DNA Research **19**, 23–35, (2012) doi:10.1093/dnares/dsr039
Advance Access Publication on 30 December 2011

Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L.

Mathilde Allegre[1,†], Xavier Argout[1,*,†], Michel Boccara[1,2], Olivier Fouet[1], Yolande Roguet[1], Aurélie Bérard[3], Jean Marc Thévenin[4], Aurélie Chauveau[3], Ronan Rivallan[1], Didier Clement[1,5], Brigitte Courtois[1], Karina Gramacho[5], Anne Boland-Augé[3], Mathias Tahi[6], Pathmanathan Umaharan[2], Dominique Brunel[3], and Claire Lanaud[1]

CIRAD, UMR 1334 AGAP, TA 108/03-34398, Montpellier Cedex 5, France[1]; University of the West Indies, Cocoa Research Unit (CRU), St Augustine, Trinidad and Tobago[2]; INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, 2, rue Gaston Crémieux, CP5724, 91057 Evry, France[3]; CIRAD, Biological Systems Department, UPR Bioagresseurs, 97387 Kourou Cedex, French Guiana[4]; Comissão Executiva de Planejamento da Lavoura Cacaueira (CEPLAC), Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00, Bahia, Brazil[5] and Centre National de la Recherche Agronomique (CNRA), B.P. 808, Divo, Côte d'Ivoire[6]

*To whom correspondence should be addressed. Fax. +33 4-67-61-56-05. Email: xavier.argout@cirad.fr

Edited by Dr. Satoshi Tabata
(Received 30 June 2011; accepted 11 October 2011)

Abstract

***Theobroma cacao* is an economically important tree of several tropical countries. Its genetic improvement is essential to provide protection against major diseases and improve chocolate quality. We discovered and mapped new expressed sequence tag-single nucleotide polymorphism (EST-SNP) and simple sequence repeat (SSR) markers and constructed a high-density genetic map. By screening 149 650 ESTs, 5246 SNPs were detected *in silico*, of which 1536 corresponded to genes with a putative function, while 851 had a clear polymorphic pattern across a collection of genetic resources. In addition, 409 new SSR markers were detected on the Criollo genome. Lastly, 681 new EST-SNPs and 163 new SSRs were added to the pre-existing 418 co-dominant markers to construct a large consensus genetic map. This high-density map and the set of new genetic markers identified in this study are a milestone in cocoa genomics and for marker-assisted breeding. The data are available at http://tropgenedb.cirad.fr.**

Key words: *Theobroma cacao*; genetic map; SNP; molecular marker

1. Introduction

Theobroma cacao L. is a diploid species ($2n = 2x = 20$) with a small genome ranging in size from 411 to 494 Mb.[1] According to Cheesman,[2] its centre of origin is at the lower eastern equatorial slopes of the Andes.

Theobroma cacao is grown as a major cash crop that provides income to 14 million small-scale farmers in more than 50 tropical countries. However, cocoa production is markedly affected by a number of major diseases caused by several *Phytophthora* species, or by *Moniliophthora perniciosa* and *Moniliophthora roreri*. Several sources of disease resistance have been identified and the search for sustainable disease resistance by cumulating the different resistance genes is one of the major challenges facing

† These authors contributed equally to this work.

T. cacao-breeding programmes.[3] The quality of chocolate is another important trait in cocoa breeding, and consumer demand for high-quality chocolate is increasing. A better understanding of the molecular and genetic bases of these traits is a key goal of cocoa genetic research.

High-density genetic maps are essential tools for trait genetic studies. Several molecular marker types have been developed in *T. cacao* in recent decades: restriction fragment length polymorphism (RFLP), microsatellites or simple sequence repeats (SSRs), random amplified polymorphic DNA, amplified fragment length polymorphism and isozymes.[4–6] Among them, only RFLP, SSR and single-nucleotide polymorphism (SNP) are co-dominant markers, and therefore more powerful for genetic analyses. Compared with RFLP, the advantage of SSR and SNP markers is that they can be revealed using high-throughput technologies with scant amounts of DNA. Semagn *et al.*[7] made a detailed comparison of the characteristics of each kind of marker. A high-density cocoa linkage map enriched with SSR genomic markers, including only co-dominant markers, was developed by Pugh *et al.*[8] More recently, that map was supplemented with 114 EST-SSRs.[9]

In recent years, the use of SNP markers has substantially increased in plant genetics such as in *Arabidopsis*,[10] grapevine,[11] wheat,[12] and also a few woody perennial species.[13–15] SNP is one of the most abundant types of DNA sequence polymorphism and the SNP markers are suitable for large-scale genome analysis using high-throughput automated genotyping techniques. SNPs have been used to construct high-resolution genetic maps[16,17] or to trace evolution, particularly in the human genome, using large-scale SNP datasets.[18,19] Knowledge of nucleotide substitution dynamics is an important basis for molecular evolutionary studies, phylogeny reconstruction and natural selection studies.[20,21] Transitions are generally observed with higher frequencies than transversions. During natural selection, transitions are better tolerated because they generate more likely synonymous mutations in protein-coding sequences than transversions.[22–25]

Of existing SNP markers, EST-SNPs (i.e. SNPs located within a gene expressed sequence) are of particular interest for studying functional genetic diversity and identifying candidate genes as the functional base of quantitative trait loci (QTLs). EST-SNPs have been developed for numerous plant models such as melon,[26,27] *Brassica rapa*,[28] barley,[29] poplar,[14] and sugarcane[30] to detect QTLs for many traits and facilitate the selection of resistant and productive plants. In *T. cacao*, a few SNPs were detected in ESTs from expression libraries representing *T. cacao*/*M. perniciosa* interactions.[31]

In our study, we discovered and mapped several hundred EST-SNP markers detected in an exhaustive collection of cocoa ESTs[32] homologous to genes with a known function. These SNP markers were supplemented by 163 new SSR markers to construct a very high-density genetic map suitable for large-scale genetic studies.

2. Materials and Methods

2.1. Plant material

SNP polymorphisms were screened in a collection of diverse germplasm representing the major part of the *T. cacao* diversity and two existing mapping populations denominated UPA402 × UF676 and F2.

The collection of diverse germplasm consisted of 249 genotypes from various genetic groups and geographical origins (Table 1). Most of these accessions are maintained at the International Cocoa Genebank (ICG) at the Cocoa Research Unit (CRU), University of the West Indies, Trinidad and Tobago.

The UPA402 × UF676 mapping population consisted of 264 individuals derived from a cross of two unrelated heterozygous tree accessions; UPA402, an Upper Amazon Forastero from Peru, and UF676, a Trinitario (Forastero × Criollo hybrid) selected in Costa Rica. This progeny was maintained by Centre National de Recherche Agronomique (CNRA) in Bingerville and Divo, Côte d'Ivoire. It was used to

Table 1. *Theobroma cacao* genotypes of various geographical origins used to screen the polymorphism of the 1536 GoldenGate SNP panel

Accession collection group name	Number of genotypes	Geographical origin
AMAZ	2	Ecuador
APA	1	Colombia
Nacional	3	Ecuador
Criollo	14	Mexico-Belize
EBC	4	Colombia
Trinitario	28	Trinidad
GU	12	French Guiana
IMC	19	Peru
LCTEEN	46	Ecuador
MORONA	3	Peru
NANAY	49	Peru
PARINARI	40	Peru
POUND	6	Peru
SC	5	Colombia
SCAVINA	8	Peru
Amelonado type	3	Brazil
SPEC	6	Colombia
Total	249	

establish genetic map as a reference in our laboratory.[4,6,8,9] We mapped the new SSR and SNP markers in this population.

The F2 second progeny of 132 individuals was obtained by selfing a hybrid between two heterozygous parents: Scavina 6, an Upper Amazon Forastero collected in Peru, and ICS1, a Trinitario selected in Trinidad. This progeny was produced by Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC) at Itabuna, Brazil.

2.1.1. *Genotypes used for EST-SNP detection and selection*

Most of the ESTs screened for SNPs had been obtained from the contrasting genotypes Scavina 6, an upper Amazon Forastero genotype from Peru and ICS1, a Trinitario selected in Trinidad, a hybrid between a Criollo from Central America and a Forastero from Lower Amazonia of Brazil.

These two genotypes, which represent the three distinct genetic origins, Upper Amazon Forastero, Lower Amazon Forastero, and Criollo, were also the parents of the F2 population from Brazil used to map SNPs.

Eleven other genotypes were involved in the construction of the cDNA libraries and SNP identification: B97-CC2, a Criollo from Belize, P7, IMC47, UPA 134, Upper Amazon Forastero genotypes from Peru, Jaca, an Upper Amazon Forastero from Brazil, GU255V, collected in French Guiana, B240 and 33−49, two Nacional genotypes from Ecuador, UF676, UF273, two Trinitario, and seedlings from a hybrid selected in Papua New Guinea.

SSRs were screened in three *T. cacao* genotypes: the two parents of the reference map (UPA402 and UF676), and the sequenced Criollo genotype (B97-61/B2).

2.2. *DNA extraction and purification*

Genomic DNA was extracted according to a protocol using MATAB buffer already described for the isolation of genomic DNA.[6] DNA was resuspended with 1 ml of TE (10 mM Tris−HCl and 1 mM EDTA, pH 8.0).

DNA was purified with the Nucleobond® PC 20 kit (Macherey-Nagel, Cat. No. 740.571.100) with the modification that steps 1 and 2 were omitted and the DNA was purified directly after its isolation. A 1 ml mixture composed of 200 μl of crude DNA (20 μg DNA maximum), 450 μl water and 350 μl S3 buffer + RNAse (buffers provided with the kit) was passed through the column (step 3). This solution was homogenized on a rocking table for at least 1 h. After precipitation of the eluate with an equal volume of isopropyl alcohol, the pellet was resuspended in 70 μl of TE.

The quality and quantity of DNA were first checked on 0.8% agarose gel, compared with a standard range, and then the Quant-iT™ PicoGreen® dsDNA Assay from Invitrogen™ was used. A quality test was performed for each sample by amplifying microsatellite markers in a PCR mixture with a high DNA concentration (100 ng DNA in a 10 μl reaction volume). The purification step was repeated when the amplification failed.

2.3. *In silico SNP discovery and verification*

A collection of 149 650 ESTs (EMBL accession number CU469588 to CU633156), corresponding to 48 594 unigenes, was produced after sequencing 56 cDNA libraries constructed from material collected from different organs, genotypes, and under different environmental conditions.[32]

SNPs were detected *in silico* and quality checked using the QualitySNP pipeline,[33] as reported in Argout *et al.*[32] and in ESTtik (http://esttik.cirad.fr).

QualitySNP uses quality information related to each EST and a haplotype-based strategy to predict reliable SNPs. In order to detect SNPs in known homologous coding sequences, we selected contigs displaying a significant similarity with proteins from a non-redundant protein sequence database (NR), with entries from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq, and as described in Argout *et al.*[32]

2.4. *Validation of SNPs via golden gate assay*

A total of 30−50 ng of genomic DNA per plant was used for Illumina SNP genotyping with the Illumina BeadArray platform at the French National Genotyping Centre (CNG, CEA-IG, Evry, France), according to the GoldenGate Assay manufacturer's protocol. Three 3-day assays were carried out to genotype the progeny samples for the 1536 SNP set revealed by the GoldenGate assay (Supplementary Table S1). The protocol was similar to that briefly described by Hyten *et al.*[34] except for the number of oligonucleotides involved in a single DNA reaction, thus comprising 4608 custom oligos assembled in the oligo pooled assays (OPA) designed by Illumina Inc. Raw hybridization, intensity data processing, clustering, and genotype calling were performed using the genotyping module in the BeadStudio/GenomeStudio package (Illumina, San Diego, CA, USA). Illumina has developed a self-normalization algorithm that relies on information contained in each array, as described by Akhunov *et al.*[35]

The clustering and genotype calling of each of the 1536 SNP markers were checked for their conformity and correct genotype distribution using known homozygous and heterozygous genotypes, included in the collection of diverse genotypes, as standards.

2.5. SSR in silico discovery and genotyping

The MIcroSAtellite identification tool (MISA http://pgrc.ipk-gatersleben.de/misa) was used to perform SSR searches, and primers were designed with Primer3 software.[36]

Primers flanking microsatellite loci were designed at each end of the scaffolds to orient and anchor them to the genetic map.

SSRs identified in the scaffolds were mapped only on the reference map established from the UPA402 × UF676 cross.

A total of 409 primer pairs (Supplementary Table S2) were defined in the 100 larger non-anchored scaffolds using Primer3 software[36] and screened for their ability to segregate in the UPA402 × UF676 progeny.

For a given SSR locus, the forward primer was designed with a 5′-end M13 tail (5′-CACGACG TTGTAAAACGAC-3′). PCR amplifications were performed in a Mastercycler ep384 thermocycler (Eppendorf, Germany) with 5 ng of purified DNA in a 10 μl final volume of buffer containing 10 mM Tris–HCl (pH 8), 50 mM KCl, 0.001% glycerol, 2.0 mM $MgCl_2$, 0.08 μM of the M13-tailed forward primer, 0.1 μM of the reverse primer, 200 μM of dNTP, 1 U of Taq DNA polymerase (Life Technologies, USA), 0.1 μM of M13 primer-fluorescent dye 6-FAM™, NED®, VIC®, or PET® (Applied Biosystems, CA, USA). The DNA and buffer were distributed in 384 plates using a Biomek NX automatic pipetting robot (Beckman Coulter, CA, USA). The touchdown PCR programme used was as follows: initial denaturation at 95°C for 5 min, followed by 10 cycles at 95°C for 45 s, Tm of 56–46°C (−1°C/cycle) for 1 min, and 72°C for 1 min 30 s. After these cycles, an additional round of 25 cycles were performed at 95°C for 45 s, Tm of 50°C for 1 min, and 72°C for 1 min, with a final elongation step at 72°C for 30 min.

PCR products were diluted specifically for each dye and pooled for multiplex SSR genotyping (revealing two SSRs having different sizes of amplified product per dye). A mixture of 15 μl of Hi-Di™ formamide (Applied Biosystems) and 0.12 μl of size marker GeneScan™ 600-LIZ-Size® Standard V2.0. (Applied Biosystems) was added to 2 μl of the diluted PCR pool. This pool was then analysed using the ABI 3500xL automatic sequencer (Applied Biosystems).

Images were analysed using Genemapper 4.0 software (Applied Biosystems) and exported as a data table.

2.6. Genetic mapping

The UPA402 × UF676 population was the result of a cross between two heterozygous cocoa clones, UPA402 (♀) an Upper-Amazon Forastero and UF676 (♂) a Trinitario. In this case, there were three segregation possibilities: loci that were homozygous for one parent and heterozygous for the other, segregation (1:1), and those that segregated in both parents (1:2:1 or 1:1:1:1).

Segregations were checked for goodness-of-fit to the expected Mendelian ratio using a chi-square test at significance levels of 0.05 and 0.01.

Individual and consensus maps were constructed using Joinmap software, version 4.0.[37]

Joinmap is able to combine data of several segregation types to construct a consensus genetic map. Here we used population type CP for the UPA402 × UF676 map, and population type F2 for the F2 population from Brazil. A lod score of 6 was used to identify 10 linkage groups (LGs) independently for each map. A consensus genetic map was established from the two distinct genetic maps. The corresponding groups were associated in pairs with JoinMap software. The Kosambi mapping function, with a lod score of 5 and a jump threshold of 3, was used to convert recombination frequencies into map distances.[38]

This consensus map combined the new EST-SNPs and genomic SSRs defined from the scaffolds, in addition to the previously mapped markers.[9] This map contained only markers with a known nucleotide sequence.

3. Results

3.1. Identification of SNPs and development of the golden gate assay

The assembly made from the 149 650 T. *cacao* EST sequences (see Materials and Methods) generated 12692 T. *cacao* contigs. The number of ESTs per contig ranged from 2 to 5102. To detect good quality *in silico* SNPs, we assumed that contigs with more than 100 members contained paralogous sequences.[13,39] We therefore first selected 4818 contigs that contained at least 4 but no more than 100 EST members. A total of 5246 SNPs were identified *in silico* in 2012 contigs.

We selected 4150 *in silico* SNPs detected in 1834 contigs that had a significant BlastX annotation similarity with known proteins of the NCBI non-redundant protein sequence database (NR) with entries from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq, and as described in Argout *et al.*[32]

3.2. SNP performance and quality

The set of 4150 *in silico* SNPs was selected in the EST contigs and the SNP-harbouring sequences were then submitted to Illumina for processing using the Illumina® Assay Design Tool (ADT). ADT generates scores for each SNP that can range from 0 to 1;

SNPs with scores >0.6 have a high probability of being converted into a successful genotyping assay. In the set of 4150 submitted SNPs, 83.5% showed a high conversion success rate (>0.6), 9.2% showed a moderate conversion success rate (between 0.4 and 0.6), and 7.3% showed either a low conversion success rate or no score. A total of 1536 SNP sites having ADT scores >0.4 and without any other SNPs within the adjacent 60 bp was selected for the OPA design (Supplementary Table S1).

3.3. Analysis of base changes

One thousand and forty-four *in silico* SNPs (68%) were transitions and 462 (32%) were transversions (Table 2). This ratio of transition/transversion SNPs tallies with the results observed in other plant species, where transition SNPs are always more frequent than transversion SNPs.

3.4. SNP polymorphism

From the 1536 SNPs, 841 (55%) with a non-ambiguous polymorphic pattern across accessions were retained as true and verified SNPs and denominated TcSNP. Of the rest, 113 (7%) failed to be genotyped, 436 (28%) had a monomorphic pattern, and 146 (10%) were polymorphic but did not show any clear fluorescent pattern suitable for reliable genotype classification.

Of the 841 polymorphic SNPs, 461 segregated in the mapping population (UPA402 × UF676) and could be mapped on the reference map. Five hundred and thirty-one were polymorphic and mapped on the F2 population map. Two hundred and thirty-nine SNP markers were segregating in both maps, thus enabling construction of a consensus map between them.

3.5. SSR polymorphism

A high-density genetic map is a key tool to order the scaffold assembly needed to generate a complete cocoa genome sequence. SSR markers were defined in the largest non-anchored scaffolds in order to improve anchoring of the *T. cacao* genome assembly provided by the International Cocoa Genome Sequencing consortium[1] on the genetic map.

From the 409 screened SSRs (Supplementary Table S2), 163 were polymorphic for the UPA402 × UF676 progeny and could be mapped.

The new SSR markers defined from scaffolds were named mTcCIR450 to mTcCIR613 to extend the previously identified SSR marker series; mTcCIR 1 to mTcCIR 291[8] from genomic DNA and mTcCIR 292 to mTcCIR 447[9] from ESTs.

Table 2. Nucleotide substitution types of the 1536 selected *in silico* SNPs

Types	Number of SNPs	Percentage		Percentage
A ↔ C	126	8	Transversion	32
A ↔ T	128	8		
C ↔ G	112	7		
T ↔ C	126	8		
T ↔ G	612	40	Transition	68
A ↔ G	432	29		

3.6. Individual genetic linkage maps

3.6.1. Map of the UPA402 × UF676 population

A new set of 624 markers with their corresponding sequences, including 461 EST-SNP and 163 new SSR markers located on scaffolds of the genome assembly,[1] were added to the reference map (Fig. 1, Supplementary Table S3).

The new UPA402 × UF676 map contained 1043 markers, including 461 EST-SNPs, 524 SSRs and 58 RFLPs (Table 3). Of the 1043 markers, 571 corresponded to gene markers. The length of this map was 751.7 cM having an average distance of 0.7 cM between adjacent markers.

Skewed segregation was observed for 118 markers (11.3%). The skewed markers were mainly located in LGs 2, 3, 6 and 10, as is shown in Fig. 3.

3.6.2. Map of the F2 population

The F2 map (Fig. 2, Supplementary Table S4) contained 531 EST-SNP markers. This map had a total length of 753.9 cM, with an average distance of 1.4 cM between neighboring markers. The marker density varied, with an average distance between neighboring markers ranging from 0.9 cM in LG 9 to 2.7 cM in LG7 (Table 4).

Skewed segregation was observed in 97 markers (18.3%). The skewed markers were mainly located in LGs 1, 3 and 4, as is shown in Fig. 3.

3.7. Consensus genetic linkage map

Two hundred and thirty-nine SNP markers were mapped in both populations.

The complete consensus map (Table 5) contained 1262 codominant markers including 681 EST-SNPs, 523 SSRs (163 scaffold-tagged-SSRs, 110 EST-SSRs, 250 SSRs from genomic DNA) and 58 RFLPs including 14 resistance gene analogues (Rgenes-RFLPs), arranged in 10 LGs corresponding to the haploid chromosome number of *T. cacao* (Fig. 3, Supplementary Table S5). Among the 1262 markers, 65% were gene-based markers, including SNPs, SSRs and RFLPs.

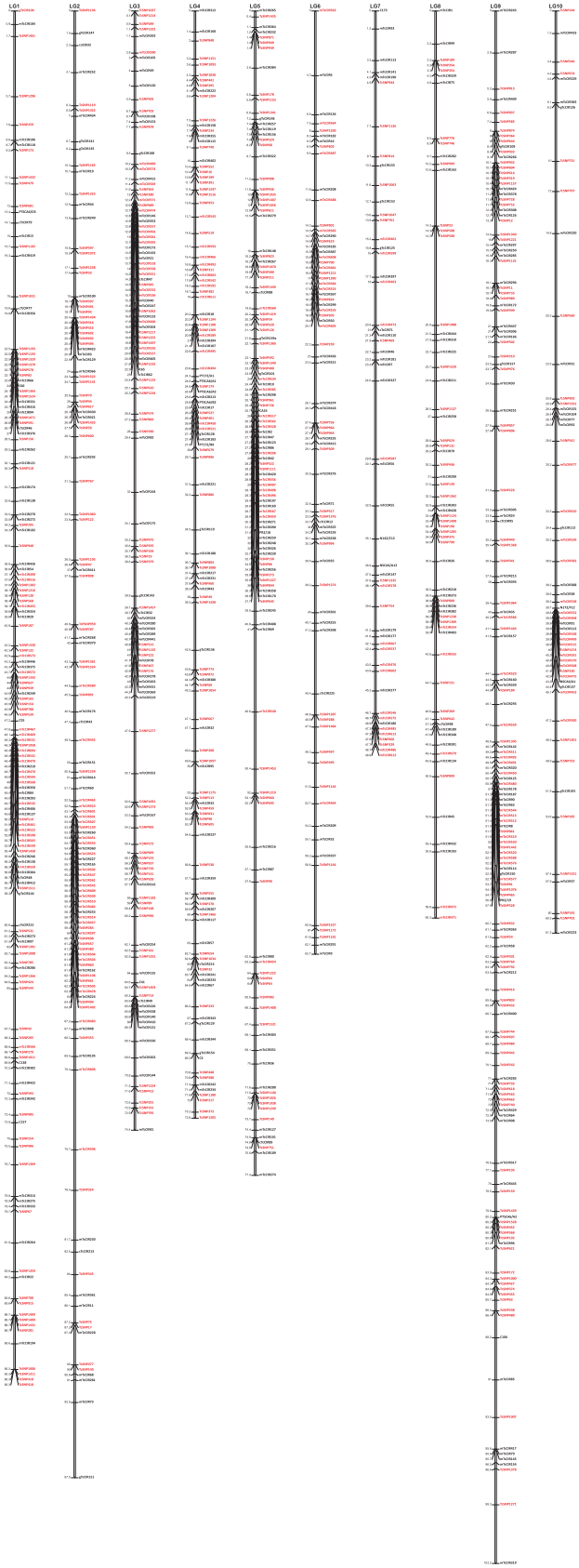

Figure 1. Genetic map constructed from an F1 progeny of 264 individuals (located in CNRA, Côte d'Ivoire) belonging to the UPA402 × UF676 cross. This map consists of 1043 markers of a known DNA sequence (461 SNPs, 524 SSRs, and 58 RFLPs), spanning 752 cM. The average distance between two markers is 0.7 cM. The new markers added to this map are printed in red.

Table 3. Distribution of each marker type in the LGs of the reference map (UPA402 × UF676)

LG	Length (cM)	Total number of markers	Average distance between markers (cM)	SNP	RFLP	Genomic SSR	SSR from scaffold	EST-SSR
LG1	90.3	150	0.6	69	8	33	22	18
LG2	97.5	126	0.8	55	6	26	28	11
LG3	74.4	126	0.6	57	6	33	18	14
LG4	75.6	120	0.6	61	11	27	14	8
LG5	77.4	121	0.6	55	6	34	15	11
LG6	62.7	73	0.9	28	2	19	12	12
LG7	48.9	51	1.0	11	6	16	16	2
LG8	60.3	64	0.9	30	1	17	5	11
LG9	103.2	154	0.7	78	6	34	17	20
LG10	61.3	54	1.1	17	6	12	16	3
Total	751.7	1043	0.7	461	58	251	163	110

SNP, single-nucleotide polymorphism; RFLP, restriction fragment polymorphism; SSR, simple sequence repeat; EST, expressed sequence tag.

The total map length was 733.6 cM, i.e. slightly shorter than previously constructed maps (782.8 cM for Pugh *et al.*[8] and 779.2 cM for Fouet *et al.*[9]). The average distance between adjacent markers on this map was 0.6 cM, and thus shorter than the 1.3 cM of the map of Fouet *et al.*[9]

The number of mapped loci varied substantially between LGs on the consensus map; from 63 in LG10 to 201 in LG1. The average distance between two markers in the different LGs ranged from 0.4 cM in LG1 to 0.9 cM in LG10.

In total, 844 new markers (681 SNP markers and 163 SSR defined in scaffolds) were mapped. These new markers were well distributed over all chromosomes allowing to fill some gaps in the previous maps, for example on chromosome 10.

4. Discussion

A large set of EST-SNP markers was generated and mapped in *T. cacao*. New SSR markers were added to these SNPs, providing an efficient tool for high-throughput genotyping of cocoa populations.

SSR markers are multiallelic and well adapted for fine analysis of population diversity structure.[40–43] In *T. cacao*, an average number of 5.8 alleles per SSR was observed by Loor Solorzano[44] after genotyping a collection of genetic resources of various genetic origins, and with a maximum of 15 alleles revealed by one SSR (mTcCIR322). This is not the case for SNPs that are only biallelic, but a higher number of SNP markers (several thousands) can be easily revealed at once using high-throughput technologies.

We used our new SNP and SSR markers to construct a very high-density genetic map. Sixty-five per cent of the markers were from within genes and the average distance between adjacent markers was 0.6 cM.

Several chromosome regions include markers with skewed segregations, particularly on LG 1, LG 3, LG 4, and LG 6. The region on LG 4 includes the locus for self-incompatibility previously identified by Crouzillat *et al.*[5] The gameto-sporophytic incompatibility system existing in *T. cacao*[45,46] could possibly explain the segregation distortion on this LG 4 region. Other factors which could explain segregation distortion, such as chromosome rearrangements in banana[47] which are responsible for highly skewed marker segregations, have not been reported in *T. cacao*.

This high-density genetic map can be used as a major tool for efficient genome-wide association studies (GWASs) in *T. cacao* populations. This method, first applied in human and animal genetics,[48–51] was also found to be highly effective for studying the determinism of useful traits in plants,[52–56] particularly in cocoa with the analysis of some recent hybrid populations.[8,57,58] GWAS is an alternative to QTL analyses in cross progenies for the purpose of studying genetic control of phenotypic traits in cocoa.

GWASs can be carried out on unrelated genetic resources such as wild or cultivated populations or germplasm collections. Large cocoa germplasm collections are maintained in many countries and characterized for useful traits. Two international cocoa collections are hosted at the International Cocoa Genebank, Trinidad (ICG,T, preserving 2300 accessions),[59] and at the Centro Agronomico Tropical de Investigacion y Ensenanza, Turrialba, Costa Rica (CATIE, preserving 1150 accessions).[60] The markers identified here will now certainly facilitate such

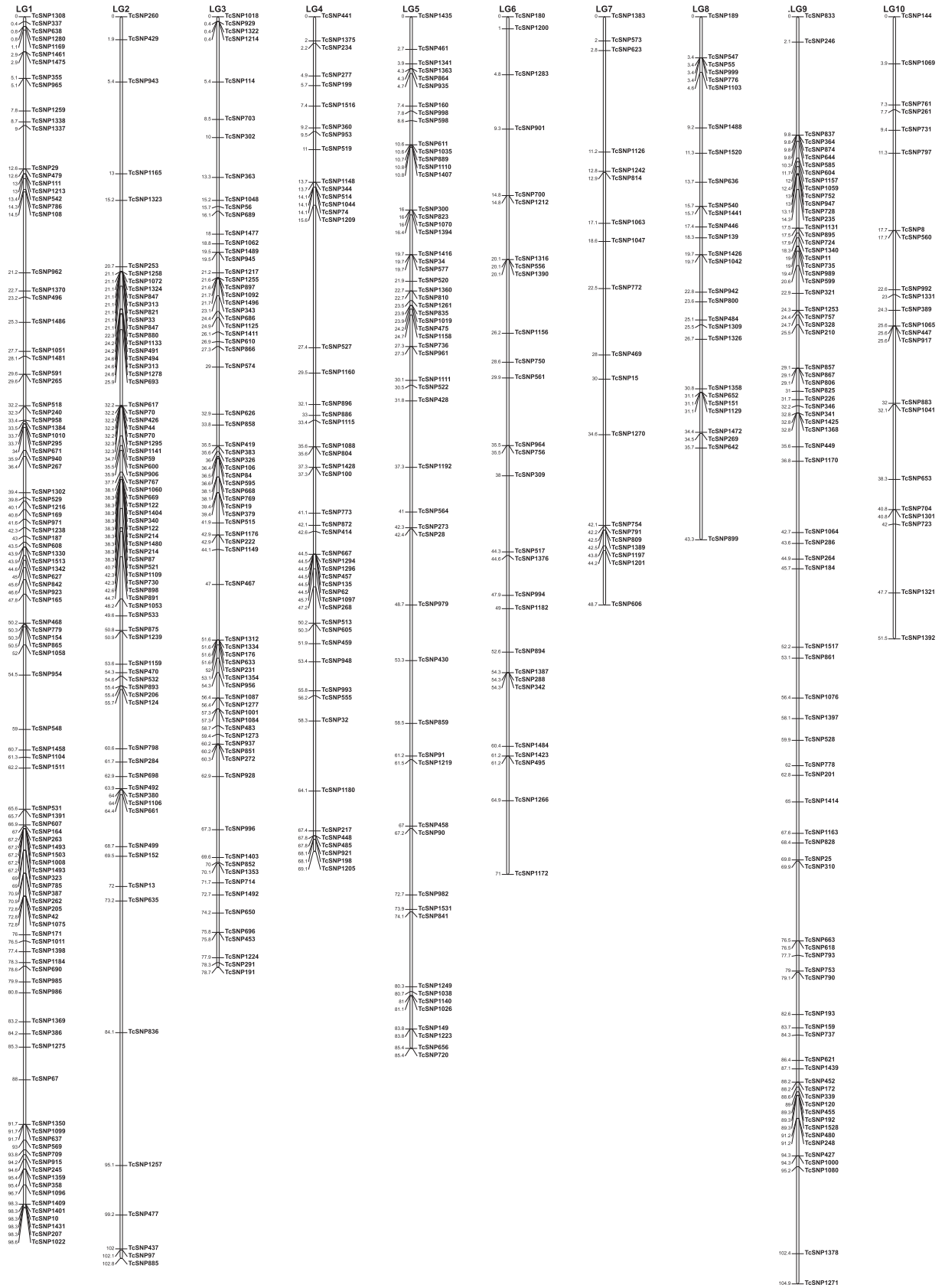

Figure 2. Genetic map constructed from an F2 progeny of 132 individuals (located at CEPLAC, Brazil) obtained by selfing of a single (Scavina 6 × ICS1) selected hybrid. This map consists of 531 SNP markers, spanning 754 cM. The average distance between two markers is 1.4 cM.

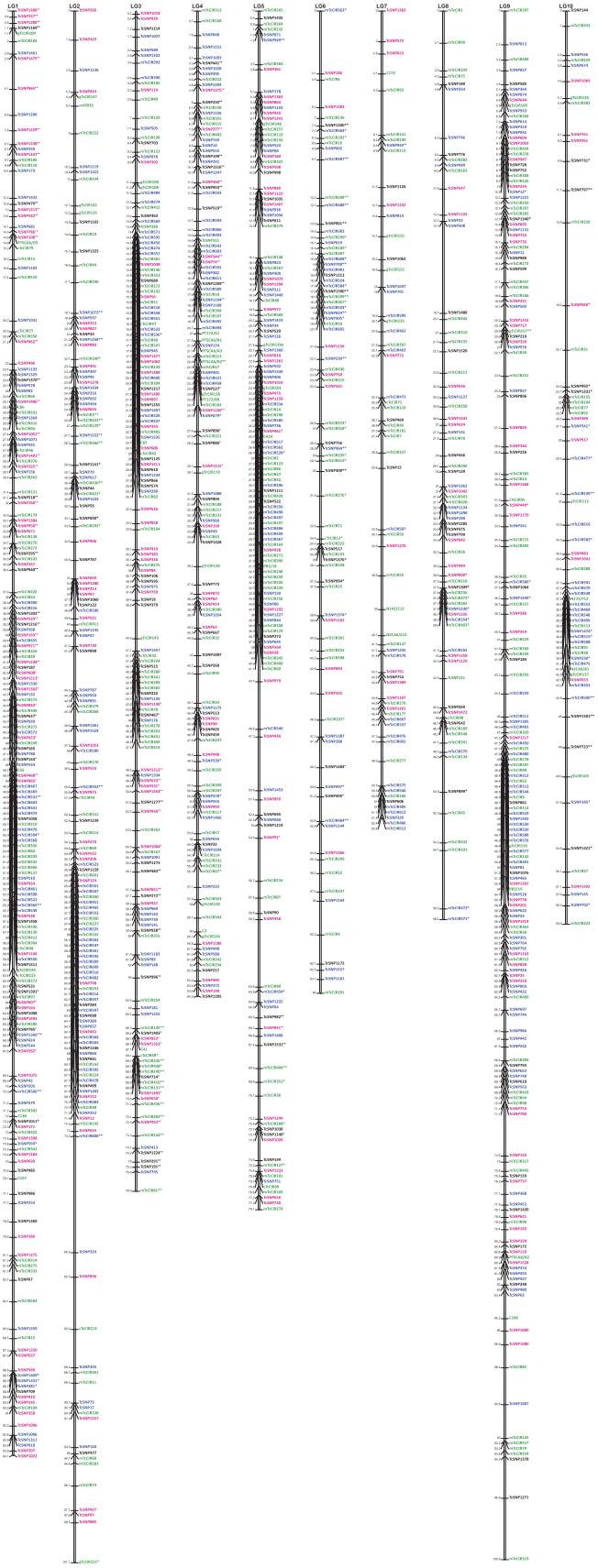

Figure 3. Consensus map of (UPA402 × UF676) and F2 progenies. Markers segregating in both progenies are indicated in black, those segregating only in (UPA402 × UF676) are printed in green (previously mapped markers) and blue (newly mapped markers). Markers segregating only in the F2 progeny are printed in pink. This consensus map consists of 1262 markers of a known DNA sequence, and it has a length of 734 cM. The average distance between two markers is 0.6 cM. Among the 1262 markers, 810 correspond to markers defined in expressed genes. Significant skewed segregations are indicated by asterisks (*$P < 0.05$, **$P < 0.01$) or dots (F2 population).

GWASs, providing added-value to this wide characterization work, thus boosting knowledge on the genetic determinants of useful cocoa traits.

Another benefit of this large set of mapped markers is the possible integration of molecular information in conventional cocoa breeding schemes using marker-assisted selection (MAS).

In cocoa, few MAS experiments are currently underway.[3,61] The efficiency of MAS in selecting *P. palmivora*-resistant cocoa plants has been reported by Lanaud *et al.*[3]

Until now, MAS studies have mainly been focused on traits controlled by a small number of genes, using only markers close to QTLs. However, such methods are of limited use for traits that are determined by a large number of genes of small effects.

Table 4. Distribution of SNP markers in the LGs of the F2 map (SCA6 × ICS1) selfing

LG	Length (cM)	Number of SNP markers	Average distance between markers (cM)
LG1	98.6	104	0.9
LG2	102.8	74	1.4
LG3	78.7	73	1.1
LG4	69.1	49	1.4
LG5	85.4	56	1.5
LG6	71.0	28	2.5
LG7	48.7	19	2.6
LG8	43.3	28	1.5
LG9	104.9	78	1.3
LG10	51.5	22	2.3
Total	753.9	531	1.4

SNP, single-nucleotide.

Substantial genome-wide molecular data can now be generated at lower cost by high-throughput technologies, such as SNP genotyping. This progress has paved the way for the development of new methods to predict genotype value via MAS. The genome-wide selection or genomic selection (GS) method was recently successfully applied in animal or plant breeding[62–66] and allows to predict phenotypes using all marker information.

The integration of molecular markers in cocoa recurrent breeding programmes[67–72] could be facilitated by the GS approach in order to accelerate genetic gains. The GS strategy seems particularly suitable for the selection of multigenic traits such as yield and disease resistance. Cumulating a large number of resistance alleles is one of the main objectives of cocoa breeding for sustainable cocoa resistance. The large set of available SNP markers could facilitate the selection of resistant and high yielding cocoa trees via GS approaches enabling the use of all genome regions tagged by SNP markers, even those with very small effects.

The search for candidate genes underlying trait variation is another major challenge for plant biologists, with the aim of gaining further insight into the mechanisms underlying trait variation, and producing tools to efficiently screen and exploit genetic resources.

The consensus map produced in this work has been used efficiently for anchoring an assembly of *T. cacao* Criollo genome sequences, and for constituting pseudomolecules.[1] Recently, two different cocoa varieties, i.e. Criollo[1] and Forastero from the Lower Amazon region (http://www.cacaogenomedb.org/), were sequenced, with 28 798 and 35 000 annotated genes, respectively. These sequences will greatly facilitate the identification of candidate genes, allowing

Table 5. Distribution of each marker type in the LGs of the consensus genetic map

LG	Length (cM)	Total number of markers	Average distance between markers (cM)	SNP	RFLP	Genomic SSR	SSR from scaffold	EST-SSR
LG1	77.1	201	0.4	120	8	33	22	18
LG2	101.1	156	0.6	85	6	26	28	11
LG3	76.9	162	0.5	91	6	33	18	14
LG4	64.2	135	0.5	75	11	27	14	8
LG5	78.1	147	0.5	81	6	34	15	11
LG6	64	81	0.8	36	2	19	12	12
LG7	52.6	62	0.8	22	6	16	16	2
LG8	59.2	73	0.8	39	1	17	5	11
LG9	100.9	182	0.6	106	6	33	17	20
LG10	59.5	63	0.9	26	6	12	16	3
Total	733.6	1262	0.6	681	58	250	163	110

SNP, single nucleotide polymorphism; RFLP, restriction fragment polymorphism; SSR, simple sequence repeat; EST, expressed sequence tag.

integration of both genetic and genomic (functional and structural) data. Overall, about 300 QTLs or marker/trait associations have already been identified in *T. cacao*. High-throughput genotyping associated with a high marker density will facilitate fine mapping of genes involved in trait variation (with GWAS or classical QTL analyses conducted on large progenies), thus allowing to refine the QTL position in the genome, while facilitating the search for candidate genes in corresponding genome sequences. Several functional studies have already been conducted in cocoa, focused mainly on genes generally expressed in specific physiological conditions or metabolisms.[73] It will be now possible to focus more specifically on the expression of genes directly responsible for trait variation after candidate gene validation.

Analysing genome evolution during domestication processes or adaptation to climate change can also help us to identify key genes underlying adaptive traits.[74] Loss of diversity generally occurs during genome evolution, and some genes are selectively involved in natural selection or domestication. A large set of SNPs defined in expressed genes, such as those identified in this study, provides a key tool for identifying selection signatures or adaptive substitutions, and then highlighting candidate genes potentially involved in the adaptation[60] or domestication processes and their corresponding molecular functions.[75] All SNPs reported in this paper were identified in orthologous genes or gene families, thus facilitating comparative genomic approaches, and benefiting from gene knowledge accumulated in other species to accelerate cocoa breeding.

5. Availability

Information on the consensus linkage map, molecular markers, and primers are available in the Map Study 'SSR_SNP_consensus_map' of the cocoa module of TropGeneDB database (http://tropgenedb.cirad.fr).

Acknowledgements: We acknowledge Chantal Hamelin for TropGeneDB support. We thank Peter Biggins for improving the English in this paper, and the reviewers for their useful comments and editing job of this paper.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

We thank the Region Languedoc-Roussillon, Valrhona and the Fondation Agropolis for their financial support in this work.

References

1. Argout, X., Salse, J., Aury, J.M., et al. 2011, The genome of *Theobroma cacao*, *Nat. Genet.*, **43**, 101–8.
2. Cheesman, E.E. 1944, Notes on the nomenclature, classification and possible relationships of cocoa populations, *Trop. Agricult.*, **21**, 144–59.
3. Lanaud, C., Fouet, O., Clément, D., et al. 2009, A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L., *Mol. Breed.*, **24**, 361–74.
4. Lanaud, C., Risterucci, A.M., Ngoran, A.K.J., et al. 1995, A genetic-linkage map of *Theobroma cacao* L., *Theor. Appl. Genet.*, **91**, 987–93.
5. Crouzillat, D., Lerceteau, E., Petiard, V., et al. 1996, *Theobroma cacao* L.: a genetic linkage map and quantitative trait loci analysis, *Theor. Appl. Genet.*, **93**, 205–14.
6. Risterucci, A.M., Grivet, L., N'Goran, J.A.K., Pieretti, I., Flament, M.H. and Lanaud, C. 2000, A high-density linkage map of *Theobroma cacao L.*, *Theor. Appl. Genet.*, **101**, 948–55.
7. Semagn, K., Bjornstad, A.A. and Ndjiondjop, M.N. 2006, An overview of molecular marker methods for plants, Afr. J. Biotechnol., **5**, 2540–68.
8. Pugh, T., Fouet, O., Risterucci, A.M., et al. 2004, A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers, *Theor. Appl. Genet.*, **108**, 1151–61.
9. Fouet, O., Allegre, M., Argout, X., et al. 2011, Structural characterization and mapping of functional EST-SSR markers in *Theobroma cacao*, *Tree Genet. Genomes*, **99**, 1–19.
10. Schmid KJ, R.S.T. 2003, Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*, *Genome Res.*, **13**, 1250–7.
11. Lijavetzky, D., Cabezas, J.A., Ibanez, A., Rodriguez, V. and Martinez-Zapater, J.M. 2007, High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology, *BMC Genomics*, **8**, 424.
12. Berard, A., Le Paslier, M.C., Dardevet, M., et al. 2009, High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.), *Plant Biotechnol. J.*, **7**, 364–74.
13. Dantec, L.L., Chagne, D., Pot, D., et al. 2004, Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences, *Plant Mol. Biol.*, **54**, 461–70.
14. Zhang, B., Zhou, Y., Zhang, L., Zhuge, Q., Wang, M.X. and Huang, M.R. 2005, Identification and validation of single nucleotide polymorphisms in poplar using publicly expressed sequence tags, *J. Integr. Plant Biol.*, **47**, 1493–9.
15. Pavy, N., Parsons, L.S., Paule, C., MacKay, J. and Bousquet, J. 2006, Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs, *BMC Genomics*, **7**, 174.
16. Cho, R.J., Mindrinos, M., Richards, D.R., et al. 1999, Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*, *Nat. Genet.*, **23**, 203–7.

17. Hoskins, R.A., Phan, A.C., Naeemuddin, M., et al. 2001, Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*, *Genome Res.*, **11**, 1100–13.
18. Tian, C., Plenge, R.M., Ransom, M., et al. 2008, Analysis and application of European genetic substructure using 300 K SNP information, *PLoS Genet.*, **4**, e4.
19. Novembre, J., Johnson, T., Bryc, K., et al. 2008, Genes mirror geography within Europe, *Nature*, **456**, 98–101.
20. Yang, Z. and Yoder, A.D. 1999, Estimation of the transition/transversion rate bias and species sampling, *J. Mol. Evol.*, **48**, 274–83.
21. Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. 1990, Phylogeny reconstruction, *Mol. Syst.*, **411**, 501.
22. Brown, W.M., Prager, E.M., Wang, A. and Wilson, A.C. 1982, Mitochondrial DNA sequences of primates: tempo and mode of evolution, *J. Mol. Evol.*, **18**, 225–39.
23. Gojobori, T., Li, W.H. and Graur, D. 1982, Patterns of nucleotide substitution in pseudogenes and functional genes, *J. Mol. Evol.*, **18**, 360–9.
24. Wakeley, J. 1994, Substitution-rate variation among sites and the estimation of transition bias, *Mol. Biol. Evol.*, **11**, 436–42.
25. Wakeley, J. 1996, The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance, *Trends Ecol. Evol. (Amst.)*, **11**, 158–62.
26. Morales, M., Roig, E., Monforte, A.J., Arus, P. and Garcia-Mas, J. 2004, Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.), *Genome*, **47**, 352–60.
27. Deleu, W., Esteras, C., Roig, C., et al. 2009, A set of EST-SNPs for map saturation and cultivar identification in melon, *BMC Plant Biol.*, **9**.
28. Li, F., Kitashiba, H., Inaba, K. and Nishio, T. 2009, A *Brassica rapa* linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits, *DNA Res.*, **16**, 311–23.
29. Kota, R., Varshney, R., Prasad, M., Zhang, H., Stein, N. and Graner, A. 2008, EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome, *Funct. Integr. Genomics*, **8**, 223–33.
30. Cordeiro, G.M., Eliott, F., McIntyre, C.L., Casu, R.E. and Henry, R.J. 2006, Characterisation of single nucleotide polymorphisms in sugarcane ESTs, *Theor. Appl. Genet.*, **113**, 331–43.
31. Lima, L.S., Gramacho, K.P., Carels, N., et al. 2009, Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao, *Genet. Mol. Res.*, **8**, 799–808.
32. Argout, X., Fouet, O., Wincker, P., et al. 2008, Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions, *BMC Genomics*.
33. Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A. 2006, QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species, *BMC Bioinformatics*, **7**, 438.
34. Hyten, D.L., Song, Q., Choi, I.-Y., et al. 2008, High-throughput genotyping with the GoldenGate assay in the complex genome of soybean, *Theor. Appl. Genet.*, **116**, 945–52.
35. Akhunov, E., Nicolet, C. and Dvorak, J. 2009, Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay, *Theor. Appl. Genet.*, **119**, 507–17.
36. Rozen, S. and Skaletsky, H. 2000, Primer3 on the WWW for general users and for biologist programmers, *Methods Mol. Biol.*, **132**, 365–86.
37. Van Ooijen, J.W. 2006, *Joinmap, Software for the Calculation of Genetic Linkage Maps*. Wageningen: The Netherlands.
38. Kosambi, D.D. 1944, The estimation of map distance from recombination values, *Ann. Eugen.*, **12**, 172–5.
39. Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J. and Edwards, D. 2003, Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data, *Plant Physiol.*, **132**, 84–91.
40. Zhang, D., Mischke, S., Johnson, E.S., Phillips-Mora, W. and Meinhardt, L. 2008, Molecular characterization of an international cacao collection using microsatellite markers, *Tree Genet. Genomes*, **5**, 1–10.
41. Lachenaud, P. and Zhang, D. 2008, Genetic diversity and population structure in wild stands of cacao trees (*Theobroma cacao* L.), in French Guiana, *Ann. For. Sci.*, **65**, 7.
42. Zhang, D., Arevalo-Gardini, E., Mischke, S., Zúñiga-Cernades, L., Barreto-Chavez, A. and Del Aguila, J.A. 2006, Genetic diversity and structure of managed and semi-natural populations of cocoa (*Theobroma cacao*) in the Huallaga and Ucayali Valleys of Peru, *Ann. Bot.*, **98**, 647–55.
43. Motamayor, J.C., Lachenaud, P., da Silva E Mota, J.W., et al. 2008, Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.), *PLoS ONE*, **3**, e3311.
44. Loor Solorzano, R.G. 2007, Contribution à l'étude de la domestication de la variété Nacional d'Equateur: recherche de la variété native et de ses ancêtres sauvages. PhD Thesis, Montpellier SUPAGRO- ED SIBAGHE- spécialité: biologie intégrative des plantes.
45. Knight, R. and Rogers, H. 1955, Incompatibility in *Theobroma cacao*, *Heredity*, **9**, 69–77.
46. Cope, F.W. 1958, Incompatibility in *Theobroma cacao*, *Nature*, **181**, 279.
47. Fauré, S., Noyer, J.L., Horry, J.P., Bakry, F., Lanaud, C. and Gońzalez de León, D. 1993, A molecular marker-based linkage map of diploid bananas (*Musa acuminata*), *Theor. Appl. Genetics*, **87**, 517–26.
48. Blott, S., Kim, J.-J., Moisio, S., et al. 2003, Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition, *Genetics*, **163**, 253–66.
49. Cardon, L.R. and Bell, J.I. 2001, Association study designs for complex diseases, *Nat. Rev. Genet.*, **2**, 91–9.

50. Liu, P., Wang, Y., Vikis, H., et al. 2006, Candidate lung tumor susceptibility genes identified through whole-genome association analyses in inbred mice, *Nat. Genet.*, **38**, 888–95.
51. Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M.E. 2002, Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping, *Genetics*, **161**, 373–9.
52. Zhang, Y.-M., Mao, Y., Xie, C., Smith, H., Luo, L. and Xu, S. 2005, Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.), *Genetics*, **169**, 2267–75.
53. Aranzana, M.J., Kim, S., Zhao, K., et al. 2005, Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes, *PLoS Genet.*, **1**, e60.
54. Jannoo, N., Grivet, L., Dookun, A., D'hont, A. and Glaszmann, J.C. 1999, Linkage disequilibrium among modern sugarcane cultivars, *Theor. Appl. Genet.*, **99**, 1053–60.
55. Wei, X., Jackson, P.A., McIntyre, C.L., Aitken, K.S. and Croft, B. 2006, Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure, *Theor. Appl. Genet.*, **114**, 155–64.
56. Agrama, H.A., Eizenga, G.C. and Yan, W. 2007, Association mapping of yield and its components in rice cultivars, *Mol. Breed.*, **19**, 341–56.
57. Marcano, M., Pugh, T., Cros, E., et al. 2007, Adding value to cocoa (*Theobroma cacao* L.) germplasm information with domestication history and admixture mapping, *Theor. Appl. Genet.*, **114**, 877–84.
58. Marcano, M., Morales, S., Hoyer, M.T., et al. 2009, A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (*Theobroma cacao* L.) population, *Tree Genet. Genomes*, **5**, 329–37.
59. Iwaro, A.D., Bekele, F.L. and Butler, D.R. 2003, Evaluation and utilisation of cacao (*Theobroma cacao* L.) germplasm at the International Cocoa Genebank, Trinidad, *Euphytica*, **130**, 207–21.
60. IBPGR Working Group on Genetics Resources of Cocoa. 1981, *Genetic Resources of Cocoa*. Rome, Italy, p. 25.
61. Schnell, R.J., Kuhn, D.N., Brown, J.S., et al. 2007, Development of a marker assisted selection program for cacao, *Phytopathology*, **97**, 1664–9.
62. Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. 2001, Prediction of total genetic value using genome-wide dense marker maps, *Genetics*, **157**, 1819–29.
63. Heffner, E.L., Sorrells, M.E. and Jannink, J.L. 2009, Genomic selection for crop improvement, *Crop Sci.*, **49**, 1–12.
64. Bernardo, R. and Yu, J. 2007, Prospects for genomewide selection for quantitative traits in maize, *Crop Sci.*, **47**, 1082.
65. Wong, C.K. and Bernardo, R. 2008, Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations, *Theor. Appl. Genet.*, **116**, 815–24.
66. Jannink, J.L., Lorenz, A.J. and Iwata, H. 2010, Genomic selection in plant breeding: from theory to practice, *Brief. Funct. Genomic.*, **9**, 166.
67. Clement, D., Eskes, A.B., Sounigo, O. and N'Goran, J. 1993, Amélioration génétique du cacaoyer en Côte d'Ivoire-Présentation d'un nouveau schéma de selection, In: *Proceedings of the 11th International Cocoa Research Conference*, pp. 18–24.
68. Paulin, D. and Eskes, A.B. 1995, Le cacaoyer: stratégies de sélection, *Plant. Recherche Dével.*, **2**, 5–18.
69. Pires, J.L., Monteiro, W.R., Pinto, L.R.M., Figueira, A., Yamada, M.M. and Ahnert, D. 1996, A proposal for cocoa breeding, In: *Proceedings of the 12th International Cocoa Research Conference*, pp. 17–23.
70. Iwaro, A.D., Singh, V., Bharath, S.M., Perez, C., Ali, L. and Butler, D.R. 2010, Germplasm enhancement for resistance to black pod disease – final stages and achievements, In: *Proceedings of the 16th International Cocoa Research Conference*, pp. 493–499.
71. Tahi, M., Lachenaud, P., N'Goran, J., et al. 2010, Second cycle de selection recurrente du cacaoyer (*Theobroma cacao* L.) en Cote d'Ivoire: bilan a mi-parcours et propositions de sorties varietales, In: *Proceedings of the 16th International Cocoa Research Conference*, pp. 3–12.
72. Lopes, U.V., Monteiro, W.R., Pires, J.L., Clement, D., Yamada, M.M. and Gramacho, K.P. 2011, Cacao breeding in Bahia, Brazil—strategies and results, *Crop Breed. Appl. Biotechnol.*, **1**, 73–81.
73. Micheli, F., Guiltinan, M., Gramacho, K.P., et al. 2010, Functional genomics of Cacao, In: Jean-Claude, K. and Michel, D., (eds.), *Advances in Botanical Research*, vol. 55. Academic Press: Londres, pp. 119–77.
74. Doebley, J.F., Gaut, B.S. and Smith, B.D. 2006, The molecular genetics of crop domestication, *Cell*, **127**, 1309–21.
75. Nielsen, R. 2005, Molecular signatures of natural selection, *Annu. Rev. Genet.*, **39**, 197–218.