

BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution

Alexis Dereeper · Romain Guyot · Christine Tranchant-Dubreuil · François Anthony · Xavier Argout · Fabien de Bellis · Marie-Christine Combes · Frederick Gavory · Alexandre de Kochko · Dave Kudrna · Thierry Leroy · Julie Poulain · Myriam Rondeau · Xiang Song · Rod Wing · Philippe Lashermes

Received: 21 December 2012 / Accepted: 14 May 2013 / Published online: 25 May 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Coffee is one of the world's most important agricultural commodities. Coffee belongs to the Rubiaceae family in the euasterid I clade of dicotyledonous plants, to which the Solanaceae family also belongs. Two bacterial artificial chromosome (BAC) libraries of a homozygous doubled haploid plant of *Coffea canephora* were constructed using two enzymes, *Hind*III and *Bst*YI. A total of 134,827 high quality BAC-end sequences (BESs) were generated from the 73,728 clones of the two libraries, and 131,412 BESs were conserved for further analysis after elimination of chloroplast and mitochondrial sequences.

Electronic supplementary material The online version of this article (doi:10.1007/s11103-013-0077-5) contains supplementary material, which is available to authorized users.

A. Dereeper · F. Anthony · M.-C. Combes · M. Rondeau · P. Lashermes (✉)
Institut de Recherche pour le Développement (IRD), UMR RPB (CIRAD, IRD, UM2), BP 64501, 34394 Montpellier Cedex 5, France
e-mail: Philippe.Lashermes@ird.fr

R. Guyot · C. Tranchant-Dubreuil · A. de Kochko
Institut de Recherche pour le Développement (IRD), UMR DIADE (CIRAD, IRD, UM2), BP 64501, 34394 Montpellier Cedex 5, France

X. Argout · F. de Bellis · T. Leroy
Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR AGAP, Avenue d'Agropolis, 34398 Montpellier, France

F. Gavory · J. Poulain
Commissariat à l'Énergie Atomique (CEA), Institut de Génétique, BP 5706, 91057 Evry, France

D. Kudrna · X. Song · R. Wing
Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

This corresponded to almost 13 % of the estimated size of the *C. canephora* genome. 6.7 % of BESs contained simple sequence repeats, the most abundant (47.8 %) being mononucleotide motifs. These sequences allow the development of numerous useful marker sites. Potential transposable elements (TEs) represented 11.9 % of the full length BESs. A difference was observed between the *Bst*YI and *Hind*III libraries (14.9 vs. 8.8 %). Analysis of BESs against known coding sequences of TEs indicated that 11.9 % of the genome corresponded to known repeat sequences, like for other flowering plants. The number of genes in the coffee genome was estimated at 41,973 which is probably overestimated. Comparative genome mapping revealed that microsynteny was higher between coffee and grapevine than between coffee and tomato or *Arabidopsis*. BESs constitute valuable resources for the first genome wide survey of coffee and provide new insights into the composition and evolution of the coffee genome.

Keywords Comparative genomics · *Coffea* · Genome · BAC library · Transposable elements · Microsatellites

Introduction

Coffee is one of the world's major crops, and is mainly cultivated in Africa, America and Asia. Smallholders with <5 ha account for about 70 % of world coffee production and more than 80 million people depend on the crop for their income. Coffee belongs to the 4th largest flowering plant family, Rubiaceae, which consists of more than 11,000 species in 660 genera (Robbrecht and Manen 2006). The Rubiaceae family belongs to the euasterid I clade of dicotyledonous plants, to which the Solanaceae family also belongs.

Although the *Coffea* genus includes more than 124 species (Davis et al. 2011), commercial coffee production relies mainly on two related species: *Coffea arabica* L. and *Coffea canephora* Pierre ex A. Froehner, which account for 65 and 35 % of world coffee production, respectively (International Coffee Organization, <http://www.ico.org>). *C. arabica* is a recent allotetraploid ($2n = 4x = 44$) species which derived from a spontaneous hybridization between two closely related diploid species, *C. eugenioides* S.Moore and *C. canephora*, whereas *C. canephora* ($2n = 2x = 22$) is an allogamous diploid tropical tree consisting of polymorphic populations and of strongly heterozygous individuals. However, producing doubled haploid plants using haploid embryos, which occur spontaneously in association with polyembryony, offers the possibility of developing completely homozygous genotypes from heterozygous parents in a single generation (Couturon and Berthaud 1982; Lashermes et al. 1994). The haploid genome of *C. canephora* is estimated to be 710 Mb in size (Noirot et al. 2003), over five times larger than that of the model plant *Arabidopsis thaliana* (125 Mb) and more than 50 % larger than that of grapevine (*Vitis vinifera*) or poplar (*Populus trichocarpa*). Both macro- and microsynteny have been reported between coffee trees (Rubiaceae) and tomato (Solanaceae), and between coffee trees and species belonging to the Rosid clade (Cenci et al. 2010, 2012; Guyot et al. 2009, 2012; Lefebvre-Pautigny et al. 2010; Mahé et al. 2007). In particular, a high level of conservation was observed between coffee and grapevine (Vitaceae). Recent phylogenetic analyses identified the Vitaceae family as the earliest diverging lineage of the Rosid clade, meaning this family can be considered as the “sister” group of all other Rosid plant species (Jansen et al. 2006).

The ability to capture and efficiently use the abundant genetic resources in coffee breeding programmes is essential for sustainable coffee production. Until recently, coffee improvement mainly relied on conventional breeding methods. Moreover, coffee is a perennial plant with a time from seed to seed of about 5 years, which makes genetic studies difficult and time consuming. While some genomic information has been generated in the last few years, it is far less than what is available for many other agricultural species. Significant advances in our understanding of the coffee genome and its biology must be achieved in the coming decades to increase coffee quality and yield, and to protect the crop from major losses caused by insect pests, diseases and abiotic stresses related to climatic change.

The availability of a large-insert genomic DNA library is indispensable for physical mapping, map-based gene cloning, and analysis of gene structure and function in most organisms including plants. Large insert genomic libraries

constructed with bacterial artificial chromosomes (BACs) are known for their high degree of genomic insert structural stability and easy handling of *Escherichia coli* host cells. The construction of a high quality BAC library, together with sequencing of BAC-ends (BESs, BAC-end sequences) refer to bidirectional end sequencing of the genomic DNA insert (with the help of universal priming sites in the cloning vector) are also two of the first steps in a whole-genome sequencing project. Although the sequences sampled may be not truly random, due to the need for specific restriction sites for the construction of a BAC library, a BES project can provide significant clues about the composition of the genome and the evolution of a given species.

The use of BESs from a large number of BACs was first proposed as a strategy for identifying overlapping clones during whole genome sequencing since BESs greatly facilitate the assembly of contigs into scaffolds (Goff et al. 2002). In spite of continued improvements in sequencing technologies and the development of whole-genome shotgun (WGS) sequencing approach for plant genomes, the BESs remain extremely helpful during the genome sequence assembly process, in particular for repeat-rich regions or genomes (Feuillet et al. 2011). Furthermore, BESs also represent a sample of the whole genome which can be used to get a first glimpse into the sequence composition of a given genome. BES analyses have been performed in a number of plants in the initial stage of genome characterisation including rice, maize, Korean ginseng, papaya, *Brassica rapa*, wheat 3B, *Musa acuminata*, white clover, *Brachypodium*, potato, tomato, citrus, apple and carrot (Cavagnaro et al. 2009; Cheung and Town 2007; Datema et al. 2008; Febrer et al. 2007; Han and Korban 2008; Hong et al. 2004, 2006; Huo et al. 2008; Lai et al. 2006; Mao et al. 2000; Messing et al. 2004; Paux et al. 2006; Terol et al. 2008). BESs are also a rich sources of genomic simple sequence repeats (SSRs) which act as reliable landmarks across the genome during genetic mapping, as reported in plant genomes such as cotton, soybean and *Brassica napus* (Cheng et al. 2009; Frelichowski et al. 2006; Shultz et al. 2007).

Here, we report the generation of coffee BESs from two BAC libraries from a homozygous doubled haploid plant. They represent approximately 13 % of the *C. canephora* nuclear genome, and provide initial insights into the content and composition of the coffee genome. Our analysis focused on microsatellite content, repeat element composition, protein-coding regions and comparative mapping of BES pairs to other sequenced plant genomes. This study provides the first glimpse of the genome of the DH200-94 accession, which is the accession chosen for the whole-genome sequencing of *C. canephora*.

Materials and methods

BAC libraries

Bacterial artificial chromosome libraries of the *C. canephora* accession DH200-94 were constructed from high molecular weight DNA isolated from 20 g of young leaf tissue by the Arizona Genome Institute (AGI, USA). The accession DH200-94 is a doubled haploid plant produced from the clone IF200 based on the haploid plants occurring spontaneously in association with polyembryony. Two restriction enzymes, *Hind*III and *Bst*YI, were used for partial digestion of mega-size DNA, cloned into the pAGIBAC1 vector (Jetty et al. 2006) and transformed in the *E. coli* strain DH10B (Invitrogen, United States). A total of 36,864 clones were picked for each library. In order to evaluate the average BAC insert size, 768 BAC clones (384 clones from each library) were randomly chosen and the corresponding DNA was extracted, digested with the rare cutter *Not*I enzyme and analysed by PFGE.

BAC-end sequencing

Bidirectional-end sequencing of the 73,728 BAC clones was performed using ABI dye-terminator chemistry at Genoscope (Evry, France) on ABI3730 sequencers (Applied Biosystems, Foster City, United States). Sequence and quality files from trace files were read by the PHRED software for base calling and quality trimming using a quality score of 20. Cloning vector sequences were removed using an internal procedure based on the Smith-Waterman algorithm. The sequence data were then filtered for sequences contaminated by *E. coli* and by plant organelle genomes based on matches with the mitochondria genome sequence of *A. thaliana* (NC001284) and *C. arabica* chloroplast (NC008535) sequences. BESs shorter than 60 bp were also removed. Sequences were deposited in EMBL-EBI Bank (accession numbers FO535330, FO538768 to FO624989, and FO624992 to FO680656).

Identification of simple sequence repeats (SSR)

A pipeline software previously developed for SSR mining (Poncet et al. 2006) was used for the identification of BESs containing microsatellites. A total of 69,066 ESTs from *C. canephora* (available at NCBI and downloaded in May 2012) were assembled into 26,483 unigenes (16,870,301 bp) using the TGICL program (Pertea et al. 2003). A similar SSR analysis was performed on these unigenes. The SSR program was also used to design the primer pairs and to check their high specificity. The parameters were set for detection of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 15, 9, 6, 5, 4, and 3 repeats, respectively.

The following primer design parameters were used: primer length from 18 to 21 bp (optimum 20 bp), PCR product size from 100 to 300 bp, optimum annealing temperature 60 °C.

Analysis of known repetitive sequences

To tentatively identify repeated sequences based on sequence similarities, we performed BLASTn (Altschul et al. 1997) and Censor (Kohany et al. 2006) analyses on the 131,412 BESs from the *C. canephora* *Bst*YI and *Hind*III libraries. We first used a BLASTn with E-value significance thresholds of $1e^{-10}$ and a minimum of 50 aligned base pairs against the Plant Repeats database (<http://plantrepeats.plantbiology.msu.edu/>) for similarity searches (Jan. 2011). This reference database contains 6,715 sequences from Brassicaceae, Fabaceae, Gramineae and Solanaceae families. According to the results obtained and based on the lack of detection sensitivity, we used Censor and the transposable element (TE) protein database containing 10,307 reference amino acids downloaded (<http://www.girinst.org/>) from Repbase (Version 17.01, Jun. 2012; 8,416 and 1,891 sequences from Class I and Class II TEs respectively) (Jurka et al. 2005) to detect similarities. We applied a minimum detected fragment length of 100 residues to be reported by Censor.

Bacterial artificial chromosome-end sequences from seven different plant species were retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/>) or from the SOL genomic web site (<http://solgenomics.net/>): *A. thaliana* (46,193); *Oryza sativa* (124,625); *V. vinifera* (229,315), *Solanum tuberosum* (140,540), *Solanum lycopersicum* (76,975), *Mimulus guttatus* (*Bst*YI library: 4,721; *Hind*III library: 3,398) and used for a similar analysis. Finally, we constructed our *C. canephora* TEs database by extracting identified sequences and by assembling them into contigs using a sequence assembly program (CAP3, Huang and Madan 1999) with default parameters.

Analysis of coding regions

For these analyses, we used both assembled and unassembled TE masked BESs, named BESnr and BESr, respectively. These were assembled using TGICL (Pertea et al. 2003) with default parameters. Similarity searches were performed using BLAST against *Coffea* ESTs and protein databases.

For functional analysis, BESnr were compared with the UniProtKB (SwissProt + TrEMBL) protein database with an E-value cutoff of $1e^{-20}$. Annotations from the two databases were merged into a single annotation file as follows: for each BES, priority was given to the SwissProt annotation, which is more reliable. When no match was found in the SwissProt database, the TrEMBL annotation

was then searched. If no match was found in either database, the BES was discarded. Gene ontology (GO) terms were assigned to the BESs present in this annotation file. Categories for the annotations were determined and visualized using BLAST2GO (Conesa and Götze 2008).

To determine the coding region, we used the BESr dataset in order to avoid biases and artifacts in estimating the coding fraction due to underestimation of the number of multigenic families. We performed a BLASTn with an E-value cutoff of $1e^{-6}$ using a *Coffea* spp. assembled unigene database, which consists of 70 % *C. arabica* ESTs and 30 % *C. canephora* ESTs (Vidal et al. 2010). The total length of matching sequences was calculated by summing the length of each match. When several HSPs (high segment pairs) were identified for a single query, thus revealing the presence of introns, the total length (HSPs + introns) was identified as the transcribed coding region.

Comparative genome mapping

Unassembled high-quality sequences not containing known repetitive sequences were used to detect potential regions of microsynteny between the coffee genome and the available complete genome sequences of *A. thaliana* (Arabidopsis Genome Initiative, 2000), *S. lycopersicum* (Tomato Genome Consortium, 2012) and *V. vinifera* (Jaillon et al. 2007). The analysis was limited to BES pairs for which both forward and reverse sequences showed similarities with coding regions. These paired BES were aligned against the genomes using tBLASTx, with $1e^{-6}$ E-value as cut off. For each sequence, only the highest scores were retained for subsequent analysis. BAC clones were used for microsynteny analysis if both ends had a high score in the target model genome. All BES pairs which produced significant hits with *V. vinifera* and *S. lycopersicum*, co-localized (i.e. within a 10–300 kb interval) or not, were finally mapped on the chromosomes of these species using a dedicated in-house Perl script. Ratios of the number of co-localized BES pairs to the number of double hits were calculated on a sliding window and mapped along the *V. vinifera* and *S. lycopersicum* chromosomes with a window length of 1 and 2 Mb, and a step size of 250 and 500 kb, respectively.

Results and discussion

Characterisation of the BAC library

With the aim of developing essential genomic resources for the genus *Coffea*, we constructed two high quality genomic BAC libraries with high molecular weight DNA from

C. canephora. We chose a double haploid tree (acc. DH200-94) to reduce the complexity of a highly heterozygous plant and consequently increase the effective coverage of the genome. To limit cloning biases associated with a single restriction enzyme-based BAC library, two different enzymes, *Hind*III and *Bst*YI, were used. Each of the two libraries contained 36,864 BAC clones that were arrayed in 96 384-well microtiter plates. The *Hind*III library, named CC_Ba, and the *Bst*YI library, named CC_Bb, contained estimated average insert sizes of 166 and 121 kb, respectively (Fig. S1, supplemental data). Since the haploid genome size of *C. canephora* is in the order of 710 Mb (Noirot et al. 2003), genome coverages of $\sim 8.6X$ and $\sim 6.3X$ were estimated for the *Hind*III and *Bst*YI libraries, respectively. The characteristics of the two high quality BAC libraries are summarised in Table S1 (supplemental data), and both resources are publicly available from the Arizona Genomics Institute Resource Center (<http://www.genome.arizona.edu/orders/>).

BAC-end sequencing

A total of 134,827 high quality BESs were generated from 73,728 BAC clones of the two *Hind*III and *Bst*YI libraries, with a success rate of 91.4 %. All generated sequences exceeded 60 bases in length with a PHRED-quality score ≥ 20 . Sequences were compared to *C. arabica* chloroplast genome and *A. thaliana* mitochondrial genome sequences. Results indicated that a total of 1,389 (1.03 %) and 28 (0.02 %) reads showed high similarity with chloroplast or mitochondrial genomes, respectively. In addition, 1,998 BESs (1.48 %) appeared to be contaminated by cloning vector sequences. The two libraries contributed a similar number of reads and genomic raw sequences. Finally, 131,412 BESs with an average length of 682 bp were retained for further analysis. This represents ~ 92 Mb of genomic sequence data, corresponding to almost 13 % of the estimated size of the *C. canephora* genome. These data are an excellent resource to provide a first view of the composition, structure and evolution of the coffee genome.

Analysis of simple sequence repeats (SSR)

Simple sequence repeats (or microsatellites) are a class of molecular markers which are often polymorphic and are widely used to produce genetic maps. In addition to ESTs, BESs have also proved to be very useful to design SSR markers in plants (Cheng et al. 2009; Frelichowski et al. 2006; Shultz et al. 2007). Depending on the search parameters used (see “Materials and methods” section), 9,875 SSRs were detected in 6.7 % of the obtained BESs. Mononucleotide motifs were the most abundant (47.8 %), followed by di- (22.9 %), and hexa-nucleotide repeats

Table 1 Distribution of simple sequence repeats in *C. canephora* BESs as a function of the length of the repeat motif

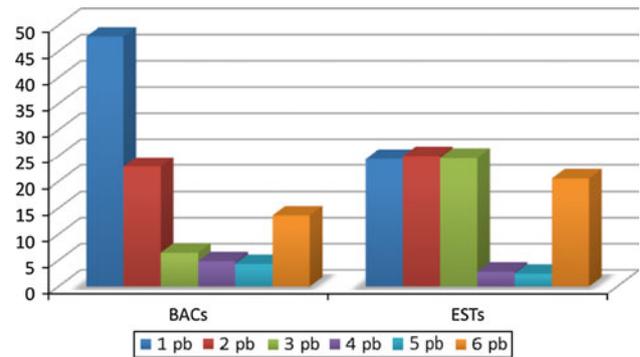
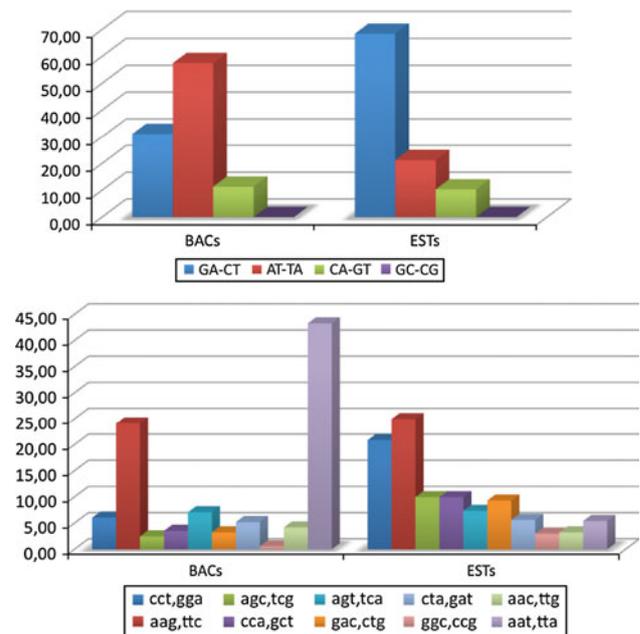
Motif	SSR number	% SSR
Mononucleotide	4,731	47.8
Dinucleotide	2,273	22.9
Trinucleotide	639	6.5
Tetranucleotide	467	4.9
Pentanucleotide	428	4.3
Hexanucleotide	1,337	13.6
Total SSRs	9,875	100.0

Table 2 Distribution of simple sequence repeats in *C. canephora* BESs as a function of the motif of di- and tri-nucleotide repeats

SSR motif	SSR number	% SSR
Dinucleotide	2,273	100
GA-CT	709	31.2
AT-TA	1,304	57.5
CA-GT	260	11.3
GC-CG	0	0
Trinucleotide	639	100
CCT-GGA	39	6
AAG-TTC	153	24.2
AGC-TCG	16	2.5
CCA-GGT	23	3.6
AGT-TCA	45	7.1
GAC-CTG	21	3.3
CTA-GAT	33	5.2
GGC-CCG	4	0.6
AAC-TTG	27	4.2
AAT-TTA	278	43.3

(13.5 %) (Table 1). Among the dinucleotide motifs, AT/TA was the most abundant (57.5 %) and no GC/CG motif was found. The most abundant trinucleotide motif was AAT/TTA (43.4 %), followed by AAG/TTC (24.2 %) and GGC/CCG was the least abundant motif (0.6 %) (Table 2).

The distribution of microsatellite motif frequencies differed between ESTs and BESs whereas their frequency was identical in BESs and ESTs (0.1 SSR per kb). All the EST sequences available at NCBI for *C. canephora* were downloaded and screened for SSRs using the same program and the same parameters. A total of 1,865 SSRs were identified. Di-, tri- and mono-nucleotide motives were the most abundant (respectively 25, 24.7 and 24.5 %) (Fig. 1). The motif frequency differed with the type of sequence (Fig. 2). Among the dinucleotide motifs, the most common repeat from the ESTs was the GA/CT motif (68.4 %) while for the trinucleotide motives, the AAG/TTC motif was the most abundant (25 %).

**Fig. 1** Distribution of simple sequence repeats in *C. canephora* BESs and ESTs as a function of the length of the repeat motif**Fig. 2** Distribution of simple sequence repeats in *C. canephora* BESs and ESTs as a function of the di and trinucleotide motifs

Our results indicated that potential SSRs are more numerous in BESs (which group non-coding and coding regions) than coding regions only as detected in ESTs. All the novel SSRs detected in this analysis are currently being used to develop and refine a dense genetic map of *C. canephora*. Because it has been proven that many microsatellite markers defined in *C. canephora* can be used in other coffee species (Poncet et al. 2007), the markers reported here will help advance comparative genomic studies in the *Coffea* genus.

Identification and characterisation of repetitive DNA elements

The 131,412 BESs were first screened for repetitive DNA sequences by BLASTn searches of the Plant Repetitive

Table 3 Number of TE sequences, length (bp) of TE sequences and number of BESs containing repeated DNA using Censor and the Repbase transposable element protein database (minimum length of detected fragment: 100 residues)

Repeated DNA Type	Number of TE in <i>Bsr</i> YI BES	Length of TE in <i>Bsr</i> YI BES (bp)	% of TE type in <i>Bsr</i> YI BES	Number of TE in <i>Hind</i> III BES	Length of TE in <i>Hind</i> III BES (bp)	% of TE type in <i>Hind</i> III BES	Number of TE in <i>Bsr</i> YI and <i>Hind</i> III BES	Length of TE in <i>Bsr</i> YI and <i>Hind</i> III BES (bp)	% of TE type in <i>Bsr</i> YI and <i>Hind</i> III
Class I	11,446	6,371,316	93.74	6,388	3,538,319	89.79	17,834	9,909,635	92.28
Class I ?	3	981	0.02	5	1,738	0.07	8	2,719	0.04
Class I DIRS	0	0	0	0	0	0	0	0	0
Class I LINE	1,207	648,611	9.88	1,219	676,606	17.13	2,426	1,325,217	12.55
Class I LINE ?	0	0	0	0	0	0	0	0	0
Class I LINE I	1	653	0	0	0	0	1	653	0.0052
Class I LINE Joekey	0	0	0	0	0	0	0	0	0
Class I LINE L1	1,205	647,587	9.86	1,217	675,561	17.10	2,422	1,323,148	12.53
Class I LINE R2	0	0	0	1	662	0.01	1	662	0.0052
Class I LINE RTE	1	371	0	1	383	0.01	2	754	0.0103
Class I LTR	10,236	5,721,724	83.83	5,164	2,859,975	72.58	15,400	8,581,699	79.69
Class I LTR ?	109	43,858	0.89	138	69,341	1.93	247	113,199	1.27
Class I LTR Bel-Pao	2	1,051	0.01	7	3,683	0.09	9	4,734	0.04
Class I LTR Copia	2,422	1,330,829	19.83	1,349	754,094	18.96	3,771	2,084,923	19.51
Class I LTR ERV	0	0	0	0	0	0	0	0	0
Class I LTR Gypsy	7,688	4,337,527	62.96	3,661	2,027,345	51.46	11,349	6,364,872	58.73
Class I LTR Retrovirus	15	8,459	0.12	9	5,512	0.12	24	13,971	0.12
Class I PLE	0	0	0	0	0	0	0	0	0
Class II	.	764	404,491	6.25	726	377,599	10,20	1,490	782,090
Class II ?	.	0	0	0	0	0	0	0	0
Class II Crypton	.	217	116,810	1.77	139	72,609	1.95	356	189,419
Class II Helitron	.	312	166,113	2.55	351	191,118	4.93	663	357,231
Class II Maverick	.	0	0	0	0	0	0	0	0
Class II TIR	.	235	121,568	1.92	236	113,872	3.31	471	235,440
Class II TIR CACTA	.	54	27,165	0.44	70	34,840	0.98	124	62,005
Class II TIR hAT	.	149	81,564	1.22	143	69,684	2.01	292	151,248
Class II TIR Merlin	.	0	0	0	0	0	0	0	0
Class II TIR P	.	1	398	0	0	0	0	1	398
Class II TIR PIF-Harbinger	.	24	9,136	0.19	20	7,685	0.28	44	16,821
Class II TIR PiggyBac	.	0	0	0	0	0	0	0	0
Class II TIR Tc1-Mariner	.	7	3,305	0.05	3	1,663	0.04	10	4,968
Class II TIR Transib	.	0	0	0	0	0	0	0	0
Total		12,210	6,775,807	100 %	7,114	3,915,918	100 %	19,324	10,691,725
TE length/ total BES length		14.96 %			8.81 %			11.91 %	

database in a way similar to the method published by Hsu et al. (2011). A relatively low number of repeated sequences were detected using this method with the exception of rRNA, which represented the majority of the repeats found (respectively 82 and 53 % of all repeated sequences in the *Bst*YI and *Hind*III genomic libraries) (Fig. S2, supplemental data). Including the detection of rRNA, only 5.6 and 2 % of BESs in the *Bst*YI and *Hind*III libraries were found to contain repeated sequences. This result indicates that the method used lacks sensitivity and suggests that most of TEs in *C. canephora* may diverge significantly at the nucleotide level from data in the reference database.

To overcome the observed lack of sensitivity, we used the RepBase database for TEs, which contains 10,307 reference sequences, and Censor (Kohany et al. 2006).

Like in other plant genomes, we found that the *C. canephora* genome contains a significant proportion of TEs (Table 3). In total, 17,834 fragments representing 11.91 % of the whole length of the BESs contained similarities with known TE peptides (Table 3). Interestingly, we observed a bias depending on whether the *Bst*YI or *Hind*III library was used for the analysis (14.96 vs. 8.81 %). Long terminal repeat (LTR)-retrotransposons (both *Ty1-Copia* and *Ty3-Gypsy* elements) were over-represented in the *Bst*YI library compared to the *Hind*III library (Table 3). Equal amounts of the other type of TEs (LINE and Class II DNA transposons) were identified in the two libraries. We compared the results with results of a similar analysis conducted on publicly available BESs from *M. guttatus* *Bst*YI and *Hind*III libraries (respectively 4,721 and 3,398 sequences). A bias in the representation of LTR-retrotransposons was also observed between the *Bst*YI and *Hind*III libraries (Fig. S3 and S4, supplemental data). We concluded that the bias we observed between *Bst*YI and *Hind*III libraries could be due to the restriction enzyme sites used to construct the two libraries. These results underlined the need to construct BAC libraries based on different restriction sites to overcome cloning artifacts due to the use of a single enzyme.

We also compared our results with a selection of other plant species. BESs from seven different reference plant genomes were retrieved and analysed in the same way as for *C. canephora* BESs (Fig. 3).

Among the 17,834 fragments containing potential TEs, a large majority (92.2 %) showed homology with Class I retrotransposons (Table 3). This class is subdivided into *Ty1-Copia* and *Ty3-Gypsy* LTR retrotransposon superfamilies. In *C. canephora* *Ty3-Gypsy* clearly outnumbered *Ty1-Copia* LTR retrotransposons with a *Ty3-Gypsy*:*Ty1-Copia* ratio of 3:1.

This ratio is similar to those of *O. sativa* (2.95:1), *Ara-bidopsis* (2.94:1), *S. tuberosum* (2.48:1) and *S. lycopersicum*

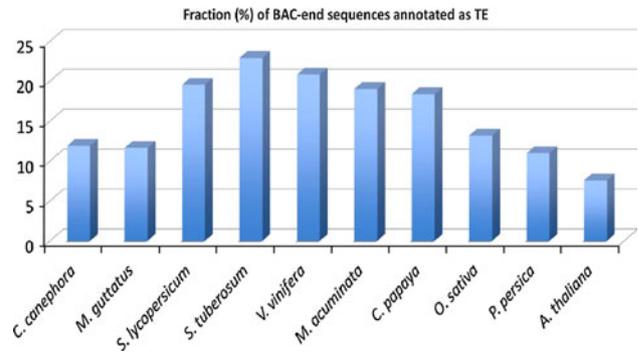


Fig. 3 Comparison between the ratios of sequence lengths similar to known TEs found in *C. canephora* BES sequences and a selection of BESs from model plant species (i.e. % total sequence lengths similar to TE/whole length of BESs). TEs were detected using Censor and proteins from REPBASE as the reference library

(2.45:1) but higher than or equal to those of *Carica papaya* (1.61:1) and *M. guttatus* (1:1); and *Prunus persica* (1:1.18) *Musa acuminata* (1:2.47) and *V. vinifera* (1:3) where *Ty1-Copia* outnumbered *Ty3-Gypsy* TEs (Fig. 4). The second most abundant TE type found in *C. canephora* was the non-LTR Retrotransposons LINE element (L1 family) (Table 3).

Taken together, our analyses of BESs against known coding sequences of TEs indicate that 11.9 % of the genome corresponds to known repeat sequences. Thus the percentage of TE in *C. canephora* appears to be similar to the percentage found for *M. guttatus* (11.6 %) and *P. persica* (10.9 %).

However, considering that in plant genomes many elements are non-coding elements (with a lack of coding sequences like in MITEs), that a large proportion of TE domains are non-coding (like the LTR domains of retrotransposons) and that the coding sequences of elements inserted a long time ago are highly altered, the proportion of TE may be significantly higher than our current estimation.

We conclude that the *C. canephora* genome present a significant proportion of TE similarly to *M. guttatus* and *P. persica* genomes. The most frequent elements were class I retrotransposons *Ty3-Gypsy* group. Considering the relative importance of LTR-Retrotransposons, the transcriptional and transpositional activities of these elements should be analysed and used to study the genetic diversity of the *C. canephora* species.

Finally, all identified TE sequences were extracted from BESs and assembled into 2,475 contigs and 4,951 singlets (for a total size of 4,610 Mb) to construct the first database of *C. canephora* TE sequences (Online resource, Table S2). Among the assembled contigs, 615 were $\geq 1,000$ bp and 61 $\geq 2,000$ bp. The longest assembled sequence was 4,704 bp showing similarities with *Ty3-Gypsy* proteins from the public database when translated. This database can be used for masking procedures in the assembly of

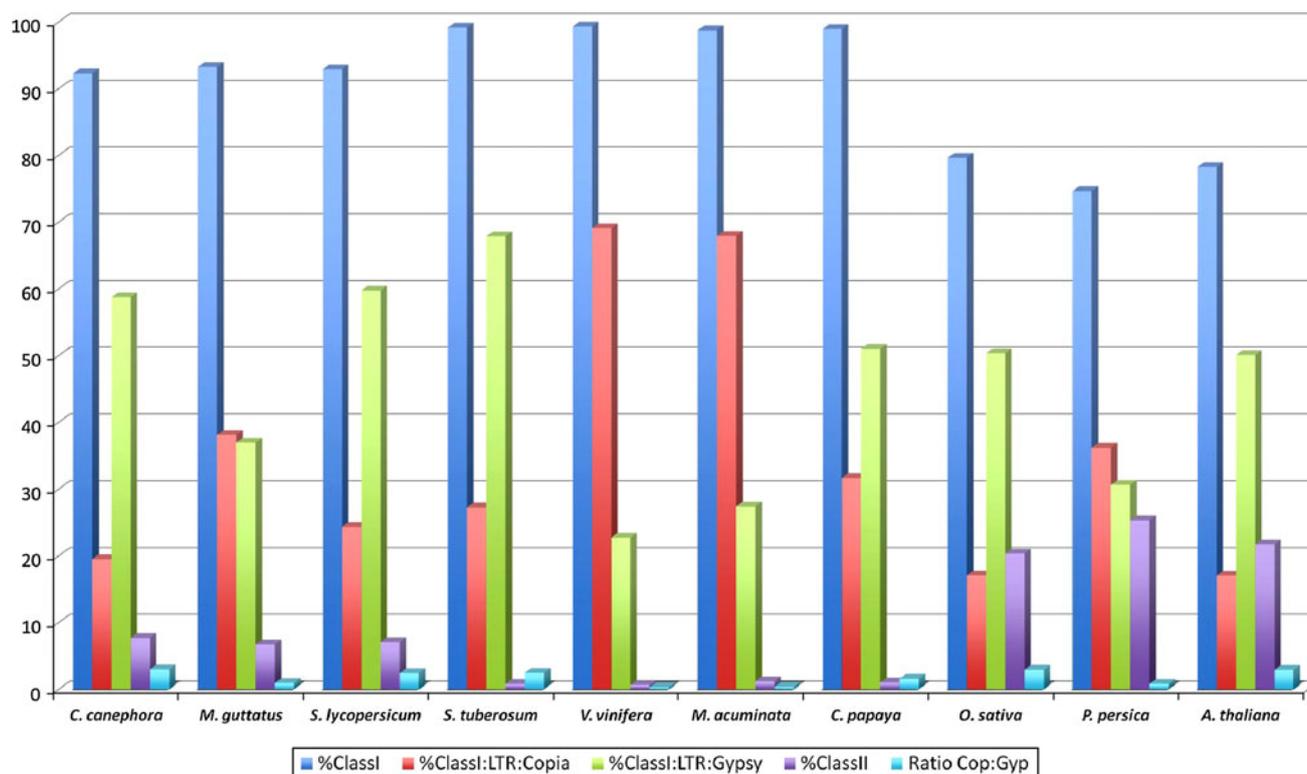


Fig. 4 Frequency of TE classes (Class I retrotransposons, *Ty1-Copia*, *Ty3-Gypsy* and Class II DNA transposons) in the BESs of *C. canephora* *Bst*YI and *Hind*III libraries and a selection of reference plant species

C. canephora genomic sequences and detailed TE characterisations.

Analysis of coding regions

BAC-end sequences were analyzed to identify coding region via homology searches. Overall, 54,501 BESs (41.5 % of BESr) displayed significant homology with *Coffea* spp. ESTs represented approximately 18 Mb of

transcribed sequences and accounted for 20.1 % of the cumulative length of BESr (Table 4). At the *C. canephora* genome scale, this represents a transcribed portion of 142.7 Mb in length (20.1 % of 710 Mb). Assuming an average gene length (3.4 kb) similar to that of *V. vinifera* (Jaillon et al. 2007) and a total coding region of 142.7 Mb, the gene contents of the *C. canephora* genome was assessed as 41,973. This estimate is likely to be biased due to the generation process of BESs or by the fact that the EST

Table 4 Summary of BLAST analyses using different BES sets against SwissProt and TrEMBL protein databases (E-value cutoff: $1e^{-20}$) and Coffee EST databases (E-value cutoff: $1e^{-6}$)

	Input		SwissProt		TrEMBL		EST	
	# seq	Cumulative length (bp)	Cumulative length (bp)	# hit	Cumulative length (bp)	# hit	Cumulative length (bp)	# hit
BESraw	137470	93491385 (13.2 %) ^a	2788658 (3.0 %) ^b	5409	5550846 (5.9 %) ^b	12047	19006885 (20.3 %) ^b	57374 (41.7 %) ^c
<i>Bst</i> YI	68542	46514577 (6.5 %) ^a	1036838 (2.2 %) ^b	2031	2501317 (5.3 %) ^b	5822	8142696 (17.5 %) ^b	27260 (39.8 %) ^c
<i>Hind</i> III	68928	46976808 (6.6 %) ^a	1751820 (3.7 %) ^b	3378	3049529 (6.5 %) ^b	6225	10864189 (23.1 %) ^b	30114 (43.7 %) ^c
BESr	131412	89686101 (12.6 %) ^a	2306934 (2.6 %) ^b	4437	4950940 (5.5 %) ^b	10768	17987533 (20.1 %) ^b	54501 (41.5 %) ^c
BESnr	89950	66996986 (9.4 %) ^a	1911377 (2.9 %) ^b	3416	3372098 (5.0 %) ^b	6497	10864189 (16.2 %) ^b	33220 (36.9 %) ^c

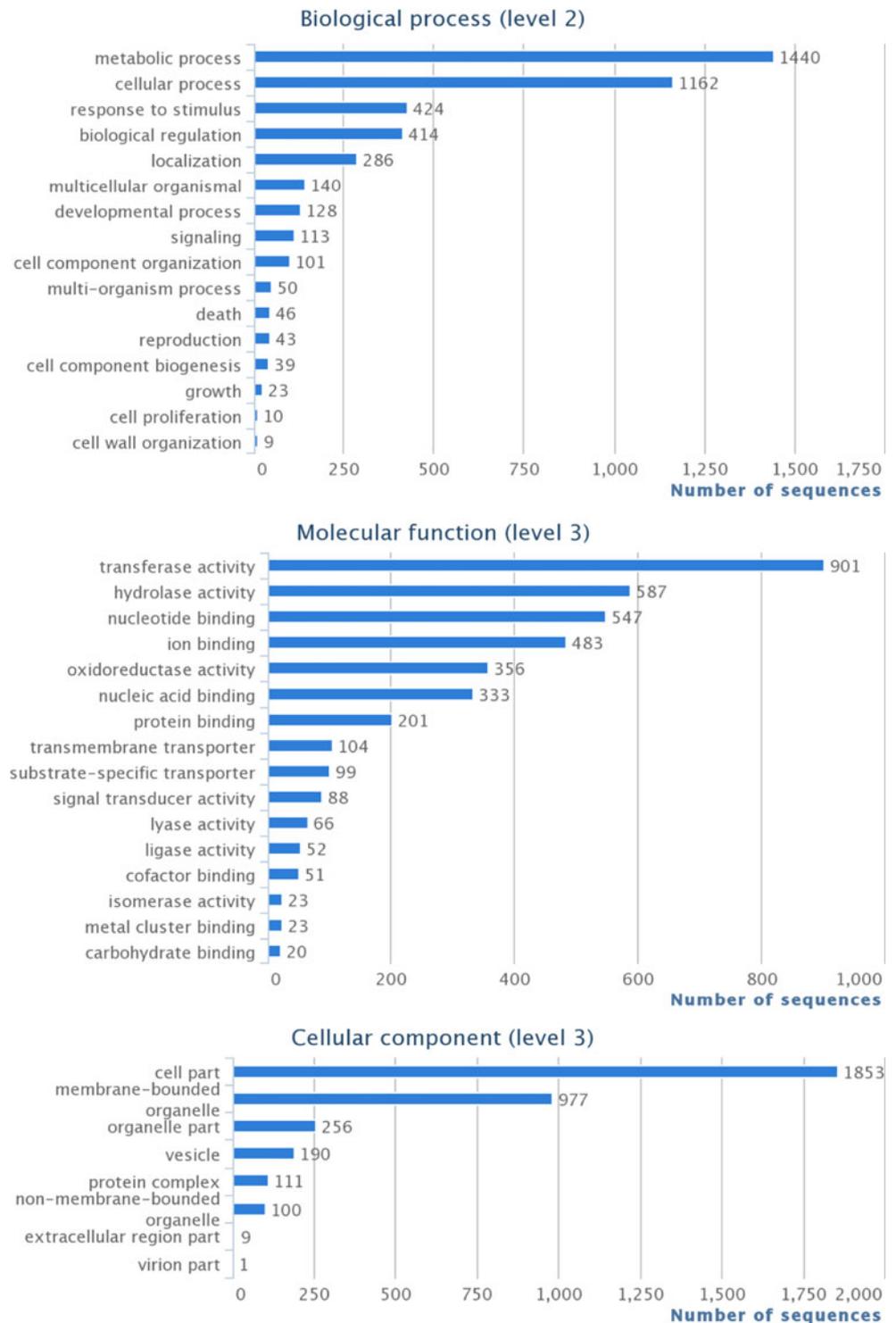
Line 2 and 3 shows BESraw results split up into BAC libraries: *Bst*YI and *Hind*III

^a Percentage compared to the genome size reference, 710 Mb

^b Percentage compared to the input cumulated length

^c Percentage compared to the initial number of BES

Fig. 5 Distribution of GO annotations of gene products predicted from the *C. canephora* BESs for biological process, molecular function and cellular component. Only major categories are presented



database used does not represent the complete transcriptome of *C. canephora*. Taking this probable overestimation into account, this estimation is in the same range as the number of protein-coding genes identified in the genome sequence of related plant species (i.e. 30,434; 34,727 and 35,004 in grapevine, tomato and potato, respectively).

Additionally, the BESraw based analysis was split up into two separate analyses for the two BAC libraries to point out potential differences in coding sequence representation. This revealed a noticeable difference between the two libraries, the *Bst*YI library showing a higher proportion of coding sequence compared to the *Hind*III library.

Table 5 Microsynteny analyses between 14,325 *C. canephora* BES pairs (showing similarities with coding regions) and *A. thaliana*, *S. lycopersicum* and *V. vinifera* whole-genome sequences

TBLASTx best hits (E-value $\leq 1e^{-6}$)	<i>A. thaliana</i>	<i>S. lycopersicum</i>	<i>V. vinifera</i>
BES pairs with single hit	5,324	5,560	5,534
BES pairs with double hits	1,383	1,752	1,642
Double hit BES pairs on the same chromosome	364	495	622
Double hit BES pairs within an interval of 10–300 kb	88	270	331

Coding sequences were annotated with GO terms. A total of 10,607 GO annotations were found and 3,812 BESs (corresponding to more than 3 % of the cumulative length of BESnr and to 4.2 % of the total number of BESnr) were characterised by at least one annotation. This result is rather low compared to results in the literature: 36 % in *Citrus* (Terol et al. 2008) or 24 % in *Brachypodium* (Huo et al. 2008), but are easily explained by the more conservative E-value threshold and protein databases we chose (UniProtKB vs. NCBIInr). Annotations were divided into three classes as follows: molecular function (45 %/3,133 sequences), biological process (29 %/2,019 sequences) and cellular component (26 %/1,803 sequences). Figure 5 shows the distribution of GO terms of gene products predicted from BES. Among the BES in the biological process categories, 32 % corresponded to proteins involved in metabolic process and 26 % were associated to cellular process. Sequences in the molecular function categories were distributed as follows: 22 % with transferase activity, 14 % with hydrolase activity and 13 % with involvement in nucleotide binding. Among the BESs in the cellular component categories, more than a half (53 %) encoded cellular proteins, 28 % membrane-bounded organelle proteins, and 5 % vesicle proteins. The distribution of GO categories in the coffee BES is comparable to previous findings in other dicot species by BES approaches (Cavagnaro et al. 2009; Han and Korban 2008) and in *Arabidopsis* by a whole genome functional annotation (Berardini et al. 2004).

Comparative genome mapping

The BAC-end sequences were also used to evaluate synteny relationships between *C. canephora* and reference plant genomes, the latter represented by *A. thaliana*, *S. lycopersicum* and *V. vinifera*. We used tBLASTx searches ($1e^{-6}$ as cut off value), with a set of BES pairs free of known TEs and containing different coding region families in both paired sequences as queries. A total of 14,325 pairs of sequences corresponding to the parameters described above were used in this analysis against reference whole-genome sequences. tBLASTx results were filtered according to the location and distance between each pair of BAC sequences and only the best hits were taken into

consideration. BES pairs mapped between intervals of 10–300 kb within the same reference plant chromosome were then considered as potential microsyntenic regions.

Table 5 shows that the *C. canephora* BESs and *V. vinifera* genome share more potential micro-syntenic regions (331) than tomato (270) and *Arabidopsis* (88) genomes, suggesting that microsynteny is higher between coffee tree and grapevine than between coffee tree and tomato. However the coffee tree and tomato genomes are more closely related than the coffee tree and *V. vinifera*. Coffee tree and tomato diverged from a common ancestor approximately 83–89 million years ago, (Wikström et al. 2001) while the divergence between coffee tree and *V. vinifera* is estimated to have occurred 114–125 million years ago. This result confirms at the genome scale the high level of conservation of genome microstructure observed between *C. canephora* and grapevine (Cenci et al. 2010; Guyot et al. 2009) and the ancestral synteny established using conserved orthologous sequences (COSII) (Guyot et al. 2012). Although distantly related, the high level of conservation between the grapevine and coffee genomes at the microstructure level suggests limited genome evolution associated with the perennial habit of these species (Cenci et al. 2013). Regarding *A. thaliana*, the limited microsynteny observed is likely the consequence of the long history of segmental duplication and the resulting genome reshuffling that occurred in *Arabidopsis* (Blanc et al. 2000; Mahé et al. 2007).

Furthermore, all BES pairs (1,642) that produced significant hits with *V. vinifera* were mapped on the 19 grapevine chromosomes as shown in Fig. 6. Distributions of BES pairs that either co-localized between intervals of 10–300 kb (331 BESs) or not, were compared. Similar comparative analysis was also performed with tomato (Fig. S5, supplemental data). A bias in the distribution of BES pairs showing significant hits along the chromosomes of both species was observed. Nevertheless regions exhibiting a high frequency of BESs showing microsynteny were detected. It is noteworthy that the potential syntenic regions identified did not appear to be uniformly distributed on the different chromosomes. The occurrences of microsynteny appear to be preferentially located in the distal part of the chromosomes; this statement is particularly obvious for tomato chromosomes whereas it is less



Fig. 6 Mapping of *C. canephora* BES pairs on grapevine (*V. vinifera*) chromosomes. Chromosome numbers are shown on the left. Coloured vertical lines indicate the positions of the *C. canephora* BES pairs mapped on the grapevine genome (TBLASTx E-value $1e^{-6}$). Red bars indicate BES pairs (1,642) mapped whatever the

interval and green bars indicate the positions of BES pairs (331) mapped within an interval of 10–300 kb (Table 5). The blue line represents the ratio between the number of BES showing microsynteny and the number of mapped BES on a 1 Mb sliding window (step size: 250 kb) along chromosomes

marked when comparing coffee and grape. This observation is likely related to the pericentric heterochromatin and distal euchromatin reported in the tomato chromosomes, resulting in a substantial higher density of genes in distal regions (Tomato Genome Consortium 2012). Although

covered by BES pairs, some regions showed a low frequency of “microsyntenic” BES pairs (see grapevine chromosomes 10 and 19 or tomato chromosomes 6, 10 and 12 in Fig. 6 and supplementary data 4), suggesting that the conservation of structure between the coffee genome and

the related genomes studied here is not uniform either along a given chromosome or from one chromosome to another.

Conclusion

The two developed BAC libraries are a genomic resource that is suitable for a broad range of applications in genetic and genomic research in coffee. In addition, the analyses and the data generated in this study provide a first glimpse of the genome constitution of *C. canephora*. Compared to reference plant genomes, a high level of microsynteny was observed between coffee tree and grapevine suggesting conservation of the microstructure of the genome. Furthermore, in relation to the ongoing *C. canephora* genome sequencing initiative, the present project appears extremely useful. While the generated paired-end sequences from BACs should considerably facilitate the scaffolding of sequence contigs, SSR identified from BESs could be used for saturating existing linkage maps and for anchoring physical and genetic maps. Moreover, the constructed TE database resulting from the present BES analysis would greatly improve the masking procedures in the assembly of *C. canephora* genomic sequences.

Acknowledgments This research was supported by a grant from the *Agence Nationale de la Recherche* (ANR; Genoplante ANR-08-GENM-022-001).

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402. doi:10.1093/nar/25.17.3389
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814):796–815
- Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol* 135(2):745–755. Epub 2004 Jun 1
- Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* 12:1093–1101. doi:10.2307/3871257
- Cavagnaro PF, Chung SM, Szklarczyk M, Grzebelus D, Senalik D, Atkins AE, Simon PW (2009) Characterization of a deep coverage carrot (*Daucus carota* L.) BAC library and initial analysis of BAC-end sequences. *Mol Genet Genomics* 281:273–288. doi:10.1007/s00438-008-0411-9
- Cenci A, Combes MC, Lashermes P (2010) Comparative sequence analyses indicate that *Coffea* (Asterids) and *Vitis* (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol Genet Genomics* 283:493–501. doi:10.1007/s11103-011-9852-3
- Cenci A, Combes MC, Lashermes P (2012) Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. *Plant Mol Biol* 78:135–145. doi:10.1007/s11103-011-9852-3
- Cenci A, Combes MC, Lashermes P (2013) Differences in evolution rates among eudicotyledon species observed by analysis of protein divergence. *J Hered*. doi:10.1093/jhered/est025
- Cheng X, Xu J, Xia S, Gu J, Yang Y, Fu J, Qian X, Zhang S, Wu J, Liu K (2009) Development and genetic mapping of microsatellite markers from genome survey sequences in *Brassica napus*. *Theor Appl Genet* 118:1121–1131. doi:10.1007/s00122-009-0967-8
- Cheung F, Town CD (2007) A BAC-end view of the *Musa acuminata* genome. *BMC Plant Biol* 7:29. doi:10.1186/1471-2229-7-29
- Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 2008:1–13
- Couturon E, Berthaud J (1982) Présentation d'une méthode de récupération d'haploïde spontanés découverts chez le *Coffea canephora* var. robusta. *Café Cacao Thé* 19(3):267–270
- Datema E, Mueller LA, Buels R, Giovannoni JJ, Visser RGF, Stiekema WJ, van Ham CHJ (2008) Comparative BAC-end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. *BMC Plant Biol* 8:34. doi:10.1186/1471-2229-8-34
- Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: *Psilanthus* (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. *Bot J Linn Soc* 167(4):357–377. doi:10.1111/j.1095-8339.2011.01177.x
- Febrer M, Cheung F, Town CD, Cannon SB, Young ND, Abberton MT, Jenkins G, Milbourne D (2007) Construction, characterization and preliminary BAC-end sequencing analysis of a bacterial artificial chromosome library of white clover (*Trifolium repens* L.). *Genome* 50:412–421. doi:10.1139/G07-013
- Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16(2):77–88
- Frelichowski JE Jr, Palmer MB, Main D, Tomkins JP, Cantrell RG, Stelly DM, Yu J, Kohel RJ, Ulloa M (2006) Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. *Mol Genet Genomics* 275:479–491. doi:10.1007/s00438-006-0106-z
- Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92–100. doi:10.1126/science.1068275
- Guyot R, de la Mare M, Viader V, Hamon P, Coriton O, Bustamante-Porras J, Poncet V, Campa C, Hamon S, de Kochko A (2009) Microcollinearity in an ethylene receptor coding gene region of the *Coffea canephora* genome is extensively conserved with *Vitis vinifera* and other distant dicotyledonous sequenced genomes. *BMC Plant Biol* 9(1):22. doi:10.1186/1471-2229-9-22
- Guyot R, Lefebvre-Pautigny F, Tranchant-Dubreuil C, Rigoreau M, Hamon P, Leroy T, Hamon S, Poncet V, Crouzillat D, de Kochko A (2012) Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* Sp.) and rosid (*Vitis vinifera*) clades. *BMC Genomics* 13:103. doi:10.1186/1471-2164-13-103
- Han Y, Korban SS (2008) An overview of the apple genome through BAC-end sequence analysis. *Plant Mol Biol* 67:581–588. doi:10.1007/s11103-008-9321-9
- Hong CP, Lee SJ, Park JY, Plaha P, Park YS, Lee YK, Choi JE, Kim KY, Lee JH, Lee J, Jin H, Choi SR, Lim YP (2004) Construction of a BAC library of Korean ginseng and initial analysis of BAC-

- end sequences. *Mol Genet Genomics* 271:709–716. doi:10.1007/s00438-004-1021-9
- Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. *Mol Cells* 22(3):300–307
- Hsu CC, Chung YL, Chen TC, Lee YL, Kuo YT, Tsai WC, Hsiao YY, Chen YW, Wu WL, Chen HH (2011) An overview of the *Phalaenopsis* orchid genome through BAC-end sequence analysis. *BMC Plant Biol* 11:3. doi:10.1186/1471-2229-11-3
- Huang XQ, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9(9):868–877. doi:10.1101/gr.9.9.868
- Huo N, Lazo GR, Vogel JP, You FM, Ma Y, Hayden DM, Coleman-Derr D, Hill TA, Dvorak J, Anderson OD, Luo MC, Gu YQ (2008) The nuclear genome of *Brachypodium distachyon*: analysis of BAC-end sequences. *Funct Integr Genomics* 8:135–147. doi:10.1007/s10142-007-0062-7
- Jaillon O, Aury J, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Huguency P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou Y, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon A, Weissenbach J, Quétier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467. doi:10.1038/nature06148
- Jansen RK, Kaitanis C, Sasaki C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6:32. doi:10.1186/1471-2148-6-32
- Jetty SS, Luo M, Goicoechea J, Wang W, Kudrna D, Mueller C, Talag J, Kim H, Sisneros N, Blackmon B, Fang E, Tomkins J, Brar D, MacKil D, McCouch S, Kurata N, Lambert G, Galbraith G, Arumuganathan K, Rao R, Walling J, Gill N, Yu Y, SanMiguel P, Soderlund C, Jackson S, Wing R (2006) The *Oryza* BAC library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries. *Genome Res* 16(1):140–147
- 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. doi:10.1159/000084979
- Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase submitter and censor. *BMC Bioinformatics* 7:474. doi:10.1186/1471-2105-7-474
- Lai CWJ, Yu Q, Hou S, Skelton RL, Jones MR, Lewis KLT, Murray J, Eustice M, Guan P, Agbayani R, Moore PH, Ming R, Presting GG (2006) Analysis of papaya BAC-end sequences reveals first insights into the organization of a fruit tree genome. *Mol Genet Genomics* 276:1–12. doi:10.1007/s00438-006-0122-z
- Lashermes P, Couturon E, Charrier A (1994) Doubled haploids of *Coffea canephora*—development, fertility and agronomic characteristics. *Euphytica* 74:149–157. doi:10.1007/BF00033781
- Lefebvre-Pautigny F, Wu F, Philippot M, Rigoreau M, Priyono, Zouine M, Frasse P, Bouzayen M, Broun P, Pétiard V et al (2010) High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. *Tree Genet Genomes* 6(4):565–577. doi:10.1007/s11295-010-0272-3
- Mahé L, Combes MC, Lashermes P (2007) Comparison between a coffee single copy chromosomal region and *Arabidopsis* duplicated counterparts evidenced high level syntenic between the coffee genome and the ancestral *Arabidopsis* genome. *Plant Mol Biol* 64:699–711. doi:10.1007/s11103-007-9191-6
- Mao L, Wood TC, Yu Y, Budiman MA, Tomkins JP, Woo S-S, Sasnowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73000 sequence tagged connectors. *Genome Res* 10:982–990. doi:10.1101/gr.10.7.982
- Messing J, Bharti AK, Karlowski KM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KFX, Wing RA (2004) Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* 101:14349–14354. doi:10.1073/pnas.0406163101
- Noirot M, Poncet V, Barre P, Hamon P, Hamon S, de Kochko A (2003) Genome size variations in diploid African *Coffea* species. *Ann Bot London* 92(5):709–714. doi:10.1093/aob/mcg183
- Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdilille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* 48:463–474. doi:10.1111/j.1365-3113X.2006.02891.x
- Perteua G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19(5):651–652. doi:10.1093/bioinformatics/btg034
- Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, de Kochko A, Hamon P (2006) SSR mining in coffee tree EST databases: potential use of EST–SSRs as markers for the *Coffea* genus. *Mol Genet Genomics* 276:436–449. doi:10.1007/s00438-006-0153-5
- Poncet V, Dufour M, Hamon P, Hamon S, de Kochko A, Leroy T (2007) Development of genomic microsatellite markers in *Coffea canephora* and their transferability to other coffee species. *Genome* 50(12):1156–1161. doi:10.1139/G07-073
- Robbrecht E, Manen JF (2006) The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of *Coptosapelta* and *Luculia*, and supertree construction based on *rbcl*, *rps16*, *trnL-trnF* and *atpB-rbcL* data. A new classification in two subfamilies, Cinchonoideae and Rubioideae. *Syst Geogr* 76:85–146
- Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA (2007) The development of BAC-end sequence based microsatellite markers and placement in the physical and genetic maps of soybean. *Theor Appl Genet* 114:1081–1090. doi:10.1007/s00122-007-0501-9
- Terol J, Naranjo MA, Ollitrault P, Talon M (2008) Development of genomic resources for *Citrus clementina*: characterization of three deep coverage BAC libraries and analysis of 46,000 BAC-end sequences. *BMC Genomics* 9:423. doi:10.1186/1471-2164-9-423
- Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641. doi:10.1038/nature11119
- Vidal RO, Costa Mondego JM, Pot D, Ambrósio AB, Andrade AC, Protasio Oeireira LF, Colombo CA, Esteves Vieira LG, Carazzolle MF, Pereira G (2010) A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. *Plant Physiol* 154:1053–1066. doi:10.1104/pp.110.162438
- Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. *Proc R Soc B Biol Sci* 268(1482):2211–2220. doi:10.1098/rspb.2001.1782