Genetic and phenotypic diversity of Coffea canephora and Coffea arabusta collections in French Guiana

15/12/2017

Suntory – CIRAD research collaboration agreement

Through the use of Sequencing-based diversity array technology – DarTseq method – on a collection of 360 C. *canephora* and 50 Arabusta composed of known and unknown accessions, 2719 polymorphic SNPs were identified. We used a multivariate analysis using SNP data from reference accessions in order to confirm and further fine-tune the genetic diversity of C. *canephora*. We identified the genetic origin of the different accessions and classified them in the genetic groups well-known.

Conclusions: The genetic characterization based on SNP markers and the phenotypic characterization of the accessions conserved in the French Guiana collection increased our knowledge on the genetic diversity of C. *canephora*. This information is considered very important for future conservation and development conditions for a breeding program.

Keywords:

Genetic diversity, DArTseq, *Coffea canephora*, phenotypic diversity, Arabusta

Genetic and phenotypic diversity of Coffea canephora and Coffea arabusta collections in French Guiana

SUNTORY – CIRAD RESEARCH COLLABORATION AGREEMENT

MILESTONE	DELIVERABLE

Diversity analysis	Report with (i) materials and methods used, (ii) results (including tables and plots), and (iii) general comments about the results
Phenotypic analysis	Report with (i) materials and methods used, (ii) results (including table), and (iii) general comments about the results
Video-Tutorial content preparation	Slides and content for the Video-tutorial
Video-Tutorial	Presentation, making, and editing of a video- tutorial

BACKGROUND

C. *canephora* is a rubiaceous plant originated from the sub-equatorial plains of Africa. It belongs to the *Coffea* genus, which comprises 124 species, originating from Africa, Madagascar, the Mascarene Islands, Asia and Oceania [1]. C. *canephora* and *Coffea* species are lowland, generally allogamous and diploids (2n = 2x = 22), with the notable exception of the highland, self-fertilizing allotetraploid (2n = 4x = 44) C. *arabica* [2]. Wild C. *canephora* plants are naturally distributed within intertropical Africa, stretching from Guinea to Uganda and from Central African Republic to Angola. Natural populations are composed of few individuals, subjected to gene flows from neighboring populations up to a few kilometers away [3, 4].

Based on former genetic studies [5-7], five main regions of wild genetically distant populations can be recognized: (i) West Africa (Guinea and Ivory Coast); (ii) Central Africa, Cameroon and Congo; (iii) the Atlantic frontage from Gabon to Angola; (iv) the Congo central basin; and (v) Uganda. The genetic diversity of C. *canephora* has been analyzed using isozyme markers, microsatellites and RFLPs [8, 13]. While these former analyses gave consistent results regarding the number and geographic origin of genetic groups, each independent work gave different names ending up with some confusion for the coffee community, suggesting the importance of precisely defining a general nomenclature. In this paper, we have therefore chosen, for clarity's sake, the use of a new unified nomenclature for the five previously referenced genetic groups of C. *canephora,* which will be explained in detail in the plant materials section.

As mentioned before, C. *canephora* genetic diversity has been analyzed using a limited number of isozyme, SSR and RFLP markers, representing only a restricted fraction of the C. *canephora* genome. In contrast to classical molecular markers, SNPs (Single nucleotide polymorphisms) are the most abundant markers, particularly in the non-coding regions of the genome. New sequencing technologies (so called Next Generation Sequencing or NGS) used jointly with different complexity reduction methods, like the ones used in RADseq (Restriction site associated DNA sequencing) [14], GBS (Genotyping by sequencing) [15] and DArTseq (Sequencing-based diversity array technology) [16] methods, enable a large-scale discovery of SNPs in a wide variety of non-model organisms. When such techniques are applied to hundreds of genotypes, they provide measures of genetic divergence and genetic diversity within the major genetic clusters that comprise crop germplasm [16]. Indeed, the recently sequenced and assembled C. *canephora* genome, representing 64% of the 710 Mb genome [17], facilitates the use of such marker technology and further analyses of the obtained data.

For this extended study of the genetic diversity of C. *canephora*, we use SNPs markers. DArTseq [18], a technique based on complexity reduction by the use of restriction enzymes targeting gene-rich regions and NGS sequencing, has been previuosly used to study the genetic diversity of C. *canephora*..

In French Guiana, C. canephora genetic resources were introduced in a collection owned by CIRAD. The origin of this germplasm is diverse and represents the main genetic groups. This information is considered very important for future conservation and development conditions for a breeding program in the country.

The specific objectives of the present study are i) to identify the genetic origin of the coffee plants kept in the CIRAD coffee germplasm, and ii) to discuss possible consequences for coffee quality and breeding.

Individuals of unknown groups were projected onto the discriminant functions found with DAPC, using the "predict" function from the package [19].

To illustrate the genetic relationships between individuals, unrooted NJ trees were constructed with the package "poppr" 2.1.0, based on a Nei's genetic distance matrix [21], modified to measure distances between individuals. Bootstrap analyses were also computed with "poppr", using 100 iterations.

2. Material and Methods

2.1. Plant Material

358 accessions from the French Guiana collection and 52 from the Nicaragua collection were analyzed in this study. Details on the accessions are given in Table 'List of Material Vgetal analyzed in the study'. The accessions named 'Active' correponded to the plants for what the country of origin and the groups is well-known.

2.2. DNA Preparation and Genotyping

Data manipulation and format conversion

In order to identify common markers between the dataset from the 2016 paper [18] and the new data obtained by CIRAD, the DArT sequences supplied by the company were mapped against C. *canephora* pseudo-molecules (http://coffee-genome.org) using the Bowtie2 algorithm [26] with the sensitive, end-to-end alignment option. The respective SAM files were processed to identify the exact

position of each SNP in the genome, based on the position within the read provided by the sequencing company. Only SNPs from mapped reads were further analyzed. The two files were compared to find the intercept using Bedtools. A final filter was carried out in order to remove shift errors on the position of the SNP and duplicated positions within each data set.

In order to transform the data from the nomenclature given by the sequencing company ("0" = Reference allele homozygote, "1" = SNP allele homozygote, "2"= heterozygote, and "-" = double null/null allele homozygote), to a format corresponding to the reference and SNP alleles represented as REF/SNP, a homemade shell script was developed and applied to the data.

2.3. Data Analysis

All the genetic statistical analyses were carried out using R, version 3.4.1. Diversity structure present in the C. canephora collection was analyzed using a Discriminan Analysis ofPrincipal Components (DAPC) multivariate analysis implemented in "adegenet" 2.1.0 as follows :

First, 113 known individuals (33 from the 2016 paper), corresponding to the previously described diversity groups, were used to model the diversity present in the panel, after centering the data. This set of accessions is thereafter referenced as "Analysis Individuals". The most probable number of groups that define the diversity evaluated was inferred using the "find.cluster" function, running successive K-means with an increasing number of clusters (k) from one to ten, and with the Bayesian Information Criterion (BIC) as the statistical measure of goodness of fit. The number of retained Principal Components (PC) to be used in the discriminant analysis was determined using the "xvalDapc" function with the default parameters. Using a threshold calculated with the median hierarchical clustering method implemented in the "snpzip" function from "adegenet", a set of alleles with the highest contribution to the between-population structure was identified.

Additional accessions of unknown groups were projected onto the discriminant functions found with DAPC, using the "predict" function from the "adegenet" package. The individuals were grouped according to the information available on their origin (supplied by the costumer), and are referenced thereafter as "Supplemental Individuals", "Clone Individuals", and "Arabusta Individuals".

To illustrate the genetic relationships between individuals, unrooted NJ trees were constructed with the package "poppr" 2.5.0, based on a Nei's genetic distance matrix, modified to measure distances between individuals. Bootstrap analyses were also computed with "poppr", using 100 iterations. The pairwise Euclidean distance between individuals and the number of loci for which individuals differ were also calculated using, respectively, the "dist" function from "adegenet" and the "dist.gene" function from the "ape" 4.1 package [21]. The Fixation index (FST) [24] between groups was calculated with the "pairwise.fst" function of the "hierfstat" 0.04-22 package [20], with the data from the first analysis individuals.

A video tutorial based on the methodology used for analysis and results obtained (including the Viedo file) was produced. The scripts for the analysis are included (file Diversity_analysis.Rmd) to be used by the students using the tutorial is also provided.

2.4 Phenotypic evaluation

Phenotypic analysis was performed at the Guiana collection, and based on observations collected on 2017. The evaluation was carried out in 257 plants. Each accession contained different number of individual (between 1 and 58).

3. Results and interpretation

Data from 105 individuals from the 2016 paper, and from 409 individuals from the 2017 datasets was analyzed. Five individuals (arabusta-427, CIRAD3, CIRAD7, CIRAD8, and CIRAD35) from the 2017 dataset were not included in the analysis as their data was not reported by the sequencing company.

From the 4,021 SNPs analyzed in the 2016 paper and the new 19,457 SNPs supplied by the sequencing company, only 2,719 were common in the two datasets and passed all the filters applied. The data retained and used for the analysis, along with their corresponding sequences and positions in the C. *canephora* reference genome are given in the "Data_2016-2017.xlsx" file.

Genetic structure of the *C. canephora* collection

Analysis Individuals

In order to interpret C. *canephora* diversity in a whole genome context, the DArTseq SNP data from the "Analysis Individuals" (113 C. *canephora* accessions, 33 from the 2016 paper and 80 from the 2017 dataset) was analyzed using a DAPC multivariate analysis.

The first six principal components of the Principal Component Analysis (PCA), which explained 27.7%, 20.3%, 5.3%, 3.3% 2.7% and 2.2% of the variance, respectively, were retained for the discriminant analysis with the DAPC function. The first four Discriminant Functions (Discriminant axes – DA) were retained afterwards.

Seven genetic clusters were identified after the analysis (Figure DAPC_analysis_individuals.pdf), with 22, 11, 15, 21, 16, 16 and 12 individuals, respectively (Table "Analysis_individuals" from the "Identified_groups_12-10.xlsx" file).

The position of the individuals on the four DAs are presented in the "Analysis_ind_coordinates" table from the "DAPC_coordinates_12-10.xlsx" file and the "DAPC_analysis_individuals.pdf" file.

Membership probabilities for each accession were calculated, and are shown in the "Membership_analysis_ind" table from the "Membership_probability_12-10.xlsx" file, and in the

"Membership_probabilities" PDF files.

In order to identify the genomic regions contributing to the population structure found in C. *canephora*, the identity of the SNPs discriminating the seven groups was determined. Respectively, 1, 7, 49, and 8 structural alleles contributing to the four discriminating functions were identified (File DifferentialSNPs_12-10.xlsx).

To obtain a more complete picture of the genetic relationships linking the C. *canephora* accessions evaluated, an unrooted NJ tree was constructed using the data from the "Analysis individuals". (File "NJ_Analysis_Individuals_12-10_fig.pdf").

The pairwise Euclidean distance calculated between the "Analysis individuals" is included in the "Euclid_dist_Analysis_ind" table from the "Euclidean_distances_12-10", while the number of loci for which individuals differ can be found in the "DifLoci_Analysis" from the "Loci_differences_between_Individuals_12_10.xlsx" file.

Finally, the F_{ST} between the seven groups found was calculated, and is presented in the "Pairwise_Nei_Fst_12-10.xlsx" file.

Supplemental Individuals

With the aim of assessing group membership of other accessions in the collection, and to identify their putative genetic origin and relationships, the DArTseq SNP data obtained from the evaluation of 347 C. *canephora* and three C. *arabica* "Supplemental Individuals" were interpolated into the DAPC analysis. The dataset is composed of 72 individuals from the 2016 paper, and 278 from the 2017 data.

The position of the "Supplemental Individuals" on the four DAs is presented in the "Suppl_ind_coordinates" table from the "DAPC_coordinates_12-10.xlsx" file and the "DAPC_an_supplemental.pdf" file. From the 350 individuals, 120 are identified as part of group 1, one as group 2, 74 as group 3, one as group 4, 108 as group 5, 16 as group 6, and 30 as group 7.

Membership probabilities for each accession were calculated, and are shown in the "Memberships_supplemental_ind" table from the "Membership_probability_12-10.xlsx" file, and in the "Membership_probability_Analisis_Supplemental_Individuals.pdf" file.

The NJ tree constructed for these individuals corresponds to the file "NJ_supplemental_12-10_fig.pdf", while the pairwise Euclidean distance calculated between the "Analysis individuals" and

the "Supplemental Individuals" is included in the "Euclid_dist_An_Suppl_ind" table from the "Euclidean_distances_12-10", and the number of loci for which individuals differ can be found in the "DifLoci_Analysis_Supp" from the "Loci_differences_between_Individuals_12_10.xlsx" file.

Clone Individuals

Assessing group membership was also carried out for 17 other C. *canephora* accessions in the collection, called "Clone Individuals", by interpolating them into the DAPC analysis.

The position of the "Clone Individuals" on the four DAs is presented in the "Clone_ind_coordinates" table from the "DAPC_coordinates_12-10.xlsx" file and the"DAPC_an_clone.pdf" file. From the 17 individuals, three are identified as part of group 3, 12 as group 5, and two as group 7.

Membership probabilities for each accession were calculated, and are shown in the "Memberships_clone_ind" table from the "Membership_probability_12-10.xlsx" file, and in the "Membership_probability_Analisis_Clone_Individuals.pdf" file.

The NJ tree constructed for these individuals corresponds to the file "NJ_an_clon_12- 10_fig.pdf", while the pairwise Euclidean distance calculated between the "Analysis individuals" and the "Clone Individuals" is included in the "Euclid_dist_Analis_clone" table from the "Euclidean_distances_12-10", and the number of loci for which individuals differ can be found in the "DifLoci_Analysis_clone" from the "Loci_differences_between_Individuals_12_10.xlsx" file.

Arabusta Individuals

Finally, group membership was assessed for 34 other C. *canephora* accessions in the collection, called "Arabusta Individuals", by interpolating them into the DAPC analysis.

The position of the "Arabusta Individuals" on the four DAs is presented in the "Arabusta_ind_coordinates" table from the "DAPC_coordinates_12-10.xlsx" file and the "DAPC_an_arabusta.pdf" file. From the 34 individuals, 8 are identified as part of group 1, six as group 5, and 20 as group 7.

Membership probabilities for each accession were calculated, and are shown in the "Memberships_arabusta_ind" table from the "Membership_probability_12-10.xlsx" file, and in the "Membership_probability_Analisis_arabusta_Individuals.pdf" file.

The NJ tree constructed for these individuals corresponds to the file "NJ_an_arabusta_12-10_fig.pdf", while the pairwise Euclidean distance calculated between the "Analysis individuals" and the "Arabusta Individuals" is included in the "Euclid_dist_Analis_arabus" table from the "Euclidean_distances_12-10", and the number of loci for which individuals differ can be found in the "DifLoci_Analysis_arabusta" from the "Loci_differences_between_Individuals_12_10.xlsx" file.

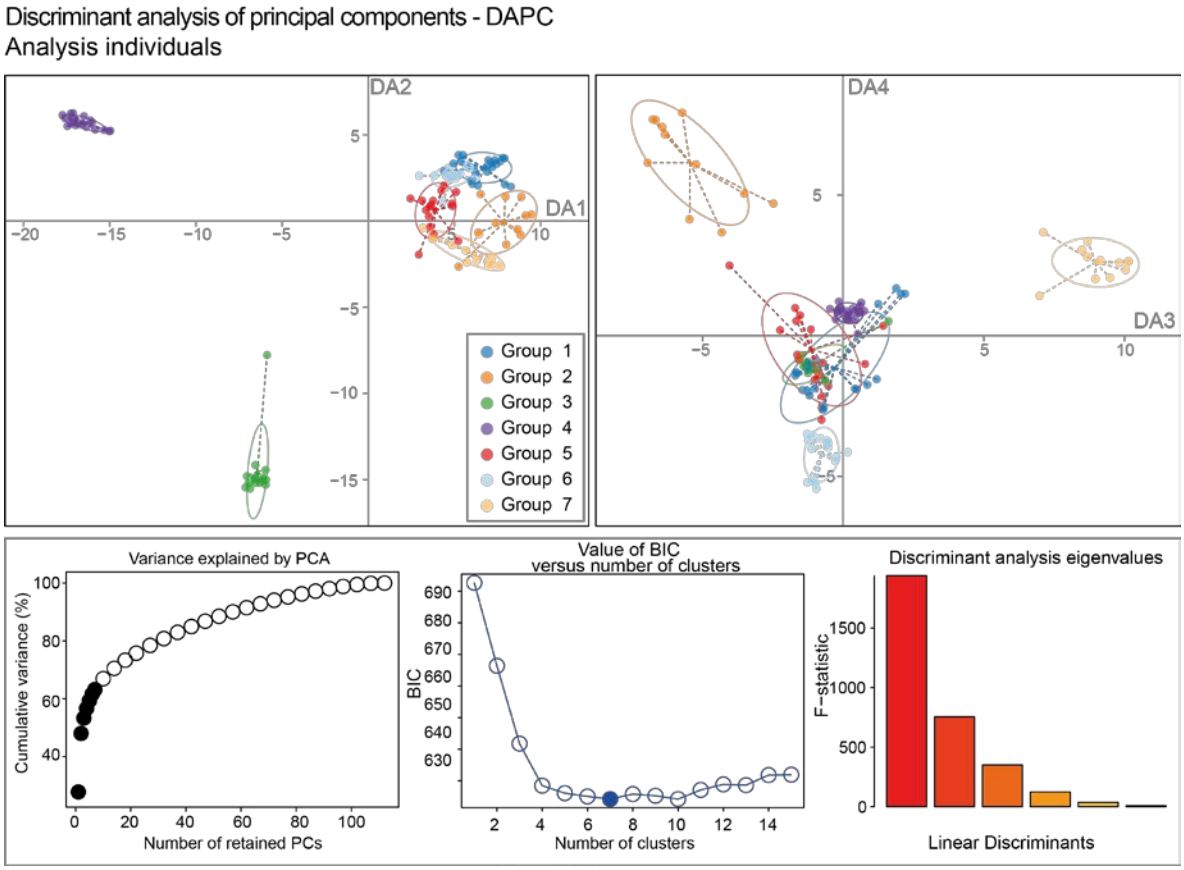


Figure 1 DAPC_analysis.

Scatter plots from the DAPC analysis carried out with 133 *C. canephora* accessions. Discriminant axes 1 and 2 (left) and 3 and 4 (right) representing the seven groups (inertia ellipses) determined by the DAPC. The percentage of cumulative variance for the retained PCA eigenvectors (black dots), the Bayesian Information Criterion (BIC) used to determine the optimal k number of clusters (blue dot), and the F-statistic of the between/within group variance ratio for the discriminant functions (colored bars) are also shown below the DAPC plot.

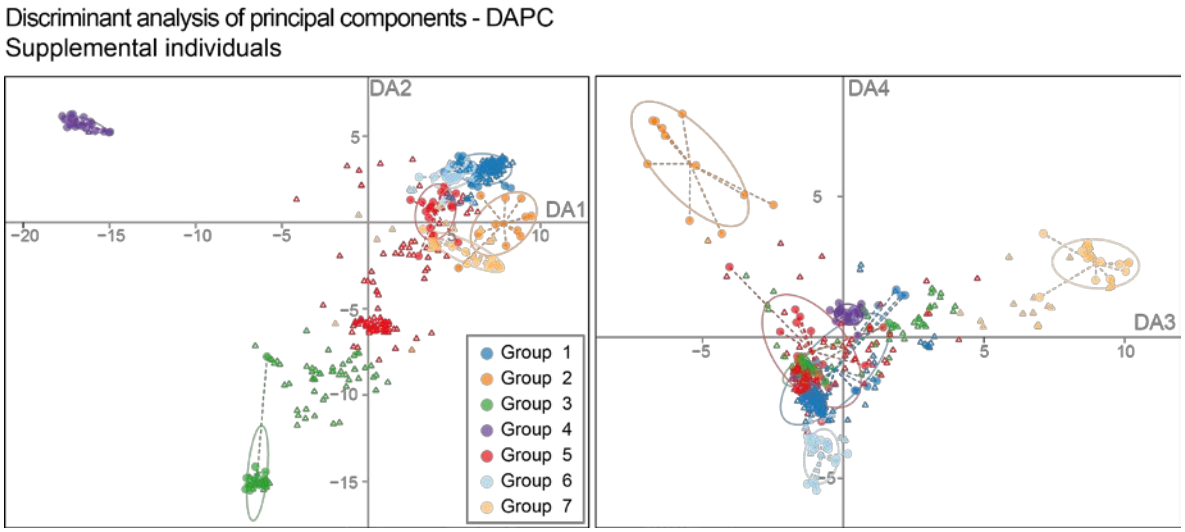


Figure 2 DAPC_an_supplemental

Scatter plots from the DAPC analysis, showing the 347 *C. canephora* and 3 *C. arabica* "Supplemental Individuals" analyzed. Discriminant axes 1 and 2 (left) and 3 and 4 (right) representing the five groups (inertia ellipses) determined by the DAPC. Empty circles represent the "Analysis Individuals" used to identify genetic groups, while empty triangles represent interpolated "Supplemental Individuals".

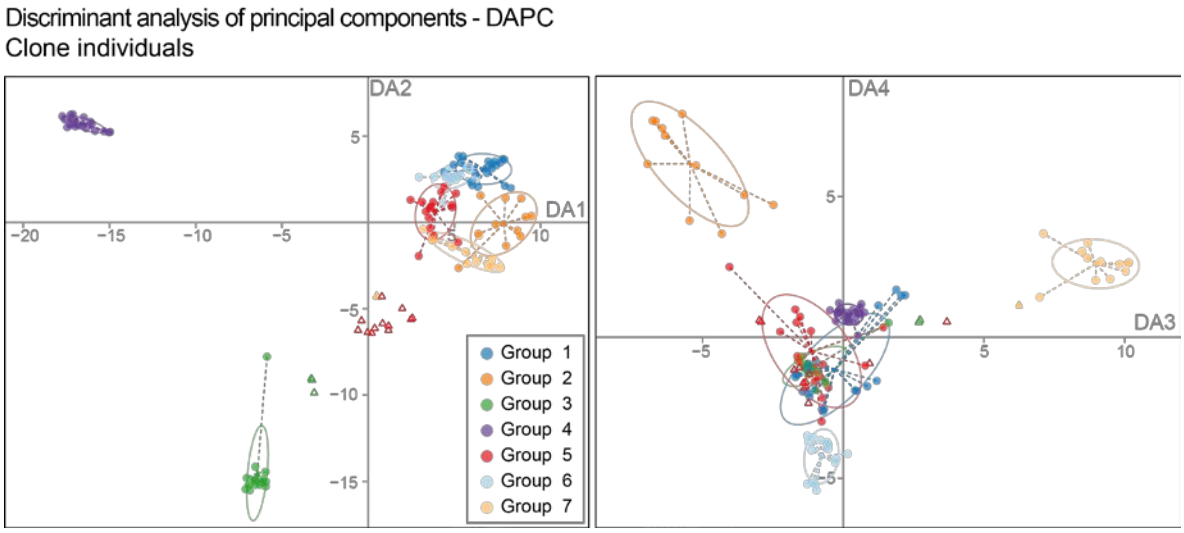


Figure 3 DAPC_an_clone

Scatter plots from the DAPC analysis, showing the 17 *C. canephora* "Clone Individuals" analyzed. Discriminant axes 1 and 2 (left) and 3 and 4 (right) representing the five groups (inertia ellipses) determined by the DAPC. Empty circles represent the "Analysis Individuals" used to identify genetic groups, while empty triangles represent interpolated "Clone Individuals".

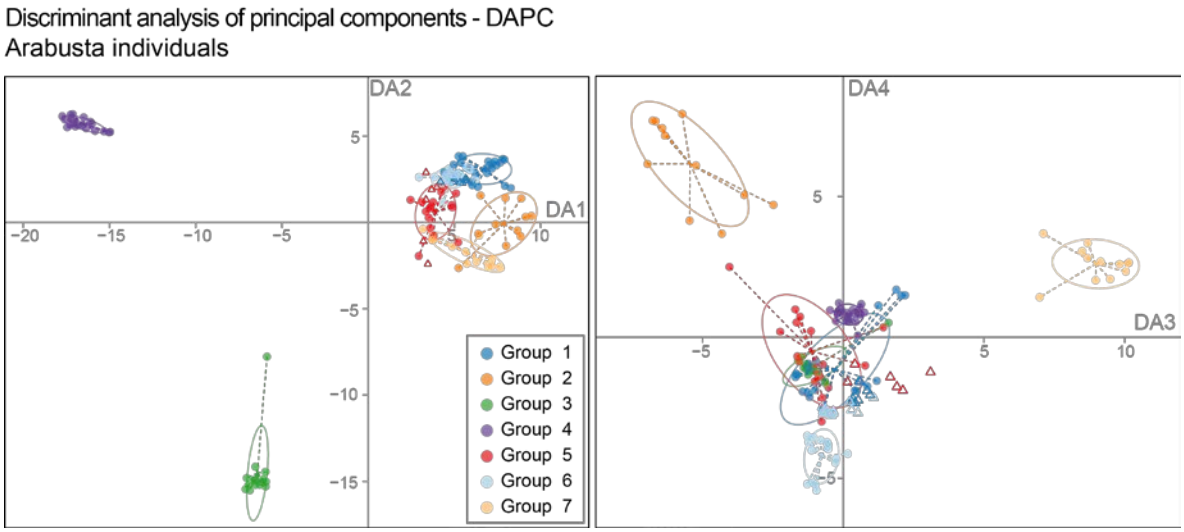


Figure 4 DAPC_an_arabusta

Scatter plots from the DAPC analysis, showing the 17 *C. canephora* "Arabusta Individuals" analyzed. Discriminant axes 1 and 2 (left) and 3 and 4 (right) representing the five groups (inertia ellipses) determined by the DAPC. Empty circles represent the "Analysis Individuals" used to identify genetic groups, while empty triangles represent interpolated "Arabusta Individuals".

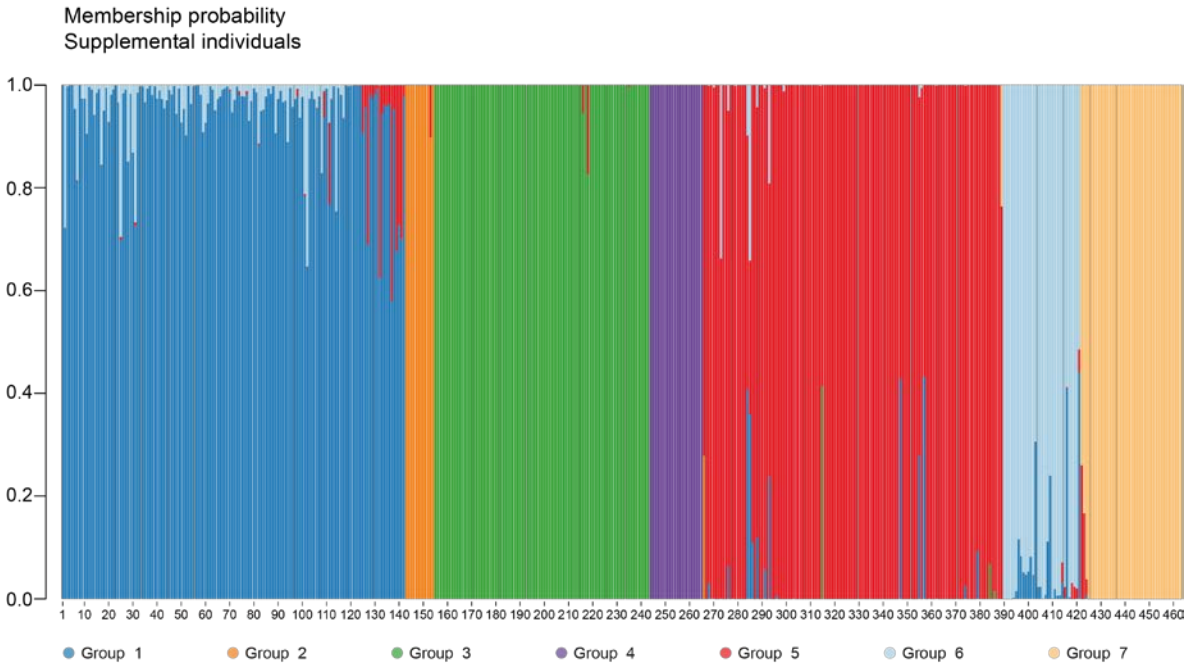


Figure 5 Membership_probability_Analisis_Supplemental_Individuals

Bar plots of the posterior membership probabilities obtained with the DAPC analysis for the "Analysis" and "Supplemental" Individuals. Names of the accessions used to identify the genetic groups are written in black, while the supplemental individuals in gray.

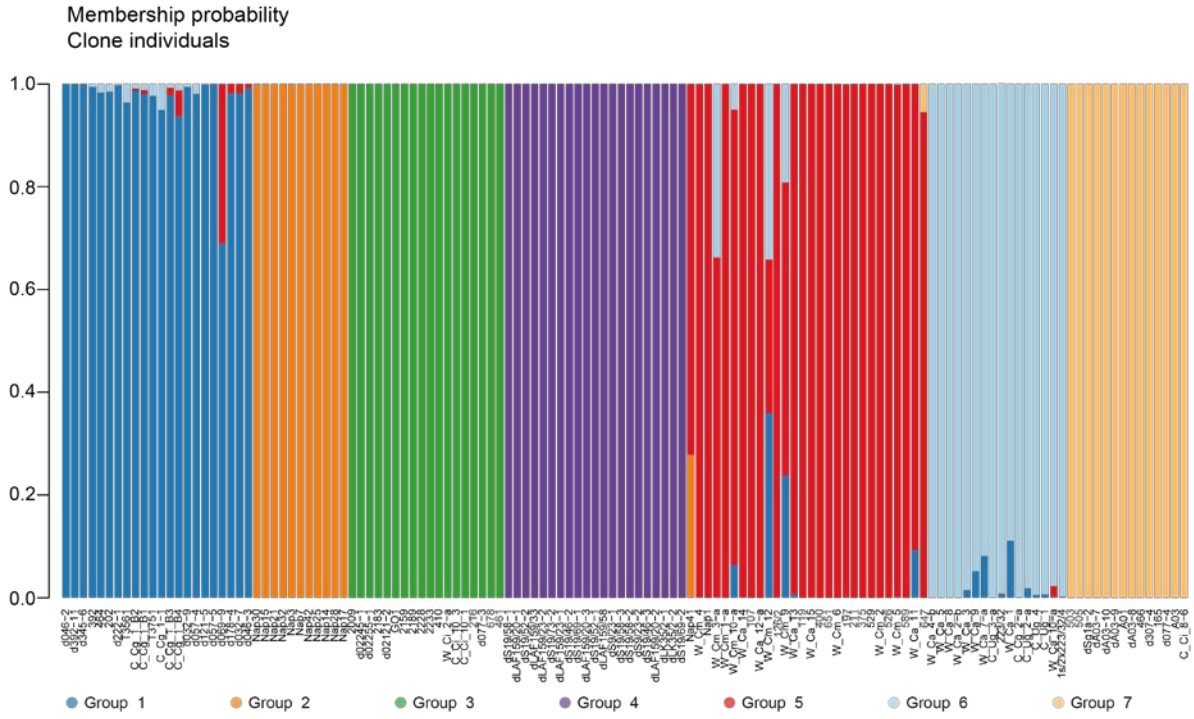


Figure 6 Membership_probability_Analisis_Clone_Individuals

Bar plots of the posterior membership probabilities obtained with the DAPC analysis for the "Analysis" and "Clone" Individuals. Names of the accessions used to identify the genetic groups are written in black, while the clone individuals in gray.

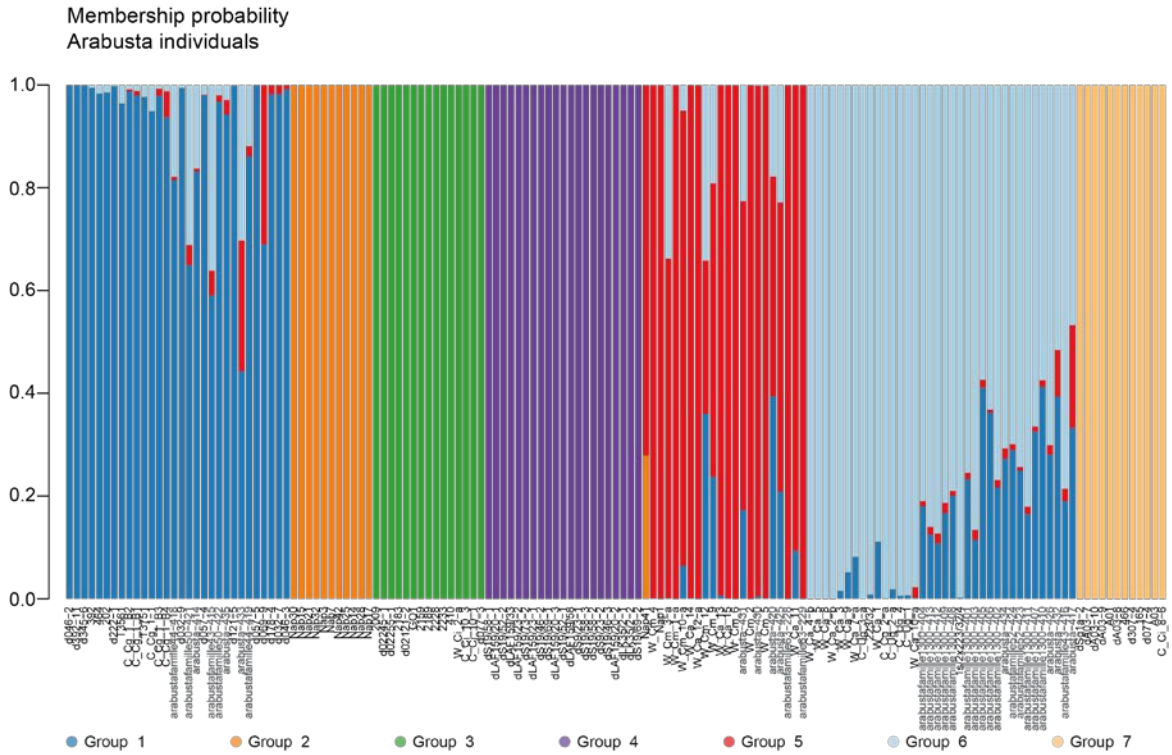


Figure 7 Membership_probability_Analisis_Arabusta_Individuals

Bar plots of the posterior membership probabilities obtained with the DAPC analysis for the "Analysis" and "Arabusta" Individuals. Names of the accessions used to identify the genetic groups are written in black, while the clone individuals in gray.

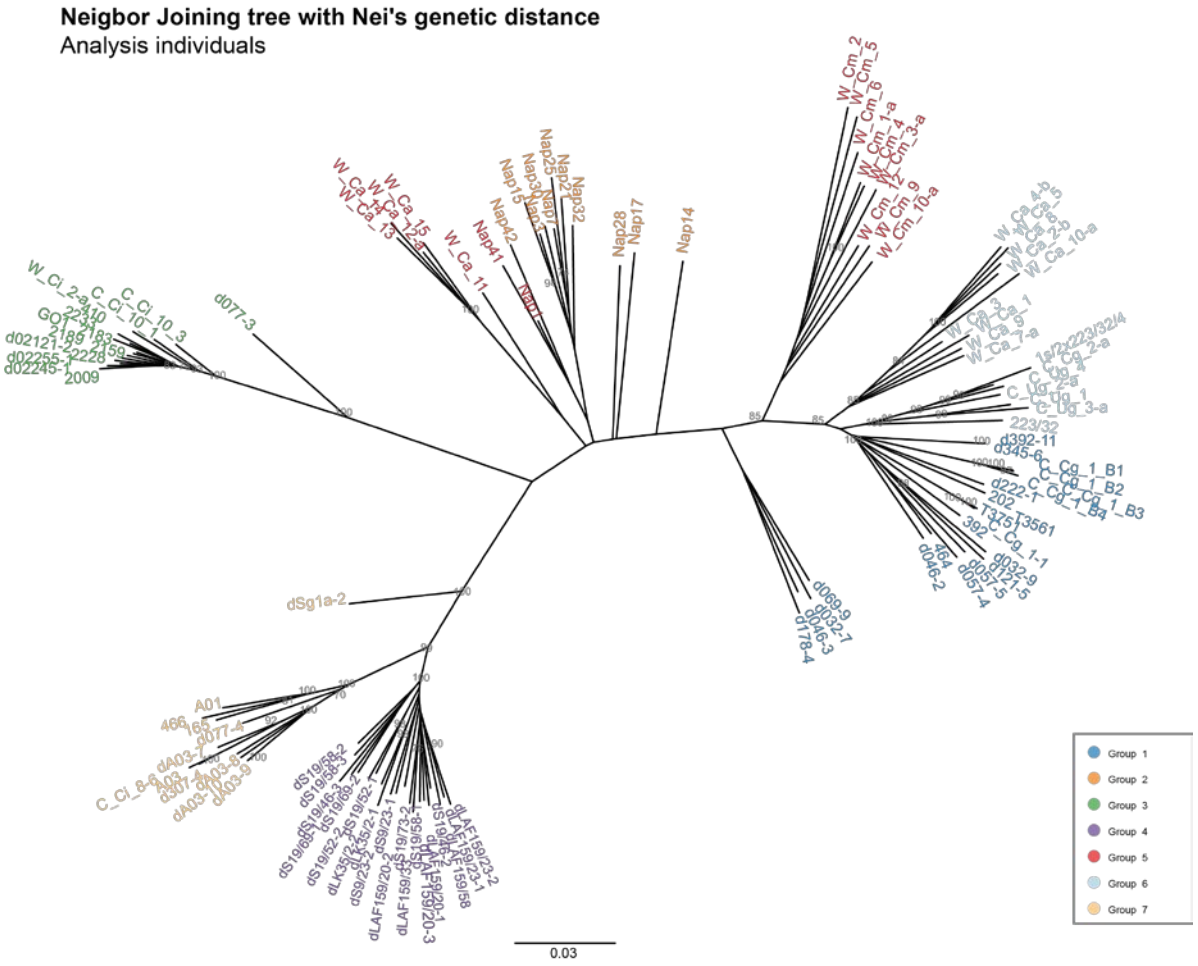


Figure 8 NJ_Analysis_Individuals

Unrooted tree using the Neighbor-joining algorithm based on Nei's genetic distances between 113 "Analysis Individuals". Only bootstrap values over 70 are shown.

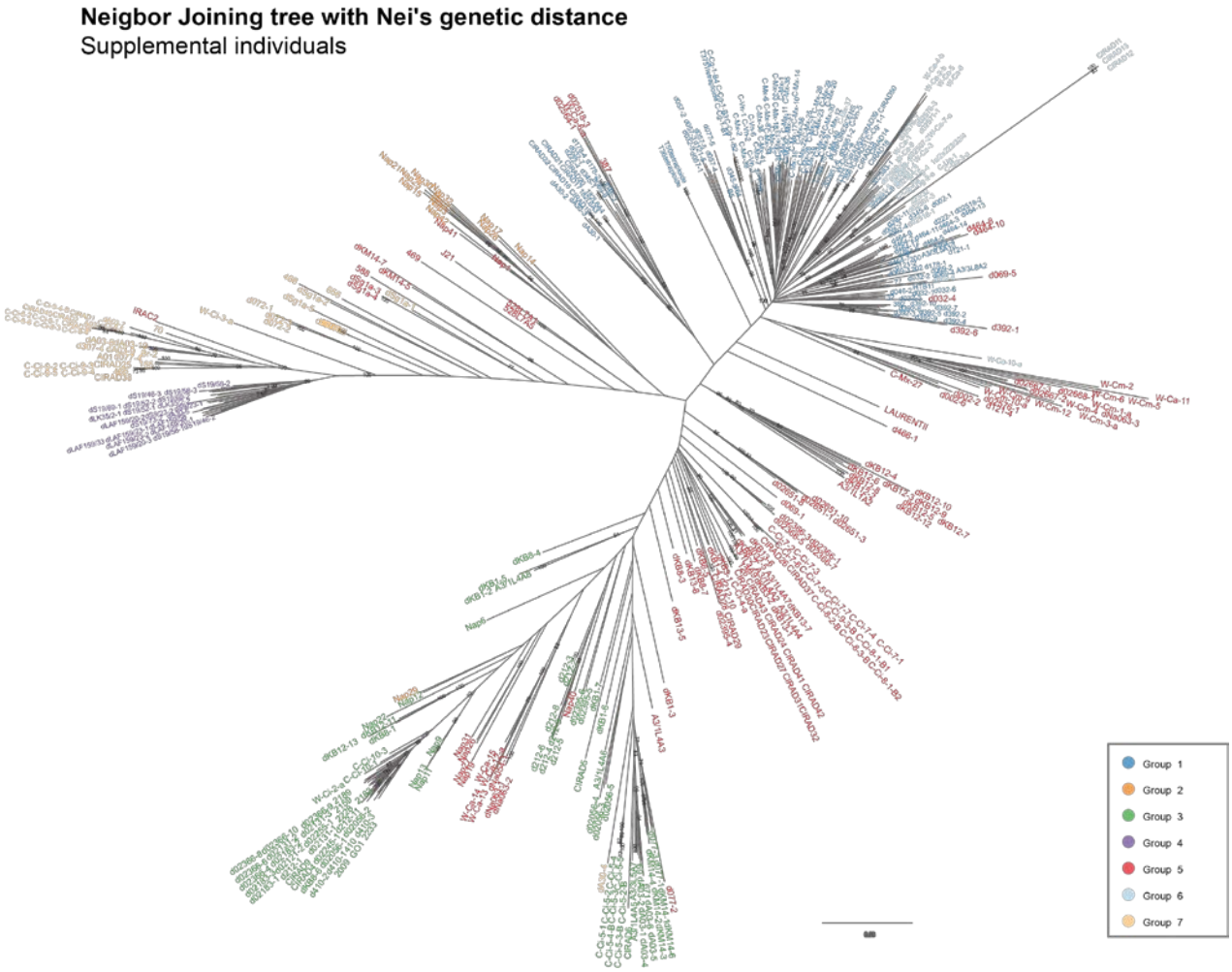


Figure 9 NJ_Analysis_supplemental

Unrooted tree using the Neighbor-joining algorithm based on Nei's genetic distances between 113 "Analysis Individuals" and 350 "Supplemental Individuals". Only bootstrap values over 70 are shown.

Figure 10 NJ_Analysis_clone

Unrooted tree using the Neighbor-joining algorithm based on Nei's genetic distances between 113 "Analysis Individuals" and 17 "Clone Individuals". Only bootstrap values over 70 are shown.

Figure 11 NJ_Analysis_arabusta

Unrooted tree using the Neighbor-joining algorithm based on Nei's genetic distances between 113 "Analysis Individuals" and 34 "Arabusta Individuals". Only bootstrap values over 70 are shown.

Phenotypic characterization

We have characterized the different genotypes by phenotypic measurements in 2017 (Phenotypic data.xls). We observed that the SG2 group (Congo-Uganda) and the Congolense B group presented largest and longer leaves, were more susceptible to drought and presented a harvest grouped.

In contrast the SG1 'Conilon' and the SG1 'Luki' have smaller leaves with a late harvest. The Nana group and the Guineans seem more tolerant to drought than the other groups.

The hybrid groups presented intermediate characterisitics.

	CAPTION
A	Putative genetic group
B	mois d'apparition des premiers fruits mûrs entre mai et octobre (2017) (notations continuent)
C	% of ripened fruits (august 2017)
D	average of ripened fruits (oct 2017)
E	width width (cm; 5 leaves/genotype)
F	lenght width (cm; 5 leaves/genotype)
G	width/lenght ratio
H	length of plagiotropic internodes (cm)
I	Note 1 (few) to 4 (strong) of fruit set
L	Note 1 to 7 : reaction drought (with 7 = highly susceptible)
M	Flowering after moderate drought (Y or N)

Average by group

GROUPE	DENOMINATION CIRAD	NOMBRE	B	C	D	E	F	G	H	I	L	M
1	SG2 'Congo-Uganda'	68	Harvest grouped July-augus	0,4	0,8	8,9	20,6	2,3	5,8	2	5,7	0,3
2	Nana C group	12	Harvest spread	0,4	0,5	7	17,7	2,5	5,6	1,5	3,8	0,2
3	Guineans	73	Harvest spread	0,1	0,3	7	17,9	2,6	5,3	2,8	4,1	0,3
4	SG1 'Luki'	22	Late	epsilon	epsilo	7	16,6	2,4	5,6	1,8	6	0
5	Hybrids	84	Harvest grouped June-July-	0,25	0,4	7,8	19	2,5	5,4	2,6	4,4	0,4
6	Congolense B	14	Harvest grouped July-augus	0,5	0,95	9,6	22,3	2,3	5,9	2,2	5,3	0,4
7	SG1 'Conilon'	21	Harvest spread	0,05	0,15	7,6	19,1	2,5	5,5	2,2	4,5	0,1
	ENSEMBLE	294		0,25	0,45	7,8	19	2,4	5,5	2,4	4,8	0,3
	min			0	0	4,9	12,7	1,8	3,1	1	1	
	max			1	1	12,2	26,1	3,3	8,5	4	7	
	écart-type			0,3	0,37	1,2	2,5	0,2	1	0,9	2,1	

Difference seems significative

Interpretation and Discussion

The different groups with different codes coulors represented the different origins of the material. Our study confirmed the genetic diversity of accessions from CIRAD collection , covering the six main groups of Canephora known for the moment : SG1, Nana, Guinean, SG1 'Luki', B, SG1 'Conilon'.

We noted (Fig. 8, 9, 10 , 11) that the groups SG2 (Blue) and B (Blue sky) (respectively Robusta Congo-Uganda and Central Africa) are closed and opposed to the SG1 groups (Luki and Conilon, respectively Purple and Yellow, Gabon).

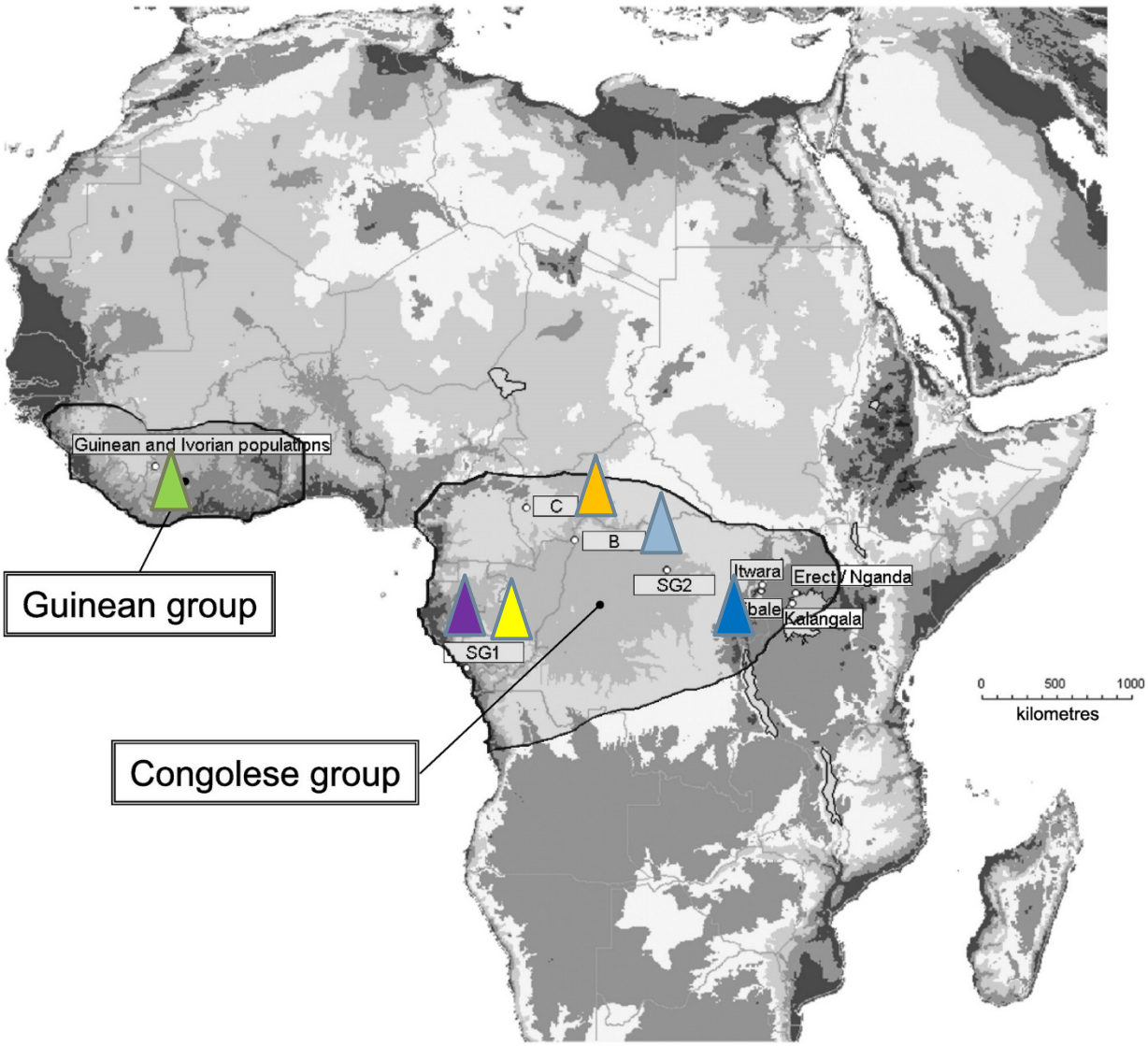
We also found that the Guinean group (green) is an isolated group. We know that the hybrids between Guinean x SG2 confers a high level of heterosis.

Finally we found a high number of hyrbids between groups. The majority of thos hybrids seem coming from SG2 x Guinean.

The Arabusta group (Fig 11) ranks close to the SG2 group or to the B group. This result was expected because the tetraploid Robusta clones used to cross the Arabica were coming from the SG2 group.

Group		Color
1	**SG2** = "Robusta Congo – Uganda" group (sometimes called SG2) corresponding to the wild populations or cultivated varieties native to Uganda and the Congo basin.	Blue
2	"**Nana**" group (sometimes called **C group**), stands for the coffee originating from the fringes of South-East Cameroon, South-West Central Africa and Northern Congo	Orange
3	"**Guinean**" Group (sometimes called D group), it is the genetic group originating from the Ivory Coast – Guinea area in West Africa	Green
4	**SG1** origin **'Luki'** eastern province of the republic of Congo	Purple
5	Hybrids intergroups	Red
6	"Robusta Congo – Central Africa" group (sometimes called **group B**)	Blue sky
7	"**Conilon"** group (sometimes called **SG1** or A) represented by Niaouli and Kouilou domesticated populations, originating from the south of Gabon	Yellow
8	Arabusta	Red (only in Fig. 11)

Our study confirmed the genetic diversity of accessions from CIRAD collection , covering the six main groups of Canephora known for the moment.



Perspectives :

Crosses between groups

Guineans	SG1 (Luki and Conilon)	5 clons x 5 clons
Guineans	SG2 (SG2 and B group)	5 clons x 5 clons
Guineans	Nana (Group C)	5 clons x 5 clons
SG1 (Luki and Conilon)	SG2 (SG2 and B group)	5 clons x 5 clons
SG1 (Luki and Conilon)	Nana (Group C)	5 clons x 5 clons

References

1. Davis AP, Tosh J, Ruch N, Fay MF: **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* 2011, **167**(4):357--377.
2. Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Perez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, Poncet V, Nagai C *et al*: **Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea)**. *Plant J* 2011, **67**(2):305-317.
3. Berthaud J: **Gene flow and population structure in *Coffea canephora* coffee populations in Africa**. In: *Genetic differentiation and dispersal in plants.* Edited by Jacquart P, Heim G, Antonovics J. Berling: Springer Verlag; 1985.
4. Montagnon C, Leroy T, Yapo A: **Caractérisation et évaluation de caféiers *Coffea canephora* prospectés dans des plantations de Côte-d'Ivoire.** *Café, Cacao, Thé* 1993, **37**(2):115-119.
5. Cubry P, de Bellis F, Avia K, Bouchet S, Pot D, Dufour M, Legnate H, Leroy T: **An initial assessment of linkage disequilibrium (LD) in coffee trees: LD patterns in groups of Coffea canephora Pierre using microsatellite analysis**. *BMC Genomics* 2013, **14**:10.
6. Montagnon C, Leroy T, Yapo A: **Diversité génotypique et phénotypique de quelques groupes de caféiers (*Coffea canephora* Pierre) en collection. Conséquences sur leur utilisation en sélection**. *Café, Cacao, Thé* 1992, **36**(3):187-198.
7. Dussert S, Lashermes P, Anthony F, Montagnon C, Trouslot P, Combes MC, Berthaud J, Noirot M, Hamon S: **Le caféier, *Coffea canephora***. In: *Diversité génétique des plantes tropicales cultivées.* Edited by P H, M S, X P, C. GJ. Montpellier: CIRAD; 1999: 175–194.

8. Montagnon C, Guyot B, Cilas C, Leroy T: **Genetic parameters of several biochemical compounds from green coffee, Coffea canephora**. *Plant Breed* 1998, **117**(6):576--578.
9. **Berthaud J**: **Les ressources génétiques pour l'amélioration des caféiers africains diploïdes : évaluation de la richesse génétique des populations sylvestres et de ses mécanismes organisateurs. Conséquences pour l'application**. Paris: ORSTOM; 1986.
10. Cubry P, De Bellis F, Pot D, Musoli P, Leroy T: **Global analysis of Coffea canephora Pierre ex Froehner (Rubiaceae) from the Guineo-Congolese region reveals impacts from climatic refuges and migration effects**. *Genet Resour Crop Evol* 2013, **60**(2):483-501.
11. Cubry P, Musoli P, Legnaté H, Pot D, de Bellis F, Poncet V, Anthony F, Dufour M, Leroy T: **Diversity in coffee assessed with SSR markers: structure of the genus Coffea and perspectives for breeding**. *Genome* 2008, **51**(1):50-63.
12. Musoli P, Cubry P, Aluka P, Billot C, Dufour M, De Bellis F, Pot D, Bieysse D, Charrier A, Leroy T: **Genetic differentiation of wild and cultivated populations: diversity of Coffea canephora Pierre in Uganda**. *Genome* 2009, **52**(7):634-646.
13. Gomez C, Dussert S, Hamon P, Hamon S, de Kochko A, Poncet V: **Current genetic differentiation of Coffea canephora Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities**. *BMC Evolutionary Biology C7 - 167* 2009, **9**(1):1-19.
14. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: **Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers**. *PLoS ONE* 2008, **3**(10):e3376.
15. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species**. *PLoS ONE* 2011, **6**(5):e19379.
16. Cruz VMV, Kilian A, Dierig DA: **Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop Lesquerella and Related Species**. *PLoS ONE* 2013, **8**(5):e64062.
17. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G *et al*: **The coffee genome provides insight into the convergent evolution of caffeine biosynthesis**. *Science* 2014, **345**(6201):1181-1184.
18. Garavito, A., Montagnon C., Guyot R, Bertrand B. (2016). **"Identification by the DArTseq method of the genetic origin of the Coffea canephora cultivated in Vietnam and Mexico."** BMC Plant Biology 16(1): 242
19. Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers**. *Bioinformatics* 2008, **24**(11):1403-1405.
20. Goudet J: **hierfstat, a package for r to compute and test hierarchical F-statistics**. *Mol Ecol Notes* 2005, **5**(1):184-186.
21. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS: **A high-performance computing toolset for relatedness and principal component analysis of SNP data**. *Bioinformatics* 2012, **28**(24):3326-3328.
22. Jombart T, Devillard S, Balloux F: **Discriminant analysis of principal components: a new method for the analysis of genetically structured populations**. *BMC Genetics C7 - 94* 2010, **11**(1):1-15.
24. Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G: **LOSITAN: A workbench to detect molecular adaptation based on a F st -outlier method**. *BMC Bioinformatics* 2008, **9**(1):1-5.
25. Kamvar ZN, Tabima JF, Grünwald NJ: **An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction**. *PeerJ* 2014, **2**(3):e281.

26. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**(4):357- 359.

Authors:

Benoît Bertrand, Bernard Perthuis, Pierre Charmetant, Thierry Leroy, Sunao Nakamura

Corrresponding author

Benoît Bertrand, CIRAD

911 Av Agropolis BP 64501 F34394 Montpellier Cedex 5 France

Tel 33 0467416273 – 0770328455

Reference: Genetic diversity analysis of *Coffea canephora*