Nucleotide Diversity of Genes Involved in Sucrose Metabolism. Towards the Identification of Candidates Genes Controlling Sucrose Variability in *Coffea* sp.

D. POT[1,2], S. BOUCHET[1], P. MARRACCINI[1,2], F. DE BELLIS[1], P. CUBRY[1], I. JOURDAN[1], L. F. P. PEREIRA[3], L. G. E. VIEIRA[2], L. P. FERREIRA[2], P. MUSOLI[4], H. LEGNATE[5], T. LEROY[1]

[1]CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) TA 80/03 Avenue d'Agropolis, 34398 Montpellier Cedex 5 France
[2]IAPAR (Instituto Agronômico do Paraná) LBI-AMG CP 481 86001-970 Londrina (PR) Brazil
[3]EMBRAPA Café / IAPAR
[4]NARO-CORI P.O. BOX 185 Mukono Uganda
[5]CNRA Ivory Coast

SUMMARY

Quality and drought stress tolerance are two important targets for *Coffea* species cultivation. Currently, efficient genetic improvement of these traits is still hampered by the lack of early and cheap predictors. In this context, identification of molecular tools linked to these traits would significantly improve breeding efficiency. Based on the available literature, different metabolisms involved in the variability of both drought tolerance and coffee quality can be proposed. Based on this information, a study was initiated in *Coffea* species, aiming at estimating nucleotide diversity of four sucrose metabolism enzymes (Sucrose Synthase, Cell Wall Invertase, acid Vacuolar Invertase and Sucrose Phosphate Synthase). The two mains objectives of this work were i) to assess the level of variability of these genes within the whole area of distribution of *Coffea canephora,* and within 15 related *Coffea* species representing the four groups of diversity of this genus, and ii) to identify polymorphisms useful for mapping and association genetic studies. Almost 200 polymorphisms (SNP, INDELS, SSR) were identified through sequencing of *Coffea canephora* genotypes. In addition, analysis of the variability of these genes between different *Coffea* species allowed the identification of 300 additional polymorphic sites. Parallel in-silico analysis of EST resources confirmed the interest of this approach towards the identification of polymorphisms in *Coffea* sp. Identification of nucleotide polymorphisms will not only provide useful markers for traditional genetic studies (genetic mapping, population genetics, association studie) but also provide criteria to infer the evolutionary history of the analysed genes. Such information will be particularly relevant to select the best candidate genes to test in future association studies.

INTRODUCTION

Coffee is one of the world's heavily traded commodities. However, little is known about the genomic control of cup quality and abiotic stresses tolerance in *Coffea sp* which are two key components of the sustainability of the coffee market. Currently, rapid genetic improvement of these traits is still hampered by the lack of early and cheap predictors, phenotypic ones being cost and time consuming to use. In this context, identification of molecular tools linked to these traits would significantly improve breeding efficiency. Among the different metabolisms involved in fruit quality development and drought tolerance, sucrose metabolism

is particularly relevant. Several studies underlined the importance of enzymes/genes of this biosynthesis pathway in drought stress response (Andersen et al., 2002; Hazen et al., 2005; Pelah et al., 1997). At the coffee quality level, sucrose has been pointed out as an important precursor because its degradation during roasting leads to allyphatic acids, hydroxymethyl furfural and furans that contribute to flavours and aromas (Grosch, 2001; Homma, 2001). In addition, part of the preference for *C. arabica* compared to *C. canephora* coffee, has been frequently attributed to sucrose content differences between these species (Guyot et al., 1996; Casal et al., 2000; Ky et al., 2001).

Identification of the genomic regions and genes controlling the variability of traits of agronomic importance is a long and difficult task. Identification of QTLs in mapping pedigrees, despite its importance, only provides partial information on the genetic control of these traits, the underlying genes and the responsible polymorphisms remaining unknown. Such lack of information often hampers the application of marker assisted breeding in conventional breeding schemes.

In this context the aim of this study was to evaluate the nucleotide diversity of genes encoding 4 enzymes of the sucrose biosynthesis pathway (Cell Wall Invertase (CWI), acid Vacuolar Invertase (VI), SUcrose Synthase (SUS) and Sucrose Phosphate Synthase (SPS)). This analysis was performed using two strategies: i) direct sequencing of *C canephora* and *C.* spp genotypes and ii) in-silico analysis of the EST resources available. In addition, analysis of the landscape of diversity of these genes from an evolutionary point of view provides information useful to identify the best candidate gene/ sites to use in future mapping and association experiments.

MATERIAL AND METHODS

Polymorphism discovery by direct sequencing

The primers used to amplify the targeted genes were designed based on the sequences available in The Brazilian Coffee Genome Project (Vieira et al., 2006; http://www.lge.ibi.unicamp.br/cafe/). After a first polymorphism discovery step in *C. canephora* based on 7 genotypes belonging to the different genetic groups (Congolese SG1, Congolese SG2, Congolese B, Congolese C, Guineans, Uganda wild and Uganda N'Ganda, (Cubry et al., 2006)) a larger sample size (35 to 70 genotypes belonging to the different genetic groups) was analysed for the most interesting genes (SUS1, SUS2 and SPS). *C. canephora* genotypes belonging to the Congolese and Guineans groups were provided by the CNRA (Centre National de Recherche Agronomique) of Ivory Coast Republic whereas the Ugandan genotypes (Wild and N'Ganda) were obtained from the NARO-CORI of Uganda. All the genotypes used for the interspecific analyses (14 species) came from the IRD (Institut de Recherche pour le Développement) collection (Montpellier, France).

In-silico polymorphism discovery

In parallel with the polymorphism discovery based on the direct sequencing strategy, a search of the polymorphisms available in the EST resources generated by the Brazilian Coffee Genome Project and the Nestlé/ Cornell project (Lin et al., 2005; http://harvest.ucr.edu/) was performed using exclusively the contigs containing at least four sequences. Only polymorphisms for which the rare allele was present at least twice were considered.

RESULTS AND DISCUSSION

Sucrose metabolism genes: Mining of the Brazilian Coffee Genome Project

As a starting point of this project, data from the Brazilian Coffee Genome Project were analysed with a particular emphasis on the genes encoding the four main enzymes of sucrose metabolism. Between 6 (for SPS) and 237 ESTs (for SUSY) were identified (Table 1). Bioinformatic analyses of these sequences allowed the identification of 7 putative genes for the CWI and two for the other genes (2 SUS, 2 SPS and 2 VI). Based on this information, primers were defined to amplify 3 CWI genes, 1 VI gene and the 4 putative genes encoding SUS and SPS proteins.

Table 1. Sucrose biosynthesis pathway: Mining of the Brazilian Coffee Genome Project.

Gene	Number of ESTs	Number of singletons	Number of contigs	Number of contigs > 4 seq	Deduced gene number[a]
Cell Wall Invertase (CWI)	22	3	6	2	7 (2)
Vacuolar invertase (VI)	9	1	1	1	2 (1)
Sucrose Phosphate Synthase (SPS)	6	0	2	0	2 (0)
Sucrose Synthase (SUS)	237	18	9	9	2 (2)

[a]*Number of genes expected based on the BLASTX results obtained using the contigs and singleton. Between parentheses: the number of genes for which contigs, with at least 4 sequences, are available.*

Assessment of *Coffea canephora* nucleotide diversity by direct sequencing

All the fragments were sequenced in at least one genotype belonging to the different genetic groups (7 genotypes). The SUS and SPS genes were then analysed using a larger sample size (35 to 70 genotypes). Sizes (in bp) of the explored regions and polymorphic sites discovered are indicated in the Table 2.

Table 2. Polymorphism discovery in *C. canephora.*

Gene	Regions explored (bp)					Polymorphic sites			
	5'UTR	Coding	3'UTR	Introns	Total	Ntot	SNP	INDELS	SSR
CWI_1	0	191	158	0	349	2	2	0	0
CWI_3	0	529	0	0	529	3	3	0	0
CWI_5	0	276	60	195	531	12	11	0	1
VI_1	0	732	0	314	1046	13	13	0	0
SUS_1	542	2183	273	1159	4157	36	31	0	5
SUS_2	0	2375	570	1865	4810	97	79	11	7
SPS_C1	0	286	377	0	663	17	17	0	0
SPS_C2	0	664	0	535	1199	1	1	0	0
Total	542	7236	1438	4068	13284	181	157	11	13

A total of 13 kb was explored leading to the identification of 181 polymorphisms including 157 Single Nucleotide Polymorphisms (SNP), 11 INsertions/DELetionS (INDELS) and 13 microsatellites (SSR). On average 1.2 SNP was detected every 100bp. When considering all the sequenced fragments, most of the SNP were located in untranslated regions (69%) and one third of the polymorphisms detected in the coding regions led to Non Synonymous

mutations (33 Synonymous [S] vs 16 Non Synonymous [NS]) (Table 3). In the context of association studies aiming at identifying the genes/sites controlling the genetic variability of agronomic traits, this class of polymorphism will be the most interesting to use as they are likely to induce modification of enzyme's activity/affinity.

Table 3. Landscape of nucleotide diversity (SNP only) in *C. canephora.*

Gene	Ntot	Intron	5'UTR	3'UTR	ExsonS	ExsonNS
CWI_1	2	–	–	2	0	0
CWI_3	3	–	–	–	2	1
CWI_5	11	6	–	1	2	2
VI_1	13	5	–	–	4	4
SUS_1	31	8	10	9	4	0
SUS_2	79	46	–	8	17	8
SPS_C1	17	–	–	12	4	1
SPS_C2	1	1	–	–	0	0
Total	157	66	10	32	33	16

In a second step, according to the results obtained through physiological and genomic studies (Geromel et al., 2006; Marraccini et al., unpubished) it has been chosen to focus our efforts on the Sucrose Synthase and Sucrose Phosphate Synthase genes, which are the ones that seem to regulate sucrose accumulation in the coffee bean.

For SUS1 and SUS2, almost the full-length genomic sequences were analysed (4.1 and 4.8 kb respectively). For both, much more Synonymous (respectively 4 and 17) than Non Synonymous polymorphisms (respectively 0 and 8) were detected, suggesting that these genes are under strong evolutionary constraints due to their key role in plant development.

For SUS1, SUS2 and SPS_C1, examination of at least 5 genotypes per genetic group allowed the analysis of the population genetic structure based on nucleotide polymorphism information. Genetic differentiations (Fst) ranging from 0.46 (SPS) to 0.54 (SUS1 and SUS2) were observed. These values correspond to the one's obtained with SSR (0.380; Cubry et al., 2006). If this analysis did not reveal, in this particular case, a discrepancy between candidate genes and neutral markers (SSR), this strategy can be extremely useful to identify genes of agronomic interest (these genes usually presenting diverging patterns of evolution compared to neutral genomic regions (Pot et al., 2005)).

Nucleotide diversity of SUS1, SUS2 and SPS in *Coffea* sp

For SUS1, SUS2 and SPS_C1, in addition to *C. canephora*, 14 species were analysed. These species belong to the 4 groups of diversity of *Coffea* and are characterized by different amounts of sucrose and present variable levels of tolerance to drought stress. For the 3 genes analysed, 311 polymorphic sites corresponding to 265 SNP, 28 INDELS and 18 SSR were detected (Table 4).

Table 4. Polymorphism discovery in *Coffea* species.

Gene	Ntot	SNP	Indels	Microsat
SUS_1	142	121	9	12
SUS_2	107	90	11	6
SPS_C1	62	54	8	0
Total	311	265	28	18

Consistently with the results reported at the *C. canephora* level, most of the SNP were located in untranslated regions (76 %), and within the coding regions the Synonymous mutations were predominant (53 S vs 10 NS) (Table 5). These results suggest that at both *C. canephora* and *C.* spp levels these genes are under high selective constraints confirming their key role in plant development.

Table 5. Landscape of nucleotide diversity (SNP only) in *Coffea species.*

Gene	Ntot	Intron	5'UTR	3'UTR	ExsonS	ExsonNS
SUS_1	121	88	–	–	26	7
SUS_2	90	47	–	23	19	1
SPS_C1	54	–	–	44	8	2
Total	265	135	0	67	53	10

In addition, this study allowed the confirmation of the origin of *C. arabica*. Lashermes et al. (1997) and Cros et al. (1998) using respectively ITS from nuclear DNA and chloroplastic sequences proposed that the two parental species of *C. arabica* could be *C. canephora* and *C. eugenioides*. The results obtained at the nucleotide level for SUS1, SUS2 and SPS_C1 confirmed this hypothesis revealing for these three genes close relationships between the haplotypes of *C. arabica* and the ones of these two species (Figure 1). It is also interesting to note that when considering the 26 sites heterozygotes in *C. arabica*, in 92 % of these cases (24), the two alleles were present in *C. canephora* and/or *C. eugenioides*. And within these 24 sites, 19 present fixed differences between the two species.

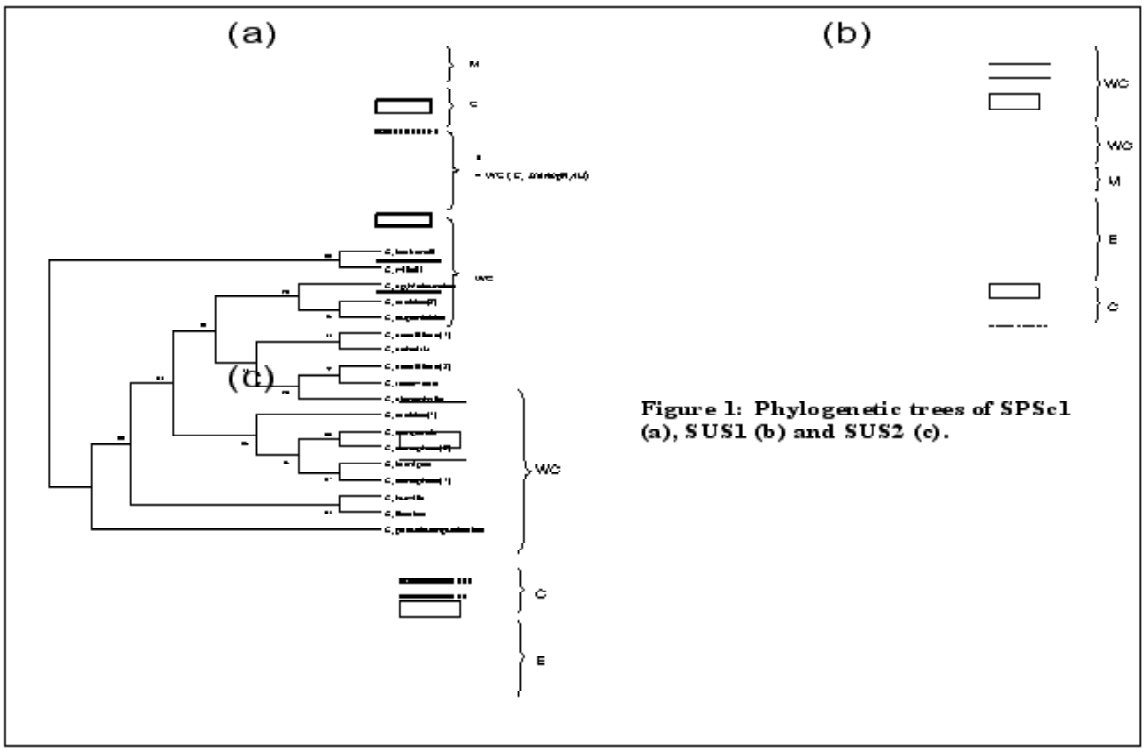

Figure 1. Phylogenetic trees of SPS_C1 (a), SUS_1 (b) and SUS_2(c).

In-silico polymorphism detection

SNP have recently become the marker of choice in genetic analyses due to their abundance and stability compared to SSR and INDELS. Although the most classical way to identify SNP is, as presented earlier, direct sequencing of amplicons, an alternative method takes advantage

of the redundancy of gene sequences generated in EST sequencing programmes. Such programmes have been developed for coffee, generating a total of 261 964 sequences (47000 from the Cornel/Nestle project and 214 964 from the Brazilian Coffee Genome Project) offering the opportunity to initiate an in-silico polymorphism discovery for the coffee species.

Table 6. Polymorphic sites detected in the Brazilian Coffee Genome Project.

Gene	Lengh	Number sequences[a]	Number of polymorphic sites[b]	Number of SNP	Number of Indels	Number of SSR	Number of Non Synonymous mutations[c]
CWI_1	1432	4 (0)	0	0	0	0	0
CWI_5	1748	6 (0)	1	1	0	0	1
VI_1	1723	8 (0)	20	20	0	0	8
SUS_2	1804	13 (0)	0	0	0	0	0
SUS_2	257	5 (0)	0	0	0	0	0
SUS_1	2985	200 (8)	50 (10)	43	4	3	4 (0)

[a]Between parentheses, the number of sequence of Coffea racemosa; [b]Between parentheses, the number of polymorphic sites corresponding to fixed differences between Coffea arabica and Coffea racemosa; [c]Between parentheses, the number of synonymous mutations corresponding to fixed differences between Coffea arabica and Coffea racemosa.

Analysis of Brazilian Coffee Genome Project allowed the identification of 64 polymorphic sites (Table 6). These polymorphisms were mainly detected in VI_1 (acid Vacuolar Invertase I) and SUS1 (Sucrose Synthase I). Out of these 64 polymorphic sites, 13 lead to NS modifications. According to the availability of *C. arabica* (which is of allotetraploid origin) and *C racemosa* sequences in the Brazilian Coffee Genome Project, comparison of the polymorphisms discovered through the direct sequencing methodology at the interspecific level and the ones discovered In-Silico was possible. Based on this comparison it appeared that 50 % (34/66) of the polymorphic sites detected by traditional sequencing and potentially present in the Brazilian Coffee Genome Project (polymorphisms located in the coding sequence and in regions available in the Brazilian Coffee Genome Project) were detected.

At the *C. canephora* level, 34 % (12/35) of the polymorphic sites detected by traditional sequencing and potentially present in the Brazilian Coffee Genome Project were detected. This result suggests that a significant part of the variability present in *C. canephora* is still present in *C. arabica*. In addition, this result underlines the interest of the analysis of *C. arabica* EST resources not only in the frame of project concerning exclusively *C. arabica* but also *C. canephora*.

Regarding the Nestle/Cornell EST resource, only SUS1 was analysed (70 sequences). Indeed for the other genes, the minimum number of sequences per contig (i.e 4) was not reached. For this gene, 25 polymorphisms were detected. Out of these, 10 had been detected by traditional sequencing (4 of them were initially considered as false positive in the traditional sequencing strategy according to their localization in low quality sequences). Out of the 10 "validated" polymorphic sites detected by the traditional sequencing strategy in the same region in *C. canephora*, 6 were detected in the Nestle/Cornell EST database (60%). When compared to the sequencing approach, 15 additional polymorphisms were found in the Nestlé/Cornell dataset. Several reasons can be proposed to explain the detection of additional polymorphic sites in-silico: i) low density of sequencing in the region analysed (1 genotype per group), ii) small sample size at the within group level…

The results obtained through the "In-Sillico polymorphism discovery" strategy revealed its importance and complementarity with the strategy of direct sequencing. This strategy could easily provide the Coffee community with a large set of polymorphic markers useful for various purposes: genetic mapping, association studies, analysis of geographic origin, certification of varieties…

CONCLUSION AND PERSPECTIVES

Analysis of nucleotide diversity of sucrose biosynthesis genes allowed the identification of several polymorphisms at the intra and interspecific levels. In addition to their interest for traditional genetic studies like genetic mapping, population genetic analysis, association studies… their pattern of diversity will also provide criteria to infer the evolutionary history of the analysed genes. Such information will be particularly relevant to select the best candidate genes to test in future association studies.

The results obtained through the In-Silico polymorphism discovery strategy confirmed the importance of this approach towards the identification of markers useful for the coffee community. A whole genome scan could be rapidly initiated using the EST resources currently available (Brazilian Coffee Genome Project and Nestlé/Cornell project); such approach which presents an extremely low cost compared to traditional sequencing efforts would provide in a short term a large set of polymorphism relevant towards the identification of molecular markers useful for marker assisted breeding of coffee sp.

ACKNOWLEDGEMENTS

This project was supported by the "Brazilian Consortium for Coffee Research and Development" (2003-2005). D. Pot and P. Marraccini received financial support from the French Embassy in Brazil (project DCSUR-BRE-4C5-008). We thank the team of "Genomic and coffee quality" of IRD (Montpellier France) and specially Dr Alexandre de Kochko for graciously providing the Coffea species samples that were used in the interspecific studies.

REFERENCES

Andersen, M.N.; Asch, F.; Wu, Y.; Jensen, C.R.; Naested, H.; Mogensen, V.O.; Koch, K.E. Soluble invertase expression is an early target of drought stress during the critical, abortion-sensitive phase of young ovary development in Maize. *Plant Physiol*. 2002, 130: 591-604.

Casal, S.; Oliveira, M.B.; Ferreira, M.A. HPLC/diode-array to the thermal degradation of trigonelline, nicotinic acid and caffeine in coffee. *Food Chem*. 2000, 68:481-485.

Cros, J.; Combes, M.C.; Trouslot, P.; Anthony, F.; Hamon, S.; Charrier, A.; Lashermes, P. Phylogenetic analysis of chloroplast DNA variation in Coffea L. Mol *Phylogenet Evol*. 1998, 9, 109-117.

Cubry, P.; Musoli, P.; Legnate, H.; Aluka, P.; Dufour, M.; Pot, D.; De Bellis, F.; Leroy, T. Genetic diversity within Coffea canephora germplasm maintained in RCI using microsatellites loci. Results and future prospects. . *Proceedings of ASIC*. 2006, Montpellier, France, Sept 11-15 2006.

Geromel, C.; Ferreira, L.P.; Guerreiro, S.M.C.; Cavalari A.A.; Pot, D.; Pereira, L.F.P.; Leroy, T.; Vieira, L.G.E.; Mazzafera, P.; Marraccini, P. Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development. In press in *Journal of experimental Botany* 2006.

Grosch, W. Volatile compounds, in *Coffee: recent developments*, ed by Clarke RJ and Vitzthum OG. Blackwell Science. 2001, Oxford, pp 68-89.

Guyot, B.; Manez, J.C.; Perriot, J.J.; Giron J.; Villain L. Influence de l'altitude et de l'ombrage sur la qualité des cafés arabica. *Plant. Rech. Dév.* 1996, 3:272-280.

Hazen, S.P.; Pathan, M.S.; Sanschez, A.; Baxter, I.; Dunn, M.; Estes, B.; Chang, H-S.; Zhu, T.; Kreps, J.A.; Nguyen, H.T. Expression profiling of rice segregating for drought tolerance QTLs using a rice genome array. *Functional and integrative genomics*. 2005, 5: 104-116.

Homma, S. Non-volatile compounds, part II, in *Coffee: recent developments*, ed by Clarke RJ and Vitzthum OG. *Blackwell Science*. 2001 Oxford, pp 50-67.

Ky, C.L.; Louarn, J.; Dussert, S.; Guyot B.; Hamon, S.; Noirot, M. Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora* P. accessions. *Food Chem*. 2001, 75:223-230.

Lashermes, P.; Combes, M.C.; Trouslot, P.. Charrier, A. Phylogenetic relationships of coffee-tree species (Coffea L.) as inferred from ITS sequences of nuclear ribosomial DNA. *Theor Appl Genet*. 1997, 97, 947-955.

Lin, C.; Mueller, L.A.; Mc Carthy, J.; Crouzillat, D.; Pétiard, V.; Tanksley, S.D. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theor Appl Genet*. 2005, 112: 114-130.

Pelah, D.; Wang, W.; Altman, A.; Shoseyov, O.; Bartels, D. Differential accumulation of water stress-related proteins, sucrose synthase and soluble sugars in *Populus* species that differ in their water stress response. *Physiol Plant*. 1997, 99: 153-159.

Pot, D.; McMillan, L.; Echt, C.; Le Provost, G.; Cato, S.; Plomion, C. Nucleotide variation in genes involved in wood formation in two pine species. *New Phytologist.* 2005, 167: 101-112.

Vieira, L.G.E.; Andrade, A.C.; Colombo, C.A.; Moraes, A.H.de.A.; Metha, A.; Carvalho de Oliveira, A.; Labate, C.A.; Marino, C.L.; Monteiro-Vitorello, C.de.B.; Monte, D. de C.; Giglioti, E.; Kimura, E.T.; Romano, E.; Kuramae, E.E.; Lemos, E.G.M.; Pereira de Almeida, E.R.; Jorge, E.C.; Albuquerque, E.V.S.; da Silva, F.R.; Vinecky, F.; Sawazaki, H.E.; Dorry, H.F.A.; Carrer, H.; Abreu, I.N.; Batista, J.A.N.; Teixeira, J.B.; Kitajima, J.P.; Xavier, K.G.; Maria de Lima, L.; Aranha de Camargo, L.E.; Pereira, L.F.P.; Coutinho, L.L.; Lemos, M.V.F.; Romano, M.R.; Machado, M.A.; Costa, M.M. do. C.; Grossi, de Sá M.F.; Goldman, M.H.S.; Ferro, M.I.T.; Tinoco, M.L.P.; Oliveira, M.C.; Van Sluys, M-A.; Shimizu, M.M.; Maluf, M.P.; Souza da Eira, M.T.; Guerreiro Filho, O.; Arruda, P.; Mazzafera, P.; Mariani, P.D.S.C.; de Oliveira, R.L.B.C.; Harakava, R.; Balbao, S.F.; Tsai, S.M.; di Mauro, S.M.Z.; Santos, S.N.; Siqueira, W.J.; Costa, G.G.L.; Formighieri, E.F.; Carazzolle, M.F.; Pereira, G.A.G. 2006. Brazilian coffee genome project: an EST-based genomic resource. *Braz. J. Plant Physiol*. 2006, 18(1):95-108.