Tree Genetics & Genomes
DOI 10.1007/s11295-011-0424-0

ORIGINAL PAPER

Identification and mapping of conserved ortholog set (COS) II sequences of cacao and their conversion to SNP markers for marker-assisted selection in *Theobroma cacao* and comparative genomics studies

David N. Kuhn · Don Livingstone III · Dorrie Main · Ping Zheng · Chris Saski · F. Alex Feltus · Keithanne Mockaitis · Andrew D. Farmer · Gregory D. May · Raymond J. Schnell · Juan C. Motamayor

Received: 28 September 2010 / Revised: 12 July 2011 / Accepted: 4 August 2011

Abstract *Theobroma cacao* (cacao) is a tree cultivated in the tropics around the world for its seeds that are the source of both chocolate and cocoa butter. Genetic marker development for marker-assisted selection (MAS) is critical for the success of cacao breeding for disease resistance and yield. To develop conserved ortholog set II (COSII) single-nucleotide polymorphism (SNP) markers for MAS in cacao, we have used three strategies and three types of cacao genetic and sequence data to identify and map 98 cacao COSII genes. The resources available at the time these studies were first undertaken dictated the strategy utilized. For the first strategy, SNPs were identified using cacao expressed sequence tags homologous to COSII sequences. Strategy II utilized a leaf transcriptome of cacao genotype "Matina 1–6" and Strategy III the genomic sequence of a 3-Mb region of "Matina 1–6" linkage group 5 associated with an important quantitative trait locus (QTL) for resistance to black pod. We have identified SNP markers for 83 of the 98 mapped COSII genes, and 19 of these SNP markers co-locate with QTLs. These COSII SNP markers, the first identified for cacao, will be used for genotyping and off-typing in cacao breeding programs and employed for genetic mapping and syntenic studies to trace co-location of genes regulating traits of importance between cacao and other species.

Keywords Chocolate · Genetic mapping · Molecular markers · Quantitative trait loci (QTL) co-localization

Communicated by A. Dandekar

Electronic supplementary material The online version of this article (doi:10.1007/s11295-011-0424-0) contains supplementary material, which is available to authorized users.

D. N. Kuhn (✉) · D. Livingstone III · R. J. Schnell · J. C. Motamayor
USDA-ARS SHRS,
Miami, FL 33158, USA
e-mail: David.Kuhn@ars.usda.gov

D. Main · P. Zheng
Department of Horticulture and Landscape Architecture,
Washington State University,
Pullman, WA 99164-6414, USA

C. Saski
Clemson University Genomics Institute,
Clemson, SC 29634, USA

F. A. Feltus
Department of Genetics and Biochemistry, Clemson University,
Clemson, SC 29634-0318, USA

K. Mockaitis
Center for Genomics and Bioinformatics, Indiana University,
Bloomington, IN 47405, USA

A. D. Farmer · G. D. May
National Center for Genome Resources,
Santa Fe, NM 87505, USA

J. C. Motamayor
Mars, Inc.,
800 High St.,
Hackettstown, NJ 07840, USA

Published online: 09 September 2011

Introduction

Theobroma cacao L. (cacao), an understory tree native to the upper Amazon, is cultivated globally in the humid tropics; it is a major source of currency for small farmers, as well as the main cash crop of several West African countries (http://faostat.fao.org). Cacao is a diploid ($2n=20$) member of the Malvaceae family with a genome size of approximately 430 Mb (Argout et al. 2011). Its cauliflorous fruits (pods) contain the seeds (beans) that are later processed by the multi-billion-dollar chocolate industry.

Since 2000, the USDA Agricultural Research Service (ARS) Subtropical Horticulture Research Station (SHRS) and Mars, Inc. have been involved in an international breeding program to improve cacao (Schnell et al. 2007). A key component of our breeding program is the utilization of marker-assisted selection (MAS) to identify disease-resistant trees from among the 30,000 genotypes currently under evaluation. An important aspect of our MAS program is to develop molecular marker assays that can be used in cocoa-producing countries to accelerate the process of selecting improved cacao cultivars with enhanced disease resistance and better yield. With the exception of Brazil and Malaysia, most cacao breeding is carried out in countries where it has been technically difficult to use microsatellite markers, requiring the majority of the genotyping to be done in the USA, France, or Great Britain. For the last 10 years, we, and others, have been developing and applying molecular markers for use in breeding improved cacao (*T. cacao*) varieties and curating the world's germplasm resources of cacao (Kuhn et al. 2003, 2006, 2008; Borrone et al. 2004, 2007; Schnell et al. 2007; Motilal and Butler 2003; Zhang et al. 2009; Motilal et al. 2010; Irish et al. 2010; Risterucci et al. 2000; Pugh et al. 2004; Lima et al. 2009). Currently, a single-nucleotide polymorphism (SNP) assay we developed is being used in breeding efforts in Ghana and is also being established in Cote d'Ivoire (Livingstone et al. 2011), the two top cocoa-producing countries in the world. SNPs are abundant and evenly distributed in genomes (Rafalski 2002) and therefore often make more effective markers than other, lower density markers such as microsatellites; utilizing SNP assays in the countries where cacao breeding occurs would make MAS for improved cacao varieties more efficient.

We were also interested in creating a set of genotyping markers for single-copy genes in cacao. Most valuable would be markers that had been mapped or could be mapped in other crop plant species in the Malvaceae, like cotton, for which extensive genetic data are available (Rong et al. 2007). Syntenic blocks (three or more conserved anchor points) of single-copy genes between the cacao and cotton genomes near a trait of interest in cotton may also be near the same trait in cacao, which would allow us to identify candidate markers in cacao from the cotton data. Such single-copy markers could also anchor a genetic recombination map of *Theobroma grandiflorum*, a close relative of cacao with interesting agronomic traits and a species that we are already using for comparative genomic studies (Kuhn et al. 2010).

Conserved ortholog set (COS) sequences seemed ideally suited for all of these purposes. COS sequences were first described by Tanksley and co-workers (Fulton et al. 2002) and were initially used in comparative genomics and phylogenetic studies. In a further refinement, Wu et al. (2006) identified 2,869 single-copy orthologs, referred to as COSII sequences, that are found in most plant species and are particularly useful for evolutionary and systematic studies. COS sequences have been studied in tree species (Krutovsky et al. 2007; Cabrera et al. 2009) and used as universal markers in the Asteraceae (Chapman et al. 2007) and for synteny mapping in tomato, eggplant (Wu et al. 2009b), pepper (Wu et al. 2009a), coffee (Lefebvre-Pautigny et al. 2010), and the Rosaceae (Cabrera et al. 2009). We therefore set out to identify and map COSII markers in cacao.

We report here the development of 83 SNP markers that are based on the coding regions of COSII sequences and the map positions of the loci they mark in the cacao genome. We describe the three different strategies used to identify them. These SNP markers have relatively high minor allele frequencies (MAF), and many are located close to important quantitative trait locus (QTL) in cacao (Brown et al. 2005; Clement et al. 2003a, b; Risterucci et al. 2003) and therefore may be useful for MAS once the favorable alleles have been identified from the populations used for the QTL mapping. They can be used for genotyping and off-typing in cacao and, because the COSII sequences they are based on are conserved among many species, they may play an important role in comparative genomics studies with other species in which they have been mapped.

Materials and methods

Plant materials

A diversity panel of 15 cacao genotypes representing the ten distinct STRUCTURE groups described by Motamayor et al. (2008) consisted of the following: "SCA6" and "U48" (Contamana group), "GU255/P" (Guiana group), "IMC51" and "AMAZ15-15" (Iquitos group), "PA120" and "PA150" (Marañon group), "Matina 1–6" (Amelonado group), "CRIOLLO13" (Criollo group), "COC3335" and "NAP30" (Curaray group), "U26" (Nacional group), "Pound5/C" and "Pound 7" (Nanay group), and "EBC-148" (Purus group). All plants were from the USDA-ARS SHRS germplasm collection in Miami except "EBC-148" (leaves of which were kindly provided by Dr. Wilbert Phillips, Centro

Agronomico Tropical de Investigacion y Ensenanza (CATIE), Turrialba, Costa Rica) and had been genotyped using microsatellite markers as described in Kuhn et al. (2008) to verify their identity.

The mapping population (Brown et al. 2005) consisted of 146 F_2 individuals from Centro de Pesquisas do Cacau (CEPEC)/Comissao Executiva do Plano da Lavoura Cacaueira (CEPLAC), Itabuna, Bahia, Brazil obtained from selfing an F_1 hybrid cultivar, "TSH516" of a cross "SCA6"×"ICS1." COSII candidates (Strategy I) that showed SNP heterozygosity in "TSH516" were identified through sequencing and selected for mapping.

DNA and RNA isolation

Leaf DNA from the 15 cacao cultivars of the diversity panel was isolated as described in Livingstone et al. (2011). To identify SNPs using Strategy I, candidate COSII primer pairs were first tested for amplification efficiency and then used to amplify leaf DNA from the diversity panel.

Leaf RNA from the diversity panel was used in the more general SNP discovery strategy that utilized the entire leaf transcriptome (Strategy II). Leaf RNA samples from the 15 cacao cultivars of the diversity panel were isolated from 1 g of mature leaves using the method of Bailey et al. (2005) with the following modifications. Leaf tissue (excluding the mid-rib) was collected from each tree and placed immediately in liquid nitrogen, then ground to a fine powder and homogenized in 15 mL of 65°C extraction buffer (Bailey et al. 2005) using a PRO200 Homogenizer with a 10×115-mm generator (PRO Scientific, Inc.). Samples were extracted twice with equal volumes chloroform (1 min homogenization with the PRO200 Homogenizer and 30 min centrifugation at 10,500×*g*). One-third volume of 8 M LiCl was then added, the sample mixed, and the RNA precipitated overnight or over the weekend at 0°C (in an ice water bath in a 4°C refrigerator).

BAC library construction

Three BAC libraries were made from the highly homozygous genotype "Matina 1–6" (Amelonado group) and used for the production of the physical map (Saski et al. 2011) used for Strategy II. The "Matina 1–6" clones used for DNA isolation were kept at greenhouse conditions and dark-treated for 12 h prior to leaf harvesting. Approximately 100 g of young, expanding leaf tissue (mid-vein removed) were harvested, washed two times with autoclaved ddH_20, and ground in liquid nitrogen with a mortar and pestle to a coarse powder. Nuclei were prepared following the methods of Luo and Wing (2003) with the following modifications: inclusion of 1% (*w*/*v*) soluble PVP-40 (Sigma-Aldrich), 0.1% (*w*/*v*) L-ascorbic acid (Sigma-Aldrich), 0.13% (*w*/*v*) sodium diethyldithiocarbamate trihydrate (Sigma-Aldrich), and 0.4% beta-mercaptoethanol (added to the nuclei isolation buffer right before use). Protein digestion and plug washing was carried out exactly as previously described (Luo and Wing 2003). To prepare high molecular weight genomic DNA fragments, plugs were macerated with a single-edge razor blade followed by partial digestion, separately, with *Hin*dIII (library TCC_Ba), *Eco*RI (TCC_Bb), and *Mbo*I (TCC_Bc). DNA size selection, electro-elution, and ligation were carried out following the methods of Luo and Wing (2003).

High information content fingerprinting (HICF) of BACs

Plates, 384-well, containing BACs were decondensed to the 96-well format robotically with the Q-bot (Genetix). Two pins were removed from the sub-plate inoculators to allow for manual insertion of control clones in the E07 and H12 positions to use in assessing data uniformity. DNA was isolated from a total of 108,288 clones from the TCC_Ba, TCC_Bb, and TCC_Bc BAC libraries by following standard alkaline lysis miniprep methods (Sambrook and Maniatis 1989) and used as substrate for HICF following the methods of Luo et al. (2003) except that approximately 0.5 μg of BAC DNA was digested with 2.0 U of *Hin*dIII, *Bam*HI, *Xba*I, *Xho*I, and *Hae*III at 37°C for 2 h. The DNA was labeled with 0.25 U of SNaP-shot (kit from Applied Biosystems, Foster City, CA, USA) at 65°C for 1 h and then precipitated with ethanol using standard procedures. The labeled DNA was then reconstituted in 9 μl of Hi-Di formamide and 0.05 μl of LIZ600 (both from Applied Biosystems, Foster City, CA, USA). BAC fingerprints were sized on an ABI3730 (Applied Biosystems, Foster City, CA, USA) using a 36-cm array and POP7 (Applied Biosystems, Foster City, CA, USA). The fingerprint profiles were processed using GeneMapper 3.7 (Applied Biosystems, Foster City, CA, USA) for sizing quality and FPMiner (Bioinforsoft) for digitized fingerprint assignment. For data quality, vector bands, clones without inserts, and restriction profiles with less than 20 or more than 200 bands were removed and remaining profiles uploaded to FPC v8.5.3 (Soderlund et al. 2000) for fingerprint contig assembly.

Physical map assembly

The initial build was done at $1e^{-80}$ and a tolerance of 3. The DQer function of FPC was used to break down all contigs with more than 10% questionable (Q) clones to reduce false joins. Further physical map refinement was performed with the Ends-to-Ends and Singles-to-Ends functions of FPC and stepwise reductions of the Sulston score cutoff values to a final score of $1e^{-50}$. Additional fingerprint contig merges were made with lower Ends-to-Ends overlap when there was additional agreement with the anchored genetic map.

Contigs merged in this fashion used Sulston score cutoffs as low as $1e^{-25}$.

PCR

Polymerase chain reaction (PCR) cycling conditions on a standard thermocycler for the Strategy I cacao COSII candidate gene primers were 94°C for 2 min; 40 cycles of 94°C for 30 s, 60°C for 1 min, 72°C for 1 min; followed by single cycles at 49°C and 72°C for 1 min each.

Source of data sets from which COSII sequences in cacao were identified

Database-identified COSII sequences

For all strategies, COSII sequences (1,086) in a file named cosii.tar.gz were retrieved from the SOL Genomics Network (SGN) website (ftp://ftp.sgn.cornell.edu/COSII/) in July 2007. The sequences available at that time were only a subset of the 2,869 COSII sequences described in Wu et al. (2006). The FASTA files were renamed with the *Arabidopsis* gene name as the first part of the gene ortholog name, e.g., At3g11210_At3g11210.1 was the renamed *Arabidopsis* ortholog and At3g11210_lgn_216577 was the renamed tomato ortholog. This allowed identification of the non-*Arabidopsis* orthologs in the parsed reports from the BLAST searches. The renamed COSII sequences were then used as query sequences in local TBLASTX searches of the cacao sequence datasets (see below).

ESTs from NCBI database (Strategy I)

In December 2007, we downloaded the ~150,000 unassembled cacao expressed sequence tag (EST) sequences from the NCBI database (http://www.ncbi.nlm.nih.gov/nucest), assembled them into longer contigs with the CAP3 assembler (Huang and Madan 1999), and used them for BLAST searches as described below. The assembled cacao ESTs are currently available from the cacaogenomedb.org website.

Transcriptome sequencing and assembly (Strategy II)

Total leaf RNA of "Matina 1–6" (800 ng) was used in library preparation optimized for Roche/454 GS FLX Titanium sequencing according to protocols developed at the Indiana University Center for Genomics and Bioinformatics (IUCGB) (Carter et al. in press; Shulaev et al. 2011). To reduce the number of high copy transcripts, the amplified dsDNA library intermediate was partially normalized using Trimmer Direct (Evrogen) protocols. Emulsion PCR and sequencing were performed according to the manufacturer (Roche/454 Life Sciences). High quality sequence reads (1,510,557) were trimmed (https://sourceforge.net/projects/estclean/) and assembled using Newbler v2.0 into 29,787 contigs of average length 725 (Std Dev=558). The reads have been deposited in the NCBI short read archive (accession No. SRA027322).

Physical map of an important QTL region of cacao linkage group 5 (Strategy III)

A near equimolar pool of the 27 BACs making up the minimum tiling path (MTP) of a 3-Mbp region of linkage group 5 (Contig 11) of the cacao physical map (Saski et al. 2011) was prepared. This 3-Mbp region was determined to have a consensus QTL for black pod resistance based on seven individual black pod resistance QTLs with effects varying from 5.6% to 11.4% (Lanaud et al. 2009), as well as QTLs for bean weight and pod index (Clement et al. 2003a, b). A 3-kb span paired-end library was synthesized from the DNA pool and sequenced using the Roche/454 GS FLX Titanium platform following the method of Rounsley et al. (2009). The sequencing yielded 654,297 reads that passed quality filters and have been deposited in the NCBI Short Read Archive (accession No. SRA027323). Reads were split using the Celera CABOG Assembler (Miller et al. 2008a) "sffToCA" program using the following non-default settings: -trim chop -clear 454 -linker titanium -insertsize 3000 300. Reads were screened for vector and *Escherichia coli* contamination using Seqclean (http://compbio.dfci.harvard.edu/tgi/software/). Next, reads were trimmed using Lucy2 (Li and Chou 2004) with Pindigo-BAC536 (HindIII Splice) as the masking vector sequence. Preprocessed reads (20× mate-pair and 10× linear based on a 3-Mbp estimate) were converted to FRG "fragment" format using the Celera CABOG script convert-fasta-to-v2.pl (with the following non-default settings for paired end libraries: -paired: -454 -mean 3000 -stddev 20). FRG files were then assembled into scaffolds using the CABOG Assembler (using the following non-default settings: overlapper=mer obtOverlapper=mer ovlOverlapper=mer unitigger=bog). Scaffolds were then ordered in compliance with BLASTN alignment ($E \leq 1e^{-75}$; percent identity≥98%) to contig 11 MTP BAC end sequences and the expected position based on the MTP. Based on this order, a pseudomolecule was constructed by concatenating the scaffolds with an insertion of 70 Ns representing a gap of unknown size between the scaffolds.

BLAST search strategy for identifying COSII sequences in the sequence databases

The three cacao sequence datasets described immediately above were formatted to allow local BLAST searches using the program formatdb from the NCBI toolbox (www.ncbi.nlm.nih.gov). The TBLASTX searches with the renamed

COSII sequences as the query were done using the program "blastall" from the NCBI toolbox. The BLAST reports were parsed with the program "blast_parser" (http://cgpdb.ucdavis.edu/BlastParser/Blast_Parser.html). For Strategies I and III, the parsed BLAST reports were manually analyzed to identify candidate COSII cacao genes using the criteria described by Wu et al. (2006). Candidates for Strategy II were identified using a Perl program based on the criteria described in Wu et al. (2006). The script first identified all the COSII query sequences that had an *e*-score of less than $1e^{-50}$ in the TBLASTX search. Next, the script identified the orthologs that had a best hit to the same contig and had a second best hit to a different contig with an *e*-value tenfold greater than the best hit. The multiple orthologs of the query sequences related to a single locus had to have the same cacao sequence as the top hit of a TBLASTX search. In addition, if the second best hit in the dataset was to a different cacao sequence, that hit had to have an *e*-score at least tenfold greater than the top hit (e.g., if the top hit was $e=-50$, the next best hit would have to be $e\geq-40$), as suggested in the initial report identifying COSII sequences (Wu et al. 2006).

SNP discovery within COSIIs, genetic mapping, and integration into physical map

Of 132 COSII candidates identified using Strategy I, 25 were chosen due to their relatively even distribution throughout the *Arabidopsis* genome as determined based on the *Arabidopsis* gene-naming convention (e.g., At1g00100 is a gene on the proximal end of chromosome 1). Intron positions were identified within the candidate sequences using the Intron Finder program available at SGN (http://solgenomics.net). Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) was used to design primers to flank at least one intron and amplify approximately 500 bps of each candidate COSII sequence, including an estimated 150 bps of the respective intron. Only those primers that amplified a single product that was less than 1 kb in length were then used to amplify DNA from the diversity panel, and the amplicons were directly sequenced as described by Kuhn et al. (2008). The resulting sequences originating from a given primer pair were aligned and examined for SNPs using the phred/phrap/polyphred/consed pipeline (Ewing et al. 1998; Ewing and Green 1998; Gordon et al. 1998; Stephens et al. 2006).

When the genotype "TSH516" was determined to be heterozygous for a SNP, the primers were used to amplify the DNA from all 146 individuals in the F_2 population. Amplicons from all individuals were directly sequenced, and SNPs called using polyphred (Stephens et al. 2006) were used as genotyping data. Genotyping by sequencing was less expensive than designing fluorescently labeled probes for the 146 individuals in the mapping population or using a bead array-based assay. The SNP genotyping data identified by sequencing were combined with 180 previously mapped simple sequence repeat and single-strand conformation polymorphism (SSCP) markers (Brown et al. 2005). Genotype data were imported into JoinMap v4 (Van Ooijen 2006) and used to create a genetic map using regression mapping with the Kosambi mapping function (Kosambi 1944).

Cacao COSII sequences identified using Strategy II were mapped by designing overgo probes to non-intron-containing regions, hybridizing the probes to the BAC libraries, and using the hybridization results to place each COSII sequence on the physical map constructed from three complementary BAC libraries of "Matina 1–6" (Saski et al. 2011). We also converted all the mapped markers from two genetic recombination maps (Brown et al. 2005; Pugh et al. 2004) into overgo probes and were able to link the genetic recombination map and the physical map using these markers as anchors, which can be seen as a CMAP view on (http://www.cacaogenomedb.org). High-quality marker sequences were processed through a Clemson University Genomics Institute (CUGI) pipeline consisting of Repeat Masker (Smit et al. 1996–2010) with the RepBase database (Jurka et al. 2005), Cross_Match, and Tandem Repeat Finder (Benson 1999). The remaining sequences were used for overgo design using the overgomaker software (http://genomeold.wustl.edu/tools/software/overgo.cgi). Manual filter hit-calling and deconvolution of the multi-dimensional pool hybridization data was accomplished using HybDecon, an open source software available at http://www.genome.clemson.edu/software/hybdecon. The filter hit-calling functionality is an improved version of Hybsweeper, a web-based Java tool first reported in 2005 (Lazo et al. 2005). In addition, a Perl-based deconvolution script, written entirely at CUGI, accompanies the tool and is launched from within the graphical interface once manual calling of positive hits is complete. The source code, a test dataset, and installation manual are all available online (http://www.genome.clemson.edu/software/hybdecon).

Cacao COSII sequences identified using Strategy III had already been mapped as the pooled BACs utilized for that strategy had been localized on linkage group 5 based on the physical map (Saski et al. 2011).

SNP discovery for exonic regions (Strategy I–III)

SNP discovery in COSII sequences of cacao identified using Strategies II and III was undertaken using the "Matina 1–6" leaf transcriptome (described above) as the reference sequence. Leaf RNA isolated from the leaves of each of the 15 members of the diversity panel was sequenced on the Illumina GAII platform at the National Center for Genome Resources (NCGR). Briefly, following quality assessment using a Bioanalyzer 2100 (Agilent Inc.,

Santa Clara, CA, USA), poly A+RNA was isolated from total RNA by two rounds of oligo-dT selection using a kit available from Invitrogen, Inc. (Santa Clara, CA, USA). mRNA was annealed to random hexamers and reverse transcribed. Following standard second-strand synthesis, end repair, and A-tailing, adapters complementary to Illumina sequencing primers were ligated to the cDNA fragments. Resultant cDNA libraries were size-fractionated on agarose gels; the 300-bp fragments were excised and then amplified in unbiased PCRs and sequenced on the Illumina GAII. Library assessment was performed using a Bioanalyzer 2100. Sequence analysis was performed using Illumina GAII sequencing instruments (Illumina, San Diego, CA, USA) and standard procedures.

Variants were identified by aligning the Illumina reads with the "Matina 1–6" leaf transcriptome using the Alpheus analysis pipeline, database, and visualization software (Miller et al. 2008b). Filtering criteria for variant detection were as follows: variant bases had a quality score of Q20 or higher, a frequency of 20% or higher at the position calling the variant allele, and two or more sequence reads calling the variant. A variant report was generated from the Alpheus output. The MAF was calculated for each locus, by determining the ratio of minor (variant) alleles to the total number of possible alleles for all 15 individuals in the diversity panel. Variants were thus linked to individual "Matina 1–6" transcriptome contigs. The COSII candidate sequences were used as the query in a local BLASTN search of the "Matina 1–6" leaf transcriptome. We used the variant report to identify SNPs for all the mapped cacao COSII markers, when possible, by identifying the cacao transcriptome contig in the reference transcriptome by BLAST search and looking up that contig in the variant report.

Statistical analysis of synteny

A chi-square test was performed to determine if the percentage of individual linkage groups from sampling *Arabidopsis* (N_{LG}=5) and tomato (N_{LG}=12) occurred with a significantly reliable frequency ($p \geq 0.05$) in ten different cacao linkage groups.

Results

Identification and mapping of cacao COSII sequences

Strategy I The TBLASTX search of the assembled NCBI EST dataset conducted using the 1,086 COSII loci and ortholog sequences as the query resulted in 132 COSII candidates with *e*-scores of <−50. We chose 25 from among the intron-containing candidates (as SNPs occur more frequently in introns) to test. The *Arabidopsis* genes that correspond to those 25 are distributed among all five chromosomes (five candidates per linkage group). Without a cacao genome reference, we targeted these 25 with the hope of finding widely distributed markers in cacao. Primers designed to flank introns were designed for all 25 and tested for amplification success. Of the 25 primer pairs tested, 13 sets yielded either no product, multiple products, or inconsistent amplification results.

The 12 primer pairs that gave consistent results were used to amplify DNA isolated from the cacao cultivars in a diversity panel consisting of 15 genetically diverse genotypes (Motamayor et al. 2008). Direct sequencing of the resulting amplicons was successful for 10 of the 12 loci, and a total of 66 SNPs were identified among them, all from the intron region. Loci that were heterozygous in "TSH516," the selfed parent of the F_2 mapping population, were Tc_At1g44446, Tc_At1g63780, Tc_At1g77470, Tc_At3g15290, Tc_At3g63490, and Tc_At4g24830. Primers for these loci were used to amplify DNA isolated from the 146 individuals of the F_2 mapping population. SNPs identified by direct sequencing of the resulting amplicons were genetically mapped (Fig. 1). Map positions of all the COSII markers identified using this first strategy are given in Table 1.

Strategy II We identified 128 candidate cacao COSII sequences as a result of conducting a TBLASTX search of a 454-sequenced "Matina 1–6" leaf transcriptome with the 1,086 COSII sequences. The 128 candidate sequences were annotated for intron position, and overgo probes were designed to avoid intron–exon junctions based on the annotated sequences. Thus, all the overgo probes would be in exon regions of the candidate genes.

We then took our 128 overgo probes for cacao COSII candidates identified from the "Matina 1–6" transcriptome of unknown map position and, through hybridization to the BAC arrays, placed them on the physical map and inferred their positions on the genetic recombination map as well, based on the hybridization results of the overgo probes for the previously mapped microsatellite and candidate gene markers. As a result of these overgo hybridizations, 81 COSII candidate gene sequences mapped to BACs that were anchored to the recombination genetic map, 19 mapped to unanchored BAC contigs, and 28 probes failed to hybridize or hybridized to more than one contig. The 81 sequences that mapped to the physical map were distributed across all ten cacao linkage groups (Table 1) and included four of the six COSII markers identified and mapped using Strategy I (Tc_At1g63780, Tc_At1g77470, Tc_At3g63490, and Tc_At4g24830). All four markers mapped to the same region on the physical map as they had on the genetic recombination map. Distribution of the mapped COSII markers across the ten chromosomes is shown in Fig. 1, and map positions are given in Table 1.

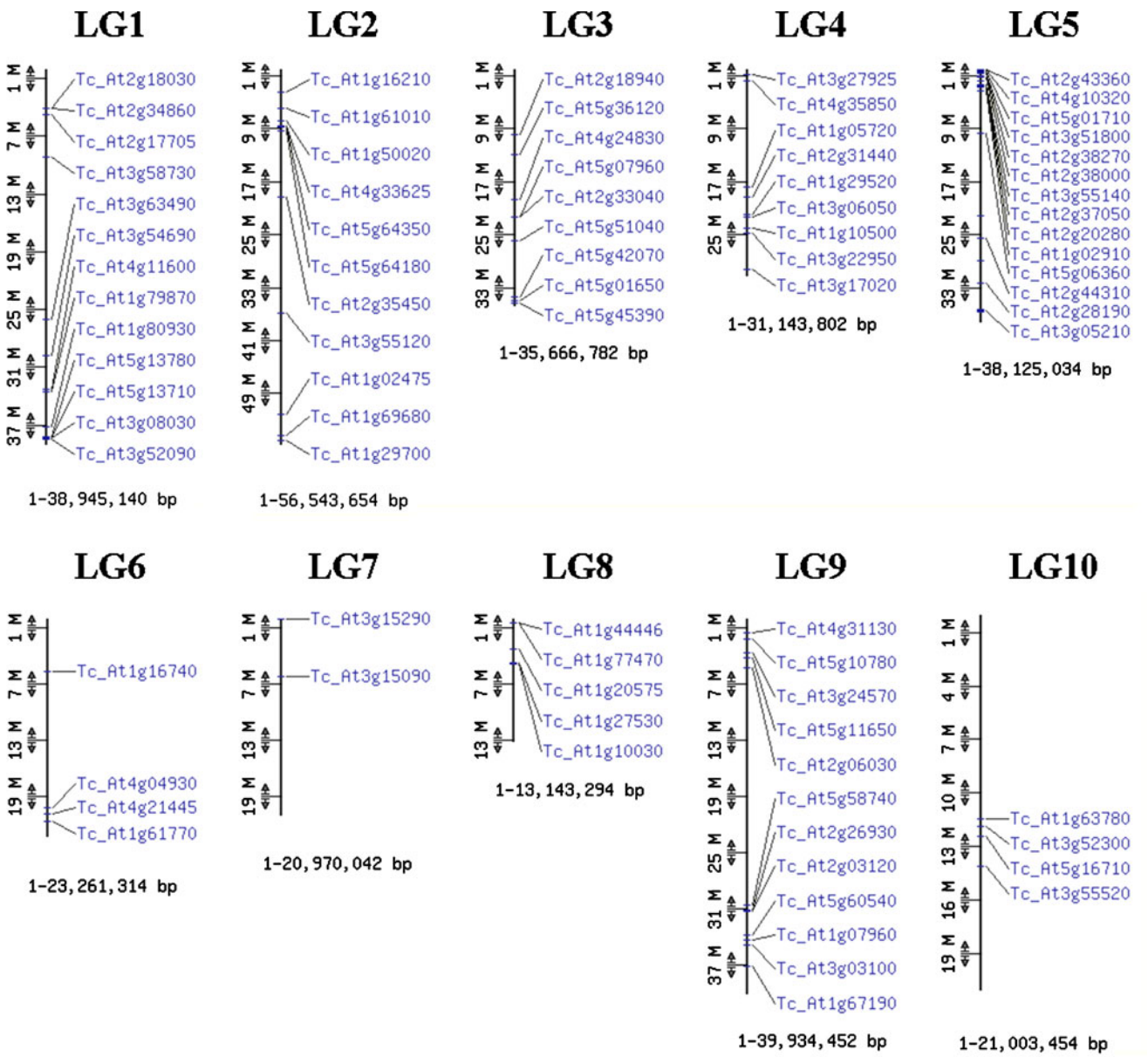

Fig. 1 Distribution of the 83 COSII SNP markers on the ten linkage groups of cacao. The ten linkage groups are as defined by mapping of microsatellite markers in Brown et al. (2005). COSII markers were placed on the physical map by hybridization of overgo probes to the BAC libraries; their positions on the physical map, in base pairs, were estimated from the minimum tiling path for each pseudomolecule (*chromosome*). *Numbers under the linkage groups* give size in base pairs for the pseudomolecule as estimated from the minimum tiling path

Strategy III Two cacao COSII sequences were identified in an important QTL region of cacao linkage group 5 using Strategy II. These same two COSII sequences, Tc_At2g43360 and Tc_At3g11210, were then validated using Strategy III. A TBLASTX search of the 3-Mbp sequence region of linkage group 5 with the 1,086 COSII sequences as query identified 17 cacao COSII sequences (including the two previously identified by hybridization), which were homologous to COSII loci from all five *Arabidopsis* linkage groups (Table 1).

The number of unique mapped sequences resulting from all three strategies is 98; six unique loci were identified using Strategy I, 77 unique loci were identified using Strategy II (another four were also identified using Strategy I), and 15 unique loci were identified from Strategy III (another two had also been identified using Strategy II).

Development of SNP markers from COSII genes

We aligned the leaf transcriptome of "Matina 1–6," one of the 15 accessions in the diversity panel to the leaf transcriptome sequences from the remaining 14 culti-

Table 1 Detailed description of the mapped cacao COSII markers

Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]	Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]	Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]
Tc_At1g79870[c]	II	1	3	0.48	Tc_At1g05720	II	4	1	0.29	Tc_At4g10320	III	5	5	0.34
Tc_At1g80930[d]	II	1	12	0.46	Tc_At1g10500	II	4	1	0.16	Tc_At5g01230	III	5	2	0.36
Tc_At2g17705	II	1	1	0.28	Tc_At1g29520[c]	II	4	1	0.40	Tc_At5g01710	III	5	3	0.22
Tc_At2g18030	II	1	3	0.19	Tc_At2g31440	II	4	3	0.42	Tc_At5g02230	III	5	4	0.43
Tc_At2g34860	II	1	9	0.38	Tc_At3g06050[c]	II	4	9	0.40	Tc_At5g06360	II	5	0	
Tc_At3g08030	II	1	0		Tc_At3g17020	II	4	5	0.24	Tc_At1g16740	II	6	1	0.13
Tc_At3g52090	II	1	1	0.34	Tc_At3g22950	II	4	7	0.27	Tc_At1g61770	II	6	5	0.42
Tc_At3g54690[c]	II	1	2	0.42	Tc_At3g27925[f]	II	4	2	0.39	Tc_At4g04930	II	6	3	0.31
Tc_At3g58730	II	1	2	0.43	Tc_At4g35850	II	4	5	0.36	Tc_At4g21445	II	6	1	0.11
Tc_At3g63490	I	1	0		Tc_At1g02910	II	5	11	0.43	Tc_At3g15090	II	7	3	0.50
Tc_At4g11600[c]	II	1	2	0.17	Tc_At1g54780	II	5	1	0.17	Tc_At3g15290	I	7	3	0.43
Tc_At5g13710	II	1	2	0.46	Tc_At1g56290	III	5	4	0.40	Tc_At1g10030[d]	II	8	4	0.35
Tc_At5g13780[d]	II	1	3	0.24	Tc_At1g65000	II	5	0		Tc_At1g20575	II	8	2	0.41
Tc_At1g02475	II	2	3	0.39	Tc_At2g20280	III	5	4	0.20	Tc_At1g27530[d]	II	8	1	0.33
Tc_At1g16210[c,d]	II	2	6	0.44	Tc_At2g28190	II	5	2	0.47	Tc_At1g44446	I	8	2	0.38
Tc_At1g29700	II	2	6	0.42	Tc_At2g36930	II	5	4	0.50	Tc_At1g77470	I	8	12	0.50
Tc_At1g50020	II	2	2	0.45	Tc_At2g37050	III	5	9	0.39	Tc_At1g07960[e]	II	9	0	
Tc_At1g61010	II	2	1	0.09	Tc_At2g37110	III	5	2	0.37	Tc_At1g67190[e]	II	9	0	
Tc_At1g69680[c]	II	2	1	0.18	Tc_At2g37240	III	5	0		Tc_At2g03120[e]	II	9	6	0.45
Tc_At2g35450	II	2	1	0.25	Tc_At2g38000	III	5	2	0.45	Tc_At2g06030	II	9	2	0.34
Tc_At3g55120[c]	II	2	2	0.37	Tc_At2g38270	III	5	7	0.28	Tc_At2g26930[e]	II	9	7	0.38
Tc_At4g33625	II	2	1	0.13	Tc_At2g41680[c]	II	5	1	0.12	Tc_At3g03100[e]	II	9	9	0.49
Tc_At5g64180	II	2	1	0.30	Tc_At2g43360	II	5	7	0.43	Tc_At3g24570	II	9	0	
Tc_At5g64350	II	2	2	0.23	Tc_At2g44310[c]	II	5	4	0.46	Tc_At4g31130	II	9	2	0.48
Tc_At2g18940	II	3	0		Tc_At3g05210	II	5	1	0.38	Tc_At5g10780	II	9	2	0.23
Tc_At2g33040	II	3	6	0.46	Tc_At3g08950	III	5	0		Tc_At5g11650	II	9	2	0.44
Tc_At4g24830	I	3	1	0.33	Tc_At3g09320	III	5	7	0.35	Tc_At5g58740	II	9	2	0.50
Tc_At5g01650	II	3	1	0.19	Tc_At3g09740	III	5	3	0.20	Tc_At5g60540[e]	II	9	4	0.47
Tc_At5g07960	II	3	0		Tc_At3g11210	II	5	3	0.49	Tc_At1g63780	I	10	3	0.38
Tc_At5g36120	II	3	5	0.46	Tc_At3g51800	III	5	3	0.47	Tc_At3g52300	II	10	3	0.49
Tc_At5g42070	II	3	0		Tc_At3g55140	III	5	5	0.50	Tc_At3g55520	II	10	1	0.25

Table 1 (continued)

Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]	Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]	Marker name	Strategy	Linkage group	v2 SNPs[a]	MAF[b]
Tc_At5g45390	II	3	0		Tc_At3g57290	II	5	2	0.41	Tc_At5g16710	II	10	3	0.48
Tc_At5g51040[e]	II	3	5	0.31	Tc_At4g02680	III	5	3	0.36					

Marker name is derived from the *Arabidopsis* ortholog. Strategy refers to which strategy was used to identify the cacao COSII ortholog. Linkage groups as defined by genetic recombination mapping in Brown et al. (2005). Co-location with quantitative trait locus (QTL) as inferred from map position of COSII ortholog and with reference to published position of QTL (see footnotes for individual QTL references). v2 SNPs refers to the number of SNPs identified for a particular ortholog from a variant report based on the cacao leaf transcriptome. Minor allele frequency (MAF) for the best SNP for each locus was determined from allele frequency in the 15 genetically diverse cultivars used in the SNP discovery project

[a] v2 SNPs: SNPs identified from the SNP discovery project by comparison with the Matina 1–6 reference transcriptome

[b] MAF: minor allele frequency of preferred SNP marker (SNP 1, ESM 1 and ESM 2)

[c] Consensus (c) black pod (BP) QTL as in Lanaud et al. (2009)

[d] Frosty pod (FP) and black pod (BP) QTL identified by Miami (m) group in Brown et al. (2007)

[e] Wet bean weight (WBW), bean length (BL), and pod number (PN) QTL as described in Clement et al. (2003b)

[f] Witches' broom (WB) QTL identified in an F_2 population created by the self fertilization of TSH516 (bz) population as in Brown et al. (2005)

vars of the cacao diversity panel, and generated a variant report from the aligned reads. Filters were applied to retain contigs that at the variant site contain: (1) polymorphism in at least two cultivars and (2) conservation with the reference sequence in at least two cultivars. SNPs were also filtered by isolation, removing those with neighboring variants within 60 nucleotides on either side. The cacao COSII gene candidates were then queried against the reference transcriptome in a BLASTN search, and the contig hits were used to look up the SNPs in the filtered SNP dataset, which only contained SNPs found in exon regions. Of the 98 total COSII loci mapped using the three strategies, 83 had corresponding SNPs in this filtered SNP dataset (Table 1). When multiple SNPs were identified at a single locus, we chose the SNP with the greatest MAF, based upon the allele frequency in the genotypes of the 15 genetically diverse cultivars used for SNP discovery, for further evaluation (Table 1). The Electronic Supplemental Material accompanying this article contains a table of all SNPs identified for the COSII loci (ESM 1) and another of the sequences flanking the SNPs (ESM 2) to allow SNP assay design for individual loci. We compared the map positions of our final group of SNPs with the known map positions of disease resistance (Brown et al. 2005; Borrone et al. 2007; Clement et al. 2003a) and horticultural (Clement et al. 2003b) QTLs in cacao. Nineteen of our SNPs were co-located with previously determined QTLs for black pod, frosty pod, witches' broom, wet bean weight, bean length, and pod number (Table 1).

Evaluation of synteny among the genomes of cacao, *Arabidopsis*, tomato, pepper, and eggplant

We then explored the use of the cacao COSII markers for comparative genomics by carrying out synteny studies. We compared our 98 mapped COSII markers in cacao to their counterparts in eggplant, pepper, and tomato (http://solgenomics.net). In total, 39 of the markers in cacao had also been mapped in tomato (Table 2), of which 13 also mapped in pepper and four in eggplant (data not shown); three markers (At2g37240, At3g03100, and At3g11210) were common to all four species. In cacao, linkage groups are arbitrarily but uniformly numbered based on the occurrence of common anchor markers, whereas linkage groups in *Arabidopsis* and tomato have been cytogenetically linked to chromosomes. In several cases, COSII markers found in one linkage group in cacao were also found in one linkage group in tomato, but their relative map positions were often very different. For example, two tomato COSII markers that were only 0.7 cM apart (At3g08030 and

Table 2 Comparison of cacao COSII markers with orthologs in *Arabidopsis* and tomato

Marker name	Cacao linkage group	Cacao map position (cM)	Tomato linkage group	Tomato map position (cM)
At3g08030	1	100	1	47.4
At3g63490	1	53	1	43.5
At3g52090	1	100	11	82.7
At3g58730	1	38	1	46.7
At2g18030	1	16	2	83.1
At2g34860	1	16	1	7.5
At1g50020	2	31	1	146.1
At3g55120	2	64	5	44
At1g16210	2	11	1	165
At5g64350	2	32	1	137.2
At5g07960	3	17	6	52.3
At4g24830	3	24	5	51
At5g51040	3	28	3	134
At1g10500	4	62	5	76
At1g29520	4	56	3	52.7
At3g06050	4	56	1	46.3
At2g37240	5	15	9	1
At5g06360	5	28	9	116
At2g41680	5	38	9	15.5
At3g11210	5	5	6	24.5
At3g09740	5	16	10	71.5
At2g36930	5	21	9	15.1
At5g02230	5	8	9	42.5
At2g38270	5	12	9	0.5
At2g43360	5	0	6	67
At1g02910	5	24	9	57.7
At4g04930	6	47	10	23.5
At3g15290	7	28	7	63.5
At1g27530	8	28	4	88.3
At1g20575	8	21	4	132.7
At1g44446	8	39	11	40.5
At1g10030	8	28	4	88.7
At1g77470	8	40	6	39.1
At1g07960	9	73	11	82.5
At4g31130	9	0	8	3.5
At5g60540	9	66	11	76.5
At2g03120	9	58	1	18.2
At3g03100	9	74	3	72.7
At5g16710	10	43	11	31.4

Marker name is *Arabidopsis* ortholog and designates linkage group (Atx) and relative position (gxxxxx). All map positions for cacao and tomato are in centimorgans (cM)

At3g58730) appeared 62 cM apart in cacao. Two cacao markers that were coincident on LG1 in cacao (Tc_At2g18030 and Tc_At2g34860) were present on separate linkage groups (LG1 and LG2) in tomato. However, there were several instances in which COSII genes occur on the same linkage group in both cacao and *Arabidopsis* or tomato with a frequency much higher than chance (Table 3). For example, all five COSII genes on LG8 in cacao are found on LG1 in *Arabidopsis* (χ^2=320, p< 0.0001), and three of those five are found on LG4 of tomato (χ^2=320, p<0.0001). Eleven of the 29 COSII genes on LG5 in cacao were located on LG2 in *Arabidopsis* (χ^2=16.08, p=0.0029). Eight of the eleven COSII genes on LG5 in cacao mapped to LG9 in tomato (χ^2=616, p<0.0001) and four of those eight were also located on LG2 in *Arabidopsis*.

Table 3 Summary of cacao COSII markers grouped by linkage group

Linkage group	Number of COSII	Number with SNPs	Number with SNPs co-located with QTL	Linkage group match with *Arabidopsis*	χ^2 value, p value	Linkage group match with tomato	χ^2 value, p value
1	13	12	5	LG3 (5/13)	17.01, 0.0019	LG1 (4/6)	408.33, <0.0001
2	11	11	2	LG1 (6/11)	59.669, 0.0001	LG1 (3/4)	533.33, <0.0001
3	9	6	1	LG5 (6/9)	108.89,<0.0001		
4	9	9	3	LG3 (4/9)	29.88, <0.0001		
5	29	23	2	LG2 (11/29)	16.08, 0.0029	LG9 (8/11)	616.333, <0.0001
6	4	4	0				
7	2	2	0	LG3 (2/2)	320, <0.0001		
8	5	5	2	LG1 (5/5)	320, <0.0001	LG4 (3/5)	320.333, <0.0001
9	12	9	4				
10	4	4	0	LG3 (2/4)	14.58, 0.0057		
Total	98	83	19				

Columns 2–4 represent the number of COSII markers in each linkage group, the number of COSII markers with identified SNPs, and the number of COSII markers with SNPs that map to a QTL region. The linkage group of *Arabidopsis* or tomato that most frequently matched the cacao markers and, in parentheses, the number of cacao orthologs out of the total for the linkage group that matched are shown in columns 5 and 7, respectively. Results of the statistical analysis involving *Arabidopsis* and tomato are shown in columns 6 and 8, respectively

Discussion

To date, most of the molecular markers used in cacao genetic mapping and diversity studies have been RAPDs, AFLPs, microsatellites, or candidate genes (Lanaud et al. 1999; Brown et al. 2008; Kuhn et al. 2005). The first three of these are frequently found outside of protein-coding regions. The lack of association with transcribed genes makes these markers less useful in comparative genomics projects and, additionally, sharing RAPD, AFLP, and microsatellite marker data is technically difficult. We were therefore interested in taking advantage of the rapid advances being made in the sequencing of genomes from various plant species to develop a set of genetic markers for cacao that were based on single-copy genes. Because single-copy conserved orthologs have been well defined and used for mapping in a variety of plants (Wu et al. 2009a, b; Lefebvre-Pautigny et al. 2010), we undertook this study of them in cacao. Our hope was that, as more COSII sequences are identified in other crop plants and co-located with existing QTLs, we could use SNP markers for those loci for MAS in cacao. We are working with the international cacao research community to develop a core set of SNPs to be used both in breeding and in genotyping of germplasm collections.

Over the course of our study, we benefitted from the availability of increasing amounts of mapping and sequencing data from our ongoing genome-sequencing project and consequently employed three different strategies to identify and map cacao COSII genes. Each strategy was comprised of two parts that varied in the source of reference sequence data used and the method by which the COSII candidates were mapped. Common to all strategies was the use of the COSII ortholog sequences to screen the sequence data for COSII candidates.

Strategy I was modeled upon our successes with developing SSCP markers from cacao ESTs (Kuhn et al. 2008) and resulted in 25 candidates, of which, six were mapped (Table 1). By design, the SNPs identified by Strategy I reside in introns. The ~150,000 cacao EST sequences utilized for this strategy represented shallow sequencing (much less than 1× coverage) of about 60 different EST libraries (~2,500 sequences/library) from many different tissues, cultivars, and environmental conditions (Argout et al. 2008). Because of the shallow sequencing, COSII sequences may be under-represented in this EST sequence library, and that is likely why Strategy 1 resulted in fewer numbers of SNP markers than the other strategies we employed.

In Strategy II, sequencing of the "Matina 1–6" leaf transcriptome was extensive (~10× average coverage), but because it only represents transcripts from a single tissue from a single genotype grown under constant environmental conditions, COSII sequences may also be under-represented. We also used very stringent conditions to identify COSII candidate genes to avoid genes that may occur more than once in the genome, which may have contributed to our identifying only ~10% of the COSII genes that we used as query sequences. Four of the COSII candidate genes that had been identified and mapped using Strategy I were also identified and mapped using Strategy II. All four of these COSII candidate genes mapped to the same locations using both Strategy I and Strategy II, thus validating these results. The Strategy II approach for mapping was much more successful than the Strategy I approach, however, as it did not require markers to be heterozygous in "TSH516," and 81 of the 128

candidates were mapped. The SNPs we identified among the Strategy II candidates were all in gene coding regions.

In Strategy III, 17 COSII markers were identified and mapped within a 3-Mbp region of cacao LG5 (Feltus et al. 2011). Two of the 17 COSII genes identified in that region had been identified using Strategy II, which allowed us to manually adjust the stringency used to accept a candidate, contributing to the greater number of candidates identified in that relatively short segment of the cacao genome when compared to Strategy I and II. Using an estimated size of 430-Mbp (Argout et al. 2011) for the cacao genome, extrapolation of these results would predict a total of some 2,437 COSII genes in cacao, a sum similar to the 2,869 COSII genes that have been identified in *Arabidopsis*. Our lower estimation is likely due to querying the sequence datasets with only the 1,086 COSII genes that were initially available to us. The map positions for the two COSII genes in the 3-Mbp region of LG5 (Tc_At3g11210 and Tc_At2g43360) determined using Strategy II correlated with their positions as determined in the sequenced pseudomolecule for that region and thus corroborated the Strategy II mapping. Based on the physical positions of the genes in the pseudomolecule and our estimate of 430-Mbp for the size of the cacao genome, we estimate that, on average, 1 cM is ~130 kb in the cacao genome, although cM size may vary based on position along the chromosome.

In total, we identified SNP markers for 83 of the 98 cacao COSII genes we mapped. The sequences associated with each SNP are 121-mers with the SNP in the middle and no other SNP identified in the 60 nucleotides on either side (ESM 2), which allows for design of TaqMan probes or allele-specific PCR primers to assay these SNPs. The 83 SNP markers are distributed across all ten cacao linkage groups (Fig. 1), and 19 of them are co-located with previously identified QTLs for black pod, frosty pod, and witches' broom resistance and wet bean weight, bean length, and pod number and may therefore be useful for MAS for these traits. SNPs marked with a "d" or "f" in Table 3 are in QTLs that our group have identified (Brown et al. 2007, 2005) and for which we have the original field data. We are currently identifying which SNP allele is associated with the favorable trait for these QTLs to allow their use in MAS. For the other populations (Clement et al. 2003a, b; Lanaud et al. 2009), we do not have access to the original field data, but breeders can score their own mapping populations with these SNP markers to identify the favorable alleles.

We had previously developed SNP markers from mapped SSCP markers, and seven of them are currently being used for cacao genotyping in Ghana (Jemmy Takrama, Cocoa Research Institute of Ghana, personal communication). For example, a Taqman SNP assay using an end-point fluorescence plate reader has been used at the Cocoa Research Institute of Ghana (CRIG) to genotype cacao clones used in the breeding program and in the clonal seed gardens from which CRIG distributes hybrid seeds to farmers (Livingstone et al. 2011). The availability of the COSII SNP markers we report herein vastly expands the number of SNP markers and the extent of the cacao genome covered by those markers, and thereby increases the resolution of the genotyping possible. These COSII SNP markers, especially those associated with important disease resistance QTLs, will be immediately useful in the cacao breeding programs at CRIG, the Centre National de Recherche Agronomique (CNRA) in Cote d'Ivoire and the International Institute of Tropical Agriculture (IITA) in Nigeria, which have recently purchased the plate readers necessary for the assay, as well as any other cacao breeding program.

In light of successful production of high resolution synteny maps relating coffee and tomato (Lefebvre-Pautigny et al. 2010) and predictions of map positions in eggplant based on striking synteny with pepper and tomato (Wu et al. 2009b), we hope to exploit synteny between cacao and cotton, with particular interest in identifying syntenic blocks of genes that are linked with traits such as disease resistance, self-incompatibility, and seed oil content and composition. Using the complete genome assembly, further analyses of syntenic blocks of genes between cacao LG5 and tomato LG9 are warranted, as tomato has extensive QTL mapping data available (http://solgenomics.net). As QTL data become available for the Rosaceae (Cabrera et al. 2009), whole genome synteny analysis based on COSII sequences as unique single copy anchors will be valuable. As more crop plant genomes are sequenced, we will examine them for regions of high conservation with cacao and, ideally, identify additional syntenic blocks of COSII genes that allow alignment of the sequenced genomes. By searching those aligned regions for similar QTLs, we hope to identify blocks of single-copy markers that are co-located with disease resistance and other important agronomic traits across a number of species. Depending on the value of the traits they are linked to, these syntenic blocks of COSII genes, identified through leveraging the information of comparative genomics, could be the best suited for inclusion in a core set of SNPs to be used by cacao breeders and by germplasm curators for genotyping. In addition, we plan to map the COSII genes identified in this study in an F_1 mapping population of *T. grandiflorum*. We would like to reveal the syntenic relationships between this species and *T. cacao* in order to identify genomic regions in both species involved with important agronomic traits. A previous marker study with *T. grandiflorum* demonstrated that *T. cacao* microsatellite markers work

with high efficiency in *T. grandiflorum* (Alves et al. 2006). *T. grandiflorum* is thus an excellent candidate for comparative genomic studies with cacao.

The most efficient way to identify and position COSII genes in cacao will be to use the complete *T. cacao* genome sequence. However, while the genome sequence of the cacao cultivar "Criollo" (B97-61/B2) was recently published (Argout et al. 2011) and a draft sequence of the genome of the cacao cultivar "Matina 1–6" was released in September 2010 (www.cacaogenomedb.org), our genome-wide analysis of COSII loci awaits further refinements in the anchoring and orientation of the draft assemblies. In the meantime, all three strategies presented herein were successfully used to identify and position COSII genes in cacao and, depending upon the genetic information available, scientists can employ any of them to efficiently identify COSII genes in their species of interest.

MAS in crop plants has been discussed for decades but does not have a large number of success stories (Collard and Mackill 2008; Hospital 2009). Reasons cited for this lack of progress range from the lack of sufficient molecular markers in previous years to the difficulty of matching MAS to the needs of breeders. The obvious advantages, especially for tree crops, are the ability to screen seedlings for favorable alleles, identify off-types in breeding material, take advantage of genetic diversity in making crosses, and share unambiguous genotype information with other breeders. Disadvantages have been that QTLs are often not stable across environments or crosses, traditional breeding practices have been successful without molecular markers, and the added expense and technical difficulty of collecting and sharing marker data. Most successful MAS to date has been with microsatellite markers on annual crops (Collard and Mackill 2008; Quraishi et al. 2009) with the notable exception of the use of COSII SNP markers to fine map a flour quality QTL in wheat (Quraishi et al. 2009). In tree crops, Rosaceae Conserved Ortholog Set (RosCOS) SNP markers have been identified and mapped in *Prunus* (peach, cherry, apricot, and almond) (Cabrera et al. 2009). However, the RosCOS SNPs have not yet been successfully co-located with QTLs. We are the first to identify and map COSII loci in cacao and to co-locate them with existing QTLs. In addition, we have identified SNPs in the exon regions of 83 of the mapped COSII loci, which can be used by breeders in their own populations to both genotype currently used breeding material and farmers' selections with improved traits and identify the alleles associated with the favorable traits in QTLs in those populations. By providing more SNP markers which have already had some success in West Africa, the most important cacao-growing area in the world, we hope that we will promote progress in using MAS in these breeding programs.

Acknowledgments We wish to acknowledge Mars, Inc. for partial funding of this project, Barbie Freeman for excellent technical support, Dr. Belinda Martineau for editing the manuscript, and Dr. J. Michael Moore for assistance with the statistical analysis of the synteny data.

References

Alves RM, Sebbenn AM, Artero AS, Figueira A (2006) Microsatellite loci transferability from *Theobroma cacao* to *Theobroma grandiflorum*. Mol Ecol Notes 6(4):1219–1221. doi:10.1111/j.1471-8286.2006.01496.x

Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Loor G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C (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 Genom 9:19. doi:10.1186/1471-2164-9-512

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Berard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C (2011) The genome of *Theobroma cacao*. Nat Genet 43(2):101–108. doi:10.1038/ng.736

Bailey BA, Strem MD, Bae HH, de Mayolo GA, Guiltinan MJ (2005) Gene expression in leaves of *Theobroma cacao* in response to mechanical wounding, ethylene, and/or methyl jasmonate. Plant Sci 168(5):1247–1258. doi:10.1016/j.plantsci.2005.01.002

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res 27(2):573–580

Borrone JW, Kuhn DN, Schnell RJ (2004) Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. Theor Appl Genet 109(3):495–507

Borrone JW, Brown JS, Kuhn DN, Motamayor JC, Schnell RJ (2007) Microsatellite markers developed from *Theobroma cacao* L. expressed sequence tags. Mol Ecol Notes 7(2):236–239

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 F-2 population using SSR markers and candidate genes. J Amer Soc Hort Sci 130(3):366–373

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(5):1851–1858. doi:10.2135/cropsci2006.11.0753

Brown J, Sautter R, Olano C, Borrone J, Kuhn D, Motamayor J, Schnell R (2008) A composite linkage map from three crosses between commercial clones of cacao, *Theobroma cacao* L. Trop Plant Biol 1(2):120–130

Cabrera A, Kozik A, Howad W, Arus P, Iezzoni AF, van der Knaap E (2009) Development and bin mapping of a Rosaceae conserved ortholog set (COS) of markers. BMC Genom 10. doi:10.1186/1471-2164-10-562

Carter J, Smith Z, Mockaitis K (in press) Library preparation for transcriptome discovery using long read 454 sequencing. In: Springer P (ed) Methods in Molecular Biology: Plant Functional Genomics, Springer, New York

Chapman MA, Chang J, Weisman D, Kesseli RV, Burke JM (2007) Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). Theor Appl Genet 115 (6):747–755. doi:10.1007/s00122-007-0605-2

Clement D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003a) Mapping QTL for yield components, vigor, and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. Genome 46(2):204–212. doi:10.1139/g02-125

Clement D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003b) Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. Genome 46(1):103–111. doi:10.1139/g02-118

Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363(1491):557–572. doi:10.1098/rstb.2007.2170

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185

Feltus FA, Saski CA, Mockaitis K, Haiminen N, Parida L, Smith Z, Ford J, Staton ME, Ficklin SP, Blackmon BP, Schnell RJ, Kuhn DN, Motamayor JC (2011) Sequencing of a QTL-rich region of the *Theobroma cacao* genome using pooled BACs and the identification of trait specific candidate genes. BMC Genomics 12(1):379. doi:10.1186/1471-2164-12-379

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14 (7):1457–1467. doi:10.1105/tpc.010479

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8(3):195–202

Hospital F (2009) Challenges for effective marker-assisted selection in plants. Genetica 136(2):303–310. doi:10.1007/s10709-008-9307-1

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9(9):868–877

Irish BM, Goenaga R, Zhang DP, Schnell R, Brown JS, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station cacao (*Theobroma cacao* L.) germplasm collection. Crop Sci 50(2):656–667. doi:10.2135/cropsci2009.06.0299

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4):462–467

Kosambi DD (1944) The estimation of map distance from recombination values. Annals of Eugenics 12:172–175

Krutovsky KV, Elsik CG, Matvienko M, Kozik A, Neale DB (2007) Conserved ortholog sets in forest trees. Tree Genet Genom 3 (1):61–70. doi:10.1007/s11295-006-0052-2

Kuhn DN, Heath M, Wisser RJ, Meerow A, Brown JS, Lopes U, Schnell RJ (2003) Resistance gene homologues in *Theobroma cacao* as useful genetic markers. Theor Appl Genet 107(2):191–202

Kuhn DN, Borrone J, Meerow AW, Motamayor JC, Brown JS, Schnell RJ (2005) Single-strand conformation polymorphism analysis of candidate genes for reliable identification of alleles by capillary array electrophoresis. Electrophoresis 26(1):112–125

Kuhn DN, Narasimhan G, Nakamura K, Brown JS, Schnell RJ, Meerow AW (2006) Identification of cacao TIR-NBS-LRR resistance gene homologues and their use as genetic markers. J Amer Soc Hort Sci 131(6):806–813

Kuhn DN, Motamayor JC, Meerow AW, Borrone JW, Schnell RJ (2008) SSCP markers provide a useful alternative to microsatellites in genotyping and estimating genetic diversity in populations and germplasm collections of plant specialty crops. Electrophoresis 29(19):4096–4108. doi:10.1002/elps.200700937

Kuhn DN, Figueira A, Lopes U, Motamayor JC, Meerow AW, Cariaga K, Freeman B, Livingstone DS, Schnell RJ (2010) Evaluating *Theobroma grandiflorum* for comparative genomic studies with *Theobroma cacao*. Tree Genet Genom 6(5):783–792. doi:10.1007/s11295-010-0291-0

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8(12):2141–2143

Lanaud C, Fouet O, Clement D, Boccara M, Risterucci AM, Surujdeo-Maharaj S, Legavre T, Argout X (2009) A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. Mol Breeding 24 (4):361–374. doi:10.1007/s11032-009-9297-4

Lazo GR, Lui N, Gu YQ, Kong X, Coleman-Derr D, Anderson OD (2005) Hybsweeper: a resource for detecting high-density plate gridding coordinates. Biotechniques 39(3):320–322, 324

Lefebvre-Pautigny F, Wu FN, Philippot M, Rigoreau M, Priyono ZM, Frasse P, Bouzayen M, Broun P, Petiard V, Tanksley SD, Crouzillat D (2010) High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. Tree Genet Genom 6(4):565–577. doi:10.1007/s11295-010-0272-3

Li S, Chou HH (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics 20(16):2865–2866

Lima LS, Gramacho KP, Carels N, Novais R, Gaiotto FA, Lopes UV, Gesteira AS, Zaidan HA, Cascardo JCM, Pires JL, Micheli F (2009) Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao. Genet Mol Res 8(3):799–808. doi:10.4238/vol8-3gmr603

Livingstone D, Motamayor J, Schnell R, Cariaga K, Freeman B, Meerow A, Brown J, Kuhn D (2011) Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. Mol Breeding 27(1):93–106

Luo M, Wing R (2003) An improved method for plant BAC library construction. In: Grotewold E (ed) Methods in molecular biology: plant functional genomics: methods and protocols, vol 236. Human, Totowa, pp 3–20

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82(3):378–389

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G (2008a) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24 (24):2818–2824

Miller NA, Kingsmore SF, Farmer A, Langley RJ, Mudge J, Crow JA, Gonzalez AJ, Schilkey FD, Kim RJ, van Velkinburgh J, May GD, Black CF, Myers MK, Utsey JP, Frost NS, Sugarbaker DJ, Bueno R, Gullans SR, Baxter SM, Day SW, Retzel EF (2008b) Management of high-throughput DNA sequencing projects: Alpheus. J Comput Sci Syst Biol 1:132

Motamayor JC, Lachenaud P, Mota J, 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(10):8

Motilal L, Butler D (2003) Verification of identities in global cacao germplasm collections. Genet Resour Crop Evol 50(8):799–807

Motilal LA, Zhang DP, Umaharan P, Mischke S, Mooleedhar V, Meinhardt LW (2010) The relic Criollo cacao in Belize—

genetic diversity and relationship with Trinitario and other cacao clones held in the International Cocoa Genebank, Trinidad. Plant Genet Resour-Charact Util 8(2):106–115. doi:10.1017/s1479262109990232

Pugh T, Fouet O, Risterucci AM, Brottier P, Abouladze M, Deletrez C, Courtois B, Clement D, Larmande P, N'Goran JAK, Lanaud C (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. Theor Appl Genet 108(6):1151–1161. doi:10.1007/s00122-003-1533-4

Quraishi UM, Abrouk M, Bolot S, Pont C, Throude M, Guilhot N, Confolent C, Bortolini F, Praud S, Murigneux A, Charmet G, Salse J (2009) Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. Func & Integ Genom 9(4):473–484. doi:10.1007/s10142-009-0129-8

Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162(3):329–333

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101(5–6):948–955

Risterucci AM, Paulin D, Ducamp M, N'Goran JA, Lanaud C (2003) Identification of QTLs related to cocoa resistance to three species of *Phytophthora*. Theor Appl Genet 108(1):168–174. doi:10.1007/s00122-003-1408-8

Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel JF, Paterson AH (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176(4):2577–2588. doi:10.1534/genetics.107.074518

Rounsley S, Marri P, Yu Y, He R, Sisneros N, Goicoechea J, Lee S, Angelova A, Kudrna D, Luo M, Affourtit J, Desany B, Knight J, Niazi F, Egholm M, Wing R (2009) De novo next generation sequencing of plant genomes. Rice 2(1):35–43

Sambrook JFE, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor, Cold Spring Harbor

Saski CA, Feltus FA, Staton ME, Blackmon BP, Ficklin SP, Kuhn DN, Schnell RJ, Shapiro H, Motamayor JC (2011) A genetically anchored physical framework for *Theobroma cacao* cv. Matina 1-6. BMC Genomics 12(1):413. doi:10.1186/1471-2164-12-413

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. Phytopathol 97(12):1664–1669

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman TL, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW Jr, Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Lopez Girona E, Zdepski A, Wang W, Kerstetter RA, Schwab W, Korban SS, Davik J, Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43(2):109–116. doi:10.1038/ng.740

Smit A, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0. Available at http://www.repeatmasker.org

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 10 (11):1772–1787

Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38(3):375–381

Van Ooijen JW (2006) JoinMap 4, software for the calculation of genetic linkage maps in experimental populations, 4th edn. Kyazma B. V, Wageningen

Wu FN, Mueller LA, Crouzillat D, Petiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174(3):1407–1420. doi:10.1534/genetics.106.062455

Wu FN, Eannetta NT, Xu YM, Durrett R, Mazourek M, Jahn MM, Tanksley SD (2009a) A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus *Capsicum*. Theor Appl Genet 118(7):1279–1293. doi:10.1007/s00122-009-0980-y

Wu FN, Eannetta NT, Xu YM, Tanksley SD (2009b) A detailed synteny map of the eggplant genome based on conserved ortholog set II (COSII) markers. Theor Appl Genet 118(5):927–935. doi:10.1007/s00122-008-0950-9

Zhang DP, Boccara M, Motilal L, Mischke S, Johnson ES, Butler DR, Bailey B, Meinhardt L (2009) Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from Upper Amazon using microsatellite DNA markers. Tree Genet Genom 5 (4):595–607. doi:10.1007/s11295-009-0212-2