weight and wet bean weight; mTcCIR12 co-locating with QTLs for wet bean weight and pod number; and mTcCIR242 co-locating with a QTL for bean length. Linkage group 9 contained three significant associations: mTcCIR205 and mTcCIR212 co-located with QTLs for pod number and wet bean weight, while mTcCIR35 did not co-locate with a QTL. Linkage group 10 contained one significant association, mTcCIR223 that co-located with a QTL for canopy width (data not shown). All Chi-squared tests (Pearson’s, likelihood ratio, Mantel-Haenszel, and Fisher’s exact test) showed significant heterogeneity among the alleles across groups (productive and unproductive) for these 17 loci. The contingency tables identified at least one allele that could be responsible for discriminating between the two groups, and for some loci there were two or three alleles that could possibly be associated. A limitation of the case-control design used in...
Fig. 1. Allele sizes in base pairs detected at the mtcCIR19 microsatellite locus in seven cacao accessions illustrating seven alleles using the computer program Genotyper ver 4.0. The major peaks in each electropherogram are the PCR amplified alleles at this locus.

this study for association analysis is its potential for confounding, leading in such a case to possible artifactual correlations of allele frequencies of loci with a categorical phenotype. Undetected heterogeneity originated by population subdivision is a major confounding factor that might cause such spurious associations. To explore this possibility as a source of confounding, we tested the hypothesis of the presence of cryptic population structure using the Bayesian clustering approach developed by Pritchard et al. (2000) and Falush et al. (2003).

**POPULATION STRUCTURE.** The model-based cluster analysis showed no evidence of population structure. The conditional probability of the observed genotypic array (x), given a number of subpopulations (K), is a direct indicator of the posterior probability of having the K number of subpopulations given the observed genotypic array (x), since they are proportional in a Bayesian setup (Falush et al., 2003; Pritchard et al., 2000). For truly subdivided populations, small probabilities for K values, less than the real number of subpopulations, and larger probabilities
approaching plateaus for K values larger than the real number of populations are expected. The natural logarithm of the conditional probability Pr(X|K) did not show any abrupt change when K was varied from one to four in this study. K was varied from one to four because at most three introductions to Hawaii have occurred. The homogeneity chi-square test did not find any significant deviation ($P = 0.05$) from the symmetric proportions ($1/K$) of the sample assigned to each subpopulation for each value of K, giving evidence in favor of a completely random assignment to each subpopulation, or equivalently to no population structure. The overall chi-square test failed to reject the null hypothesis of Hardy–Weinberg equilibrium ($P = 0.05$) for 54 microsatellite markers, showing that the population was in H-W equilibrium for most loci, which is an assumption required for the Structure software to perform properly.

Table 3. Genetic variation within the three putative parental populations and the productive and unproductive seedlings of *Theobroma cacao*, across 19 microsatellite loci. $P_{0.05}$ = proportion of polymorphic loci when the most frequent allele does not exceed 95%; $A$ = mean number of alleles per locus; $H_{nb}$ = unbiased gene diversity (Nei, 1987); $H_{obs}$ = observed heterozygosity. Standard deviations are indicated in parentheses.

| Population          | Sample size | $P_{0.05}$ | $A$  | $H_{nb}$         | $H_{obs}$         |
|---------------------|-------------|------------|------|------------------|-------------------|
| Waimanalo           | 20          | 1.00       | 6.95 | 0.699(0.168)     | 0.703(0.198)      |
| Productive seedlings| 99          | 1.00       | 8.58 | 0.686(0.154)     | 0.687(0.159)      |
| Unproductive seedlings | 50       | 1.00       | 7.79 | 0.686(0.138)     | 0.597(0.123)      |
| Kunia               | 76          | 0.737      | 2.00 | 0.311(0.209)     | 0.309(0.209)      |
| Malama Kai          | 08          | 0.737      | 1.84 | 0.329(0.226)     | 0.387(0.316)      |

Table 4. Results of the cacao population assignment test with 20 random individuals from Kunia and the complete Waimanalo population compared to 75 populations from Central and South America. These populations were characterized using 12 microsatellite markers and the values are the number of individuals assigned to the population in which their genotype was more likely to occur.

| Hawaii populations | Comparison populations |
|-------------------|------------------------|
|                   | Trinitario             | Upper Amazon       |
|                   |                       |                    |
|                   |                       |                    |
|                   |                       |                    |  Chalmers and Allen collections$^a$ |
|                   |                       |                    |                |
| Kunia ($n = 20$)  | 20 | 0 | 0 | 0 | 0 |
| Waimanalo$^b$ ($n = 20$) | 3 | 6 | 2 | 3 | 4 | 2 |

$^a$Collected in the upper Amazon by F.J. Pound (Pound, 1943).
$^b$Collected by W.S. Chalmers and J.B. Allen in the upper Amazon (Chalmers, 1968).

*Individuals are assigned to a single group, although due to the high degree of hybridization, very close frequency values are obtained for more than one group.*
Table 5. Most likely parents with the number of productive and unproductive seedlings, as determined by Cervus parentage analysis of *Theobroma cacao* populations.

| Candidate parent | Location    | Productive seedlings (no.) | Unproductive seedlings (no.) |
|------------------|-------------|-----------------------------|-----------------------------|
| TC02342          | Waimanalo   | 4                           | 3                           |
| TC02343          | Waimanalo   | 9                           | 8                           |
| TC02344          | Waimanalo   | 3                           | 0                           |
| TC02345          | Waimanalo   | 4                           | 0                           |
| TC02346          | Waimanalo   | 2                           | 3                           |
| TC02347          | Waimanalo   | 5                           | 1                           |
| TC02348          | Waimanalo   | 1                           | 1                           |
| TC02349          | Waimanalo   | 11                          | 3                           |
| TC02350          | Waimanalo   | 3                           | 1                           |
| TC02351          | Waimanalo   | 14                          | 1                           |
| TC02352          | Waimanalo   | 3                           | 3                           |
| TC02353          | Waimanalo   | 15                          | 7                           |
| TC02354          | Waimanalo   | 0                           | 3                           |
| TC02355          | Waimanalo   | 3                           | 5                           |
| TC02356          | Waimanalo   | 4                           | 2                           |
| TC02357          | Waimanalo   | 3                           | 1                           |
| TC02358          | Waimanalo   | 1                           | 2                           |
| TC02359          | Waimanalo   | 2                           | 3                           |
| TC02360          | Waimanalo   | 4                           | 2                           |
| TC02361          | Waimanalo   | 2                           | 0                           |
| UF273A4          | Miami       | 4                           | 0                           |
| ICS1             | Miami       | 2                           | 0                           |
| SCA6             | Miami       | 0                           | 1                           |

**Discussion**

The PCA clearly demonstrated that the productive and unproductive seedlings at the Waialua planting originated from the UH research station at Waimanalo (Fig. 2). This was unexpected, as anecdotal information suggested the source of these seedlings was Kunia. However, in support of the PCA, the lower allelic diversity in the Kunia and Malami Ki populations is characteristic of Trinitario types (Motamayor et al., 2003), while bean size and flavor characteristics of the Waialua material are more similar to upper Amazon Forastero (UAF) types. Three trees were identified from the parentage analysis as having contributed a high number of productive seedlings: TC2349, TC2351, and TC2353 (Table 5). These trees contain 16, 14, and 13 favorable alleles, respectively, associated with productivity at the 17 significant microsatellite loci, and will be further evaluated as parents. The number of seedlings that have mismatching loci with candidate parents indicates that at least some of the plants from UH–Waimanalo have died since the seed was harvested and planted at Waialua.

The reintroduction of cacao to Hawaii occurred in the 1980s when Hawaii Vintage Chocolate as a joint venture with Kekela Enterprises Inc., Hershey Foods Corp., and Amfac Hawaii imported seed. The Malami Ki planting originated from seed sent from the Hershey farm in Belize. This material was selected from the “hybrid seed” populations produced in CATIE, Costa Rica, and is Trinitario in type. Open-pollinated seed was taken from Malami Ki and planted at Kunia. Seed was also obtained from the Taganan estate plantation on the island of Mindanao in the Philippines, and a nursery was established at Keauw on the island of Hawaii. The Waimanalo material originated from the Keauw planting. The Taganan records sent with the seeds indicate that the material was UAF × Trinitario hybrid seeds, and our analysis confirms this (Table 4). The assignment test indicates the population in which an individual’s expected genotype frequency is the highest; therefore, individuals with a high degree of hybridization will be assigned to a single group even though very close frequency values are obtained for more than one group. For most Waimanalo individuals very similar frequency values were obtained for a number of UAF populations as well as for the Trinitario population. This suggests that these individuals are the result of several cycles of hybridization and contain alleles from Trinitario types and from multiple UAF populations. Since the late 1950s these UAF populations have been used extensively in cacao breeding programs including Malaysia, which was the most likely source of the seeds planted at Taganan. For example, the three parental trees, TC2349, TC2351, and TC2353, have alleles found in Iquitos Mixed Calabacillo, Parinari, Scavina, and Trinitario populations, and this allelic diversity can only be explained by intercrossing among groups over several generations.

High tonnage and superior quality will be necessary for cacao production in Hawaii to be profitable. Selection from the Waimanalo population is expected to result in higher yielding clones. The productive seedlings were all selected as having produced over 40 pods per year. If we estimate average bean weight at 1 g with average number of beans per pod at 40, then each productive tree will produce 1.6 Kg per year of dried beans. The planting density being recommended in Hawaii is 2.5 x 2.5-m spacing with 1680 trees per hectare, giving an estimated yield of 2.7 t·ha⁻¹. Yields of up to 5 t·ha⁻¹ have been reported using high density planting with clonal material in Malaysia, while yields of 300 Kg·ha⁻¹ are common in West Africa (Lim and Phang, 1989).

Clement et al. (2003a, 2003b) detected QTLs for bean traits and ovule number in one study, and for yield components, vigor, and resistance to *P. palmivora* in a second study using two Trinitario clones (‘DR’ and ‘SS’), and one UAF clone (‘IMC78’) all crossed to a common male parent (‘Catongo’) and evaluated for 9 years. QTL analysis was carried out using composite interval mapping, and significance thresholds were defined using the permutation method of Churchill and Doerge (1994). In the first study (Clement et al., 2003a), 35 significant QTLs were discovered, but only five were common among the families and considered stable. These included one QTL on linkage group 2 and four QTLs on linkage group 4. In a second study, Clement et al. (2003b) evaluated the same families for different productivity traits. Twenty-one significant QTLs were identified: two on linkage group 1, three on linkage group 2, one on linkage group 3, eight on linkage group 4, three on linkage group 5, one on linkage group 7, two on linkage group 9, and one on linkage group 10. The microsatellites used in our investigation were not evenly spaced across the genome, but rather were concentrated on linkage groups 1, 2, 4, 5, and 9 where the greatest number of QTLs had been reported (Table 2).

Of the 17 loci with alleles associated with the productive seedlings, one was located on linkage group 1, five on linkage group 2, two on linkage group 3, five on linkage group 4, three on linkage group 9, and one on linkage group 10 (Fig. 3). The number of alleles for the 17 significant loci associated with productivity ranged from four to 13 and a single allele could be associated with the productive seedlings at four loci, while two alleles were associated for 11 loci, and three alleles at two loci.

One of the primary causes of LD is the expansion of populations from a small number of founders, such as has occurred with the population planted at Waialua. Pods were collected from the parental plants at Waimanalo, which were seedlings of the original introductions from the Philippines. The extent and nature of LD in
Fig. 3. Microsatellite markers and productivity QTL locations for linkage groups 2, 4, and 9. Microsatellite markers found to be associated with productivity traits by Armitage’s trend test indicated in boxes. Linkage map comparisons by LG indicate location of microsatellite and RFLP markers on the consensus linkage map of UPA 402 × UF 676 (Pugh et al., 2004) in blue; RFLP, AFLP, and microsatellite markers in progeny of IMC78 × Catongo in red; S52 × Catongo in green; and DR1 × Catongo in black (Clement et al., 2003a, 2003b). Productivity QTL locations indicated by vertical bars to the right of the linkage group as follows: bean length (BL, red), trunk circumference (TC, violet), canopy width (CW, pink), pod weight (PW, orange), shape index (SI, blue), wet bean weight (WBW, green), pod number (PN, brown), ovule number (OV, yellow), dry bean weight (SW, black). ** indicates significant association with productive seedlings. ++ indicates location based on F2 map (Brown et al., 2003).
cacao is unknown, but cacao is an out-crossed species with a small genome. New cacao cultivars are selected from crosses between heterozygous parents and fixed by vegetative propagation using grafting. Population admixture can cause spurious associations between a marker and a phenotype. While some of the associations found in this study may be false positives, they are not due to subgroups within the population. Based on the results from the population structure analysis no subgroups were detected, and HW equilibrium was found for 54 of the 65 loci. However, preliminary results from a Venezuelan population with Criollo and Forastero ancestors indicated that LD was significant for genetic distances <15 cM (Lanaud, 2004), which is within the size range of all QTLs identified by Clement et al. (2003a, 2003b).

Association mapping is an alternative approach to the traditional family-based mapping (F1, F2, backcross). It holds great promise in long lived perennial plants where the production of recombinant inbred lines is not possible and where phenotypic evaluation can require five to eight years of mature plant data collection. Although significant correlations of particular alleles for 17 loci and productivity variation was found in this study, the inference is conditional on the precision achieved by discrete data collected for two years. Additional quantitative trait data that includes yield and yield components collected for several years will be required to determine if these associations are real. If these associations are repeatable, the procedure will have utility in screening germplasm collections and in the way we accomplish MAS in cacao.

**Literature Cited**

Armitage, P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Brown, J.S., D. Kuhn, J.C. Motamayor, J. Borron, U. Lopes, and R. Schnell. 2003. A preliminary genomic map based on microsatellite markers and resistance gene homologues and stress-related WRKY genes in an F2 population of *Theobroma cacao* L. Proc. 14th Intl. Cocoa Res. Conf, Accra, Ghana, 13–18 Oct. 2003.

Brzustowski, J. 2002. Doh assignment test calculator. 1 Oct. 2003. <http://www2.biology.ualberta.ca/jbrzusto/Doh.php>.

Chalmers, W.S. 1968. Cacao collecting trip to the Oriente of Ecuador (November 1968). Univ. West Indies, Cacao Res. Unit, Port of Spain, Trinidad and Tobago.

Churchill, G.A and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Clement, D. 2001. QTL mapping of traits of agronomic interest in *Theobroma cacao* L. Natl. Inst. of Agronomy Paris-Grignon, France, PhD Thesis. (in French).

Clement, D., A.M. Risterucci, J.C. Motamayor, J. N’Goran, and C. Lanaud. 2003a. Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. Genome 46:103–111.

Clement, D., A.M. Risterucci, J.C. Motamayor, J. N’Goran, and C. Lanaud. 2003b. Mapping QTL for yield components, vigor, and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. Genome 46:204–212.

Crawford, D. L. 1937. Hawaii’s Crop Parade: A review of uses products derived from the soil in the Hawaiian Islands past and present. Advertiser Publishing Co., Honolulu, Hawaii. p. 73–75.

Cuatrecasas, J. 1964. Cacao and its allies: A taxonomic revision of the genus *Theobroma*. Bul. U.S. Natl. Museum 35:379–614.

Falush, D., M. Stephens, and J.K. Prichard. 2003. Inference of population structure: Extensions to linked loci and correlated allele frequencies. Genetics 164:1567–1578.

Flint-Garcia, S., J.M. Thornerisney, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Hunter, R.J. 1990. The status of Cacao (*Theobroma cacao*, Sterculiaceae) in the western hemisphere. Econ. Bot. 44(4):425–439.

Lanaud, C., P. Hammon, and C. Duperray. 1992. Estimation of the nuclear DNA content of *Theobroma cacao* by flow cytometry. Café Cacao Thé 36:3–8.

Lanaud, C., A.M. Risterucci, I. Piepertti, M. Falque, A. Bouet, and P.J.L. Lagoda. 1999. Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol. Ecol. 8:2141–2152.

Lanaud, C. 2004. Genetic Origin and movement of cocoa varieties: Consequences on genome analyses and modern breeding strategies. Amer. Assn. Advancement Sci. Annu. Conf., Seattle, Wash. 14 Feb. 2004.

Lim, D.H.K. and T.Y. Phang. 1989. Advances in cocoa planting densities/patterns with special reference to high density systems. Malaysian Cocoa Board/Malaysian Cocoa Growers Conference Workshop on Cocoa Agricultural Research, Kuala Lumpur. 16–25 July 1989.

Lockwood, R. 2003. Who needs clothing? INGENIC Nwslt. 7:2–5.

Marshall, T.C., J. Slate, L. Kruuk, and J.M. Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7:639–655.

Motamayor, J.C., A.M. Risterucci, M. Heath, and C. Lanaud. 2003. Cacao domestication II: Progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–330.

Motilal, L.A., D.R. Butler, and V. Mooleedhar. 2002. Verification in global cacao germplasm collections. INGENIC Nwslt. 7:4–8.

Nei, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

Paetkau, D., W. Calvert, I. Sterling, and C. Strobeck. 1995. Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. 4:347–354.

Pound F.J. 1943. Cocoa and witches’ broom disease (*Marasmius perniciosus*) of South America with notes on other species of *Theobroma*. Yuille’s printery, Port of Spain, Trinidad and Tobago.

Pritchard, J.K., M. Stephens, and P. Gonnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pugh T., A.M. Fount, P. Bottrier, M. Abouladze, C. Deletrez, B. Courtois, D. Clement, P. Larmande, J. N’Goran, and C. Lanaud. 2004. A new cacao linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. Theor. Appl. Gen. 108:1151–1161.

SAS Institute. 2002. The CASECONTROL procedure, p. 35–44. In: SAS/Genetics user’s guide. SAS Inst., Cary, N.C.

Saunders, J.A. 2002. The USDA program for *Theobroma cacao*: Molecular genomics, disease resistance and IPM strategies. Manufacturing Confectioner 82(9):109–118.

Slate, J., T. Marshall, and J.M. Pemberton. 2000. A retrospective assessment of the accuracy of the paternity inference program Cervus. Mol. Ecol. 9:801–808.

Thornsberry, J., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. Dwarf polymorphisms associate with variation in flowering time. Nature Genet. 28:286–289.

Warren, J.M. and A.J. Kennedy. 1991. Cocoa breeding revisited. Cocoa Growers’ Bul. 44:18–24.

Wright, S. 1978. Evolution and the genetics of populations. Vol. 4. Variability within and among natural populations. Univ. of Chicago Press, Chicago.