Anchor Markers for Comparative Mapping within the *Coffea* Genus

V. PONCET, P. HAMON, A. CAYREL, M.B. SAUVAGE DE SAINT MARC, T. BERNARD, S. HAMON, M. NOIROT

IRD, UMR 1097 Diversité et Génomes des Plantes Cultivées (DGPC), 911 avenue Agropolis BP 64501, 34394 Montpellier cedex 5, France

SUMMARY

In order to unravel the *Coffea* genomic reorganization, we focused on comparative mapping within the *Coffea* genus and between *Coffea* and related families such as Solanaceae. Prerequisites for comparative mapping are a genetic linkage map for each species and a common set of DNA markers that can be used to align the maps. With this objective, we started developing anchor markers corresponding to two classes: 1) anonymous genomic sequences such as SCAR markers derived from AFLPs as well as microsatellite markers, and 2) coding genomic sequences (derived from ESTs: e.g. EST-SSR, COS). For all markers analysed in this study, we obtained good transferability among *Coffea* species even when considering species belonging to different phylogenetic clades. These locus-specific markers will facilitate assembly of a panel of "anchor" PCR-based markers for the study of similarities and differences in the structure and function between *Coffea* genomes. Furthermore, we have defined an initial set of 54 highly conserved, single copy genes—COS markers—which can be used as markers for comparative mapping between the tomato and coffee genomes. These comparative maps will not only shed light on the nature of genome evolution, but will also facilitate comparative mapping studies of qualitative and quantitative traits.

INTRODUCTION

Coffee trees belong to the *Rubiaceae* family and originated from the intertropical forest of Africa, Madagascar, and Mascarene islands. All species share the same chromosomic number, x = 11, and are diploid, except *C. arabica*, which has 44 chromosomes.

Phylogenetic relationships based on cpDNA (Cros et al., 1998) and rDNA ITS-2 (Lashermes et al., 1997) analyses highlighted a strong geographical differentiation between East African, West and Central African and Malagasy species.

This genetic differentiation is supported by results on the fertility of the F1 interspecific hybrid (Louarn, 1992). Indeed, considering only African species, fertility observed on the basis of pollen viability or seed set is much higher within than between geographical clusters. With intra-cluster combinations, fertility can reach up to 90%, while it is very low in inter-cluster combinations, with complete sterility often occurring.

In all cases, the sterility of interspecific hybrids is correlated with the number of univalents in pollen mother cells and could be due to a lack of chromosome pairing during meiosis.

Moreover, despite a same number of chromosomes for diploid species, the amount of DNA per genome 2C (qDNA) ranges from 0.93 pg (*C. racemosa*) to 1.78 pg (*C. humilis*) (Cros et al., 1994; Noirot et al., 2003). Interestingly, this amount is correlated with the geographical

origin of the species. For African diploid species, the amount of qDNA increases from East to West Africa.

How can we explain differences in chromosome pairing during meiosis and differences in qDNA between species? Two nonexclusive hypotheses could be put forward: i)- gain/loss of species specific noncoding DNA including repetitive DNA and transposable elements; ii)- genomic reorganization following the size increase/decrease of genomes during speciation.
In order to test the genomic reorganization hypothesis, we first focused on comparative mapping within the *Coffea* genus and between *Coffea* and related families such as Solanaceae (Fulton et al., 2002). With this objective, we started developing anchor markers corresponding to two classes: 1)- anonymous genomic sequences such as SCAR markers derived from AFLPs as well as microsatellite markers, and 2) coding genomic sequences (derived from ESTs).

SEQUENCE-CHARACTERIZED AMPLIFIED REGION (SCAR) MARKERS

Genetic maps based on interspecific BC1 progenies have been previously produced by our research team (Ky et al., 2000; Coulibaly et al., 2003). These maps were constructed with AFLP markers specific to the donor species. We first considered the [(*C. pseudozanguebariae* x *C. liberica* var *dewevrei*) x *C. liberica* var *dewevrei*] cross (Ky et al., 2000) and analyzed specific sequences of the *C. pseudozanguebariae* (PSE) genome relative to *C. liberica* var *dewevrei* (DEW) produced from *Eco*RI and *Mse*I digestion (Poncet et al., 2004).

SCAR markers were derived from these PSE AFLP sequences. Finally, 23 AFLP bands were cloned, successfully sequenced, analyzed and the corresponding primer pairs tested on other species to determine whether they could amplify homologous sequences across *Coffea* species.

We noted two key features when considering the nature of these sequences. First, no similarities between PSE sequences were observed, indicating that none of them corresponded to disperse repetitive DNA. Then comparison of PSE sequences to public database sequences revealed significant similarity with five sequences (BLASTx and BLASTn, E-value less than e^{-04}), with two of them being expressed proteins from *Arabidopsis*.

The study of the base composition of the PSE AFLP sequences showed an AT content distribution ranging from 49.2% to 75.2% with two modes (58.6% and 71.4%), while for *Coffea* expressed sequences (EST) the AT content distribution was unimodal with a mean of 55.0%. Interestingly, the two AFLP sequences which showed some similarity with *A. thaliana* proteins, had an AT content of around 55%. An AFLP sequence showing similarity to repetitive DNA, had an AT content of 61.4%.

These results are in line with the hypothesis whereby the main part of AFLP bands produced from *Eco*RI and *Mse*I digestion more likely correspond to noncoding sequences. Indeed, due to i)-the relative importance of noncoding *vs.* coding sequences in eukaryote genomes and ii)-the DNA methylation insensitivity of *Eco*RI, access to noncoding sequences is more frequent.

In the next step, 23 SCAR primers were designed and tested on genomic DNA (gDNA) of nine *Coffea* species representative of the natural distribution of coffee trees, and gDNA of *A. thaliana* (THA). Fifteen primer pairs efficiently amplified at least one *Coffea* species and four of them gave good amplification with the nine *Coffea* species tested. *Arabidopsis thaliana* was only amplified with one SCAR primer (Table 1). The amplification success was found to be independent of the species considered and did not reflect the relatedness between

the species as defined by Lashermes et al. (1997). In general, amplification of a single fragment of the expected size was obtained with the PSE genotype.

However, despite this high transferability potential, we observed a very low level of polymorphism, even between species. Consequently, only two SCAR primers could be mapped on the [(*C. canephora* x *C. pseudozanguebariae*) x *C. canephora*] progeny and one was used for the [(*C. liberica* x *C. canephora*) x *C. canephora*] mapping.

This result is not specific to the *Coffea* genome and it is well known that during SCAR development the loss of the original polymorphism often poses experimental challenges. Indeed, the original polymorphism was found to be mainly located at the restriction sites. It was not always possible to design primers covering the corresponding region. Nevertheless, when SCARs are produced from AFLPs linked to QTLs of interest, the amount of variation they could reveal should be further studied, for example through enzymatic digestion of the PCR products (cleaved amplified polymorph sequences, CAPS).

Table 1. Amplification with SCAR primers across nine *Coffea* species and *Arabidopsis thaliana* and analyses on mapping progenies.

Species[a]	PSE	DEW	CAN	CON	EUG	HET	MIL	MOL	ARA	THA	Mapping potential[b]
Amplification (%)	10/21 (47.6)	5/21 (23.8)	8/21 (30.9)	9/21 (42.8)	7/21 (33.3)	10/21 (47.6)	7/21 (33.3)	8/21 (30.9)	8/21 (30.9)	1/21 (4.8)	
Weak amplification (%)	2/21 (9.5)	3/21 (14.3)	5/21 (23.8)	3/21 (14.3)	4/21 (19)	-	4/21 (19)	3/21 (14.3)	3/21 (14.3)	0	
Total	12	8	13	12	10	10	10	11	11	1	5

[a]: *PSE = C. pseudozanguebariae, DEW = C. liberica var dewevrei, CAN = C. canephora, CON = C. congensis, EUG = C. eugnenioides, HET = C. heterocalyx, MIL = C. millotii, MOL = C. sp Moloundou, ARA = C. arabica, THA = Arabidopsis thaliana.*

[b] : *mapping analyses: at least one mapping progeny concerned*

MICROSATELLITE MARKERS FROM ANONYMOUS GENOMIC SEQUENCES

Microsatellite primer sets were designed on *C. arabica* sequences obtained from Genbank/EMBL databases (Combes et al., 2000; Rovelli et al., 2000) or on *C. canephora* sequences (Dufour et al., 2001). Primers obtained from Baruah et al. (2003) were also evaluated. All of them were tested for amplification and their ability to reveal polymorphism on several *Coffea* species. Indeed, as expected, a previous microsatellite study carried out on *C. canephora* (CAN) and *C. pseudozanguebariae* (PSE) revealed a high level of genetic diversity (measured with PIC* value) within species (Poncet et al., 2004). Furthermore, both PIC distributions were bimodal but there was no correlation between the two PIC sets. This result is important since it indicates that it is not possible to predict the polymorphism level of one given species from the polymorphism observed in another species (Poncet et al., 2004).

Finally, out of 355 primer pairs available, only 53 could be effectively used for CP mapping while 33 have to be tested on the progeny (Table 2).

---

* ($PIC_i = 1 - \sum_{j=1}^{n} P_{ij}^2$, where P_{ij} is the frequency of the jth allele for the ith marker and summed over n alleles)

Table 2. Results obtained with the three sets of primer pairs tested. Only results related to the ((*C. canephora* x *C. pseudozanguebariae*) x *C. canephora*) (CP) mapping are reported.

Primer origin	# primer pairs	No PCR product	analyzed on CP progeny	to test on CP progeny
C. arabica[1]	116	48 (41.4%)	25 (21.6%)	-
C. canephora[2]	230	115 (50%)	26 (11.3%)	33 (14.3%)
C. arabica[3]	9	-	2	-

[1]: *Poncet et al. (2004) based on sequences from Combes et al. (2000) and Rovelli et al. (2000)*
[2]: *based on sequences from Dufour et al. (2001)*
[3]: *from Baruah et al. (2003)*

Likewise, some of these primers were tested on another interspecific progeny ((*C. canephora* x *C. heterocalyx*) x *C. canephora*) and 34 couples were easily readable on the BC1 progeny (Coulibaly et al., 2003).

SSR MARKERS DERIVED FROM *COFFEA* EST SEQUENCES

EST databases provide a valuable resource for the development of SSR-markers, which are associated with transcribed genes. Two *C. canephora* cDNA libraries were produced from two organs, i.e. leaves and fruits and sequencing is ongoing. Currently, two sets of 5814 fruit EST sequences and 3112 leaf EST sequences are available in our Coffee database (Table 3).

In a first step, nonnuclear sequences were eliminated. Then, search of microsatellite motifs was performed on the remaining sequences using a modified version of Tandem Repeat Finder (http://tandem.bu.edu) (C. Tranchant, christine.tranchant@mpl.ird.fr).

Table 3. Summary of *Coffea canephora* EST libraries (in progress).

	Leaf library	Fruit library
Total sequences	3112	5814
Clean sequences	2709 (87 %)	5706 (98 %)
Mean size	575 bp	599 bp
Potential Unigenes (singletons + contig)	1859 (1520 (56 %) + 339)	3436 (2552 (45 %) +884)
	4852 (3551 + 1301)	
Microsatellites	446 (~10%)	

In a second step, sequences including microsatellites were compared to public database sequences using the BLASTx algorithm. The search was conducted in July 2004 using the default parameters. Similarities were considered significant when the E-value was less than e^{-30}.

In a third step, the 25 more significant E-values were chosen for primer designing (Primer 3 software). Out of these 25 primer pairs, 23 gave an amplification with at least one of the seven *Coffea* species involved in the mapping populations. Since only diagnostic loci – distinguishing the two species of an interspecific cross – can be considered, intraspecies polymorphism was assessed. Finally, between 8 and 12 primer pairs could be used for mapping purposes i.e. 34.8 to 52.2% of the primers tested (Table 4). These numbers are higher than that obtained from anonymous genomic SSR sequences.

Table 4. Results concerning the 23 primer pairs giving a PCR product and obtained on the five mapping progenies currently studied.

PCR amplification / 25 pairs	Mapping progeny	Useful for mapping / 23
PSE: 20 (80%)	CP	11 (47.8%)
HET: 18 (72%)	CH	11 (47.8%)
DEW: 21 (84%)	PD	12 (52.2%)
LIB: 23 (92%)	LC	9 (39.1%)
EUG: 22 (88%)	CE	8 (34.8%)
CAN: 23 (92%)		

CP : (C. canephora x C. pseudozanguebariae) x C. canephora
CH : (C. canephora x C. heterocalyx) x C. canephora
PD : (C. pseudozanguebariae x C. liberica var dewevrei) x C. liberica var dewevrei
LC : (C. liberica x C. canephora) x C. canephora
CE : (C. canephora x C. eugenioides) x C. canephora

SSR MARKERS DERIVED FROM *LYCOPERSICON* EST SEQUENCES

Coffee belongs to Rubiaceae and is closely related to the Solanaceae family, as these two families are included in the Asterid I class. We could thus take advantage of the sequencing of the tomato genome thanks to the generation of shared markers. Among tomato (*Lycopersicon esculentum*) sequences, EST sequences including microsatellite motifs were identified and some of them mapped on the *L. esculentum* LA925 x *L. pennellii* LA716 linkage map (http://www.sgn.cornell.edu/cgi-bin/mapviewer/mapTop.pl?map_id=1). We selected EST markers which mapped with a LOD score of at least 2 and then looked for similarity with our *Coffea* EST sequences using BLASTn algorithms. Of the 32 *Lycopersicon* EST sequences retained, only one had good similarity (E-value = e^{-138}), with a *Coffea* EST sequence similar to an histone H3.2 protein. Thus, except for this sequence, only heterologous primers previously designed and available on the **Solanaceae Genomics Network** website (http://www.sgn.cornell.edu) could be used and tested for amplification on the seven *Coffea* species involved in mapping progenies.

Among the 32 primer pairs tested, 8 did not give any amplification. Between 11 and 21 were not useful for mapping purposes and 12.4 % to 50 % will be analyzed to determine their progeny mapping potential (Table 5). A lower number of primers are useful for mapping purposes compared to the results given in Table 4. These results are not surprising since these SSR primers are derived from heterologous sequences.

Table 5. Results concerning the 24 primer pairs giving a PCR product and obtained on the five mapping progenies studied.

Mapping progeny	Amplification but not useful for mapping	Total # for mapping (%)
CP	11 (45.8%)	12 (49.9)
CH	16 (66.6%)	8 (38.5)
PD	12 (50%)	12 (50)
LC	21 (87.5%)	3 (12.4)
CE	20 (83.3%)	4 (16.6)

CP : (C. canephora x C. pseudozanguebariae) x C. canephora
CH : (C. canephora x C. heterocalyx) x C. canephora
PD : (C. pseudozanguebariae x C. liberica var dewevrei) x C. liberica var dewevrei
LC : (C. liberica x C. canephora) x C. canephora
CE : (C. canephora x C. eugenioides) x C. canephora

COS MARKERS DERIVED FROM *LYCOPERSICON-ARABIDOPSIS-COFFEA* CONSERVED ORTHOLOG SET SEQUENCES.

To overcome the problem of sequence divergence between *Coffea* and the Solanaceae family and to define a set of shared markers, we screened our EST database for sequences that could correspond to COS (conserved orthologous set) markers. These markers are derived from the identification of unique genes which are highly conserved between plant species – initially tomato – and arabidopsis and their functions have been conserved throughout evolution.

In a first step, we considered the 1025 conserved ortholog set (COS) sequences as defined by Fulton et al. (2002). Out of them, 311 were effectively mapped on the tomato genome with a LODscore ≥ 2. Considering only this set of sequences, their comparison to our *Coffea* EST sequences using BLASTn revealed close similarity with 54 of them (E-value $< e^{-10}$ with a mean value of e^{-43}). Moreover, the corresponding 54 markers are distributed all over the 12 tomato chromosomes. Using BLASTx, a putative function was attributed to 42 sequences (e-value <-30). Finally, this set of 54 EST sequences would be a first initial set of COS markers to map on our progenies.

In a second step, multiple alignments were performed to identify conserved and variable regions of the EST sequences. ORFs (open reading frames) were identified.

An analysis is currently under way to take the location of conserved regions and of ORFs into account. The primer pairs will be designed and tested for amplification on the seven *Coffea* species. When possible, they will be used for *Coffea* interspecific mapping projects.

CONCLUSION

For all markers analysed in this study, we obtained good transferability among *Coffea* species even when considering species belonging to different phylogenetic clades such as CAN, EUG, and PSE. The amplification success was found to be independent of the species considered and did not reflect the between-species relatedness. This is in agreement with the recent speciation scenario noted within the *Coffea* genus.

All these locus-specific markers will facilitate assembly of a panel of "anchor" PCR-based markers for comparative mapping studies in coffee trees and for marker-assisted selection. Moreover, microsatellites associated to ESTs as well as COS markers have the advantage of detecting unique expressed regions of the genome. They will facilitate the study of similarities and differences in the structure and function between *Coffea* genomes.

Furthermore, we have defined an initial set of 54 highly conserved, single copy genes – COS markers – which can be used as markers for comparative mapping between the tomato and coffee genomes.

These comparative maps will not only shed light on the nature of genome evolution, but will also facilitate comparative mapping studies of qualitative and quantitative traits.

REFERENCES

Baruah A, Naik P, Hendre S, Rajkumar R, Rajendrakumar P, Aggarwal RK (2003) Isolation and characterization of nine microsatellite markers from Coffea arabica L., showing wide cross-species amplifications. Molecular Ecology Notes 3:647-650

Combes MC, Andrzejewski S, Anthony F, Bertrand B, Rovelli P, Graziosi G, Lashermes P (2000) Characterization of microsatellite loci in *Coffea arabica* and related coffee species. Molecular Ecology 9:1178-1180

Coulibaly I, Revol B, Noirot M, Poncet V, Lorieux M, Carasco-Lacombe C, Minier J, Dufour M, Hamon P (2003) AFLP and SSR polymorphisme in a *Coffea* interspecific backcross progeny ((*C. heterocalyx* x *C. canephora*) x *C. canephora*. Theoretical and Applied Genetics 107:1148-1155

Cros J, Combes MC, Trouslot P, Anthony F, Hamon S, Charrier A, Lashermes P (1998) Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. Molecular Phylogenetics and Evolution 9:109-117

Cros J, Gavalda MC, Chabrillange N, Recalt C, Duperray C, Hamon S (1994) Variations in the total nuclear DNA content in African *Coffea* species (Rubiaceae). Cafe, Cacao, The 38:3-10

Dufour M, Hamon P, Noirot M, Ristrerucci AM, Brottier P, Vico V, Leroy T (2001) Potential use of SSR markers for *Coffea* spp. genetic mapping. 19th Int. Sci. Colloq. on Coffee, Trieste, Italy

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 14:1457-1467

Ky CL, Barre P, Lorieux M, Trouslot P, Akaffou S, Louarn J, Charrier A, Hamon S, Noirot M (2000) Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (*Coffea* sp.). Theoretical and Applied Genetics 101:669-676

Lashermes P, Combes MC, Trouslot P, Charrier A (1997) Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. Theoretical and Applied Genetics 94:947-955

Louarn J (1992) La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origine africaine (Genre *Coffea* L. sous-genre *Coffea*). Montpellier II, Sciences et Techniques du Languedoc, Montpellier, France

Noirot M, Poncet V, Barre P, Hamon P, Hamon S, De Kochko A (2003) Genome size variations in diploid African Coffea species. Ann Bot (Lond) 92:709-714

Poncet V, Hamon P, Minier J, Carasco-Lacombe C, Hamon S, Noirot M (2004) SSR cross-amplification and variation within coffe trees (*Coffea* spp.). Genome In press

Poncet V, Hamon P, Sauvage de Saint Marc M-B, Bernard T, Hamon S, Noirot M (2004) Base composition of *Coffea* AFLP sequences and their conservation within the genus. Journal of Heredity In press

Rovelli P, Mettulio R, Anthony F, Anzueto F, Lashermes P, Graziosi G (2000) Microsatellites in *Coffea arabica* L. In: Roussos S (ed) Coffee Biotechnology and Quality. Kluwer Academic Publishers, Netherlands, pp 123-133

Poncet Valérie, Hamon Perla, Cayrel A., Sauvage de Saint Marc M.B., Bernard T., Hamon Serge, Noirot Michel (2005)

Anchor markers for comparative mapping within the *Coffea* genus

In: ASIC 2004: 20th international conference on coffee science = ASIC 2004 20ème colloque scientifique international sur le café = ASIC 2004 : 20 internationales wissenshaftliches kolloquim über kaffe = ASIC 2004 : 20° coloquio cientifica internacional sobre el café

Paris: ASIC, 560-566. ASIC 2004: International Conference on Coffee Science, 20

ISBN 2-900212-19-7