Tree Genetics & Genomes (2015) 11: 19
DOI 10.1007/s11295-015-0839-0

ORIGINAL PAPER

Assessing microsatellite linkage disequilibrium in wild, cultivated, and mapping populations of *Theobroma cacao* L. and its impact on association mapping

J. Conrad Stack · Stefan Royaert · Osman Gutiérrez · Chifumi Nagai · Ioná Santos Araújo Holanda · Raymond Schnell · Juan-Carlos Motamayor

Received: 29 September 2014 / Revised: 18 December 2014 / Accepted: 14 January 2015 / Published online: 7 February 2015

Abstract Linkage disequilibrium (LD) measured over the genomes of a species can provide important indications for how future association analyses should proceed. This information can be advantageous especially for slow-growing, perennial crops such as *Theobroma cacao*, where experimental crosses are inherently time-consuming and logistically expensive. While LD has been evaluated in cacao, previous work has been focused on relatively narrow genetic bases. We use microsatellite marker data collected from a uniquely diverse sample of individuals broadly covering both wild and cultivated varieties to gauge the LD present in the different cacao diversity groups and populations. We find that genome-wide LD decays far more rapidly in the wild and primitive diversity groups of cacao as compared to those representing cultivated varieties. The impact that such differences can have on association analyses is demonstrated using phenotypic data on pod color and genotypic data from two cacao populations with contrasting patterns of LD decay. Our results indicate that the more rapid LD decay in wild and primitive germplasm can lead to higher-resolution mapping intervals when compared to results from cultivated germplasm. Through simulations, we demonstrate how future association mapping analyses, comprising of cacao samples with a wild or primitive background, will likely exhibit lower LD and would be more suitable for fine-scale association mapping analyses. As many traits targeted by cacao breeders are found exclusively in wild and primitive germplasm, association mapping in wild cacao populations holds significant promise for cacao improvement through marker-assisted breeding and emphasize the need to further explore the natural diversity of Amazonian cacao.

Keywords *Theobroma cacao* · Linkage disequilibrium · Association mapping · Haplotyping

Communicated by D. Grattapaglia

Electronic supplementary material The online version of this article (doi:10.1007/s11295-015-0839-0) contains supplementary material, which is available to authorized users.

J. C. Stack · R. Schnell · J.-C. Motamayor (✉)
Mars, Incorporated, McLean, VA, USA
e-mail: juan.motamayor@effem.com

S. Royaert
Mars Center for Cocoa Science, CP55 Itajuípe, Bahia 45625-000, Brazil

O. Gutiérrez
Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA

C. Nagai
Hawaii Agriculture Research Center, Kunia, HI, USA

I. S. A. Holanda
Departamento de Ciências Vegetais, Universidade Federal Rural do Semi-Arido, BR 110—Km 47, Bairro Pres. Costa e Silva, Mossoró, RN CEP 59.625-900, Brazil

Introduction

Genetic studies of many plant species have demonstrated how qualitative or quantitative trait loci (QTL) can be accurately mapped by crossing two (highly) homozygous parents who differ for one or more traits of interest (Collard et al. 2005; Collard and Mackill 2008). The extent to which a QTL can be detected in these analyses, however, is typically a function of the size of the mapping populations resulting from these crosses as well as the degree of homozygosity (due to inbreeding) of the chosen parents, and how well they segregate for the trait of interest and their respective QTL (Mackay and Powell 2007; Mackay et al. 2009). This family-based experimental arrangement, commonly referred to simply as linkage mapping, has been successfully employed to study genotype–phenotype associations in many different plant and animal species. However, in perennial crops the time and cost of such analyses can put severe practical limits on both the number of progeny that can be maintained in a mapping population and the homozygosity of the parents, as generations of inbreeding prior to their crossing become infeasible. These limitations are especially manifest in slow-growing, tropical crops such as cacao (*Theobroma cacao* L.; Bartley 1994, 2005).

T. cacao L. is a perennial, mostly outbreeding tree species that is cultivated between latitudes 20°N and 20°S in tropical regions around the world. The seeds (referred to as beans) contained within the fruit (pods) of cacao trees are the primary ingredient in chocolate and are used extensively in the cosmetics industry. Cacao trees undergo a relatively long maturation time before starting to produce pods, 4–5 years, and they are intolerant to low temperatures, which has largely restricted experimental breeding trials developing tropical regions where logistics can be challenging (Bartley 1994; 2005). The linkage mapping studies that have been carried out in cacao have concentrated on mapping QTL within F1 or F2 mapping populations. While this work has been valuable in identifying some QTL in the cacao genome (see Lanaud et al. 2009; Motamayor et al. 2013; Royaert et al. 2011; Schnell et al. 2007), its effect on cacao breeding and research has been limited by a number of factors. These include small mapping populations derived from relatively heterozygous parents, funding inconsistencies, and the pervasive mislabeling of elite cacao varieties. As a result, the stability of identified QTL is not well-understood as many have not been properly studied in different genetic backgrounds or in different environments.

A complementary approach to linkage mapping is linkage disequilibrium (LD) mapping, or association mapping, where the former can be viewed as a special case of the latter. Instead of relying on data from a specific experimental cross or pedigree, association mapping methods exploit the natural genetic variability available in the species or in a particular population, which reflect some number of natural crosses between their members. Under certain conditions, these methods can be far more precise in identifying molecular markers strongly associated with a phenotype of interest. They have been widely used in human genetics, where experimental breeding trials are not ethically feasible, to locate disease-related genes (Cantor et al. 2010; Frazer et al. 2009). They are being increasingly used in plant genetics as well, especially as researchers look to germplasm collections for the genetic variation needed to deal with challenging issues in sustainability and food security, such as emerging disease threats and environmental change (Mackay and Powell 2007; McCouch et al. 2013). For cacao, as a perennial species, association mapping is particularly appealing as large populations of trees, comprising a wide range of genetic variation, exist in of germplasm collections and commercial plantations (Bartley 1994, 2005). Before association mapping studies can be effectively planned and carried out in cacao, it is first necessary to understand the extent and structure of LD.

LD is an observed correlation, or non-random association, between the alleles present at two or more genetic loci. Linkage, in this context, refers to the common phenomenon that the alleles observed at loci in proximity to one another which are more likely to be jointly inherited during meiosis than the alleles of loci farther apart. Assuming that recombination events occur between any two consecutive genomic bases with equal likelihood, in a randomly mating population with no migration, the LD present in a population's genome will progressively break down as recombination events accumulate over many generations and effectively decouple even tightly linked (proximal) loci. Although, in reality, other factors including migration, population (sub)structure, selection, and genetic drift also affect the level of LD in various ways (Gupta et al. 2005; Mackay and Powell 2007). Association mapping attempts to leverage the breakdown of LD between loci, regardless of its origin, to identify loci that are associated with a trait of interest. When the decay of LD is rapid as the distance between markers increases, markers which are strongly associated with the trait of interest are more likely to either be or be near the causal genetic variation (Flint-Garcia et al. 2003; Mackay and Powell 2007). In this sense, association mapping analyses can also provide higher resolution of genotype–phenotype associations when compared to linkage mapping on F1 or F2 mapping populations, where LD is generally higher over longer chromosome segments as only a few, fixed-number of generations (and recombination events) separate sampled individuals.

A large-scale characterization of LD present in cacao has not been reported. An analysis by Marcano et al. (2007) is one of few that highlights the decay of LD in two specific admixed cacao populations. The demographic histories of these two populations reflect continual, if inconsistent, cultivation (Marcano et al. 2007; Motamayor et al. 2008). While cultivated cacao variation comprises most of cacao's global biomass, they appear to derive from narrow genetic bases (Bartley 2005; Motamayor et al. 2003, 2002). Other diverse populations of cacao exist, however, as part of germplasm collections, commercial operations, and wild populations (see Turnbull and Hadley 2014) for reference). Some of these populations and collections have been genotyped with both Simple Sequence Repeat (microsatellite) and Single Nucleotide Polymorphic (SNP) markers, but analyses have focused mainly on the conservation of genetic diversity (Boza et al. 2013; Irish et al. 2010; Motilal et al. 2011; Zhang et al. 2007, 2012). Characterizing the extent of LD in these various populations is an important step toward understanding how association mapping could be best employed in cacao.

One of the most extensive sets of genetically diverse cacao samples was assembled, genotyped, and analyzed by Motamayor et al. (2008). In that study, the authors extensively describe the population structure and domestication history of cacao, identifying ten major clusters of genetic diversity representing cultivated, primitive, and wild varieties. We build upon their work, using the microsatellite genotype data presented in their study to understand the structure of LD in these ten diversity groups, as well as in two recently divergent hybrid populations. The patterns of LD are shown to be notably different for samples that were assigned to wild versus cultivated diversity groups. In wild diversity groups, LD decays rapidly with marker distance whereas in cultivated diversity groups LD decays slowly. We demonstrate how these differences in sample-wide LD can practically impact the resolution of association mapping analyses. Using microsatellite haplotypes and phenotypic data on pod coloration, a well-understood cacao phenotype (Marcano et al. 2008; Motamayor et al. 2013), we show how association mapping analyses are far less precise in an F1-mapping population from Brazil (referred to as MP01) compared to a semi-cultivated population from Hawaii (referred to as Hawaii). Finally, given the existence of varying levels of LD in wild and cultivated cacao varieties, we highlight how careful sample selection will be a crucial factor in the efficacy of future association mapping analyses.

Materials and methods

Microsatellite data

With the exception of the data obtained from the MP01 mapping population, the microsatellite genotypes used in this study have been presented and thoroughly discussed in previous work, mostly in regard to the extensive genetic diversity present in cacao and its possible origins as a domesticated crop (Motamayor et al. 2008; Schnell et al. 2005). The genetic material used to determine microsatellite genotypes was extracted from leaf tissue samples collected from each individual cacao tree. Detailed DNA extraction and genotyping protocols for the diversity group, hybrid, and Hawaii population data sets are described in the original studies where they are presented (Motamayor et al. 2008; Schnell et al. 2005) and the protocols used to extract and genotype the MP01 population were identical to those used on the diversity group and hybrid population sample sets. The microsatellite molecular markers themselves, shown in Fig. 1, are well-established, still in use, and detailed in previous studies (Brown et al. 2005; Lanaud et al. 1999; Pugh et al. 2004; Schnell et al. 2005). While some of the microsatellite loci used were characterized by a single repeated di- or tri-nucleotide element, most contained non-repeat units or multiple repeating units. For this reason, all microsatellite data are represented as sequence lengths, rather than counts of a repeated element. The physical location of each marker loci was determined by comparing microsatellite flanking sequences and their primers against a recently published, high-resolution cacao genome (Motamayor et al. 2013) using both BLAST and e-PCR (Altschul et al. 1990; Schuler 1997). All base-pair positions reported in this study indicate a physical location on this reference genome, which was built from the Matina 1–6 cultivar, the most common cultivated type of cacao and belonging to the Amelonado diversity group (Aikpokpodion et al. 2009; Efombagn et al. 2008; Motamayor et al. 2013, 2003). These physical locations are consistent with their locations on a previously published composite genetic map (Brown et al. 2008). The location of each marker is shown in Fig. 1 and summary statistics for the diversity groups, hybrid, MP01, and Hawaii populations are presented in Table 1. Table S1 provides an extended set of summary statistics.

Diversity groups and hybrid populations

Genotype data from 96 multiallelic microsatellite markers that were spread across all 10 linkage group in the cacao genome were available for 778 individuals (Fig. 1a and Table 1). These samples constitute representatives of the ten proposed diversity groups as well as two recently diverged hybrid populations created by human-mediated dispersal and cross-breeding. The samples were partitioned into these ten primary diversity groups and two hybrid populations by an analysis of the genotype data with the software STRUCTURE (Motamayor et al. 2008). Potential subgroups had also been identified within each of the top level groups by separately applying STRUCTURE to only the genotypes from those groups.

The leaf material originated from wild or primitive (i.e., geographically indigenous and cultivated on a small scale) cacao populations, as well as from tropical research institutions (i.e., Instituto Nacional Autónomo de Investigaciones Agropecuarias, Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Mars Center for Cocoa Science) and farms spanning Central America, Brazil, Peru, Columbia, French Guiana, Ecuador, Venezuela, and Ghana (Figure S1) that were sampled from 1937 to 2005 (Bartley 2005). The Amelonado, Criollo, and Nacional diversity groups comprise the most common and well-known cultivated varieties of cacao, such as "West African Amelonado" and the namesake Criollo and Nacional varieties from Central America and Ecuador, respectively (Motamayor et al. 2008). The Iquitos, Nanay, Purus, Marañón, Guiana, Contamana, and Curaray diversity groups, named after the Amazonian regions

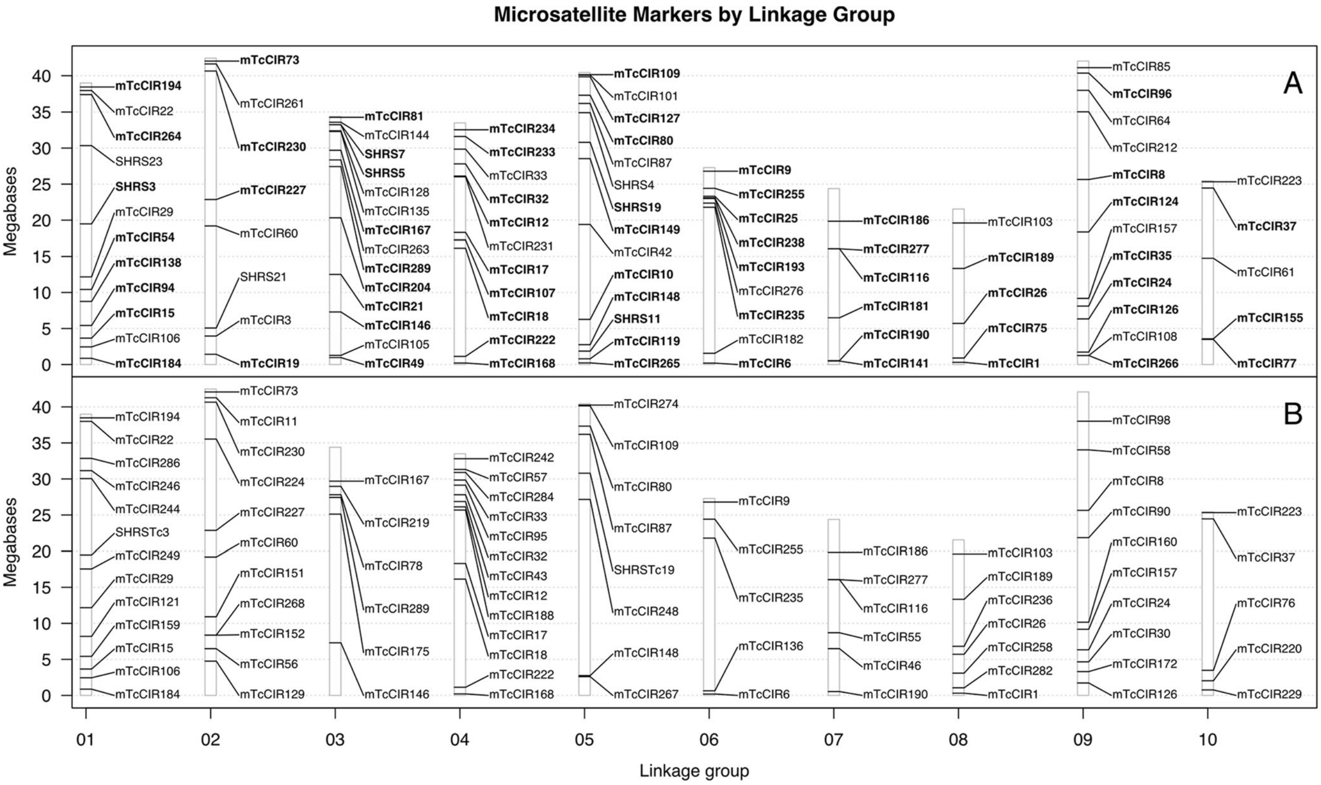

Fig. 1 The physical location of the 96 and 84 microsatellite markers used to genotype the full structured and mapping population (**a**) and the Hawaii population (**b**), respectively. All ten linkage groups are displayed on the *x* axis and their lengths (in Mega base-pairs) on the *y* axis. Microsatellite loci that were also used to genotype the MP01 mapping population have *bolded* labels (*A*)

where most of their constitutive samples were collected, represent wild and primitive cacao varieties, collected from Peru, Ecuador, Brazil, Colombia, and French Guiana (Bartley 2005; Wadsworth et al. 2003). Cacao samples that were assigned to the Trinitario hybrid population are the results of a historic admixture among Criollo and Amelonado cultivars on the island of Trinidad, where varieties from both groups were separately introduced two to three centuries ago to be cultivated as a cash crop. The "EET" (Estacion Experimental Tropical) hybrid population represents a relatively recent human-mediated introgression of wild or primitive genotypes from the Upper Amazon region (i.e., Nacional, Curaray, Marañón, Nanay, Iquitos, Contamana, and Purus diversity groups) with Trinitario varieties. These clones are the product of breeding efforts carried out in Ecuador throughout the last century and have been artificially selected for traits such as disease resistance and cocoa bean yield (Schnell et al. 2007).

MP01 mapping population

MP01 is a full-sibling mapping population established in Brazil earlier this decade from a cross between two genetically dissimilar cacao cultivars, "TSH 1188" (mother), a Trinitario Select Hybrid, and "CCN 51" (father), which segregate for a number of traits (Motamayor et al. 2013). The latter parent is a cultivar with a genetic background in which Amelonado, Criollo, and Iquitos are heavily represented (Boza et al. 2014). A randomly chosen subset of the progeny from this cross along with the two parents (N=284) were genotyped at 67 microsatellite markers, a subset of the 96 microsatellites that were used for the diversity and hybrid groups (Fig. 1a). While these data have not been published, the protocol used to genotype the samples at these microsatellite loci is identical to the one used in (Royaert et al. 2011; S. Royaert, personal communication).

Hawaii population

A population of open-pollinated cacao trees (N=151) growing in Waialua, Oahu, Hawaii with a Trinitario × Upper Amazon genetic background was genotyped at 84 microsatellite loci (Schnell et al. 2005). The progenitors of this population, a small founder population, were established by Dole Food Co. Hawaii in 1998 from local cacao seeds. These local seeds were originally transported to Hawaii from CATIE, Costa Rica, and on the basis of previous work (Schnell et al. 2005), are believed to be synthetic varieties (following (Smith 2004), varieties derived from random mating within a population being targeted for improvement via mass selection) with a genetic background indicative of Trinitario × Upper Amazon hybrids and a population demographic history involving some degree of human-mediated selection,

Table 1 Descriptive statistics of the genetic diversity groups and hybrid, MP01, and Hawaii populations

	Amelonado	Criollo	Nacional	Trinitario	EET	Guiana	Iquitos	Nanay	Marañón	Purus	Contamana	Curaray	MP01	Hawaii
Type	Cultivated			Hybrid		Primitive				Wild			Mapping	Emerging
No. Sampled Individuals	63	39	36	15	28	51	75	121	130	74	59	87	282	151
Pct. of loci, Polymorphic	89	81	97	99	100	69	100	100	99	99	100	100	100	100
Pct. of alleles missing	2.22	7.05	1.87	1.67	1.78	2.21	2.20	1.38	2.72	2.75	4.41	2.48	0.00	1.20

Included in this table are the number of samples available for each diversity group or population, the percentage of microsatellite loci that had two or more alleles (i.e., polymorphic), and the percentage of genotypes that were missing

isolation, and open pollinations. The sampled Hawaii population has expanded through natural pollination over the course of 1–2 generations since its establishment on Waialua, based on the average generation time of cacao, before leaf samples were collected for genotyping in 2004. Structural analyses indicated no evidence of population structure among the sampled individuals, who are presumed to belong to a single diversity group (K=1, where values of 1–5 were tested; Schnell et al. 2005). Slightly fewer than half of these loci overlap with those genotyped in the diversity groups (Fig. 1b).

Haplotype phasing

Linkage disequilibrium values were calculated from the haplotypes (i.e., gametes) rather than the genotypes of individuals. When haplotypes (i.e., gametic phases) are known or can be accurately inferred, the joint frequencies of alleles on the same haplotype are more comprehensively reflected in the data and can, in turn, be used to obtain more accurate estimates of the LD present throughout the genomes of a particular population (Slatkin 2008). The major limitation to estimating haplotype-based LD is that gametic phases are not regularly observed and their statistical inference is inherently uncertain. We attempt to incorporate this uncertainty into our estimates of LD by employing a bootstrapping procedure (discussed below) where LD is calculated from sets of haplotype pairs sampled according to their relative probability. Haplotype phase probabilities associated with each individual's genotype, on a given linkage group, were inferred using PHASE v2.1.1 (Stephens and Scheet 2005; Stephens et al. 2001). PHASE is generally considered to be the most accurate phasing software and is most appropriate for use with our datasets given the relatively low numbers of genetic markers per linkage group (Browning and Browning 2011). We were, however, unable to phase linkage group (LG) 05 for the diversity group and hybrid population genotypes with PHASE due the program running out of memory or not finishing after 2 months of running on a modern computing cluster. Based on the results of our sensitivity analyses (see below), the structure of LD on LG 05 was congruent with the patterns observed for the other LGs. To avoid over-complicating this manuscript and its findings, however, analyses of LG 05 for the diversity groups and hybrid populations are relegated to the supporting information. For each of the other linkage groups, genotype data were partitioned by both diversity group (the Trinitario and EET hybrid samples were phased together) and phased separately. While there are some indications that subdividing samples like this for haplotype phasing can lead to slightly less accurate haplotypes compared to phasing all samples together, the later grouping can bias downstream LD significance tests toward higher type-I error rates (Balding 2006). In addition to inferring haplotype phase, PHASE also imputes missing alleles using the pool of available alleles at a particular locus.

These pools of alleles were generally smaller when the genotype data were partitioned by their diversity group (or population, in the case of the MP01 and Hawaii samples, which were phased separately). We carried out sensitivity tests and found that sample grouping does lead to a slight difference in downstream LD values, although the overall correlation between the LD values calculated from the results of the different sample groupings was high (Spearman's ρ=0.87; *p* value <0.0001). A more extensive description and discussion of these sensitivity tests can be found in the supporting information (Appendix S1).

The number of iterations PHASE's Markov chain was set to run for differed depending on the number of markers on a given linkage group as well as on previous runs in which the chain did not converge. Most analyses needed to be run for, at minimum, 1,000 iterations (doubled on the final run of the algorithm, −X2 argument) with a thinning interval of 1,000 and a burn-in of 250 iterations. Convergence statistics and diagnostic criteria were assessed using the R package coda (v0.16.1; Plummer et al. 2006). The Markov Chain Monte Carlo runs were considered acceptable if the effective sample size was at least 500 and passed one of two diagnostic criteria, based on Heidelberger and Welch's and Geweke's convergence diagnostics implemented in the coda package (Geweke 1991; Heidelberger and Welch 1983). Unacceptable analyses were rerun, typically for more iterations. The infinite allele model option was used in PHASE (−d1 argument) as the default stepwise mutation model was not appropriate for our data set, given the complex nature of our microsatellite polymorphisms. The general model of haplotype reconstruction that we used ignores recombination as well as the assumption that linkage disequilibrium decays with distance (−MS argument). This was done so that no strong prior assumptions would be made about LD decay, although our sensitivity tests on smaller subsets of the data indicate that the choice of this model did not have a strong influence on LD values (Appendix S1). Initial segment sizes were set to be small (−l3 argument) and a different random seed value was used every time the algorithm was run. All other parameters assumed their default values. See Appendix S1 for more information on PHASE parameters and convergence diagnostics for each haplotype phasing analysis.

Sensitivity analyses were carried out to gauge the effects of modeling parameter choices and phasing program. All haplotype phasing analyses were rerun using an alternative phasing program, Beagle (v3.3.2), which produces point estimates of haplotype phases by progressively building hidden Markov models using expectation-maximization (Browning 2008; Browning and Browning 2007). While the authors of this program suggest that it is likely not as accurate as PHASE when a small number of makers (<100) are used (Browning and Browning 2011), the LD values that we obtained from Beagle haplotypes were largely similar to those obtained from PHASE (Spearman's ρ=0.85; *p* value <0.0001). These sensitivity tests are extensively discussed in Appendix S2.

Linkage disequilibrium

The linkage disequilibrium between pairs of alleles at two different loci was represented by the squared sample correlation (r^2) for multiallelic loci. The LD between two loci was then calculated as the mean of all r^2 estimates, weighted by the respective allele frequencies at each locus involved in each r^2 calculation (Hedrick 1987; Zaykin et al. 2008). The statistical significance of observed weighted r^2 values was estimated using a correlation-based test statistic, T_2, with an approximately χ^2 distribution (detailed in Zaykin et al. 2008), and confirmed by a permutation-based test (Weir 1996). After all T_2-based *p* values were calculated for all data sets, they were adjusted to control the false discovery rate due to multiple comparisons (Benjamini and Hochberg 1995). Permutations were made by randomizing, without replacement, the genotypes of each individual at a given locus, for all loci. Null distributions of weighted LD values were calculated for all pairs of loci using 1,000 permutations of the data set. If the LD value observed for the non-permuted data set was greater than or equal to the 99th percentile of their counterpart null distribution, they were considered significant. In practice, these two methods of evaluating the statistical significance of LD produced nearly identical results as suggested in (Zaykin et al. 2008).

LD calculations were generally made using the most likely haplotype pair for each individual in a given data set. The variability associated with haplotype inference was incorporated into our results through a bootstrapping procedure where LD values were calculated for a number (*N*=1,000) of resampled data sets. These resampled data sets were created by randomly selecting a haplotype pair for each individual, based on their relative probabilities (contained in the _pairs file output by PHASE). For each pair of marker loci, a 95 % equal-tail credibility interval of all resampled LD values was recorded. The most likely LD values and their associated credibility intervals are depicted in Figs. 2 and 3 as a function of the physical distance between the microsatellite loci being compared (referred to as LD decay plots). Also shown in these figures are blue trend lines representing first degree LOESS curves (Cleveland et al. 1992; smoothing α=0.35) fitted to the LD values which were significant according to the permutation tests. These lines were added to help facilitate the comparison of LD decay trends between the different diversity groups and populations.

All LD calculations, haplotype determination (see below), permutation tests, and resampling were carried out in R (Team 2013) using our own implementations (see Script S1).

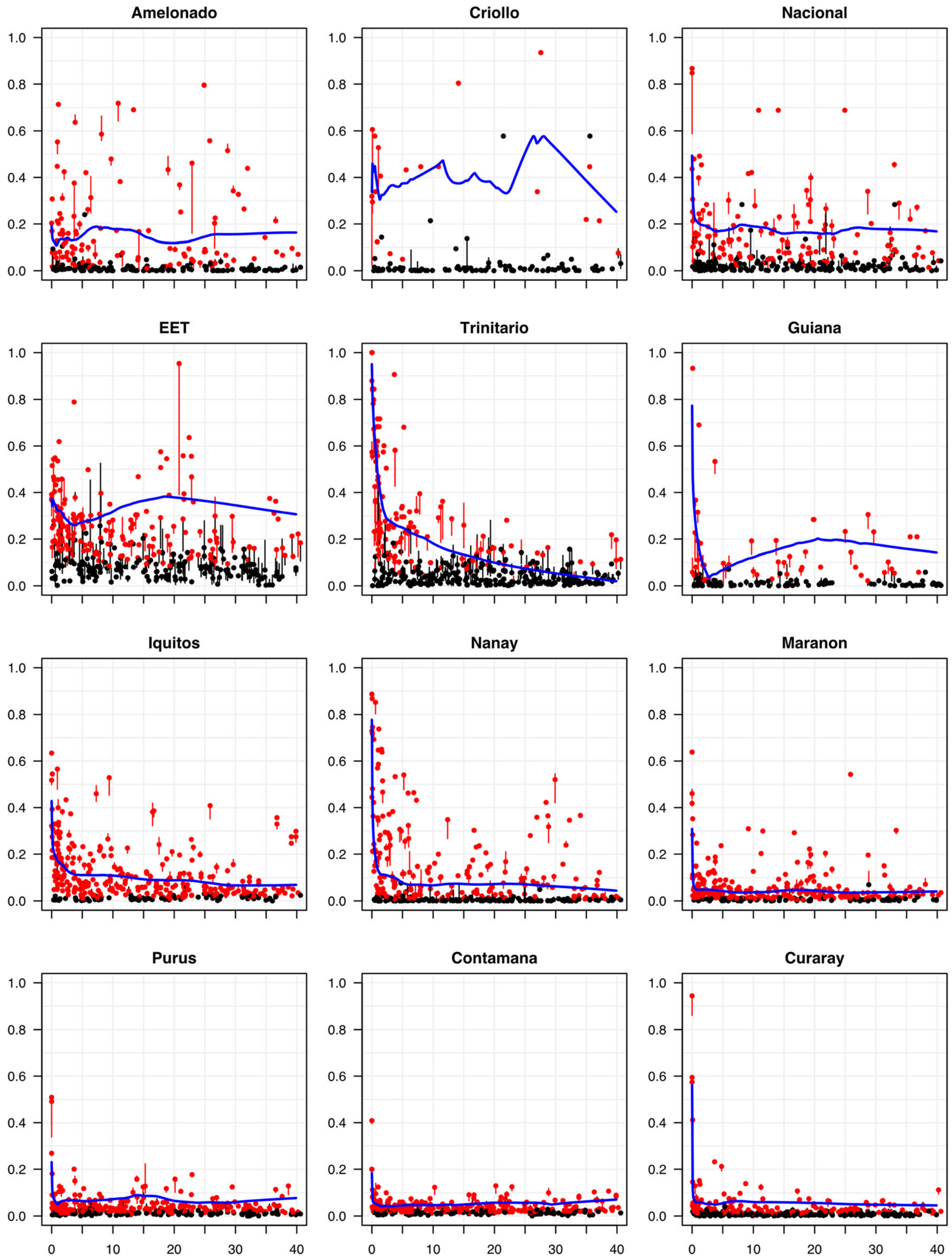

Fig. 2 Pairwise LD values plotted against the physical distance between markers for the ten primary diversity groups are shown along with the Trinitario and EET hybrid populations. The *points* on each plot indicate the weighted r^2 value (y axis) and physical distance (x axis, in Megabases) between a pair of loci on the same chromosome. *Red points* are LD values that were determined to be statistically significant (see the "Materials and methods" section), while the values represented by *black points* were not. The *vertical bars* that are visible represent the 95 % credibility intervals for LD values based on resampling the haplotype pairs (see see the "Materials and methods" section) returned by PHASE haplotype phasing analyses. *Blue trend lines* represent first degree LOESS curves (smoothing α=0.35) computed using only the significant (i.e., red) points. These *lines* are solely meant to help the reader intuit the general trend of LD's decay. All linkage groups except LG 05, where PHASE analysis failed, are presented for each diversity and hybrid population

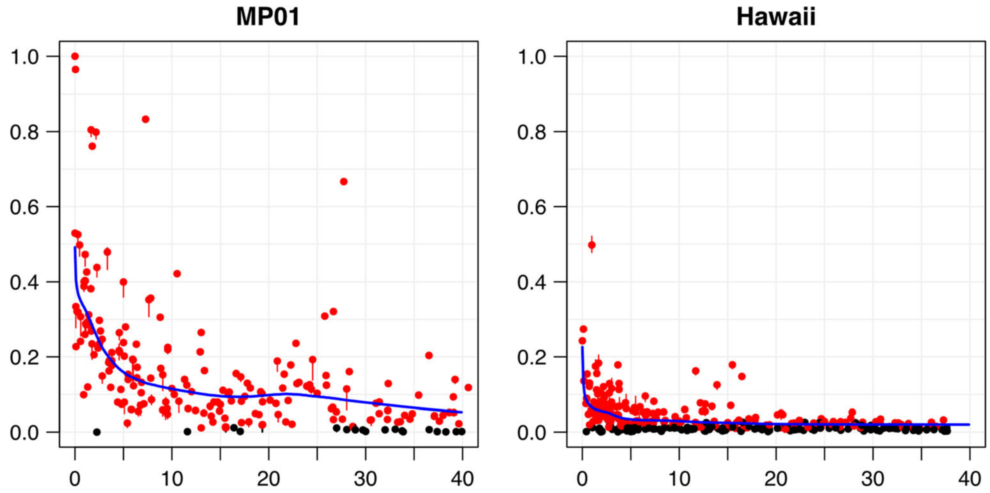

Fig. 3 Pairwise LD values (*y* axis) plotted against the physical distance between markers (in Megabases, *x* axis) for the MP01 mapping population and Hawaii population. These plots are the same stylistically as those shown in Fig. 2. In contrast to Fig. 2, however, all ten linkage groups are represented in each of the two plots (LG05 is not represented in Fig. 2). The MP01 and Hawaii populations were genotyped at fewer and different loci (see Fig. 1), and as a result, PHASE analyses were successfully carried out on all linkage group

Pod color phenotypes

For a majority of the individual cacao trees from both the MP01 and Hawaii populations, observations were made about the coloration of the fruits (referred to as pods) they produced. This trait was chosen for association analysis because its genetic mechanism is thought to be well-understood, having been thoroughly explored and mapped as part of recent research (Motamayor et al. 2013). Cacao pod color is determined by the amount of red pigment produced during pod development that mixes with its base coloration, which is green. While the eventual coloration effect of the red pigmentation is influenced by the starting shade of green, pods can be characterized as exhibiting a red or green phenotype (Bartley 2005). Because of the concealing effects of the red pigment, the red phenotype is considered to be dominant (Motamayor et al. 2013).

Cacao trees from the MP01 mapping population were assigned either a red or green pod color phenotype by researchers who observed multiple pods from each tree. These observations were made over the course of multiple growing seasons in order to confirm the stability of the phenotypes. Of the 284 individuals from the MP01 which were genotyped at the microsatellite loci described above, 234 also had pod color phenotypes (MP01: N_{green}=66, N_{red}=168). Pod color phenotypes of the trees of the Hawaii population were also judged from multiple pods per tree, collected during different growing seasons. Phenotypes were recorded for all 151 trees genotyped, but 11 tree were excluded from the association analysis because the pod color phenotype they exhibited was ambiguous (Hawaii: N_{green}=115, N_{red}=25).

Association analyses

For both the MP01 and Hawaii populations, associations between haplotypes and pod color phenotypes were made using the methods described in Schaid et al. (2002) and implemented in the haplo.stats R package (v1.6.8; Sinnwell and Schaid 2013). Pod color phenotypes were converted to a binary data type where 0 = green and 1 = red. The package function haplo.score was adapted to use the posterior haplotype probabilities reported by PHASE (Script S1) and this function was then used within the haplo.score.slide function, which runs association tests between phenotype and haplotypes comprising a contiguous subset of loci. These contiguous subsets of loci are determined by a sliding window over the loci of a given linkage group with a user-defined window length. Window lengths of 2–4 loci were evaluated and all accurately identified the known location of the pod color QTL. The haplotype effect was set as “dominant” as described in (Motamayor et al. 2013). Null distributions of the score statistics were simulated ($1,000,000\leq N_{\mathrm{permuations}}\leq 5,000,000$) for each association test. These were used as a part of each association test to compute a global *p* value, which indicates the strength of the association between pod color and the set of contiguous loci as a whole, and haplotype *p* values, which indicate strength of the association between pod color and individual haplotypes at these loci. These simulated *p* values were adjusted for multiple comparisons—numbering around 5,000—using Bonferroni’s method.

To demonstrate more generally how sample selection impacts the degree of linkage disequilibrium present in sampled genotypes, LD decay curves were calculated on genotypes resampled from the ten diversity groups. The number of genotypes resampled was $N_{\mathrm{resampled}}$=140, matching the number of phenotyped samples in the Hawaii population ($N_{\mathrm{resampled}}$= 234, equivalent to the number of phenotyped samples in MP01, was also used with no noticeable difference). The percent of resampled genotypes from Cultivated vs. Wild/ Primitive diversity groups (Table 1) was varied between

5 and 95 % in increments of 5 %. For each of these percentages, 1,000 sets of genotypes of size $N_{\mathrm{resampled}}$ were resampled from the diversity group genotypes. First-degree LOESS curves were fit to each set of resampled genotypes and then these curves were aggregated to produce a median LOESS curve along with 95 % equal-tail credibility intervals. The aggregate curves for each Cultivated vs. Wild percentage are shown in Fig. 5.

Results

Overall, there are large and meaningful differences in the genomic LD patterns exhibited by the different populations of cacao, which likely reflect their different domestication and demographic histories. A general decrease of LD with marker distance is observed in most instances, but the rate of LD decay and its stability differ significantly between cultivated, wild, and primitive diversity groups. The decay trends of LD also appear to differ by linkage group. These differences are, however, difficult to interpret given the broad range of demographic histories represented by the different diversity groups as well as likely differences in their relative mutation rates (Schlötterer 2000).

Cultivated diversity groups

The Amelonado, Criollo, and Nacional diversity groups (forming the "cultivated" superset) exhibited high r^2 values, which persist over extended physical distances. In the Amelonado group, for example, statistically significant r^2 values remain high between loci that are separated by up to 30 Mbp (Fig. 2). In the Criollo group, relatively few LD calculations were possible due to the large number of monomorphic loci (Tables 1 and S1), which is symptomatic of low allelic richness and high homozygosity. The few significant, non-zero values available, however, indicate that LD is maintained at a high level regardless of locus separation. In the Nacional group, LD between loci is maintained across all levels of marker separation, but at a slightly lower level on average than in Amelonado and Criollo. The high and persistent LD observed within these cultivated diversity groups is consistent with observations about the domestication history of cultivated cacao. Historical records and genetic evidence suggest that cultivated cacao derived from a relatively narrow genetic basis, reflecting some degree of artificial selection and inbreeding (Bartley 2005; Loor Solorzano et al. 2012; Motamayor et al. 2008, 2002). These factors are all associated with general increases in genome-wide LD and at least partly explain the overall higher levels of LD in the genomes sampled from the cultivated diversity groups (Gupta et al. 2005).

Given the extent of LD in these diversity groups, fine-scale association mapping seems impractical. Additionally, LD as a function of marker distance is erratic enough that low-resolution association analyses involving fewer genetic markers could be difficult in practice. Inconsistent correlations between different markers and casual genetic loci could lead to spurious results that would make it difficult to clearly identify and delineate associated regions.

Wild and primitive diversity groups

In general, the genome-wide LD observed in the Iquitos, Nanay, Purus, Marañón, Guiana, Contamana, and Curaray diversity groups decays far more rapidly than their cultivated counterparts; however, the stability of LD decay varies between them. On the basis of these LD decay functions, these diversity groups can be broadly categorized as either exhibiting rapid, consistent decay (Purus, Contamana, Curaray) or moderate, inconsistent decay (Iquitos, Nanay, Marañón, Guiana).

In his definitive review of cacao genetic diversity and its origins, Bartley (2005) draws a distinction between wild and primitive cacao based on the degree of human involvement involved in their establishment. For cacao groups they identified as "primitive," evidence strongly suggests that they were established or maintained by human action (e.g., trees planted along roughly straight lines), but no evidence of human involvement was observed when samples were collected. In contrast, so-called "wild" groups refer to those that were established and developed naturally. The partitioning of the non-cultivated diversity groups based on their LD presented above strongly correlates with the evidence available to classify them as either primitive or wild according to Bartley's criteria (Bartley 2005). Thus, the Purus, Contamana and Curaray diversity groups form a "wild" superset, while Iquitos, Nanay, Marañón, and Guiana form a "primitive" superset.

The wild diversity groups all exhibit very low overall LD, which rapidly decays within 1–2 Mbps to an r^2 value around 0.1 (Fig. 2). With microsatellite genotypes sampled from these diversity groups, genotype–phenotype associations could be localized in the genome with relatively high precision using an association mapping approach. QTL with a smaller effect on (i.e., less highly correlated with) the trait of interest could also be identified. Detecting smaller QTL would, however, require a substantial increase in the marker density to roughly one marker per Megabase (assuming a sample size equivalent to those shown in Table 1). The number of microsatellite markers developed for cacao diversity or linkage analyses to date ($N<500$) might not be sufficient in number or uniformity of coverage.

The trends of LD decay observed in the primitive diversity groups are also shown to be relatively rapid, with the LOESS trend of r^2 dropping to a value of 0.1 within 5–10 Mb. In contrast to the wild diversity groups, individual LD values appear to deviate more greatly from these trends, sporadically rising above 0.2 as the distance between marker loci increases (Fig. 2). The unusual trend of LD in the Guiana diversity group as indicated by the blue LOESS curve is likely an artefact resulting from the small number of significant LD values used in the regression analysis. Similar to Criollo, the samples from Guiana diversity group are monomorphic at a relatively large number of loci (Table 1). The association mapping situation for these diversity groups is similar to the wild groups, with some additional caveats. Based on trends of LD decay, samples taken from primitive diversity groups appear as good candidates for genome-wide association mapping. Due to their slightly higher level of LD across chromosomes fewer markers would be necessary to identify associations, but this higher level of background LD "noise" would likely preclude the detection of smaller-effect QTL (e.g., the QTL of complex traits). The instances in the primitive diversity group sample of high LD between marker loci that are separated by moderate or large distances would be a complicating factor, however. Such instances appear most frequently between markers on LGs 03 and 09 (Fig. S2) and could be a reflection of their cultivated origins. In association analyses, these highly aberrant LD values (with respect to their LOESS trends) could manifest as additional QTL for a particular trait, which would be difficult to rule out as spurious.

While Motamayor et al. (2008; Efombagn et al. 2008) was able to distinguish between wild and primitive diversity groups using a genome-wide set of microsatellite genotypes, they are historically difficult to segregate based on phenotypic and passport data (as noted in Bartley 2005). Without extensive genetic pre-screening of candidate populations therefore, samples chosen to represent either wild or primitive diversity groups for the purpose of association mapping are likely to represent a mix of both. The LD present in the samples reflecting a mixture of both wild and primitive diversity groups is demonstrated in Fig. 5. These patterns as discussed further below.

Hybrid populations

The Trinitario hybrid population appeared on the island of Trinidad in the 1750s (Bartley 2005) from natural crosses between the Criollo and Amelonado diversity groups (Motamayor et al. 2003). Per the standard farming practices of cacao at the time, subsequent generations of the Trinitario population were also the results of open, sexual reproduction among the trees. The LD decay observed in the samples from this population is very gradual with increasing marker distance, dropping below 0.1 at around 30 Mbps (Fig. 2). This pattern of decay in LD is comparable to that observed in the primitive diversity group, which is consistent with their histories of human-mediated cultivation (Bartley 2005). In contrast, the LD decay in the EET hybrid population strongly resembles the LD patterns present in the samples of cultivated diversity groups, with an overall high level of LD across all distances (Fig. 2). The samples from the EET hybrid population are the products of relatively few generations of human-mediated crosses made within the last century. These crosses were attempts by breeders to introgress disease resistance phenotypes from wild Upper Amazon clones into local, high yielding breeding stock with a Trinitario genetic background (Loor et al. 2009). It is difficult to make any generalizations about association analyses using the EET population as the number of samples was very small (Table 1). Also, the small sample size is reflected in the substantial amount of uncertainty present in LD estimates (Fig. 2).

LD decay and association mapping in two cacao populations

To demonstrate the relationship between LD decay and association mapping in cacao in a practical setting, two association mapping analyses were carried out using genotype data from two distinct populations. The LD in the MP01 population decays slowly with marker distance, but sporadic high LD values are observed at larger distances, particularly on LG 05 (Figs. 3 and S2). This slow decay of LD is not surprising given the single generation (and single meiosis event) separating the majority of samples, the progeny, from their parents. While the exact genealogical history of the samples constituting the Hawaii population is unknown, they are believed to have originated from a genetically diverse set of founder trees, mating through natural pollination (Schnell et al. 2005). LD decays rapidly in this population to a level below 0.1 and, with a few exceptions, remains below this level as marker distance increases (Fig. 3). This pattern of LD decay is comparable to those of the wild diversity groups, notably Purus and Contamana (Fig. 2).

The Pod color phenotype in cacao have been mapped to LG 04 in various genetic backgrounds, including the MP01 population (Motamayor et al. 2013), using both family-based linkage and association mapping approaches (Brown et al. 2007; Marcano et al. 2008; Motamayor et al. 2013). In the most recent of these studies, extensive analysis of the trait led to the identification of four SNP variants (between 20,878,891 and 20,879,148 bp) within a MYB transcription factor gene (*TcMYB113*), which are strongly believed to affect pod color differences between cacao varieties (Motamayor et al. 2013). On the basis of this and the earlier work, pod color in cacao is considered to be a monogenic trait, attributable to a narrow genetic region with a strong phenotypic effect (i.e., explains a large degree of phenotypic variation).

Separate association mapping analyses were carried out for the MP01 and Hawaii populations, correlating their genotype data with phenotypic data collected on pod color. In the MP01 population, the first 8 out of 9 total microsatellite loci on LG04 were highly associated with the pod color phenotype (Fig. 4a).

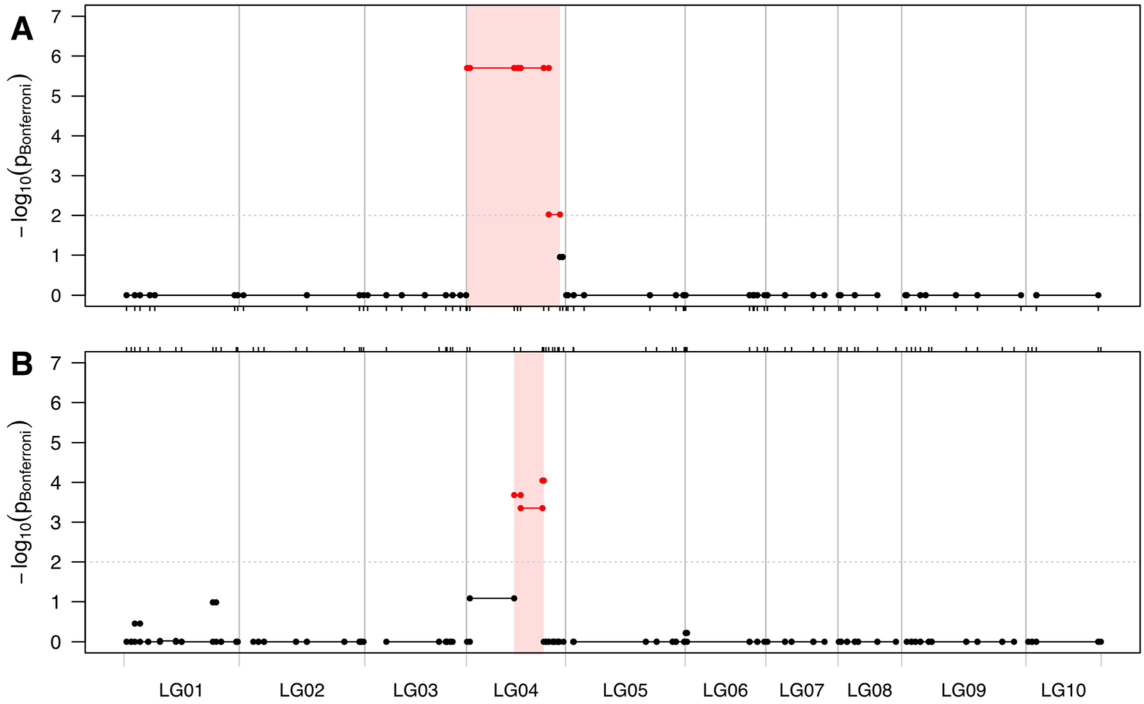

Fig. 4 Associations between haplotypes and pod color phenotypes for two populations. Association analyses are shown for the MP01 (**a**) and Hawaii (**b**) populations using a sliding window length of two loci. In both populations, pod color maps (correctly) to linkage group (LG) 04. The $-\log_{10}$ of the *p* values (after adjustment for multiple comparisons) are on the *y* axes while the physical positions of the microsatellite markers are on the *x* axes. Association values on the *y* axis that are greater or equal to 2.0 (i.e., adjusted *p* value≤0.01) are colored *red* along with the microsatellite loci they represent. A *transparent red band* spans the physical range of the loci on linkage group 4 that are non-randomly associated with pod color. In the Hawaii population (**b**), this range is far narrower than on the one observed in the MP01 population (**a**) and indicates the location of the genetic variation responsible for pod color with greater precision

These loci were located in a genetic region on LG04 ranging from 0.2 to 31.6 Mb. In the Hawaii population, 4 out of 13 total microsatellite loci on LG04 were highly associated with the pod color phenotype (Fig. 4b). These loci delimit a genetic region on LG04 from 16.1–26.1 Mb, or roughly 4.7–5.3 Mb from the putative causal genetic variation, which is consistent with the LD decay seen in Fig. 3.

The results are highly consistent with both earlier analyses of this trait and the patterns of genome-wide LD exhibited by both samples sets. The association analyses carried out using both the MP01 and Hawaii population samples were accurate in the sense that strong associations were observed exclusively on LG04, where the putatively causal genetic has been identified. There is also a clear consistency between the alleles present in MP01 and Hawaii populations that were associated with pod color and previous work, which identified the genetic variation responsible for red pod color as having likely originated in the Criollo diversity group (Motamayor et al. 2013). In both populations, haplotypes strongly associated with red pod coloration (the dominant phenotype) were observed most frequently in Criollo, Contamana, and Guiana diversity groups (Table 2). The Contamana diversity group is closely related to the Criollo group (Motamayor et al. 2008) and red pod color has also been reported on the Guiana diversity group (Lachenaud and Motamayor 2004). See Appendix S3 for figures depicting allele counts by diversity group for loci associated with red pod color.

The results from these association analyses clearly demonstrate how the resolving power of association is a function of the degree of LD in a population. LD in the MP01 population persists over long physical distances (Fig. 3) and the presence of extended blocks of LD is not uncommon (Appendix S4). Given that MP01 is a (relatively large) full-sib mapping population, these patterns are not unusual, but the practical effect of this relatedness is clearly highlighted in the degree of precision observed in the association mapping analysis (Fig. 4a). The slow decay of LD observed in LG04 (Fig. S2) indicates that too few generations have passed to enable recombination to decouple adjacent loci by breaking up haplotypes. In contrast, the association analysis for the Hawaii population provides higher resolution, despite the fact that far fewer individuals were analyzed compared to the MP01 population (140 vs. 234). Of the 13 markers on LG04 genotyped in the Hawaii population, only 4 were associated with pod color, including the two markers (mTcCIR18 and mTcCIR17) that are closest to the putative causal allele. The Hawaii population was genotyped at more markers on LG 04 (13 vs. 9); however, the distributions of loci used for each population are qualitatively similar (Fig. 1). Given the known location of the causative genetic variation (~20.8 Mb on LG 04), the only additional

Table 2 Results from haplotype score tests which were run on the genotypic data from the Hawaii and MP01 populations and pod color phenotypes

	Haplotype score tests			Haplotypes of microsatellite loci on LG04					Frequency of haplotype in each diversity group									
	p-value	test statistic	n.slide	mTcCIR18	mTcCIR107	mTcCIR17	mTcCIR188	mTcCIR12	Am.	Cri.	Nac.	Iqu.	Nan.	Pur.	Mar.	Gui.	Cont.	Cur.
Hawaii	6.00E-10	6.18	2	335		271			0	0	0.28	0.12	0.02	0	0.01	**0.47**	0.06	0.06
	4.00E-09	5.87	2			281	129		0	**0.85**	0.01	0	0	0.02	0	0	0	0.12
	3.00E-10	6.3	2				129	187	0	**1.00**	0	0	0	0	0	0	0	0
	1.30E-05	4.3	3	331		281	129		0	**1.00**	0	0	0	0	0	0	0	0
	3.50E-12	6.9	3	335		271	114		0	0	0.28	0.12	0.02	0	0	**0.47**	0.06	0.06
	4.00E-11	6.6	4	335		271	114	211	0	0	0	0	0	0	0	0	**1.00**	0
MP01	3.00E-13	7.3	3	331	111	281			0	**1.00**	0	0	0	0	0	0	0	0
	7.00E-28	−10.94	3	333	117	271			0	0	0	0.16	**0.73**	0	0	0.10	0	0
	4.00E-14	8.6	3	331	111	281			0	**1.00**	0	0	0	0	0	0	0	0
	2.00E-11	6.6	4	331	111	281		187	0	**1.00**	0	0	0	0	0	0	0	0
	2.00E-12	−7	4	333	117	271		199	0	0	0	0.10	**0.82**	0	0	0.09	0	0
	1.50E-14	7.6	4	331	117	281		211	(haplotype not found in any diversity group)									
	2.20E-18	−8.75	4	333	117	271		203	0	0	0	**1.00**	0	0	0	0	0	0

Each row displays haplotypes that were significantly associated with either red (positive valued test statistic) or green (negative valued test statistic) pod color. These haplotypes contain 2–4 markers (n. slide column), which were the ranges we tested. The Hawaii population samples were not genotyped at mTcCIR107 and the MP01 population was not genotyped at locus mTcCIR188. The haplotype frequencies are displayed on the right side of the table and indicate the diversity groups (abbreviated column names) they represent. The highest frequencies for each haplotype is shown in bold with shading to indicate the pod color phenotype it is associated with.

marker, mTcCIR188, that improves the resolution of the Hawaii association analysis does so by less than 0.5 Mb. Thus, we believe that is unlikely that the overall higher resolution of the association in the Hawaii samples versus the MP01 samples is a product of the different numbers of microsatellite markers. It is also does not seem that markers unique to one population or the other are a contributing factor as 6 of 8 markers associated with pod color in MP01 are also present in the Hawaii genotype data.

The trade-off between marker density and the power to detect associations is also manifest in these results. The associations between pod color and genotype data in the MP01 population are generally orders of magnitude larger than those observed in the Hawaii population (Fig. 4). For phenotypes with less of their variation attributable to genotype, the rapid decay of LD in the Hawaii population would likely necessitate a higher density of markers covering LG 04 to detect associations.

Future association analyses in cacao

Given that clear differences can be observed in the association mapping analyses using the MP01 and Hawaii populations, it is possible to make some generalizations about the efficacy of association mapping using the similar genotype data from the wild, primitive, and cultivated diversity groups. While passport information can often be used to classify cacao trees as cultivated or not, Bartley (2005) discusses the difficulties of distinguishing between wild and primitive cacao trees. For this reason, we analyzed LD decay patterns of a large number of simulated data sets consisting of different proportions of samples from wild-and-primitive and cultivated diversity groups.

These LD decay trends, shown in Fig. 5, reflect the general trends in LD decay among the diversity groups, where wild and primitive samples exhibited much more rapid LD decay than their cultivated counterparts. When samples are predominantly selected from cultivated diversity groups (Fig. 5, top-most line), the LD is relatively high, remaining well-above 0.10 across linkage groups, more closely resembling the LD decay of the MP01 and cultivated diversity group samples. As the percentage of genotypes sampled from wild and primitive diversity groups increases, the general trend of LD decay becomes more similar to the Hawaii and wild and primitive diversity group samples. Figure 5 also demonstrates the strong impact that these differences could have on future association analyses, and as a corollary, the importance of sample selection. The broader pattern of shifting LD decay trends with different compositions of sampled varieties suggests that association mapping analyses would be more efficacious when using samples representing traditionally wild and primitive cacao diversity groups rather than samples from cultivated groups.

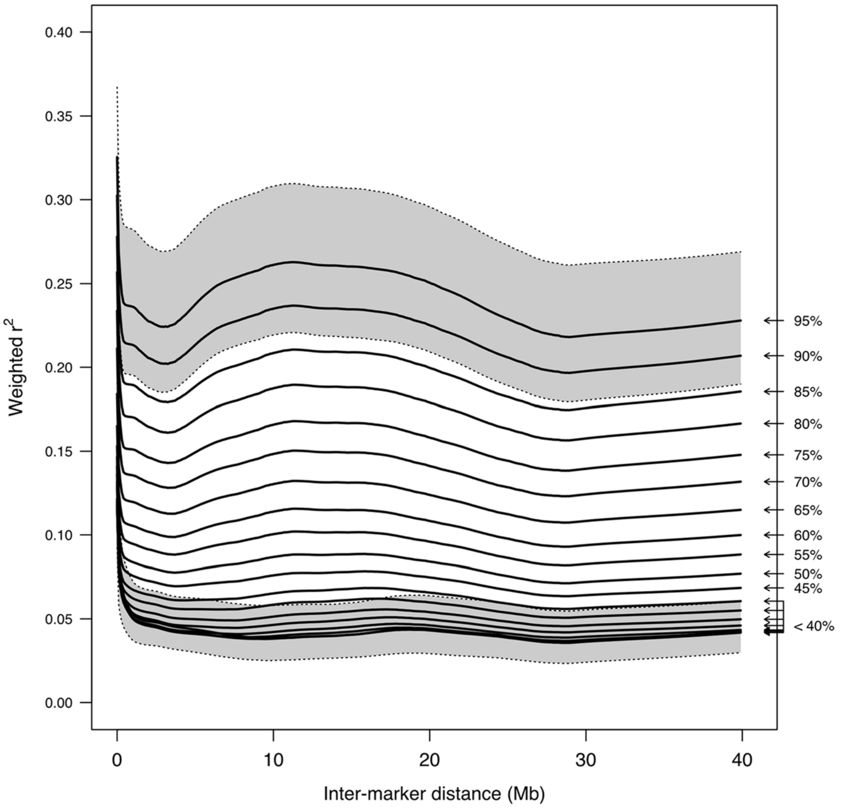

Fig. 5 Aggregate LOESS curves for datasets with different proportions of wild versus cultivated samples. The axes on this figure are similar to those used in Figs. 2 and 3 where the *y* axis shows weighted r^2 values and the *x* axis shows the distance between microsatellite markers in Megabases (Mb). Multiple aggregate LOESS curves are plotted and the percentage values on the right-hand *y* axis indicate the proportion of individuals resampled from cultivated diversity groups. The 95 % equal-tail credibility intervals for the top-most (95 %) and bottom-most (5 %) curves are also shown as *shaded bands*, bracketed by *dotted lines*, in order to indicate, very generally, the variance of LD decay among the resampled data sets

Discussion

There is a substantial amount of variation among perennial crop species regarding the patterns of linkage disequilibrium observed in cultivated versus wild varieties. The rapid decay of LD in the wild and primitive diversity groups of cacao and the slow or non-existent decay in cultivated groups (Fig. 2) resemble similar patterns observed for peach (Li et al. 2013) and cherry (Arunyawat et al. 2012). In contrast, wild and cultivated varieties of grapevine exhibit very similar patterns of genome-wide LD (Myles et al. 2011). A similar study in *Coffea canephora* (coffee) showed overall lower and rapidly decaying LD and higher genetic diversity in cultivated populations when compared to wild varieties (Civetta et al. 2009; Cubry et al. 2013). The domestication processes that underlie these genetic differences between cultivated crops and their wild cousins are generally not well understood. Many complex factors, such as the length of juvenile phases, natural population structures, and modes of reproduction in natural and cultivated populations, can impact the genetic diversity present in different populations and therefore how they should be treated for the purposes of association mapping analyses.

For *T. cacao*, we find that wild and primitive varieties are highly suited for population-level association mapping analyses. This result seems especially auspicious considering the substantial amount of genetic diversity in cacao that has only been observed in wild and primitive varieties (Motamayor et al. 2008). Many such varieties are currently maintained in germplasm collections around the world, where phenotypes have been recorded for a number of traits. Efforts are currently on-going as well to genotype many germplasm collections (e.g., Boza et al. 2013; Irish et al. 2010; Ji et al. 2013; Motilal et al. 2012; Motilal et al. 2011; Zhang et al. 2012). While these efforts have focused mainly on genotyping a large number of individuals at only a few (SNP) loci for purposes of conservation, these genotype data could be a valuable aid in selecting samples for association analyses. Thoughtful mining these germplasm in more depth could provide crucial insights into the genetic architecture underlying agricultural traits of immediate importance, such as disease resistance and robustness to changing climate conditions (McCouch et al. 2013).

Some additional caveats need to be noted regarding our findings, however. The genetic simplicity of the pod color phenotype was useful for the purposes of this study. Evidence suggests, however, that other phenotypes of cacao are likely products of far more complex interactions between genes, regulatory elements and the environment. Pod color also has a very strong genetic component compared to other cacao phenotypes (e.g., Brown et al. 2007, 2005). While this makes it difficult to make specific recommendations about marker density, it does suggest that association analyses will require a higher density to maintain the power to detect associations. Future genome-wide analyses will likely rely on high-density SNP arrays. Also, cultivated cacao, like many crops subject to artificial selection for particular agronomic traits, the degree of inbreeding observed is typically greater than that observed in their wild counterparts. The traditional varieties Criollo, Amelonado, and Nacional are self-compatible (Bartley 2005) and most cultivated cacao inherited some proportion of these traditional varieties (Motamayor et al. 2003). The differences among cultivated and wild accessions, reflected in our analysis of LD, appear to be relatively consistent with this logic. However, exceptions to this general pattern have been observed or at least suspected (Bartley 2005) and it is possible that indiscriminate sampling from either wild or primitive or cultivated populations will result in a sample without the presupposed, desired properties regarding LD. As our understanding of the mating systems and demographic behaviors of cacao (and its pollination vectors) continues to grow, the impact of this potential issue should become clearer.

The overall objective of this study was to build a more comprehensive picture of the genetic variability in different cacao populations and to highlight how population-level association mapping analyses might be employed to help improve cacao as a crop. Wild and primitive cacao varieties hold a great deal of promise for cacao genetics and are currently under-utilized in this field. As the cost of genotyping decreases, it will be possible to explore these varieties in more depth, and through association studies, could significantly add to our understanding of the relationship between cacao genotype and various phenotypes.

Acknowledgments JCM originally conceived the research presented here and contributed to the study design. JCS contributed to the study design, ran all analyses and wrote the R scripts. SR and ISAH collected phenotype and microsatellite genotype data for the MP01 population. CN, OG and RS contributed phenotype and microsatellite genotype data for the Hawaii population. JCS, OG, RS and JCM and contributed to the writing of this manuscript.

Ethical standards The authors declare that all experiments described herein comply with the law of each governmental unit in which they were carried out.

Conflict of interest The funder (Mars, Incorporated) provided support in the form of salaries for authors JCS, SR, RS and JCM, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the "Acknowledgement" section.

Data archiving statement Genotype data for the Diversity group, Hybrid, Hawaii and MP01 cacao accessions, along with accompanying phenotype data, are available for download at http://www.cacaogenomedb.org/microsatellites.

References

Aikpokpodion PO et al (2009) Genetic diversity assessment of sub-samples of cacao, Theobroma cacao L. collections in West Africa using simple sequence repeats marker. Tree Genet Genomes 5:699–711

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arunyawat U, Capdeville G, Decroocq V, Mariette S (2012) Linkage disequilibrium in French wild cherry germplasm and worldwide sweet cherry germplasm. Tree Genet Genomes 8:737–755

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791

Bartley BGD (1994) International Workshop on Cocoa Breeding Strategies

Bartley BGD (2005) The utilization of the genetic resources (1 ed). CABI Publishing, Wallingford, UK

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300

Boza EJ et al (2013) Genetic diversity, conservation, and utilization of Theobroma cacao L.: genetic resources in the Dominican Republic. Genet Resour Crop Evol 60:605–619

Boza EJ et al (2014) Genetic characterization of the cacao cultivar CCN 51: its impact and significance on global cacao improvement and production. J Am Soc Hortic Sci 139(2):219–229

Brown JS, Phillips-Mora W, Power EJ, Krol C, Cervantes-Martinez C, Motamayor JC, Schnell RJ (2007) Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in Theobroma cacao L. Crop Sci 47:1851–1858. doi:10.2135/cropsci2006.11.0753

Brown JS, Sautter RT, Olano CT, Borrone JW, Kuhn DN, Motamayor J, Schnell RJ (2008) A composite linkage map from three crosses between commercial clones of cacao, Theobroma cacao L. Trop Plant Biol 1:120–130

Brown JS, Schnell RJ, Motamayor JC, Lopes U, Kuhn DN, Borrone JW (2005) Resistance gene mapping for witches' broom disease in Theobroma cacao L. in an F2 population using SSR markers and candidate genes. J Am Soc Hortic Sci 130:366–373

Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124:439–450

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12:703–714

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86:6–22. doi:10.1016/j.ajhg.2009.11.017

Civetta A et al (2009) Genetic differentiation of wild and cultivated populations: diversity of Coffea canephora Pierre in Uganda. Genome/National Research Council Canada=Genome/Conseil national de recherches Canada 52:634–646

Cleveland WS, Grosse E, Shyu WM (1992) Local regression models Statistical models in S:309–376

Collard B, Jahufer M, Brouwer J, Pang E (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica 142:169–196

Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans Royal Soc B: Biol Sci 363:557–572

Cubry P et al (2013) An initial assessment of linkage disequilibrium (LD) in coffee trees: LD patterns in groups of Coffea canephora Pierre using microsatellite analysis. BMC Genomics 14:10

Efombagn IBM et al (2008) Genetic diversity and structure of farm and GenBank accessions of cacao (Theobroma cacao L.) in Cameroon revealed by microsatellite markers. Tree Genet Genomes 4:821–831

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Annu Rev Plant Biol 54:357–374. doi:10.1146/annurev.arplant.54.031902.134907

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251. doi:10.1038/nrg2554

Geweke J (1991) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Federal Reserve Bank of Minneapolis, Research Department

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology 57:461–485. doi:10.1007/s11103-005-0257-z

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

Heidelberger P, Welch PD (1983) Simulation run length control in the presence of an initial transient. Oper Res 31:1109–1144

Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station Cacao (L.) Germplasm. Collect Crop Sci 50:656–667

Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P, Meinhardt LW (2013) Genetic diversity and parentage in farmer varieties of cacao (Theobroma cacao L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genet Resour Crop Evol 60:441–453

Lachenaud P, Motamayor J (2004) Red pods in progenies from the Euleupousing River in French Guiana. INGENIC Newsletter 9:12–15

Lanaud C et al (2009) A meta-QTL analysis of disease resistance traits of Theobroma cacao L. Mol Breed 24:361–374. doi:10.1007/s11032-009-9297-4

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJ (1999) Isolation and characterization of microsatellites in Theobroma cacao L. Mol Ecol 8:2141–2143

Li X-W et al (2013) Peach genetic resources: diversity, population structure and linkage disequilibrium. BMC Genet 14:84

Loor R et al (2009) Tracing the native ancestors of the modern Theobroma cacao L. population in Ecuador. Tree Genet Genomes 5:421–433

Loor Solorzano RG et al (2012) Insight into the wild origin, migration and domestication history of the fine flavour Nacional Theobroma cacao L. variety from Ecuador. PLoS One 7:e48438. doi:10.1371/journal.pone.0048438

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63. doi:10.1016/j.tplants.2006.12.001

Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics 10:565–577. doi:10.1038/nrg2612

Marcano M et al (2008) A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (Theobroma cacao L.) population. Tree Genet Genomes 5:329–337. doi:10.1007/s11295-008-0185-6

Marcano M et al (2007) Adding value to cocoa (Theobroma cacao L.) germplasm information with domestication history and admixture mapping. Theor Appl Genet 114(5):877–884

McCouch S et al (2013) Agriculture: feeding the future. Nature 499:23–24

Motamayor J et al (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol 14:R53

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

Motamayor JC, Lachenaud P, da Silva e Mota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). PLoS ONE 3:e3311

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity 89:380–386

Motilal LA, Zhang D, Umaharan P, Boccara M, Mischke S, Sankar A, Meinhardt LW (2012) Elucidation of genetic identity and population structure of cacao germplasm within an international cacao genebank. Plant Genetic Resources 10:232

Motilal LA, Zhang D, Umaharan P, Mischke S, Pinney S, Meinhardt LW (2011) Microsatellite fingerprinting in the International Cocoa Genebank, Trinidad: accession and plot homogeneity information for germplasm management. Plant Genet Resour 9:430–438

Myles S et al (2011) Genetic structure and domestication history of the grape. Proc Natl Acad Sci 108:3530–3535

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R news 6:7–11

Pugh T et al (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 108:1151–1161. doi:10.1007/s00122-003-1533-4

Royaert S et al (2011) Identification of marker-trait associations for self-compatibility in a segregating mapping population of Theobroma cacao L. Tree Genet Genomes 7:1159–1168

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434

Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365–371

Schnell RJ, Kuhn DN, Brown JS, Olano CT, Phillips-Mora W, Amores FM, Motamayor JC (2007) Development of a marker assisted selection program for cacao. Phytopathology 97:1664–1669. doi:10.1094/PHYTO-97-12-1664

Schnell RJ, Olano CT, Brown JS, Meerow AW, Cervantes-Martinez C, Nagai C, Motamayor JC (2005) Retrospective determination of the parental population of superior cacao (Theobroma cacao L.) seedlings and association of microsatellite alleles with productivity. J Am Soc Hortic Sci 130:181–190

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res 7:541–550

Sinnwell JP, Schaid DJ (2013) Haplo. stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous, 1.6.3 edn

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. doi:10.1038/nrg2361

Smith SE (2004) Breeding synthetic cultivars. In: Encyclopedia of Plant and Crop Science. pp 205–206

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462. doi:10.1086/428594

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Team RC (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Turnbull C, Hadley P (2014) International Cocoa Germplasm Database (ICGD). NYSE Liffe/CRA Ltd./University of Reading, UK. http://www.icgd.rdg.ac.uk/

Wadsworth R, Ford C, Turnbull C, Hadley P (2003) International cocoa germplasm Database v. 5.2 Euronext liffe/University of Reading, UK

Weir BS (1996) Genetic data analysis 2: methods for discrete population genetic data Sinauer, Massachusetts

Zaykin DV, Pudovkin A, Weir BS (2008) Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180:533–545. doi:10.1534/genetics.108.089409

Zhang D, Boccara M, Motilal L, Butler DR, Umaharan P, Mischke S, Meinhardt L (2007) Microsatellite variation and population structure in the "Refractario" cacao of Ecuador. Conserv Genet 9:327–337. doi:10.1007/s10592-007-9345-8

Zhang D et al (2012) Genetic diversity and spatial structure in a new distinct Theobroma cacao L. population in Bolivia. Genet Resour Crop Evol 59:239–252